模範演技の点数に対してスケーリングと差分を用いた演技スポーツ映像自動採点モデルの評価

篠田　拓樹†　　青野　雅樹††

† 豊橋技術科学大学大学院工学研究科 情報・知能工学専攻　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
†† 豊橋技術科学大学 情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
E-mail: †{shinoda.hiroki.vo,masaki.aono}@tut.jp

あらまし　従来，体操のような演技を評価するスポーツでは，人間が演技と映像を見て採点を行ってきた．もし，演技スポーツで，リアルタイムに高精度な採点が行えれば，自動採点が実現できると考えられる．これまで，自動採点を目指した研究では，採点対象演技と模範演技を比較し，演技間の点数差をモデル化する方法が提案されてきた．しかし，演技間の点数差だけでは，差が大きい場合の採点が難しくなると考えられる．そこで，本研究では，採点対象演技映像と模範演技映像に，Attention 構造を用いて特徴比較を行い，得られた特徴量から模範演技点数にスケーリングと差分の演算を適用し，採点順位の高精度化を目指した．評価には，飛び込み競技データセットを用い，一部の評価指標で SOTA に匹敵する結果を得た．

キーワード　動画認識，スポーツ，自動採点，Action Quality Assesment，Attention

1　はじめに

近年，スポーツ分野では，審判の補助や選手の分析などに映像解析技術が用いられている．例として，テニスやバレーボールでは，コート内にボールが入っているかを判定するために，ハイスピードカメラによってボールを追跡する技術が用いられている．また，サッカーでは，専用のカメラでコート全体を撮影することで得られた映像から，選手の動きを追跡し，走行距離をはじめとしたトラッキングデータを取得している．

そのなかでも，従来，体操やフィギュアスケートなどの演技を評価するスポーツでは，数名の審判員が，生の演技と映像を見ることで，専門的な知識のもと，採点を行なってきた．もし，このような演技を評価するスポーツにおいて，リアルタイムに高精度な採点を行うことができれば，自動採点を実現することができると考えられる．そこで本研究では，演技映像を入力として演技の採点を行う手法を提案し、採点精度を高めることを目的とする。

演技スポーツの自動採点のように，特定のアクションを評価する研究分野として，Action Quality Assessment(AQA) と呼ばれる分野がある．AQA は，従来の映像認識や分類に比べて，難しいタスクである．例として，飛び込みの演技を挙げる．飛び込みの演技では,「踏切」,「技」,「入水」といった基本的な流れは同じだが,「踏切の仕方」や「回転」,「ひねり」,「入水時の水しぶき」といった細かな動作に違いがみられる．このような細かな動作を捉えるためには，映像全体の特徴を用いるだけでは難しく，演技や動作の部分ごとの特徴を用いる必要がある．そのため，AQA は主にスポーツ分野と医療分野で研究が進んでおり，その応用分野も広い．スポーツでは，体操や飛び込みなどの演技映像からスコアを回帰する研究 [1], [3], [6], [9], [12], [13], [14] が行われている．医療分野では，縫合をはじめとした医療行為に対して評価を行う研究 [4] が行なわれている．

特に，スポーツの自動採点を目的とした研究では，採点対象の演技と模範となる演技を比較し，演技間の点数差をモデル化する方法 [1], [12], [13] が提案されてきた．しかし，演技間の点数差だけでは，差が大きい場合の採点が難しくなると考えられる．そこで，本研究では，採点対象の演技映像と模範となる演技映像を比較し，模範となる演技の点数にスケーリングと差分の演算を適用するモデルを提案する．

また，従来研究 [12] では，演技を部分ごとに分けて表現を得るために，セグメンテーション構造を用いた手法が提案されている．従来研究 [12] では，映像の部分ごとに「1.5 Twist」や「Entry」などのアクションラベルが付与されているデータセットを用いている．しかし，AQA では，部分ごとのラベルがついていないデータセット [7] があり，そのようなラベルがないデータセットでの学習が行えないという問題点が挙げられている．そのため，本研究では，従来研究 [12] で用いられていたセグメンテーション構造の代わりに，Self-Attention 構造を導入し，映像の部分ごとの関連性を捉えることで，特徴づけを行う．

本研究では，提案モデルを評価するために飛び込み競技を取り上げ，飛び込みの演技映像を集めた AQA-7 データセット [7] を用いて性能評価実験を行った．結果として，一部の評価指標で SOTA に匹敵する結果を得た．

本研究の貢献をまとめ，以下に示す．

- スケーリングと差分を導入し，差が大きい場合の精度を高める．また，分割された演技映像に対して Self-Attention を適応する構造を含む Clip-level Attention を導入し、時間の切替をつかむ．
- 提案モデルを AQA-7 [7] データセットの飛び込み競技

映像を用いて評価し，SOTA に匹敵する結果を示した．
- 採点対象の演技と比較対象の演技の点数差が大きい場合に，一部の指標で，点数差分だけのモデルよりも，提案モデルの方が性能が高いことを示した．

2 先行研究

先行研究として，Action Quality Assessment(AQA) における近年の研究と，評価に用いられるデータセットを中心に述べる．なお，近年までの AQA における研究の全体像は，Wang et al. [11] にまとめられている．

