時空間敵対的生成ネットワークにおける U-Net Discriminator の中間層特徴の融合による動画像異常検知

○ 橋本 慧志[1]，工藤 謙一[2]，高橋 孝幸[3]，梅田 和昇[4]

○ Satoshi HASHIMOTO[1], Kenichi KUDO[2], Takayuki TAKAHASHI[3] and Kazunori UMEDA[4]

1:中央大学大学院理工学研究科，hashimoto@sensor.mech.chuo-u.ac.jp
2:中央大学研究開発機構，kkudo32h@tamajs.chuo-u.ac.jp
3:プリマハム株式会社開発本部，takayuki.takahashi@primaham.co.jp
4:中央大学理工学部，umeda@mech.chuo-u.ac.jp

＜要約＞　近年，深層学習を用いた日常生活や製造業の現場等における映像の異常を捉える試みが検討されている．特に，敵対的生成ネットワーク（GAN: generative adversarial networks）の活用が盛んであるが，非効率的である問題や，単純な差分ベースの手法ではノイズの影響を受ける問題がある．本稿では，時空間敵対的生成ネットワークを用いた新たな教師なし学習による動画像異常検知手法を提案する．提案手法は U-Net Discriminator の中間層特徴を活用するため，Discriminator の注視領域に限定した効率的かつ高精度な異常検知が可能である．

＜キーワード＞ 教師なし学習，敵対的生成ネットワーク，異常検知，動画像

1. 序論

近年の深層学習の発展に伴い，日常生活における異常をとらえる動画像異常検知に関する研究が盛んに行われている[1-7]．異常検知においては一般に異常な事象の発生が希有であるため教師データの収集が困難である．そのため，正常データのみを用いた教師なし学習を行い，獲得した分布から逸脱したものを異常と定義するアプローチがよく用いられる．動画像の異常検知は，従来ではよく HOG 特徴などのハンドメイドな特徴に基づき識別器を教師なしで学習する手法や，スパース表現を学習する手法を用いて行われていた．しかしこうした手法では，動画像の異常検知というチャレンジングなタスクを十分に解決することはできていなかった．近年では動画像異常検知の手法は主に時空間ネットワーク（STN: spatio temporal networks）を用いた手法[8-12]と，appearance 特徴と motion 特徴に分離してモデル化する手法[13-17]の 2 つに大別される．前者は，入力の動画像を再構成する Encoder-Decoder 型のモデルを基本とする．Luo ら[9]は，STN を用いて，Encoder-Decoder ベースの動画像の再構成誤差による異常検知手法を提案している．後者は，pix2pix[18]を用いて appearance 特徴と motion 特徴間のドメイン変換を学習する．特にこちらは大きく成果を上げている．Ravanbakhsh ら[13]は，pix2pix を用いて optical flow と frame 画像間の関係性をモデル化して異常検知を行っている．また，最近の手法では敵対的生成ネットワーク（GAN: generative adversarial networks）の活用が盛んであり[13-17]，動画像異常検知の精度向上に貢献している．しかし，こうした既存手法の多くに共通して以下の 2 点の課題がある．1 つ目は，効率的でない点である．STN を用いた手法は動画像の再構成モデルであり，入力と等しいタイムステップ長を有する動画像出力を得るが，推論時には直近の frame 画像のみで検知する場合が多い．また，GAN を用いた手法の多くは推論時に Discriminator を無視する[19]ため，この点からも効率的とは言えない．2 つ目はノイズの問題である．単純な frame 画像全体の差分ベースで異常検知を行う場合，差分時に発生するノイズの影響を受けやすく，異常領域以外にもノイズが混入することで，異常検知の性能が低下することが懸念される．本稿では，以上の背景を踏まえ，時空間敵対的ネットワークを用いた新たな動画

像異常検知手法を提案する．提案手法は，動画像入力に対する frame 予測型のモデルであり，従来手法では無視される Discriminator の中間層特徴を活用する．Discriminator の注視領域を融合することでノイズの影響を軽減し，効率的かつ高精度な異常検知が可能である．本稿の貢献は次のとおりである．

- frame 予測型の時空間敵対的生成ネットワークを構築し，効率的な異常検知手法を確立する[7].
- U-Net Discriminator の中間層特徴の効率的な活用により，差分時のノイズの問題を改善する．

本稿では UCSD データセット[5]と Avenue データセット[33]を用いて SoTA との比較を行い，その有効性を確認した．特に，中間層特徴を融合することで異常検知の性能が大きく向上した．

