社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

長期的分類と短期的分類を統合した現実世界での異常検出

渡邉祐大† 岡部誠† 原田泰典†† 鹿島直二††

† 静岡大学
†† 中部電力株式会社
E-mail: †watanabe.yudai.16@shizuoka.ac.jp, m.o@acm.org,
‡Harada.Yasunori@chuden.co.jp, Kashima.Naoji@chuden.co.jp

あらまし 動画から異常を検出するための簡単で効果的な弱教師あり学習手法を提案する．我々は動画から得られる全セグメントを入力とし，その動画が正常か異常かを判定するモデルを提案する．Self-attention 機構を導入することで，入力された全セグメントを解析し，正常/異常の判定に重要な特徴を抽出させる．提案手法はセグメント単位でなく動画単位の学習を行うため MIL が不要となる．その結果，手法の実装が簡単となり，また，学習時に指定すべきハイパーパラメータの数も減るため，扱いやすさが向上する．推論時には，対象の動画を複数の短い動画に分割し，分割された各動画に対して提案手法を適用することで短時間あたりの正常/異常を検出する．3 つのベンチマーク・データセット（UCF-Crime，ShanghaiTech，XD-Violence）を用いて，フレームレベルの検出精度を評価したところ，提案手法は最先端の手法と同等の精度が達成できることが分かった．また，提案手法は他手法とのアンサンブルを行うことで更なる精度向上が実現できる．我々は，短期的な情報に基づく新しい異常検出器を開発するために，セグメントから双方向 LSTM によって抽出される特徴量を用いて MIL を行う．長期的な検出器と短期的な検出器の結果の平均をとることで，最先端の手法を上回る検出精度が達成できることを示す．
キーワード 動画からの異常検出，監視カメラ，弱教師あり学習，アンサンブル

Real-world Anomaly Detection by Integrating Long-Term and Short-Term Classifications

Yudai WATANABE†, Makoto OKABE†, Yasunori HARADA††, and Naoji KASHIMA††

† Shizuoka University
†† Chubu Electric Power Co., Inc.
E-mail: †watanabe.yudai.16@shizuoka.ac.jp, m.o@acm.org,
‡Harada.Yasunori@chuden.co.jp, Kashima.Naoji@chuden.co.jp

Abstract We propose a simple and effective weakly supervised learning method for detecting anomalies in videos. We propose a model to determine whether a video is normal or abnormal by using all the segments obtained from the video as input, and introducing a self-attention mechanism to analyze all the input segments and extract features that are important for determining normal/abnormal. The proposed method does not require MIL because it learns per video instead of per segment. As a result, the implementation of the method is simple and the number of hyperparameters to be specified during training is reduced, which improves the ease of use. During inference, the target video is divided into multiple short videos, and the proposed method is applied to each divided video to detect normal/abnormal per short time. We evaluated the frame-level detection accuracy of the proposed method on three benchmark datasets (UCF-Crime, ShanghaiTech, and XD-Violence), and found that the proposed method can achieve the comparable accuracy as the state-of-the-art methods. In addition, the accuracy can be further improved by ensembling with other methods. We perform MIL on features extracted from segments by bidirectional LSTM to develop a novel anomaly detector for a short-term segment. We show that by averaging the results of the long-term and short-term detectors, we can achieve better detection accuracy than state-of-the-art methods.
Key words Video anomaly detection, surveillance camera, weakly supervised learning, ensemble

1. はじめに

街中に設置されている防犯カメラの台数は世界中で増加を続けており，また工場や発電所などの大規模施設でも複数台の監視カメラを用いて安全確認が行われている．しかし，これらの映像全てを人間が目で見て確認することは難しいため，人間に代わって人工知能が映像を解析し，自動的に異常事象を検出できる技術の開発が急務である．異常事象はめったに観測されないため，正常状態のみを学習データとして用いる手法が多く存在する [1]〜[3], [5]．推論時には入力動画が学習済みの正常状態からどれくらい乖離しているかという基準で正常か異常か判断する．しかし，こういった手法は動画の見た目や速度の違いなどの低レベルな特徴量に基づいた異常検出しかできない．そこで近年，正常な動画と異常な動画の両方を含む弱い教師データセットを用いて異常検出器を学習する手法が提案されている [4], [19].

弱い教師データセットでは動画単位で正常/異常のラベルが付いている．即ち，正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．このようなデータセットを用いることで，動画中のフレーム毎に正常/異常のラベル付けをする必要がなくなり，ラベル付けの労力を削減できる．

既存手法では連続する複数のフレームを 1 つの短期的なセグメントとして扱い，多重インスタンス学習 (MIL) を用いてセグメント毎の正常/異常を判定している．MIL では各動画を 1 つのバッグとして扱い，バッグの中には対応する動画から得られた全セグメントが入っている．異常スコアが最も高いセグメントをバッグから取り出し，そのセグメントが正常な動画から来たものであれば 0 を出力するように，あるいは，そのセグメントが異常な動画から来たものであれば 1 を出力するように異常検出器を学習する [19].

しかし，近年成功している既存手法を観察すると，単一のセグメントだけを用いるのではなく，動画の長期的な情報を効率良く扱うことで高い精度を達成していることが分かった．例えば，top-k 戦略を用いる手法では，異常スコアの上位 k 個のセグメントを用いて MIL を実施している [20], [28]．RTFM [20] では時間方向の畳み込み演算と top-k 戦略を組み合わせることでセグメントの長期的な時間変化を考慮し，高精度な異常検出器を実現した．また，Sapkota ら [28] は Distributionally Robust Optimization を導入することによって，パラメータ k を選択する必要がないモデルを提案した．

これらの手法に触発され，我々は動画の長期的特徴を用いる異常検出器 *Video Classifier* を提案する．提案手法は動画から得られる全セグメントを入力とし，その動画が正常か異常かを判定するモデルを提案する．学習段階において，提案手法はセグメント単位でなく動画単位の学習を行うため MIL が不要となる．その結果，学習手法の実装が簡単となる．入力された全セグメントを self-attention 機構を導入して解析し，正常/異常の判定に重要な特徴量を抽出する．我々の提案手法はハイパーパラメータの数が少なく扱いやすい．Sapkota ら [28] と同様に学習に用いるセグメントの数 (k) をユーザが指定する必要もないが，一方で複雑なアルゴリズムも必要としない簡単なモデルである．

推論時には動画単位ではなく，より短い時間間隔で正常/異常の判断をしたい．そのため，対象の動画を複数の短い動画に分割し，分割された各動画に対して提案手法を適用することで短時間あたりの正常/異常を検出する．

