

ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN



TRẦN MINH TÙNG

PHÁT HIỆN SỰ KIỆN BẤT THƯỜNG TRONG VIDEO
(Abnormal Event Detection in Videos)

Ngành: Khoa học máy tính

Mã số: 948010

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC
TS. NGUYỄN TẤN TRẦN MINH KHANG
PGS.TS. NGUYỄN VĂN TÂM

TP. HỒ CHÍ MINH - NĂM 2025

Công trình được hoàn thành tại: **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN - ĐẠI HỌC QUỐC GIA TP.HCM.**

Cán bộ hướng dẫn khoa học:

TS. Nguyễn Tấn Trần Minh Khang

PGS.TS. Nguyễn Văn Tâm

Phản biện 1:

.....

Phản biện 2:

.....

Phản biện 3:

.....

Phản biện độc lập 1:

.....

Phản biện độc lập 2:

.....

Luận án được bảo vệ trước Hội đồng chấm luận án họp tại: Trường Đại học Công nghệ thông tin - Đại học Quốc gia TP. HCM.

Vào lúc: . . . giờ . . . ngày . . . tháng . . . năm 2025.

Có thể tìm hiểu luận án tại thư viện:

- Thư viện Quốc gia Việt Nam.
- Thư viện Đại học Quốc gia TP.HCM.
- Thư viện Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM.

Mục lục

Chương 1. GIỚI THIỆU BÀI TOÁN.....	1
1.1 Cơ sở nghiên cứu	1
1.2 Mục đích, mục tiêu và nội dung nghiên cứu.....	1
1.2.1 Mục đích nghiên cứu.....	1
1.2.2 Mục tiêu nghiên cứu	2
1.2.3 Nội dung nghiên cứu.....	2
1.3 Phát biểu bài toán phát hiện sự kiện bất thường trong video	2
1.3.1 Phát biểu bài toán tổng quát.....	2
1.3.2 Phát biểu bài toán của luận án.....	3
1.4 Độ phức tạp nghiên cứu.....	4
1.5 Thách thức của bài toán nghiên cứu	4
1.6 Ý nghĩa khoa học và thực tiễn của luận án	4
1.7 Bố cục của luận án.....	5
Chương 2. NGHIÊN CỨU TỔNG QUAN.....	5
2.1 Các khái niệm	6
2.2 Phương pháp luận cho bài toán phát hiện sự kiện bất thường trong video.....	6
2.3 Các phương pháp phát hiện sự kiện bất thường trong video	6
2.4 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video với camera có góc quay cố định.....	7
2.4.1 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video trên không.....	7
2.4.2 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video.....	7
Chương 3. ĐỀ XUẤT PHƯƠNG PHÁP PHÁT HIỆN SỰ KIỆN BẤT THƯỜNG TRONG VIDEO TRÊN KHÔNG VỚI	

NGŨ CẢNH GIAO THÔNG.....	8
3.1 Phương pháp đề xuất	8
3.1.1 Các công thức sử dụng trong phương pháp đề xuất.....	9
3.1.2 Bộ mã hóa Transformer theo không gian - thời gian (Spatio - Temporal Transformer Encoder)	11
3.1.3 Bộ mã hóa Transformer theo không gian (Spatial Transformer Encoder).....	12
3.1.4 Bộ mã hóa Transformer theo thời gian (Temporal Transformer Encoder).....	12
3.1.5 Cơ chế chú ý chéo theo thời gian (Temporal Cross - Attention)	12
3.1.6 Phát hiện khung hình có chứa sự kiện bất thường dựa trên dự đoán khung hình tương lai	14
3.2 Ưu điểm và hạn chế của phương pháp đề xuất	15
Chương 4. XÂY DỰNG BỘ DỮ LIỆU VIDEO TRÊN KHÔNG VỚI NGŨ CẢNH GIAO THÔNG.....	16
4.1 Mô tả bộ dữ liệu UIT-ADrone.....	16
4.2 Thu thập dữ liệu.....	16
4.3 Thống kê bộ dữ liệu	17
4.4 Gán nhãn cho bộ dữ liệu	17
Chương 5. THỰC NGHIỆM VÀ KẾT QUẢ.....	20
5.1 Các bộ dữ liệu chuẩn dùng để thực nghiệm.....	20
5.2 Độ đo đánh giá.....	20
5.3 So sánh với các phương pháp tối tân hiện tại	21
5.4 Nghiên cứu cắt lát.....	24
Chương 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....	24
6.1 Kết quả đạt được.....	24

6.2	Các đóng góp chính của luận án	24
6.3	Các công trình nghiên cứu đã công bố.....	25
6.3.1	Các công bố chính của luận án	25
6.3.2	Các công bố khác có liên quan đến luận án.....	25
6.3.3	Các đề tài nghiên cứu khoa học đã tham gia	26
6.4	Hướng phát triển.....	26

Chương 1. GIỚI THIỆU BÀI TOÁN

1.1 Cơ sở nghiên cứu

Trong những năm gần đây, bài toán phát hiện sự kiện bất thường trong video đối mặt với những thách thức mới. Các sự kiện bất thường như va chạm giao thông, cướp giật, hỏa hoạn, lái xe sai phần đường, xe đi lùi trên đường cao tốc và tội phạm trên đường phố gần đây đã trở thành chủ đề ngày càng được quan tâm trong các hệ thống giao thông thông minh [13], [115], [134], [154]. Đây vẫn là một bài toán rất khó khăn do phải xử lý đồng thời một lượng lớn dữ liệu video giám sát và ngày càng gia tăng theo thời gian, đặc biệt trong môi trường giao thông vì các cảnh quay video chứa nhiều đối tượng và phương tiện chuyển động đồng thời, điều kiện ánh sáng, độ phân giải thấp, các đối tượng có kích thước khác nhau và dữ liệu chứa sự kiện bất thường chỉ chiếm một số lượng rất nhỏ so với dữ liệu bình thường. Nhờ những tiến bộ trong các kiến trúc mạng, thuật toán huấn luyện, tài nguyên tính toán và công nghệ trí tuệ nhân tạo (Artificial Intelligence - AI), phát hiện sự kiện bất thường trong video giao thông được quay bằng máy bay không người lái đã và đang được triển khai để tăng cường phạm vi bao phủ cũng như giám sát an toàn và an ninh công cộng. So với video giám sát từ camera có góc quay cố định (fixed camera), video được quay bằng phương tiện bay không người lái có nhiều thách thức hơn như sau: camera chuyển động thay vì camera tĩnh; đối tượng có kích thước nhỏ so với kích thước đối tượng trong camera tĩnh; độ phân giải không gian thay đổi do các đối tượng không gian địa lý được quay ở các độ cao bay khác nhau, điều kiện thời tiết và ánh sáng thay đổi. Do đó, vẫn còn nhiều khó khăn và thách thức cho nhiệm vụ phát hiện sự kiện bất thường trong video từ góc nhìn trên không cũng như sự khan hiếm bộ dữ liệu cho bài toán phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông.