以下では，最初に関連技術について示す．次に，提案手法を示す．さらに，検証実験について示し，最後に結論と今後の展望を述べる．提案手法の概要を図 1 に示す．手法の詳細は 3 章で示す．

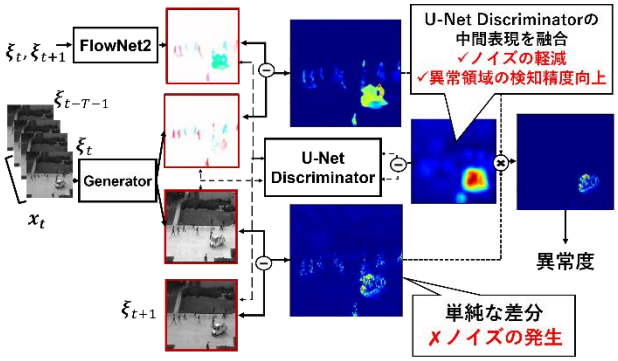

図 1 提案手法概要．異常検知モデルは Generator と U-Net Discriminator からなり，正常データのみの教師なし学習を行う．Generator は入力動画像の 1frame 先の画像と optical flow を推定する．異常検知には，Generator の推定結果と真値との差分画像に対して U-Net Discriminator の中間特徴を重みづけしたマップを用いる．単純な差分画像にはノイズが散見されるが，中間特徴を融合することで異常領域の検知精度の向上が期待できる．

2. 関連技術

2.1 敵対的生成ネットワーク

敵対的生成ネットワーク（GAN: generative adversarial networks）は，Goodfellow[20]らによって提案された生成モデルの一つである．GAN は，生成器（Generator）と識別器（Discriminator）の 2 つのモデルからなり，互いに騙し合うように学習する．具体的には，式（1）に示す最小最大化問題を最適化することで，学習データの分布p_xに一致するように生成分布p_gを獲得する．

$$min_{Gen} max_{Dis} \ \mathbb{E}_{x\sim p_x} log[Dis(\boldsymbol{x})] + \mathbb{E}_{z\sim p_z} log[1 - Dis(Gen(\boldsymbol{z}))] \quad (1)$$

ここで，Genは Generator，Disは Discriminator，$\boldsymbol{x}$は入力データ，$\boldsymbol{z}$は潜在空間からサンプリングされるノイズである．Generator はノイズ$\boldsymbol{z}$を入力としてデータの分布p_xに存在するようなデータ$Gen(\boldsymbol{z})$を生成する．一方，Discriminator はデータの分布p_xに実在する$\boldsymbol{x}$もしくは Generator により生成された$Gen(\boldsymbol{z})$を入力として，それぞれが本物か偽物かを識別する．さらに，Radford ら[21]によって Deep Convolutional Generative Adversarial Networks（DCGAN）が提案され，高品質な画像生成が可能となった．DCGAN は Generator 及び Discriminator に Convolutional Neural Networks（CNN）を採用し，各層に Batch Normalization を用いる等の特徴を有しており，これにより単純な GAN よりも高品質で高解像度な画像を生成することが可能となった．

2.2 U-Net GAN

U-Net GAN は，Schonfeld ら[22]によって提案された手法である．図 2 に手法の概要を示す．GAN における Discriminator に U-Net[23]を用いる点が最大の特徴である．典型的な GAN においては，Discriminator は入力画像全体に対してその真偽を一意に判定する識別モデルとして用いられる．しかし，U-Net GAN では Discriminator の出力は入力と同じサイズを有しており，pixel 単位で入力の真偽を識別する．つまり，画像の局所的な再現品質を Generator にフィードバックすることができる．U-Net GAN は画像の生成タスクにおける SoTA な手法である．

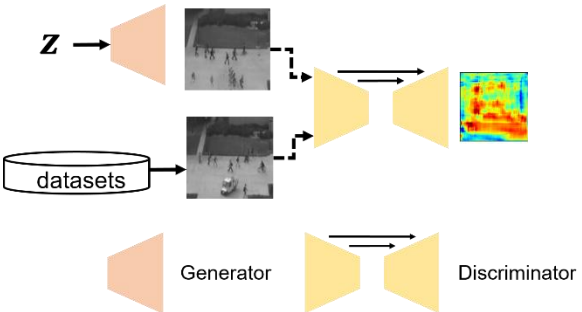

図 2 U-Net GAN[22]