2.1 Action Quality Assesment

AQA では，演技映像から様々な方法で特徴を抽出し，点数を回帰する方法が提案されてきた．Pan et al. [6] では，採点対象演技から関節に関する複数の特徴量を抽出し，点数の回帰に用いる手法が提案された．特徴量として，I3D [2] で抽出した映像全体の特徴量と関節の動きの特徴量，隣り合う間接の動きの特徴量，隣り合う間接の空間差分の特徴量，隣り合う関節の時間差分特徴量を抽出している．Farabi et al. [3] では，パラメータを増やさずに特徴量を集約する新たな手法が提案されている．ニューラルネットワークを用いて重みを計算し，特徴量の重み付き和を計算する Weight-Decider(WD) が提案されている．Tang et al. [9] では，人間の審判員が採点を行うときに含まれる不確実性を考慮するため，点数を確率分布で出力する，Uncertainty-aware Score Distributions Learning(USDL) が提案された．また，複数の異なる出力をする USDL を用いて採点を行う，Multi-path Uncertainty-aware Score Distributions Learning(MUSDL) も提案されている．Zhang et al. [14] では，モデルに点数の不確実性を含めるために，平均と分散を出力し，その値から得られる確率分布を用いて採点を行う Distribution AutoEncoder(DAE) を提案した．この手法では，学習方法の一部として，VAE [5] の Reparameterization Trick を用いている．また，USDL [9] や CoRe [13] と組み合わせたモデル (DAE-USDL, DAE-CoRe) が提案されている．

近年では，採点対象の演技映像のみを用いた手法ではなく，模範となる演技映像と比較する手法が用いられている．このような，演技映像同士を比較する手法は Contrastive Regression Framework と呼ばれ，Yu et al. [13] によって提案された．Yu et al. [13] では，採点対象の演技映像と模範となる演技映像を入力として，I3D [2] を用いて映像の特徴を抽出し，木構造の回帰器によって模範演技との点数差を回帰する Contrastive Regression(CoRe) が提案された．Xu et al. [12] は，Contrastive Regression Framework の考え方に Temporal Segmentation Attention(TSA) と呼ばれる構造を追加し，点数差を回帰するモデルを提案した．TSA では，演技映像を時間軸でセグメンテーションし，採点対象演技と模範演技のセグメンテーションされた特徴量を Transformer Decoder によって比較する Cross-Attention が行われる．Bai et al. [1] では，Contrastive Regression Framework の考え方を各映像のクリップごとの特徴に対して行い，点数差をクラスと実数値で回帰する Temporal Parsing Transformer(TPT) が提案された．また，最適化に，部分表現を差別化するための Sparcsity Loss と，クエリが異なる時間領域を持つための Ranking Loss を導入している．

2.2 AQA データセット

AQA では，主に 4 つのデータセット [4], [7], [8], [12] がモデルの性能評価に用いられている．AQA-7 [7] は，飛び込み，体操，スキービッグエア，スノーボードビッグエア，シンクロ飛び込み 3m，シンクロ飛び込み 10m，トランポリンの合計 7 種目の演技スポーツ映像を集めたデータセットである．映像は，7 種目合計で 1189 ビデオが含まれており，803 ビデオがトレーニングデータ，306 ビデオがテストデータとして提供されている．MTL-AQA [8] は，個人やシンクロ，男女，3m や 10m の高さなどの様々な飛び込み映像を集めたデータセットである．合計で 1412 ビデオが含まれており，1059 ビデオがトレーニングデータ，353 ビデオがテストデータとして提供されている．また，ラベルとして演技点数のほかに，演技のアクションクラスや難易度が付与されている．JIGSAWS [4] は，縫合，針を通す，糸を結ぶという 3 つの医療行為映像を集めたデータセットである．3 種類合計で 103 ビデオが含まれている．FineDiving [12] は，飛び込み映像を集めたデータセットである．合計で 3000 ビデオが含まれている．また，ラベルとして演技点数のほかに，52 クラスの演技アクション，29 クラスのサブアクション，23 個の難易度と非常に細かなラベルが付与されている．

3 提案手法

本章では，本研究で提案するモデルについて述べる．3.1 節では，提案モデルのベースとなるモデルと提案モデルの全体像について述べる．3.2 節では，提案モデルに含まれる Clip-level Attention について述べる．提案モデルの全体像を図 1 に示す．

3.1 モデルの全体像

演技スポーツの自動採点を目的とした研究 [1], [12], [13] では，採点対象演技映像 V_n と模範演技映像 V_m を用いて演技を比較し，演技間の点数差を回帰するモデルが提案されている．従来研究のモデルでは，まず，採点対象演技映像 V_n と模範演技映像 V_m をクリップし，採点対象演技映像集合 $C_n = \{c_t^n \in \mathbb{R}^D\}_{t=1}^T$ と模範演技映像集合 $C_m = \{c_t^m \in \mathbb{R}^D\}_{t=1}^T$ に分割する．ここで，分割する映像数を T とした．この演技映像集合 C_n, C_m を入力として，演技間の点数差分 Δs を回帰する．最終的に，出力された差分値 Δs を模範演技の点数 s_m に加算することで予測点数 $\hat{s_n}$ を出力する．モデル式は以下のように表される．

$$\Delta s = \mathcal{R}(\mathcal{F}(C_n), \mathcal{F}(C_m)) \tag{1}$$

$$\hat{s_n} = s_m + \Delta s \tag{2}$$

ここで，式中の $\mathcal{F}$ は，各クリップから特徴量を抽出する特徴量抽出器を表し，式中の $\mathcal{R}$ は，抽出された特徴量から点数差を出力する回帰器を表す．多くの研究では，特徴量抽出器 $\mathcal{F}$ には I3D [2] が用いられ，回帰器 $\mathcal{R}$ には Attention を用いた構

