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

○橋本 慧志[1]　工藤 謙一[2]　高橋 孝幸[3]　梅田 和昇[4]
1:中央大学大学院理工学研究科 〒112-8551 東京都文京区春日 1-13-27
2:中央大学研究開発機構 〒112-8551 東京都文京区春日 1-13-27
3:プリマハム株式会社開発本部 〒300-0841 茨城県土浦市中向原 635
4:中央大学理工学部 〒112-8551 東京都文京区春日 1-13-27
hashimoto@sensor.mech.chuo-u.ac.jp

概要：　近年，日常生活や製造業の現場等における深層学習を用いた動画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，非効率的であることや，不安定性が課題となっている．本稿では，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新たな動画像異常検知手法を提案する．提案手法はフレーム予測型のモデルであり，Patch-Discriminator の高レベル特徴を活用するため，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

1. 序論

近年の深層学習の発展に伴い，日常生活や製造業の現場等における監視カメラ画像の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[1-6]．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，正常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近年では主に時空間ネットワーク（STN: spatio temporal networks）を用いた手法[7-11]と，appearance 特徴と motion 特徴に分離してモデル化する手法[12-15]の 2 つに大別される．Luo ら[7]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．Ravanbakhsh ら[12]は，pix2pix を用いて optical flow とフレーム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）の活用が盛んであり[12-15]，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する課題として以下の 3 点が挙げられる．1 つ目は非効率性である．STN を用いた手法は動画像の再構成モデルであるが，推論時には直近のフレーム画像のみで異常度を計算する場合が多い．また，GAN を用いた手法の多くは推論時に Discriminator を無視する[16]．2 つ目は不安定性である．敵対的学習を行うモデルは一般的に学習が不安定であり，学習の各段階において性能が変化する．そのため，手法の再現性に課題がある[17]．3 つ目はノイズの問題である．motion 特徴の取得において推定される optical flow にはノイズがしばしば発生することから，性能に悪影響が及ぶ．

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．本稿の貢献は次のとおりである．

- フレーム予測型の時空間敵対的生成ネットワークを構築し，従来手法と比較して効率的な異常検知手法を確立する．
- GAN を用いた手法の課題である不安定性について，各種安定化手法の実装により改善する．
- Patch-Discriminator の高レベル特徴の活用により，効率的で高精度な異常検知が可能となる．

以下では，最初に関連技術について示す．次に提案手法のフレームワークを示す．さらに，検証実験に関して示し，最後に結論と今後の展望を述べる．

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[18]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$\boldsymbol{p_x}$に一致するように生成分布$\boldsymbol{p_g}$を学習により獲得する．

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

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[19]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 においては 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 Vanilla な GAN よりも高品質で高解像度の画像を生成することが可能となった．

2.2. 関連研究

動画像の異常検知においても GAN の活用は盛んである．図 1 に示す Ravanbakhsh らの手法[12]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 とフレーム画像間のドメイン変換を pix2pix の枠組みで学習する．optical flow$O$をフレーム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との間の差分を求め，最終的に融合することで，異常検知を行う．また，Fの差分算出に関しては Naïve に pixel 単位で差分をとるのではなく，AlexNet[20]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．一方，図 2 に示す Liu らの手法[15]は，pix2pix をフレーム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[21]を用いてF_tとの間の 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．