3 つのベンチマーク・データセット（UCF-Crime, ShanghaiTech, XD-Violence）を用いて，フレームレベルの検出精度を評価したところ，提案手法は最先端の手法と同等の精度が達成できることが分かった．また，提案手法は他手法とのアンサンブルを行うことで更なる精度向上が実現できることも分かった．アンサンブルに用いる目的で，我々は短期的な情報に基づく新たな異常検出器も開発した．この異常検出器は単一のセグメントから双方向 long short-term memory(LSTM) によって抽出される特徴量を用いて MIL を行うことで得た．長期的な検出器と短期的な検出器の結果の平均をとることで，最先端の手法を上回る検出精度が達成できることを示す．

2. 関連研究

異常検出器を学習するためには学習データセットが必要である．しかし，現実世界で観測できるのはほとんどが正常状態であり，異常事象は滅多に観測されることがない．そのため，多くの異常検出手法が正常状態のみを学習する教師なし学習のアプローチで開発されてきた．推論時には入力動画が学習済みの正常状態からどれくらい乖離しているかという基準で正常か異常か判断する．学習には混合ガウスモデルなどの統計的なモデルが用いられてきた [14]．深層学習を用いるアプローチでは，autoencoder を用いる手法 [1]〜[3], [5]，GAN を用いる手法 [15] などが提案されている．画像における異常検出では，画像内からランダムに小パッチを切り貼りすることで異常データを作成して学習に用いる手法も提案され，高い精度が報告されている [16]．しかし，こういった手法は基本的に動画の見た目や速度の違いなどの低レベルな特徴量に基づいた異常検出しかできない．

そこで近年, より高レベルな異常検出器を開発するために, 正常な動画と異常な動画の両方を含む弱い教師データセットがいくつか公開されており, それらを用いて異常検出器を学習する手法が数多く提案されている [4], [6], [7], [9], [13], [19], [20], [22], [28]. 通常，フレーム単位の異常検出器を学習させる場合には動画中の全てのフレームに対してラベルを付ける必要があるが，ラベル付けのコストが高くなるという問題がある．そこで Sultani らは動画単位で正常/異常のラベルが付けられた弱い教師データセットを提案した [19]．学習においては動画を 1 つのバッグと考え，そのバックの中から最も異常スコアの高いセグメントを選び出して学習に用いる MIL を導入することで異常検出器を開発した．

弱い教師データセットを用いる多くの異常検出手法が Sultani らの MIL を用いた学習手法 [19] に基づいている．MIL に基づく手法では，異常な動画中の正常な部分を選んで学習してし

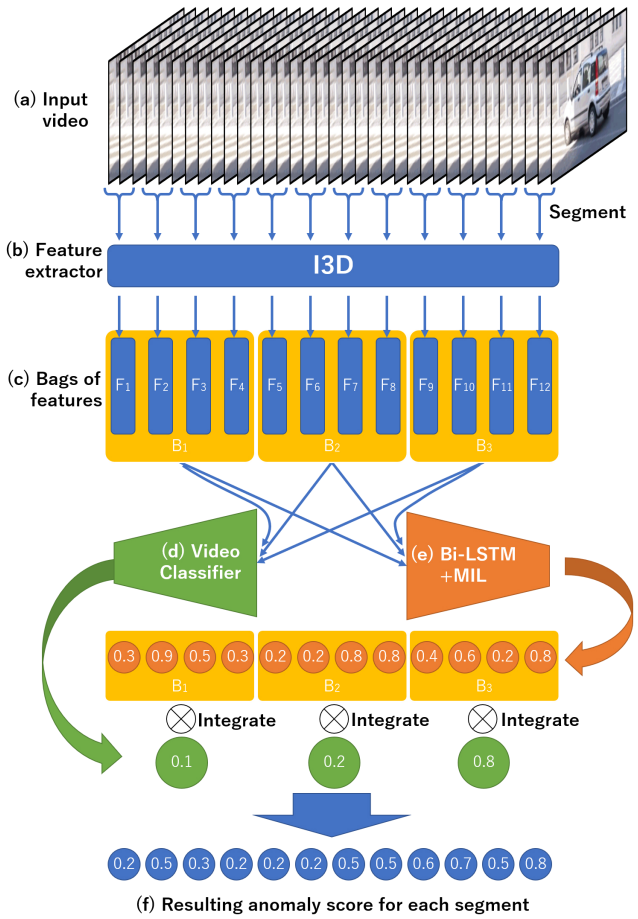

図 1　提案手法における推論時の概要図．(a) 連続する 16 フレームを 1 つのセグメントとし，入力動画をセグメントに分割する．(b) 各セグメントは特徴抽出器である I3D [17] を通して特徴ベクトルに変換される．(c) 複数の特徴ベクトル (図では 4 つ) を 1 つのバッグにまとめる．(d) 各バッグを *Video Classifier* モデルに入力することで，バッグ毎の異常スコアを計算する．(e) 各バッグを *Bi-LSTM+MIL* モデルに入力することで，バッグ内のセグメント毎の異常スコアを計算する．(f) *Video Classifier* によるスコアと *Bi-LSTM+MIL* によるスコアを統合する (ここでは平均をとっている) ことで最終的なセグメント毎の異常スコアを計算する．

Fig. 1　Overview diagram of the proposed method during inference. (a) The input video is divided into segments, with 16 consecutive frames as one segment. (b) Each segment is converted into a feature vector through I3D [17] as the feature extractor. (c) Multiple feature vectors (four in this figure) are put into one bag. (d) Each bag is input to the *Video Classifier* model to calculate the anomaly score for each bag. (e) We also input each bag into the *Bi-LSTM+MIL* model to calculate the anomaly score for each segment in the bag. (f) The *Video Classifier* and *Bi-LSTM+MIL* scores are integrated (averaged here) to compute the final per-segment anomaly score (f).

