Scaled Background Swap: Video Augmentation for Action Quality Assessment with Background Debiasing

XIN ZHANG, Hangzhou Dianzi University, China, Key Laboratory of Complex Systems Modeling and Simulation Ministry of Education, China, Zhoushan Tongbo Marine Electronic Information Research Institute, Hangzhou Dianzi University, China, and Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics, China
HONGZHI FENG, Hangzhou Dianzi University, China
M. SHAMIM HOSSAIN*, Department of Software Engineering, College of Computer and Information Sciences King Saud University, Saudi Arabia
YINZHUO CHEN, Hangzhou Dianzi University, China
HONGBO WANG, Hangzhou Dianzi University, China
YUYU YIN†, Hangzhou Dianzi University, China, Key Laboratory of Complex Systems Modeling and Simulation Ministry of Education, China, and Zhoushan Tongbo Marine Electronic Information Research Institute, Hangzhou Dianzi University, China

Action Quality Assessment (AQA) has become crucial in video analysis, finding wide applications in various domains, such as healthcare and sports. A significant challenge faced by AQA is the background bias due to the dominance of the background in videos. Especially, the background bias tends to overshadow subtle foreground differences, which is crucial for precise action evaluation. To address the background bias issue, we propose a novel data augmentation method named Scaled Background Swap. Firstly, the background regions between different video samples are swapped to guide models focus toward the dynamic foreground regions and mitigate its sensitivity to the background during training. Secondly, the video's foreground region is up-scaled to further enhance models' attention to the critical foreground action information for AQA tasks. In particular, the proposed Scaled Background Swap method can effectively improve models' accuracy and generalization by prioritizing foreground motion and swapping backgrounds. It can be flexibly applied with various video analysis models. Extensive experiments on AQA benchmarks demonstrate that Scaled Background Swap method achieves better performance

*Corresponding author.
†Corresponding author.

Authors' addresses: Xin Zhang, zhangxin@hdu.edu.cn, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018 and Key Laboratory of Complex Systems Modeling and Simulation Ministry of Education, Hangzhou, China, 310018 and Zhoushan Tongbo Marine Electronic Information Research Institute, Hangzhou Dianzi University, Zhoushan, China, 316104 and Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics, Kunming, 650221, Yunnan, China; Hongzhi Feng, fenghongzhi@hdu.edu.cn, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018; M. Shamim Hossain, mshossain@ksu.edu.sa, Department of Software Engineering, College of Computer and Information Sciences King Saud University, Riyadh, 12372, Saudi Arabia; Yinzhuo Chen, chenyinzhuo@hdu.edu.cn, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018; Hongbo Wang, whongbo@hdu.edu.cn, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018; Yuyu Yin, yinyuyu@hdu.edu.cn, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018 and Key Laboratory of Complex Systems Modeling and Simulation Ministry of Education, Hangzhou, China, 310018 and Zhoushan Tongbo Marine Electronic Information Research Institute, Hangzhou Dianzi University, Zhoushan, China, 316104.

ACM 1551-6865/2025/5-ART
https://doi.org/10.1145/3737461

than baselines. Specifically, the Spearman's rank correlation on datasets AQA-7 and MTL-AQA reaches 0.8870 and 0.9526, respectively. The code is available at: https://github.com/Emy-cv/Scaled-Background Swap.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Background Swap, Foreground Up-scale, Data Augmentation, Background Bias, Action Quality Assessment

1 INTRODUCTION

Action Quality Assessment (AQA) aims to evaluate the quality of specific actions in videos, gaining increasing attention in the scientific community in recent years [6, 11, 51, 55]. Specifically, AQA has significant potential in real-world applications, such as health care and motion analysis. For example, AQA can effectively assist evaluating the quality of surgeries in hospital, and help to improve the action scoring process in sports events such as Olympic diving. Due to rapid advancement of deep learning [21, 23, 29], significant progress has been made for video analysis and understanding, particularly in the area of action recognition [10, 28, 42], which are closely related with AQA tasks.

Generally, AQA tasks fall into three categories: regression scoring [40, 46], grading [40], and pairwise ranking [7, 26]. Specifically, regression scoring is prevalent in sports, where referees provide actual scores for videos in AQA datasets from major sporting events. Grading is common in medical skill assessment tasks, where an operator's performance is categorized into distinct levelslike expert, intermediate, and novice. Pairwise ranking tasks assess action quality by selecting two videos from a video library. The AQA tasks can be formed as learning the label matrix with C_N^2 possible combinations on a dataset with N videos. Then the performance of AQA model is measured through pairwise ranking accuracy. In this work, we explore the task of action quality assessment using the regression scoring approach.

Existing works in AQA field primarily focus on the design of regression heads [46] and feature representation [40], which have shown promising results. The crucial factor in evaluating action quality is the foreground moving object (person) in the video, rather than the background. However, a commonly overlooked issue is that two moving instances in different videos of the same scenes often share similar backgrounds. Moreover, the background usually occupies a significant portion of the entire video. These two factors lead to **background bias** [3, 16, 25, 41] as shown in Fig. 1. Therefore, when comparing a pair of videos with background bias, models tend to pay more attention to the background and overlook the importance of the moving object. Especially, compared with traditional video analysis tasks, AQA tackles a more intricate aspect of video by focusing on subtle distinctions between different video instances. In addition, the individuals in the videos often engage in actions within similar background environments, further increasing the challenge of background bias in AQA task.

To address this background bias challenge, we propose a novel data augmentation method that guides the model to focus more attention on dynamic foreground information of the video. Our inspiration stems from FAME [5], which adopts a video data augmentation method to integrate foreground and background information. Notably, FAME primarily targets action recognition, emphasizing a detailed understanding of videos. In contrast, the AQA task focuses on evaluating the overall quality of actions. Therefore, directly applying FAME to video action quality assessment may not yield satisfactory results. Therefore, we propose a novel data augmentation method to swap scaled background regions between different videos, which can effectively enhance the accuracy and generalization of AQA models.

Note that the motivation in this work is to preserve the maximum amount of motion information in the video (i.e., foreground, such as the athlete in diving) and enhance the information important for AQA tasks, while replacing the irrelated background information in the video. To achieve this, a set of foreground bounding boxes (proposed regions) are firstly generated using annotation tools or automatic detection models. The area inside

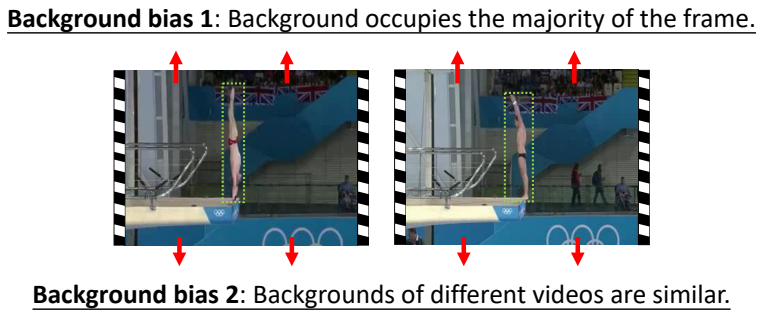

Fig. 1. Background bias in video analysis. Frames from two distinct videos in the diving action category of the AQA-7 dataset [32] show two athletes performing different actions. It is evident that the backgrounds in these videos are with significant portion and nearly identical. This phenomenon can easily lead to the background bias issue in AQA models.

the bounding boxes serve as the foreground, which is the region the model should focus on. The remaining parts serve as the background. In the same batch of data, foregrounds and backgrounds from different instances are merged to obtain the final video for the model input.