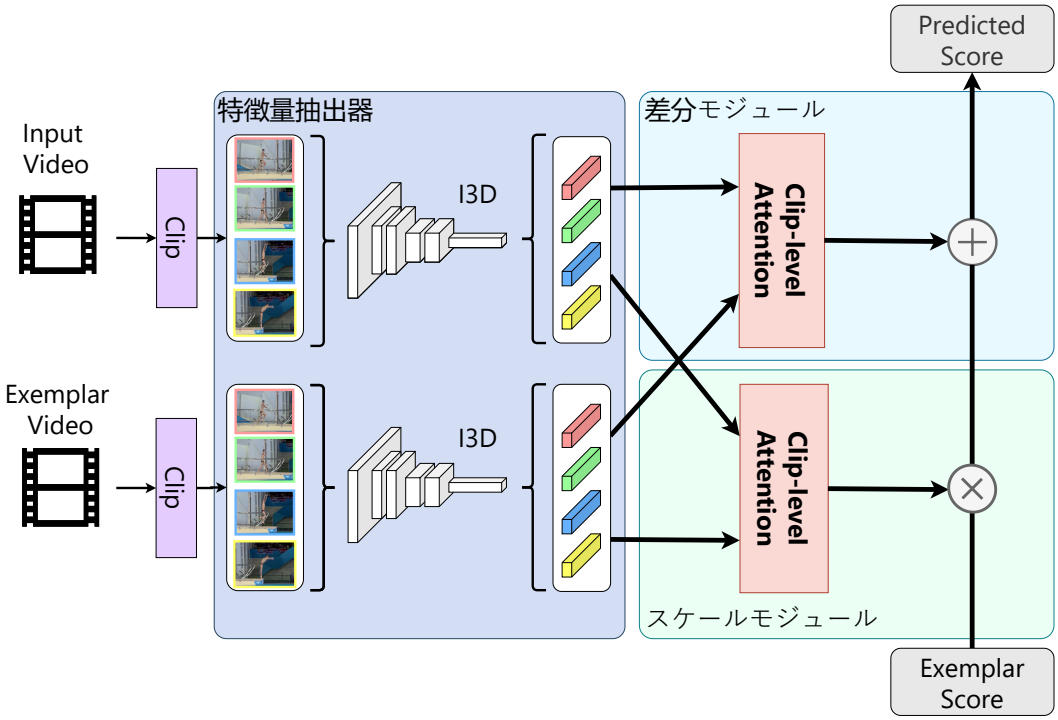


図 1: モデル全体図

造 [10], [12] や木構造 [13] などが用いられる．

本研究では，先行研究の構造をもとに点数差のみではなく，スケール λ と差分 Δs を用いて，点数を出力するモデルを提案する．提案モデルでは，先行研究と同様に，採点対象演技映像 V_n と模範演技映像 V_m をクリップし，採点対象演技映像集合 $C_n = \{c_t^n \in \mathbb{R}^D\}_{t=1}^T$ と模範演技映像集合 $C_m = \{c_t^m \in \mathbb{R}^D\}_{t=1}^T$ に分割する．この演技映像集合を入力として，演技間のスケール λ と点数差 Δs を回帰する．最終的に，模範演技の点数 s_m にスケール λ を乗算し，差分 Δs を加算することで予測点数 $\hat{s_n}$ を出力する．モデル式は以下のように表される．

$$\lambda = \mathcal{R}_\lambda(\mathcal{F}(C_n), \mathcal{F}(C_m)) \tag{3}$$

$$\Delta s = \mathcal{R}_\delta(\mathcal{F}(C_n), \mathcal{F}(C_m)) \tag{4}$$

$$\hat{s_n} = s_m \times \lambda + \Delta s \tag{5}$$

ここで，式中の $\mathcal{F}$ は各クリップから特徴量を抽出する特徴量抽出器を，$\mathcal{R}_\lambda$ は抽出された特徴量からスケール λ を出力する回帰器（スケールモジュール）を，$\mathcal{R}_\delta$ は抽出された特徴量から点数差分 Δs を出力する回帰器（差分モジュール）を表す．提案モデルでは，特徴量抽出器 $\mathcal{F}$ に I3D [2] を用い，回帰器 $\mathcal{R}_\lambda$, $\mathcal{R}_\delta$ には，Clip-level Attention を用いた．

3.2 Clip-level Attention

先行研究 [1], [12] では, 回帰器に Transformer ベースの構造が用いられている．提案モデルに含まれる Clip-level Attention の詳細を図 2 に示す．本研究では，先行研究 [12] で利用されている Transformer 構造に含まれる Source-Target Attention 構造を用いることで，採点対象映像と模範映像間での特徴量の比較を行う．また，Self-Attention 構造を導入し，映像ごとにクリップ特徴量間の関連性を捉える．