まったときに学習に悪影響を及ぼすという問題がある．この問題に対処するために，Zhong らは弱い教師データセットを誤ったラベルを含んだデータセットであると捉え，グラフ畳み込みニューラルネットワークを用いて誤ったラベルを修正して学習するというアプローチを提案した [7]．また，Zaheer らは attention 機構と動画毎のクラスタリングロスを用いて異常な動画中の異常な部分をより強調して学習に用いる手法を提案した [9]．また，Tian らは時間方向の畳み込み演算による特徴抽出と，異常スコアの高い上位 k 個のセグメントを選択して学習に用いるという top-k 戦略を導入した [20]．また，Sapkota らは distributionally robust optimization を導入することでパラメータ k を選択する必要のないモデルを提案した [28]．近年成功しているこれらの手法を観察してみると，ほとんどの手法が動画の長期的な時間変化を解析して特徴抽出をしている．そこで我々は，self-attention 機構を用いて動画を解析することで，動画の長期的な特徴を抽出して学習に用いる手法を提案する．学習は動画単位で行うが，推論時は動画を分割した短時間の動画毎に異常検出する．提案手法は MIL に基づく手法と比較して実装が簡単であり，ユーザが指定すべきハイパーパラメータが少なく扱いやすい．にも拘らず，既存手法と比較して同等のフレームレベルの検出精度を達成できる．

3. 提 案 手 法

我々は動画の長期的な情報を扱う異常検出器 *Video Classifier* と，短期的なセグメントに基づく異常検出器 *Bi-LSTM+MIL* を提案する．これらの異常検出器を別々に学習し，2 つのモデルの結果を統合することで長期的な時間的特徴と短期的な時間的特徴のどちらにも注目して異常検出を行う手法を提案する．図 1 に提案手法における推論時の概要図を示す．

弱い教師データセットを $\mathcal{D} = \{(\mathbf{V}_i, y_i)\}$ とする．ただし，$\mathbf{V}_i$ は学習データセット中の i 番目の動画，y_i は $\mathbf{V}_i$ に付けられたラベルである．$y_i = \{0, 1\}$ で 0 ならば正常，1 ならば異常を表す．正常とラベル付けされた動画では全フレームを通して正常状態のみが記録されている．一方で，異常とラベル付けされた動画には正常状態と異常状態のフレームが混在している．$\mathbf{V}_i$ は T 個のセグメントに分割され，各セグメントは特徴抽出器によって D 次元の特徴ベクトル $\mathbf{F}_{i,j}$ に変換されている．ただし，$\mathbf{F}_{i,j}$ は $\mathbf{V}_i$ の j 番目の特徴ベクトルを表す．特徴抽出器は全ての実験を通して，Kineticts データセット [29] によって学習済みの I3D [17] を用いた．

3.1 長期的特徴を用いる検出器: Video Classifier

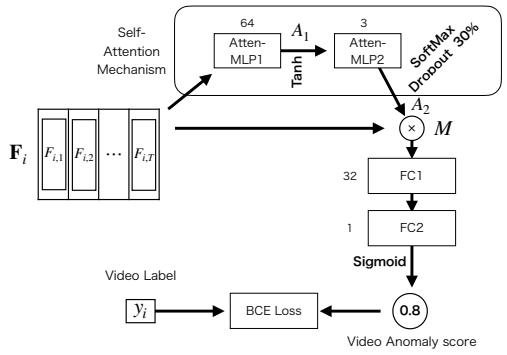

図 2　*Video Classifier* の概要図

Fig. 2　Overview diagram of *Video Classifier*

Video Classifier は self-attention 機構 [18] と 2 層の全結合層からなるシンプルなモデルである．概要図を図 2 に示す．

$\mathbf{F}_i \subset \mathbb{R}^{D \times T}$ は特徴ベクトル $\{\mathbf{F}_{i,1}, \mathbf{F}_{i,2}, \cdots, \mathbf{F}_{i,T}\}$ を結合して

得られる行列とする．Self-attention 機構では，入力された $\mathbf{F}_i$ は Atten-MLP1 の多層パーセプトロンによって $64 \times T$ の行列 $\mathbf{A}_1$ に変換される．この時，活性化には Tanh 関数を用いる．$\mathbf{A}_1$ は Atten-MLP2 の多層パーセプトロンによって $3 \times T$ の行列 $\mathbf{A}_2$ に変換される．この時，活性化には softmax 関数を用いる．更に 30% のドロップアウト正則化 [21] も行う．

Self-attention 機構から得られた重み行列 $\mathbf{A}_2$ を用いて $\mathbf{M} = \mathbf{F}_i\mathbf{A}_2^t$ を計算する．$\mathbf{M}$ の形を $D \times 3$ 次元のベクトルに変形した後，2 層の全結合層 FC1(32 ユニット) と FC2(1 ユニット）を経て動画の異常スコアに変換される．FC1 の活性化関数は恒等関数，FC2 の活性化関数は sigmoid 関数である．損失関数にはバイナリクロスエントロピー (BCE) 関数を用いた．

Video Classifier は Tian らの RTFM [20] などで用いられている top-k 戦略に触発されて考案した手法である．RTFM では，$\mathbf{F}_i$ に対し multi-scale temporal network を用いて時間方向の解析を行った後，上位 k 個の特徴ベクトルを選択し，学習に用いている．我々は上位の特徴ベクトルを選択するという過程をニューラルネットワークに任せるために self-attention 機構を導入している．Self-attention 機構の導入によって，上位何個の特徴ベクトルを学習に使うかのパラメータ k をユーザが指定する必要がなくなり，扱いやすい手法となっている．

3.2 Video Classifier による推論

Video Classifier を適用して異常検出したい動画を $\mathbf{V}^e$ とする．まず，連続する 16 フレームを 1 つのセグメントとし，$\mathbf{V}^e$ を N 個のセグメント $\{\mathbf{V}_1^e, \mathbf{V}_2^e, \cdots, \mathbf{V}_N^e\}$ に分割する．各セグメント $\mathbf{V}_i^e$ は I3D [17] によって D 次元の特徴ベクトル $\mathbf{F}_i^e$ に変換される．特徴ベクトルの集合を $\mathbf{F}^e = \{\mathbf{F}_1^e, \mathbf{F}_2^e, \cdots, \mathbf{F}_N^e\}$ とする．次に分割サイズを l とし，$\mathbf{F}^e$ を l 個のセグメント毎に $m = N/l$ 個のバッグに分割する．即ち，

$$\begin{aligned}\mathbf{F}^e &= \{\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m\} \\ &= \{\{\mathbf{F}_1^e, \cdots, \mathbf{F}_l^e\}, \{\mathbf{F}_{l+1}^e, \cdots, \mathbf{F}_{2l}^e\}, \cdots, \{\mathbf{F}_{N-l+1}^e, \cdots, \mathbf{F}_N^e)\}\}\end{aligned} \tag{1}$$

となる．次に，$\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ を 1 つずつ *Video Classifier* に入力とすることで推論結果 $\mathbf{S}^v = \{s_1^v, s_2^v, \cdots, s_m^v\}$ を得る．s_i^v は $\{\mathbf{V}_{(i-1)l+1}^e, \cdots, \mathbf{V}_{il}^e\}$ の各セグメントに対する異常スコアとして用いられる．フレームレベルの異常検出の結果を作る必要がある際は，得られたスコアを各セグメントの全フレームに割り当てる．*Video Classifier* はシンプルなモデルだが，フレームレベルの異常検出において，最先端の手法と同等の精度が達成できることを 4 章で示す．

