Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2022.230

PHÁT HIỆN BẤT THƯỜNG: DỰ ĐOÁN KHUNG HÌNH TƯƠNG LAI KẾT HỢP SOTA OPTICAL-FLOW MODEL VÀ CẢI TIẾN HÀM LỖI DỰA TRÊN FENCEGAN

Võ Trung Hiếu*, Võ Duy Nguyên và Nguyễn Tấn Trần Minh Khang
Trường Đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh
**Người chịu trách nhiệm về bài viết: Võ Trung Hiếu (email: 18520758@gm.uit.edu.vn)*

Thông tin chung:
Ngày nhận bài: 10/06/2022
Ngày nhận bài sửa: 11/07/2022
Ngày duyệt đăng: 14/07/2022

Title:
Anomaly detection: Future Frame Prediction combining sota optical-flow model and error function modification based on FenceGAN

Từ khóa:
Dự đoán khung hình, dự đoán bất thường trên video, hàm lỗi GAN, luồng quang học, phát hiện bất thường

Keywords:
Anomaly detection, frame prediction, GAN error function, optical-flow, video anomaly prediction

ABSTRACT

Anomaly detection is a problem that has received much attention in recent years with many high-precision methods born. Future Frame Prediction is a novel approach to video anomaly detection. This method has the main idea of using videos containing only normal frames to train the model. In the test phase, the model can take as input a video containing an anomaly frame, the anomaly detection is based on the difference between the frame reconstructed from the normal data and the actual frame containing any abnormality, usually in test video. In this paper, the model was improved by changing the optical-flow estimator in the Future Frame Prediction model and modify the error function of this model based on the FenceGAN work to increase the anomaly detection performance. The results after fine-tuning showed that the model improves on average 0.5% accuracy on standard datasets such as UCSD Ped1, UCSD Ped2, Avenue.

TÓM TẮT

Phát hiện bất thường là một bài toán được quan tâm nhiều trong những năm gần đây với nhiều phương pháp ra đời có độ chính xác cao. Future Frame Prediction là phương pháp tiếp cận mới lạ đối với bài toán phát hiện bất thường trong video. Phương pháp này có ý tưởng chính là sử dụng các video chỉ chứa khung hình bình thường để huấn luyện mô hình. Giai đoạn kiểm tra, mô hình có thể nhận đầu vào là video có chứa khung hình bất thường, việc phát hiện bất thường dựa trên sự khác biệt giữa khung hình được tái tạo từ dữ liệu bình thường và khung hình thực tế có chứa bất thường trong video kiểm tra. Trong bài báo này, mô hình được cải thiện bằng cách thay đổi thành phần ước lượng optical-flow trong mô hình Future Frame Prediction và sửa đổi hàm lỗi dựa trên công trình FenceGAN nhằm tăng hiệu suất phát hiện bất thường. Kết quả sau khi tinh chỉnh, mô hình cải thiện độ chính xác trung bình 0,5% trên các bộ dữ liệu chuẩn như UCSD Ped1, UCSD Ped2, Avenue.

1. GIỚI THIỆU

1.1. Bất thường trong video

1.1.1. Định nghĩa về bất thường

Một số công trình trước đây từng đề cập đến khái niệm bất thường như Chandola et al. (2009), Chalapathy et al. (2019), Zhu et al. (2020), Ramachandra et al. (2020)… Trong đó, định nghĩa về bất thường của Ramachandra et al. (2020) không chỉ bao quát các định nghĩa trước mà còn bổ sung thêm thông tin ngữ nghĩa về không gian và thời gian khi xác định sự kiện bất thường. Cụ thể,

Ramachandra và các cộng sự đã cho rằng bất thường trong video có thể được coi là sự xuất hiện của đối tượng bất thường, thuộc tính chuyển động bất thường hoặc sự xuất hiện của đối tượng có thuộc tính hoặc chuyển động không thường gặp ở các địa điểm hoặc thời điểm xác định. Với định nghĩa này, trong công trình cũng đã nêu ra những khó khăn và thách thức đến từ dữ liệu của bài toán. Cụ thể, bản chất khái niệm bất thường không cố định và không nhất quán ở những ngữ cảnh hoặc những thời điểm khác nhau.

1.1.2. Không nhất quán khái niệm bất thường theo không gian và thời gian

Trong các khảo sát về bất thường mà chúng tôi tham khảo, việc không nhất quán và không ổn định của bất thường trong video cũng là một vấn đề lớn được quan tâm.

Khi nói về sự không nhất quán của bất thường theo không gian, nghĩa là khi thay đổi ngữ cảnh video nhận định về bất thường có thể không còn đúng nữa. Giả sử video ghi hình tại một công viên hoặc một địa điểm công cộng có cảnh xung đột, bạo lực, đánh nhau,… ngữ cảnh này chính là cảnh bất thường. Mặt khác, nếu video được ghi hình tại một giải đấu võ, boxing,… thì hành động xung đột, đánh nhau không hẳn là một cảnh bất thường. Tóm lại, khi ngữ cảnh video bị thay đổi, định nghĩa bất thường cũng không còn đúng với ngữ cảnh mới. Nói cách khác, bất thường trong một hoặc một số ngữ cảnh không hẳn sẽ là bất thường ở tất cả các ngữ cảnh khác và ngược lại

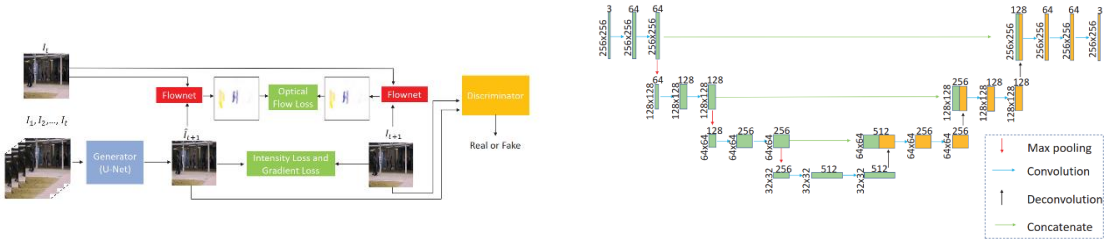
Đối với việc không nhất quán bất thường theo thời gian, giả sử một trung tâm mua sắm hoạt động từ 8h – 23h mỗi ngày trừ các ngày cuối tuần, trung tâm này có camera an ninh ghi hình tại quầy thu ngân. Việc khách hàng ra vào quầy thu ngân trong giờ làm việc là sự kiện bình thường (không phải bất thường). Tuy nhiên, ngoài giờ làm việc hoặc ở thời điểm mà trung tâm mua sắm đóng cửa, việc ghi lại cảnh có người ra vào quầy thu ngân lại là sự kiện bất thường. Tóm lại, cùng một ngữ cảnh, nếu thời gian thay đổi thì khái niệm bất thường cũng không còn chính xác tuyệt đối nữa. Nói cách khác, để xác định đúng sự kiện bất thường cần chú ý đến không chỉ là ngữ cảnh trên video mà còn cần đặc biệt lưu ý đến thời điểm diễn ra bất thường để tránh nhầm lẫn cho việc phát hiện bất thường từ đó tránh sai sót cho các kiến trúc mô hình được xây dựng cho bài toán này.