1.2 Mục đích, mục tiêu và nội dung nghiên cứu

1.2.1 Mục đích nghiên cứu

Nghiên cứu và đề xuất phương pháp phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông. Bên cạnh đó, hướng đến phát triển các ứng dụng thực tế để tự động phát hiện sự kiện bất thường trong hệ thống giám sát giao thông nhằm đưa ra phản ứng kịp thời cho các cơ quan thực thi pháp luật trong các thành phố thông minh

để ngăn chặn và hạn chế hành vi và hoạt động tội phạm và sự cố đe dọa đến an toàn và an ninh của các phương tiện và những người tham gia giao thông, từ đó cải thiện thời gian ứng phó trong các tình huống khẩn cấp và có thể mang lại hiệu quả cao trong việc duy trì trật tự giao thông, đặc biệt là giao thông đô thị ở các thành phố lớn với mật độ và phương tiện tham gia giao thông đông đúc.

1.2.2 Mục tiêu nghiên cứu

Mục tiêu thứ nhất: Nghiên cứu, khảo sát và hệ thống hóa các cách tiếp cận và phương pháp nghiên cứu, các bộ dữ liệu chuẩn công khai, hiệu suất các phương pháp tối tân hiện tại, các độ đo phổ biến cho bài toán phát hiện sự kiện bất thường trong video, thảo luận chuyên sâu về các thách thức và ứng dụng của bài toán của bài toán nghiên cứu.

Mục tiêu thứ hai: Đề xuất phương pháp phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông.

Mục tiêu thứ ba: Thu thập và xây dựng bộ dữ liệu cho bài toán phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông tại TP.HCM, Việt Nam. Bộ dữ liệu này được công bố công khai phục vụ cộng đồng nghiên cứu vì mục đích phi lợi nhuận.

1.2.3 Nội dung nghiên cứu

Nội dung 1: Nghiên cứu, khảo sát các bài toán và các phương pháp giải quyết bài toán phát hiện sự kiện bất thường trong video.

Nội dung 2: Nghiên cứu và đề xuất phương pháp phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông.

Nội dung 3: Thu thập và xây dựng bộ dữ liệu cho bài toán phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông tại TP.HCM, Việt Nam.

1.3 Phát biểu bài toán phát hiện sự kiện bất thường trong video

1.3.1 Phát biểu bài toán tổng quát

Qua các nghiên cứu đã công bố, bài toán tổng quát cho nhiệm vụ phát hiện sự kiện bất thường trong video được phát biểu như sau:

- Dữ liệu vào: Video.
- Dữ liệu ra là một trong bốn dạng sau:
 - + Xác định video có sự kiện bất thường hay không? [22, 128, 129].
 - + Các đoạn video (video segments) chứa sự kiện bất thường [115, 145, 149].
 - + Các khung hình (frame) có chứa sự kiện bất thường [64, 77, 84,

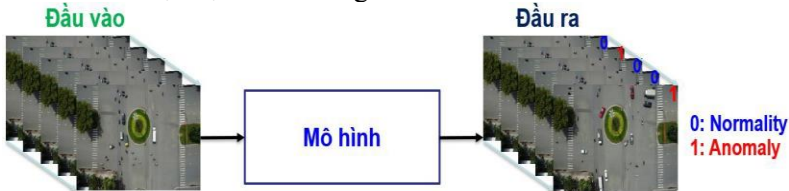
129, 140].

+ Các khung hình và vùng bất thường [19, 37, 100, 120, 122].

1.3.2 Phát biểu bài toán của luận án

Bài toán nghiên cứu của luận án là bài toán phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông được phát biểu và trình bày như Hình 1.1.

- Dữ liệu vào: Video.
- Dữ liệu ra: Chuỗi các khung hình với nhãn tương ứng.
 - + 0: Không có sự kiện bất thường.
 - + 1: Có sự kiện bất thường.



Hình 1.1: Bài toán phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông.

Nguyên lý của bài toán phát hiện sự kiện bất thường ở mức khung hình trong video trên không theo chiến lược học không giám sát (unsupervised learning) là học các mẫu bình thường từ các khung hình video và tự động xác định khung hình có sự khác biệt so với các khung hình đã học trước đó là khung hình có chứa sự kiện bất thường trong giai đoạn kiểm tra. Nghĩa là mô hình chỉ huấn luyện với dữ liệu bình thường và đánh giá bằng cách so sánh khung hình tương lai được tạo sinh với các khung hình liền kề trước đó và nhãn thực tế của nó trong giai đoạn kiểm tra. Bên cạnh đó, mô hình có thể mở rộng do có khả năng quát hóa sang các miền và môi trường khác nhau. Ngoài ra, mô hình mạnh mẽ trước các sự kiện bất thường mới vì chỉ học các mẫu dữ liệu bình thường nên mô hình có thể phát hiện ra các sự bất thường chưa từng thấy trước đó.

Phạm vi của bài toán nghiên cứu:

- Ngữ cảnh: Video trên không với ngữ cảnh giao thông.
- Phát hiện sự kiện bất thường ở mức khung hình. Trong đó, khung hình có chứa sự kiện bất thường được gán nhãn là 1, ngược lại được gán nhãn là 0.
- Góc quay từ trên cao với không gian và trường nhìn rộng hơn.
- Camera chuyển động với độ cao từ 50m – 70m.

- Điều kiện thời tiết và ánh sáng thuận lợi.

1.4 Động lực nghiên cứu

- Góp phần tăng cường khả năng khai thác dữ liệu giám sát giao thông tại Việt Nam.

- Lấy cảm hứng từ các nghiên cứu mới nhất và thành công gần đây của các phương pháp tối tân hiện tại [11, 15, 19] đối với các bài toán liên quan trong lĩnh vực thị giác máy tính bao gồm: phân lớp ảnh [99, 170], phân loại video [101], phát hiện đối tượng [102, 103, 150] và phát hiện sự kiện bất thường trong video [77, 135, 139].