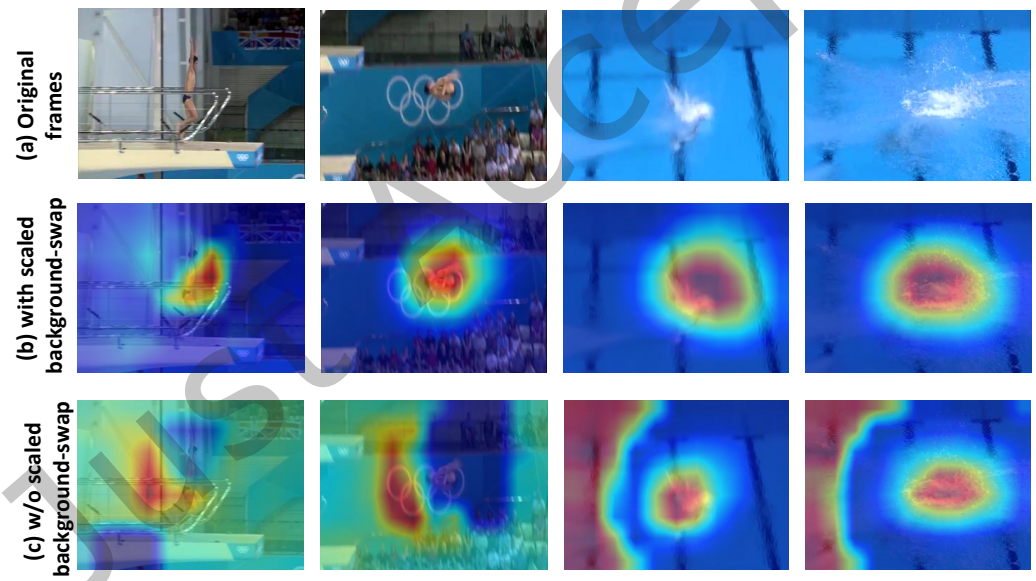

Fig. 2. Visualization [37] of feature maps with and without the proposed Scaled Background Swap data augmentation method. Red indicates regions with high attention, while blue represents regions with low attention. The backbone model for AQA is CoRe [46] method. (a) Original frames. (b) Results with scaled Background Swap data augmentation, where the model's attention is predominantly concentrated on the athletes. (c) Results without Scaled Background Swap data augmentation, where the model's attention focuses on both the athletes and the surrounding background. The visualization results demonstrate that the proposed data augmentation method significantly alleviates the background bias issue.

Fig. 2 illustrates the visualizations of feature maps with and without the proposed scaled Background Swap data augmentation method. We adopt CoRe [46] as the backbone model for AQA task. Fig. 2(b) presents the visualization of CoRe with scaled Background Swap, and Fig. 2(c) presents the visualization of CoRe without scaled Background Swap. The red areas indicate regions where the model pay more attention on, while blue areas indicate the opposite. After applying the proposed scaled Background Swap data augmentation method, the model focuses on the athlete without being influenced by the background. In contrast, without scaled Background Swap, the attentive region both includes the athlete and the background. By comparing Fig. 2(b) and Fig. 2(c), the effectiveness of the proposed scaled Background Swap data augmentation method can be demonstrated.

To validate the effectiveness of the proposed method, extensive experiments are conducted on two benchmarks, including AQA-7 dataset [32] and MTL-AQA dataset [33]. The experimental results on AQA tasks demonstrate that the proposed scaled Background Swap method enables the model to extract action features more effectively from videos. Our contributions are summarized as follows:

- We propose an effective and flexible data augmentation method, Scaled Background Swap, which facilitates the model in learning action features rather than background features within videos.
- We apply our proposed data augmentation method on different AQA models and have successfully enhanced the performance of the original models.
- We conducted extensive experiments to demonstrate the effectiveness of Scaled Background Swap data augmentation method, which achieves state-of-the-art (SOTA) performance on both the AQA-7 and MTL-AQA datasets.

The remainder of the paper is structured as follows. The existing related work is presented in Section 2. Section 3 describes the proposed model in detail. Section 4 provides detailed experimental results and analysis. We conclude our work and discuss the future work in Section 5.

2 RELATED WORK

2.1 Action Quality Acessment

Model based on Handcrafted Features. Before deep learning based methods, AQA tasks primarily rely on handcrafted features. Gordon [14] first proposed video-based AQA tasks, analyzing the feasibility of AQA techniques using gymnastics scoring as an example. Ilg et al. [18] introduced a spatiotemporal deformation model for instance comparison by establishing correspondences at both the level of global action sequences and individual motion elements. A graph-based action recognition and assessment network [2] is proposed by aligning dynamic skeleton sequence. Pirsiavash et al. [35] introduced a frequency domain analysis method for quality assessment by combining both basic visual features and advanced posture features. A generic method for online evaluation of motion quality is introduced by utilizing Kinect data. It involves reducing the dimensionality of noisy data through robust nonlinear manifold learning with the Markov assumption. However, these methods often fail to fully capture the rich visual and temporal cues required for AQA and have poor generalization capabilities across different domains.

Model based on Deep Learning. Typical neural networks like 2D-CNN, 3D-CNN, and LSTM are employed to learn and integrate features from videos, followed by performing evaluation tasks. AQA frameworks based on deep learning methods comprise modules for extracting video features and modules for evaluation. The structure of the evaluation module is closely related to the evaluation task. These methods can be categorized into two groups including design of network architecture and loss function.

Some studies improve the AQA perfomance with effective network architectures to extract more discriminative features. In the work [45], self-attention LSTM and multi-scale convolutional skip LSTM models are employed to predict TES and PCS in figure skating by capturing both detailed and global sequence features from long-term videos. S3d [43] divides the diving process into four stages including start, jump, descent, and water entry, and

employs four separate P3D [36] models to extract features. P3D has limited spatiotemporal modeling capabilities due to its separate convolutions in time and space dimensions, and its Generalizability is also poor due to its hybrid design. A graph-based joint relationship model [30] is proposed by analyzing the motion and correlations of human joints with joint commonality module and joint difference module. A novel recursive neural network [31] with a growing self-organizing structure is introduced to capture body motion information and conduct the task of matching. Wang et al. [40] proposed a feature aggregation technique known as the tubular self-attention (TSA-Net), which effectively creates comprehensive spatiotemporal context details by leveraging sparse feature interactions. Notably, the TSA-Net employs masks generated from tracking bounding boxes, inserting them into intermediate layers of I3D to guide the attention of AQA model. The proposed Scaled Background Swap method enhances data augmentation based on tracking bounding boxes to steer the model's attention more effectively. In work [20], actions are modeled as the structured process, which encodes action components within dense trajectories via an LSTM network. In summary, while these methods bring substantial improvements to feature extraction and performance prediction in AQA, they also introduce challenges related to computational demand, scalability, generalizability, and dependency on specific data characteristics.

In AQA task, effective loss function design facilitates better performance. An end-to-end framework is proposed in work [24] with C3D as the feature extractor, combining MSE loss with ranking loss. C3D-AVG-MTL framework [33] is proposed for AQA task in a multitask learning scenario. It includes three parallel prediction tasks: AQA score regression, comment generation, and action recognition. A method proposed by Tang et al. [38] learns the distribution of scores with uncertainty awareness and incorporates difficulty into the modeling process and more realistically simulating the scoring process. Gedamu et al. [12] introduced a fine-grained spatio-temporal parsing network. This network consists of an intra-sequence action parsing module capturing fine-grained subaction features and a spatiotemporal multiscale transformer module for learning long-range dependencies across multiple scales. Overall, these methods illustrate a trend towards feature-rich models in AQA, which offers significant improvements in performance.However, they also bring challenges related to complexity, computational demand, and training efficiency.

2.2 Copy-paste Data Augmentation

As one kind of data augmentation [19, 27] methods, copy-paste [13, 50] combines information from different instances and is considered a simple and effective approach in object perception learning. Several techniques fall under this category.