まず，I3D [2] によって抽出されたクリップ特徴量 F_n, F_m を入力とし，Self-Attention(SA) によって，映像ごとにクリップ特徴量間の関連性を捉える．

$$F'_n = \mathrm{SA}(F_n) \tag{6}$$

$$F'_m = \mathrm{SA}(F_m) \tag{7}$$

次に，Self-Attention 構造から出力されたクリップ特徴量 F'_n, F'_m を，Source-Target Attention(STA) に入力し，入力映像と模範映像間での特徴量の比較を行う．

$$F''_n = \mathrm{STA}(F'_n, F'_m) \tag{8}$$

$$F''_m = \mathrm{STA}(F'_m, F'_n) \tag{9}$$

最後に，各映像のクリップ特徴量 F''_n, F''_m を平均した後に結合し，MLP に入力する．

$$f_n = \mathrm{Mean}(F''_n) \tag{10}$$

$$f_m = \mathrm{Mean}(F''_m) \tag{11}$$

$$f_{(n,m)} = \mathrm{Concat}([f_n, f_m]) \tag{12}$$

$$y = \mathrm{MLP}(f_{(n,m)}) \tag{13}$$

得られた出力 y は，スケール λ，差分 Δs として扱われる．

4 実　　験

4.1 実験設定

4.1.1 データセット

本研究では，モデルの性能を比較するために，AQA-7 [7] に含まれる Diving のデータを用いる．Diving は，飛び込み競技の演技映像と点数がセットになったデータセットである．1 つ

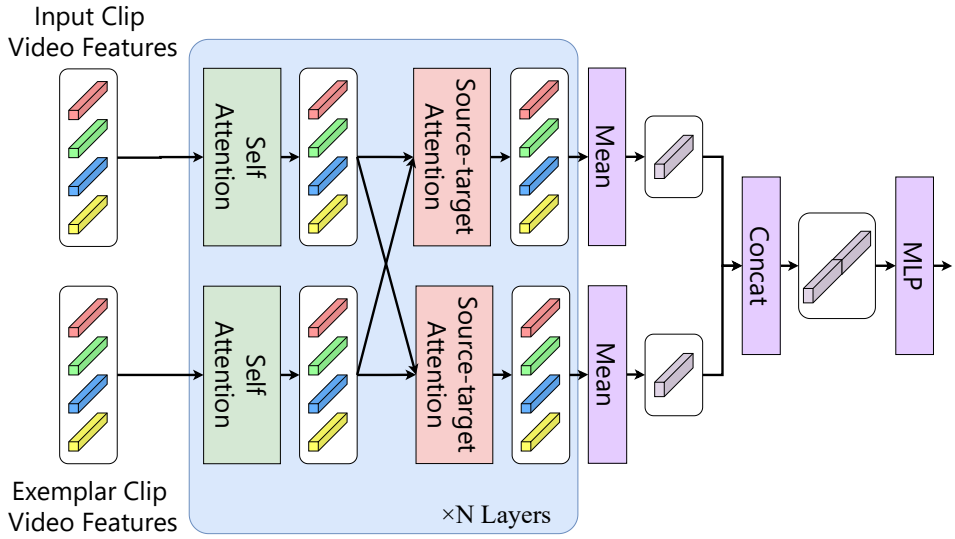


図 2: Clip-level Attention

の演技映像は 103 フレームで構成され，トレーニング用に 300 videos，テスト用に 70 videos の合計 370 videos が含まれている．

4.1.2 評価指標

評価指標には，順位相関係数と $R\text{-}l_2$ 距離を用いる．

順位相関係数 順位相関係数は，式 (14) で表される．ここで，式中の $p_i, \hat{p_i}$ は，それぞれ，i 番目の採点対象映像の順位と比較映像の順位を表す．順位相関係数 ρ の値が，1.0 に近いほど，良い性能であることを表す．

$$\rho = \frac{\sum_{i=1}^{N}(p_i - \bar{p})(\hat{p_i} - \bar{\hat{p}})}{\sqrt{\sum_{i=1}^{N}(p_i - \bar{p})^2 \sum_{i=1}^{N}(\hat{p_i} - \bar{\hat{p}})^2}} \tag{14}$$

$R\text{-}l_2$ 距離 $R\text{-}l_2$ 距離は，式 (15) で表される．ここで，$s_n, \hat{s_n}$ は，それぞれ，n 番目の採点対象映像の点数と比較映像の点数を表す．また，s_{max}, s_{min} は，それぞれ，最も高い点数と最も低い点数を表す．$R\text{-}l_2$ の値が，小さいほど良い性能であることを表す．

$$R\text{-}l_2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|s_n - \hat{s_n}|}{s_{max} - s_{min}}\right) \tag{15}$$