- Sự phổ biến của thiết bị bay không người lái (drone) trong lĩnh vực giao thông và các lĩnh vực khác bao gồm: quân sự, chăm sóc y tế từ xa, cứu hộ cứu nạn, nông nghiệp thông minh, dịch vụ giao hàng, kiểm tra cơ sở hạ tầng,... [165, 166, 167, 168].

1.5 Thách thức của bài toán nghiên cứu

Các thách thức của bài toán phát hiện sự kiện bất thường trong video trên không bao gồm:

- Môi trường giao thông phức tạp (the complex traffic environment) [11, 27, 89, 115, 134, 154].

- Camera chuyển động, góc quay và độ cao thay đổi [11, 27, 89, 115, 134, 154].

- Đối tượng và phương tiện tham gia giao thông có kích thước nhỏ và đa dạng [44, 45, 47, 135].

- Sự đa dạng các sự kiện bất thường trong thế giới thực [3, 4, 9, 11, 126].

- Tính sẵn có của dữ liệu (availability of data) cho bài toán phát hiện sự kiện bất thường trong video còn hạn chế [87, 135, 180].

1.6 Ý nghĩa khoa học và thực tiễn của luận án

Các nội dung nghiên cứu của luận án có đóng góp về mặt khoa học và tiềm năng ứng dụng của bài toán nghiên cứu trong thực tế. Cụ thể:

- Luận án đề xuất phương pháp phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông bằng cách dự đoán khung hình tương lai dựa trên kiến trúc Vision Transformer với chiến lược học không giám sát thông qua việc khai thác biểu diễn đặc trưng không gian - thời gian trên các khung hình liên tiếp trong quá khứ để dự đoán khung hình tương lai. Bằng cách khai thác và kết hợp hiệu quả đặc trưng không gian - thời gian, phương pháp đề xuất dự đoán các khung hình tương lai với các đặc trưng không gian - thời gian có chất

lượng phong phú hơn và hiệu quả hơn thông qua việc huấn luyện chỉ với các mẫu dữ liệu bình thường và phát hiện sự kiện bất thường ở mức khung hình trong giai đoạn kiểm tra. Hơn nữa, đề tận dụng tối đa sức mạnh của cơ chế tự chú ý, luận án cũng tiến hành áp dụng cơ chế tự chú ý chéo (cross-self-attention) để nâng cao hiệu suất của mô hình đề xuất. Thực nghiệm được tiến hành và đánh giá trên các bộ dữ liệu chuẩn công khai và so sánh với các phương pháp tối tân hiện tại (state-of-the-arts) để xác thực tính hiệu quả của phương pháp đề xuất.

- Luận án tiến hành khảo sát, nghiên cứu và phân loại các bài toán bất thường trong video gồm: phát hiện sự kiện bất thường, phân loại bất thường, dự đoán sự kiện bất thường và định vị sự kiện bất thường. Bên cạnh đó, hệ thống hóa các cách tiếp cận và các phương pháp nghiên cứu, các bộ dữ liệu chuẩn, các độ đo đánh giá cho bài toán phát hiện sự kiện bất thường trong video giám sát. Ngoài ra, thảo luận chuyên sâu các ưu điểm, hạn chế của các phương pháp tối tân hiện tại và gợi mở các hướng nghiên cứu tiếp theo cho bài toán phát hiện sự kiện bất thường trong video trên không.

- Luận án thu thập và xây dựng một bộ dữ liệu bất thường mới, có tên là UIT- ADrone cho bài toán phát hiện sự kiện bất thường trong video với ngữ cảnh giao thông được quay bằng thiết bị bay không người lái tại thành phố Hồ Chí Minh, Việt Nam. Bộ dữ liệu được công bố công khai dành cho mục đích nghiên cứu tại địa chỉ <https://uit-together.github.io/datasets/UIT-ADrone/>

1.7 Bố cục của luận án

Luận án được bố cục gồm các chương sau: **Chương 1:** Giới thiệu bài toán; **Chương 2:** Nghiên cứu tổng quan; **Chương 3:** Đề xuất phương pháp phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông; **Chương 4:** Xây dựng bộ dữ liệu video trên không với ngữ cảnh giao thông; **Chương 5:** Thực nghiệm và kết quả; **Chương 6:** Kết luận và hướng phát triển.

Chương 2. NGHIÊN CỨU TỔNG QUAN

2.1 Các khái niệm

- Sự kiện (an event) là một hoạt động phức tạp xảy ra tại một không gian và thời gian cụ thể liên quan đến con người, sự tương tác của người này với người khác hoặc các đối tượng khác nhau [172, 173, 174].

- Bất thường (an anomaly) có thể được xác định là các hoạt động chưa xác định hoặc các sự kiện không nhìn thấy, có khác biệt so với những gì được mong đợi, phổ biến hoặc bình thường [1, 11, 13, 15, 16, 17].



Hình 2.1: Phân loại các bài toán liên quan đến bất thường trong video [CT.2].

Bảng 2.1: Danh sách các bộ dữ liệu chuẩn cho các bài toán bất thường và một số công trình nghiên cứu liên quan [CT.2].

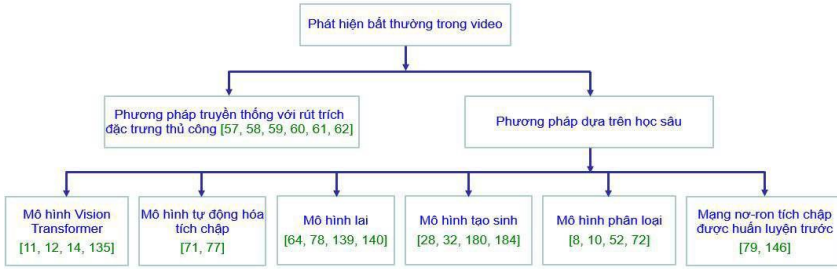
STT	Bộ dữ liệu	Loại bài toán	Công trình công bố
1	UMN [120], UCSD [122], ShanghaiTech [125], Street Scene [127]	Anomaly Detection	[77, 139, 140]
2	UCF-Crime [126], XD-Violence [128]	Anomaly Classification	[162, 163]
3	YNC [124], Pittsburgh [160], SIMCD-Prediction [161], NWPU Campus [142]	Anomaly Prediction	[142, 147, 148, 161]
4	UMN [120], UCSD [122]	Anomaly Localization	[100, 120, 122]

2.2 Phương pháp luận cho bài toán phát hiện sự kiện bất thường trong video

Trong phần này luận án tập trung trình bày cách tiếp cận chính của các phương pháp dựa trên học sâu cho bài toán phát hiện sự kiện bất thường trong video theo chiến lược học có giám sát (supervised learning), học giám sát yếu (weakly-supervised learning), học không giám sát (unsupervised learning) và học tự giám sát (self-supervised learning). Hình 2.1 phân loại các bài toán liên quan đến bất thường trong phân tích video gồm: (1) Phát hiện bất thường (Anomaly detection); (2) Phân loại bất thường (Anomaly classification); (3) Dự toán bất thường (Anomaly prediction); và (4) Định vị bất thường (Anomaly localization). Mỗi bài toán có các bộ dữ liệu chuẩn và các công trình nghiên cứu liên quan được trình bày trong Bảng 2.1.

