Semi-Supervised Action Quality Assessment with Self-Supervised Segment Feature Recovery

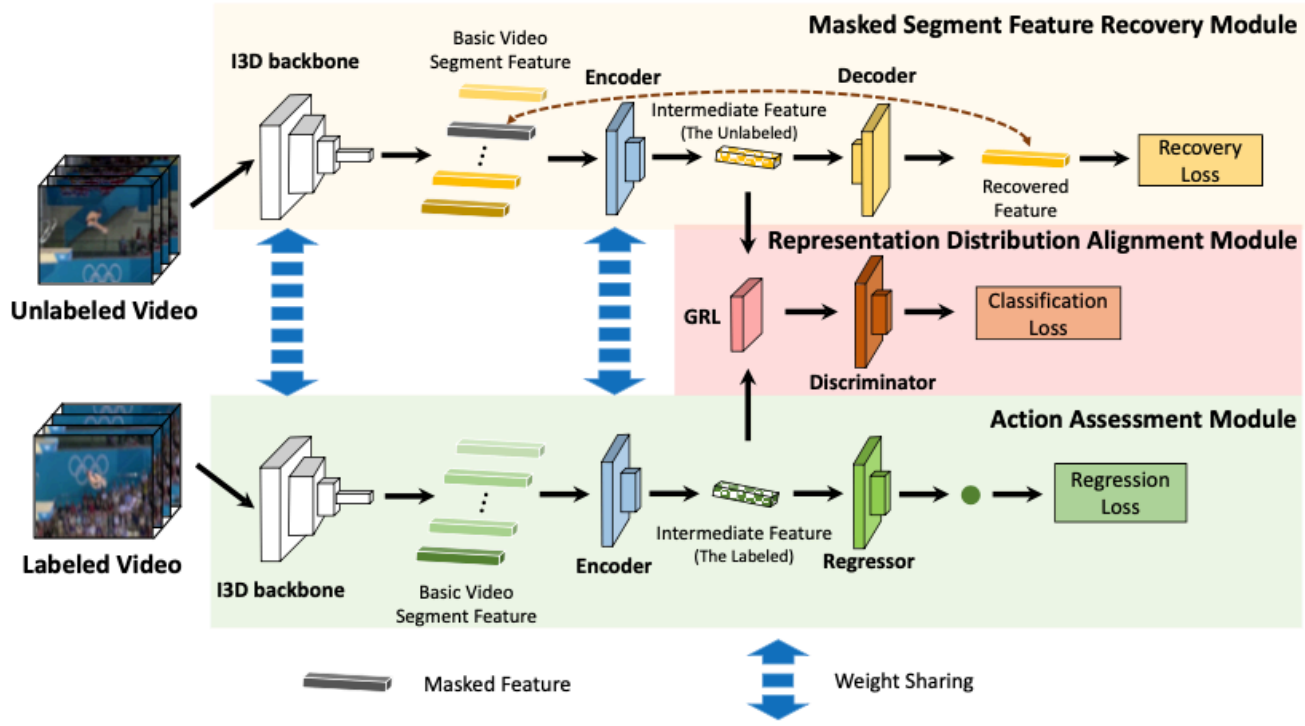
1、 Motivation

过往的AQA方法太依赖于多规模标注的数据集，却几乎没有工作关注半监督的方法。

从直觉上，一个动作的表现会收到每个stage的制约。比如说，如果运动员跳水跳出了很大的水花，那么肯定是前面出现了失误。而如果他前面一段表现很好，中间一段未知，而最终有很大的水花，那么他中间肯定出现了失误。而预测这种中间动作的表现有助于模型在无标签的情况下学会对动作质量进行建模。

因此，作者的方法是首先用一个encoder和decoder预测被mask住的feature来让模型学会动作质量表示。然后用有标签数据学会如何回归分数。最后用一个对抗学习模块将有标签和无标签数据的特征分布拉近。

2、Approach



2.1 problem formulation

Our goal is to learn to assess the quality of the action performance by exploiting both the labeled and the unlabeled samples.

