Sử dụng mạng sinh ảnh (GAN) để phát hiện ảnh bất thường trong bộ dữ liệu ảnh chụp nhà trạm

Lê Ngọc Thiện
Email: thienlengoc.engineer@gmail.com

Nho Minh Tú
Email: @.com

Nguyễn Hoàng Việt
Email: @.com

***Tóm tắt nội dung*—Trong thực tế ở Viettel, bài toán phát hiện ảnh bất thường có thể được áp dụng để giải quyết rất nhiều vấn đề, một trong số đó là kiểm tra các ảnh chụp nhà trạm. Các phương pháp phát hiện bất thường truyền thống cần nhiều nhân lực và tài nguyên nhưng vẫn không thể đáp ứng các yêu cầu về độ chính xác theo thời gian thực. Đồng thời những thuật toán học máy phân loại truyền thống (SVM, Decision Trees,..) khó có thể phát hiện bất thường ở các mẫu chưa từng được huấn luyện. Trong nghiên cứu này, chúng tôi sẽ tập chung triển khai và đánh giá việc ứng dụng các mô hình học sâu - mô hình Đối sinh (GANs) cụ thể là GANomaly thuộc để phát hiện ảnh bất thường trên tập dữ liệu nhà trạm thực tế ở Viettel. Báo cáo này cũng so sánh kết quả của mô hình sinh ảnh (GANs) với một mô hình học sâu phát hiện bất thường thuộc nhóm phương pháp khác là STFPM. Kết quả của nghiên cứu này cho thấy năng lực phát hiện bất thường của GANomaly, đồng thời cũng làm nổi bật các điểm mạnh và hạn chế của nó.**

***Index Terms*—anomaly detection, GANomaly, GANs, STFPM, MVTec AD**

I. Giới thiệu

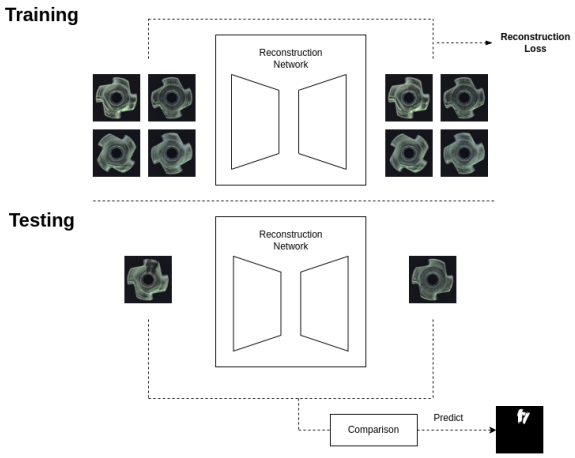
Theo định kỳ, nhân viên nhà trạm sẽ đi chụp ảnh nhà trạm nhằm phục vụ mục đích bảo dưỡng nhà trạm. Và cùng với sự phát triển của các mô hình học sâu đặc biệt là trong xử lý hình ảnh, chúng tôi đã khảo sát việc ứng dụng các mô hình học sâu cho nhiệm vụ kiểm tra ảnh nhân viên chụp để phát hiện các đối tượng bất thường. Sự bất thường của ảnh có thể được ví dụ trong các trường hợp sau: chụp sai hướng dẫn, vật thể trong ảnh bị hỏng hóc, thiếu sót. Tập dữ liệu để huấn luyện của bài toán phát hiện bất thường có đặc điểm chỉ thiên về một lớp là các ảnh được gán nhãn bình thường do việc thu thập mẫu và gán nhãn cho lớp bất bình thường sẽ tốn chi phí lớn để có thể thu thập và cũng không thể bao quát được tất cả các trường hợp bất thường có thể xảy ra. Vậy nên, trong nghiên cứu này chúng tôi sẽ tập trung xem xét đến các mô hình thuật toán học không giám sát để giải quyết bài toán trên.

Dựa trên các thuật toán đã được phát triển và thử nghiệm trên tập dữ liệu MVTec AD [1], các mô hình học không giám sát trong lĩnh vực này hiện tại có thể được chia thành hai nhóm phương pháp chính là: trích xuất đặc trưng (feature-embedding) và tái tạo ảnh (reconstruction). Trong nghiên cứu này, chúng tôi đã chọn mô hình mạng Đối sinh (GANs) thuộc phương pháp tái tạo ảnh (reconstruction) là GANomaly để phát triển một mô hình phát hiện bất thường trên tập dữ liệu MVTec AD và ảnh nhà trạm thực tế của Viettel. Đồng thời, một mô hình thuộc phương pháp trích xuất đặc trưng (feature-embedding) là STFPM cũng được triển khai để có thể đánh giá được điểm mạnh và các hạn chế của mô hình đã đề xuất.

