The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos

Hazel Doughty Walterio Mayol-Cuevas Dima Damen
University of Bristol, Bristol, UK
<Firstname>.<Surname>@bristol.ac.uk

Abstract

We present a new model to determine relative skill from long videos, through learnable temporal attention modules. Skill determination is formulated as a ranking problem, making it suitable for common and generic tasks. However, for long videos, parts of the video are irrelevant for assessing skill, and there may be variability in the skill exhibited throughout a video. We therefore propose a method which assesses the relative overall level of skill in a long video by attending to its skill-relevant parts.

Our approach trains temporal attention modules, learned with only video-level supervision, using a novel rank-aware loss function. In addition to attending to task-relevant video parts, our proposed loss jointly trains two attention modules to separately attend to video parts which are indicative of higher (pros) and lower (cons) skill. We evaluate our approach on the EPIC-Skills dataset and additionally annotate a larger dataset from YouTube videos for skill determination with five previously unexplored tasks. Our method outperforms previous approaches and classic softmax attention on both datasets by over 4% pairwise accuracy, and as much as 12% on individual tasks. We also demonstrate our model's ability to attend to rank-aware parts of the video.

1. Introduction

Skill determination is the problem of assessing how well a subject performs a given task. Automatic skill assessment from video will enable us to explore the wealth of online videos capturing daily tasks, such as crafts and cooking, for training humans and intelligent agents - *which video should a robot imitate to prepare you scrambled eggs for breakfast?*

For long videos, previous approaches make a naive assumption; the same level of skill is exhibited throughout the video, and thus skill can be determined in any (or all) of its parts [7, 23, 29, 38, 40]. Take for example the task of 'tying a tie'; draping the tie around the neck or straightening the tie may be uninformative when determining a subject's skill, however the way the subject crosses one side over and pushes the tie into the loop are key. Additionally, there may be variation in skill across the video: when comparing two videos, one subject may perform better at neatly crossing the tie but worse at pulling through the loop.

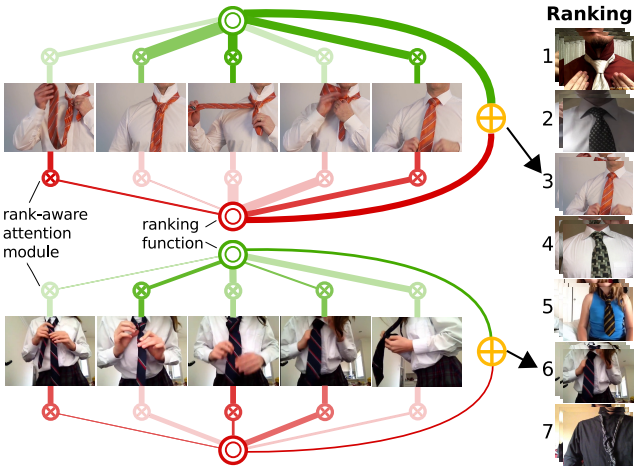

Figure 1. Rank-aware attention for skill ranking. We determine a video's rank by using high (green) and low (red) skill attention modules, which determine each segment's influence to the rank. Both modules are fused (orange) for an overall skill assessment of the video. Line opacity indicates the attention value for a segment and the line thickness indicates the score.

Accordingly, we consider skill determination to be a fine-grained video understanding problem, where it is important to first localize relevant temporal regions to distinguish between instances [25]. We target skill determination for common tasks, where ranking videos [2, 7, 21] is more suitable than estimating an objective score [23, 27, 40]. For many tasks, objective scores would be hard to articulate or find expert bodies to certify. Instead, crowd-sourcing can obtain a ranking on any task, which is consistent through consensus of judgment. Therefore, we devise a Siamese CNN over temporal segments, including attention modules adapted from [22], which we train to be rank-aware using a novel loss function. This is because relevance may differ depending on the skill displayed in the video - e.g. mistakes may not appear in higher-ranked videos. When trained with

our proposed loss, these modules specialize to separately attend to parts of the video informative for high skill or sub-standard performance (see Fig. 1).

While temporal attention has previously been used to indicate relevance in long videos [22, 25], no prior work has proposed to learn rank-aware temporal attention. Our **main contribution** is that we address the challenges of fine-grained video ranking by demonstrating the need for rank-aware temporal attention and propose a model to learn this effectively. We additionally contribute a new skill determination dataset, by collecting and annotating 5 tasks from YouTube, each containing 100 videos. In total, our dataset is 26 hours of video, twice the size of existing skill determination datasets, with videos up to 10 minutes in length. We outperform our previous effort as well as alternative attention baselines on EPIC Skills [7] and our newly collected dataset, BEST, and present a comprehensive evaluation of the contribution of rank-aware attention.

The rest of the paper is organized as follows. Section 2 reviews the related work. We introduce our proposed method in Section 3 and our new dataset in Section 4. Section 5 presents quantitative and qualitative results of our method, followed by the conclusion in Section 6.

2. Related Work

In this section, we first review skill determination works in video, both task-specific and widely applicable methods. We then review works proposing attention modules, specifically temporal attention, for a variety of problems.

Skill Determination. Several seminal works attempted skill determination in video [13, 14, 37]. Gordon [13] was the first to explore the viability of automated skill assessment from videos, as well as identifying appropriate tasks for analysis, with a case study on skill assessment of gymnastic vaults from skeleton trajectories. Despite the importance of automatic skill assessment from video for training and guidance [5, 1], following works remain limited [2, 7, 23, 27, 29, 35, 38, 40, 41]. These works demonstrate good performance by focusing on features specific to the task, such as skeleton trajectory in diving [27] or entropy between repeated sutures in surgery [40]. Parallel efforts instead perform skill determination from non-visual sensors such as inertial measurement units [8, 9, 21, 33, 39].