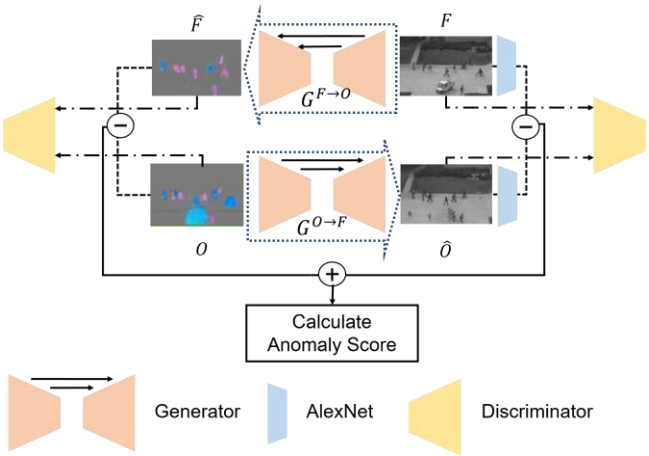

図 1 Ravanbakhsh らの手法

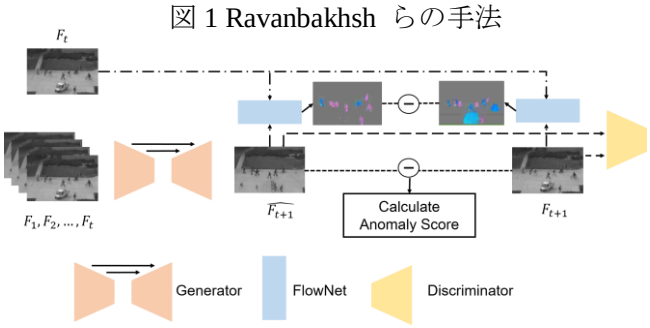

図 2 Liu らの手法

しかし，こうした既存手法の多くに共通する課題が存在する．一つ目は効率性である．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 を用いた手法の学習においては当然ながら Discriminator が必要であるが，異常の算出には用いられない．また，AlexNet や FlowNet 等の別のタスク向けの学習済みモデルを必要とするので，一つの GAN モデルで完結せず，この点からも効率的とは言えない．二つ目は不安定性である．敵対的学習は最小最大化問題を最適化する問題設定

であるが，これはそもそも複雑な問題であり，Generator と Discriminator が理想的な学習を行うことは難しい．具体的な問題として，Discriminator の損失が発散する，Discriminator 側が一方的に収束する場合がある．こうした学習の不安定性は予期できないため，学習の段階により性能が変化する可能性がある．例えば，通常の Autoencoder ベースの異常検知手法では，Autoencoder は再構成損失を最小化すればよいので，基本的に学習の推移と同時に異常検知性能も向上する．一方，GAN を用いた手法は不安定なため，学習の初期で高い異常検知性能を示すこともあれば，その逆もあり得る．ゆえに，手法の再現性という面では課題が残っている．3 つ目はノイズの問題である．動画像異常検知手法においては motion 特徴として一般的に optical flow が用いられる．しかし，optical flow の計算においてはノイズがしばしば発生することから，性能に悪影響が及ぶ恐れがある．また，optical flow を用いた手法は，動画像中の長期的な動きを含めた特徴を正確にモデル化することはできない．

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に，フレーム予測型の時空間敵対的生成ネットワークを用いる．また，敵対的学習の不安定性を改善するために各種安定化手法を用いる．さらに，Patch-Discriminator の出力を融合することで高レベルな特徴を活用し，効率的な異常検知を確立する．提案手法の概要図を図 3 に示す．