II. Related works

A. *Phương pháp tái tạo ảnh (sử dụng mạng sinh ảnh) trong bài toán phát hiện bất thường*

Đây là nhóm các mô hình dựa trên việc tự đạo tạo bộ mã hóa (encoder) và bộ giải mã (decoder) để tái tạo hình ảnh nhằm phát hiện những điểm bất thường.



Hình 1. Cấu trúc của các mô hình phát hiện bất thường bằng phương pháp tái tạo.

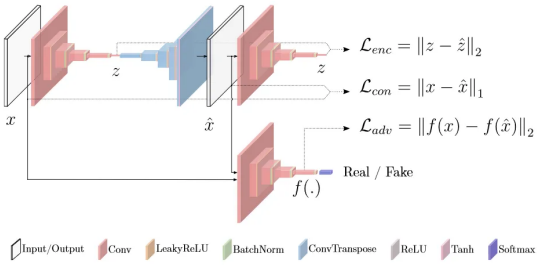
Cấu trúc của các mô hình thuộc phương pháp tái tạo ảnh được mô tả trong hình 1. Trong quá trình đào tạo, các hình ảnh bất thường sẽ được gửi đến một mạng tái tạo ảnh (reconstruction network) và hàm mất mát của quá trình này sẽ được sử dụng để cải thiện khả năng tái tạo của mạng. Trong quá trình kiểm thử, để có thể đưa ra dự đoán, mô hình sẽ so sánh ảnh gốc với ảnh được tái tạo. Các mô hình trong phương pháp này chủ yếu sẽ khác nhau ở việc xây dựng các mạng tái tạo (reconstruction network) ví dụ như: Autoencoder, GANs, Transformer, Diffusion,...

B. *GANomaly*

Vào năm 2014, Goodfellow đã giới thiệu mạng Đối sinh (GANs) lần đầu tiên [2]. Sau đó GANs được biết đến rộng rãi với việc có thể tạo ra các hình ảnh chân thực bằng cách huấn luyện song song một cặp mạng sinh và phân biệt (generator and discriminator). Sau đó, GANs cũng được ứng dụng để

làm mạng sinh ảnh trong bài toán phát hiện bất thường. Và GANomaly [3] cũng đã được Samet Akcay giới thiệu vào năm 2018 dựa trên ý tưởng này.

GANomaly là một mô hình phát hiện bất thường sử dụng kiến trúc của mạng Đối sinh (GANs) để học phân phối của dữ liệu. GANomaly thực hiện một thay đổi so với các thuật toán thuộc nhóm phương pháp tái tạo ảnh là đưa ra dự đoán dựa trên việc so sánh không gian đặc trưng tiềm ẩn (latent space) thu được từ lần mã hóa đầu tiên và không gian đặc trưng tiềm ẩn thu được ở lần mã hóa thứ hai.



Hình 2. Kiến trúc của GANomaly.

Trong quá trình đào tạo, mạng sẽ chỉ được huấn luyện với đầu vào là những mẫu thuộc lớp bình thường, nên thì theo giả định bộ tái tạo sẽ chỉ học được phân phối của các mẫu bình thường. Ở quá trình suy luận, nếu đầu vào là các mẫu bất thường, khoảng cách giữa hai không gian đặc trưng tiềm ẩn sẽ lớn do không được học trong quá trình đào tạo. Khi khoảng cách giữa hai không gian đặc trưng tiềm ẩn của một ảnh đầu vào vượt qua một ngưỡng nhất định, thì nó được xác định là một mẫu bất thường.

Có thể chia cấu trúc của GANomaly thành 3 thành phần chính:

- **Bộ tái tạo (generator)**, thực chất là một bộ mã hóa tự động (autoencoder) được dùng để học cách tái tạo lại đầu vào.
- **Bộ phân biệt (discriminator)**, được xử dụng để phân biệt ảnh đầu vào và ảnh sau khi được tái tạo bởi bộ sinh (generator). Qua đó có thể cải thiện chất lượng của kết quả của bộ tái tạo.
- **Bộ mã hóa (encoder) thứ hai**, ánh xạ ảnh sau khi tạo bởi phần sinh thành một không gian các đặc trưng tiềm ẩn (latent space) - z'.