Several datasets have been introduced in prior work [7, 11, 23, 27, 35]. MIT Dive [27] and UNLV datasets [23] only include short video clips ($<$ 5s), whilst the remaining [11, 7, 27] are small scale datasets. Fis-V [35] contains 500 figure skating videos, however this is not publicly available. We test on our previous dataset, EPIC-Skills [7], as this includes the JIGSAWS [11] dataset re-annotated for ranking alongside 3 other tasks. We also present a new dataset for skill assessment from longer videos (avg length = 188s), consisting of 500 videos across 5 daily-living tasks.

To assess skill in long videos, different approaches have been proposed. One is to first localize pre-selected events specific to the task [2], such as shooting or passing the ball in a basketball game. Alternatively, global features from the entire video have been used [27, 29, 38, 40], such as skeleton trajectories [27], features averaged across the video [23], or from randomly sampled segments in our previous work [7]. The only work to use attention in long videos is [35] for figure-skating. They use a self-attentive LSTM and a multi-scale skip LSTM to learn local (technical movements) and global (performance of players) scores respectively. This method uses a regression framework specifically for predicting the components of figure skating scores, not appropriate for common tasks.

We differ from all previous works in that we train a model to attend to skill-relevant parts of a video; learnable thus applicable to any task. We use a convolutional network with temporal segments and propose a novel *rank-aware* loss function. We do not use LSTMs due to the reported issues with maintaining information over longer videos [30, 32], and inferior performance compared to non-recurrent networks in many sequence-based tasks [3, 12, 32].

Attention Modules. Attention is increasingly used in fine-grained recognition, as intelligently weighting input is key to distinguishing between similar categories. This is a common problem in image recognition [10, 31] where attention can localize discriminative attributes in the object of interest. For instance, Fu et al. [10] present RA-CNN to recursively zoom into the most discriminative image region with an inter-scale ranking loss. Singh et al. [31] adapt the spatial transformer network [15] into a Siamese network to perform relevant attribute ranking. Similarly, in person re-identification from video, attention [16, 19, 34] is utilized to select the frames with the best view of identifying attributes.

Attention has also been adopted in the video domain for action recognition [25, 26] and localization [17, 28, 22, 24], including for weakly supervised localization from video-level labels [22, 24]. Pei et al. [25] combine an attention module with a gated recurrent network to classify actions in untrimmed video. Piergeovanni et al. [26] present temporal attention filters to discover latent sub-events in activities. Nguyen et al. [22] use attention filters within a CNN to identify a sparse set of video segments which minimize a video's classification loss. They use this in combination with class-specific attention from the activations to localize target actions. We build on the class-agnostic attention filters used in this work for our rank-aware attention (Sec 3.3).

Using class-specific attention is a common technique in existing temporal attention works [22, 24]. In this work, we propose the first model to train rank-specific (which we call rank-aware) attention, and demonstrate that it outperforms rank-agnostic attention and existing methods.

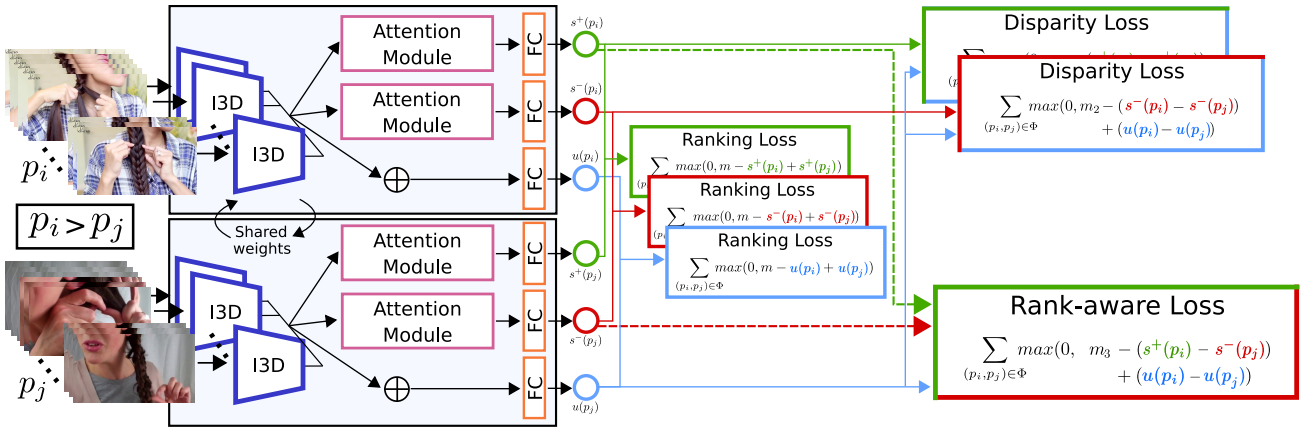

Figure 2. Rank-Aware Attention Network. Given a ranked pair of videos (p_i, p_j) where p_i exhibits higher skill: each video is uniformly split into segments. Extracted features (I3D) are passed into a pair of attention modules to produce video-level representations for the ranking functions (FC layers). Each ranking function produces a score s^+ (green) or s^- (red). Additionally, a uniformly weighted video representation produces a third ranking score u (blue). Three types of losses are defined: the ranking loss maximizes the margin (green-to-green, red-to-red, blue-to-blue) between the pair of ranked videos, the disparity loss ensures attention branches outperform uniform (green-to-blue, red-to-blue) and the final loss optimizes the attention modules to become rank-aware (green-to-red).

3. Rank-Aware Attention Network

In this section, we re-formulate the skill determination problem in long videos. We then detail the combination of training losses used to achieve rank-aware attention.