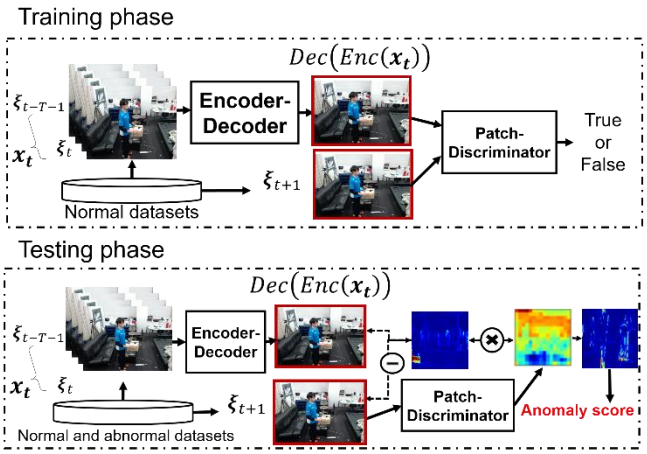

図 3 提案手法

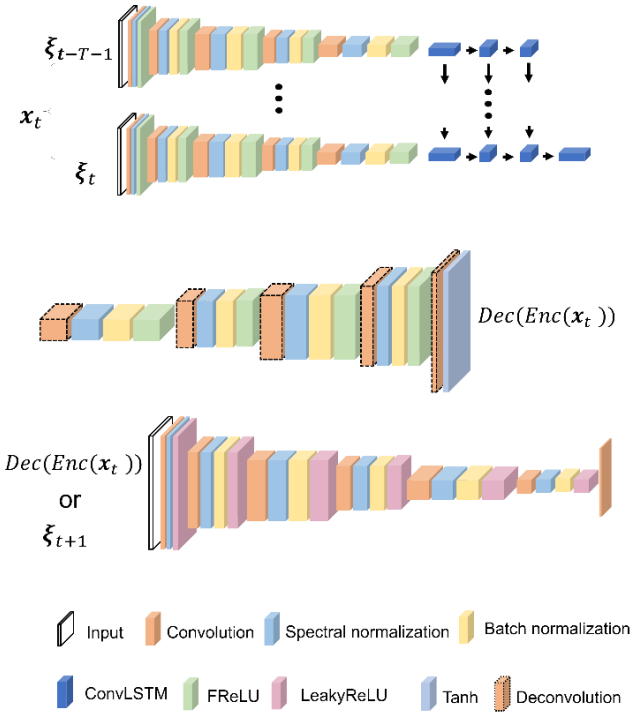

図 4 モデル詳細．上段から Encoder，Decoder，Discriminator．

3.2. モデルの学習

モデルの詳細を図 4 に示す．我々のモデルは，Encoder，Decoder，Discriminator の 3 つから構成されている．Encoder は入力の動画像から畳み込み層と Convolutional-LSTM[22]を用いて特徴を抽出し，Decoder は Encoder により抽出された特徴を用いて逆畳み込み層により動画像の 1 フレーム先のフレーム画像を予測する．Discriminator は生のフレーム画像か Decoder により予測されたフレーム画像かを識別する．ここで，Discriminator は[23]で提案された Patch-Discriminator を採用する．モデルが解くべき最適化問題は式（2）で表される．

$$min_{Enc,Dec} max_{Dis} \quad \lambda L_{recon} + L_{gan} \tag{2}$$

ここで，Encは Encoder，Decは Decoder を表す．また，L_{recon}は真値と予測結果間の予測損失，L_{gan}は敵対的損失，λは予測損失に対する重みづけの定数である．各損失は式（3）及び（4）に示す．

$$L_{recon} = \mathbb{E}_{x \sim p_x} \left\| \xi_t - Dec\left(Enc(x_t)\right) \right\|_1 \tag{3}$$

$$L_{gan} = \mathbb{E}_{x \sim p_x} \log[Dis(\xi_T)] + \mathbb{E}_{z \sim p_z} \log\left[1 - Dis\left(Dec\left(Enc(x_t)\right)\right)\right] \tag{4}$$

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$\boldsymbol{\xi}_{t-T-1}, \boldsymbol{\xi}_{t-T-2}, \ldots, \boldsymbol{\xi}_t$から構成される．$\boldsymbol{\xi}$は各フレームの画像である．さらに，提案手法では各種安定化手法として，[24]を参考に，TTUR[25], Self-Attention[26]，Spectral normalization[27]を導入する．学習の最適化手法はAdabound[28]を用いる．Encoder と Decoder の各層の活性化関数は FReLU[29]，Discriminator のそれは LeakyReLU を用いる．なお，Decoder の最終層には活性化関数として tanh を用いる．

