動画像における行動認識での自己教師あり学習の有効性

植田, 崇弘 / UEDA, Takahiro

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
65
(開始ページ / Start Page)
1
(終了ページ / End Page)
5
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030734

動画像における行動認識での自己教師あり学習の有効性

EFFECTIVENESS OF SELF-SUPERVISED LEARNING FOR ACTION RECOGNITION IN VIDEOS

植田崇弘
Takahiro UEDA
指導教員　平原誠

法政大学大学院理工学研究科応用情報工学専攻修士課程

In supervised learning, achieving high accuracy often requires a large labeled dataset; however, the associated human labeling costs can be high. The aim of this study is to improve recognition accuracy using a small labeled dataset in the context of action recognition in videos by employing self-supervised learning for feature extraction. In the pretraining phase, three-dimensional convolutional neural networks (3DCNN) are trained on frame order prediction to learn temporal features in video sequences. The pretrained parameters are used to initialize the 3D-CNN, which is then fine-tuned for action recognition as the main task. The action recognition is analyzed using training dataset sizes of 500, 2,500, 5,000, 10,000, 20,000, and 30,050 instances. For small training dataset sizes of 500, 2,500 and 5,000 instances, the pretrained model shows a significant improvement in recognition accuracy compared to the model without pretraining, confirming the effectiveness of self-supervised learning.

***Key Words** : Action Recognition, Self-supervised Learning, Neural Networks, Fine-tuning*

1.　はじめに

近年，画像や動画を対象とした行動認識は，盛んに研究されており実生活にも幅広く活用されている．例えば，IDT(Improved Dense Trajectories)を用いた研究[1]では91.5%の認識率を，LSTM[2]を用いた研究では 93.3%の認識率を示しており，人物の行動をかなり正確に捉えることが可能になってきた．高い認識率を達成している多くの先行研究では，教師あり学習が用いられている．教師あり学習では，人間が各データに対してラベルを付与する必要がある．ラベル付きデータを用いることで，モデルが正しい答えを学習し高い認識率を得ることにつなげている．しかし，認識率を高めるためには大規模なラベル付きデータセットが必要となり，その作成にあたる人的コストが膨大なものになってしまうという問題がある．ラベルの品質も認識率に直接影響を与えることから，高品質なラベル付けが必要となり，人的コストは高まるばかりである．

この教師あり学習の問題を軽減するために，本研究では事前学習に自己教師あり学習[3]を用いた手法を提案する．自己教師あり学習を用いてラベル無しデータセットから事前に特徴抽出することで，小規模のラベル付きデータセットでも認識率の低下を抑えられる可能性がある．近年，自己教師あり学習を用いた研究は自然言語処理分野[4]や画像処理分野[5]などで幅広く利用されており，ラベル付きデータ数を増やすことなく高い認識率を維持することに成功している．本研究では，動画像における行動認識において，自己教師あり学習によるラベル無しデータセットからの特徴抽出を事前に試み，少量のラベル付きデータセットでも高い認識率を持つ 3D-CNN を用いたモデルの構築を目標としている．

2.　自己教師あり学習

自己教師あり学習は目的とするタスク（メインタスク）とは別のタスク（事前タスク）を用いて，メインタスクに有効な特徴を獲得する手法である．事前タスクとしてラベル無しデータセットから自動でラベル付けできるタスクを設定する必要がある．メインタスクでは，事前学習した重みのうち出力層から「遠い」重みを固定し，出力層に「近い」重みのみを学習するというのが一般的である．