3.1. Problem Formulation

We propose a pairwise ranking supervised learning approach for skill determination. In this setup the training set comprises of all pairs of videos, P, where each pair $(p_i, p_j) \in P$, has been annotated such that video p_i displays more skill than p_j. Such pairwise annotations can be acquired for any task using crowd-sourcing (see Sec. 4). The aim is then to learn a ranking function $f(\cdot)$ for an individual task such that

$$f(p_i) > f(p_j) \qquad \forall (p_i, p_j) \in P \tag{1}$$

For long videos, **previously** we assumed these pairwise skill annotations can be propagated to any part of the video [7]. Given p_{it} is the t^{th} video segment, $t \in [0, T)$, skill annotations were propagated so that,

$$f(p_{it}) > f(p_{jt}) \qquad \forall t \in [0, T); (p_i, p_j) \in P \tag{2}$$

Another approach to deal with long videos [23, 36], is to use a uniform weighting of feature vectors to learn a video level ranking. This assumes all parts of the video are equally important for skill assessment, i.e. $u(p_i) > u(p_j)$ where,

$$u(p_i) = f(\frac{1}{T}\sum_t p_{it}) \tag{3}$$

In this work, we believe these assumptions do not hold. First, some parts of the video may not exhibit any difference in skill, or may even show reversed ranking - where the overall better video has segments exhibiting less skill. Second, non-uniform pooling should better represent the video's overall skill by increasing the weight for segments more pertinent to a subject's skill. Third, comparing corresponding video chunks (p_{it}, p_{jt}) assumes tasks are performed in a set order, at the same speed. We deviate from these assumptions, and instead aim to jointly learn temporal attention $\alpha(\cdot)$, alongside ranking function $r(\cdot)$ such that

$$s(p_i) > s(p_j); \qquad s(p_i) = r(\sum_t \alpha(p_{it}) p_{it}) \tag{4}$$

While $\alpha(\cdot)$ is a standard attention module for relevance, we observe that the segments most crucial to determining skill may differ depending on the subject's skill; a low-skill subject may perform certain actions (e.g. mistakes) not performed by a high-skill subject and vice-versa. Therefore, we propose to train two general attention modules to produce scores s^+, s^-, for all pairs $(p_i, p_j) \in P$, such that:

$$s^+(p_i) > s^+(p_j); \quad s^-(p_i) > s^-(p_j); \quad s^+(p_i) \gg s^-(p_j) \tag{5}$$

In particular, $s^+(p_i) \gg s^-(p_j)$, encourages the two attention modules to diverge, such that one attends to segments which display a high skill (α^+) and the other to low skill (α^-), along with differing ranking functions g, h:

$$s^+(p_i) = g(\sum_t \alpha^+(p_{it}) p_{it}) \tag{6}$$

$$s^-(p_i) = h(\sum_t \alpha^-(p_{it}) p_{it}) \tag{7}$$

3.2. Rank-Aware Attention and Overall Network

We show our overall architecture in Fig. 2. The Siamese network takes a video pair (p_i, p_j) and splits each into T

segments of uniform length. The features from all segments $\{p_{it}\}$ are then passed to three branches. Within each branch, we first obtain a video level representation from all segments either weighted by our learned attention functions $\alpha^+(\cdot)$ and $\alpha^-(\cdot)$ (Sec. 3.3), or through uniform weighting $\frac{1}{T}\sum_t^T p_{it}$. Three ranking functions are then learned (one per branch) $g(\cdot), h(\cdot)$ and $f(\cdot)$ with a fully connected (FC) layer to produce corresponding scores per video s^+ (Eq. 6), s^- (Eq. 7) and u (Eq. 3). The FC layers are separate for each weighting function, but shared by both sides of the Siamese network. These scores are then evaluated by different loss types: ranking loss, disparity loss and rank-aware loss, each of which is explained below.

For each branch, a margin **ranking loss** function ensures p_i is ranked higher that p_j,

$$L^+_{rank} = \sum_{(p_i,p_j)\in P} max(0, m - s^+(p_i) + s^+(p_j)) \quad (8)$$

where $s^+(p_i)$ is the final score of video p_i from the high-skill attention module and m is a constant margin. The ranking loss is defined similarly for the low-skill and uniform weighting branches:

$$L^-_{rank} = \sum_{(p_i,p_j)\in P} max(0, m - s^-(p_i) + s^-(p_j)) \quad (9)$$

$$L^u_{rank} = \sum_{(p_i,p_j)\in P} max(0, m - u(p_i) + u(p_j)) \quad (10)$$

While the need for uniform weighting may not be obvious, we empirically noted that ranking using the attention module frequently falls into local-minima during training. The learned attention weights for such a local-minimum perform worse than uniform weighting. We avoid this by introducing an attention **disparity loss**, which explicitly encourages an attention branch to outperform uniform:

$$L^+_{disp} = \sum_{(p_i,p_j)\in P} max(0, m_2 - (s^+(p_i) - s^+(p_j)) + (u(p_i) - u(p_j))) \quad (11)$$

Here, m_2 is a separate margin from m specific to this loss. For a video pair (p_i, p_j), this loss encourages the difference between scores $(s^+(p_i), s^+(p_j))$ to be greater than the difference between scores $(u(p_i), u(p_j))$, thereby encouraging the attention module to produce video-level representations better at distinguishing between the skill displayed in the two videos than uniform weighting. This loss alone could instead cause the performance of $f(\cdot)$ to degrade, however by jointly optimizing with Eq. 10 this is avoided. An analogous loss L^-_{disp} is defined for the low-skill branch.