2.3 Các phương pháp phát hiện sự kiện bất thường trong video

Luận án tiến hành nghiên cứu và hệ thống hóa các kỹ thuật và phương pháp tiếp cận khác nhau cho nhiệm vụ phát hiện sự kiện bất thường trong video thành hai loại dựa trên cuộc cách mạng học sâu: (1) Các phương pháp truyền thống với rút trích đặc trưng thủ công; và (2) Các phương pháp dựa trên học sâu. Phân loại các phương pháp phát hiện sự kiện bất thường trong video được minh họa trong Hình 2.2.



Hình 2.2: Phân loại các phương pháp phát hiện sự kiện bất thường trong video [CT.2].

2.4 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video

2.4.1 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video với camera có góc quay cố định

Các bộ dữ liệu chuẩn công khai liên quan bài toán phát hiện sự kiện bất thường trong video với camera có góc quay cố định được trình bày trong Bảng 2.2.

Bảng 2.2: Danh sách các bộ dữ liệu bất thường trong video [CT.2].

STT	Bộ dữ liệu	Năm	Độ phân giải	#Khung hình	#Loại sự kiện bất thường	Môi trường
1	UMN [120]	2006	320 x 240	7.740	3	indoor outdoor
2	Subway Entrance [121]	2008	512 x 384	71.845	66	indoor
3	Subway Exit [121]	2008	512 x 384	72.401	19	indoor
4	UCSD Ped1 [122]	2013	238 x 158	14.000	40	outdoor
5	UCSD Ped2 [122]	2013	360 x 240	4.560	12	outdoor
6	CUHK Avenue [123]	2013	640 x 360	30.652	14	indoor outdoor
7	NYC [124]	2014	720 x 480	1.500.000	10	outdoor
8	Pittsburgh [160]	2014	360 x 240	106.000	6	outdoor
9	ShanghaiTech [125]	2017	856 x 480	315.398	130	outdoor
10	Street Scene [127]	2017	1280 x 720	203.257	205	outdoor
11	UCF-Crime [126]	2018	320 x 240	13.000.000	13	indoor outdoor
12	XD-Violence [128]	2020	854 x 480	217.617	6	indoor outdoor
13	IITB Corridor [141]	2020	1920 x 1080	483.566	10	outdoor
14	LAD [129]	2021	640 x 360	1.120.000	14	indoor outdoor
15	SIMCD-Prediction [161]	2022	640 x 360	360.000	6	outdoor
16	NWPU Campus [142]	2023	720 x 1280	1.466.073	28	outdoor

2.4.2 Các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video trên không

Các bộ dữ liệu chuẩn công khai liên quan bài toán phát hiện sự kiện bất thường trong video trên không được trình bày trong Bảng 2.3.

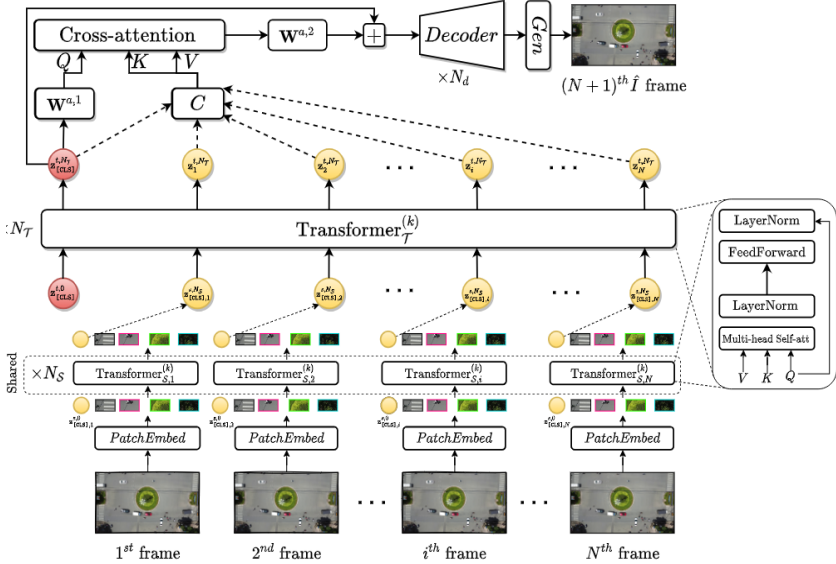
Bảng 2.3: Danh sách các bộ dữ liệu bất thường trong video trên không [CT.3].

STT	Bộ dữ liệu	Năm	Loại dữ liệu	Độ phân giải	#Khung hình	#Loại sự kiện bất thường
1	Mini Drone [130]	2015	Thực tế	224 x 224	23.295	3
2	UTT Drone [131]	2021	Thực tế	227 x 227	14.021	3
3	Brutal Running [132]	2021	Tổng hợp	1920 x 1080	1.000	1
4	AU-AIR-Anomaly [133]	2021	Thực tế	1920 x 1080	32.823	4
5	Drone-Anomaly [135]	2022	Tổng hợp	640 x 640	87.488	-

Chương 3. ĐỀ XUẤT PHƯƠNG PHÁP PHÁT HIỆN SỰ KIẾN BẤT THƯỜNG TRONG VIDEO TRÊN KHÔNG VỚI NGŨ CẢNH GIAO THÔNG

3.1 Phương pháp đề xuất

Để khai thác hiệu quả đặc trưng không gian - thời gian, luận án đề xuất một mạng dự đoán khung hình tương lai dựa trên kiến trúc Vision Transformer cho nhiệm vụ phát hiện sự kiện bất thường ở mức khung hình trong video được quay bằng máy bay không người lái với ngữ cảnh giao thông theo chiến lược học không giám sát. Hình 3.1 mô tả kiến trúc tổng thể của phương pháp đề xuất [143].



Hình 3.1: Phương pháp đề xuất cho bài toán phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông dựa trên kiến trúc Vision Transformer theo chiến lược học không giám sát [CT.1].