4.1.3 実験詳細

特徴量抽出器として用いた I3D [2] は Kinetics [2] によって事前学習を行った．エポック数は 400 エポック，バッチサイズを 4 とした．最適化手法には，Adam を用い，weight decay は 0 に設定した．学習率は，I3D [2] を 1e-03 とし，それ以外を 1e-04 とした．損失関数には，平均二乗誤差 (Mean Squared Error) を用いた．I3D [2] に映像を入力する前処理として，103 フレームで構成される 1 つの入力映像を 10 個のクリップに分割した．1 つのクリップには，16 フレームが含まれ，クリップ間で重複するフレームが存在する．演技の点数にも前処理として正規化を行い，点数の範囲を 0 100 とした．学習に用いるデータは，訓練用と検証用を 8:2 の割合で分割し，訓練用 240 データ，検証用 60 データとした．先行研究 [1], [12], [13] と同様に，テスト時には，1 つの入力映像に対して 10 回分の模範演技を用いて推論を行い，出力された点数を平均して，最終的な点数とした．各モジュールでは，Clip-level Attention 構造を 3 層積み重ねた．層の数は，2 層，3 層，5 層で試した結果，評価指標が最も良かった 3 層を採用した．また，MLP は先行研究 [12] の構造を参考にし，線形層を 3 層積み重ねている．各層は，それぞれ，1 層目が入力 2048 次元, 出力 256 次元，2 層目が入力 256 次元, 出力 64 次元，3 層目が入力 64 次元, 出力 1 次元で構成されている．活性化関数は，ReLU を用いている．

4.2 実験結果

実験結果を表 1 に示す．ここで，提案モデルの実験結果と AQA-7 データセット [7] の飛び込みにおける従来研究との比較を行う．従来研究としては，CoRe [13]，DAE-CoRe [14]，TPT [1] を挙げた．実験結果を表 1 に示す．まず，順位相関係数では，DAE-CoRe [14] よりも 0.0028 高い結果となった．一方で，SOTA である TPT [1] よりは，0.0018 低い結果となった．次に，$R\text{-}l_2(\times 100)$ では，SOTA である TPT [1] よりも 0.23 大きい結果となった．

表 1: AQA-7 の Diving における性能比較結果

モデル	順位相関係数 ↑	$R\text{-}l_2(\times 100)$ ↓
CoRe [13]	0.8824	0.64
DAE-CoRe [14]	0.8923	–
TPT [1]	**0.8969**	**0.53**
ours	0.8951	0.76

4.3 Ablation Study

4.3.1 スケール・差分モジュールの性能検証

ここでは，スケールモジュールと差分モジュールの性能を検証する．モジュールの性能比較結果を表 2 に示す．表 2 において，× は対象のモジュールを除いていることを表し，○ は対象のモジュールを用いていることを表す．差分モジュールを除い

た baseline1 では，提案モデルよりも，相関係数で 0.9610 低く，$R\text{-}l_2(\times 100)$ で 3.26 大きくなった．スケールモジュールを除いた baseline2 では，提案モデルよりも，相関係数で 0.0255 低く，$R\text{-}l_2(\times 100)$ で 0.10 大きくなった．まとめると，スケールモジュール，差分モジュールを除くと，順位相関係数，$R\text{-}l_2$ 距離ともに，性能が低下する結果となった．

表 2: スケール，差分モジュールの性能検証結果

モデル	スケール	差分	順位相関係数 ↑	$R\text{-}l_2(\times 100)$ ↓
baseline1	×	○	-0.0659	4.02
baseline2	○	×	0.8696	0.86
ours	○	○	0.8951	0.76

4.3.2 Clip-level Attention の性能検証

ここでは，スケールモジュールと差分モジュールに含まれる，Clip-level Attention の性能を検証する．Clip-level Attention 構造の性能検証を表 3 に示す．表 3 において，× は Clip-level Attention を除き，3 層の MLP に置き換えていることを表し，○ は Clip-level Attention を用いていることを表す．

スケールモジュールと差分モジュール，両モジュールの Clip-level Attention を置き換えた baseline1 では，提案モデルよりも順位相関係係数で 0.0452 低く，$R\text{-}l_2(\times 100)$ で，0.21 大きくなった．スケールモジュールの Clip-level Attention を置き換えた baseline2 では，提案モデルよりも順位相関係係数で 0.0463 低く，$R\text{-}l_2(\times 100)$ で，0.28 大きくなった．差分モジュールの Clip-level Attention を置き換えた baseline3 では，提案モデルよりも順位相関係係数で 0.0534 低く，$R\text{-}l_2(\times 100)$ で，0.09 大きくなった．まとめると，各モジュールの Clip-level Attention を除くと，順位相関係数，$R\text{-}l_2$ 距離ともに，性能が低下する結果となった．

表 3: Clip-level Attention の性能検証結果

モデル	モジュールごとの Clip-level Attention の有無		順位相関係数 ↑	$R\text{-}l_2(\times 100)$ ↓
	スケール	差分		
baseline1	×	×	0.8499	0.97
baseline2	×	○	0.8488	1.04
baseline3	○	×	0.8417	0.85
ours	○	○	0.8951	0.76

4.3.3 Self-Attention の性能検証

ここでは，Clip-level Attention に含まれる Self-Attention の性能を検証する．Clip-level Attention は，スケールモジュールと差分モジュールに用いられているため，それぞれの Self-Attention を除くことで評価指標の変化を比較する．Self-Attention の性能検証結果を表 4 に示す．表 4 において，× は Self-Attention を除くことを表し，○ は Self-Attention を用いることを表す．