Trong quá trình đào tạo, hàm mục tiêu của mô hình là kết hợp của ba hàm mất mát sau:

- **Adversarial Loss**: là khoảng cách L2 giữa biểu diễn đặc trưng của ảnh gốc x và biểu diễn đặc trưng của ảnh được tạo bởi bộ tái tạo generator $x' = G(x)$. Trong hàm mất mát này, $f(x)$ là đầu ra của bộ phân loại của bộ phân biệt (discriminator):

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p_x} \left\| f(x) - \mathbb{E}_{x \sim p_x} f(G(x)) \right\|_2 . \quad (1)$$

- **Contextual Loss**: là khoảng cách L1 giữa đầu vào ban đầu x và hình ảnh được tạo $G(x)$. Hàm mất mát này giúp hàm mục tiêu có thêm thông tin về ngữ cảnh của đầu vào:

$$\mathcal{L}_{con} = \mathbb{E}_{x \sim p_x} \left\| x - G(x) \right\|_1 . \quad (2)$$

- **Encoder Loss**: là khoảng cách L2 giữa các không gian đặc trưng tiềm ẩn của đầu vào $z = G_E(x)$ và của hình ảnh được tạo bằng bộ sinh $z' = E(G(x))$. Với hàm mất mát này, bộ tái tạo (generator) sẽ được hướng dẫn để hiểu cách mã hóa các đặc trưng của các mẫu bình thường:

$$\mathcal{L}_{enc} = \mathbb{E}_{x \sim p_x} \left\| G_E(x) - E(G(x)) \right\|_2 . \quad (3)$$

Và hàm mục tiêu được xác định với việc thêm trọng số vào các hàm mất mát trên:

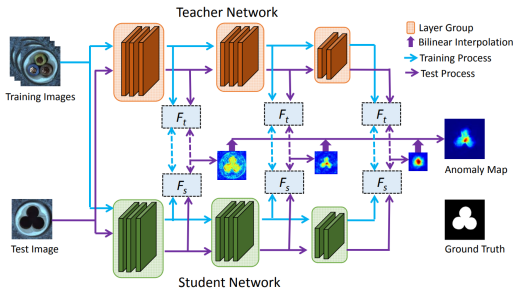
$$\mathcal{L} = w_{adv}\mathcal{L}_{adv} + w_{con}\mathcal{L}_{con} + w_{enc}\mathcal{L}_{enc} \quad (4)$$

Trong quá trình kiểm thử, độ bất thường của ảnh thử nghiệm được định nghĩa theo hàm Encoder Loss:

$$\mathcal{A}(\hat{x}) = \left\| G_E(\hat{x}) - E(G(\hat{x})) \right\|_1 \quad (5)$$

C. *STFPM*

STFPM [4] được giới thiệu vào năm 2021, là một cách tiếp cận bài toán phát hiện bất thường đơn giản nhưng đạt hiệu suất cao.



Hình 3. Kiến trúc của STFPM.

Kiến trúc của phương pháp này gồm hai mạng là giáo viên và học sinh. Mạng giáo viên sẽ là một mạng mạnh mẽ đã được đào tạo trước trên các tập dữ liệu lớn, về phân loại hình ảnh (ví dụ: ResNet-18 được đào tạo trên ImageNet) và mạng học sinh sẽ có cùng kiến trúc với mạng giáo viên nhưng chưa được trải qua quá trình đào tạo. Trong quá trình đào tạo của mô hình này, mạng giáo viên sẽ "truyền đạt", hướng dẫn cho mạng học sinh các kiến thức về trích chọn đặc trưng của ảnh với đầu vào là các hình ảnh được gán nhãn bình thường. Và trong quá trình suy luận, các đặc điểm về mẫu bất thường được trích xuất từ mạng học sinh và giáo viên sẽ tương đương nhau, trong khi các đặc điểm được trích xuất từ ảnh bất thường sẽ khác biệt. Bằng cách so sánh bản đồ đặc trưng (feature maps) được tạo bởi hai mạng, ta có thể tạo ra được bản đồ các điểm bất thường từ đó xác định được ảnh bất thường trong tập dữ liệu.