2.3　動画像異常検知

動画像の異常検知においても GAN の活用は盛んである. 図 3 に示す Ravanbakhsh らの手法[13]は, 動画像をそのままモデル化して再構成ベースの異常検知を行う STN を用いた手法とは対照的に, optical flow と frame 画像間のドメイン変換を pix2pix の枠組みで学習する. optical flow O を frame F に変換する Generator を $G^{O\to F}$, その逆を $G^{F\to O}$ として, この 2 つの Generator からの出力 $\hat{F}$, $\hat{O}$ に対してそれぞれ F, O 間の差分を求め, 最終的に融合することで, 異常検知をしている. また, F の差分算出に関しては, Naïve に pixel level で差分をとるのではなく, AlexNet[24]の中間表現を用いている. これは, 単純な pixel 単位の差分ベースで異常マップを算出すると意味的な情報量が少ないことが経験則的に確認されていることに起因する. 図 4 に示す Liu らの手法[16]は, pix2pix を frame 予測に応用している. Generator は入力の複数フレーム $F_1, F_2, \dots, F_t$ に対してその最終フレームの 1 つ先 F_{t+1} を予測する. さらに, 真値 F_{t+1} と予測結果 $\widehat{F_{t+1}}$ それぞれに対して Flownet[25]を用いて F_t との optical flow を推論し, その差分が一致するように学習時に制約を課している. 推論時には, フレームの予測誤差を用いる.

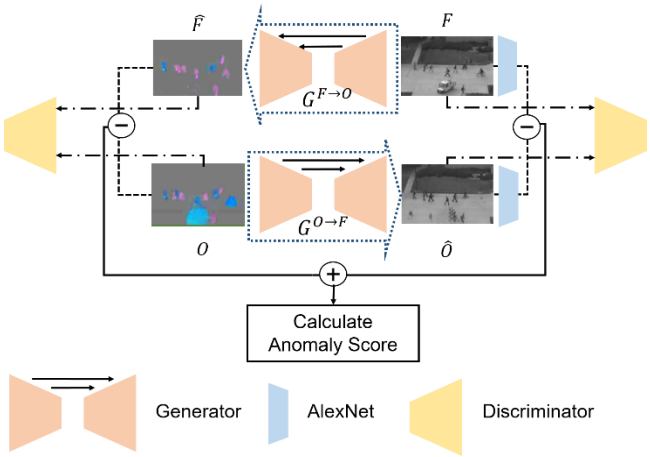

図 3 Ravanbakhsh[13] らの手法

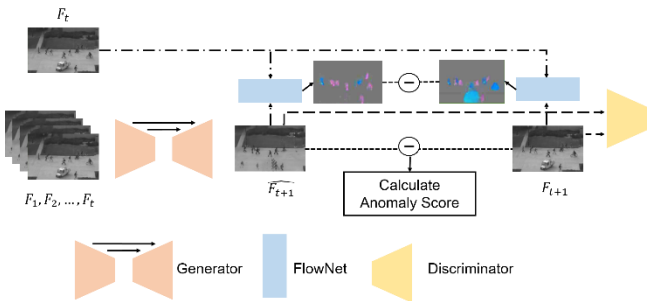

図 4 Liu らの手法[16]

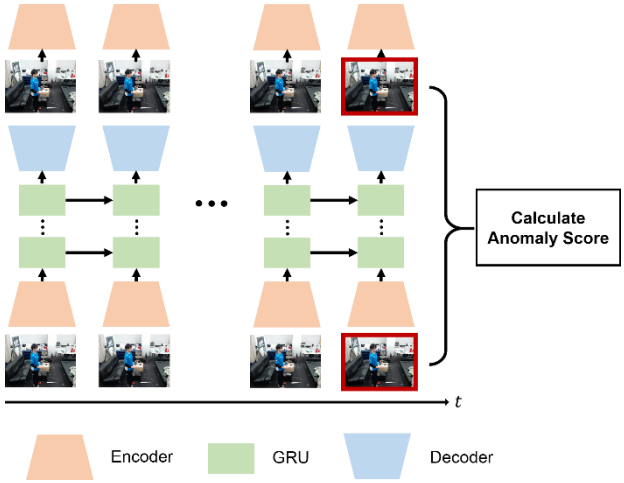

図 5 動画像再構成モデルの例[8]

しかし, こうした既存手法の多くに共通して課題が存在する. 一つは効率性である. STN を用いた手法は動画像の再構成モデルであり, 図 5 に示すように推論時には直近の frame 画像のみで検知する場合が多い. また, GAN を用いた手法の多くは推論時に Discriminator を無視している. pix2pix 等の GAN を用いた手法においては当然 Discriminator が必要であるが, 推論時には無視されるため, 異常の算出には用いられていない[19]. もう一つはノイズの問題である. 単純な frame 画像全体の差分ベースで異常検知を行う場合, 差分時に発生するノイズの影響を受けやすく, 性能が低下することが懸念される.

