POSE-GUIDED PROPORTIONAL STRUCTURE ASSESSMENT FOR 2D ILLUSTRATED CHARACTERS

Yu-Jyuan Lin[1], Chian-Ru Lin[2], Si-Xian Hsu[2]
[1]Department of Information Management, National Taiwan University
[2]Department of Electrical Engineering, National Taiwan University
{b13705012, b13901007, b13901009}@ntu.edu.tw

ABSTRACT

Structural errors in human anatomy, particularly proportion and composition, are common challenges for beginners in 2D character illustration. Current research primarily focuses on content generation rather than structural correction and diagnosis. This study proposes a novel evaluation framework for detecting abnormal proportions in illustrated figures. Our method combines a Spatial Transformer Network for geometric alignment with a Siamese CNN for proportion-aware similarity learning, and uses Grad-CAM to visualize matching results as guidance for beginner illustrators. This framework provides an interpretable and objective approach for assessing proportional errors in 2D illustrations, offering practical guidance for improving structural accuracy in manga character drawing.

1 INTRODUCTION

For beginners in anime illustration, mastering human anatomy and skeletal structure poses a significant challenge. Novices often fall into the trap of local depiction, focusing excessively on details while neglecting the overall spatial arrangement. This frequently results in unbalanced skeletal structures or collapsed perspectives in the final artwork. Unlike realistic drawing, which adheres to fixed anatomical standards, 2D anime creation allows for a certain degree of stylistic variation. This flexibility makes the criteria for incorrect anatomy ambiguous and difficult to define.

In the field of computer vision, current research on 2D anime characters focuses primarily on colorization, line art extraction, or generation, with less exploration into drawing tutorials or skeletal correction. While mature technologies for pose extraction and rehabilitation action assessment exist for natural images, relevant literature in the anime domain remains scarce. To address this, we simulate proportional abnormalities and diverse compositions using scaling, translation, and warping techniques to model common learner errors.

To address this gap, we propose a proportionality-aware skeleton analysis framework based on an Siamese convolutional neural network with the spatial transformer network (STN) [2](Jaderberg et al., 2015) as an alignment module coupled with gradient-weighted Class Activation Mapping (Grad-CAM) [4](Selvaraju et al., 2017)visualization. During inference, a user-provided skeleton with potential proportional errors is automatically matched against a database of correctly proportioned reference skeletons. The input skeleton and its most similar reference are then passed through the two branches of the Siamese CNN to compute their cosine similarity, after which Grad-CAM is applied to identify the regions most responsible for the similarity score. Consequently, the resulting heatmap highlights the proportionally correct regions of the input skeleton while suppressing activation in distorted areas, offering intuitive and interpretable feedback for learners.

2 RELATED WORK

2.1 POSE-GUIDED MATCHING

We recognize that our task—detecting abnormal proportions in illustrated figures—shares the same high-level idea as pose assessment in sports and rehabilitation, a well-developed research field. The

following section outlines related work that is applicable to our scenario with certain adjustments and modifications.

One relevant paper focuses on aiding patients in rehabilitation exercises [3](Qiu et al., 2022). The rehabilitation exercise of interest is the eight-section brocade, a traditional Chinese practice composed of only eight key postures. Their training data consist of frames captured from videos of patients and a video of a specialist performing the eight-section brocade. Only frames around the key posture points are extracted, so the dataset can be categorized into eight classes. Images of the specialist are treated as "standard," while images of patients are treated as "learner."

The paper introduces a new idea for pose matching, namely pose-guided matching, which aims to provide objective and accurate scores, feedback, and guidance to patients when their poses are compared with the standard poses. More specifically, the authors propose a pair-based Siamese Convolutional Neural Network (SCNN), abbreviated ST-AMCNN, to realize pose-guided matching. They simplify multi-stage pose matching by merging the alignment and matching modules into a one-stage task, such that only one loss function is required, reducing computational complexity. Building upon the Spatial Transformer Networks (STN) used as the alignment module, they propose a new Attention-based Multi-Scale Convolution (AMC) to match different posture parts (i.e., multi-scale). Finally, Gradient-weighted Class Activation Mapping (Grad-CAM) is adopted to visualize the matching results for the learner.