III. THIẾT LẬP THÍ NGHIỆM

Trong phần này, chúng tôi sẽ giới thiệu về các tập dữ liệu đã được dùng để thử nghiệm và đánh giá phương pháp đã đề xuất, đồng thời là cách thiết lập quá trình huấn luyện cũng như kiểm thử.

A. *Dữ liệu*

a) MVTec AD: [23] là tập dữ liệu dữ liệu được thu thập bởi MVTec software GmbH và được sử dụng rộng rãi để đánh giá các mô hình phát hiện bất thường. Tập dữ liệu được chia thành 15 tập dữ liệu con của 15 vật thể. Mỗi tập dữ liệu con đều có đầy đủ tập huấn luyện với các hình ảnh được gán nhãn là bình thường và tập kiểm thử với các hình ảnh bình thường cũng như của các trường hợp bất thường khác nhau. Chúng tôi đã chọn ra tập dữ liệu của 4 trong số 15 vật thể để tiến hành thử nghiệm đó là thảm (Carpet), lưới (Grid), quả phỉ (Hazelnut), điện trở (Transistor).

b) Tập dữ liệu nhà trạm thực tế của Viettel: dữ liệu được Viettel cung cấp bao gồm hai tập là tập huấn luyện và tập kiểm thử. Tập huấn luyện bao gồm 200 hình ảnh được gán nhãn là bình thường. Tập kiểm thử có 99 hình ảnh, trong đó 59 ảnh được gán nhãn bình thường và 40 ảnh được gán nhãn là bất bình thường được chia thành hai nhóm là chụp sai hướng và chụp nhầm đối tượng.

B. *Tiền xử lý dữ liệu*

a) MVTec AD:

C. *Thiết lập mô hình*

Chúng tôi đã triển khai huấn luyện và kiểm thử các mô hình trên NVIDIA RTX 3060 - 6GB graphics memory.

a) GANomaly: Đầu tiên chúng tôi sẽ giới thiệu về tham số của mạng GANomaly đã được chúng tôi sử dụng để huấn luyện trên tập dữ liệu nhà trạm của Viettel. Theo khảo sát kích thước các ảnh có trong tập dữ liệu, các ảnh có tỉ lệ kích thước là 4:3 chiếm đa số. Vì vậy, chúng tôi đã quyết định sẽ huấn luện GANomlay với ảnh đầu vào được thay đổi kích thước về 86:64. Và sau đó là thiết kế lại các tầng tích chập (convolutional layers) và các tầng tích chập chuyển vị (transposed convolutional) để phù hợp với đầu vào mới này của dữ liệu.

Trong quá trình huấn luyện, mạng Đối nghich được tối ưu hóa bằng trình tối ưu hóa Adam với learning rate khởi tạo là $lr = 0.0006$, momentumns $\beta_1 = 0.5$, $\beta_2 = 0.999$. Mô hình được tối ưu dựa trên hàm mục tiêu $\mathcal{L}$ (được định nghĩa ở phương trình (4)) với các trọng số $w_{bce} = 1, w_{rec} = 50, w_{enc} = 1$, được chọn dựa trên đề xuất của nhóm tác giả GANomaly.

b) STFPM: Chúng tôi lựa ResNet18 là kiến trúc của mạng giáo viên và mạng học sinh, và mạng giáo viên được khởi tạo với bộ trọng số đã được huấn luyện trên tập dữ liệu ImageNet. Tương tự như với GANomaly, chúng tôi đã lữa chọn đầu vào của mô hình STFPM sẽ được thay đổi kích thước về 172:128. Đồng thời lựa chọn kích thước của bản đồ đặc trưng là 86:64. Mô hình được tối ưu hóa với trình tối ưu hóa SGD cùng bộ tham số khởi tạo: $learningrate = 0.4, momentum = 0.9, weightdecay = 0.0001$

D. *Độ đo đánh giá*

Do đây bản chất là một bài toán phân loại ảnh, nên chúng tôi đã sử dụng những phép đo sau để đánh giá mô hình.

*a) **AUC-ROC**:* là một phép đo đánh giá hiệu suất của một mô hình phân loại. AUC-ROC được tính dựa trên đường cong ROC, đồ thị biểu thị mối quan hệ giữa tỷ lệ true positive (TPR) và tỷ lệ false positive (FPR) tại các ngưỡng quyết định khác nhau và cung cấp một cái nhìn tổng quan về hiệu suất tổng thể của mô hình phân loại mà không phụ thuộc vào một ngưỡng quyết định cụ thể.