Mixup [49] blends two samples randomly in a certain proportion, leading to classification results that are proportionally distributed as well. CutMix [47] removes a portion of the image without filling it with zeros. Instead, it fills the pixel values randomly from other regions within the training set. The classification results align with a specific distribution ratio. Differs from Mixup and CutMix, InstaBoost [9] doesn't paste instances from other images but segments the target from the original image, then crops and pastes it to nearby areas with slight scale and rotation. The method proposed in [8] cuts images with object content by a segmentation network and extracts foreground regions. These images are then pasted onto random background images with transformations such as scale, rotation, occlusion, and added noise. The augmented images are further trained outside the original dataset. Simple Copy-Paste [13] randomly selects two images and applies random scale jitter, flipping, and pasting a randomly chosen subset of the target from one image to a random position in the other image. Despite resulting in seemingly unreasonable outcomes, experimental results show improved model accuracy.

Similarly to CutMix, FAME [5] pastes a part of one instance onto another. However, FAME utilizes the motion trends of objects in the video to guide foreground extraction, ensuring that the synthesized video sample contains object motion information, unlike CutMix, which applies random patching. Our proposed Scaled Background

Swap model in this paper addresses the problem of action quality assessment, which differs from classification tasks like image classification or action recognition. AQA imposes stricter requirements on the model, demanding the learning of motion details within videos. While FAME includes object motion information in augmented videos, it may not fully capture the entire motion object, negatively impacting action quality assessment. Therefore, the method in this paper uses single-object tracking bounding boxes as foreground extraction assistance. The resulting video, after data augmentation, can replace background information and fully capture the object's motion information throughout the sequence.

3 METHOD

To address the issue of the model overly relying on background information for quality assessment, we propose the Scaled Background Swap method to perform data augmentation. Specifically, the dynamic foreground regions (e.g. athletes) in the video are extracted by bounding boxes with a scale ratio and the background regions are swapped with the background of other videos. With the proposed data augmentation strategy, AQA model is guided to focus on the dynamic regions within the video. Specifically, before inputting the video into the model, the data is augmented using the proposed Scaled Background Swap method, where the background of one video is replaced with the background from another video within the same batch.

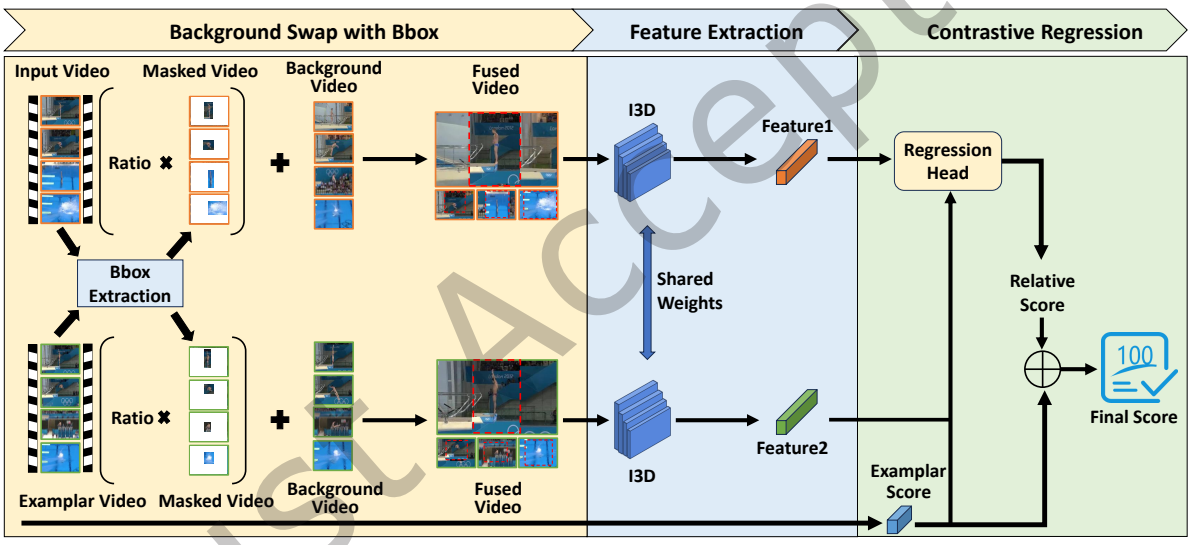

Fig. 3. Architecture of Scaled Background Swap method. Input videos include the video to be assessed and an exemplar video with a reference score. In the Background Swap with Bbox module, a single-object tracking tool is used to annotate the bounding boxes of the athlete in each frame. Then the scaled foreground masks are generated, which specify the foreground regions to be overlaid onto the background of the other video, resulting in an augmented video. Two I3D models extract spatiotemporal features from the two videos, which are concatenated into a vector along with the reference score. The vector is then input into an AQA regression head for predicting relative score [46], which is summed with the reference score to get the final score.

The network architecture is illustrated in Fig. 3. It takes as input a pair of action videos, and there are L frames in each video. Specifically, each video is divided into C segments. Then, a single-object tracker or a single-object tracking annotation tool is employed to obtain bounding boxes of all the moving objects in each frame of video. Specifically, to make sure the objects are completely enclosed in the box, we up-scale the box with a fixed

ratio. Pixels within the bounding boxes are considered as foreground, while the remaining pixels are treated as background. Then background regions of two videos are swapped to facilitate the model to pay more attention on the foreground region.

The augmented videos are then fed into an AQA model. Here, we employ CoRe [46] as the AQA model, where I3D (Inflated 3D ConvNet) is first utilized to extract features of both the input video and the exemplar video. The obtained features, along with the exemplar score, are then input into a group-aware regression tree (GART) for calculating the relative score of the two videos. The final predicted score for the input video is obtained by combining the relative score and the exemplar score. Note that the proposed Scaled Background Swap method can also enhance performance when integrated with other AQA models. We use the CoRe [46] method, which includes a Feature Extraction module and a Group-aware Contrastive Regression module (shown in Fig. 3), as the baseline AQA model in our architecture.

3.1 Scaled Background Swap

In the field of self-supervised representation learning, FAME [5] addresses background bias using the copy-paste data augmentation method. However, experimental results demonstrate that directly applying FAME to action quality assessment not only fails to improve the model's accuracy but also results in a certain degree of accuracy degradation. The visualization of FAME's results is shown in Fig. 4 and reveals that it cannot fully capture the motion of the main object throughout the entire video. This limitation has a significant negative impact on action quality assessment. AQA in real-world often requires observing the performance of moving objects throughout the entire motion sequence. For example, in diving competition, judges need to consider key moments such as take-off, mid-air rotations, and entry into the water to provide an objective score. Therefore, we propose a novel data augmentation method to improve the model's accuracy in action quality assessment tasks.

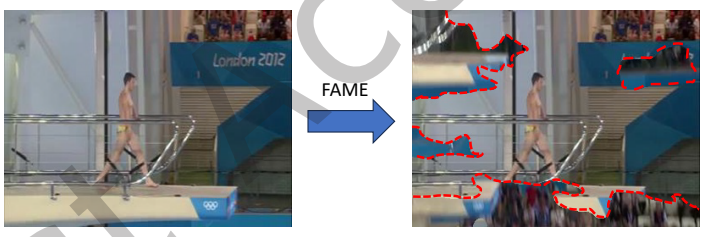

Fig. 4. Visualization of FAME [5]. It can be observed that FAME's foreground extraction algorithm is relatively coarse and fails to accurately capture the athletes in the video. Additionally, FAME applies the same mask to all frames. For action quality assessment, a comprehensive analysis of the athlete's performance at each stage is needed. This uniform foreground extraction approach actually decreases the model's accuracy.

The proposed Scaled Background Swap method leverages single-object tracking bounding boxes to extract foreground region. The videos obtained after data augmentation can replace background information while capturing the object's motion information throughout the sequence. The approach intends to preserve the foreground regions from the original videos and shuffle the backgrounds among different videos. Specifically, Scaled Background Swap initially divides the video into foreground (dynamic) region and background (static) region and then combines the foreground and background from different videos with each other.