自己教師あり学習における事前タスクは，メインタスクを解くにあたって必要な特徴を捉えるものになっている必要があり，これまでに様々な手法が提案されている．

例えば自然言語処理分野では，Devlin らの研究[4]にお

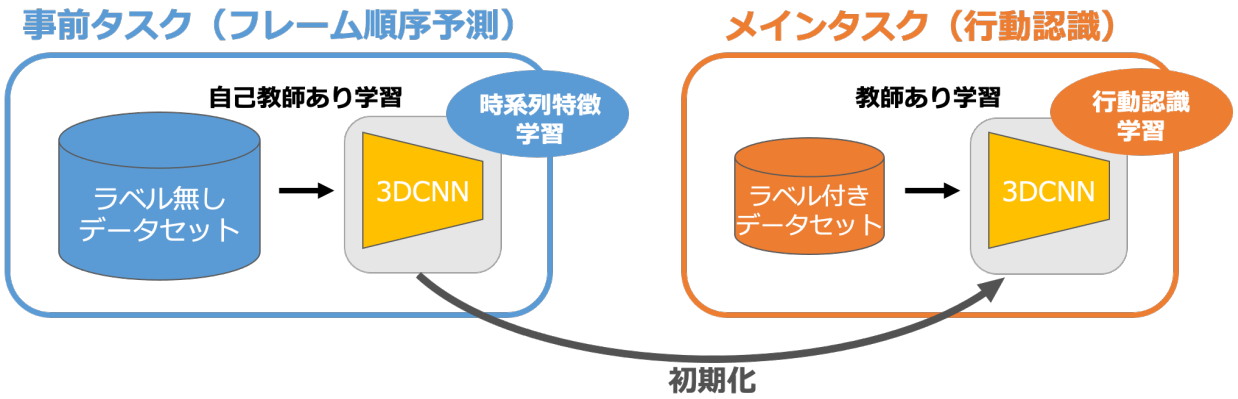

図 1　本研究の概略

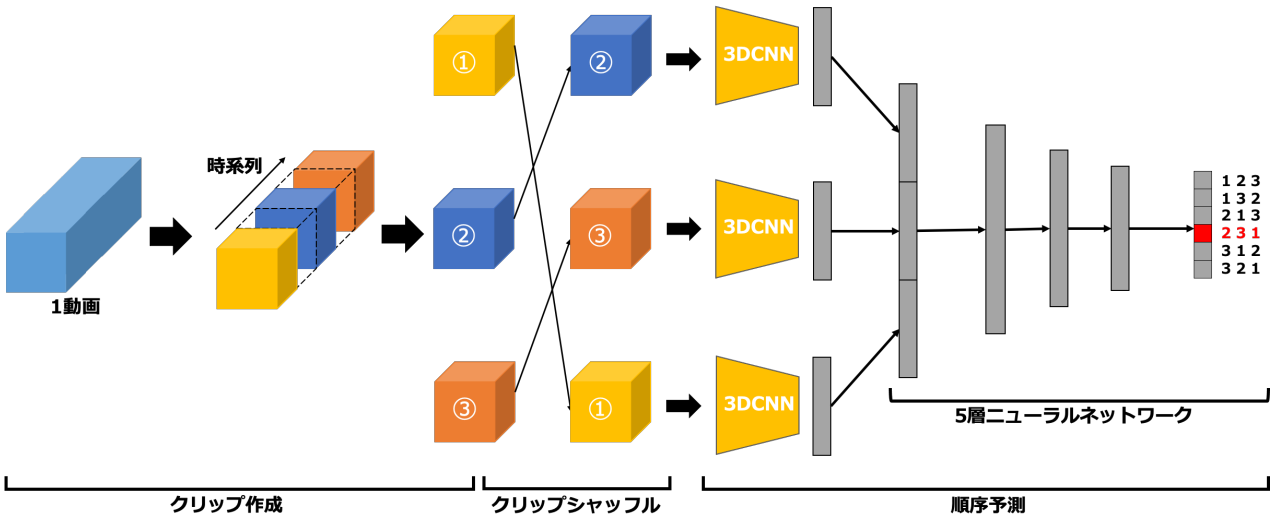

図 2　フレーム順序タスク

いて 2 つの事前タスクが提案されている. 1 つ目の事前タスクは，1 つの文章中の単語をランダムに「MASK」トークンに置き換え，この置き換えられた単語を推測するタスクになっている．正解ラベルには，「MASK」で置き換える前の元の単語を与える．この事前タスクを通じて，一文の単語の並びを考慮しながら単語の意味を学習することができる．2 つ目の事前タスクは，2 つの文章が与えられた時に，これらが連続しているか否かを予測するタスクになっている．このタスクを通じて，文章間の関係を学習することができる．