3.　提案手法

3.1　概要

提案手法では, 動画像を効率的にモデル化するため, frame 予測型の時空間敵対的生成ネットワークを用いる. Liu ら[16]は, 異常検知とは, 期待されていない事象の識別であるため, 過去の動画フレームから将来の動画フレームを予測し, その予測値と真値とを比較して異常検知を行うのが自然であると主張しており, 我々もこのアイデアを踏襲する. 提案手法は入力の動画像に対して 1frame 先の画像を予測し, その結果と真値との差分を用いて異常検知を行う. さらに, U-Net Discriminator[23]の中間層特徴を融合することで効率的な異常検知手法を確立する. Discriminator は画像の真偽を識別するモデルであるが, 異常検知すべき領域は偽に近いと考えられ, その中間層特徴は異常な領域を注視すると考えられる. これを融合することで単純な差分画像に発生しうるノイズの影響を回避することが期待できる.

3.2　提案モデル

我々のモデルは, 図 6 に示す通り, Encoder, Decoder O, Decoder F, Discriminator の 4 つからなる．Encoder と Decoder O，Decoder F からなるモデルをそれぞれ Generator O(G_O)，Generator F(G_F) と定義する．Encoder は入力の動画像から畳み込み層と Convolutional LSTM[26] を用いて特徴を抽出し，Decoder F はそれを用いて逆畳み込み層により frame 画像を予測し, Decoder O は optical flow を予測する．G_F の構造には, Liu ら[16]のように U-Net を用いることも選択できるが，U-Net の持つ skip 構造により，入力に含まれる異常な情報が伝播してしまう可能性が危惧される．そのため，Lee ら[27]が採用したモデル構造を参考に，予測型のモデルへと拡張した．一方，G_O には U-Net を採用し，典型的な条件付き GAN の戦略に従った構造とした．これは，G_O が解くべき問題が画像変換タスクであり，条件付き GAN は[18]で実証されたように画像変換に適しているからである．Discriminator には Schonfeld ら[23]により提案された U-Net Discriminator を採用し，真の optical flow か Decoder により予測された optical flow かを pixel-level と frame-level とで識別する．これは，U-Net Discriminator がマルチレベルで真偽判定を行うほか, 後述する Consistency Regularization を採用しているため，既存手法と比較したときに安定的な学習となることが期待でき，モデルの予測品質の向上に寄与すると考えるからである．

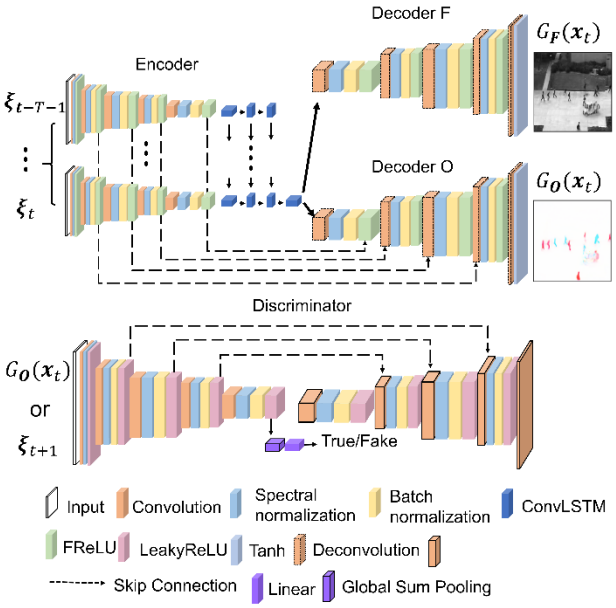

図 6　モデル構造．上段：Generator，下段：U-Net Discriminator.

