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

Video Anomaly Detection Using Spatio-Temporal Adversarial Networks with Unsupervised Learning

精密工学専攻　39 号　橋本慧志

Satoshi Hashimoto

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[1-3]. 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. 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[4-7]と，appearance 特徴と motion 特徴に分離してモデル化する pix2pix[8]ベースな手法[9-13]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[4]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 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[9]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[6,9-13]，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，optical flow の時間情報量の少なさである．Optical flow は隣接フレーム間の速度情報を有するが，特定すべき異常の特性によっては適切な特徴量といえない．2 つ目はノイズの問題である．大半の手法[3-7,9-13]は Naïve な frame 画像全体の差分ベースで異常検知を行うが，データの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．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，前述の 2 つのアプローチを統合し，STN を用いて任意の時間情報量を有する動画像をモデル化しつつ pix2pix の枠組みで optical flow の変換を学習する．また，従来手法では無視される Discriminator の中間層特徴を活用し，その注視領域を融合することでノイズの影響を軽減する．そのため，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．

- pix2pix ベース手法の optical flow による時間情報量の少なさを，STN を組み合わせることで改善する．
- 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[14]と Avenue データセット[15]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 Fig. 1 に示す．手法の詳細は 3 章で示す．

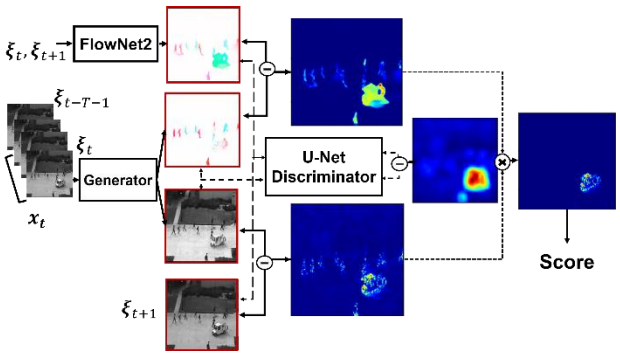

Fig. 1 Overview of the proposed method.

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[16]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．

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

ここで，Genは Generator，Disは Discriminator，xは入力データ，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$Gen(z)$を生成する．一方，Discriminator はデータの分布p_xに実在するxもしくは Generator により生成された$Gen(z)$を入力として，それぞれが本物か偽物かを識別する．

2.2 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．Fig. 2 に示す Ravanbakhsh らの手法[9]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[17]の中間表現を用いている．Fig. 3 に示す Liu らの手法[12]は，pix2pix を frame 予測に応用している．Generator は入力の複数フレーム$F_1, F_2, \ldots, F_t$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[18]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．

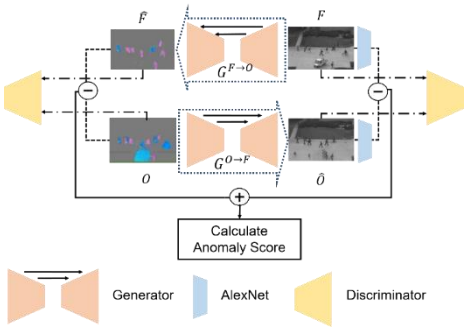

Fig. 2 The method of Ravanbakhsh et al. [9]

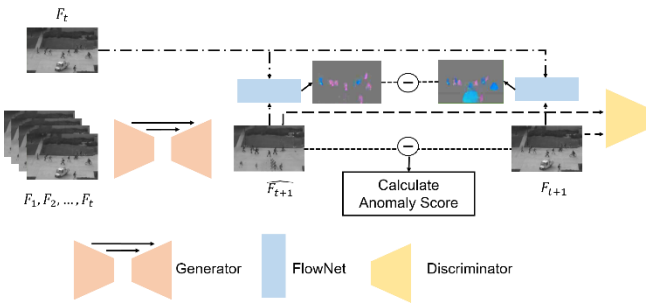

Fig. 3 The method of Liu et al. [12]