This paper provides a workflow that we consider plausible to transplant to our scenario. However, several challenges arise when adapting this standard operating procedure from a rehabilitation assistant for patients to an illustration assistant for art beginners. First, it is impossible to obtain many poorly illustrated figures paired with corrected versions drawn by experienced artists, so we must create a method for generating (standard, learner) counterparts in our scenario. Second, we do not aim to fit all standing-pose skeletons to a single standard pose, as this would violate the artist's intention. Instead, we aim to preserve most aspects of the original pose (e.g., facing direction, stance width) while correcting only proportion errors, making our task significantly more challenging than in the eight-section brocade case.

2.2 Pose Estimation

We also examine a paper on illustrated character skeleton generation and dataset construction [1](Chen & Zwicker, 2022). This work expands existing datasets and proposes a transfer-learning model to address the scarcity of pose-estimation data for illustrations. In addition to achieving high-accuracy skeleton extraction for illustrated figures, the authors further apply their state-of-the-art character pose estimator to the novel task of pose-guided illustration retrieval. They construct a retrieval system that searches for illustrated characters in similar poses by performing a simple nearest-neighbor search of euclidean distances over normalized keypoints. The original purpose of this system is to serve as a practical reference tool for artists, who commonly rely on reference drawings during the illustration process.

This paper provides us with a reliable skeleton generator and a large dataset equipped with a retrieval mechanism, thereby enabling more effective construction of a dataset tailored to our task.

3 Methodology

3.1 Overview

We construct a dataset of manga character skeletons for training and evaluation. Because it is not feasible to collect a sufficient number of poorly illustrated figures paired with corrected versions drawn by experienced artists, we generate our own ground truth. In this setup, the skeleton of a well-illustrated figure serves as the standard, and warped variants of this skeleton serve as the learners. Additional details regarding dataset generation are provided in the experiment section.

Our model is trained on paired images using a Siamese CNN as the primary architecture. Each image pair is processed in the way that only the learner skeleton passes through a Spatial Transformer Network (STN) module while the standard skeleton directly to the Siamese CNN. The Siamese

CNN outputs a pair of corresponding embeddings, which are used to compute a single loss value. This loss is then backpropagated to jointly update both the CNN parameters and the STN parameters.

For the overall model design, we adopt the ST-AMCNN structure described in [3](Qiu et al., 2022). This framework employs STN for alignment, a Siamese CNN for matching, and Grad-CAM for generating visualized correction guidance. We retain most of the original structure and the hyperparameters due to the conceptual similarity between our task and the pose-assessment task. Nonetheless, several components are modified to better accommodate our application scenario and dataset characteristics

3.2 Spatial Transformer Network (STN)

Before the matching stage, the two skeletons being compared must be aligned to the same position, size, and orientation. This ensures that the Siamese CNN focuses on differences in posture rather than irrelevant variations in scale, location, or rotation. The structure of the STN is shown in Fig. 1.

The STN consists of three main components: a localization network, a grid generator, and a grid sampler. The localization network is a CNN that produces a parameter vector θ Figure. 2. The grid generator then uses θ to construct a transformation matrix A_θ(Eq. 1a 1b). This matrix maps the target coordinates back to the learner's coordinate space; given this mapping, the reverse transformation can also be obtained. The training objective of the STN is therefore to warp the input skeleton into a centered, upright position.

3.3 Siamese CNN

We implement a Siamese CNN architecture composed of two Attention-based Multi-Scale Convolutional Networks (AMCNN). AMCNN contains a bundle of three convolution kernels with different sizes, designed to capture features at multiple spatial scales. The extracted features are then passed into an attention module, which resembles the Squeeze-and-Excitation (SE) block proposed in SENet. This attention mechanism encourages the network to focus on posture-relevant features while suppressing irrelevant ones.

3.4 Grad-CAM

To inform the learner which specific parts of the skeleton require correction, we apply Grad-CAM to generate a heatmap. Grad-CAM takes the output of the last convolutional layer and computes the cosine similarity between the two outputs. This visualizes the comparison result of the two input poses from the CNN's perspective. In the heatmap, red regions indicate that the corresponding parts of the learner skeleton are more similar to the standard skeleton.

3.5 Loss Function

Here, we used two loss function method for identify the model. We implement Cosine Embedding Loss and evaluate its performance. It is important to note that the loss function penalizes not only mismatches between the two skeletons, but also misalignment of the input skeleton produced by the STN.