The input video is represented as $X \in R^{C \times T \times H \times W}$, where C, T, H and W respectively denote the number of channel, frame, height and width. To extract the foreground, we utilize a semi-automatic object tracking annotation

tool DarkLabel[1] to generate bounding boxes for the moving objects in the video. It annotates bounding boxes for the starting and ending frames and then generates bounding boxes of all intermediate frames by interpolation. While other approaches, such as deep learning-based single-object tracking models, can also be employed to generate video bounding boxes automatically. It should be noted that achieving precise object tracking is not the primary focus of our method, and the bounding boxes merely provide a rough region of the video foreground. The generated foreground bounding boxe set is denoted as $B = \{b_l\}_{l=1}^{l=L}$, where $b_l = (x^l_{\min}, y^l_{\min}, x^l_{\max}, y^l_{\max})$ is the bounding box of the frame l. $(x^l_{\min}, y^l_{\min})$ is the coordinate of the top-left corner of the bounding box, and $(x^l_{\max}, y^l_{\max})$ is the coordinate of the bottom-right corner of the bounding box. Then, a corresponding foreground mask M^l_{fg} is generated using bounding box b_l, which is formulated as follows,

$$\left[M^l_{fg}\right]_{ij} = \begin{cases} 1, & x^l_{\min} \le i \le x^l_{\max} \text{ and } y^l_{\min} \le j \le y^l_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where (i, j) is a coordinate within frame l. If a pixel (i, j) is within the foreground bounding box of the frame l, it is assigned the value 1. Otherwise, it is assigned 0.

Subsequently, we upscale the foreground bounding box with a factor $ratio$. Specifically, the dimensions (height and width) of bounding box b_l are scaled up by the factor $ratio$. We denote the upscaled foreground bounding box by $sb_l = (\widetilde{x}^l_{\min}, \widetilde{y}^l_{\min}, \widetilde{x}^l_{\max}, \widetilde{y}^l_{\max})$. $(\widetilde{x}^l_{\min}, \widetilde{y}^l_{\min})$ is the coordinate of the top-left corner of the upscaled bounding box, and $(\widetilde{x}^l_{\max}, \widetilde{y}^l_{\max})$ is the coordinate of the bottom-right corner of the upscaled bounding box. The coordinates are calculated as follows,

$$\begin{cases} \widetilde{x}^l_{\min} = \max\left(\frac{x^l_{\min}+x^l_{\max}}{2} - \frac{x^l_{\max}-x^l_{\min}}{2} \times \text{ratio}, 0\right) \\ \widetilde{x}^l_{\max} = \min\left(\frac{x^l_{\min}+x^l_{\max}}{2} + \frac{x^l_{\max}-x^l_{\min}}{2} \times \text{ratio}, W\right) \\ \widetilde{y}^l_{\min} = \max\left(\frac{y^l_{\min}+y^l_{\max}}{2} - \frac{y^l_{\max}-y^l_{\min}}{2} \times \text{ratio}, 0\right) \\ \widetilde{y}^l_{\max} = \min\left(\frac{y^l_{\min}+y^l_{\max}}{2} + \frac{y^l_{\max}-y^l_{\min}}{2} \times \text{ratio}, H\right) \end{cases} \tag{2}$$

The Eq. (2) ensures that the upscaled foreground bounding box should be within the frame l. Then, the corresponding upscaled foreground mask SM^l_{fg} is calculated by the following formula:

$$[SM^l_{fg}]_{ij} = \begin{cases} 1, & \text{if } \widetilde{x}^l_{\min} \le i \le \widetilde{x}^l_{\max} \text{ and } \widetilde{y}^l_{\min} \le j \le \widetilde{y}^l_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The parameter $ratio$ is a hyperparameter to determine the scaling ratio of the foreground. The introduction of the $ratio$ hyperparameter aims to address the following concerns: if foreground masks are generated solely based on the original bounding box, the extracted foreground might be incomplete. This limitation arises because the generated bounding boxes only provide an approximation of the foreground region rather than an exact contour. Moreover, it has been noted that videos featuring small dynamic objects (such as those found in snowboarding videos) tend to undergo extensive content replacement. This excessive replacement is considered undesirable.

Finally, we generate the augmented video X_{fusion} for the input video X and the exemplar video Y. X_{fusion} is composed of the upscaled foreground in X and background in Y. The augmented video X_{fusion} is formulated as

$$X_{fusion} = X \otimes SM_{fg} + Y \otimes (1 - SM_{fg}), \tag{4}$$

where $\otimes$ represents element-wise multiplication.

Fig. 5 visually illustrates the swapping augmentation process of our proposed method. The figure consists of two rows, each showcasing a video from the same batch (assuming a batch of size 2). The green boxes represent

[1]DarkLabel could be downloaded at https://github.com/darkpgmr/DarkLabel

the foreground bounding boxes annotated by DarkLabel, which outline specific objects or regions of interest within the videos. To generate the augmented video, the dimensions of the green boxes are scaled by a factor ratio, resulting in red boxes that depict the scaled foregrounds. The foregrounds are then superimposed onto each other's backgrounds, effectively swapping the original backgrounds. This process generates an augmented video through the extraction and swapping of foregrounds.

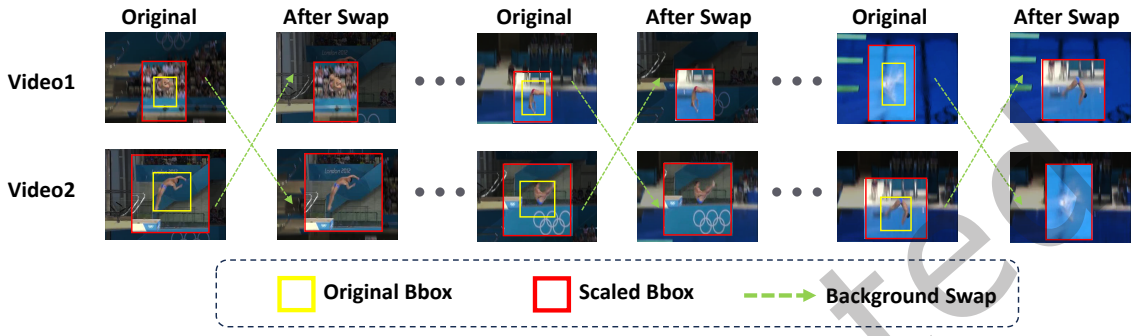

Fig. 5. Visualization of the proposed Scaled Background Swap. Yellow boxes represent the original bounding boxes, while red boxes are generated by scaling the yellow boxes with a specified ratio. The region outside the red boxes represents the background of the video. The swapping of backgrounds between two videos generates the fused videos. To more clearly illustrate the swapping augmentation process, we darken the background areas to achieve higher contrast between the foreground and background.

3.2 Bounding Boxes Extraction

There are three methods to generate the foreground bounding boxes mentioned in the previous section. The first method involves simply selecting a rectangular region at the center of the video as the foreground. The ablation studies in the experimental section demonstrate that this straightforward foreground extraction method significantly enhances model's performance.

The second method involves using a video annotation tool called DarkLabel[2] . DarkLabel uses a linear interpolation algorithm to achieve semi-automatic annotation. In practice, annotating three to four bboxes per video is generally sufficient, with the remaining bboxes generated through linear interpolation.

The third method is based deep learning technique, which commonly employs the VOT [39] method to annotate bounding boxes. VOT tracks the target's position in each subsequent frame after being provided with the initial target position in the first frame. However, since VOT requires fine-tuning to adapt to the new dataset before produing satisfactory bbox, its annotation efficiency is not as high as that of DarkLabel. Moreover, the bboxes generated by VOT do not provide any advantages over those produced by DarkLabel, so we choose DarkLabel over VOT as the annotation tool.