さらに画像処理分野では, Doersch らの研究[5]において対照学習という事前タスクが提案されている．まず，データセットの各画像に対して 2 つの新たな変形画像を作成する．そして，作成した変形画像の中からランダムに 2 つ選択し，その 2 つが同じ画像からの変形画像かどうかを判断するタスクになっている．この事前タスクを通じて，類似画像の判別に強い学習を実現している．

3.　3D-CNN（3D Convolutional Neural Network）

3D-CNN とは，動画像などの 3 次元データを扱うための畳み込みニューラルネットワークである．入力動画の空間情報（2D）と時系列情報（1D）をまとめ，3D 情報として畳み込み，プーリングを行う．これにより，空間情報だけでなく時系列情報も考慮して動画の特徴を獲得することができる．時系列情報を考慮しない単純な CNN での行動認識[6]では 65.4%の認識率にとどまっているのに対し，3D-CNN を用いた手法 [7]では 95.6%と認識率を格段に高めることに成功している．

4.　提案手法

本研究では図 1 に示すように，事前タスクで自己教師あり学習を用いた特徴抽出を行い，メインタスクで教師あり学習を用いた行動認識を行う．教師あり学習で問題となっている人的コストの軽減を目的とするため，メインタスクで扱うラベル付きデータセットは小規模から大規模まで扱い，事前タスクで扱うラベル無しデータセットは大規模とする．事前タスクとしてフレーム順序予測タスクを課し，動画の特徴を自己教師あり学習させる．その後，3D-CNN を学習済みパラメータで初期化し，メインタスクとなる行動認識で教師あり学習によりファインチューニングする．

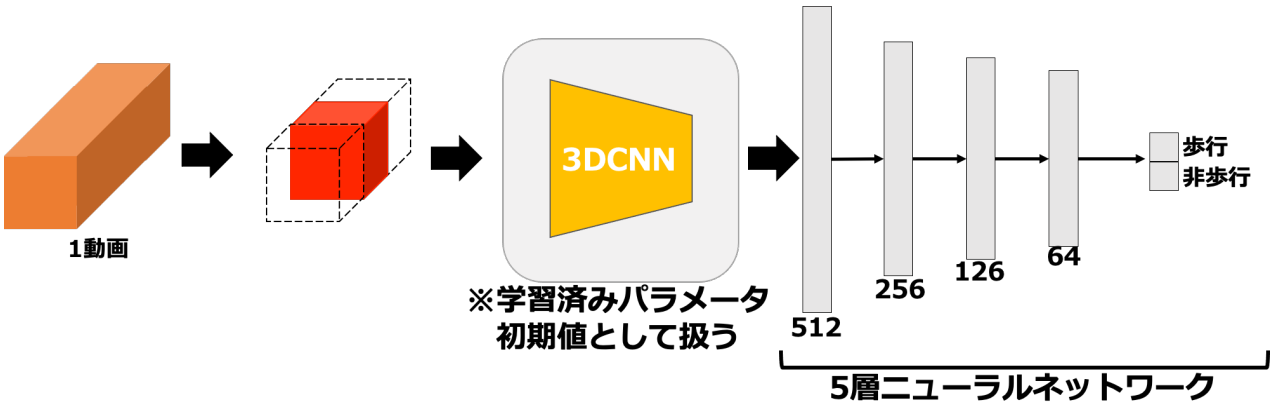

図 3　行動認識タスク

4.1. フレーム順序タスク

本研究のメインタスクは行動認識であることからその実現に必要な特徴は，動画の時系列特徴であると考える．本研究では，時系列特徴を抽出するための事前タスクとして，フレーム順序予測タスクを用いる．

フレーム順序予測タスクは，図 2 に示すように，大きく分けて 3 つの手順で構成される．

まず，自己教師あり学習の教師信号を自動作成するためのクリップ作成を行う．クリップ作成では，一つの動画からランダムな開始時刻で始まる 16 フレームのクリップ動画 3 つを互いに重ならないように作成し，それぞれに元動画の時系列順で番号を振る．