To compute similarity between the two embeddings, we use a cosine embedding loss. Because cosine similarity measures directional alignment, it is more sensitive to posture orientation than Euclidean distance. The embedding loss is defined as:

$$L(x_1, x_2, y) = \begin{cases} 1 - \cos(x_1, x_2), & y = 1, \\ \max(0, \cos(x_1, x_2) - m), & y = -1, \end{cases}$$

where x_1 and x_2 are the embeddings corresponding to the standard and learner skeletons, $y = 1$ indicates that the pair belongs to the same class, $y = -1$ indicates different classes, and m is the margin.

To avoid penalty, the model must learn to produce highly similar embeddings for same-class pairs and sufficiently different embeddings for different-class pairs.

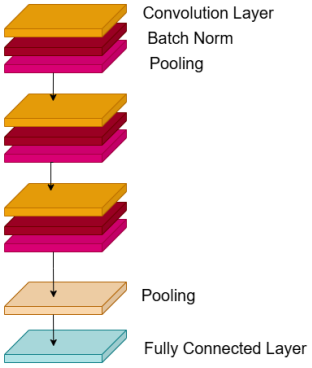

Figure 1: Localization Network of STN Module

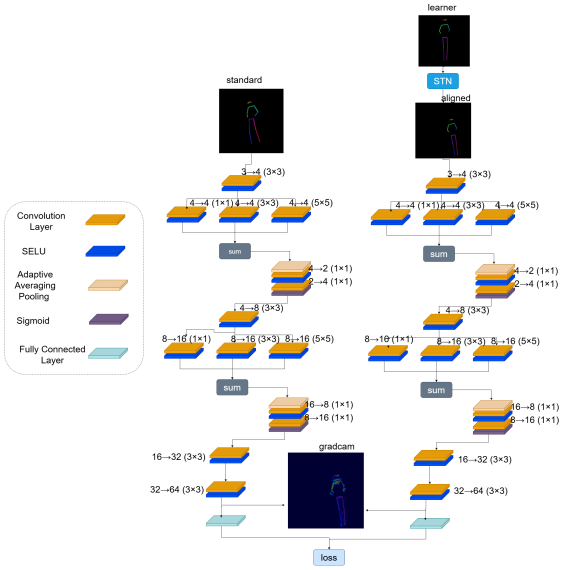

Figure 2: Architecture of our Training Model

$$\begin{bmatrix} x_i^s \\ y_i^s \\ 1 \end{bmatrix} = A_\theta \begin{bmatrix} x_i^t \\ y_i^t \\ 1 \end{bmatrix} = \begin{bmatrix} s_x \cos\theta & -s_x \sin\theta & t_x \\ s_y \sin\theta & s_y \cos\theta & t_y \end{bmatrix} \begin{bmatrix} x_i^t \\ y_i^t \\ 1 \end{bmatrix} \tag{1a}$$

$$s_x = 1 + \alpha s_x^{raw}, \qquad s_y = 1 + \alpha s_y^{raw} \tag{1b}$$

Equation 1a, 1b: Affine transformation of STN. s_x, s_y are the scaling parameters, α is the scaling factor, t_x, t_y are translation parameters, θ is the rotation parameter. A_θ is the affine matrix, (x_i^s, y_i^s) is the source coordinate, (x_i^t, y_i^t) is the target coordinate.

3.6 Evaluation

To assess whether the model correctly identifies proportional errors, we construct test skeletons in which the disproportion is applied exclusively to one of four regions: the left half, right half, upper half, or lower half. For each test sample, we compute the cosine similarity score between the distorted skeleton and its corresponding standard skeleton, and then apply Grad-CAM to visualize which regions contribute most strongly to the score. Since the distorted region should not resemble the standard skeleton, an ideal model should produce weak or no activation on the distorted half, while highlighting structural correspondences on the undistorted half.

A prediction is therefore considered correct when the Grad-CAM activation on the undistorted region exceeds the activation on the distorted region, either in intensity or spatial extent. The final accuracy metric is computed as the proportion of test samples that satisfy this criterion.

Based on the proposed STN–Siamese CNN–Grad-CAM architecture, our experiments aim to evaluate whether the model can reliably identify skeletons with incorrect proportions, and whether spatial alignment via the STN further improves this capability. This section presents the construction of the dataset, the implementation details of our training setup, the comparative performance of models with and without the STN module, and the effects of varying key hyperparameters.