3.3 短期的特徴を用いる検出器: Bi-LSTM+MIL

i 番目の正常な動画を $\mathbf{V}_i^n$ とすると，$\mathbf{V}_i^n$ から得られる T 個の特徴ベクトルのバッグを $\mathbf{F}_i^n = \{\mathbf{F}_{i,1}^n, \mathbf{F}_{i,2}^n, \cdots, \mathbf{F}_{i,T}^n\}$ とする．同様に，j 番目の異常な動画を $\mathbf{V}_j^a$ とすると，$\mathbf{V}_j^a$ から得られる T 個の特徴ベクトルのバッグを $\mathbf{F}_j^a = \{\mathbf{F}_{j,1}^a, \mathbf{F}_{j,2}^a, \cdots, \mathbf{F}_{j,T}^a\}$ とする．

Bi-LSTM+MIL は 2 層の双方向 LSTM と 3 層の全結合層からなるモデルである．概要図を図 3 に示す．

双方向 LSTM は，特徴ベクトルのバッグ ($\mathbf{F}_i^n$ または $\mathbf{F}_j^a$) に含まれる特徴ベクトルを前から順番に読み込んでいく順方向

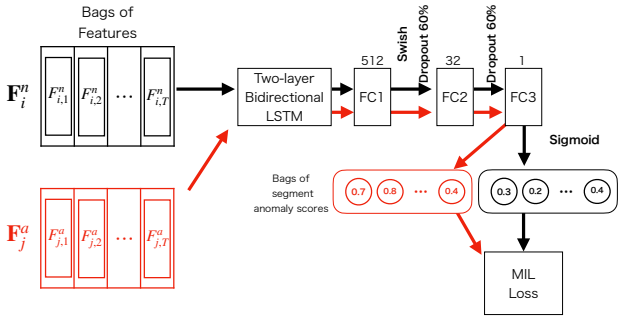

図 3　*Bi-LSTM+MIL* の概要図

Fig. 3　Overview diagram of *Bi-LSTM+MIL*

LSTM と，後ろから逆順に読み込んでいく逆方向 LSTM の 2 つからなるモデルである．双方向 LSTM では順方向 LSTM からの出力と逆方向 LSTM からの出力を結合して最終的な出力とする．LSTM の隠れ状態の次元数を 256 とした．従って，1 つのバッグを入力としたとき，2 層の双方向 LSTM からは 512 次元の特徴ベクトルが T 個出力される．各特徴ベクトルを 3 層の全結合層に入力することで，その特徴ベクトルに対応する異常スコアが得られる．3 層の全結合層は，512 ユニットの全結合層，32 ユニットの全結合層，1 ユニットの全結合層からなる．活性化関数は 1 層目が swish 関数 [27]，2 層目が恒等関数，3 層目が sigmoid 関数である．2 層目と 3 層目の直前で 60% のドロップアウト正則化 [21] を行う．損失関数は既存手法 [19] と同様の MIL ロスを用いた:

$$\begin{aligned}L(\mathbf{F}_i^n, \mathbf{F}_j^a, \mathbf{W}) = \underbrace{max(0, 1 - \max_{\mathbf{F}_j^a} f(\mathbf{F}_{j,k}^a) + \max_{\mathbf{F}_i^n} f(\mathbf{F}_{i,k}^n))}_{①} \\ + \underbrace{\lambda_1 \sum_{k=1}^{T-1} (f(\mathbf{F}_{j,k}^a) - f(\mathbf{F}_{j,k+1}^a))^2}_{②} + \underbrace{\lambda_2 \sum_{k=1}^{T} f(\mathbf{F}_{j,k}^a)}_{③} + \lambda_3 \|\mathbf{W}\|_F\end{aligned} \tag{2}$$

関数 f は *Bi-LSTM+MIL* モデルを表し，$\mathbf{W}$ は 3 層の全結合層の重みを表し，$\|\cdot\|_F$ はフロベニウスノルムを表す．$\mathbf{F}_j^a$ 内で最高の異常スコアを与える特徴ベクトルを $\mathbf{F}_{j,p}^a$，$\mathbf{F}_i^n$ 内で最高の異常スコアを与える特徴ベクトルを $\mathbf{F}_{i,q}^n$ としたとき，① は $f(\mathbf{F}_{j,p}^a) > f(\mathbf{F}_{i,q}^n)$ となるようにモデルを学習するための項である．② は異常スコアがセグメント間で滑らかに変動するように制約する項，③ は検出される異常の数を抑えるための L^2 正則化項である．

Bi-LSTM+MIL は *Video Classifier* や他の既存手法とのアンサンブルの目的に考案した手法である．他の手法とのアンサンブルで高精度を実現するためには，アンサンブルするモデルに多様性を持たせたい．我々の知る限り双方向 LSTM と MIL を組み合わせた手法は他に存在しなかったので 1 つのバリエーションとして開発した．また，映像情報のみでなく音声情報が含まれているようなマルチモーダルなデータセットを扱う場合に，双方向 LSTM が音声情報の時間的変化を上手く捉えることができるのではないかという期待も，導入の動機である．

3.4 Bi-LSTM+MIL の学習

深層学習においては通常，学習データセット内の全データに対して一通り学習を行うことを 1 エポックとし，1 エポックの終了時点で精度評価を行い，良い精度が得られれば重みの保存を行う．しかし，*Bi-LSTM+MIL* の学習過程を観察すると，各学習のステップでランダムに学習データをサンプルしてミニバッチを作って学習を行い，精度評価 (と重みの保存) は高い頻度で実施した方が良いモデルが見つかりやすいことが分かった．実際，1 エポックの終了時点で精度評価を行っていた場合より，学習データのランダムサンプリングと高頻度での精度評価を行った方が，AUC が 1% 以上向上した．この事実に対する我々の考察は，MIL は外れ値に大きな悪影響を受けるのではないか，ということである．外れ値とは，式 (2) の $\max_{\mathbf{F}_j^a} f(\mathbf{F}_{j,p}^a)$ において正常なのに異常として選ばれてしまったセグメントや，異常っぽく見えるのに正常とラベル付けされたセグメントなどである．尚，*Video Classifier* の学習については，通常通りの 1 エポック終了時点で精度評価と重みの保存を行っても精度の低下は見られない．この事実も MIL を使わない *Video Classifier* の扱いやすさを示している．

3.5 2 つのモデルの統合

Video Classifier モデルによる推論結果と *Bi-LSTM+MIL* モデルによる推論結果の平均をとって統合することで，それぞれのモデルを単体で用いる場合よりも高精度な結果を得ることができる．