3.3　学習フェーズ

我々の時空間敵対的生成ネットワークは，以下の 2 つの損失 L_G，L_D を交互に最小化することで最適化される．

$$L_G = \lambda_f L_{frame} + \lambda_o L_{opt} + L^G_{D_{enc}} + L^G_{D_{dec}} \quad (2)$$

$$L_D = L_{D_{enc}} + L_{D_{dec}} + \lambda_c L_{consist} \quad (3)$$

G は Generator，D_{enc}, D_{dec} はそれぞれ Discriminator の Encoder module，Decoder module を表す．また，L_{frame}, L_{opt} はそれぞれ真値と予測結果間の画像，optical flow の予測損失，λ_f, λ_o は予測損失に対する重みづけの定数である．λ_c は Consistency Regularization に対する重みづけの定数である．
L_G の各項は以下のとおりである．

$$L_{frame} = \|\boldsymbol{\xi_{t+1}} - G_F(\boldsymbol{x_t})\|_1 \quad (4)$$

$$L_{opt} = \|\boldsymbol{o_{t+1}} - G_O(\boldsymbol{x_t})\|_1 \quad (5)$$

$$L^G_{D_{enc}} = -\mathbb{E}_{\boldsymbol{\xi}\sim \boldsymbol{p_\xi}}\left[log\left(1 - D_{enc}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])\right)\right] - \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_x}}[log(D_{enc}([G_F(\boldsymbol{x_t}), \boldsymbol{o}_{t+1}]))] \quad (6)$$

$$L^G_{D_{dec}} = \sum_{i,j} log[\,1 - D_{dec}([\boldsymbol{\xi}_t, \boldsymbol{o}_{t+1}])]_{i,j}] + \sum_{i,j} log\,[D_{dec}([G_F(\boldsymbol{x_t}), \boldsymbol{o}_{t+1}])]_{i,j}] \quad (7)$$

入力 $\boldsymbol{x_t}$ はある時刻 t における固定長 T を有する部分時系列 $\boldsymbol{\xi}_{t-T-1}, \boldsymbol{\xi}_{t-T-2}, \ldots, \boldsymbol{\xi}_t$ から構成される．$\boldsymbol{\xi}$ は各フレームの画像である．$\boldsymbol{o}_{t+1}$ は時刻 $t+1$ における optical flow である．$[D_{dec}(\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1})]_{i,j}$ および $[D_{dec}(G_F(\boldsymbol{x_t}), \boldsymbol{o}_{t+1})]_{i,j}$ は，ピクセル (i,j) における Discriminator の出力結果を表す．$[\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]$ は，$\boldsymbol{\xi}_{t+1}$ と $\boldsymbol{o}_{t+1}$ のチャンネル方向の結合を意味する．さらに，L_D の各項は以下のとおりである．

$$L_{D_{enc}} = -\mathbb{E}_{\boldsymbol{\xi}\sim \boldsymbol{p_\xi}}\,[\,log(D_{enc}([\boldsymbol{\xi}_t, \boldsymbol{o}_{t+1}]))] - \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}}[log(1 - D_{enc}([G_F(\boldsymbol{x_t}), \boldsymbol{o}_{t+1}])] \quad (8)$$

$$L_{D_{dec}} = -\mathbb{E}_{\boldsymbol{\xi}\sim \boldsymbol{p_\xi}}[\sum_{i,j} log\,[D_{dec}([\boldsymbol{\xi_t}, \boldsymbol{o}_{t+1}])]_{i,j}] - \mathbb{E}_{\boldsymbol{x}\sim \boldsymbol{p_x}}[\sum_{i,j} log\,[1 - D_{dec}([G_F(\boldsymbol{x_t}), \boldsymbol{o}_{t+1}])]_{i,j}] \quad (9)$$

$$L_{consist} = ||D_{dec}\big(mix([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}], G_F(\boldsymbol{x_t}), M)\big) - mix\big(D_{dec}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]), D_{dec}\big(G_F(\boldsymbol{x_t})\big), M\big)||^2 \quad (10)$$

ここで，$L_{consist}$は[23]で導入された Cutmix[28]ベースの Consistency Regularization を表す．この正則化は，十分に訓練された Discriminator からの出力は，画像のクラス及びドメイン変換があっても等しくあるべきであるという考えに基づいている． mixは式（11）で計算できる．

$$mix([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}], [G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}], M) = M \odot [\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}] + (\mathbf{1} - M) \odot [G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}] \quad (11)$$

ここで，$M \in \{0,1\}^{W \times H}$は，画素 (i,j) が真の画像 (1) もしくは(0)かを示す 2 値マスク，$\mathbf{1}$ は 1 で満たされた 2 値マスク， $\odot$は要素毎の乗算を表す．なお，学習の最適化手法は AdaBelief[29]を，Generator の各層の活性化関数は FReLU[30]，Discriminator のそれは LeakyReLU を用いる．また，学習を安定的に行うために，Generator と Discriminator の各層に Spectral normalization[31] を導入する． optical flow は FlowNet2[32]により推定する．