4 EXPERIMENTS AND RESULTS

4.1 DATASET

We construct a dataset of manga-style skeletons specifically designed to ensure that the STN learns meaningful geometric transformations while the subsequent Siamese CNN learns to discriminate proportional correctness. Because the original dataset contains only a limited number of clean, well-proportioned skeletons, we treat each of these as a standard reference and generate two complementary categories of training samples. The first category consists of eight proportion-preserving transformations that encourage the STN to learn alignment-related transformations without altering the underlying anatomy. The second category consists of nine proportion-changing distortions, enabling the CNN to learn explicit differences between correct and incorrect proportions. Together, these two categories constitute approximately equal portions of the dataset, allowing the model to simultaneously acquire robust transformation invariance and proportion-awareness.

The training data are built from three canonical poses: standing with open arms, standing with hands on hip, and kneeling(Fig. 3). We follow the pipeline described in Transfer Learning for Pose Estimation of Illustrated Characters [1](Chen & Zwicker, 2022). For each pose, we first use the provided pose-retrieval model to obtain candidate images whose pose similarity meets our criterion, and then apply the released pose-estimation model to extract skeletons. To prevent confusion during training, we manually filter out skeletons that deviate substantially from the intended pose cluster, along with samples in which the character is facing away from the viewer. After filtering, we obtain 156 skeletons for the standing-with-open-arms pose, 141 for standing-with-hands-on-hip, and 120 for kneeling.

From each original skeleton, we generate a total of 17 transformed variants, resulting in 18 samples per original skeleton. Eight of these transformations preserve overall body proportions and are designed to train the STN to compensate for global misalignments such as translation, rotation, and scale variation. We apply two independent scaling augmentations to every original skeleton: a random down-scaling to 80–100% of its original size and a random up-scaling to 100–120%. In addition, each skeleton undergoes two independent rotation augmentations: a clockwise rotation randomly sampled between 0° and 30°, and a counterclockwise rotation randomly sampled between 0° and 30°. To simulate positional shifts without causing the skeleton to exceed image boundaries, we first down-scale the skeleton by 20% and then translate it by 60–90 pixels upward, downward, leftward, or rightward. All proportion-preserving transformations maintain the relative ratios between body parts.

To expose the model to disproportionate skeletons, we apply 9 proportion-changing transformations to each sample. These operations act on halves of the body and widen or narrow their width or height asymmetrically. The scaling values are chosen so that the deformations are visually noticeable while remaining anatomically plausible. In addition to asymmetric scaling, we incorporate a swirl deformation in which the original image is warped by rotating pixels around the image center with a maximum angle of roughly 30° that decays smoothly to zero at radius R. To preserve limb straightness in the final representation, the swirl deformation is applied to the image first, after which we run pose estimation to obtain the distorted skeleton.

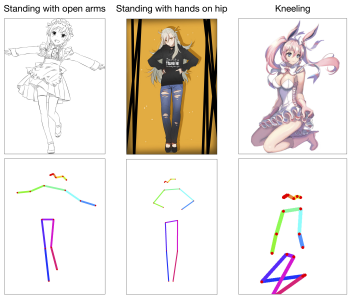

Figure 3: The sample of dataset

4.2 Implementation Details

All models are trained on a combination of Google Colab L4, Kaggle P100, and Kaggle T4×2 GPUs. Unless otherwise mentioned, we use the Adam optimizer with 0.001 learning rate and batch size 32, for 100 epochs. The dataset is split into training, validation, and test sets in a 7:1:2 ratio.

4.3 Analysis of STN Transformations

To understand the influence of the STN on this task, we analyze the skeletons produced by the STN and the corresponding affine-transformation parameters. Both visual inspection and parameter analysis indicate that the STN consistently learns to adjust translation and scale: skeletons that differ substantially in position or size are often shifted or rescaled toward a narrower range of configurations. In contrast, the rotation components of the learned affine matrices remain close to the identity transformation, even when the input has been rotated by up to 30°, suggesting that the STN makes limited use of rotational degrees of freedom under our training objective.

We further observe that the STN tends to translate skeletons toward the corners of the image, and in some cases, parts of the skeleton extend beyond the visible frame. Importantly, however, the Grad-CAM heatmaps produced from these outputs remain structurally aligned with the underlying skeletons: high-activation regions consistently follow limb segments and joints regardless of whether the skeleton is centered or partially shifted. Consequently, the lack of explicit centering does not compromise the interpretability or reliability of the heatmaps for proportion-related evaluation. Overall, the STN behaves as a coarse normalizer of size and position rather than a strict geometric aligner, and its irregular translations do not prevent the model from focusing on the structural cues relevant to proportionality.