スケールモジュールと差分モジュール，両モジュールの Self-Atttention を除いた baseline1 では，提案モデルよりも相関係数で 0.0518 低く，$R\text{-}l_2(\times 100)$ では，0.21 大きくなった．スケールモジュールの Self-Atttention を除いた baseline2 では，提案モデルよりも相関係数で 0.0381 低く，$R\text{-}l_2(\times 100)$ では，0.05 大きくなった．差分モジュールの Self-Atttention を除いた baseline3 では，提案モデルよりも相関係数で 0.0248 低く，$R\text{-}l_2(\times 100)$ では，0.22 大きくなった．まとめると，Clip-level Attention に含まれる Self-Attention を除くと，順位相関係数，$R\text{-}l_2$ 距離ともに，性能が低下する結果となった．

表 4: Self-Attention の性能検証結果

モデル	モジュールごとの Self-Attention の有無		順位相関係数 ↑	$R\text{-}l_2(\times 100)$ ↓
	スケール	差分		
baseline1	×	×	0.8433	1.02
baseline2	×	○	0.8570	0.81
baseline3	○	×	0.8708	0.98
ours	○	○	0.8951	0.76

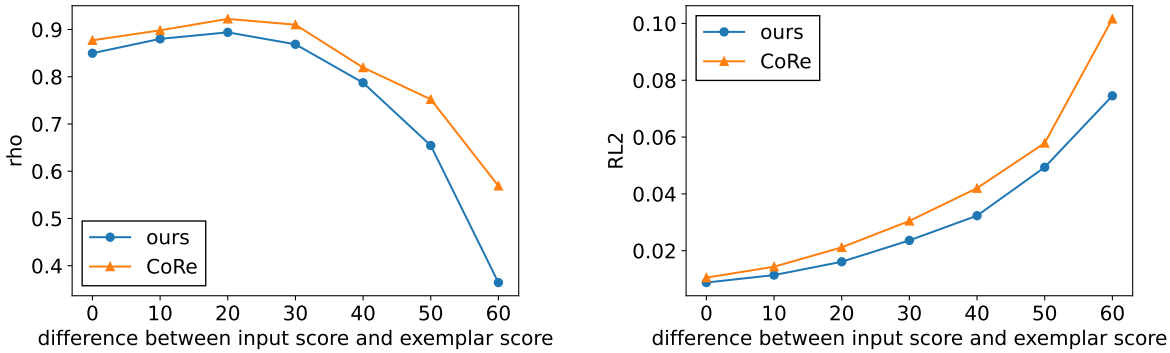
4.4 採点対象演技と模範演技の点数差とモデルの性能比較

ここでは，採点対象となる演技と模範となる演技の実際の点数差と，モデルの性能の変化を比較する．採点対象演技と模範演技の点数差とモデルの性能比較を図 3 に示す．比較対象として，採点対象演技と模範演技の点数差で回帰するモデルである CoRe [13] を用いた．図 3a では，順位相関係数を y 軸にとり，図 3b では，$R\text{-}l_2$ 距離を y 軸にとっている．x 軸は，モデルを評価するデータセットの採点対象演技と模範演技の点数差がその値以上であることを表す．

図 3 から，順位相関係数では，点数差が大きくなるほど性能が低下していくことが分かる．また，モデルを比較すると CoRe [13] が最も高い値となっていることがわかる．$R\text{-}l_2$ 距離でも，順位相関係数と同様に，点数差が大きくなるほど性能が低下していくことが分かる．また，モデルを比較すると点数差が大きくなり性能が低下しているが，提案モデルが最も高い値となっていることがわかる．まとめると，採点対象となる演技と模範となる演技の実際の点数差が大きい場合においては，順位相関では CoRe の方が精度が良くなった一方で，$R\text{-}l_2$ 距離では提案モデルの方が精度が良くなった．このことから，点数差が大きい模範演技を持ってきた場合には，提案モデルが実際の点数に近い値を出しているが，出力された点数による順位のわずかな違いによって，順位相関係数が悪くなっているのではないかと考えられる．

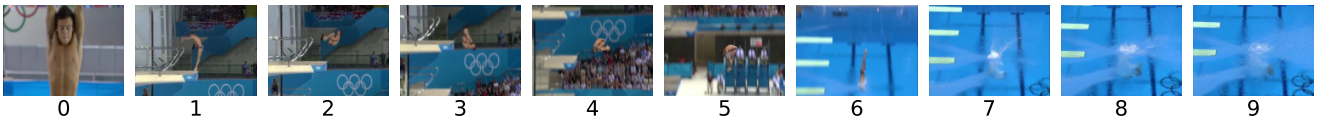
4.5 Clip-level Attention の可視化例

ここでは，Clip-level Attention に含まれる Self-Attention と Source-Target Attention の可視化を行う．可視化例として取り上げる採点対象演技と模範演技を図 4 に示す．図 4 では，映像を分割した 10 クリップ，それぞれの 1 枚目の画像とクリップ番号を可視化している．図 5 に Clip-level Attention1 層目の Attention の重みを可視化した結果を示す．図 5a は，スケールモジュールでの Clip-level Attention を可視化し，図 5b は，差分モジュールでの Clip-level Attention を可視化してい