1.2. Phương pháp dự đoán khung hình bất thường

Dự đoán bất thường trong video là một nhánh của phát hiện bất thường, được sử dụng rộng rãi trong việc xây dựng hệ thống giám sát an ninh tự động ở những nơi công cộng như đường phố, sân bay, siêu thị, trung tâm vui chơi giải trí,... (Ramachandra et al., 2020). Phương pháp Future Frame Prediction cho phép xác định sự bất thường ở những khung hình cụ thể (mức khung hình) hoặc các vùng cụ thể của khung hình (mức không gian) có chứa sự kiện bất thường trong video bằng cách so sánh các khung hình chỉ chứa sự kiện bình thường được tạo ra và khung thực tế của video. Khung hình bất thường được chỉ định phụ thuộc vào ngưỡng được đặt ra trong việc so sánh khung hình thực tế và khung hình được dự đoán. Trên hết, việc phát hiện bất thường còn phụ thuộc rất nhiều vào ngữ cảnh của mỗi tập dữ liệu.

Phương pháp này được xây dựng dựa trên kiến trúc GAN (Liu et al., 2018) trong Hình 1(a) với U-Net là Generator. Kiến trúc đặc biệt của U-Net (Hình 1(b)) này cho phép tái tạo khung hình có cùng kích thước với khung hình đầu vào. Ý tưởng chính của mô hình Future Frame Prediction là tạo ra khung hình không bất thường bằng cách chỉ học các video không chứa khung hình bất thường. Nói cách khác, mô hình chỉ học các đặc trưng của khung hình bình thường để có thể tái tạo ra khung hình không chứa các sự kiện bất thường và từ đó tạo tiền đề cho việc phát hiện khung hình bất thường ở bước kiểm tra.

Ở bước huấn luyện, thành phần "Generator" của mô hình kết hợp với optical-flow trong việc dự đoán khung hình để nâng cao chất lượng của việc tái tạo khung hình và đảm bảo tính logic chuyển động của các đối tượng giữa các khung hình trước và sau. Các khung được tạo ra sẽ được phân loại bởi "Discriminator" để đánh giá các khung được tạo này là thật hay giả. Ở bước kiểm thử, mô hình sẽ nhận vào video có chứa các khung hình bất thường, nhiệm vụ của mô hình là cố gắng tạo ra khung hình không chứa sự kiện bất thường. Khung hình này được so sánh với khung hình thực tế để xác định xem khung hình thực tế có chứa sự kiện bất thường hay không dựa trên ngưỡng xác định bất thường đã được xác định từ trước. Trong phạm vi bài báo này, thành phần ước lượng optical-flow từ FlowNet2 (2017) được chuyển sang phương pháp hiện đại hơn là RAFT (2020), đồng thời chỉnh sửa hàm lỗi của mô hình từ hàm lỗi GAN cơ sở sang ý tưởng hàm lỗi của FenceGAN nhằm cải thiện độ chính xác AUC của mô hình Future Frame Prediction.



A. Mô hình Future Frame Prediction **B. Kiến trúc mạng U-Net được điều chỉnh bởi Liu et al. (2018)**

Hình 1. Ảnh minh hoạ mô hình của Future Frame Prediction với U-Net là thành phần Generator

(Ghi chú: Đầu vào (a) là 1 loạt t frame ảnh liên tiếp (trong bài báo t=5). Trong đó, 4 frame ảnh đầu sẽ là đầu vào cho Generator (b) để dự đoán frame cuối cùng. Frame ảnh cuối cùng đầu vào được xem như "ground-truth" để đánh giá kết quả dự đoán khung hình.)

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Các nghiên cứu liên quan

2.1.1. Phát hiện bất thường trong video ở mức khung hình và mức không gian

Phát hiện bất thường trong video có hai mức chính là mức khung hình (temporal localization) và mức không gian (spatial localization) (Ramachandra et al., 2020). Trong đó, mức khung hình sẽ phát hiện sự kiện bất thường cho dù bất kỳ khung hình nào trong video có chứa sự kiện bất thường hoặc sự kiện mới lạ không được học trong bước đào tạo. Trong khi đó, phát hiện mức không gian sẽ phân đoạn vùng bất thường sau khi phát hiện khung hình bất thường.

Nói cách khác, ở mức khung hình, việc phát hiện sự bất thường không quan tâm đến hình thức và vị trí sự kiện bất thường xảy ra trong khung hình video. Thay vào đó, việc phát hiện các sự kiện bất thường trong video giám sát dưới mức khung hình sẽ mở ra khả năng áp dụng các mô hình giám sát yêu cầu phát hiện đủ nhanh và cảnh báo kịp thời trong hệ thống giám sát an ninh.

2.1.2. Mô hình ước lượng optical-flow

Optical-flow có nhiệm vụ ước tính đặc trưng chuyển động theo thời gian trên mỗi pixel giữa các khung hình video. Đây cũng là một bài toán rất lâu chưa có lời giải tối ưu. Các hệ thống tốt nhất bị hạn chế bởi những khó khăn, bao gồm các vật thể chuyển động nhanh, vật cản, chuyển động mờ và bề mặt kém ngữ nghĩa (Teed et al., 2020).