两类数据的特征可以表示为：$V^l = [v_1^l, v_2^l, \ldots, v_t^l, \ldots, v_T^l]$和$V^u = [v_1^u, v_2^u, \ldots, v_t^u, \ldots, v_T^u]$。

2.2 Learning Action assessment on labeled videos

对于有标签的数据，将clip特征送入一个encoder以后用fc进行回归得到分数。这跟过往的方法基本一致。

2.3 learning temporal dependencies on the unlabeled videos

对于无标签数据，作者用了自监督的方法进行学习。

对于无标签视频的特征$V^u = [v_1^u, v_2^u, \ldots, v_t^u, \ldots, v_T^u]$，从中随机选择第k个segment的feature，将其mask起来得到$\ddot{V}^u$。然后通过跟有监督数据共享权值的encoder得到中间特征$\ddot{f}^u$。在这之后，尝试用一个decoder还原被mask住的特征，并用Mean-Absolute-Error loss来控制模型的学习。(decoder被设计为双层带ReLU激活的FC)

$$\mathcal{L}_{rcvr} = \frac{1}{U}\sum_{u=1}^{U}|\Phi(\ddot{f}^u) - v_k^u|,$$

2.4 aligning the representation distributions

由于sampling bias和不同的损失设计，有标签和无标签数据通过encoder以后得到的特征可能会存在不对齐的现象。因此，作者引入了一个对抗训练的机制将两者对齐。

假设有标签数据得到的中间特征分布为为class1，无标签的为class0。用一个discriminator $D(.)$尝试对中间特征进行判别，而encoder尝试混淆discriminator。因此，这个min-max对抗优化问题可以写为：

$$\min_E \max_D \{\frac{1}{2U}\sum_{f^u}(\log(1 - D(f^u)) + \log(1 - D(\ddot{f}^u))) + \frac{1}{L}\sum_{f^l}\log(D(f^l))\},$$

在训练的时候，作者使用了梯度反转层来训练encoder和decoder。因此，最终的adversarial loss可以写成：

$$\mathcal{L}_{adv} = \sum_{f^i \in \{f^u, \ddot{f}^u, f^l\}} -[(1 - c_i)\log(1 - D(f^i)) + c_i \log(D(f^i))] \tag{7}$$

where c_i is the class label of a sample which is 1 for a labeled sample and 0 for an unlabeled sample. By optimizing the

综上，S^4AQA最终的损失可以写为：

$$\mathcal{L} = \mathcal{L}_{reg} + \lambda_1\mathcal{L}_{rcvr} + \lambda_2\mathcal{L}_{adv},$$

3、Experiment

3.1 comparison with SOTA

TABLE I

THE TEST SP.CORR ON THE MTL-AQA DATASET WITH 10%/40% LABELS OF TRAINING SET.THE METHODS MARKED WITH * ARE THOSE TRAINED WITHOUT USING AN END-TO-END TRAINING STRATEGY.

Method	10% of labeled data	40% of labeled data
SVR [9]	0.427	0.565
USDL* [8]	0.530	0.646
C3D-AVG-STL* [42]	0.561	0.632
C3D-AVG-MTL* [42]	0.584	0.656
COREG [32]	0.487	0.526
Pseudo-labels [35]	0.622	0.716
VAT [23]	0.635	0.724
S^4L [24]	0.621	0.721
S^4AQA (Ours)	**0.676**	**0.746**

TABLE II

THE TEST SP.CORR ON THE RHYTHMIC GYMNASTICS DATASET WITH 40% OF LABELS OF TRAINING SET. AVG DENOTES THE AVERAGE SPEARMAN'S RANK CORRELATION ACROSS MULTIPLE ACTION TYPES.

Method	Ball	Clubs	Hoop	Ribbon	Avg
SVR [9]	0.175	0.243	0.261	0.309	0.248
ACTION-NET [6]	0.196	**0.403**	0.319	0.305	0.308
COREG [32]	0.230	0.338	0.331	0.268	0.292
Pseudo-labels [35]	0.183	0.330	0.346	0.305	0.292
VAT [23]	0.208	0.355	0.345	0.292	0.301
S^4L [24]	0.209	0.325	0.324	0.290	0.288
S^4AQA (Ours)	**0.248**	0.388	**0.372**	**0.357**	**0.342**

TABLE III

THE TEST SP.CORR ON THE JIGSAWS DATASET WITH 50% OF LABELS OF TRAINING SET. AVG DENOTES THE AVERAGE SPEARMAN'S RANK CORRELATION ACROSS MULTIPLE ACTION TYPES.