Using the loss functions defined so far, the two learned attention modules $\alpha^+(\cdot), \alpha^-(\cdot)$ are indistinguishable. They attend to skill-relevant segments to form video-level representations and $g(\cdot)$ and $h(\cdot)$ perform the ranking. We finally optimize these filters to achieve the desired response with our proposed **rank-aware loss**:

$$L_{rAware} = \sum_{(p_i,p_j)\in P} max(0, m_3 - (s^+(p_i) - s^-(p_j)) + (u(p_i) - u(p_j))) \quad (12)$$

With Eq. 12, we ensure s^+ attends to higher skill parts of the better video p_i while s^- attends to video parts with lower skill from p_j. To optimize for rank-aware attention, we use a larger margin m_3 compared to single branches m_2. The overall training is then conducted by combining the losses:

$$L_R = \sum_{i=\{+,-,u\}} L^i_{rank} + \sum_{i=\{+,-\}} L^i_{disp} + L_{rAware} \quad (13)$$

As training iterates through pairs in P, the same video will be considered higher skill in one pair and lower in another (e.g. $(p_i, p_j) \in P, (p_j, p_k) \in P$). The network accordingly optimizes the *shared weights* so as to learn rank-aware attention modules.

When **testing** the network, a single video is evaluated and its rank is assigned through its ranking score:

$$R(p_i) = s^+(p_i) + s^-(p_i) \quad (14)$$

Note that in training we learn $s^+(\cdot)$ and $s^-(\cdot)$ such that $s^+(p_i) > s^+(p_j)$ and $s^-(p_i) > s^-(p_j)$ which implies $s^+(p_i) + s^-(p_i) > s^+(p_j) + s^-(p_j)$. Although $\alpha^-(\cdot)$ attends to low-skill segments, the overall score s^- reflects the correct ranking of the videos. We do not include $u(p_i)$ as the attention alone should be sufficient (shown in Fig. 5).

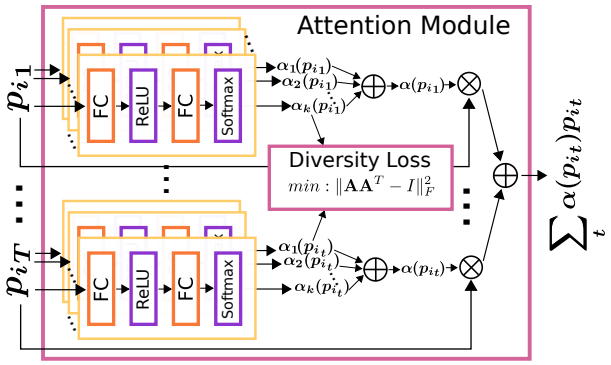

Figure 3. The attention module consists of K attention filters, each outputting a scalar weight per segment, used to produce the weighted video-level feature.

3.3. Multi-filter Attention Module

Our attention modules $\alpha^+(\cdot)$ and $\alpha^-(\cdot)$ each take a set of T video segments and learn a weighting of these segments informative for skill ranking. As the attention modules have the same structure, we will refer to the generic attention module $\alpha(\cdot)$ for simplicity. We show the architecture of the attention module in Fig. 3. The attention module consists of K filters, each comprised of two FC layers,

the first followed by a ReLU activation function, the second followed by a softmax. This is based on the attention filter used in [22] with a softmax activation instead of sigmoid. Filters are combined to achieve segment level attention:

$$\alpha(p_{it}) = \sum_{k=1}^{K} \alpha_k(p_{it}) \tag{15}$$

where α_k refers to the kth attention filter for the attention module $\alpha(\cdot)$, and importantly $\sum_{t=1}^{T} \alpha_k(p_{it}) = 1$ for each of the K filters. We include multiple attention filters to encourage a module to attend to multiple skill-relevant subtasks in the long videos; a single filter typically focuses on only one element of the task [20]. To regularize the K filters, we use a diversity loss. We define the K x T attention matrix relating to video p_i as:

$$\mathbf{A}_i = \begin{bmatrix} \alpha_1(p_{i1}) & \alpha_1(p_{i2}) & \dots & \alpha_1(p_{it}) \\ \alpha_2(p_{i1}) & \alpha_2(p_{i2}) & \dots & \alpha_2(p_{it}) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_k(p_{i1}) & \alpha_k(p_{i2}) & \dots & \alpha_k(p_{it}) \end{bmatrix} \tag{16}$$

and use the following **diversity loss**:

$$L_{div} = \sum_{(p_i,p_j)\in P} \|\mathbf{A}_i\mathbf{A}_i^T - \mathbf{I}\|_F^2 + \|\mathbf{A}_j\mathbf{A}_j^T - \mathbf{I}\|_F^2 \tag{17}$$

where $\mathbf{I}$ is the identity matrix and $\|\cdot\|_F^2$ denotes the Frobenius norm. Similar losses have been used successfully in other applications, such as text embedding [18] - here we use it to regularize temporal attention in video. In our network, this loss encourages each filter to learn a different aspect of the video. Without such a loss, all filters attend to the same most discriminative part in the video, rendering more than one filter redundant. This loss also encourages filters to be sparse and pick the few most informative segments. We assess the effect of multiple filters in Section 5.

Note that the diversity loss is within an attention module; diversity is not enforced between modules. Attentions are allowed to overlap and do so when the segment is relevant for different skill levels. Our overall training loss is:

$$L_R = \sum_{i=\{+,-,u\}} L^i_{rank} + \lambda \sum_{i=\{+,-\}} L^i_{div} + \sum_{i=\{+,-\}} L^i_{disp} + L_{rAware} \tag{18}$$

4. Tasks and Datasets

We evaluate our model on our previous dataset, EPIC-Skills [7]. It consists of four distinct tasks: surgery (knot-tying, needle passing, and suturing) from [11], dough-rolling from [6] as well as self-recorded drawing (two drawings) and chopstick-using. Every (sub-)task consists of up to 40 videos, with pairwise annotations indicating the ranking of videos in a pair. A limitation of this dataset is that each task is collected in a single environment with the same perspective and only minor variations in the background. We therefore collect and annotate a new skill determination dataset over twice as large, from online videos and thus with a variety of individuals, environments, and viewpoints.