FlowNet 2.0: Sự tiến hoá của việc ước lượng optical-flow bằng kiến trúc học sâu

FlowNet2 (Ilg et al., 2017) là phiên bản phát triển của FlowNet. Chúng là mạng được sử dụng trong nhiệm vụ ước tính optical-flow với hiệu suất cao. Bằng cách xếp chồng các biến thể FlowNet, FlowNet2 đã đạt được hiệu suất cao và được coi là hiện đại tại thời điểm được công bố. Kiến trúc này cũng được Liu et al. (2018) sử dụng cho mô hình Future Frame Prediction của mình.

RAFT: Biến đổi hồi quy tất cả cặp vùng cho việc ước lượng optical-flow

RAFT (Recurrent All-Pairs Field Transforms) (Teed et al., 2020) là mô hình ước lượng optical-flow hiện đại (Hình 2(a)). Nghiên cứu của Ilg et al. (2017) đã cố gắng cải tiến FlowNet cơ sở trở thành FlowNet2 (Hình 2(b)) bằng việc xếp chồng hai biến thể FlowNetC và FlowNetS kết hợp với FlowNet cơ sở tiêu chuẩn và thành phần warping để cải thiện chất lượng đầu ra của optical-flow. Ilg et al. (2017) cho rằng, việc kết hợp các biến thể FlowNet và tinh chỉnh dữ liệu huấn luyện đồng nhất cho biến thể FlowNetC và FlowNetS sẽ giúp cải thiện lớn về chất lượng của optical-flow. Trong bài báo của mình, họ đã chứng minh FlowNet2 làm giảm đến 50% lỗi khi ước lượng optical-flow và trở thành phương pháp state-of-the-art tại thời điểm ra mắt năm 2017.

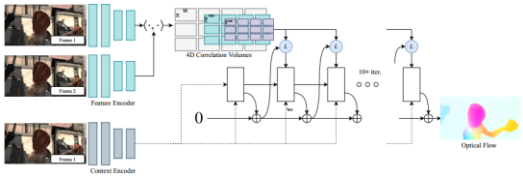
RAFT là mô hình ước lượng optical-flow SoTA tính tới thời điểm hiện tại và được ra mắt vào năm 2020. Mô hình cho hiệu suất tốt nhất trên bộ dữ liệu KITTI (Geiger et al., 2013) và cả Sintel. Khác với FlowNet2, Teed et al. (2020) đã đề xuất RAFT với ý tưởng chính là dùng kiến trúc mô hình mới với 3 giai đoạn: trích xuất đặc trưng, tính độ tương quan "visual", thực hiện vòng lặp lớn để tối ưu kết quả mô hình. Điều đặc biệt ở RAFT là sử dụng khối tương quan 2D và 4D thay vì là chỉ tính toán giá trị tương quan. Mô hình RAFT cho hiệu suất cao hơn nhiều so với FlowNet2 trên các bộ dữ liệu tiêu chuẩn. Cụ thể với lỗi đầu cuối (EPE) là 1.609 trên

bộ dữ liệu Sintel (clean) giúp giảm lỗi khoảng 60% so với FlowNet2 có EPE là 3.960.

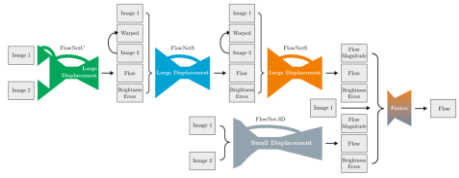
Việc thay thế FlowNet2 bằng RAFT trong nhiệm vụ ước lượng optical-flownhằm cải thiện hiệu suất dự đoán khung hình và tối ưu hàm lỗi trong quá trình huấn luyện. EPE thấp hơn trên các bộ dữ liệu tiêu chuẩn so với FlowNet2, RAFT mang đến tiềm năng tối ưu kết quả ước lượng optical-flow cục bộ và tối ưu hàm mục tiêu tổng thể của mô hình, từ đó làm tăng độ chính xác AUC của việc phát hiện bất thường.

2.1.3. Mô hình GAN

Generative adversarial networks (GAN) (Creswell et al., 2018) đã được chứng minh tính hữu ích của mình trong việc tái tạo hình ảnh với nhiều công trình liên quan. Kiến trúc mạng này bao gồm 2 thành phần, Generator (G) và Discriminator (D) được thể hiện trong Hình 1(a). Điều đó cho phép mô hình học cách tái tạo lại khung hình trong video bằng G và cải thiện chất lượng tái tạo bằng D. Tuy nhiên, việc tạo khung hình trong video RGB đòi hỏi các yếu tố như độ phân giải, chi tiết, màu sắc và đặc biệt là chuyển động của các đối tượng giữa các khung hình trước và sau phải logic với nhau về hướng chuyển động, màu sắc, độ sáng,... của các đối tượng trong khung hình. Vì vậy, Liu et al. (2018) đã đề xuất một phương pháp có tên là Future Frame Prediction bằng cách sử dụng cơ sở của GAN và kết hợp optical-flow được dự đoán bởi pre-trained của FlowNet để cải thiện chất lượng dự đoán khung hình (Hình 1(a)). Ý tưởng chính của phương pháp này là sử dụng một chuỗi t hình ảnh liên tục làm đầu vào của cả mô hình. Trong đó, họ sử dụng t-1 hình ảnh đầu tiên để tái tạo khung hình thứ t và phân loại hình ảnh được tạo này là thực hay không theo D. Ở bước tái tạo khung hình thứ t, khung hình được tạo ra xem như là khung hình được dự đoán và khung hình này được dùng để so sánh với khung hình thứ t thực tế. Độ lỗi tổng thể của mô hình bao gồm nhiều độ lỗi phụ như: lỗi tái tạo khung hình (flow loss, intensity loss, gradient loss, generation loss), lỗi phân loại khung hình được tái tạo (discriminator loss). Hàm lỗi tổng thể vẫn còn khá phân tán, chưa được các tác giả tập trung vào việc tối ưu. Với kì vọng cải thiện hiệu suất mô hình, một số hàm lỗi được tinh chỉnh thành phần để tối ưu hiệu suất tái tạo khung hình, từ đó tăng hiệu suất phát hiện khung hình bất thường.