Method	Suturing	Needle Passing	Knot Tying	Avg
USDL [8]	0.439	0.351	0.680	0.505
Pseudo-labels [35]	0.445	0.501	0.714	0.566
VAT [23]	0.524	0.526	0.749	0.612
S^4L [24]	0.455	0.529	0.730	0.585
S^4AQA (Ours)	**0.533**	**0.552**	**0.813**	**0.655**

COREG效果不好作者分析是因为他是非参数模型，比较难学。S4L效果表现不好是因为里面用了很多视频旋转的数据增强，这使得模型比较难学习

action-net里面有复杂的注意力机制和GCN，因此取得了不错的表现。而作者的模型仅用简单的设计就取得了很不错的效果。

3.2 Ablation study

主要验证mask recovery module和distribution alignment module的作用。

TABLE IV

THE ABLATION STUDY RESULTS ON THE MTL-AQA DATASET. THE TEST SP.CORR IS REPORTED ON THE MTL-AQA DATASET WITH 10%/40% LABELS. AA, $MSFR$ AND RDA REPRESENT OUR ACTION ASSESSMENT MODULE, OUR MASKED SEGMENT FEATURE RECOVERY MODULE AND OUR REPRESENTATION DISTRIBUTION ALIGNMENT MODULE, RESPECTIVELY.

Module	10% of labeled data	40% of labeled data
AA	0.618	0.703
AA+RDA	0.634	0.726
AA+MSFR	0.661	0.732
AA+MSFR+RDA	**0.676**	**0.746**

TABLE V

THE ABLATION STUDY RESULTS ON THE RHYTHMIC GYMNASTICS DATASET. THE TEST SP.CORR IS REPORTED ON THE RHYTHMIC GYMNASTICS DATASET WITH 40% OF LABELS.

Module	Ball	Clubs	Hoop	Ribbon	Avg
AA	0.192	0.322	0.327	0.289	0.283
AA+RDA	0.204	0.327	0.324	0.290	0.287
AA+MSFR	0.236	0.378	0.361	0.338	0.329
AA+MSFR+RDA	**0.248**	**0.388**	**0.372**	**0.357**	**0.342**

看起来MSFR模块应该是起到主要作用的，而RDA模块又可以进一步提升性能。其实可以发现，仅仅是AA+MSFR就已经打败了其他模型。

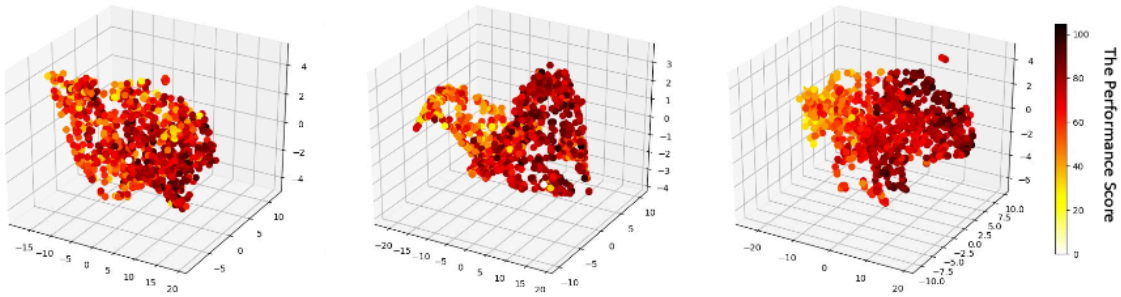


Fig. 4. The t-SNE visualization of representation distribution of unlabeled data on the MTL-AQA dateset with 10% of labels learned by AA, $AA+MSFR$ and $AA+MSFR+RDA$ models. The darker the color of the sample is, the higher its ground truth score will be. With the help of our proposed modules, the unlabeled videos with different scores are more uniformly scattered into different regions of representation space, making the representation of unlabeled videos more discriminative.

将无标签的数据根据分数可视化出来，可以发现随着模块的加入，特征点的分布越来越合理。

两个疑问：1）为什么AA+MSFR效果都已经很好了还想得到要加入RDA？2）可视化的图是怎么画出来的?