3.4 推論フェーズ

次に，推論フェーズについて述べる．入力$\boldsymbol{x}_t$に対する異常度$a(\boldsymbol{x}_t)$を式(12)のように定義する．

$$a(\boldsymbol{x}_t) = |||\boldsymbol{\xi}_{t+1} - G_F(\boldsymbol{x}_t)| \odot |\boldsymbol{o}_{t+1} - G_O(\boldsymbol{x}_t)| \odot D_{dec}([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}])\boldsymbol{map}||_1 \quad (12)$$

$$\boldsymbol{map}(\boldsymbol{N}, \dots, \boldsymbol{M}) = |F_N([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]) - F_N([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])| \odot \cdots \odot |F_M([\boldsymbol{\xi}_{t+1}, \boldsymbol{o}_{t+1}]) - F_{\mathrm{M}}([G_F(\boldsymbol{x}_t), \boldsymbol{o}_{t+1}])| \quad (13)$$

ここで，入力$\boldsymbol{x}_t$はある時刻tにおける固定長Tを有する部分時系列$\boldsymbol{\xi}_{t-T-1}, \dots, \boldsymbol{\xi}_t$から構成される． $\boldsymbol{\xi}, \boldsymbol{o}$はそれぞれ各フレームの画像，optical flow を表す．$\boldsymbol{map}$は U-Net Discriminator の Encoder，D_{enc}の任意の第$N, \dots, M$層の中間層特徴$F_N, \dots, F_M$の差を乗算し，入力のサイズにリサイズしたものである．我々がこの U-Net Discriminator の中間層特徴$\boldsymbol{map}$を融合するのは，Discriminator が注視する高レベル特徴を活用し，重みづけすることで，差分時に発生するノイズの軽減が期待できるからである．異常算出に用いる最終的なスコア$S(\boldsymbol{x}_t)$は式（14）を用いて正規化することで求められる．

$$S(\boldsymbol{x}_t) = \frac{a(\boldsymbol{x}_t)}{max(a(\boldsymbol{x}_{1 \dots m}))} \quad (14)$$

ここで，mはテストデータの総数である．これを用いて異常検知を行う．

4. 検証実験

4.1 概要

本稿では，動画像異常検知において一般的な公開データセットである UCSDped2[5]と， Avenue[33]を用いて実験を行った．図 7 にデータセットの例を示す．UCSDped2 は 16clip の訓練データ，12clip のテストデータからなる．図 7 右側に示すように，正常データは通常のスピードで歩行する様子が収録されている．一方異常データは自転車での走行，自動車の侵入などの様子が収録されている． Avenue は 16clip の訓練データ，21clip のテストデータからなる．図 7 左側に示すように，定点の監視カメラ画像を収録したものとなっており，正常データは通常のスピードで歩行する様子が収録されている．一方異常データは走る，荷物を投げる等の通常から逸脱した様子が収録されている．

これら 2 つのデータセットに対して，Frame-level の Receiver Operating Characteristic （ROC） 曲線に対する AUROC によるモデルの定量的評価を行った．なお，AUROC で評価を行うため，異常度の閾値に関する議論は行わない．

図 7 UCSDped2(左側)と Avenue(右側)の例．上段が正常データ，下段の赤色矩形領域が異常を表す．

4.2 実験設定

実験で用いたハイパパラメータを示す. Generator, Discriminator の学習率はそれぞれ 2e-4, 2e-5, タイムステップ T は 4，バッチサイズは 1 とした．画像はすべてグレースケールに変換した上，256×256 にリサイズした． 予測損失の重みλ_f, λ_oはそれぞれ 100,200 とし，Consistency Regularization の重みλ_cは 10 とした．演算には NVIDIA GeForce TITAN GPU を用い，実装には深層学習ライブラリの PyTorch を用いた．

4.3 UCSDped2 の結果

表 1 に UCSDped2 に対する定量的結果を示す．また，図 8 にモデルの入出力とその差分画像，中間層特徴とそれらを融合した異常マップを示す．AUROC を用いた定量的評価により，従来手法よりも AUROC の値が向上したことから提案手法の有効性を確認した．特に，U-Net Discriminator の中間層特徴***map***を融合することで，大きく性能が向上したことが確認できた．また，図より，***map***を融合することで Naïve な差分画像のノイズを軽減し，異常個所を重視した異常マップを得られることが定性的にも確認できた．