*b) **F2-score**:* : là giá trị trung bình điều hòa giữa recall và precision. Tuy nhiên F2-score chú trọng hơn vào việc đạt được recall cao so với precision. Có nghĩa là phép đo này sẽ đánh giá việc bỏ sót những trường hợp là bất thường trong thực tế sẽ nguy hiểm hơn so với việc dự đoán nhầm một ảnh là bất thường. Phép đo này rất phù hợp với bài toán này, vì việc bỏ sót những đối tượng bất thường có thể gây thiệt hại rất nghiêm trọng.

$$F2 = 5 * \frac{precision * recall}{4 * precision + recall}$$

*c) **Accuracy**:* tính toán tỷ lệ các dự đoán đúng (true predictions) trên tổng số dự đoán. Phép đo này được bổ sung để khắc phục hạn chế của F2-score là chỉ tập trung vào recall, precision mà bỏ qua các yếu tố như True Positive.

IV. KẾT QUẢ

A. *Bộ dữ liệu MVTec AD*

Do dữ liệu trong tập kiểm thử của các tập trong bộ dữ liệu MVTec AD bị mất cân bằng (số lượng mẫu bất thường nhiều hơn mẫu bình thường) nên chúng tôi chỉ xem xét các thử nghiệm với phép đo AUC-ROC.

Bảng I
KẾT QUẢ THỰC NGHIỆM TRÊN TẬP DỮ LIỆU MVTEC AD

Category	AUC-ROC	
	STFPM	GANomaly
Carpet	0.9855	0.8138
Grid	0.9758	0.8312
Hazelnut	0.9868	0.8539
Transistor	0.9375	0.8433

Theo như kết quả ở Bảng 1, ta có thể thấy hiệu suất của STFPM được thể hiện rất tốt và vượt trội hoàn toàn so với GANomaly. Điều có thể dễ dàng nhận thấy do STFPM là mô hình mạnh mẽ được tiền huấn luyện trên một tập dữ liệu lớn. Trong khi đó, khả năng phân loại ảnh của GANomaly kém hơn do nó không thể trích xuất được các đặc trưng mang tính ngữ nghĩa cao (high level semantic features).

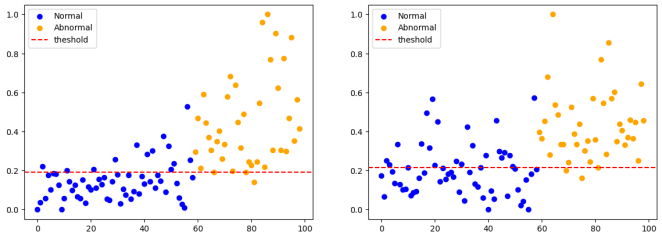
B. *Bộ dữ liệu nhà trạm*

Do thuật toán GANomaly sẽ suy luận kết quả dựa trên sự khác biệt giữa hai vetor đặc trưng tiềm ẩn z và z' nên việc lựa chọn kích thước của vector này ảnh hưởng rất lơn đến hiệu suất của mô hình. Vậy nên chúng tôi đã tiến hành thử nghiệm với 3 kích thước khác nhau của vector đặc trưng tiềm ẩn là 256, 512, 640 và thu được kết quả như trong bảng II.

Bảng II
KẾT QUẢ THỰC NGHIỆM TRÊN TẬP DỮ LIỆU NHÀ TRẠM CỦA VIETTEL

Độ đo	STFPM	GANomaly		
		256	512	640
AUC-ROC	0.9326	0.8275	0.8445	0.7864
F2-score	0.9198	0.8333	0.8676	0.8
Accuracy	0.8485	0.6666	0.7576	0.6565

Theo như kết quả ở Bảng 2, có thể thấy rằng mô hình GANomaly đạt kết quả tốt nhất với kích thước vector tiềm ẩn là 512. Cũng như trên tập dữ liệu MVTec AD, mô hình GANomaly đã đạt hiệu suất khá khả quan, tuy nhiên vẫn chưa thể bằng mô hình STFPM và đáp ứng được kỳ vọng để có thể áp dụng vào thực tế.