3.3. 異常検知

入力$\boldsymbol{x_t}$に対する異常度$a(\boldsymbol{x_t})$を式（5）のように定義する．

$$a(\boldsymbol{x_t}) = \left\| \left| \boldsymbol{\xi}_{t+1} - Dec\left(Enc(\boldsymbol{x_t})\right) \right| \boldsymbol{patch} \right\|_1 \qquad (5)$$

$\boldsymbol{patch}$は Patch-Discriminator からの出力$Dis(\boldsymbol{\xi}_{t+1})$と，成分がすべて 1 で$Dis(\boldsymbol{\xi}_{t+1})$と同じサイズの行列$A$との差のマップ$A - Dis(\boldsymbol{\xi}_{t+1})$をフレーム画像のサイズにリサイズしたものである．よって，$\boldsymbol{patch}$は最終的に各成分が 1 に近いほど偽物，0 に近いほど本物であることを意味するマップとなる．さらに，異常算出に用いる最終的なスコア$S(\boldsymbol{x_t})$は式（6）を用いて正規化する．

$$S(\boldsymbol{x_t}) = \frac{a(\boldsymbol{x_t})}{max\left(a(\boldsymbol{x}_{1 \ldots m})\right)} \qquad (6)$$

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 Avenue[30]と，我々が独自に用意した労働災害データセット[11]を用いて実験を行った. 図 5 にデータセットの例を示す. 前者は 16clip の訓練データ，21clip のテストデータからなる．図に示すように，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．後者は 5clip の訓練データ，7clip のテストデータからなる．工場内の定点の監視カメラ画像を収録している．正常データは予め定められたルールに従った運搬，歩行等の作業の様子を収録している．一方異常データは走る，規則違反の運搬行動等，ルールから逸脱した様子を収録している．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．

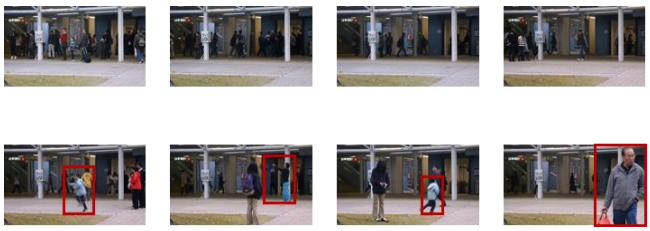

図 5 Avenue データセットの正常データ（上図）及び異常データ（下図）．赤色矩形領域が異常．

4.2. 実験条件

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 1e-3，8e-3，タイムステップTは 4 とし，予測損失の重みλは 100 とした．なおこれらの値は事前に複数回のチューニングを行った上で決定した．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. Avenue の結果

表 1 に Avenue に対する定量的結果を示す. なお，Ours only STN は$\boldsymbol{patch}$による重みづけをしない場合の結果である．また，図 6 に異常マップを示す．

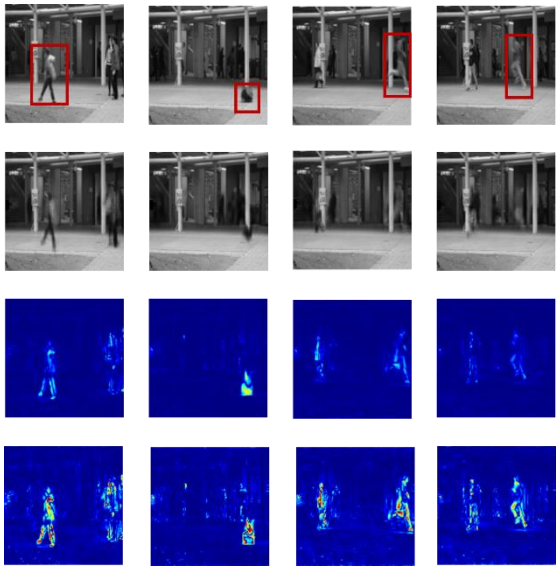

図 6 異常マップ

表 1 Avenue に対する結果

	Ours	Ours without patch	Future frame detection[15]	GRU-AE GAN[11]	SRNN[7]	MLAD[14]	Conv-AE[4]
AUROC↑	**0.873**	0.868	0.851	0.829	0.817	0.715	0.702