表 1 各データセットの結果．w/o ***map*** は Frame の差分画像のみの結果，w/ ***map*** は式(12)に示すように ***map*** を融合して得られた結果である．

Method	AUROC↑	
	UCSDped2	Avenue
Luo et al.[9]	0.922	0.817
Ravanbakhsh et al.[13]	0.935	N/A
Liu et al.[16]	0.951	0.851
Nguyen et al.[17]	0.962	0.872
Ours w/o ***map***	0.688	0.719
Ours w/ ***map***(1)	0.947	0.884
Ours w/ ***map***(2,3)	**0.964**	**0.894**

4.4 Avenue の結果

表 1 に Avenue に対する定量的結果を示す. 図 9 にモデルの入出力とその差分画像，中間層特徴とそれらを融合した異常マップを示す． Avenue に対しても，提案手法の有効性を定量的及び定性的に確認した．しかし，図 9 の 2 列目を見るに，UCSDped2 の結果を比較するとモデルの optical flow の予測精度が悪いことがわかる．また，2 列目下段の差分画像では，1 列目上段に示す赤色矩形の異常領域以外の領域においても差分値が大きくなっていることが分かる．optical flow の予測精度が低下した要因としては，Avenue が一定時間立ち止まる人間の様子を収録しているため，UCSDped2 と比較するとより複雑で難しい動きを含んでいるからと考えられる．

しかし，提案手法の中間層特徴の融合により，最終的な異常マップにおいてはノイズが除去できているため, 性能に大きく影響はないものと考えられる．

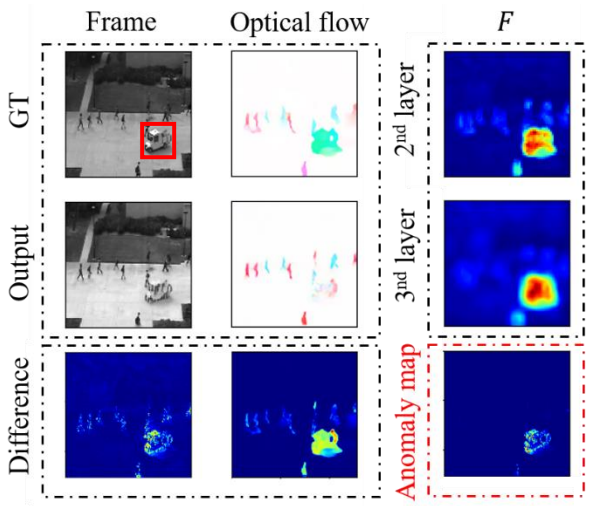

図 8 UCSD の結果．1 列目は Frame 画像であり，上段から真値，予測結果，それら 2 つの差分画像である．2 列目は optical flow の真値，予測結果，それら 2 つの差分画像である．3 列目は，上段からそれぞれ第 2 層，第 3 層の中間層特徴の差分である．最後の段は融合により得られる最終的な異常マップである．

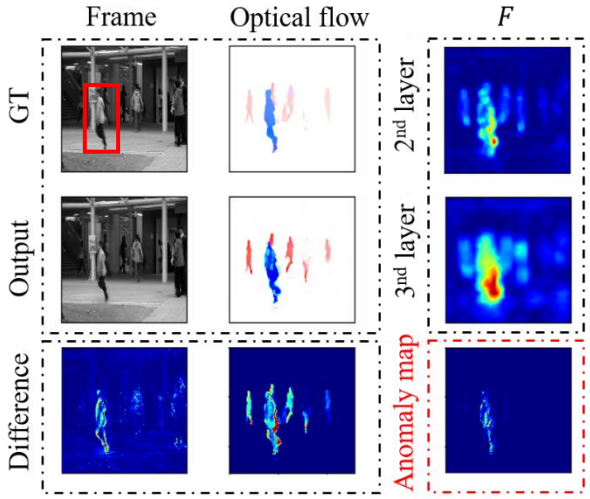

図 9 Avenue の結果．各列と段は図 8 と同様.

5. 結論

本稿では，時空間敵対的生成ネットワークを用いた新たな動画像異常検知手法を構築した．提案手法は，Discriminator の中間層特徴を活用するため，その注視領域に限定した効率的かつ高精度な異常検知を可能とした．UCSDped2，Avenue データセットに対して，AUROC を用いた定量的な評価を行い，既存手法を上回る結果を確認した. 今後の展望として，pixel-level での異常検知性能の検証や, 産業での応用可能性を検討している．

参考文献

[1] Waqas Sultani, et al., “Real-world Anomaly Detection in Surveillance Videos,” CVPR, 2018.