A. Mô hình kiến trúc của RAFT tinh gọn hơn, mô hình gồm 3 giai đoạn: Trích xuất đặc trưng, tính độ tương quan "visual", thực hiện vòng lặp lớn để tối ưu kết quả mô hình (Teed et al., 2020)



B. Mô hình kiến trúc của Flownet2 với kiến trúc lớn, xếp chồng các biến thể của FlowNet để cải thiện chất lượng optical-flow (Ilg et al., 2017)

Hình 2. Kiến trúc mô hình Flownet2 và RAFT

2.1.4. FenceGAN

Vấn đề hàm lỗi và việc tối ưu nó trong kiến trúc GAN là một vấn đề được quan tâm khá nhiều, đặc biệt đối với bài toán tái tạo khung hình. Do đặc trưng mô hình gồm nhiều thành phần, trong đó thành phần Generator (G) và Discriminator (D) được xem là hai thành phần đối ngẫu lẫn nhau. Do đó, việc tối ưu các siêu tham số cũng như hàm lỗi để hai thành phần này đạt được hiệu suất tốt vẫn là một thách thức rất lớn. Liu et al. (2018) đã dựa trên hàm lỗi cơ sở của GAN để tính toán lỗi cho thành phần G và D.

Hàm lỗi gốc của Generator trên GAN:

$$\mathcal{L}_{G_\theta}^{GAN}(G_\theta, D_\phi, \mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N}\left[log\left(1 - D_\phi(G_\theta(\boldsymbol{z}_i))\right)\right]$$

Hàm lỗi gốc của Discriminator trên GAN:

$$\mathcal{L}_{D_\phi}^{GAN}(G_\theta, D_\phi, \mathcal{X}, \mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\left(D_\phi(\boldsymbol{x}_i)\right) - \log\left(1 - D_\phi(G_\theta(\boldsymbol{z}_i))\right)\right]$$

Trong đó, $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ là tập dữ liệu gồm N điểm dữ liệu và mỗi điểm dữ liệu $\boldsymbol{x}_i$ là một ảnh hoặc khung hình. Nhiệm vụ của GAN là cố gắng tạo ra một bản sao thật nhất có thể của từng điểm dữ liệu x_i. Với mỗi điểm dữ liệu được xem như thuộc về một phân phối p_{data} của dữ liệu thực. GAN sau khi được huấn luyện sẽ nhận đầu vào là một điểm dữ liệu ngẫu nhiên thuộc về p_{noise} và được xem như thuộc về một phân phối nhiễu. Điểm dữ liệu nhiễu này được đưa vào GAN để tái tạo lại khung hình bằng những đặc trưng dữ liệu đã được học trước đó.

Tập điểm dữ liệu nhiễu chính là $\mathcal{Z} = \{z_1, z_2, \dots, z_N\}$ trong công thức hàm lỗi gốc của GAN được đề cập bên trên. Tuy nhiên, việc cố gắng tạo ra điểm dữ liệu thật nhất thuộc phân phối dữ liệu thực thường gặp phải tình trạng quá khớp với dữ liệu thực. Khi đó G sẽ đạt được hiệu suất tái tạo khung hình quá tốt dẫn đến kết quả phân loại của D kém đi và điều này được coi là không hiệu quả cho một hệ thống của mô hình phát hiện bất thường. Điều đó dẫn đến hàm lỗi gốc của GAN truyền thống không trực tiếp phù hợp với mục tiêu phát hiện bất thường. Do đó, Ngo et al. (2019) đã đề xuất Fence GAN với đóng góp chính là sửa đổi hàm lỗi của GAN sao cho các mẫu được tái tạo nằm trong ranh giới của phân phối dữ liệu thực cho mục tiêu phát hiện bất thường. Các sửa đổi mục tiêu của chúng là để Generator tái tạo ra các mẫu xung quanh ranh giới của $\mathcal{X}$, mà các tác giả ký hiệu là $\delta\mathcal{X}$. Điều đó cho phép Discriminator tại cuối quá trình đào tạo, vẽ một ranh giới "chặt chẽ" xung quanh $\mathcal{X}$. Vì vậy, Discriminator sau đó có thể được sử dụng như một mô hình phân loại một lớp hoặc một mô hình phát hiện bất thường. Thay vì sử dụng hàm lỗi đơn giản cho thành phần Generator, Ngo et al. (2019) đã đề xuất hai hàm lỗi thay thế đó là Encirclement Loss (EL) và Dispersion Loss (DL). $\mathcal{L}_{G_\theta}^{GAN}(G_\theta, D_\phi, \mathcal{Z})$ khi đó sẽ là tổng của EL và DL với hằng số $\beta \in \mathbb{R}^+$ để khuếch đại DL. Công thức hàm lỗi mới cho Generator theo FenceGAN được tổng hợp như sau:

Encirclement Loss:

$$EL(G_\theta, D_\phi, \mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\left|\alpha - D_\phi(G_\theta(z_i))\right|\right)\right]$$

Các tác giả của FenceGAN muốn G tạo ra các điểm $G_\theta(z)$ nằm bên trong ranh giới $\delta\mathcal{X}$, vì vậy họ đề xuất hàm lỗi thành phần cho G và được gọi là Encirclement Loss. Trong đó, $\alpha \in (0,1)$ được sử dụng cho EL như một siêu tham số. Trong quá trình thay thế hàm lỗi gốc $\mathcal{L}_{G_\theta}^{GAN}(G_\theta, D_\phi, \mathcal{Z})$ của mô hình Future Frame Prediction, nhiều thực nghiệm được tiến hành và đánh giá, cuối cùng $\alpha = 0{,}5$ được chọn cho tất cả các bộ dữ liệu với hiệu suất đạt được là cao nhất.

Dispersion Loss

$$\mu = \frac{1}{N}\sum_{i=1}^{N} G_\theta(z_i)$$

$$DL(G_\theta, \mathcal{Z}) = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\left(\|G_\theta(z_i) - \mu\|_2\right)}$$

Hàm lỗi thành phần kết hợp với EL cho hàm lỗi Generator của FenceGAN được gọi là Dispersion Loss. Trong đó, $\mu = (\mu_1, \mu_2, \dots, \mu_d)$ là center của mẫu dữ liệu vừa được tái tạo được tính theo công thức $\mu = \frac{1}{N}\sum_{i=1}^{N} G_\theta(z_i)$ với mục đích để các điểm dữ liệu được tạo ra từ Generator có thể phủ được toàn bộ ranh giới $\delta\mathcal{X}$. Hàm lỗi mới của G kí hiệu là $\mathcal{L}_{\text{generator}}^{\text{FGAN}}$ được viết lại dưới đề xuất của Ngo et al. (2019) như sau:

$$\mathcal{L}_{\text{generator}}^{FGAN}(G_\theta, D_\phi, \mathcal{Z}) = EL + \beta \times DL$$

Trong đó, β là hằng số khuếch đại DL với $\beta \in \mathbb{R}^+$ và đóng vai trò là một siêu tham số tương tự như α. Trong tinh chỉnh mô hình Future Frame Prediction, kết quả cho thấy DL không thật sự phù hợp và mang lại kết quả cải thiện cho mô hình. Vì vậy trong phần tinh chỉnh hàm lỗi, DL được bỏ qua hay nói cách khác giá trị cho β được chọn là 0.