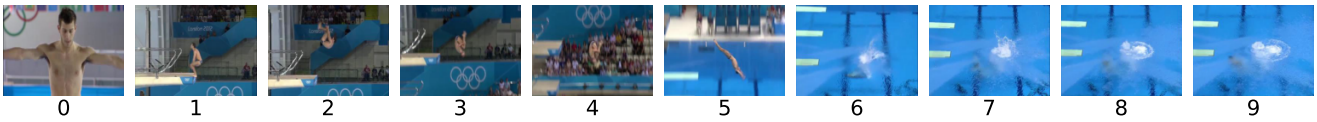


(a) 順位相関係数　　(b) R-l2 距離

図 3: 採点対象演技と模範演技の点数差によるモデルの性能の変化



(a) 採点対象演技映像



(b) 模範演技映像

図 4: 採点対象演技と模範演技の映像例

る．図 5a，5b のそれぞれの画像は，上の行の画像が採点対象の演技，下の行の画像が模範演技を表し，左の列の画像が Self-Attention，右の列の画像が Source-Target Attention を表している．また，それぞれの図の縦軸は入力となる映像のクリップ番号を表し，横軸は比較対象になる映像のクリップ番号を表す．ここで，図 4 と図 5 における，採点対象演技と模範演技，クリップ番号は，それぞれ対応している．

Self-Attention では，演技内の動作のまとまり (踏切，回転，入水) ごとに，クリップ特徴量が異なる重みをもつことで，演技内の動作の切り替えをつかむことができると考え，各映像自身のクリップ間の特徴量を比較している．

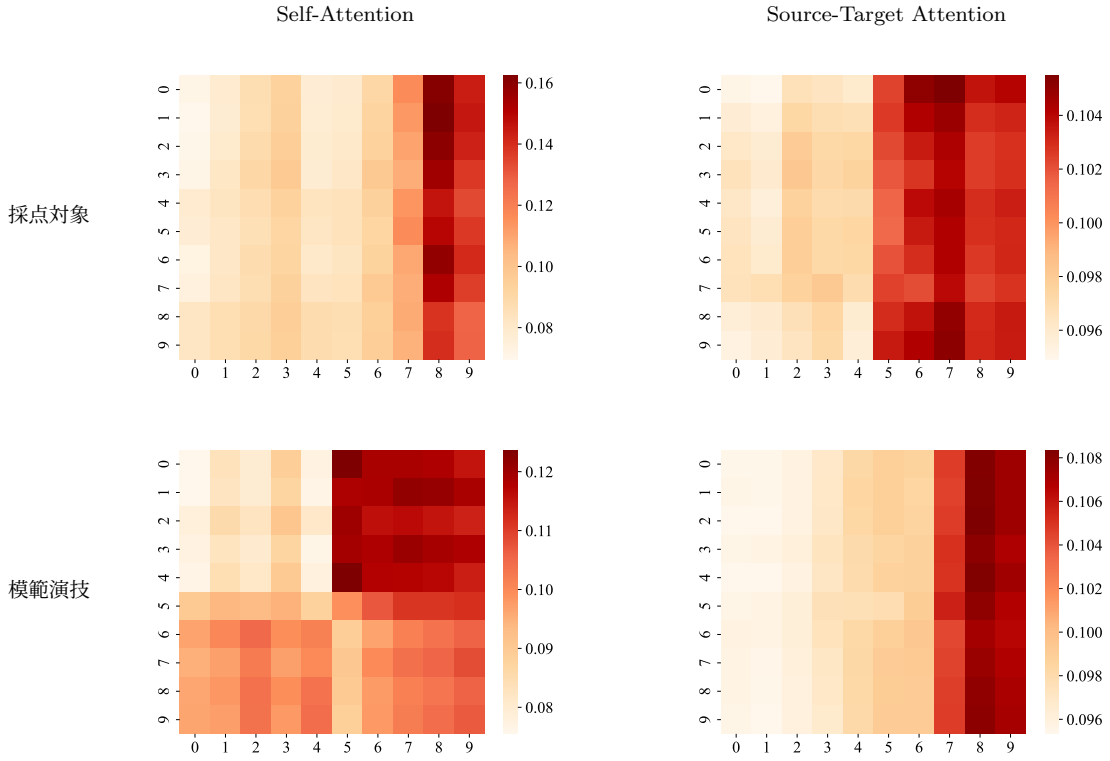
想定していた動作として近くなった例として，図 5a のスケールモジュール内の模範演技における Self-Attention を挙げる．図 5a では，縦軸の 0〜4 クリップ目，5 クリップ目，6〜9 クリップ目で値の反応が異なっている．ここで，図 4b の模範演技画像では，0〜4 クリップ目は飛び込み開始から空中での演技中の映像であり，5 クリップ目は入水前後の映像，6〜9 クリップ目では入水後の水しぶきの映像となっている．これらのことを踏まえて，演技映像と Self-Attention の可視化結果を比較すると，演技内の大きな部分ごとで Attention の反応が異なっていると考えられる．このように，全ての Self-Attention ではないが，Self-Attention によって演技の部分ごとに特徴を捉えていると考えられる可視化結果が得られた．