Table 1: The accuracy evaluation results of Section 4.4 and 4.5.

Class of distortion (half × scale)	**STN**	**STN_bend**	**Deterministic**	**Deterministic_bend**
Right × 0.5	15.79%	19.32%	21.43%	35.37%
Right × 1.3	6.25%	14.81%	20.48%	22.08%
Left × 0.7	19.00%	26.32%	34.29%	9.78%
Left × 1.5	10.26%	13.58%	15.91%	10.00%
Lower × 0.5	35.96%	9.20%	36.25%	16.51%
Lower × 1.3	23.16%	15.49%	52.75%	39.51%
Upper × 0.7	13.48%	16.67%	2.60%	11.11%
Upper × 1.5	9.64%	20.29%	3.41%	4.88%
Aggregated accuracy by distortion direction				
Right	22.04%	34.13%	41.91%	57.44%
Left	29.26%	39.90%	50.19%	19.78%
Lower	59.11%	24.59%	89.00%	56.02%
Upper	23.12%	36.96%	6.01%	15.99%
Standard Deviation	21.24%	6.59%	34.06%	22.49%
Average	16.69%	16.96%	23.39%	18.65%

4.4 Effect of STN-Based Alignment

To evaluate the impact of the STN on skeleton alignment and proportionality judgment, we compare the STN-based model with a deterministic alignment strategy. In the deterministic method, the central body axis—defined as the line segment connecting the midpoint of the shoulder keypoints to the midpoint of the pelvis keypoints—is transformed to a fixed position, orientation, and length for all skeletons. Because this alignment removes variations in translation, rotation, and global scale, only proportion-changing transformations are retained when constructing the baseline dataset, resulting in 10 variants per original skeleton. This baseline dataset bypasses the STN entirely and is fed directly into the Siamese CNN. The accuracy evaluation is shown in Table 1.

Although the axis-aligned baseline achieves a higher mean accuracy, its performance exhibits extremely large variance across distortion types. The accuracy difference between detecting lower-

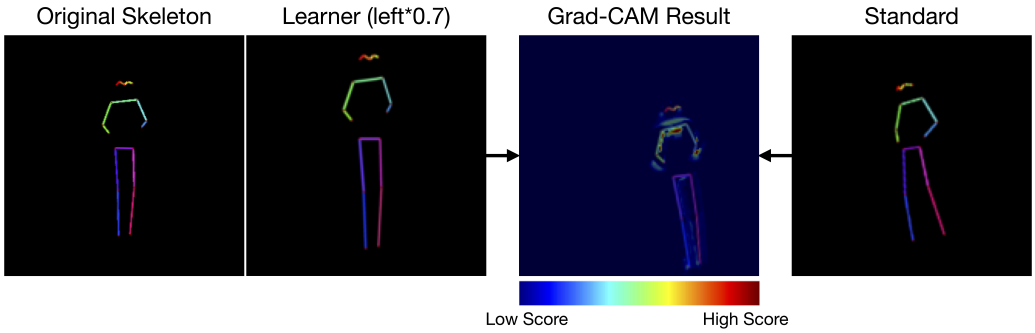

Figure 4: Example pairs of input poses and their corresponding Grad-CAM result

half and upper-half distortions reaches approximately 83%. This indicates that the baseline Siamese CNN tends to focus its activations on the upper body, even when the upper region is the one being distorted. As a result, the high accuracy observed for certain distortion directions (e.g., lower-half distortions) does not reliably reflect the model's actual ability to identify disproportion; rather, it reflects a strong directional bias in its attention pattern.

In contrast, the STN-based model shows a substantially reduced gap—about 32%—between upper- and lower-half distortions, demonstrating that the STN mitigates this bias and produces more balanced activation patterns.

Compared with the left–right distortion results, the discrepancy between upper and lower distortions is particularly pronounced. In the axis-aligned baseline, Grad-CAM activations consistently concentrate in the upper part of the skeleton, regardless of whether the upper or lower half was distorted. This suggests that the model relies heavily on shoulder-level features when comparing skeletons, implying that these features change little under vertical stretching or compression. One likely explanation is that upper-body deformations occur in a direction perpendicular to the shoulder axis, and thus the shoulder structure remains relatively unaffected; as a consequence, the model cannot rely on this region to detect vertical proportional errors.