3.1.1. Các công thức sử dụng trong phương pháp đề xuất

Trước tiên, luận án trình bày Công thức (3.1) và Công thức (3.2) nhằm mô tả phương pháp ANDT từ công trình [135]. Cụ thể, phương pháp ANDT đã áp dụng kiến trúc Vision Transformer trong bài toán phân loại ảnh để giải quyết hiệu quả bài toán phát hiện bất thường trong video. Sau đó, luận án sẽ trình bày các mở rộng từ các công thức này trong phương pháp đề xuất cho bài toán phát hiện sự kiện bất thường ở mức khung hình trong video trên không. Cụ thể, cho một chuỗi N_f khung hình RGB $\mathbf{I} = \{I_1, I_2, \dots, I_{N_f}\}$, trong đó $I_i \in \mathbb{R}^{H \times W \times 3}$, bộ mã hóa Transformer (Transformer Encoder) được xây dựng theo các công thức sau:

$$\begin{aligned} \mathbf{p}_i &= \text{PatchEmbedding}(I_i), \\ \mathcal{P} &= \text{Concatenate}\left(\left[\mathbf{p}_i\right]_{i=1}^{N_f}\right), \\ \mathbf{z}^{pe} &= \mathcal{P} + \text{PositionalEncoding}(\mathcal{P}), \\ \mathbf{z}^{e,0} &= [\mathbf{z}_{[\text{CLS}]}^0 \parallel \mathbf{z}^{pe}] \end{aligned} \quad (3.1)$$

$$\mathbf{z}^{e,k} = \text{Transformer}^{(k)}(\mathbf{z}^{e,k-1}), k = 1, 2, \dots, N_e \quad (3.2)$$

Trong đó $\text{PatchEmbedding}(\cdot)$ thực hiện chuyển đổi khung hình RGB thành chuỗi các ma trận ẩn (patch embedding vector) bằng cách chia nhỏ khung hình đầu vào thành các vùng ảnh có kích thước cố định (patch) thông qua một mạng nhúng patch (patch embedding network) và thu được $\mathbf{p} \in \mathbb{R}^{p^2 \times c}$ với kích thước patch là p . Các embedding vector từ N_f khung hình được nối (concatenate) với nhau để tạo thành

$\mathcal{P} \in \mathbb{R}^{N_f \times (p^2 \times c)}$. Thành phần $\text{PositionalEncoding}(\cdot)$ cung cấp thông tin vị trí 2D cho $\mathcal{P}$ và sau đó tạo ra $\mathbf{z}^{pe}$. Mã thông báo [CLS] nhằm mục đích tóm tắt thông tin từ các mã thông báo khác trong quá trình huấn luyện (ví dụ, $\mathbf{z}_{[\text{CLS}]}^{e,0}$) được thêm vào $\mathbf{z}^{pe}$ tạo thành $\mathbf{z}^{e,0} \in \mathbb{R}^{1+N_f \times (p^2 \times c)}$. Sau đó, mô hình áp dụng một loạt khối transformer (transformer blocks) trong kiến trúc Vision Transformer, trong đó N_e là số tầng Transformer. Mỗi khối transformer được ký hiệu là $\text{Transformer}^{(k)}(\cdot)$ gồm nhiều tầng tự chú ý nhiều đầu (multi-head self-attention layers), với số lượng đầu (head) được thiết lập là 8 để tạo ra không gian ẩn mức cao (high-level latent space) với mục đích là giảm số chiều và tạo một vector không gian ẩn giống khung hình gốc nhất có thể. Tiếp theo, tầng chuẩn hóa (normalization layer) và kết nối tắt (skip connection) được áp dụng nhằm giảm mất mát dữ liệu trong khối mã hóa và giải mã. Sau đó, mạng lan

truyền thuận (FeedForward Network - FFN) được sử dụng và tăng chuẩn hóa khác tạo ra đầu ra (output). Mỗi khối biến đổi thứ k tạo ra một chuỗi cấp cao $\mathbf{z}^{e,k} \in \mathbb{R}^{1+N_f \times (p^2 \times c)}$.

Đáng chú ý, mã thông báo [CLS] được tạo bởi khối transformer cuối cùng, ví dụ, $\mathbf{z}_{[\text{CLS}]}^{e,N_e} \in \mathbb{R}^{d_{\text{model}}}$, có nguồn gốc ban đầu được sử dụng cho nhiệm vụ phân lớp ảnh dựa trên kiến trúc Vision Transformer. Cụ thể, số lượng đầu chú ý (attention heads) trong mỗi tầng transformer (transformer layer) có kích thước patch là p và d_{model} lần lượt được khởi tạo là 8, 32 và 768. Để phù hợp với phương pháp đề xuất dựa vào dự đoán khung hình tương lai, bộ giải mã yêu cầu một thiết kế đặc biệt để tái tạo lại mã thông báo này dưới dạng khung hình RGB, được tính toán theo các công thức sau:

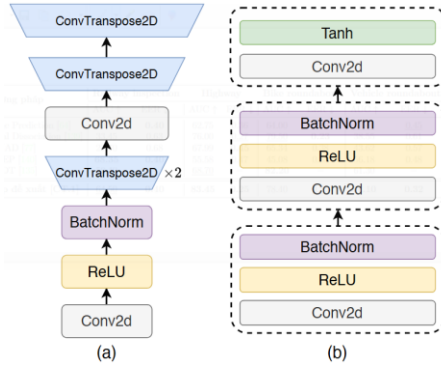
$$\text{Reshape}\left(\text{Projection}\left(\mathbf{z}_{[\text{CLS}]}^{e,N_e}\right)\right) \quad (3.3)$$

$$\mathbf{z}^{d,i} = \text{DecoderBlock}^{(i)}\left(\mathbf{z}^{d,i-1}\right), i = 1, 2, \dots, N_d \quad (3.4)$$