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，optical flow の時間情報量の少なさである．成果を上げている pix2pix ベースの手法で用いる Optical flow は隣接フレーム間の速度情報を有するが，特定すべき異常の特性によっては中長期の時間情報を扱える方が望ましい．2 つ目はノイズの問題である．大半の手法[3-7,9-13]は Naïve な frame 画像全体の差分ベースで異常検知を行うが，データのノイズの影響を受けやすく，異常検知の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，STN を用いて動画像をモデル化しつつ optical flow の変換を学習する．また，Liu ら[12]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 先の画像と optical flow 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[19]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．

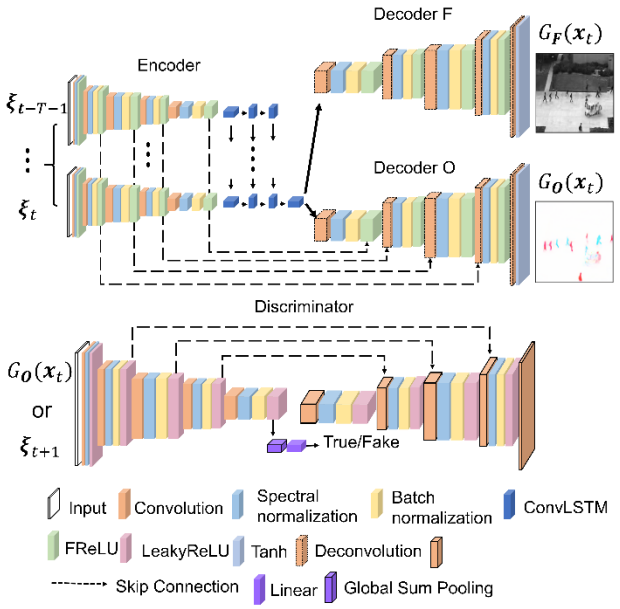

Fig. 4 Overview of our model. The upper figure shows the generator and the lower figure shows the U-Net discriminator.

3.2 提案モデル

我々のモデルは，Fig. 4 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[20]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[12]のように U-Net[21]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する可能性が危惧される．そのため，Lee ら[6]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，Discriminator には Schonfeld ら[19]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．

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．

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

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

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

ここで，$L_{consist}$は[19]で導入された Cutmix[22]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[23]を用いる．optical flow は FlowNet2[26]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．

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

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

$\boldsymbol{map}$は U-Net Discriminator の Encoder，D_{enc}の任意の第$N, \ldots, M$層の中間層特徴$F_N, .., F_M$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．

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

ここで，mはテストデータの総数である．

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[14]と，Avenue[15]を用いて実験を行った．Fig. 5 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．Fig. 5 左側に示すように，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．Fig. 5 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．

Fig. 5 Examples of UCSDped2(left)and Avenue(right). The red rectangle means abnormal region.

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 実験結果

Table 1 にそれぞれのデータセットに対する定量的結果を示す．また，Fig. 6 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$\boldsymbol{map}$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，Fig. 6 より，$\boldsymbol{map}$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確認できた．

Table 1 Results of each dataset.

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**

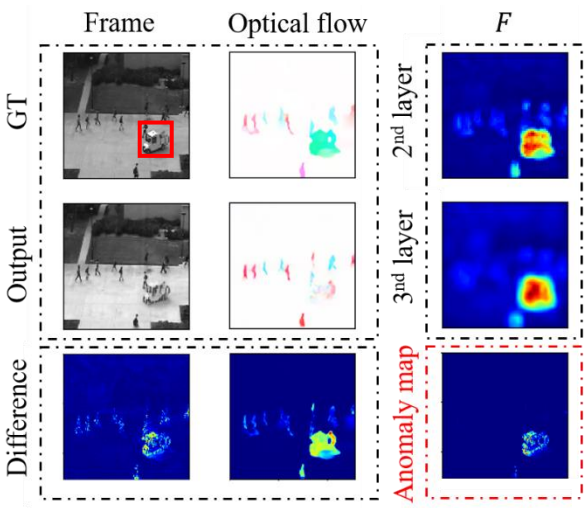

Fig. 6 Results of UCSD. The first row is the Frame image, which consists of the true value, the prediction result, and their difference image. The second row is the optical flow true value, prediction result, and their difference image. The third row shows the differences of the middle layer features of the second and third layers. The red rectangle area shows the final anomaly map obtained by fusing ***map***.