3.3 further analysis

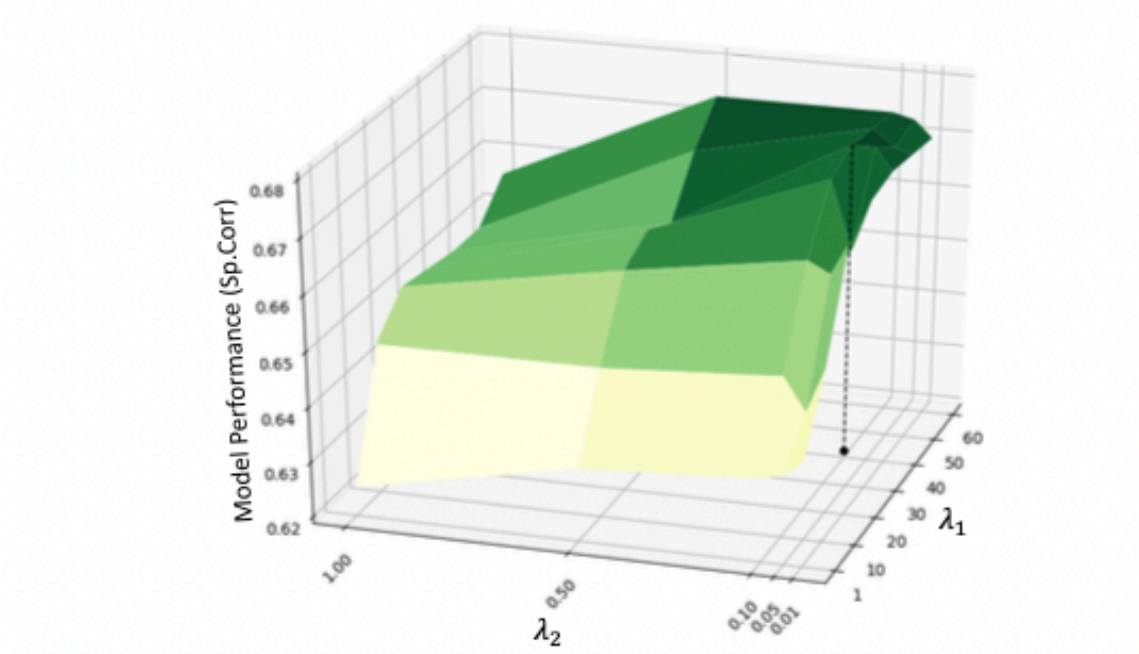


Fig. 5. The test Sp.Corr on the MTL-AQA dataset with 10% labels of training set under different weight combinations (λ_1, λ_2) of overall training loss.

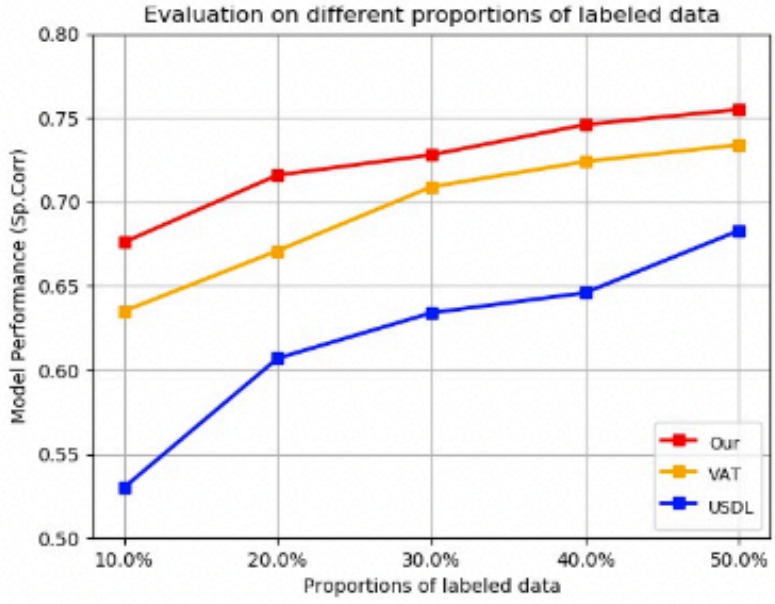


Fig. 6. The experimental comparison results of our model and other baselines under different proportion of labeled data on the MTL-AQA test set.

损失前面加的系数有没有什么直觉?