	Task	#Videos	#Pairs	%Pairs	Av. Length (s)
EPIC-Skills	Chopstick Using	40	536	69%	46 ± 17
	Dough Rolling	33	181	34%	102 ± 29
	Drawing	40	247	65%	101 ± 47
	Surgery	103	1659	95%	92 ± 41
BEST	Scramble Eggs	100	2112	43%	170 ± 113
	Tie Tie	100	3843	77%	81 ± 47
	Apply Eyeliner	100	3743	76%	122 ± 105
	Braid Hair	100	3847	78%	179 ± 91
	Origami	100	3237	65%	386 ± 193

Table 1. Comparing EPIC-Skills with BEST: #videos, #of pairs and average and standard deviation of video length.

4.1. BEST Dataset

We collect and annotate the Bristol Everyday Skill Tasks (BEST) 2019 dataset consisting of five skill tasks with 100 videos per task, publicly available[1]. This dataset gives us an opportunity to test on a larger variety of skill tasks with more and longer videos per task from varied environments.

Video Collection. We selected five tasks which can be completed using various methods and may be challenging for novices: scrambling eggs, braiding hair, tying a tie, making an origami crane and applying eyeliner. The tasks selected are deliberately varied in their content and also differ from the tasks in EPIC-Skills as this allows a more thorough testing of the proposed model.

To obtain 100 videos per task, we first retrieve the top-400 videos from YouTube using the task name as a query. We then ask AMT workers to answer questions about each video to determine its suitability for our dataset. These ensure the selected videos contain the relevant task, are good quality videos, contain a clear view of the task and the complete performance of the task with minimal edits. We also ask AMT workers for their initial opinion of the skill of the person performing the task: 'Beginner', 'Intermediate' or 'Expert'. This initial labelling ensures we select sufficient beginner videos before pairwise annotations.

As only a portion of the YouTube video may contain the desired task, we annotate the start and end of the relevant activity via AMT, using the same approach for annotations from [4]. We use the agreement of 4 workers.

Pairwise Annotation. As in [7], we ask AMT workers to watch videos in a pair simultaneously and select the video

[1] https://github.com/hazeld/rank-aware-attention-network

which displays more skill. The pair is taken as ground-truth only if all four workers agree on a pair's ordering. It is unnecessary to annotate all possible pairs. Instead, we annotate 40% of the possible pairings, where each video appears in an equal number of pairs. We remove the need for exhaustive annotation by utilizing the transitive nature of skill ranking to obtain pairs outside of the original 40%. We then perform a second round of annotations for pairs of a similar rank, to ensure our dataset contains challenging pairs.

The number and percentage of pairs per task is shown in Table 1, along with the average video length per task. Our dataset is considerably larger than our previous effort EPIC-Skills in terms of both videos and annotated pairs.

5. Experiments

We first describe the implementation details of our network. We then present results on the two datasets alongside baselines and analyze the contribution of the various components in our method with an ablation study.

5.1. Implementation Details

We uniformly sample 400 stacks of 16 frames, at 10fps, for each video. Images are re-scaled to have a height of 256 pixels then centre cropped to 224×224. We extract features using I3D, pre-trained on Kinetics [3]. To prevent overfitting we augment the features by adding noise $\mathcal{N}(0, 0.01^2)$ per dimension as in [22]. All models are trained using the Adam optimizer with a batch size of 128 and learning rate of 10^{-4} for 2000 epochs. For stable training, we iteratively optimize the network's parameters. We first fix the attention module parameters and optimize the ranking FC layer weights using L_{rank} losses (Eq 8, 9, 10). We then fix the ranking FC layer weights and optimize the attention module weights, using the remaining losses (L_{div}, L_{disp} and L_{rAware}). In all experiments, we set the weight of λ (Eq. 18) to 0.1, $m_1 = 1$ (Eq. 8), $m_2 = 0.1$ (Eq. 11) and $m_3 = 0.3$ (Eq. 12).

5.2. Quantitative Results

Evaluation Metric We evaluate tasks individually and report pairwise accuracy (% of correctly ordered pairs) and mean task accuracy for each dataset. For EPIC-Skills we use the four-fold cross validation training and test splits provided with the dataset [7]. For BEST we use a single 75%:25% split per task (provided with release), as the number of pairs is larger. Our test set consists exclusively of pairs where neither video is present in the training set.

Baselines and Attention. In Table 2 we show the results of our method in comparison with different baselines.

We outperform our previous work [7] by 4.3% and 5.4% on EPIC-Skills and BEST respectively. We also use four baselines for various temporal attention approaches. The first temporal attention baseline selects only the **last segment** of the video as skill-relevant. It could be argued that this segment, displaying the final outcome of the task, is sufficiently informative to attend to across tasks, however this performs particularly poorly on BEST. We also use **uniform weighting** and **softmax attention** as temporal attention baselines. For softmax attention we use our method with a single attention branch only optimized by L_{rank}. Importantly, our proposed method shows an improvement over both uniform weighting and standard softmax attention, particularly for BEST with longer videos. Interestingly, we see the inclusion of softmax attention decreases the accuracy for both datasets from a naive uniform weighting of segments (-4.3% and -0.7%). Although softmax attention achieves higher accuracy than uniform for several tasks, we found softmax attention to be highly inconsistent. To compare to existing temporal attention methods, we adapt the class agnostic attention from Sparse Temporal Pooling Network (STPN) [22] into a pairwise ranking framework. While this method works well for action localization, in a ranking framework it performs worse than both our method and uniform sampling.

Method	EPIC Skills	BEST
Who's Better [7]	76.0	75.8
Last Segment	76.8	61.0
Uniform Weighting	78.8	73.6
Softmax Attention	74.5	72.3
STPN [22]	74.3	70.0
Ours (Rank Aware Attention)	**80.3**	**81.2**

Table 2. Results of our method in comparison to baseline. Our final method outperforms every baseline on both datasets.