なお, 図 6 は上段からテストデータ, 予測結果, Naïve な差分マップ，***patch***を融合した結果である．

AUROC を用いた定量的評価により，提案手法の有効性を確認した．また，***patch***を融合することによる有効性も確認した．

4.4. 独自データの結果

表 2 に独自データセットに対する結果を示す．定量的評価により，提案手法の有効性を確認した．しかし，独自データセットの結果を見るに，数値的改善は現れたものの，現場に適用可能な実用的性能には至っていない．これは，実環境においては公開データセットにはない複雑な状況が存在することに起因すると考えられる．具体的には，今回我々が独自で用意したデータには，工場内に複数台の機材や用具があるため，膨大なパターンの正常が存在する．更に作業員によって異なる作業の癖等が存在する．そのため，実環境での運用においては領域を限定する等の工夫が必要である．

表 2 独自データセットの結果

	Ours	Ours without patch	GRU-AE GAN	Conv-AE
AUROC↑	**0.754**	0.750	0.728	0.571

5. 結論

本稿では，動画像を効率的にモデル化するため，フレーム予測型の時空間敵対的生成ネットワークを提案した．敵対的学習の不安定性を改善するため各種安定化手法を用い，Patch-Discriminator の出力を融合することで高レベルな特徴を活用し，効率的な異常検知を確立した．Avenue データセットにおいて，既存手法を上回る結果を確認した．しかし，独自データセットにおいては十分な精度を得たとは言えない．今後の展望として，領域を限定するために物体検知手法との融合を検討するほか，異常の局所的な検知のためのフレームワークを導入する．

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

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

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

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

[11] 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.

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

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

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

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

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

[17] Muhammad Zaigham Zaheer, et al., “Old is Gold: Redefining the Adversarially Learned One-Class Classifier Training Paradigm,” CVPR 2020.

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

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

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

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

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

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

[24] Andrew Brock, et al., “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” ICLR, 2019.

[25] Martin Heusel, et al., “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” NeurIPS, 2017.

[26] Han Zhang, et al., “Self-Attention Generative Adversarial Networks,” ICML, 2019.

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

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

[29] MuhLiangchen Luo, et al., “Adaptive Gradient Methods with Dynamic Bound of Learning Rate,” ICLR, 2019.

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

橋本慧志：2019 年 3 月中央大学理工学部精密機械工学科を卒業，2019 年 4 月より中央大学理工学研究科精密工学専攻博士前期課程において動画像異常検知の研究に従事．日本機械学会会員．

工藤謙一：1985 年日本大学大学院農学研究科農業工学専攻修士課程修了．プリマハム株式会社，東京大学大学院工学系研究科精密機械工学専攻助教，北里大学獣医学部動物資源科学科教授，（地独）青森県産業技術センター工業部門所長，2017 年より中央大学シニア URA．専門はメカトロニクス，農業情報工学．精密工学会，日本冷凍空調学会会員．

高橋孝幸：1984 年北海道大学農学部農業工学科卒業，同年プリマハム株式会社入社．開発部門にて新商品や製造ライン合理化の機械設備開発に従事．

梅田和昇：1989 年東京大学工学部精密機械工学科卒業，1994 年東京大学大学院工学系研究科精密機械工学専攻博士課程修了，同年中央大学理工学部精密機械工学科専任講師，2006 年同教授，現在に至る．ロボットビジョン，画像処理の研究に従事．博士（工学）．日本ロボット学会，精密機械工学会，日本機械学会，電子通信情報通信学会，IEEE 等の会員．