Source-Target Attention では，採点対象演技の特徴量と模範演技の特徴量における違いを強調できると考え，両演技の特徴量を比較している．

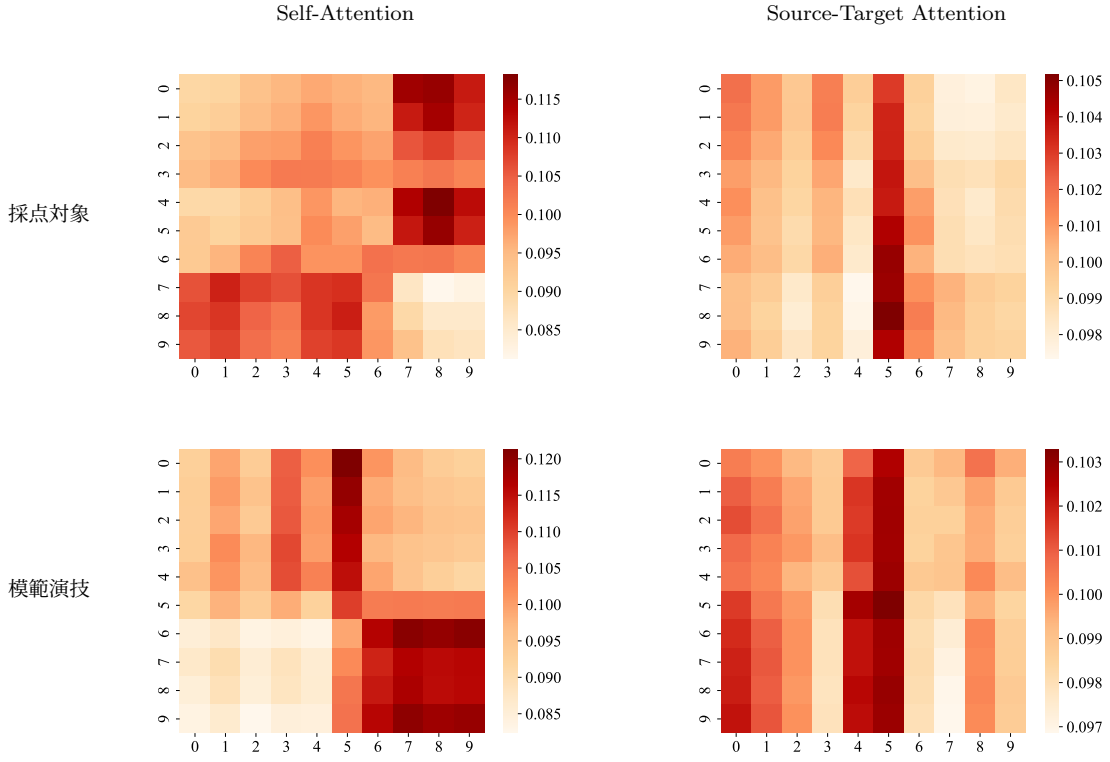
想定していた動作として近くなった例として，図 5b の差分モジュール内の採点対象演技における Source-Target Attention を挙げる．図 5b では，縦軸 (採点対象のクリップ番号) のすべてのクリップで，横軸 (模範演技のクリップ番号) の 5 クリップ目の値が高くなっている．ここで，図 4 の採点対象演技と模範演技の 5 クリップ目を比較すると，図 4a は空中での降下中の映像であるが，図 5b は入水前後の映像となっており，他のクリップに比べて，違いが目立つことが分かる．これらのことを踏まえて，演技映像と Source-Target Attention の結果を比較すると，違いの大きいクリップで高い反応を示していると考えられる。このように，全ての Source-Target Attention ではないが，Source-Target Attention では，演技特徴量間の違いを捉えていると考えられる可視化結果が得られた．

5 ま と め

本研究では，採点対象の演技映像と模範となる演技映像を比較し，模範となる演技の点数にスケールと差分の演算を適用するモデルを提案した．また，従来の研究で用いられていたセグメンテーション構造の代わりに，Self-Attention 構造を用いた Clip-level Attention を導入し，映像の部分ごとの関連性を捉えることを試みた．結果として，飛び込み競技のデータセットである AQA-7 において，SOTA に匹敵する結果を示した．スケールモジュールを導入することで，模範演技と採点対象演技の点数差が大きい場合に，差分を回帰するモデルよりも，一部



(a) スケールモジュール



(b) 差分モジュール

図 5: 採点対象映像と模範映像に対する 1 層目の Attention 可視化例

の評価指標で精度が保たれることを示した.

文　　献

[1] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European Conference on Computer Vision*, pp. 422–438. Springer, 2022.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and*

Pattern Recognition, pp. 6299–6308, 2017.

[3] Shafkat Farabi, Hasibul Haque Himel, Fakhruddin Gazzali, Bakhtiar Hasan, Md Kabir, Moshiur Farazi, et al. Improving action quality assessment using resnets and weighted aggregation. *arXiv preprint arXiv:2102.10555*, 2021.

[4] 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*, Vol. 3, 2014.

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[6] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6330–6339, 2019.

[7] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1468–1476. IEEE, 2019.

[8] Paritosh Parmar and Brendan Tran Morris. 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*, pp. 304–313, 2019.

[9] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9839–9848, 2020.

[10] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4902–4910, 2021.

[11] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *2021 International Conference on Networking Systems of AI (INSAI)*, pp. 1–9. IEEE, 2021.

[12] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2949–2958, 2022.

[13] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7919–7928, 2021.

[14] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. Auto-encoding score distribution regression for action quality assessment. *arXiv preprint arXiv:2111.11029*, 2021.