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, ノイズと時間情報量の課題に着目し, 従来手法では無視されていた Discriminator の中間層特徴の活用, pix2pix ベース手法と STN ベース手法を統合することでこれらを改善した. 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, Chen Chen, Mubarak Shah, Real-world anomaly detection in surveillance videos, In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6479-6488, 2018.

[2] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, Xiao Bai, Self-Trained Deep Ordinal Regression for End-to-End Video Anomaly Detection. In 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12173-12182, 2020.

[3] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, Larry S. Davis, Learning Temporal Regularity in Video Sequences, In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733-742, 2016.

[4] Weixin Luo, Wen Liu, Shenghua Gao, A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework, In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 341-349, 2017.

[5] Weixin Luo, Wen Liu, Shenghua Gao, Remembering history with convolutional lstm for anomaly detection, In 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 439-444, 2017.

[6] Sangmin Lee, Hak Gu Kim, Yong Man Ro. Stan: Spatiotemporal adversarial networks for abnormal event detection, In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1323-1327, 2018.

[7] Lin Wang, Fuqiang Zhou, Zuoxin Li, Wangxia Zuo, Haishu Tan, Abnormal Event Detection in Videos Using Hybrid Spatio-Temporal Autoencoder, In International Conference on Information Processing (ICIP), pp. 2276-2280, 2018.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967-5976, 2017.

[9] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, Nicu Sebe. Abnormal event detection in videos using generative adversarial nets, In 2017 IEEE International Conference on Image Processing (ICIP), pp. 1577-1581, 2017.

[10] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, Nicu Sebe, Training adversarial discriminators for crosschannel abnormal event detection in crowds, In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1896-1904, 2019.

[11] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, Dinh Phung, Robust Anomaly Detection in Videos Using Multilevel Representations, In 2019 AAAI Conference on Artificial Intelligence, pp. 5216-5223, 2019.

[12] Wen Liu, Weixin Luo, Dongze Lian, Shenghua Gao, Future frame prediction for anomaly detection a new baseline, In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6536-6545, 2018.

[13] Trong Nguyen Nguyen, Jean Meunier, Anomaly Detection in Video Sequence with Appearance-Motion Correspondence, In 2019 IEEE International Conference on Computer Vision (ICCV), pp. 1273-1283, 2019.

[14] Vijay Mahadevan, Weixin Li, Viral Bhalodia, Nuno Vasconcelos, Anomaly detection in crowded scenes, In 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975-1981, 2010.

[15] Cewu Lu, Jianping Shi, Jiaya Jia, Abnormal event detection at 150 fps in matlab, In 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2720-2727, 2013.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In 2014 Neural Information Processing Systems (NeurIPS), pp. 2672-2680, 2014.

[17] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, Imagenet classification with deep convolutional neural networks, In 2012 Neural Information Processing Systems (NeurIPS), pp. 1097-1105, 2012.

[18] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox, Flownet: Learning optical flow with convolutional networks. In 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2758-2766, 2015.

[19] Edgar Schonfeld, Bernt Schiele, Anna Khoreva, A U-Net Based Discriminator for Generative Adversarial Networks, In 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8207-8216, 2020.

[20] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, In 2015 Neural Information Processing Systems (NeurIPS), pp. 802–810, 2015.

[21] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, In 2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234-241, 2015.

[22] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, Youngjoon Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, In 2019 IEEE International Conference on Computer Vision (ICCV), pp. 6023-6032, 2019.

[23] Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar Tatikonda, Nicha Dvornek, Xenophon Papademetris, James S. Duncan, AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients, In 2020 Neural Information Processing Systems (NeurIPS), 2020.

[24] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, Thomas Brox, FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2462-2470, 2017.