4 EXPERIMENTS

We have conducted extensive experiments about the proposed method as well as baselines on two datasets, i.e., AQA-7 [32] and MTL-AQA [33], to validate its effectiveness. Especially, the combination of the proposed data augmentation method with AQA model CoRe [46] achieves state-of-the-art (SOTA) results for AQA task. The following key questions guide our experimental design:

[2]DarkLabel could be downloaded at https://github.com/darkpgmr/DarkLabel

- RQ1: What is the performance of Scaled Background Swap in comparison to SOTA methods?
- RQ2: How does scale *ratio* affect performance?
- RQ3: How do different foreground annotations affect the proposed method?
- RQ4: How does Scaled Background Swap perform when employed to different AQA models?

4.1 Experimental Settings

In this section, the datasets and the implementations are introduced in details.

Datasets. Two real-world datasets, including AQA-7 [32] and MTL-AQA [33], are adopted in this work. Specifically, AQA-7 dataset comprises videos from seven different actions, totaling 1189 samples. Among these, 803 videos are utilized as training set, while the rest 303 videos are used as testing set. To facilitate comparison with baseline models, we have excluded the trampoline category due to the lengthy video content. Besides, the MTL-AQA dataset comprises 1412 videos of diving, which belong to 16 different projects. Additionally, MTL-AQA contains detailed scores from every judge, diving difficulty, and the live commentary. Following the dataset splitting strategy in [33], 1059 videos are utilized as training set, and 353 samples are utilized as testing set.

Implementation Details. We use I3D as a feature extractor, which is pretrained on the Kinetics dataset. Adam optimizer is employed with weight decay set as 0. Besides, in the regression trees, the corresponding learning rate is set as 1×10^{-4}. For the I3D network, the learning rate is set as 1×10^{-3}. For datasets AQA-7 and MTL-AQA, 103 frames are extracted from each video and divided into 10 mutually overlapping segments. Each segment contains 16 frames. For AQA-7 dataset, exemplary videos are chosen from the same action category. For example, if the input video belongs to the "gym_vault" category, an exemplary video is randomly selected from the training set of the "gym_vault" category. For MTL-AQA dataset, exemplary videos are selected based on the difficulty (DD) information provided in the annotation file, ensuring videos from the same category and difficulty level are chosen. DarkLabel is used for generating bounding boxes for all videos in the dataset.

4.2 Quantitative Comparison with SOTA Methods

In this section, we quantitatively compare our method with state-of-the-art (SOTA) baselines (RQ1) on datasets AQA-7 [32] and MTL-AQA [33]. The Spearman's rank correlation metric S_ρ and relative L_2 distance metric R_{l2} are adopted for evaluation.

In Table 1, we present results for seven different sports categories in AQA-7 dataset with two metrics. Compared with CoRe [46], we achieve performance improvements of 1.9%, 4.5%, 9.6%, 11.1%, 2.1%, and 4.9% across different sport categories evaluated by Spearman's rank correlation. Furthermore, results of the relative L_2 distance (scaled by 100) decrease by 7.8%, 4.5%, 26.7%, 23.5%, -21.9%, 9.3% for different action categories. The average performance improvements for metrics S_ρ and R_{l2} are 5.6% and 16.5%, respectively. The average Spearman's rank correlation is calculated using the Fisher z-transform. Performance improvement strongly validates that the proposed method is effective in AQA tasks.

To intuitively show the differences between our method and the baseline CoRe [46] method, we visualize the prediction results on the AQA7 dataset in the form of scatter plots, as shown in Fig. 6. Compared with the baseline (left in Fig. 6), our method (right in Fig. 6) demonstrates significantly higher precision. The black line indicates the correct scores. The predictions of our method are clustered around the black line, whereas the results of the baseline are more dispersed. Our proposed Scaled Background Swap method can effectively guide the model to focus on foreground information and improve the performance. Specifically, our method achieves more accurate predictions than the baseline in 61.72% of cases. Moreover, our prediction score error is less than 5 in 71.29% of cases, whereas the baseline CoRe is only 66.67%.

To further validate the robustness and effectiveness of our proposed method, we conducted experiments on the MTL-AQA dataset with fixed bounding boxes. Table 2 illustrates the performance of SOTA methods and

Table 1. Performance comparison on AQA-7 dataset. Δ represents the improvement of CoRe [46] after applying our proposed data augmentation method. The best result is denoted with **bold** text. The second best result is denoted with underline text.

Sp. Corr(ρ)	Gym Vault	Diving	BigSnow.	BigSki.	Sync. 10m	Sync. 3m	Avg. Corr.
C3D-LSTM [34]	0.5636	0.6047	0.5029	0.4593	0.6927	0.7912	0.6165
JRG [30]	0.7358	0.7630	0.5405	0.6006	0.9254	0.9013	0.7849
GDLT [44]	0.7806	0.8735	0.6122	0.6380	0.9049	0.9078	0.8164
USDL [38]	0.7570	0.8099	0.7109	0.6538	0.8878	0.9166	0.8102
CoFInAI [53]	0.7685	0.8652	0.5990	0.6119	0.9334	0.9073	0.8194
AdaST [52]	0.7687	0.8065	0.5921	0.7306	0.9353	0.9620	0.8443
HGCN [54]	0.7725	0.8871	0.6487	0.6701	0.9507	0.9174	0.8450
TSA-Net [40]	0.8004	0.8379	0.6962	0.6657	0.9334	0.9493	0.8476
TECN [22]	**0.8156**	0.8604	0.5755	0.7314	0.9417	0.9432	0.8504
DAE-CoRe [48]	0.7786	0.8923	0.6842	0.7102	0.9129	0.9506	0.8520
SSPR [17]	0.7593	0.8766	0.6308	0.7222	0.9435	0.9508	0.8538
TPT [1]	0.8043	0.8969	0.6965	0.7336	**0.9545**	0.9456	0.8715
CoRe [46]	0.7746	0.8824	0.6624	0.7115	0.9078	0.9442	0.8401
Ours	0.8096	**0.8996**	**0.7361**	**0.7802**	0.9523	**0.9645**	**0.8870**
Δ	0.0350↑	0.0172↑	0.0737↑	0.0687↑	0.0445↑	0.0203↑	0.0469↑
R-ℓ_2(×100)	Gym Vault	Diving	BigSnow.	BigSki.	Sync. 10m	Sync. 3m	Avg. R-ℓ_2
DAE-CoRe [48]	2.32	0.85	4.58	6.93	3.04	2.69	3.40
CoFInAI [53]	2.07	1.50	5.39	4.57	2.56	0.56	2.78
HGCN [54]	3.61	1.01	3.70	5.61	1.47	1.19	2.77
USDL [38]	2.09	0.79	4.94	4.82	2.14	0.65	2.57
GDLT [44]	2.35	0.79	4.17	4.36	2.17	0.79	2.44
TPT [1]	**1.69**	**0.53**	3.30	2.89	**1.33**	**0.33**	**1.68**
CoRe [46]	1.78	0.64	3.87	3.67	2.35	0.41	2.12
Ours	1.70	0.59	**2.96**	**2.69**	2.13	0.50	1.77
Δ	0.08↓	0.05↓	0.91↓	0.98↓	0.22↓	0.09↑	0.35↓

our approach on the MTL-AQA dataset. Our experimental setup follows the CoRe framework. Regarding the generation of bounding boxes, we adopt a fixed bounding box strategy, wherein a rectangle centered at the video's midpoint is used as the foreground bounding box, with both length and width set to 80% of the original video dimensions. With this simple and straightforward bounding box generation strategy, the proposed method still achieves the best performance in metric of Spearman's rank correlation value (0.9526).