Trong ngữ cảnh này, tầng $\text{Projection}(\cdot)$ đóng vai trò như tầng chiếu (projection layer) chuyển đổi đầu vào từ số chiều của d_{model} sang số chiều cao hơn d_h , được thiết lập là 65, 536. Sau đó, $\text{Reshape}(\cdot)$ được áp dụng để tái cấu trúc vector kết quả thành một tensor 3D. Kết quả là $f^{d,0}$ có dạng $\mathbb{R}^{256 \times 16 \times 16}$. N_d là số khối giải mã được thiết lập là 2. Bộ giải mã Transformer (Transformer Decoder) được trình bày ở Hình 3.2 (a) gồm khối giải mã $\text{DecodingBlock}^{(1)}$ và khối giải mã $\text{DecodingBlock}^{(2)}$ nhận đầu vào là các đặc trưng có độ phân giải thấp từ bộ mã hóa Transformer. Trong đó, khối giải mã đầu tiên $\text{DecodingBlock}^{(1)}$ gồm một chồng gồm hai tầng tích chập (a stack of two convolution layers) để tinh chỉnh các đặc trưng không gian, theo sau là hàm kích hoạt ReLU và tầng chuẩn hóa BatchNorm nhằm ổn định quá trình huấn luyện. Hơn nữa, khối giải mã này còn chứa một tầng tích chập chuyển vị thực hiện tăng độ phân giải ảnh lên 2 lần. Mặt khác, khối giải mã thứ hai $\text{DecodingBlock}^{(2)}$ gồm hai tầng tích chập để tăng cường khối đặc trưng và hai tầng tích chập chuyển vị (transpose convolution layers) tiếp tục tiếp tục tăng cường chất lượng ảnh. Công thức (3.3) và (3.4) từ nay sẽ được ký hiệu ngắn gọn là $\text{Decoder}(\cdot)$. Cuối cùng, khung hình tương lai được tạo sinh theo Công thức (3.5) và so sánh với khung hình thực tế để phát hiện sự kiện bất thường:

$$\hat{I} = \text{Gen}(z^{d,N_d}). \quad (3.5)$$

Trong giai đoạn kiểm tra, chỉ có trình tạo (Generator) trong mô hình đề xuất được sử dụng để dự đoán các khung hình trong tương lai. Trình tạo được ký hiệu là $\text{Gen}(\cdot)$ được minh họa ở Hình 3.2 (b) dùng để tái tạo khung hình dự đoán tương lai và z^{d,N_d} biểu diễn vector ẩn cuối cùng do bộ giải mã tạo ra sau khi xử lý qua N_d khối giải mã. Trình $\text{Gen}(\cdot)$ được xây dựng như một chồng gồm ba tầng tích chập (layer convolution). Hai tầng đầu tiên đi kèm với hàm kích hoạt ReLU và tầng chuẩn hóa BatchNorm. Hàm kích hoạt Tanh() được sử dụng cho tầng tích chập thứ ba. Cuối cùng, tái tạo khung hình dự đoán tương lai $\hat{I} \in \mathbb{R}^{H \times W \times 3}$. Vì vậy, Transformer Decoder trong mô hình gồm các khối giải mã với các tầng tích chập và hàm kích hoạt có nhiệm vụ phóng to ảnh có kích thước như ảnh đầu vào bằng cách sử dụng các lớp tích chập ngược ConvTranspose2D để phóng từ vector đơn [CLS] mã thông báo lên khung hình $\hat{I} \in \mathbb{R}^{H \times W \times 3}$.



Hình 3.2: Minh họa Bộ giải mã Decoder (a) và Gen (b).

3.1.2. Bộ mã hóa Transformer theo không gian - thời gian (Spatio - Temporal Transformer Encoder)

Để nâng cao khả năng trích xuất đặc trưng không gian - thời gian dựa trên kiến trúc Vision Transformer, luận án tiến hành triển khai một chồng (stack) gồm hai bộ mã hóa Transformer (Transformer Encoder). Chi tiết hơn, khối mã hóa đầu tiên, được gọi là Bộ mã hóa biến đổi không gian (Spatial Transformer Encoder) chỉ tập trung trích xuất đặc trưng không gian từ mỗi khung hình, trong khi bộ mã hóa thứ hai được gọi là Bộ mã hóa biến đổi thời gian (Temporal Transformer Encoder) có nhiệm

vụ nội suy đặc trưng thời gian giữa các khung hình. Trong phương pháp đề xuất, mỗi khung hình được xử lý riêng lẻ bằng bộ mã hóa biến đổi không gian có mã thông báo [CLS] riêng, được gọi là mã thông báo không gian [CLS].

3.1.3. Bộ mã hóa Transformer theo không gian (Spatial Transformer Encoder)

Cho một chuỗi N_f khung hình RGB $\mathbf{I} = \{I_1, I_2, \dots, I_{N_f}\}$ với mỗi khung hình ký hiệu là I_i , trong đó $I_i \in \mathbb{R}^{H \times W \times 3}$, bộ mã hóa i^{th} được tính toán lần lượt theo Công thức (3.6), Công thức (3.7) và Công thức (3.8):

$$\begin{aligned} \mathbf{p}_i &= \text{PatchEmbedding}^{(i)}(I_i), \\ \mathbf{z}_i^{s,pe} &= \mathbf{p}_i + \text{PositionalEncoding}^{(i)}(\mathbf{p}_i), \\ \mathbf{z}_i^{s,0} &= [\mathbf{z}_{[\text{CLS}],i}^{s,0} \parallel \mathbf{z}_i^{s,pe}] \end{aligned} \quad (3.6)$$

$$\mathbf{z}_i^{s,k} = \text{Transformer}_{S,i}^{(k)}(\mathbf{z}_i^{s,k-1}), k = 1, 2, \dots, N_S \quad (3.7)$$

$$\mathbf{Z}^s = \text{Concatenate} \left(\left[\mathbf{z}_{[\text{CLS}],i}^{s,N_S} \right]_{i=1}^{N_f} \right) \quad (3.8)$$

Trong đó $\text{Transformer}^{(k)}(\cdot)$ biểu thị khối mã hóa biến đổi không gian k -th cho khung hình thứ i , được thiết kế tương tự như bộ mã hóa biến đổi được trình bày trong Công thức (3.2).