[2] Yaxiang Fan, et al., “Video Anomaly Detection and Localization via Gaussian Mixture Fully Convolutional Variational Autoencoder,” arXiv, 2018.

[3] Guansong Pang, et al., “Self-trained Deep Ordinal Regression for End-to-End Video Anomaly Detection,” CVPR, 2020.

[4] Mahmudul Hasan, et al., “Learning Temporal Regularity in Video Sequences,” CVPR, 2016.

[5] Vijay Mahadevan, et al., “Anomaly detection in crowded scenes,” CVPR, 2010.

[6] Radu Tudor Ionescu, et al., “Unmasking the abnormal events in video,” ICCV, 2017.

[7] 橋本 慧志, 工藤 謙一, 高橋 孝幸, 梅田 和昇, “時空間敵対的生成ネットワークを用いた教師なし学習による動画像異常検知”, ビジョン技術の実利用ワークショップ ViEW2020, IS1-13, 2020.

[8] 橋本 慧志, 工藤 謙一, 高橋 孝幸, 梅田 和昇, “GAN を活用した動画像異常検知手法の構築と労働災害防止へ向けた応用の検討”, 精密工学会画像応用技術専門委員会サマーセミナー2020, 2020.

[9] Weixin Luo, et al., “A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework,” ICCV, 2017.

[10] Weixin Luo, et al., “Remembering history with convolutional lstm for anomaly detection,” ICME, 2017.

[11] Asim Munawar, et al., “Spatio-Temporal Anomaly Detection for Industrial Robots through Prediction in Unsupervised Feature Space,” WACV, 2017.

[12] Lin Wang, et al., “Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder,” ICIP, 2017.

[13] Mahdyar Ravanbakhsh, et al., “Abnormal Event Detection in Videos using Generative Adversarial Nets,” ICIP, 2017.

[14] Mahdyar Ravanbakhsh, et al., “Training Adversarial Discriminators for Cross-channel Abnormal Event Detection in Crowds,” WACV, 2019.

[15] Hung Vu, et al., “Robust Anomaly Detection in Videos Using Multilevel Representations,” AAAI, 2019.

[16] Wen Liu, et al., “Future Frame Prediction for Anomaly Detection –A New Baseline,” CVPR, 2018.

[17] Trong Nguyen Nguyen, et al.,” Anomaly Detection in Video Sequence with Appearance-Motion Correspondence,” ICCV, 2019.

[18] Phillip Isola, et al., “Image-to-Image Translation with Conditional Adversarial Networks,” CVPR, 2017.

[19] Mohammad Sabokrou, et al., “AVID: Adversarial Visual Irregularity Detection,” arXiv, 2018.

[20] Ian J. Goodfellow, et al., “Generative Adversarial Networks,” NeurIPS, 2014.

[21] Alec Radford, et al., “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” ICLR, 2016.

[22] Edgar Schonfeld, et al., “A U-Net Based Discriminator for Generative Adversarial Networks,” CVPR, 2020.

[23] Olaf Ronneberger, et al., “U-Net: Convolutional Networks for Biomedical Image Segmentation,” MICCAI, 2015.

[24] Alex Krizhevsky, et al., “Imagenet classification with deep convolutional neural networks,” NeurIPS, 2012.

[25] Philipp Fischer, et al., “FlowNet: Learning Optical Flow with Convolutional Networks,” ICCV, 2015.

[26] Xingjian Shi, et al., “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” NeurIPS, 2015.

[27] Sangmin Lee, et al., “STAN: Spatio-Temporal Adversarial Networks for Abnormal Event Detection,” ICASSP, 2018.

[28] Sangdoo Yun, et al., “Cutmix: Regularization strategy to train strong classifiers with localizable features,” ICCV, 2019.

[29] Juntang Zhuang, et al., “AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients,” NeurIPS, 2020.

[30] Ningning Ma, et al., “Funnel Activation for Visual Recognition,” ECCV, 2020.

[31] Takeru Miyato, et al., “Spectral Normalization for Generative Adversarial Networks,” ICLR, 2018.

[32] Eddy Ilg, et al., “FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks,” CVPR, 2017.

[33] Cewu Lu, et al., “Abnormal Event Detection at 150 FPS in MATLAB,” ICCV, 2013.