Video Classifier による推論結果 $\mathbf{S}^v = \{s_1^v, s_2^v, \cdots, s_m^v\}$ は 3.2 章で述べた方法を用いて得ることができる．*Bi-LSTM+MIL* による推論結果 $\mathbf{S}^b$ については，式 (1) における $\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_m$ を 1 つずつ *Bi-LSTM+MIL* に入力とすることで得る．ただし，

$$\begin{aligned}\mathbf{S}^b &= \{\mathbf{s}_1^b, \mathbf{s}_2^b, \cdots, \mathbf{s}_m^b\} \\ &= \{\{s_1^b, \cdots, s_l^b\}, \{s_{l+1}^b, \cdots, s_{2l}^b\}, \cdots, \{s_{N-l+1}^b, \cdots, s_N^b\}\}\end{aligned}$$

である．最後に *Video Classifier* の結果 $\mathbf{S}^v$ と *Bi-LSTM+MIL* の結果 $\mathbf{S}^b$ の平均を次のように計算し，最終的な各セグメントに対する推論結果 $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}$ を得る．

$$s_{(i-1)l+j} = \frac{s_i^v + s_{(i-1)l+j}^b}{2}, (i = 1, 2, \cdots, m \ , j = 1, 2, \cdots, l) \tag{3}$$

4. 実　　験

4.1 データセット及び評価指標

提案手法の評価を行うため，異常検出のための 3 つの弱い教師データセット（UCF-Crime データセット [19]，ShanghaiTech データセット [23]，XD-Violence データセット [4]）で実験を行い，いくつかの既存手法との検出精度比較を行った．

UCF-Crime データセットは，現実世界における監視カメラ映像を集めた大規模なデータセットである．このデータセットは 13 種類の異常を含んでいる．動画の本数は計 1900 本であり，動画の総時間は 128 時間である．1900 本の動画のうち，1610 本が学習用の動画で，290 本がテスト用の動画である．学習用の各動画には動画単位で正常/異常のラベルが付けられている．テスト用の各動画にはフレーム単位で正常/異常のラベルが付けられている．

ShanghaiTech データセットは，大学内に設置された固定カメラで撮影された映像を集めた，中規模なデータセットである．このデータセットには，13 箇所の固定カメラによって撮影された 437 本の動画が含まれている．437 本の動画のうち，307 本が正常な動画で，130 本が異常を含んだ動画である．オリジナルのデータセットは正常な動画のみを学習して異常を検出する，即ち，教師なし学習による異常検出器の開発のためのデータセットであった．しかし，Zhang ら [6] は弱教師あり学習のためのデータセットとして使えるように動画単位のラベル付けを行った．我々は Zhang ら [6] と同様の手順で弱い教師データセットを構築し，実験を行った．

XD-Violence データセットは，映画や動画共有サイトの動画など，様々の種類の動画を含んだ大規模なデータセットである．動画の本数は計 4754 本であり，動画の総時間は 217 時間である．4754 本の動画のうち，爆発や戦闘や暴動などの 6 種類の異常を含んだ動画が 2405 本あり，正常な動画は 2349 本である．UCF-Crime データセットと同様，学習用の各動画には動画単位のラベルが，テスト用の各動画にはフレーム単位のラベルが付けられている．

評価指標．UCF-Crime データセットと ShanghaiTech データセットについては，既存研究 [2], [4]〜[9], [19], [20], [22] と同様に，フレーム単位の異常検出精度に基づいて計算された Receiver Operating Characteristic (ROC) 曲線における Area Under the Curve (AUC) を評価指標とする．AUC の値が大きいほど，高精度な異常検出器であることを表している．XD-Violence データセットについては，既存研究 [4], [20] と同様に Average Precision (AP) を評価指標とする．AP の値が大きいほど，高精度な異常検出器であることを表している．

4.2 実装の詳細

Video Classifier と *Bi-LSTM+MIL* は PyTorch [24] を用いて開発し，評価実験を行った．最適化アルゴリズムには Radam [25] を用い，学習率は 0.001 とした．バッチサイズは 64 とした．ハイパーパラメータについては，*Video Classifier* モデルについては，T を 32 とした．*Bi-LSTM+MIL* モデルについては，T を 64，λ_1 と λ_2 を 0.00008，λ_3 を 0.001 とした．推論時には分割サイズ l を 32 とした．UCF-Crime データセットと ShanghaiTech データセットについては既存手法 [20], [22] と同様に，各動画に対し 10-crop オーギュメンテーションを行った．XD-Violence データセットについては既存手法 [4] と同様に，各動画に対し 5-crop オーギュメンテーションを行った．

4.3 *Video Classifier* モデルの分析

我々は *Video Classifier* に双方向 LSTM を加えること，また，*Video Classifier* から self-attention 機構を取り除くことに関する実験を行った．文章分類の文脈において，Lin ら [18] のモデルが高い分類精度を達成しているが，双方向 LSTM の後に self-attention 機構が用いられている．これを参考に我々も *Video Classifier* に双方向 LSTM を加えたモデルを用い，異常検出の精度評価を行った．具体的には，3.1 章で提案した *Video Classifier*

では $\mathbf{F}_i$ を直接 Atten-MLP1 に入力している (図 2) のに対し，双方向 LSTM を加えたモデルでは $\mathbf{F}_i$ をまずは双方向 LSTM に入力し，双方向 LSTM からの出力を Atten-MLP1 に入力する．

実験結果を表 1 に示す．UCF-Crime データセットと ShanghaiTech データセットでは I3D によって得られた特徴量を用いた．XD-Violence データセットは音声情報を含んでいるので，I3D に加え，VGGish によって得られた特徴量を用いた．UCF-Crime データセットと ShanghaiTech データセットについては，双方向 LSTM を用いず self-attention 機構のみを用いたモデル（3.1 章で提案したモデル）が最も高い検出精度を達成した．XD-Violence データセットについては，双方向 LSTM と self-attention 機構の両方を用いたモデルが最も高い検出精度を達成した．これは XD-Violence データセットの各動画は音声情報を含んでいるため，音声情報に関して双方向 LSTM が効果的に働いたのではないかと推察する．

Bi-LSTM	Self-Attention	UCF-Crime	ShanghaiTech	XD-Violence
✓		81.72	92.90	75.92
	✓	83.96	95.40	75.43
✓	✓	83.28	94.06	78.75

表 1　*Video Classifier* モデルについて，双方向 LSTM の追加及び self-attention 機構の除去に関する実験の結果．

Table 1　Results of experiments on the addition of bidirectional LSTM and the removal of the self-attention mechanism for *Video Classifier* model.