The quantitative results in Table 1 and Table 2 demonstrate that the proposed method effectively directs the model's attention to the foreground by swapping the background regions, consequently achieving high-quality

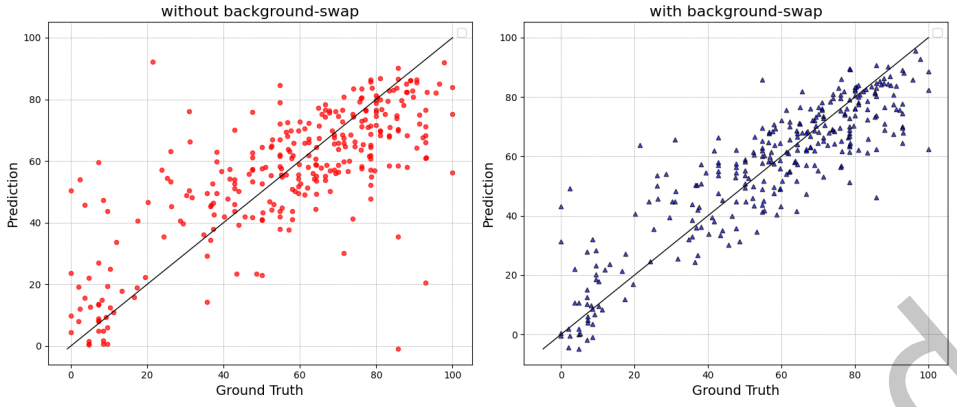

Fig. 6. Comparison of the baseline method CoRe [46] (left) and our method (right) on AQA7 dataset in scatter plot. Each point in the figure represents a video in the test set. The black line indicates the label score.

action assessment results. Our approach is not sensitive to the boundaries of the foreground region. The key to the effectiveness of the method lies in swapping with up-scaled foreground region.

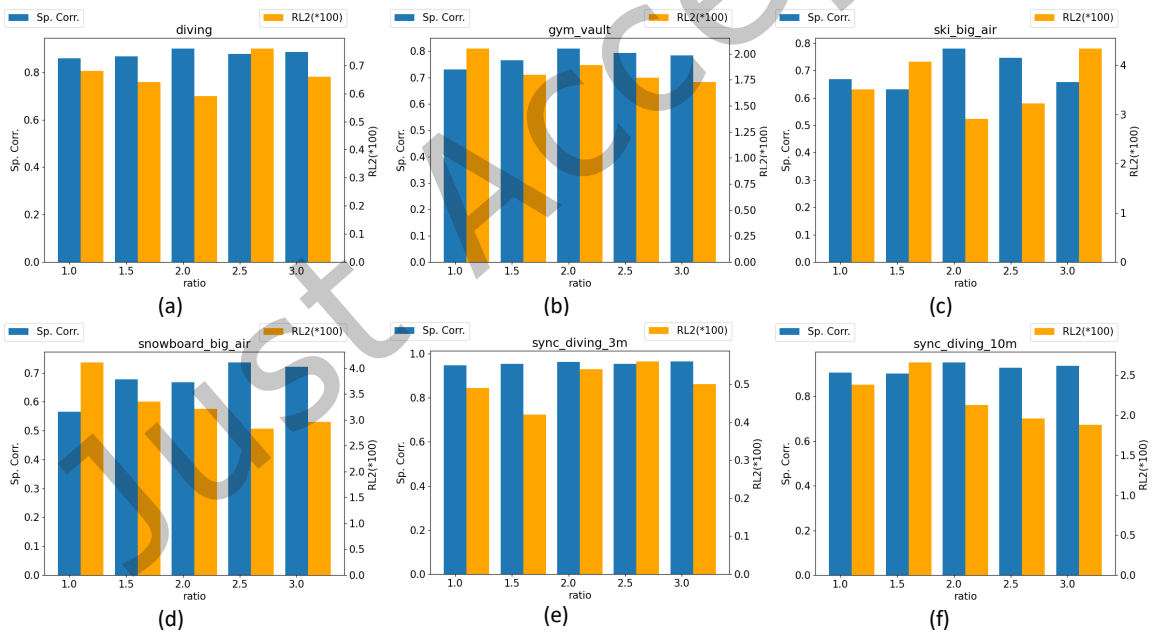

Fig. 7. Different Scale Ratios on the AQA-7 Dataset. (a) to (f) show the results of six different action categories. The horizontal axis represents the scale ratio, and the vertical axis represents quantitative metric values. The blue bars represent Spearman's rank correlation, while the orange bars represent the relative L2 distance. Higher Spearman's rank correlation indicates better model performance, and lower relative L2 distance signifies better model performance.

Table 2. Performance comparison on MTL-AQA dataset. Δ represents the improvement of CoRe [46] after applying our proposed data augmentation method. The best result is denoted with **bold** text. The second best result is denoted with underline text.

Method	Sp. Corr(ρ)
C3D-LSTM [34]	0.8489
MSCADC-MTL [33]	0.8612
TECN+OS [22]	0.8745
C3D-AVG-MTL [33]	0.9044
USDL [38]	0.9066
TECN+ES [22]	0.9095
MUSDL [38]	0.9273
MUSDL+PECop [4]	0.9372
GDLT [44]	0.9395
CoFInAI [53]	0.9461
DAE [48]	0.9497
CoRe+PECop [4]	0.9520
HGCN [54]	0.9522
CoRe [46]	0.9512
Ours	**0.9526**
Δ	0.0014↑

4.3 Ablation Study

Scale Ratio (RQ2). The scale *ratio* to adjust the size of the foreground plays a crucial role in improving the precision of the action score results. We conduct ablation study with different settings of the scale ratio for six action categories in AQA-7. The results are presented in Fig. 7. The horizontal axis represents the scale ratio. The vertical axis represents quantitative metric values. The scale ratio varies from 1 to 3, with intervals of 0.5. The blue bars represent Spearman's rank correlation, while the orange bars represent the relative L2 distance. Higher Spearman's rank correlation indicates better model performance, and lower relative L2 distance signifies better model performance. Considering the six action categories, it can be observed from Fig. 7 that when the ratio is set to 2 or 2.5, Spearman's rank correlation is the highest. Additionally, relative L2 distance demonstrates good accuracy when the ratio is set to 2 or 2.5 (except for sync-diving-3m). However, setting the ratio to 1 does not achieve optimal accuracy, as explained in Fig. 8 below.

In Fig. 8, samples from the action categories of gym vault, ski big air, and snow big air are presented. Each set of images shows the effects after Background Swap for scale ratios set to 1 and 2. When the scale ratio is set to 1, there is a larger deviation from the real scores. This is because as the ratio decreases, the proportion of the foreground in the entire video becomes smaller. When most of the content in the input video comes from the background of another video, the model finds it more challenging to capture information from the foreground, which constitutes only a small portion. Therefore, setting a reasonable scale ratio has a significant impact on the accuracy of the score.