Đối với Discriminator, các tác giả của FenceGAN cũng thêm một vài thay đổi nhỏ trong hàm lỗi của D với việc thêm vào hằng số γ trong công thức :

$$\mathcal{L}_{\text{discriminator}}^{FGAN}(G_\theta, D_\phi, \mathcal{X}, \mathcal{Z}) =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[-\log\left(D_\phi(\boldsymbol{x}_i)\right) - \gamma\log\left(1 - D_\phi(G_\theta(\boldsymbol{z}_i))\right)\right]$$

Các tác giả của FenceGAN đề cập đến việc dùng hằng số γ như một siêu tham số để khiến hàm lỗi hoạt động tốt hơn trên bài toán phát hiện bất thường. Với đặc thù của mô hình GAN, việc huấn luyện D là một bài toán đánh đổi, nhất là với các tiếp cận dự đoán khung hình cho việc phát hiện bất thường. D sẽ phải tập trung cho việc phân loại tốt ở lớp ảnh giả (được tái tạo bởi G) hoặc lớp ảnh thật (ảnh groundtruth từ bộ dữ liệu). Với hằng số $\gamma \in (0,1]$ trong công thức FenceGAN, điều này có nghĩa là khi γ nhỏ hơn 1, D sẽ tập trung hơn về việc phân loại các điểm dữ liệu thực một cách chính xác, do đó ranh giới quyết định của nó ít có khả năng bị uốn cong vào miền X, cho phép G ước lượng δX tốt hơn. Trong quá trình thực nghiệm, γ được điều chỉnh theo kinh nghiệm cho từng tập dữ liệu riêng biệt.

2.2. Phương pháp đề xuất

Trong bài báo này, việc thay thế mô hình thành phần ước lượng optical-flow từ FlowNet2 sang RAFT được đề xuất nhằm ~~làm~~ tăng chất lượng tái tạo khung hình. Mô hình dự đoán optical-flow là một thành phần quan trọng trong kiến trúc của Future Frame Prediction. Optical-flow được xem như yếu tố ràng buộc về chuyển động được nhắc đến

trong công trình của Liu et al. (2018), yếu tố này biểu diễn sự ràng buộc chuyển động theo thời gian giữa các khung hình trước và sau. Việc dựa trên yếu tố này để tối ưu lỗi tái tạo về mặt chi tiết các đối tượng trong khung hình sẽ giúp tăng tính đảm bảo logic về chuyển động theo thời gian của các đối tượng giữa các khung hình được tái tạo so với các khung hình thực tế trước đó. Một trong những đóng góp chính của phương pháp Future Frame Prediction là bổ sung thành phần ước lượng optical-flow cho mô hình GAN cơ sở nhằm tăng chất lượng tái tạo khung hình từ "Generator". Do đó, việc thay đổi mô hình ước lượng optical-flow FlowNet2 sang mô hình tiên tiến hơn là RAFT được kỳ vọng giúp cải thiện hiệu suất của việc tái tạo khung hình, từ đó làm tăng độ chính xác của quá trình phát hiện bất thường.

Một đóng góp quan trọng khác của bài báo này là đề xuất chỉnh sửa hàm lỗi của mô hình Future Frame Prediction từ hàm lỗi cơ sở của GAN truyền thống sang hàm lỗi dựa trên công trình FenceGAN của Ngo et al. (2019). Theo đó, đề xuất cải tiến và sửa đổi hàm lỗi của phương pháp Future Frame Prediction được đưa ra dựa trên ý tưởng của FenceGAN để tìm ra cấu hình phù hợp nhất của việc tinh chỉnh này cho từng bộ dữ liệu. Cụ thể, việc tinh chỉnh và cài đặt lại hàm lỗi mới hoàn toàn theo FenceGAN được thực hiện để đánh giá. Tuy nhiên, kết quả gần như không cải thiện. Nguyên nhân đến từ việc FenceGAN được phát triển cho bài toán phát hiện bất thường trên ảnh đơn và hơn hết là các bộ dữ liệu ảnh đơn giản như MNIST, CIFA10,… Mặc dù so với GAN, đóng góp thay đổi hàm lỗi theo FenceGAN giúp cải thiện đáng kể độ chính xác của phát hiện bất thường trên các bộ dữ liệu ảnh đơn tiêu chuẩn này. Tuy nhiên, đặc thù những bộ dữ liệu này chỉ chứa những đối tượng đơn và hoàn toàn không tồn tại ràng buộc chuyển động giữa các khung hình như trên video. Song, đối với phương pháp Future Frame Prediction lại là phương pháp dự đoán khung hình ở thời điểm tương lai trong video, bằng việc chỉ học đặc trưng của sự kiện bình thường ở tập huấn luyện để tái tạo khung hình tương lai không chứa sự kiện bất thường ở tập kiểm thử. Việc phát hiện khung hình ở tập kiểm thử có chứa bất thường hay không bằng cách tính toán lỗi giữa khung hình được tái tạo và khung hình thực tế trong video ở tập kiểm thử để đưa ra quyết định dựa vào một ngưỡng bất thường được đặt ra từ ban đầu. Do đó, nếu áp dụng trực tiếp hàm lỗi của FenceGAN cho phương pháp Future Frame Prediction là không hợp lí. Thay vào đó, i ý tưởng của FenceGAN được sử dụng để tinh chỉnh hàm lỗi sao cho thật sự phù hợp với phương pháp này. Cụ thể, đối với 2 hàm lỗi thành phần của G (Generator) trên FenceGAN, ý tưởng từ hàm lỗi Encirclement Loss được sử dụng, siêu tham số alpha được dùng vào hàm lỗi gốc của phương pháp Future Frame Prediction nhằm tạo ra $\delta\mathcal{X}$ cho từng bộ dữ liệu trong quá trình huấn luyện với mục tiêu để khung hình sẽ được tái tạo trong miền của $\delta\mathcal{X}$. Ý tưởng của hàm lỗi D (Discriminator) trên FenceGAN cũng được sử dụng để tinh chỉnh cho hàm lỗi gốc D của phương pháp Future Frame Prediction của Liu et al. (2018). Theo đó, hàm lỗi D được viết lại nhằm mục tiêu để D phân biệt tốt khung hình được tái tạo (fake frame) thay vì nhắm vào việc phân loại chính xác khung hình thật như hàm lỗi gốc trên phương pháp của Liu et al. (2018). Các hàm lỗi trước và sau khi tinh chỉnh được viết lại như sau:

Hàm lỗi gốc phương pháp Future Frame Prediction của Liu et al. (2018):

$$L_{adv}^{\mathcal{G}}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, 1\right)$$

$$L_{adv}^{\mathcal{D}}(\hat{I}, I) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(I)_{i,j}, 1\right) + \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, 0\right)$$

Hàm lỗi phương pháp Future Frame Prediction của Liu et al. (2018) sau khi tinh chỉnh dựa trên ý tưởng FenceGAN của Ngo et al. (2019):

$$L_{adv}^{\mathcal{G}\_Fence}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, \boldsymbol{\alpha}\right)$$

$$L_{adv}^{\mathcal{D}\_Fence}(\hat{I}, I) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(I)_{i,j}, 0\right) + \gamma \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, 1\right)$$

Trong quá trình thực nghiệm, nhiều giá trị khác nhau cho $\boldsymbol{\alpha}$ và $\boldsymbol{\gamma}$ với $\boldsymbol{\alpha} \in [0,1]$ và $\boldsymbol{\gamma} \in (0,1)$ được thử nghiệm theo công bố của Ngo et al. (2019) về việc chọn giá trị cho các siêu tham số cho hàm lỗi mới để tìm ra bộ giá trị tốt nhất. Kết thúc quá trình chỉnh sửa hàm lỗi và thực nghiệm, giá trị $\boldsymbol{\alpha}$ và $\boldsymbol{\gamma}$ được chọn dựa trên ý tưởng và công bố của FenceGAN, theo đó, Ngo et al. (2019) đã công bố kết quả cải thiện tốt nhất khi chọn $\boldsymbol{\alpha} = 0{,}5$ và $\boldsymbol{\gamma} = 0{,}1$. Kết quả là bộ siêu tham số này cũng cho ra kết quả tốt nhất trong số các thử nghiệm của nghiên cứu. Việc thử nghiệm thêm các bộ siêu tham số khác để tìm ra cấu hình tối

ưu nhất cho tinh chỉnh này được đề cập trong phần KẾT QUẢ VÀ THẢO LUẬN – Kết quả thảo luận.

2.3. Thực nghiệm và đánh giá

2.3.1. Các bộ dữ liệu bất thường tiêu chuẩn được sử dụng để đánh giá

UCSD Pedestrian

Bộ dữ liệu UCSD Pedestrian (Mahadevan et al., 2018) gồm 2 bộ dữ liệu con là UCSD Ped1 và UCSD Ped2 và đều là dữ liệu dạng video, chứa cảnh quay trên đoạn đường cố định bao gồm người đi bộ trên đường, các phương tiện giao thông và các đám đông. Bộ UCSD Ped1 chứa 34 video training và 12 video testing, độ phân giải là 158 x 238 với 200 frame ảnh ở mỗi video. Bộ UCSD Ped2 là 16 video training, 12 video testing với độ phân giải 240 x 360 và dao động từ 120 – 200 frame/ video. Ngoại trừ người đi bộ bình thường, các phương tiện và chuyển động bất thường của cả phương tiện và con người đều được quy định là bất thường trên bộ dữ liệu này. Bộ UCSD Ped1 bao gồm 14.000 khung hình (6.800 khung hình ở tập huấn luyện, 7.200 khung hình ở tập kiểm thử) trong khi bộ UCSD Ped2 chứa tổng cộng 4.560 khung hình (2.500 khung hình ở tập huấn luyện, 2.010 khung hình ở tập kiểm thử). Bộ dữ liệu UCSD cung cấp ground-truth ở cả mức khung hình lẫn mức không gian (pixel).

CUHK Avenue

Bộ dữ liệu Avenue (Lu et al., 2013) được thu thập trong CUHK Campus Avenue với tổng số 30.652 khung hình (gồm 15.328 khung hình ở tập huấn luyện và 15.324 khung hình ở tập kiểm thử). Bộ Avenue chứa 16 video ở tập huấn luyện và 21 video ở tập kiểm thử với tổng số 47 sự kiện bất thường, bao gồm ném đồ vật, lảng vảng và chạy. Bộ dữ liệu tồn tại một số thách thức để đánh giá như có sự rung nhẹ camera giám sát và sự xuất hiện của một số điểm bất thường nhỏ. Các video tập huấn luyện chỉ có các sự kiện bình thường trong khi video thử nghiệm bao gồm các sự kiện bình thường và bất thường. Ground-truth được cung cấp ở mức khung hình và cả mức pixel để đánh giá hiệu suất của các phương pháp phát hiện bất thường cơ sở và cả những phương pháp SoTA.

2.3.2. Độ đo đánh giá kết quả

ROC/AUC

Đường cong ROC (Receiver Operation Characteristic) thể hiện sự cân bằng giữa tỷ lệ dương tính thực (true positive) và tỷ lệ dương tính giả (false positive) đối với một mô hình dự đoán hoặc phân loại sử dụng ngưỡng xác suất phân loại. Đường cong ROC thích hợp sử dụng khi các quan sát được cân bằng giữa mỗi lớp, trong khi các đường cong như precision hay recall phù hợp để đánh giá cho các tập dữ liệu không cân bằng. Biểu đồ ROC biểu thị tỷ lệ báo động sai so với tỷ lệ báo động chính xác. Chi tiết về biểu đồ ROC được đề cập trong công trình của Narkhede et al. (2018).