4.4　UCF-Crime データセットの実験結果

表 2 に UCF-Crime データセットにおけるフレーム単位での AUC 性能を示す．*Video Classifier* モデル，または *Bi-LSTM+MIL* モデルを単独で用いた場合でも，MIL に基づく既存手法と同等か，もしくはそれよりも高い検出精度を達成できている．同じ I3D RGB の特徴量を用いている MIST [22] や Wu ら [4] の手法と比較すると，*Video Classifier* 単体でも高い検出精度を示している．シンプルな手法にも関わらず，*Video Classifier* は MIL に基づく手法と遜色ない検出精度が達成可能であることが分かる．また，*Video Classifier* と *Bi-LSTM+MIL* を統合した手法 (Ours) は，既存手法の中で最も検出精度の高かった Tian ら [20] の RTFM よりも 0.42% 高い検出精度を達成している．

4.5　ShanghaiTech データセットの実験結果

表 3 に ShanghaiTech データセットにおけるフレーム単位での AUC 性能を示す．*Video Classifier* モデル，または *Bi-LSTM+MIL* モデルを単独で用いた場合でも，MIL に基づく既存手法と同等か，もしくはそれよりも高い検出精度を達成できている．*Video Classifier* と *Bi-LSTM+MIL* を統合した手法 (Ours) は，既存手法の中で最も検出精度の高かった Tian ら [20] の RTFM と比較して 0.93% 劣る結果となった．しかし，シンプルなモデルである *Video Classifier* が既存のいくつかの手法と比較しても高い検出精度を示しており，*Video Classifier* の有用性を示している．

4.6　XD-Violence データセットの実験結果

表 4 に XD-Violence データセットにおけるフレーム単位での AP 性能を示す．4.3 節でも述べたように XD-Violence データセットは音声情報を含んでいるため，self-attention 機構の前に双方向 LSTM を用いた *Video Classifier* モデルの方が検出精度が高い．そのため双方向 LSTM を用いた *Video Classifier* モデルとの統合も行った (表 4 の “Ours†”)．I3D RGB 特徴量を用いた結果では *Video Classifier* モデル，*Bi-LSTM+MIL* モデル共に Wu ら [4] や RTFM [20] と比較して劣る結果となった．I3D RGB 特徴量に加え音声情報を用いた場合は，*Video Classifier* モデル，*Bi-LSTM+MIL* モデル共に Wu ら [4] の検出精度を超えた．また，統合したモデルを用いた場合は，マルチモーダルな

Supervision	Method	Feature Type	AUC(%)
Unsupervise	SVM Baseline	-	50.00
	Conv-AE [5]	-	50.60
Weakly Supervised	Sultani et al. [19]	C3D RGB	75.41
	Zhang et al. [6]	C3D RGB	78.66
	Motion-Aware [8]	PWC Flow	79.00
	GCN-Anomaly [7]	TSN RGB	82.12
	CLAWS Net [9]	C3D RGB	83.03
	Wu et al. [4]	I3D RGB	82.44
	MIST [22]	I3D RGB (Fine)	82.30
	RTFM [20]	C3D RGB	83.28
	RTFM [20]	I3D RGB	**84.30**
	Video Classifier		83.96
	Bi-LSTM+MIL	I3D RGB	83.14
	Ours		**84.72**

表 2　UCF-Crime データセットにおけるフレーム単位の AUC 性能比較．**青**が最も高い数値，**赤**が二番目に高い数値を表す．“Ours” は *Video Classifier* と *Bi-LSTM+MIL* を統合した手法を表す．

Table 2　Comparison of frame-level AUC performance on the UCF-Crime dataset. **Blue** is the highest value and **red** is the second highest value. “Ours” represents the integration of *Video Classifier* and *Bi-LSTM+MIL*.

Supervision	Method	Feature Type	AUC(%)
Unsupervised	Conv-AE [5]	-	60.85
	Mem-AE [2]	-	50.60
	VEC [12]	-	74.80
Weakly Supervised	GCN-Anomaly [7]	TSN RGB	84.44
	Zhang et al. [6]	I3D RGB	82.50
	CLAWS Net [9]	C3D RGB	89.67
	AR-Net [13]	I3D RGB&Flow	91.24
	MIST [22]	I3D RGB (Fine)	94.83
	RTFM [20]	C3D RGB	91.51
	RTFM [20]	I3D RGB	**97.21**
	Video Classifier		95.40
	Bi-LSTM+MIL	I3D RGB	94.06
	Ours		**96.28**

表 3　ShanghaiTech データセットにおけるフレーム単位の AUC 性能比較．**青**が最も高い数値，**赤**が二番目に高い数値を表す．“Ours” は *Video Classifier* と *Bi-LSTM+MIL* を統合した手法を表す．

Table 3　Comparison of frame-level AUC performance on the ShanghaiTech dataset. **Blue** is the highest value and **red** is the second highest value. “Ours” represents the integration of *Video Classifier* and *Bi-LSTM+MIL*.

Supervision	Method	Feature Type	AP(%)
Unsupervised	SVM baseline	-	50.78
	Hasan et al. [11]	-	30.77
Weakly Supervised	Sultani et al. [19]	C3D RGB	73.20
	Wu et al. [4]	I3D RGB	75.68
	RTFM [20]		77.81
	Video Classifier		75.13
	Bi-LSTM+MIL		73.30
	Ours		74.99
	Wu et al. [4]	I3D RGB +VGGish	78.64
	Pang et al [26]		**81.69**
	VideoClassifire		75.43
	VideoClassifire†		78.75
	Bi-LSTM+MIL		80.54
	Ours		81.06
	Ours†		**82.13**

表 4　XD-Violance におけるフレーム単位の AP 性能比較．**青**が最も高い数値，**赤**が二番目に高い数値を表す．“Ours” は *Video Classifier* と *Bi-LSTM+MIL* を統合した手法を表す．“Ours†” は *Video Classifier* に双方向 LSTM を追加したモデルと *Bi-LSTM+MIL* を統合した手法を表す．

Table 4　Comparison of frame-level AP performance on the XD-Violance dataset. **Blue** is the highest value and **red** is the second highest value. “Ours” represents the integration of *Video Classifier* and *Bi-LSTM+MIL*. “Ours†” represents the integration of a model that adds bidirectional LSTM to *Video Classifier* and *Bi-LSTM+MIL*.