In general the baselines struggle on BEST as they are affected by the lengthy videos and increase in irrelevant parts, while last segment is affected by variations in environment and viewpoint. By focusing on key segments indicative of skill, our method is able to combat these difficulties and gain a larger increase on this dataset.

Ablation Study. In Fig. 4 we perform a per-task ablation study, testing the individual contributions of the components of our loss function (Eq. 13). The inclusion of the diversity loss increases the result by 2% for both datasets. It is particularly useful for Drawing (+7.3%) and Tie Tie (+6%), as videos in these tasks consistently have many skill-relevant segments.

From Fig. 4 we see training the attention module alongside the uniform weighting with the disparity loss improves the results further. L_{disp} encourages the network to learn attention better at discriminating between videos than the uniform weighting and decreases the sensitivity to initial-

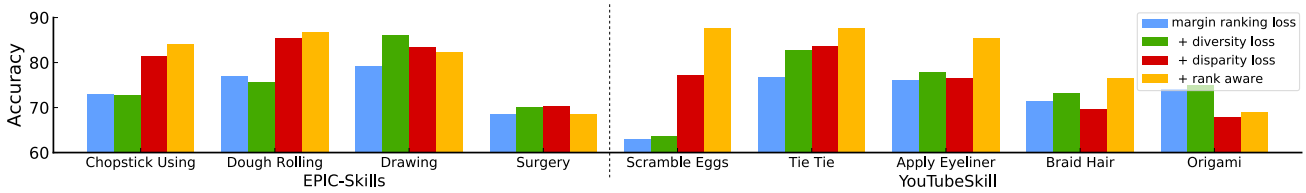

Figure 4. Ablation study of loss functions on all tasks. In general each additional loss term gives an improvement, the most significant improvement being the rank-aware loss which gives an average 5% improvement for BEST.

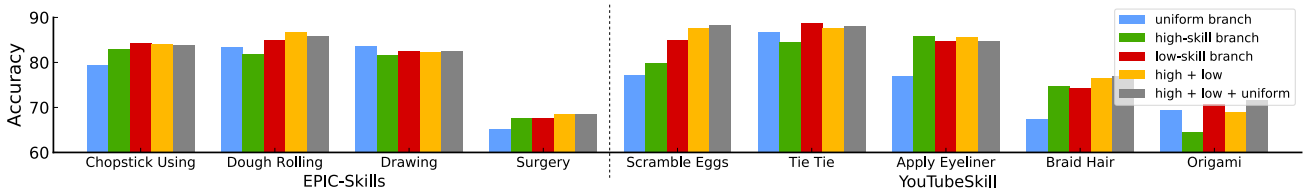

Figure 5. Contribution of different branches in the network. The addition of L^{+}_{disp} and L^{-}_{disp} cause both the high and low skill branches to perform better than uniform in most tasks. These branches offer complementary information causing an improvement in our final result.

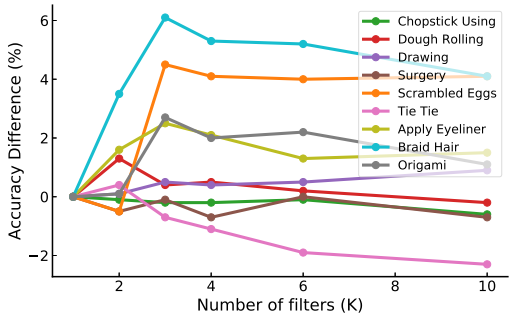

Figure 6. We test the number of filters (K) for all tasks. The number of filters causes a clear increase in many tasks, with the majority of tasks peaking at $K = 3$

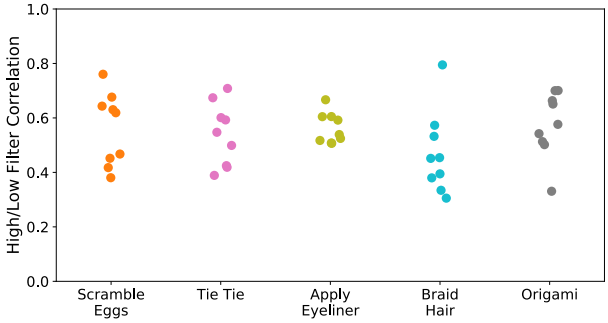

Figure 7. We test correlation of high and low skill filters for all tasks, to check they attend to different video segments.

ization. In tasks like Chopstick Using and Scramble Eggs, where attention optimized with only the ranking loss performs similarly to uniform, this can help significantly.

Our final rank-aware loss further improves the results, particularly for BEST (average improvement of 5%). This is especially true for Scramble Eggs and Apply Eyeliner (+10.4% and +8.8% respectively). These tasks contain more instances of subtasks specific to subjects with higher or lower skill, as can be seen in Section 5.3.

We note three exceptions to this trend: Drawing, Surgery and Origami. Surgery maintains a similar score throughout the ablation test and has the lowest final score of all tasks. We believe this is due to the I3D features not being able to capture the difference between the fine-grained detail of surgical motions of different abilities. Drawing and Origami both drop with the addition of L_{disp}. In Drawing the attention branch struggles to be better at separating videos than the uniform branch, indicating most segments are relevant for determining skill. In Origami, the uniform weighting has poor performance due to the visual subtlety of placing neat folds in the paper. Therefore, optimizing the attention branch to be better than uniform does not improve training.

Branch Contribution. Having trained our model with the overall loss, we now assess skill ranking using single or multiple branch scores. From Fig. 5 we see we are able to learn high and low skill branches which are both more informative than uniform. This is particularly true for tasks such as Chopstick Using and Scramble Eggs which see little improvement with attention until the disparity loss is introduced (Fig. 4). Within tasks, the performance of high and low skill branches can vary. We can see this for Tie Tie, with the low-skill branch performing best (+4.3%). Here, the presence of hesitation in lower-ranked videos proves effective for skill ranking.