Hình 4. Độ bất thường của các ảnh trong tập kiểm thử nhà trạm Viettel a) STFPM b) GANomaly

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã tiến thành thử nghiệm thuật toán sinh ảnh GANomaly để áp dụng vào bài toán phát hiện bất thường. Kết quả cho thấy GANomaly vẫn chưa thể áp dụng để giải quyết các bài toán thực tế.

Kết quả của các thí nghiệm trên có thể chưa phản ánh được đầy đủ khả năng của các mạng GANomaly trong bài toán phát hiện bất thường, do điểm mạnh của các thuật toán ở phương pháp tái tạo ảnh là ở việc chúng có thể so sánh ở cấp độ pixel - điều mà đã không được đề cập trong nghiên cứu này. Bên cạnh đó, do giới hạn của phần cứng trong việc thí nghiệm, chúng tôi đã phải điều chỉnh kích thước đầu vào của GANomaly về kích thước khá nhỏ 86:64 nên đã phần nào đó làm mất đi thông tin của ảnh. Tuy nhiên nó cũng chỉ ra rằng việc huấn luyện các mô hình tái tạo ảnh ví dụ như GANomaly cho bài toán phát hiện bất thường cần nhiều tài nguyên tính toán hơn một số phương pháp có hiệu suất cao ví dụ như STFPM. Ngoài ra, quá trình để huấn luyện GANs cũng là một thách thức lớn. Để có thể đảm bảo được hai phần mạng sinh (generator) và mạng phân loại (discriminator) có thể ổn định học song song cùng với nhau thì cần phải lựa chọn các siêu tham số một cách cẩn thận sao cho tốc độ học của hai mạng này đồng đều nhau. Vì vậy ở các nghiên cứu trong tương lai, chúng tôi sẽ nghiên cứu ứng dụng các mạng tái tạo ảnh khác ưu việt hơn như Transformer, Diffusion trong bài toán phát hiện bất thường.

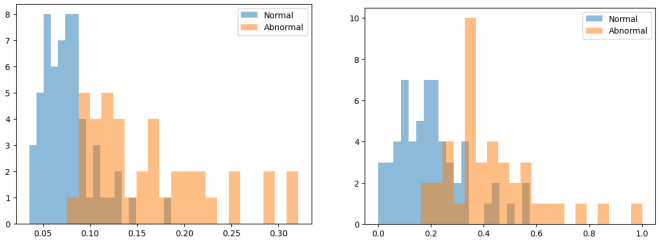
TÀI LIỆU

[1] Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C. (2021). The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. International Journal of Computer Vision, 129(4), 1038–1059.

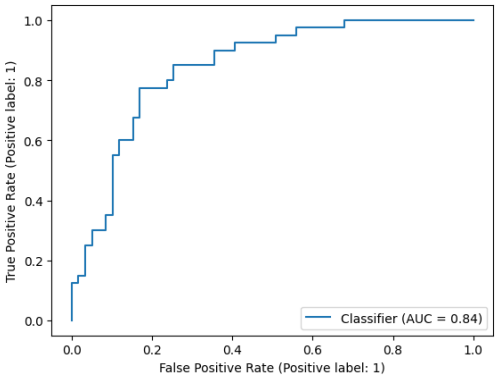
[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672–2680).

[3] Akcay, S., Atapour-Abarghouei, A., Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14 (pp. 622-637). Springer International Publishing.

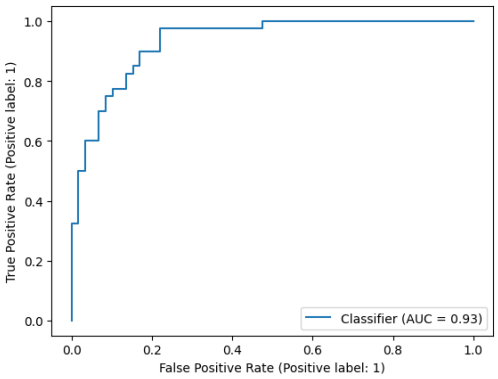
[4] Wang, G., Han, S., Ding, E., Huang, D. (2021). Student-teacher feature pyramid matching for anomaly detection. arXiv preprint arXiv:2103.04257.



Hình 5. Histogram độ bất thường các ảnh trong tập kiểm thử nhà trạm Viettel a) STFPM b) GANomaly



Hình 6. ROC curve của GANomaly trên tập kiểm thử nhà trạm.



Hình 7. ROC curve của STFPM trên tập kiểm thử nhà trạm.