Dataset	Method	Feature Type	AUC(%)
UCF-Crime	RTFM [20] +Video Classifier	I3D RGB	85.21
	RTFM [20] +Bi-LSTM+MIL		84.24
	RTFM [20] +Ours		85.11
ShanghaiTech	RTFM [20] +Video Classifier	I3D RGB	97.14
	RTFM [20] +Bi-LSTM+MIL		97.40
	RTFM [20] +Ours		97.48

表 5　既存手法と統合した場合の AUC 性能

Table 5　AUC performance when integrated with existing methods

動画の異常検出に特化した手法である Pang ら [26] の手法と比較して 0.44% 高い検出精度を示した．

4.7　既存手法との統合

Video Classifier は長期的な情報を扱う異常検出器であるが，短期的なセグメントに基づく異常検出器 *Bi-LSTM+MIL* と統合することで高精度な異常検出が可能となることを示した．しかし，これら以外の手法とも統合することが可能である．表 5 に既存手法と統合した場合の AUC 性能を示す．UCF-Crime データセットにおける実験では，RTFM [20] と *Video Classifier* との統合により，RTFM [20] 単独での検出と比較して 0.91% の精度向上を達成できた．ShanghaiTech データセットにおける実験では，RTFM と *Video Classifier* と *Bi-LSTM+MIL* の 3 つの手法の統合により，RTFM 単独で検出した場合と比較して 0.27% の精度向上を達成できた．

4.8　異常の種類ごとの識別性能

UCF-Crime データセットにおける，異常の種類別の AUC 性能

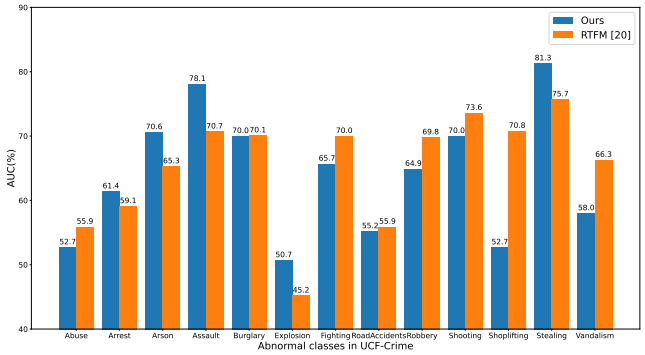

図 4　UCF-Crime データセットにおける異常種類別 AUC 性能

Fig. 4　AUC performance by anomaly classes in the UCF-Crime dataset

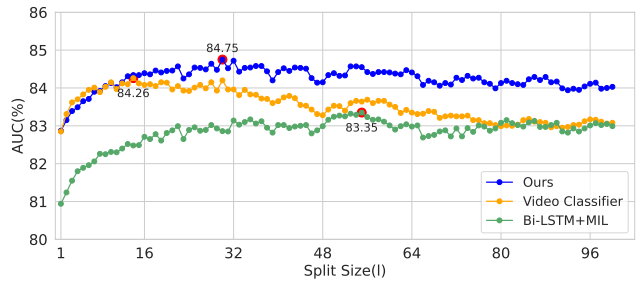

図 5　UCF-Crime データセットにおける推論時の動画分割サイズ l と AUC 性能の関係

Fig. 5　Relationship between the split size l and AUC performance on the UCF-Crime datasets during inference

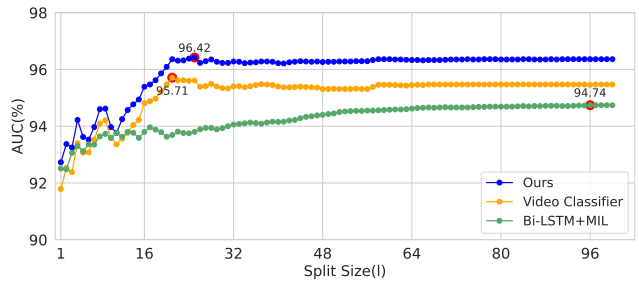

図 6　ShanghaiTech データセットにおける推論時の動画分割サイズ l と AP 性能の関係

Fig. 6　Relationship between the split size l and AP performance on the ShanghaiTech dataset during inference

を図 4 に示す．提案手法は Arrest, Arson, Assault, Burglary, Road-accident, Explosion, Stealing の 7 種類の異常において RTFM [20] と比較して高い，もしくは同等の検出精度を達成することができた．特に Arson, Assault, Explosion, Stealing の 4 種類については 5% 近い検出精度の向上を達成できた．これら 4 種類は，物体や人物の動きを長期的に解析できなければ検出できない異常であるため，この結果は提案手法が上手く長期的な特徴を捉えられたことを示しているのではないかと推察する．また，Abuse, Burglary, Robbery の 3 種類については RTFM と同等の精度となった．Shoplifting, Vandalism の 2 種類については RTFM に劣る結果となった．提案手法は長期的な分類と短期的な分類を統合するため，これら 2 種類に見られるような瞬間的

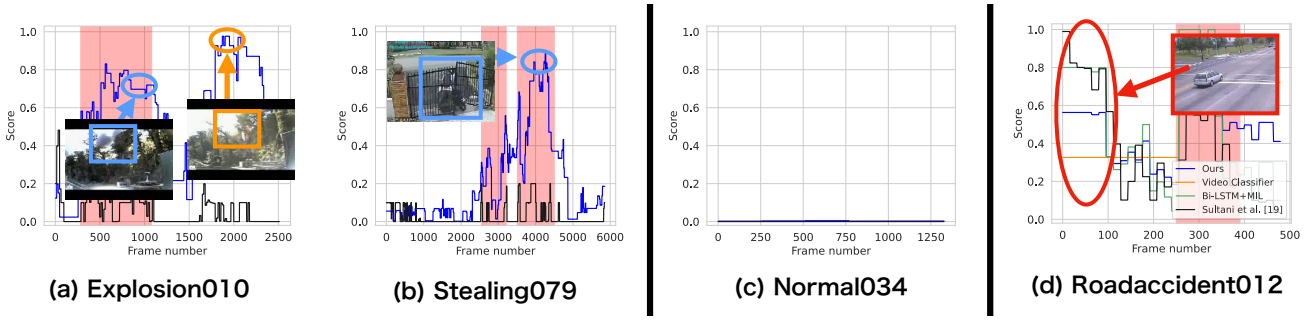

図 7　UCF-Crime データセットにおける提案手法と既存手法の異常スコアの可視化．青い線が提案手法，黒い線が Sultani ら [19] の手法の異常スコア．赤いブロックは正解を示す．青い丸は正しく予測されたフレームを，赤の丸は誤って予測されたフレームを，オレンジの丸は誤って付けられているラベル部分を示す．

Fig. 7　Visualization of anomaly scores of the proposed method and the existing method on the UCF-Crime dataset. The blue line shows the anomaly scores of the proposed method, and the black line shows the anomaly scores of Sultani et al.'s method [19]. The red blocks are ground truths of anomalous events. Blue circles indicate correctly predicted frames, red circles indicate incorrectly predicted frames, and orange circles indicate incorrectly labeled frames.