The fusion of high and low skill branches further improves the result (EPIC-Skills +2.9% and BEST +3.2%). In many tasks the branches offer complementary information, as each branch can attend to separate video segments, specific to either high or low skill (see Sec 5.3).

Number of Filters. In Fig. 6 we test the effect of K, the number of filters per attention module (Sec. 3.3). The previous sections report results using K=3. This shows a small improvement over one filter in the majority of tasks. However, with K>3 the accuracy does not increase further, as additional less-informative segments are included.

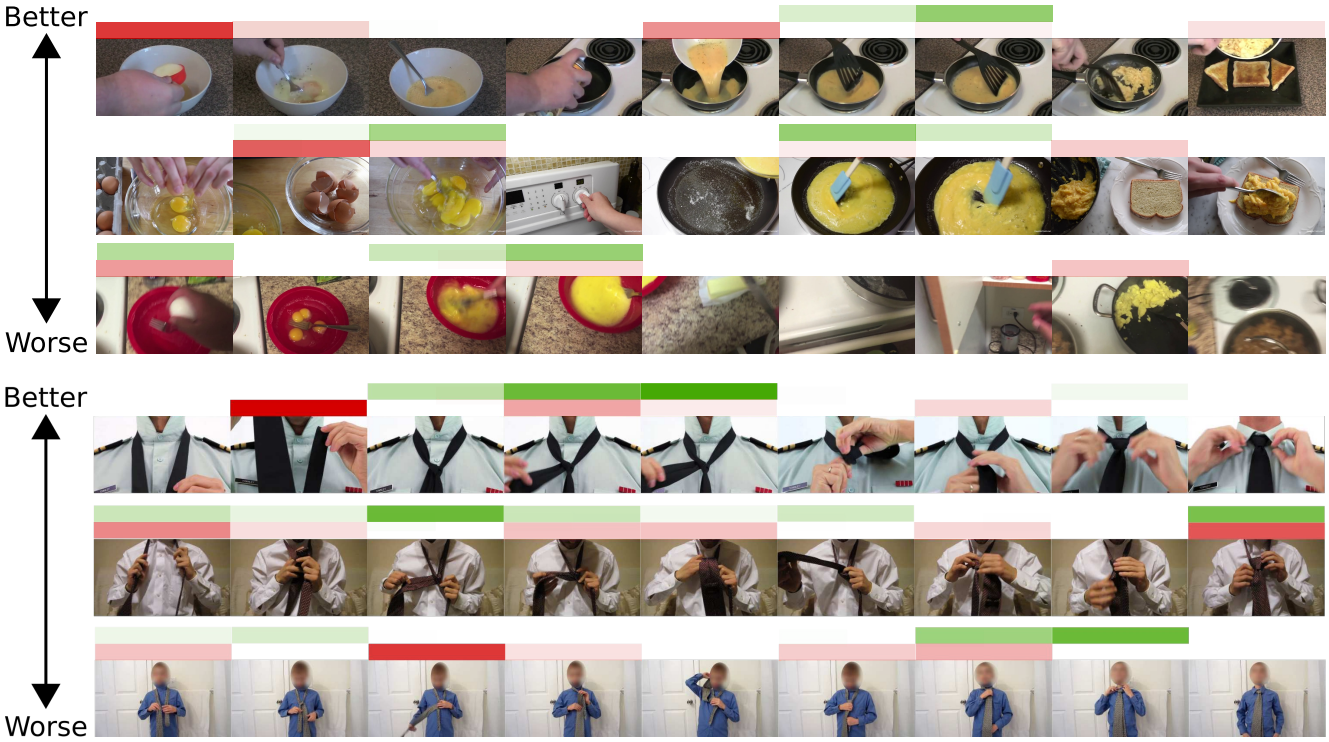

Figure 8. Attention values of the high-skill (green) and low-skill (red) modules with the corresponding video segments for examples from 'Scramble Eggs' and 'Tie Tie'. The intensity of the color indicates the attention value. We show the predicted ranking from both branches.

We also compared two rank-aware attention modules, with 3 filters each, to a single standard (i.e. rank-agnostic) module containing 6 attention filters. Results demonstrate a clear advantage of our rank-aware modules. For BEST, 81.2% accuracy drops to 75.0% without our novel loss.

Filter Correlation. To ensure our high and low skill filters are attending to different video segments we plot the correlation of pairs of filters between high and low attention modules, averaged over all videos for BEST. From Fig. 7 we can see most filter pairs have low correlation, demonstrating these are attending to different segments. There are some cases where filters have a higher correlation (Braid Hair at $\rho = 0.8$) as it can be helpful for at least one of the high and low skill filters to attend to the same segments when relevant at all levels of skill.

5.3. Qualitative Results

In Fig. 8 we show attention weights with corresponding frames for the Scramble Eggs and Tie Tie tasks. Firstly, the figure shows we are able to filter out irrelevant segments using attention, for instance turning on the stove-top and opening the cupboard in 'Scramble Eggs'. Secondly, we can see our rank-aware attention allows the modules to focus on different aspects of the video. In the Scramble Eggs task the high-skill module consistently focuses on whisking the eggs and stirring the mixture in the pan, while the low-skill module attends to adding milk/cream to the eggs and pouring. For 'Tie Tie' the high skill module gives a strong weighting to segments displaying a tight inner knot and straightening the tie before folding across, while the low-skill module focuses mainly on hesitation and repetition. We also observe cases where the filters attend to segments seemingly irrelevant to skill; in Scramble Eggs the low-skill module attends to segments containing bread. Video results are included in the supplementary material.