Moreover, more than two-thirds of our dataset consists of standing poses, in which the legs occupy a large visual area and align with the vertical direction. Because both upper- and lower-half distortions alter the relative leg length, the model becomes particularly sensitive to geometric changes in the lower body. Overlay visualizations further confirm that kneeling poses—whose leg segments do not exhibit a strong alignment with any single dominant axis—do not show the same activation concentration in the upper body. This reinforces the observation that the bias arises from the interaction between vertical deformations and the predominance of upright, vertically oriented poses in the dataset.

4.5 Effect of Bent Skeletons

In the original dataset, all skeletons are strictly straight, which reflects typical usage scenarios but may limit the diversity of geometric patterns seen during training. To investigate whether the introduction of curvature can improve the model's ability to extract robust skeletal features, we modify the dataset by removing the original swirl-based disproportion and adding four additional bending transformations. Each bending transformation is implemented using a swirl-like deformation with a different maximum rotation angle, namely 15°, 30°, 45°, or 60°. These bent skeletons account for approximately 18.2% of the total samples in the STN-based setting and 28.6% in the axis-aligned setting, and are treated as auxiliary training data rather than the primary focus.

For the STN-based model, incorporating bent skeletons keeps the overall accuracy at a similar level and even yields a slight improvement in some categories. Since these samples are processed by the STN along with all others, the model is encouraged to handle variations not only in size and position but also in local curvature. This additional variety helps the STN–Siamese CNN pipeline learn more

generalizable skeletal representations, and the model becomes better at extracting features that are truly related to body proportions rather than tied to a single canonical pose.

In contrast, the axis-aligned baseline suffers a decrease in accuracy of about 5% after bent skeletons are introduced. In this setting, bending is applied before the deterministic alignment of the central axis. As a result, skeletons with strong curvature remain visually skewed even after alignment, while the baseline model has no mechanism—unlike the STN-equipped model—to further adjust or correct these residual distortions. These highly bent and misaligned examples act as outliers in the training data and confuse the Siamese CNN, which in turn degrades its ability to judge proportional correctness.

5 Conclusion

In this paper, we presented a proportionality-aware pose assessment framework for illustrated characters and realized it through an STN–Siamese CNN architecture equipped with Grad-CAM visualization. By constructing a balanced dataset that includes both proportion-preserving and proportion-changing transformations, our method enables the STN to learn meaningful spatial alignment while simultaneously guiding the Siamese CNN to capture structural differences between anatomically correct and distorted skeletons. During inference, our system automatically matches a learner's skeleton to a correctly proportioned reference and generates intuitive heatmap-based feedback that highlights correct regions while suppressing activations on distorted areas. Experimental results demonstrate that our approach provides more balanced assessment compared with a deterministic baseline, effectively reducing directional bias and maintaining performance even when additional geometric variations. Given the lack of tools for structural guidance in anime-style illustration, we believe that our proposed framework offers a promising direction for interpretable proportion analysis and has significant potential for future development in educational and artistic applications.

6 Workload Distribution

Table 2: Workload Distribution of Team Members

Name	**Main Responsibilities**
Si-Xian Hsu	Datasets construction, experiments design, model training, result analysis, paper and poster drafting
Yu-jyuan Lin	Survey, datasets construction, model training, model exploration, paper and poster drafting
Chian-ru Lin	Survey, model building, model training, paper and poster drafting

References

Shuhong Chen and Matthias Zwicker. Transfer learning for pose estimation of illustrated characters. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 793–802, 2022.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Yuhang Qiu, Jiping Wang, Zhe Jin, Honghui Chen, Mingliang Zhang, and Liquan Guo. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 72:103323, 2022.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

A APPENDIX

A.1 LOW PERFORMANCES INTERPRETAION

Although the overall accuracies of the experiments are not high, this outcome is not unexpected given that our application setting is substantially more challenging than the original rehabilitation setting, particularly in terms of dataset construction and the difficulty of the inference task. We propose several possible interpretations for these results.

The first factor is dataset complexity. Our dataset includes much higher diversity (different learner forms, and more than one standard example for each class) in order to reflect real art needs. This larger variance makes the model harder to train compared to the original rehabilitation dataset.

The second factor is inference task difficulty. Instead of comparing a learner's pose to its unique standard, a learner must be compared to a random standard, because in realistic art practice there is no fixed standard pose. This makes it harder for the model to identify which parts have twisted from the original structure.