次に，クリップシャッフルで 3 つのクリップ動画をランダムに並び替え，並び替えられた順番を自己教師あり学習での教師信号とする．したがって，分類の数は 6(=3!) 通りとなる．ここで，クリップ動画数を 3 つにしているのは分類数を膨大にし過ぎないためである．例えばクリップ動画を 7 つ作成した場合，分類数は 7!=5,040 通りとなり非常に困難なタスクになってしまう．事前タスクでは，特徴抽出に重点を置いているためタスク自体は容易に遂行可能である必要があると考える．本研究では，事前タスクをできる限り簡単にするため作成するクリップ動画数は 3 つとする．

最後に，自動作成した教師信号をもとにフレーム順序予測を行う．作成したクリップ動画一つ一つの特徴を抽出するために 3D-CNN を用いる．なお，図 2 に示す 3 つの 3D-CNN の実体は同じであり，同じ重みを共有する．抽出された特徴（3D-CNN の出力）はすべて連結し，中間層（3 層）を通して出力層に送られる．出力は 3 つのクリップ動画の順序予測のため，分類の数 6 と同じ 6 次元とする．このフレーム順序予測を行うことにより，クリップ動画の空間情報および動画間の順序を学習する過程で，時系列特徴を抽出することが期待できる．

4.2. 行動認識タスク

事前タスクで時系列特徴を抽出した後，メインタスクで行動認識を学習する．本研究では，事前タスクの有効

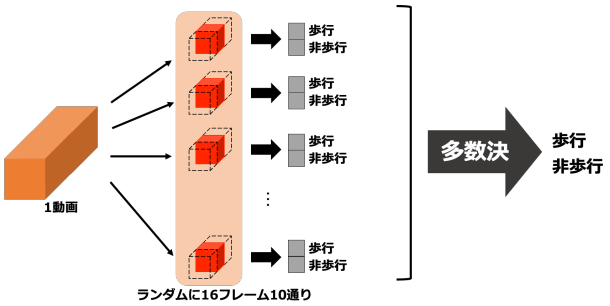

図 4　認識決定手法

性を示すことを目的とするため，図 3 に示すように行動認識問題は簡易的なものとし「歩行」動画か「非歩行」動画かの二値分類とする．まず，事前タスクでの学習済みパラメータで 3 D-CNN を初期化する．次に，一つの動画からランダムな開始時刻で始まる 16 フレームのクリップ動画を作成し， 3D-CNN で学習していく．3D-CNN からの出力は，5 層ニューラルネットワークの入力とし，中間層（3 層）を通して 2 次元出力する．訓練フェーズでは，訓練データそれぞれに対して以上の作業を 10 回試行し，学習した．テストフェーズでは，図 4 に示すように，まず，テストデータの 1 動画に対しクリップ動画を 10 個作成し，それぞれの行動認識結果を出力する．その後，10 個の認識結果の多数決で動画全体に対する最終的な認識結果とする．

5.　実験

5.1. データセット

本研究では，動画に対する認識問題で幅広く利用されている大規模なデータセット kinetics[8]を用いる．各動画は YouTube 上の公開動画から取得され 10 秒間（600 フレーム）のカラー動画になっている．「歩く」，「走る」のような基本的な行動から，「窓を掃除する」，「楽器を演奏する」など人間の行動に焦点が当てられたラベルが付与されている．400 種類の行動ラベルが割り振られており，1 つの行動ごとに少なくとも 400 の動画が含まれている．本研究では，「歩行」動画か「非歩行」動画かの二値分類を行うため，kinetics データセットにおいて「歩

く」とラベル付けされたデータを「歩行」動画，それ以外のラベル付けがされたデータを「非歩行」動画とした．

5.2. フレーム順序予測

事前タスクであるフレーム順序予測では，1つの動画から連続16フレームのクリップ動画を3つ作成する．その際，クリップ動画が連続し似た動画になることを少しでも防ぐようにクリップ動画間に60フレーム以上の間隔をあけて作成した．そのため元動画としては，最低でも168フレームが必要となる．今回扱うデータは60fpsの10秒動画（600フレーム）となるので十分可能な作成方法である．