な異常に対しての分類が難しいのではないかと推察する．また，RTFM は人物の動きの特徴を上手く捉えられる手法であることが分かる．

4.9　分割サイズの変化と検出精度への影響

我々は 3.2 章で分割サイズ l を定義した．提案手法は推論時に N 個の特徴ベクトルを $m = N/l$ 個のバッグに分けるため，各バッグには l 個の特徴ベクトルが入っている．分割サイズ l の値によって検出精度が変化するので，UCF-Crime データセットと ShanghaiTech データセットを用いて，l の変化が検出精度にどのように影響を与えるかを調査した．UCF-Crime データセットにおける結果を図 5 に，ShanghaiTech データセットにおける結果を図 6 に示す．赤で囲まれた点は各手法における最も高い検出精度を表す．*Video Classifier*，*Bi-LSTM+MIL* 共に，l が 20 前後になるまで検出精度が上がっている．どちらの手法も時間的な特徴量を必要としており，ある程度の分割サイズが必要なことがわかる．*Video Classifier* だけに注目すると，どちらのデータセットにおいても最も検出精度の高い分割サイズにおいて既存の最先端の手法と同等の検出精度となっている．*Bi-LSTM+MIL* だけに注目すると，どちらのデータセットにおいても l がある程度大きいときに最も高い検出精度を示している．これは LSTM を用いることで長期的な時間的特徴を捉えることができていることを示している．2 つのモデルを統合した手法の結果に注目すると，分割サイズ l がある程度大きいときは，分割サイズ l の値に関わらず高い精度で異常が検出できることがわかる．2 つのモデルの統合を行うことで，分割サイズ l の値に左右されない，安定した異常検出器が得られていることが分かる．

4.10　定 性 分 析

UCF-Crime データセットにおいて，提案手法が予測した異常スコアを図 7 に示す．比較のために同じ特徴量を用いて学習した Sultani ら [19] のモデルによる異常スコアを示した．また，図 7-d については *VIdeo Classifier* モデル単体，*Bi-LSTM+MIL* モデル単体による異常スコアも示す．提案手法は長期的な分類と短期的な分類を同時に行なっているため，図 7-a の動画のようなセグメント単位では雲にしか見えないような爆発や，図 7-b の動画のような，実際には窃盗行為なのだが，一見すると所有者がバイクを出そうとしているようにしか見えないような異常を検出することに成功している．また，図 7-d の動画には大量の車が急にフレームインしてくるような状況があるが，MIL に基づく *Bi-LSTM+MIL* モデルや Sultani らの手法では異常として検出してしまっている．一方で，*Video Classifier* モデルは正常と認識できている．提案手法は 2 つの手法の統合によって，こういった場合の異常スコアを抑えることができることを示している．

5.　結　　論

我々は，self-attention 機構を用いて動画の長期的な特徴を解析して学習する *Video Classifier* モデルを提案した．*Video Classifier* は MIL に基づく手法ではなく，実装が簡単で，既存手法と比べてハイパーパラメータも少なく扱いやすいモデルである．シンプルなモデルにも関わらず，最先端の手法と比較して遜色ない検出精度が達成できることを示した．また，短期的なセグメントに基づく異常検出器である *Bi-LSTM+MIL* も提案し，*Video Classifier* との統合によって最先端の手法を上回る検出精度が達成できることを示した．

文　　献

[1] Wen Liu, Weixin Luo, Dongze Lian and Shenghua Gao, "Future Frame Prediction for Anomaly Detection - - A New Baseline", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[2] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh and Anton van den Hengel, "Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection", In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019

[3] Trong Nguyen Nguyen and Jean Meunier, "Anomaly Detection in Video Sequence with Appearance-Motion Correspondence", In

IEEE/CVF International Conference on Computer Vision (ICCV), 2019

[4] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu and Zhiwei Yang, "Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision", In *European Conference on Computer Vision (ECCV)*, 2020

[5] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K, Roy-Chowdhury and Larry S Davis,"Learning temporal regularity in video sequences",In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2016

[6] J. Zhang, L. Qing, J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection",In *IEEE International Conference on Image Processing (ICIP)*, 2019

[7] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li and Ge Li, "Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019

[8] Yi Zhu and Shawn Newsam, "Motion-aware feature for improved video anomaly detection",In *The British Machine Vision Conference (BMVC)*, 2019

[9] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid and Seung-Ik Lee, "CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection", In *European Conference on Computer Vision (ECCV)*, 2020

[10] Shaojie Bai, J. Zico Kolter and Vladlen Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling", In *arXiv preprint arXiv*:1803.01271, 2018

[11] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury and Larry S. Davis, "Learning Temporal Regularity in Video Sequences", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2016

[12] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin and Marius Kloft, "Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events", In *the 28th ACM International Conference on Multimedia (ACM MM '20)*, 2020

[13] Boyang Wan, Yuming Fang, Xue Xia and Jiajie Mei, "Weakly Supervised Video Anomaly Detection via Center-guided Discriminative Learning", In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2020

[14] A. Basharat, A. Gritai and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008

[15] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery", In *International Conference on Information Processing in Medical Imaging*, 2017

[16] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon and Tomas Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[17] Joao C and Andrew Z, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017

[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 *The International Conference on Learning Representations (ICLR)*, 2017

[19] Sultani W, Chen C and Shah M, "Real-world anomaly detection in surveillance videos", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018

[20] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans and Gustavo Carneiro,"Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning", In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", In *Journal of Machine Learning Research*, 2014

[22] Jia-Chang Feng, Fa-Ting Hong and Wei-Shi Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection", In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021

[23] Weixin Luo, Wen Liu and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework", In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai and Soumith Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", In *Advances in Neural Information Processing Systems 32*, 2019

[25] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao and Jiawei Han, "On the Variance of the Adaptive Learning Rate and Beyond", In *International Conference on Learning Representations (ICLR)*, 2020

[26] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu and Yan-Xiong Li, "VIOLENCE DETECTION IN VIDEOS BASED ON FUSING VISUAL AND AUDIO INFORMATION", In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021

[27] Prajit Ramachandran, Barret Zoph and Quoc V. Le, "Swish: a Self-Gated Activation Function", In *arXiv preprinet arXiv*:arXiv:1710.05941v1, 2020

[28] Hitesh Sapkota, Yiming Ying, Feng Chen and Qi Yu, "Distributionally Robust Optimization for Deep Kernel Multiple Instance Learning", In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*,2021

[29] 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.