3.1.4. Bộ mã hóa Transformer theo thời gian (Temporal Transformer Encoder)

Tiếp theo để rút trích đặc trưng thời gian, chuỗi mã thông báo không gian [CLS] trải qua quá trình huấn luyện tương tự và được đưa vào Bộ mã hóa biến đổi thời gian để tinh chỉnh và học các phụ thuộc thời gian giữa các đặc trưng không gian của các khung hình. Quá trình này được tính toán theo Công thức (3.9) và Công thức (3.10):

$$\begin{aligned} \mathbf{z}^{t,pe} &= \mathbf{Z}^s + \text{PositionalEncoding}(\mathbf{Z}^s), \\ \mathbf{z}^{t,0} &= [\mathbf{z}_{[\text{CLS}]}^{t,0} \parallel \mathbf{z}^{t,pe}] \end{aligned} \quad (3.9)$$

$$\mathbf{z}^{t,j} = \text{Transformer}_{\mathcal{T}}^{(j)}(\mathbf{z}^{t,j-1}), j = 1, 2, \dots, N_{\mathcal{T}} \quad (3.10)$$

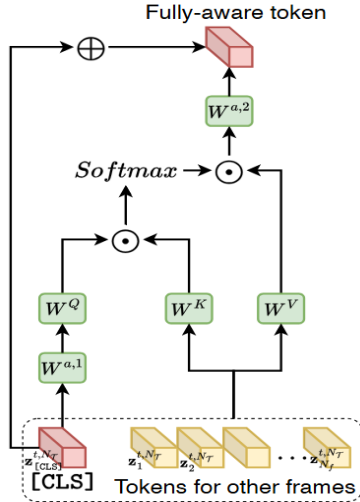
3.1.5. Cơ chế chú ý chéo theo thời gian (Temporal Cross - Attention)

Trong bài toán phát hiện sự kiện bất thường video trên không với

ngữ cảnh giao thông, cơ chế Temporal Cross - Attention cho phép mô hình học đồng thời các mối quan hệ phụ thuộc không gian trong một khung hình và các mối quan hệ phụ thuộc thời gian giữa các khung hình để theo dõi chuyển động của đối tượng theo thời gian bằng cách chú ý đến các đặc trưng từ nhiều khung hình để đảm bảo tính nhất quán theo thời gian giữa các khung hình liên tiếp. Do đó, để khai thác triệt để tất cả các mã thông báo thời gian cùng với mã thông báo [CLS], luận án giới thiệu cơ chế chú ý chéo theo thời gian (Temporal Cross - Attention) được truyền cảm hứng từ công trình đã công bố gần đây [170] và được tính toán theo Công thức (3.11), Công thức (3.12) và được trình bày ở Hình 3.3:

$$\begin{aligned} \mathbf{z}_{[\text{CLS}]}^{t', N_T} &= \mathbf{z}_{[\text{CLS}]}^{t, N_T} \odot \mathbf{W}^{a,1}, \mathbf{W}^{a,1} \in \mathbb{R}^{d \times d} \\ \mathbf{z}^{t', N_T} &= \mathbf{z}^{t, N_T} \setminus \mathbf{z}_{[\text{CLS}]}^{t, N_T} \\ \mathbf{z}^{t,c} &= [\mathbf{z}_{[\text{CLS}]}^{t', N_T}, \|\mathbf{z}^{t', N_T}\|], \mathbf{z}^{t,c} \in \mathbb{R}^{(N_f+1) \times d} \end{aligned} \quad (3.11)$$

$$\begin{aligned} \mathbf{q} &= \mathbf{z}_{[\text{CLS}]}^{t', N_T} \odot \mathbf{W}^q, \mathbf{k} = \mathbf{z}^{t,c} \odot \mathbf{W}^k, \mathbf{v} = \mathbf{z}^{t,c} \odot \mathbf{W}^v, \\ att &= \text{softmax}\left(\frac{\mathbf{qk}^T}{\sqrt{d_k}}\right), att \in \mathbb{R}^{1 \times (N_f+1)}, \\ \mathbf{z}^{t'} &= \mathbf{z}^{t,c} + (\mathbf{v} \odot att) \odot \mathbf{W}^{a,2}, \mathbf{W}^{a,2} \in \mathbb{R}^{d \times d} \\ \mathbf{z}^{t'} &= \mathbf{z}^{t'} + \mathbf{z}_{[\text{CLS}]}^{t, N_T}. \end{aligned} \quad (3.12)$$



Hình 3.3: Minh họa cơ chế chú ý chéo theo thời gian.

3.1.6. Phát hiện khung hình có chứa sự kiện bất thường dựa trên dự đoán khung hình tương lai

Phát hiện khung hình có chứa sự kiện bất thường dựa trên dự đoán khung hình tương lai: Một chồng các tầng giải mã (stack of decoder layers) được tính toán theo Công thức (3.3), Công thức (3.4) và một mạng tạo sinh được mô tả bởi Công thức (3.5) được áp dụng cho $\mathbf{z}'$ để thu được khung hình dự đoán tương lai $\hat{I} \in \mathbb{R}^{H \times W \times 3}$.

Hàm mất mát (Loss Function): Để huấn luyện bộ biến đổi không gian - thời gian nhằm tái tạo khung hình dự đoán $\hat{I}$ càng gần với khung hình thực tế càng tốt, hàm mất mát $L2$ được áp dụng và tính toán như sau:

$$\begin{aligned} \mathbf{I} &= \{I_i\}_{i=1}^{N_f}, I_i \in \mathbb{R}^{H \times W \times 3}, \\ \hat{I}_{N_f+1} &= \text{Gen}\left(\text{Decoder}(\text{Transformer}_{S,T}(\mathbf{I}))\right), \\ \mathcal{L}_{L2} &= \|\hat{I}_{N_f+1} - I_{N_f+1}\|_2 \end{aligned} \quad (3.13)$$

Trong đó $\text{Transformer}_{S,T}(\cdot)$ trình bày ngắn gọn về mô hình đề xuất của luận án được mô tả từ Công thức (3.6) đến Công thức (3.12), và N_f là số khung hình đầu vào. $\hat{I}_{N_f+1}$ biểu diễn khung hình dự đoán tương lai, trong khi I_{N_f+1} biểu diễn khung hình thực tế có gán nhãn bất thường (ground-truth), ký hiệu là $(N_f + 1)^{th}$. Rõ ràng, với một tập các khung hình chứa sự kiện bất thường trong tập kiểm tra được gán nhãn giúp đánh giá tính hiệu quả và độ tin cậy của mô hình tốt hơn. Hàm mất mát này tính toán chênh lệch giá trị trung bình bình phương sai số (Mean Squared Error - MSE) [108, 135, 171] được áp dụng để xác định giá trị khác biệt cũng như chất lượng ảnh giữa khung hình tương lai được dự đoán với nhãn thực và các khung hình liên tiếp trước đó. Cụ thể, MSE được tối ưu hóa và tính toán theo công thức (3.14) như sau:

$$\text{MSE}(\hat{I}_{N_f+1}, I_{N_f+1}) = \frac{1}{H \times W \times C} \sum_{i=1}^H \sum_{j=1}^W \sum_{c=1}^C \left(\hat{I}_{N_f+1}(i, j, c) - I_{N_f+1}(i, j, c) \right)^2 \quad (3.14)$$

Trong đó H và W là chiều rộng và chiều cao của khung hình, $C=3$ là số kênh màu RGB của khung hình video. Như vậy, khi có sự kiện bất thường xảy ra, sự khác biệt giá trị pixel tại các vùng bất thường trong khung hình dự đoán so với khung hình thực dẫn đến giá trị sai số MSE giữa hai khung hình này sẽ tăng lên. Từ quan sát này, chọn một ngưỡng và xác định các khung hình trong video có hoặc không có chứa sự kiện bất thường. Một cách trực quan là tối thiểu hóa giá trị

MSE của mô hình trong quá trình huấn luyện và trong giai đoạn kiểm tra giá trị MSE càng cao cho thấy mô hình không dự đoán được khung hình tương lai một cách chính xác, có thể là do khung hình có chứa sự kiện bất thường.