本研究の3D-CNNは，224×224×16×3（縦方向，横方向，時間軸方向，カラーチャンネル数）を入力とする．畳み込みフィルターサイズおよびプーリングフィルターサイズは3×3×4×3としている．全6層構造を持ち，512次元の特徴量を出力するネットワークになっている．3つのクリップ動画それぞれを3D-CNNに入力したときの特徴量を連結して512×3次元ベクトルとし，中間層（3層）を通して分類結果を6次元出力する．

本研究では，事前タスクで扱うラベル無しデータセットは大規模なものとして考えるため，kineticsデータセットからすでに付与されているラベルに関係なく，訓練データ30,050件をランダムに選んだ．テストデータについては，行動認識のテストデータにも用いるため，「歩行」動画600件と「非歩行」動画600件を訓練データと重ならないようにランダムに選んだ．

テストデータに対するフレーム順序予測を10回行ったところ，平均正解率は，77.2%（SD=0.69）であった．

5.3. 行動認識

本研究のメインタスクは，kineticsデータセットにある400種類の行動を認識するのではなく，「歩行」動画か「非歩行」動画かの二値分類を行った．メインタスクでは，本研究の事前タスクの有効性を調査するため，事前学習ありモデルと事前学習なしモデルの2つのモデルを作成し，結果の比較を行った．事前学習ありモデルは，3D-CNNの構造が事前タスクのものと同様であるため，事前タスクで学習したパラメータの値で初期化した．3D-CNNで出力された特徴量を5層ニューラルネットワークの入力とし分類結果を2次元出力する．一方，事前学習なしモデルでは，すべてのパラメータをランダムな数値で初期化して学習を行った．本研究では，事前学習ありモデルと事前学習なしモデルそれぞれに対しパラメータ調整を行った．その結果として，両モデルともに3D-CNNの畳み込みフィルターサイズ，プーリングフィルターサイズ，5層ニューラルネットワークの中間層の数が同じ構造となっている．

本研究のメインタスクでは，少量のラベル付きデータにおける事前学習ありモデルの行動認識率が多量のラベル付きデータにおける事前学習なしモデルの行動認識率に対抗しうるか調査する．そのため，本研究では訓練データ数を500件，2,500件，5,000件，10,000件，20,000件，30,050件とし学習した．本研究で用いるデータセットでは，「歩行」動画は1,530件となっているため，各訓練データ数において「歩行」動画と「非歩行」動画の数に偏りが生じる場合がある．データ数500件および2,500件の場合は，訓練データ30,050件から「歩行」動画と「非歩行」動画それぞれを半数ずつ(500件の場合250件，2,500件の場合1,250件)ランダムに選んだ．データ数5,000件，10,000件および20,000件の場合は，訓練データの「歩行」動画1,530件をすべて使用し，残りの件数（5,000件の場合3,470件，10,000件の場合8,470件，20,000件の場合18,470件）を「非歩行」動画からランダムに選んだ．データ数30,050件の場合は，訓練データすべてを利用した．テストデータについては，訓練データ数に関係なくこれまでと同様に「歩行」動画600件と「非歩行」動画600件を訓練データと重ならないようにランダムに選んだ．

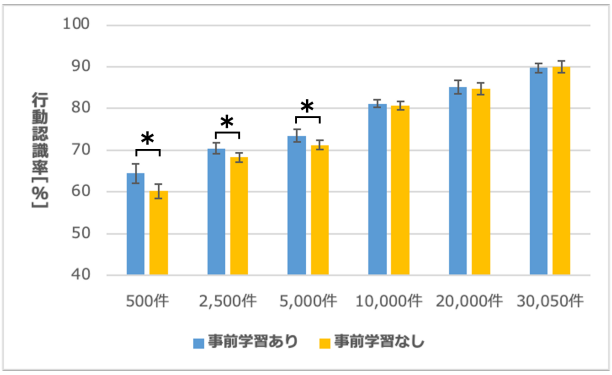

図5　訓練データ数ごとの行動認識率

それぞれの訓練データ数で学習した事前学習ありモデルおよび事前学習なしモデルについて，テストデータに対する行動認識を5回行った場合の平均行動認識率(±1SD)を図5に示す．