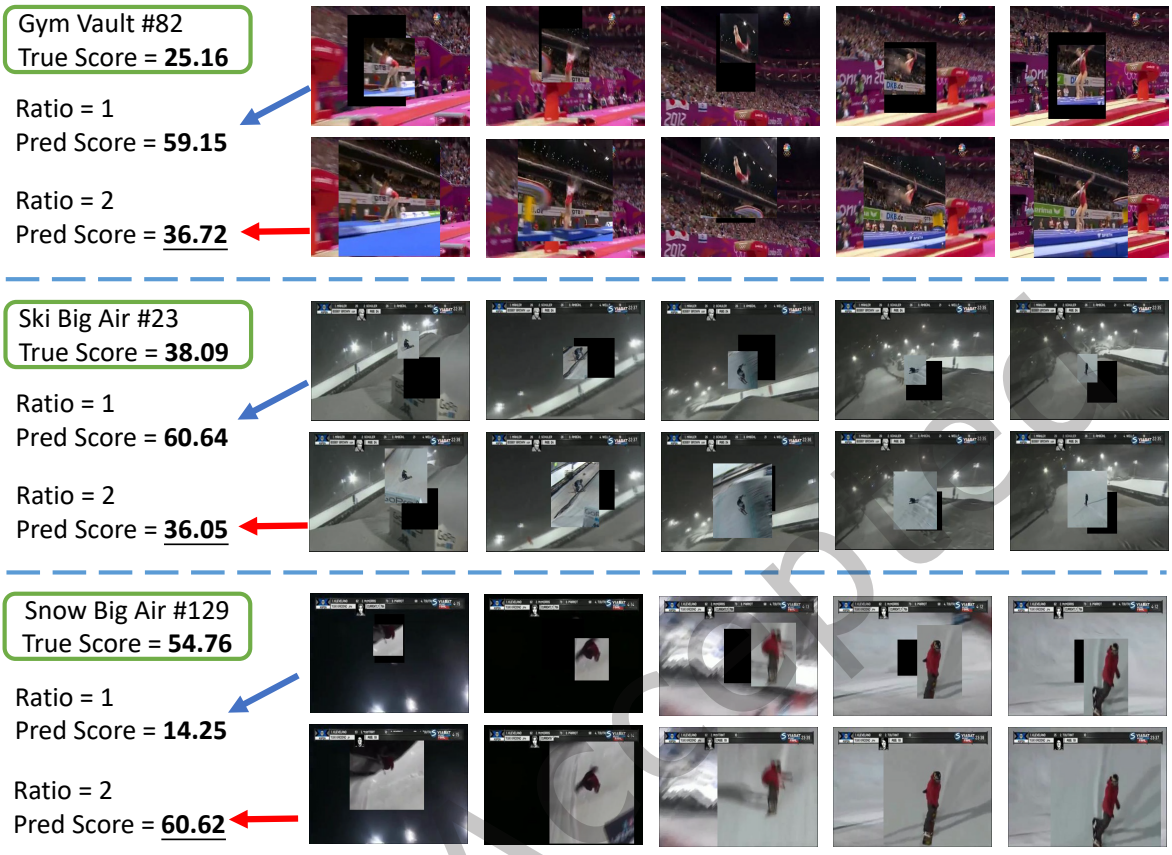

Fig. 8. Cases with Different Scale Ratio Settings: Instances from three different action categories (Gym Vault, Ski Big Air, Snow Big Air) are provided for $ratio = 1$ and $ratio = 2$. With $ratio = 1$, the foreground in the video occupies only a small portion, with most of the video content coming from another video. This situation results in the proposed method being overly aggressive, leading to significant deviations from actual scores. However, when $ratio = 2$, the foreground not only includes the athlete but also retains some background from the original video. This results in more accurate predictions.

Foreground Sources (RQ3). The Background Swapping data augmentation method proposed in this paper is highly flexible. As long as a foreground region is provided, our method can be applied to enhance the network's ability of extracting features from important foreground regions. The source of the foreground region can be diverse, such as manual annotation, a fixed region at the video center, or a single-object tracker. In this work, DarkLabel is employed for manual annotation. This tool only requires marking bounding boxes in keyframes, and it automatically completes all frames between two keyframes using interpolation. In practice, annotating three to four bounding boxes for a 103-frame AQA-7 video is sufficient. For comparison, a fixed region at the video center is also used as the foreground. Additionally, single-object trackers based on deep learning are mature, and they can also be employed to extract bounding boxes. However, since single-object trackers require annotating the bounding box in the first frame and involve a certain inference time, the cost of obtaining bounding boxes is higher than using DarkLabel for annotation.

Table 3. Influence of Different Types of Bbox.

Foreground Type	Sp. Corr(ρ)	R-ℓ_2
Baseline	0.8401	2.12
Fixed	0.8523	2.01
DarkLabel	**0.8870**	**1.77**

As shown in Table 3, "Baseline" represents the experimental results of CoRe, and "Fixed" denotes using a rectangle centered at the video center as the foreground, with both length and width set to 80% of the original video dimensions. It can be observed that, without any manual annotation, simply selecting a fixed rectangle as the foreground already yields a significant improvement compared with CoRe. Moreover, when providing more accurate foreground through manual annotation, the model's performance shows a substantial improvement. This validates that the proposed method is highly effective and flexible.

4.4 Analysis of Generalizability

To assess the generalization capability of our method, we integrate our Scaled Background Swap method into three action quality assessment networks, including USDL [38], MUSDL [38], and DAE [48] (RQ4). We maintain the default configurations specified in their respective papers to ensure fairness. We then apply our Scaled Background Swap method as data augmentation process before feeding the data into these models. The results, as illustrated in Fig. 9, show improvements in Spearman's rank correlation by 2.04%, 1.31%, and 0.86% for USDL, MUSDL and DAE models, respectively. Additionally, the error metric relative L2-distance R-L_2 decreases by 13.8%, 15.4%, and 11.4% across these models. These results confirm that our method can effectively and generally improve the performance of action quality assessment networks. It offers effective plug-and-play data augmentation across different models.

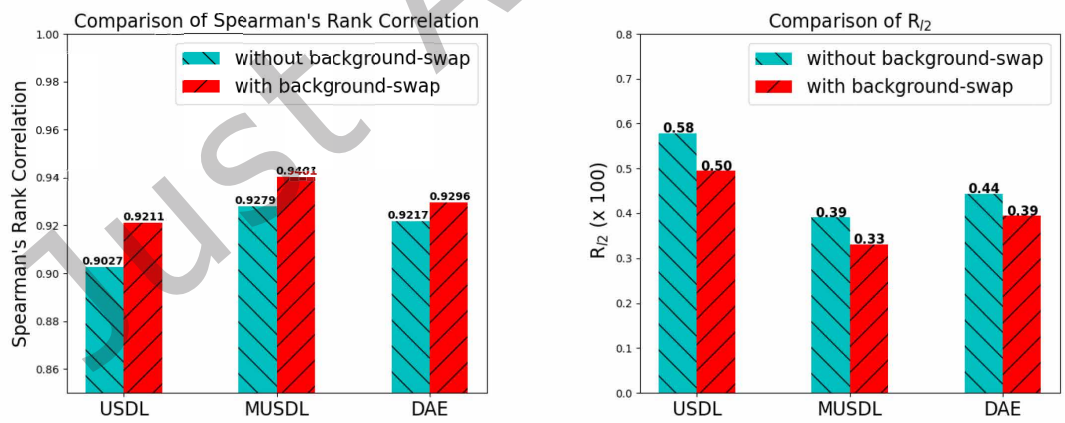

Fig. 9. Performance of the proposed Scaled Background Swap with different AQA models. The left is Spearman's rank correlation with and without the proposed scaled Background Swap method. The right shows relative L2 distance R-L_2. Note that the higher Spearman's rank correlation is the better, while the lower R-L_2 is the better.

5 CONCLUSION

This paper proposed a novel data augmentation method (named as Scaled Background Swap) for action quality assessment (AQA) task. Our method can effectively address the background bias issue particularly in video understanding field. Foreground and background regions are extracted within a video using foreground bounding boxes. These regions are then swapped across a batch of videos, resulting in fused videos inputted into the model for score prediction. We conducted extensive experiments on two AQA datasets, which show the effectiveness of the designed Scaled Background Swap method. The results show the potential to mitigate the bias towards backgrounds and improve the performance of AQA models.

For future work, we will extend the application of the proposed method to diverse video analysis domains, expanding its potential impact. Additionally, we aim to investigate the integration of this data augmentation technique with current state-of-the-art model architectures, such as the Transformer [15] and multi-modal [56], seeking to achieve further improvements in performance and generalization.