Area Under Curve (AUC) (Narkhede et al., 2018) là đại lượng vô hướng của diện tích vùng dưới đường cong biểu đồ ROC được dùng để đánh giá độ chính xác của mô hình phân loại đặc biệt là mô hình phân loại nhị phân như bài toán phát hiện bất thường. Giá trị AUC càng cao, mô hình càng đáng tin cậy. Trong bài toán phát hiện sự kiện bất thường trong video ở mức khung hình, AUC đại diện cho độ tốt và hiệu suất của mô hình mang lại khi phân loại các khung hình bất thường và bình thường trong video.

Peak Signal-to-Noise Ratio (PSNR)

PSNR (Mathieu et al., 2015) là độ đo để đánh giá chất lượng khung hình được tái tạo so với khung hình thực tế trên từng pixel thay vì sử dụng các độ lỗi MSE (Mean Squared Error) truyền thống. Mathieu et al. (2015) đề xuất dùng độ đo này cho bài toán dự đoán khung hình và được xem là cách đánh giá khung hình dự đoán tốt hơn các độ đo trước đây như MSE. Nếu giả sử I là khung hình thực tế và $\hat{I}$ là khung hình được dự đoán thì công thức tính PSNR là:

$$\text{PSNR}\ (I, \hat{I}) = 10\log_{10} \frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=0}^{N} \left(I_i - \hat{I}_i\right)^2}$$

Trong đó

- I là khung hình thực tế,
- $\hat{I}$ là khung hình được dự đoán,
- $max_{\hat{I}}$ là biến động lớn nhất của ảnh được dự đoán $\hat{I}$. Ví dụ: Nếu $\hat{I}$ là ảnh mức xám với giá trị điểm ảnh giao động trong đoạn [0, 1] thì giá trị của $max_{\hat{I}}$ là 1, với ảnh màu giá trị điểm ảnh trong đoạn [0, 255] thì giá trị của $max_{\hat{I}}$ sẽ là 255.

Trong quá trình huấn luyện, để tiện cho việc theo dõi chất lượng tái tạo khung hình, mã nguồn cũng được tích hợp thêm độ đo Peak Signal-to-Noise Ratio (PSNR), theo đó tại iteration bất kì, PSNR càng cao thì khung hình ứng với iteration đó càng được tái tạo tốt. Dựa trên PSNR, ta có thể phán đoán sớm được thời điểm mô hình xảy ra overfitting hoặc varnishing để kịp thời thay đổi cấu hình hoặc kiến trúc mô hình cho phù hợp.

2.3.3. Cấu hình thực nghiệm

Trong nghiên cứu này, thực nghiệm trong nghiên cứu này được thiết lập trên cấu hình GPU RTX 2080Ti với mã nguồn được implement trên nền tảng pytorch. Đường dẫn đến repository github được sử dụng như mã nguồn phương pháp gốc: https://github.com/feiyuhuahuo/Anomaly_Predicti on.

Tuy nhiên, với implement này, ta chạy trên bộ avenue trước khi tinh chỉnh

2.3.4. Kết quả thực nghiệm.

Để đánh giá hiệu quả của việc thay thế thành phần ước lượng optical-flow, ta tiến hành thay đổi thành phần này từ FlowNet2 của phương pháp gốc sang RAFT – phương pháp được xem như SoTA cho việc ước lượng optical-flow tính tới thời điểm hiện tại 2022. Đồng thời, ta giữ nguyên các thành phần và yếu tố khác của mô hình để kết quả thực nghiệm là trung thực nhất cho việc đánh giá. Kết quả thực nghiệm đánh giá đượcthống kê ở Bảng 1. Để kết quả đánh giá được khách quan, cấu hình thực nghiệm, các tham số như learning rate, batchsize được cố định. Sau đó, kiến trúc gốc được chạy lại để ghi nhận kết quả trước và sau khi tiến hành chạy theo ý tưởng tinh chỉnh.

Bảng 1. Kết quả chạy thực nghiệm khi thay thế FlowNet2 bằng RAFT trên các bộ dữ liệu tiêu chuẩn bằng độ đo AUC (%)

Dataset	FlowModel	PaperAUC	OurAUC
Ped1	2SD	83,10	-
Ped1	RAFT	-	83,56
Ped2	2SD	95,40	-
Ped2	RAFT	-	95,07
Avenue	2SD	84,90	-
Avenue	RAFT	-	84,45

Thực nghiệm tinh chỉnh thay đổi mô hình ước lượng optical-flow RAFT và chỉnh sửa hàm lỗi mô hình theo FenceGAN như đã đề xuất trên các bộ dữ liệu tiêu chuẩn được được tiến hành và liệt kê ở phần Phương pháp đề xuất. Kết quả thực nghiệm được liệt kê ở Bảng 2 với độ đo đánh giá là AUC.

Trong quá trình thực nghiệm ý tưởng tinh chỉnh của mình, tổng lỗi generator của mô hình được khuếch đại bằng việc đề xuất dùng thêm hằng số *amp* để tăng trọng số lỗi của optical-flow. Nguyên nhân là do thay thế FlowNet2 bằng phương pháp SoTA của bài toán ước lượng optical-flow là RAFT. Độ lỗi của việc ước lượng optical-flow sau khi thay thế FlowNet2 sang RAFT sẽ giảm đi rất nhiều. Do đó, để tránh tình trạng "over-fitting" khi thay đổi mô hình ước lượng optical-flow, hằng số khuếch đại độ lỗi optical-flow được bổ sung thêm ở giai đoạn tối ưu mô hình.Việc khuếch đại lỗi optical-flow được đề xuất nhằm buộc mô hình phải tăng chất lượng tái tạo khung hình nhiều hơn để phù hợp với mô hình optical-flow mới. Kết quả thực nghiệm đề xuất cũng được liệt kê chi tiết ở Bảng 2. Với đề xuất thêm hằng số khuếch đại *amp*, kết quả có cải thiện hơn so với việc chỉ kết hợp việc thay thế thành phần ước lượng optical-flow RAFT và tinh chỉnh hàm lỗi theo FenceGAN.