6. 考察

図5より，訓練データ数が500件，2,500件および5,000件の場合において，事前学習ありモデルの行動認識率は，事前学習なしモデルの行動認識率に対して有意に向上していることがわかる．訓練データ数が少ない場合，十分な行動認識の精度を見込むことができない場合が多く，本研究の事前タスクであるフレーム順序タスクは有効に働くと考えられる．

しかし，訓練データ数が10,000件，20,000件，30,050件と多量になった場合，事前学習ありモデルは事前学習なしモデルと同等であった．これは，データ数を増やすことによって事前学習の有効性が失われていくことを示唆する．

なお，行動認識率の更なる向上には2つほど考えられる．1つ目は，事前タスクの精度向上である．本研究の過

程で，事前タスクの精度を高めると，事前学習ありモデルの行動認識率も高まる傾向が見受けられた．本研究での事前タスクの精度は，8 割弱に留まっている．この事前タスクの精度をさらに高めることで，行動認識率の向上を期待できる．2 つ目は，学習回数の見直しである．図 5 で示している行動認識率は，訓練データそれぞれが 10 回選ばれて学習した結果となっている．訓練データそれぞれが選ばれる回数を 40 回にした際に，テストデータに対する行動認識率が高まる傾向が見受けられた．本研究では，検証データを設けずに学習したため，学習回数に対して最適であるかどうかが曖昧である．なので，検証データを設け最適な学習回数を見出すことでさらに行動認識率を高められると期待できる．

7. 今後の展望

本論文では，動画像を入力データとした行動認識モデルにおいて，自己教師あり学習による特徴抽出を事前に試み，少量のラベル付きデータでも高い認識率を達成する行動認識モデルを提案した．しかし，少量のデータ数での事前学習ありモデルの行動認識率は，多量のデータ数の事前学習なしモデルの行動認識率に対して劣っている結果になった．更なる行動認識率向上を目指して，前項で述べた手法を試み，さらにはモデルの構造変更も視野に入れると良いだろう．

画像処理分野での自己教師あり学習で有効な対照学習[9]を事前タスクの特徴抽出として用いる手法や，モデルに関しても 3D-CNN よりさらに高いパフォーマンスを持つとされる Transformer を用いた行動認識手法[10]などが提案されているため，これらを試し行動認識率の向上に有効な手法を探し出すことも今後の課題である．

謝辞

本研究を進めるにあたり，熱心なご指導と適切なご助言を賜りました修士論文指導教員の平原誠准教授に心から感謝の意を表します．

参考文献

1) Limin Wang, Yu Qiao, Xiaoou Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” Proc. IEEE conference on computer vision and pattern recognit., pp.4305-4314, 2015.
2) Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, Tieniu Tan, “An attention enhanced graph LSTM network for skeleton-based action recognition,” Proc. IEEE/CVF conference computer vision pattern recognit. (CVPR), pp.1227-1236, 2019.
3) Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, Jie Tang, ”Self-supervised learning : Generative or contrastive,” Proc. IEEE Trans. Knowl. Data Eng., early access, Jun 22, 2021. DOI:10.1109/transinfj.2021.3090866.
4) Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, ”BERT : Pre-training of deep bidirectional transformers for language understanding,” Proc. 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies volume 1, pp.4171-4186, 2019.
5) Carl Doersch , Andrew Zisserman , ”Multi-task self-supervised visual learning,” Proc. IEEE/CVF international conference on computer vision , pp.2051-2060, 2017.
6) Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Fei-Fei Li, “Large-scale video classification with convolutional neural networks,” Proc. IEEE conference on computer vision and pattern recognit. (CVPR), pp. 1725-1732, 2014.
7) Joao Carreira, Andrew Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” Proc. IEEE conference on computer vision and pattern recognit. (CVPR), pp.4724-4733, 2017.
8) Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman, “The kinetics human action video dataset,” arXiv preprint arXiv: 1705.06950, 2017.
9) Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, Yin Cui, ”Spatiotemporal contrastive video representation learning,” Proc. IEEE/CVF conference computer vision and pattern recognit. (CVPR), pp.6964-6974, 2021.
10) 水野 颯介，柳井 啓司，”Transformer を用いた人物行動検出，” 信学技報，vol.121，no.427，PRMU2021-85，pp.157-162, 2022.