6. Conclusion

In this paper we have presented a new model for rank-aware attention, trained using a novel loss function. Our rank-aware loss enables us to learn the most informative segments to attend to in relation to the skill shown in the video. We also use the disparity loss to directly optimize the attention to pick more informative segments than the uniform distribution, solving the instability in optimizing the standard softmax attention in ranking. We have tested this method on two datasets, one of which we introduce in this paper, and show our method achieves state-of-the-art results for skill determination, with an average performance of over 80% in both datasets. Future work involves exploring applications of the attention segments to improve people's skill in a task, as well as transfer learning to unseen tasks.

Acknowledgements: Access to the BEST dataset and annotations available from authors' webpages. Supported by an EPSRC DTP and EPSRC GLANCE (EP/N013964/1).

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[2] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. Am I a Baller? Basketball Performance Assessment From First-Person Videos. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017. 1, 2

[3] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? a New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 2, 6

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*, September 2018. 5

[5] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio Mayol-Cuevas. You-Do, I-Learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *British Machine Vision Conference (BMVC)*, 2014. 2

[6] Fernando De la Torre, Jessica Hodgins, Javier Montano, Sergio Valcarcel, R Forcada, and J Macey. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database. *Robotics Institute, Carnegie Mellon University*, 2009. 5

[7] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 5, 6

[8] Mahtab J Fard, Sattar Ameri, R Darin Ellis, Ratna B Chinnam, Abhilash K Pandya, and Michael D Klein. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 14(1), 2018. 2

[9] Germain Forestier, François Petitjean, Pavel Senin, Fabien Despinoy, and Pierre Jannin. Discovering Discriminative and Interpretable Patterns for Surgical Motion Analysis. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 136–145. Springer, 2017. 2

[10] Jianlong Fu, Heliang Zheng, and Tao Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[11] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014. 2, 5

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional Sequence to Sequence Learning. *arXiv preprint arXiv:1705.03122*, 2017. 2

[13] Andrew S Gordon. Automated Video Assessment of Human Performance. In *Proceedings of AI-ED*, pages 16–19, 1995. 2

[14] Winfried Ilg, Johannes Mezger, and Martin Giese. Estimation of Skill Levels in Sports Based on Hierarchical Spatio-Temporal Correspondences. In *Joint Pattern Recognition Symposium*, pages 523–531. Springer, 2003. 2

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2017–2025, 2015. 2

[16] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

[17] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. VideoLSTM Convolves, Attends and Flows for Action Recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018. 2

[18] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A Structured Self-Attentive Sentence Embedding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 5

[19] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2017. 2

[20] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. Attention Clusters: Purely Attention Based Local Feature Integration for Video Vlassification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[21] Anand Malpani, S Swaroop Vedula, Chi Chiung Grace Chen, and Gregory D Hager. Pairwise Comparison-Based Objective Score for Automated Skill Assessment of Segments in a Surgical Task. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 138–147. Springer, 2014. 1, 2

[22] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 5, 6

[23] Paritosh Parmar and Brendan Tran Morris. Learning to Score Olympic Events. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 76–84. IEEE, 2017. 1, 2, 3

[24] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised Temporal Activity Localization and Classification. In *European Conference on Computer Vision (ECCV)*, September 2018. 2

[25] Wenjie Pei, Tadas Baltrusaitis, David MJ Tax, and Louis-Philippe Morency. Temporal attention-gated model for robust sequence classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. 1, 2

[26] AJ Piergiovanni, Chenyou Fan, and Michael S Ryoo. Learning Latent Sub-events in Activity Videos Using Temporal Attention Filters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 2

[27] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the Quality of Actions. In *European Conference on Computer Vision*, pages 556–571. Springer, 2014. 1, 2

[28] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action Recognition using Visual Attention. *arXiv preprint arXiv:1511.04119*, 2015. 2

[29] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video Based Assessment of OSATS Using Sequential Motion Textures. In *International Workshop on Modelling and Monitoring of Computer Assisted Interventions (M2CAI) workshop*, 2014. 1, 2

[30] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[31] Krishna Kumar Singh and Yong Jae Lee. End-to-End Localization and Ranking for Relative Attributes. In *European Conference on Computer Vision (ECCV)*, pages 753–769. Springer, 2016. 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017. 2

[33] Ziheng Wang and Ann Majewicz Fey. Deep Learning with Convolutional Neural Network for Objective Skill Evaluation in Robot-assisted Surgery. *arXiv preprint arXiv:1806.05796*, 2018. 2

[34] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Where-and-When to Look: Deep Siamese Attention Networks for Video-based Person Re-Identification. *arXiv preprint arXiv:1808.01911*, 2018. 2

[35] Chengming Xu, Yanwei Fu, Zitian Cheng, Bing Zhang, Yu-Gang Jiang, and Xiangyang Xue. Learning to Score Figure Skating Sport Videos. *arXiv preprint arXiv:1802.02774*, 2018. 2

[36] Ting Yao, Tao Mei, and Yong Rui. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[37] Qiang Zhang and Baoxin Li. Video-based Motion Expertise Analysis in Simulation-based Surgical Training Using Hierarchical Dirichlet Process Hidden Markov Model. In *Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval*, pages 19–24. ACM, 2011. 2

[38] Qiang Zhang and Baoxin Li. Relative Hidden Markov Models for Video-based Evaluation of Motion Skills in Surgical Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1206–1218, 2015. 1, 2

[39] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Mark A Clements, and Irfan Essa. Automated Assessment of Surgical Skills Using Frequency Analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 430–438. Springer, 2015. 2

[40] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, and Irfan Essa. Video and Accelerometer-Based Motion Analysis for Automated Surgical Skills Assessment. *International journal of computer assisted radiology and surgery*, 13(3):443–455, 2018. 1, 2

[41] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Thomas Ploetz, Mark A Clements, and Irfan Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery*, 11(9):1623–1636, 2016. 2