Bảng 2. Kết quả chạy thực nghiệm khi thay thế FlowNet2 bằng RAFT kết hợp tinh chỉnh hàm lỗi theo FenceGAN bằng độ đo AUC (%) trên các bộ dữ liệu tiêu chuẩn

	α	γ	*amp*	Original	Our
Ped1	0.5	0.1	2.0	83.10	83,49
Ped1	0.5	0.1	5.0	83.10	83,66
Ped2	0.5	0.1	2.0	95.40	95,92
Ped2	0.5	0.1	5.0	95.40	95,85
Avenue	0.5	0.1	2.0	84.90	85,08
Avenue	0.5	0.1	5.0	84.90	85,32

3. KẾT QUẢ VÀ THẢO LUẬN

Đóng góp chính trong bài báo này là thay đổi thành phần kiến trúc mô hình, cụ thể là pretrained model dùng để ước lượng optical-flow. Với RAFT là mô hình ước lượng optical-flow được xem như SoTA tại thời điểm đề xuất ý tưởng cải tiến trong bài báo này. Theo đó, việc thay thế từ FlowNet2 sang RAFT không thật sự cho kết quả cải thiện vượt trội hay quá nổi bật. Tuy nhiên, trên bộ Ped1 và Ped2, kết quả có cải thiện so với việc dùng FlowNet2 0,45% trên bộ Ped1. Việc thay đổi này được kỳ vọng giúp tối ưu tốt hơn hàm lỗi mô hình của Liu et al. (2018) sau khi được tinh chỉnh và sửa đổi theo FenceGAN. Kết quả sau khi kết hợp giữa RAFT và hàm lỗi sửa đổi theo FenceGAN có cải thiện hơn so với ban đầu. AUC tăng từ 0,38% đến 0,56% trên tất cả các bộ dữ liệu bao gồm Ped1, Ped2, Avenue. Do đặc trưng của dữ liệu video bất thường thường phụ thuộc rất lớn vào ngữ cảnh và mục tiêu giám sát, những sự kiện là bất thường ở ngữ cảnh này lại có thể không phải là bất thường ở một ngữ cảnh khác. Vì vậy, việc tinh chỉnh để tìm ra cấu hình tối ưu cho từng bộ dữ liệu đặt ra thách thức về tài nguyên tính toán và thời gian là rất lớn, đặc biệt là với phương pháp có kiến trúc tương đối lớn như Future Frame Prediction. Do hạn chế về tài nguyên và thời gian thực nghiệm, chúng tôi đề xuất hướng cải tiến đạt hiệu quả cải thiện các bộ dữ liệu tiêu

chuẩn nhằm tạo tiền đề cho các nghiên cứu sau này có thể tìm ra cấu hình tốt nhất cho từng bộ dữ liệu với đề xuất của chúng tôi, độ chính xác từ đó có thể cao hơn công bố của bài báo này.

4. KẾT LUẬN

Với kết quả cải thiện đạt được trung bình là *0,5%* trên 3 bộ dữ liệu tiêu chuẩn Ped1, Ped2 và Avenue, trong tương tai có thể thực nghiệm và bổ sung thêm kết quả trên một số bộ dữ liệu khác, trong đó có bộ dữ liệu vừa được công bố năm 2021 về dữ liệu camera an ninh ghi lại các sự kiện bất thường tại Việt Nam do Vo et al. (2021) công bố. Nghiên cứu này hi vọng sẽ tạo tiền đề cơ sở cho các nghiên cứu phát triển sau này liên quan đến phát hiện bất thường nói chung và dự đoán bất thường nói riêng có thể cải thiện được độ tin cậy và độ chính xác của các mô hình phát hiện bất thường để các mô hình này có thể sớm áp dụng vào các vấn đề thực tế, giúp con người giải quyết được nhiều khó khăn nhất là đối với bài toán giám sát an ninh.

LỜI CẢM TẠ

Nghiên cứu được thực hiện tại Phòng thí nghiệm Truyền thông Đa phương tiện (MMLab), Trường Đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh.

TÀI LIỆU THAM KHẢO

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3), 1-58. https://doi.org/10.1145/1541880.1541882

Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010, June). *Anomaly detection in crowded scenes.* In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 1975-1981). https://doi.org/10.1109/CVPR.2010.5539872

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research, 32*(11), 1231-1237. https://doi.org/10.1177/0278364913491297

Lu, C., Shi, J., & Jia, J. (2013). *Abnormal event detection at 150 fps in matlab*. In Proceedings of the IEEE international conference on computer vision (pp. 2720-2727). https://doi.org/10.1109/ICCV.2013.338

Mathieu, M., Couprie, C., & LeCun, Y. (2015). *Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440*.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). *Flownet 2.0: Evolution of optical flow estimation with deep network*s. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2462-2470). https://doi.org/10.1109/CVPR.2017.179

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE* Signal *Processing Magazine, 35*(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202

Liu, W., Luo, W., Lian, D., & Gao, S. (2018). *Future frame prediction for anomaly detection–a new baseline*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6536-6545). https://doi.org/10.1109/CVPR.2018.00684

Narkhede, S. (2018). Understanding auc-roc curve.. *Towards Data Science, 26*(1), 220-227.

Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly* detection*: A survey. arXiv preprint arXiv:1901.03407*. https://doi.org/10.1145/3394486.3406704

Ngo, P. C., Winarto, A. A., Kou, C. K. L., Park, S., Akram, F., & Lee, H. K. (2019, November). *Fence GAN: Towards better anomaly detection*. In 2019 IEEE 31St International Conference on tools with artificial intelligence (ICTAI) (pp. 141-148). IEEE. https://doi.org/10.1109/ICTAI.2019.00028

Ramachandra, B., Jones, M., & Vatsavai, R. R. (2020). *A survey of single-scene video anomaly detection*. IEEE transactions on pattern analysis and machine intelligence. https://doi.org/10.1109/TPAMI.2020.3040591

Teed, Z., & Deng, J. (2020, August). *Raft: Recurrent all-pairs field transforms for optical flow*. In European conference on computer vision (pp. 402-419). Springer, Cham. https://doi.org/10.1007/978-3-030-58536-5_24

Zhu, S., Chen, C., & Sultani, W. (2020). *Video anomaly detection for smart surveillance*. In Computer Vision: A Reference Guide (pp. 1-8). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03243-2_845-1

Vo, D. T., Tran, T. M., Vo, N. D., & Nguyen, K. (2021, December). *UIT-Anomaly: A Modern Vietnamese Video Dataset for Anomaly Detection*. In 2021 8th NAFOSTED Conference on Information and Computer Science (NICS) (pp. 352-357). IEEE. https://doi.org/10.1109/NICS54270.2021.9701556