ACKNOWLEDGMENTS

This research was supported by the Researchers Supporting Project no RSP2025R32, King Saud University, Riyadh, Saudi Arabia. This research was also supported by Zhejiang Province High-Level Talent Special Support Program-Leading Talent of Technological Innovation under No. 2022R52043, the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant No. GK259909299001-019, the National Natural Science Foundation of China under Grant No. 62402151, Zhejiang Provincial Natural Science Foundation of China under Grant No. ZCLMS25F0201, the Foundation of Yunnan Key Laboratory of Service Computing under Grant No. YNSC24124.

REFERENCES

[1] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. 2022. Action quality assessment with temporal parsing transformer. In *European conference on computer vision*. Springer, Springer Nature Switzerland, Cham, 422–438.

[2] Oya Çeliktutan, Ceyhun Burak Akgul, Christian Wolf, and Bülent Sankur. 2013. Graph-based analysis of physical exercise actions. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*. Association for Computing Machinery, New York, NY, USA, 23–32.

[3] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[4] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. 2024. Pecop: Parameter efficient continual pretraining for action quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 42–52.

[5] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. 2022. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9716–9726.

[6] Yuning Ding, Sifan Zhang, Liu Shenglan, Jinrong Zhang, Wenyue Chen, Duan Haifei, Bingcheng Dong, and Tao Sun. 2024. 2M-AF: A Strong Multi-Modality Framework For Human Action Quality Assessment with Self-supervised Representation Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1564–1572.

[7] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. 2019. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Institute of Electrical and Electronics Engineers (IEEE), United States, 7862–7871.

[8] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*. 1301–1310.

[9] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. 2019. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 682–691.

[10] Zhanzhou Feng, Jiaming Xu, Lei Ma, and Shiliang Zhang. 2024. Efficient video transformers via spatial-temporal token merging for action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 4 (2024), 1–21.

[11] Honghao Gao, Si Yu, Muddesar Iqbal, and Mohsen Guizani. 2024. ResFNN: Residual Structure-Based Feedforward Neural Network for Action Quality Assessment in Sports Consumer Electronics. *IEEE Transactions on Consumer Electronics* 70, 4 (2024), 6653–6663.

[12] Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. 2023. Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing* 32 (2023), 6386–6400.
[13] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2918–2928.
[14] Andrew S Gordon. 1995. Automated video assessment of human performance. In *Proceedings of AI-ED*, Vol. 2.
[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.
[16] Yun He, Soma Shirakabe, Yutaka Satoh, and Hirokatsu Kataoka. 2016. Human action recognition without human. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, Springer International Publishing, Cham, 11–17.
[17] Feng Huang and Jianjun Li. 2024. Assessing action quality with semantic-sequence performance regression and densely distributed sample weighting. *Applied Intelligence* 54, 4 (2024), 3245–3259.
[18] Winfried Ilg, Johannes Mezger, and Martin Giese. 2003. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 523–531.
[19] Farkhund Iqbal, Ahmed Abbasi, Abdul Rehman Javed, Ahmad Almadhor, Zunera Jalil, Sajid Anwar, and Imad Rida. 2024. Data augmentation-based novel deep learning method for deepfaked images detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 11 (2024), 1–15.
[20] Seong Tae Kim and Yong Man Ro. 2017. Evaluationnet: Can human skill be evaluated by deep networks? *arXiv preprint arXiv:1705.11077* abs/1705.11077 (2017).
[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
[22] Ming-Zhe Li, Hong-Bo Zhang, Li-Jia Dong, Qing Lei, and Ji-Xiang Du. 2023. Gaussian guided frame sequence encoder network for action quality assessment. *Complex & Intelligent Systems* 9, 2 (2023), 1963–1974.
[23] Shenglan Li, Rui Yao, Yong Zhou, Hancheng Zhu, Jiaqi Zhao, Zhiwen Shao, and Abdulmotaleb El Saddik. 2024. Motion-aware Self-supervised RGBT Tracking with Multi-modality Hierarchical Transformers. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 12 (2024), 1–23.
[24] Yongjun Li, Xiujuan Chai, and Xilin Chen. 2018. End-to-end learning for action quality assessment. In *Pacific Rim Conference on Multimedia*. Springer, Springer International Publishing, Cham, 125–134.
[25] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 513–528.
[26] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. 2019. Manipulation-skill assessment from videos with spatial attention network. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 0–0.
[27] Ziyan Li, Jianfei Yu, Jia Yang, Wenya Wang, Li Yang, and Rui Xia. 2024. Generative Multimodal Data Augmentation for Low-Resource Multimodal Named Entity Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 7336–7345.
[28] Shuang Liang, Wentao Ma, and Chi Xie. 2024. Relation with Free Objects for Action Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2024), 1–19.
[29] Saksham Mittal, Mohammad Wazid, Devesh Pratap Singh, Ashok Kumar Das, and M Shamim Hossain. 2025. A deep learning ensemble approach for malware detection in Internet of Things utilizing Explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence* 139 (2025), 109560.
[30] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. 2019. Action assessment by joint relation graphs. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6331–6340.
[31] German I Parisi, Sven Magg, and Stefan Wermter. 2016. Human motion assessment in real time using recurrent self-organization. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 71–76.
[32] Paritosh Parmar and Brendan Morris. 2019. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1468–1476.
[33] Paritosh Parmar and Brendan Tran Morris. 2019. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 304–313.
[34] Paritosh Parmar and Brendan Tran Morris. 2017. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 20–28.
[35] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. 2014. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, Springer International Publishing, Cham, 556–571.

[36] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*. 5533–5541.
[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[38] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. 2020. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9839–9848.
[39] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. 2016. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1420–1429.
[40] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. 2021. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM international conference on multimedia*. Association for Computing Machinery, New York, NY, USA, 4902–4910.
[41] Yang Wang and Minh Hoai. 2018. Pulling actions out of context: Explicit separation for effective combination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7044–7053.
[42] Xuezhi Xiang, Xiaoheng Li, Xuzhao Liu, Yulong Qiao, and Abdulmotaleb El Saddik. 2024. A GCN and Transformer complementary network for skeleton-based action recognition. *Computer Vision and Image Understanding* 249 (2024), 104213.
[43] Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran. 2018. S3d: Stacking segmental p3d for action quality assessment. In *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 928–932.
[44] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. 2022. Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3232–3241.
[45] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. 2019. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology* 30, 12 (2019), 4578–4590.
[46] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. 2021. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7919–7928.
[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.
[48] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. 2024. Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications* 36, 2 (2024), 929–942.
[49] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb
[50] Linjuan Zhang, Kong Aik Lee, Lin Zhang, Longbiao Wang, and Baoning Niu. 2024. CPAUG: Refining Copy-Paste Augmentation for Speech Anti-Spoofing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 10996–11000.
[51] Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. 2022. Semi-Supervised Action Quality Assessment With Self-Supervised Segment Feature Recovery. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 9 (2022), 6017–6028. https://doi.org/10.1109/TCSVT.2022.3143549
[52] Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. 2023. Adaptive stage-aware assessment skill transfer for skill determination. *IEEE Transactions on Multimedia* 26 (2023), 4061–4072.
[53] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. 2024. CoFInAl: enhancing action quality assessment with coarse-to-fine instruction alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 196, 9 pages. https://doi.org/10.24963/ijcai.2024/196
[54] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. 2023. Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 12 (2023), 7749–7763.
[55] Kanglei Zhou, Yue Ma, Hubert P. H. Shum, and Xiaohui Liang. 2023. Hierarchical Graph Convolutional Networks for Action Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 12 (2023), 7749–7763. https://doi.org/10.1109/TCSVT.2023.3281413
[56] Yong Zhou, Zeming Xie, Jiaqi Zhao, Wenliang Du, Rui Yao, and Abdulmotaleb El Saddik. 2024. Multi-Modal LiDAR Point Cloud Semantic Segmentation with Salience Refinement and Boundary Perception. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 10 (2024), 1–20.