3.2 Ưu điểm và hạn chế của phương pháp đề xuất

- **Ưu điểm:**

- Phương pháp đề xuất của luận án huấn luyện mô hình theo cách tiếp cận không giám sát nên giảm sự phụ thuộc vào dữ liệu bất thường có gán nhãn và có thể áp dụng cho các nguồn cấp dữ liệu có độ phân giải cao và môi trường khác nhau vì dữ liệu bất thường chỉ chiếm số lượng rất ít so với dữ liệu bình thường trong các tình huống thực tế.

- Huấn luyện và trích xuất đặc trưng chuyên biệt: phương pháp đề xuất chia nhỏ bài toán thành các tác vụ riêng lẻ (rút trích đặc trưng không gian, rút trích đặc trưng thời gian, và tích hợp đặc trưng không gian-thời gian) nên dễ quản lý hơn, giúp đào tạo ổn định và hội tụ tốt hơn so với rút trích đặc trưng không gian - thời gian đồng thời.

- Tính linh hoạt: có khả năng thêm các đặc trưng khác trong tương lai bao gồm âm thanh (audio), văn bản (text).

- Khả năng mở rộng của mô hình: có khả năng thay thế, hiệu chỉnh độc lập các thành phần rút trích đặc trưng hiện tại bằng các đặc trưng tốt hơn trong tương lai mà không làm ảnh hưởng đến toàn bộ kiến trúc của mô hình đề xuất.

- Việc áp dụng cơ chế chú ý chéo (cross-attention mechanism) giúp tăng cường hợp nhất các đặc trưng không gian - thời gian đã rút trích và cho phép tương tác tốt hơn giữa các đặc trưng không gian - thời gian có liên quan giữa các khung hình khác nhau nhằm phát hiện các sự kiện bất thường liên quan đến những thay đổi theo thời gian giữa các khung hình trong video và giúp mô hình hiểu, phân tích thông tin cục bộ và toàn cục về ngữ cảnh giao thông tốt hơn.

- Hiệu suất phát hiện sự kiện bất thường tăng lên đáng kể.

- **Hạn chế:**

- Độ phức tạp tăng lên: việc rút trích riêng biệt các đặc trưng không gian - thời gian và tích hợp các đặc trưng này làm tăng thêm độ phức tạp cho kiến trúc của mô hình.

- Thời gian huấn luyện mô hình và tài nguyên tính toán: thời gian huấn luyện có thể lâu hơn và cần nhiều tài nguyên hơn do phải tối ưu hóa các thành phần và tương tác giữa chúng.

Chương 4. XÂY DỰNG BỘ DỮ LIỆU VIDEO TRÊN KHÔNG VỚI NGŨ CẢNH GIAO THÔNG

4.1 Mô tả bộ dữ liệu UIT-ADrone

Để giải quyết vấn đề khan hiếm của các bộ dữ liệu được quay bằng thiết bị bay không người lái cho nhiệm vụ phát hiện sự kiện bất thường trong thế giới thực trong lĩnh vực giao thông, luận án tiến hành thu thập và xây dựng một bộ dữ liệu phát hiện sự kiện bất thường trong video trên không, có tên là UIT-ADrone. Các cảnh (scene) được chọn để quay là các vòng xuyên trong môi trường giao thông thực tế, đây là một trong những nơi phổ biến để quay những người và phương tiện tham gia giao thông với nhiều chuyển động và sự kiện đồng thời trong tình huống thực tế. Hơn nữa, mật độ và sự tương tác giữa những người tham gia giao thông rất cao tại các giao lộ này, đặc biệt là xe gắn máy. Ngoài ra, các sự kiện bất thường chủ yếu xảy ra trong một khoảng thời gian ngắn trong video nên tốn nhiều công sức, chi phí và thời gian để thu thập với các sự kiện bất thường được liệt kê ở Bảng 4.1.

4.2 Thu thập dữ liệu

Luận án sử dụng thiết bị bay không người lái để tiến hành thu thập dữ liệu với ngữ cảnh giao thông trong thực tế cho bài toán phát hiện sự kiện bất thường tại TP.HCM, Việt Nam. Thiết bị bay được sử dụng là DJI MAVIC MINI 2, quay ở tốc độ 30 khung hình/giây (fps), với độ phân giải 1920×1080 pixel và quay ở độ cao từ 50m đến 70m vào các thời điểm khác nhau trong ngày với điều kiện thời tiết và môi trường thuận lợi. Hơn nữa, Hình 4.2 minh họa các cảnh (scene) ở vòng xuyên tại ba vị trí trong bộ dữ liệu do luận án tiến hành thu thập và xây dựng.

Bảng 4.1: Thống kê bộ dữ liệu UIT-ADrone [CT.3].

STT	Loại sự kiện bất thường	Số lượng sự kiện bất thường
1	Người đi bộ sang đường không đúng vạch	80
2	Đi bộ xuống lòng lề đường	344
3	Chạy xe không đúng vòng xuyên	686
4	Lái xe trên vỉa hè	145
5	Rẽ trái/phải không đúng luật	28
6	Đỗ xe dưới lòng lề đường	225
7	Chở hàng cồng kềnh	144
8	Đỗ xe trên vỉa hè	68
9	Lái xe ngược chiều	214
10	Ngã xe máy	1



Hình 4.2: Trục quan hóa các cảnh giao thông trong bộ dữ liệu UIT-ADrone [CT.3].

4.3 Thống kê bộ dữ liệu

Bộ dữ liệu UIT-ADrone gồm 51 video với tổng cộng 206.191 khung hình được trích xuất gồm các sự kiện bất thường khác nhau. Toàn bộ video dài khoảng 6,50 giờ được quay trong các tình huống giao thông phức tạp trong thế giới thực với 10 loại sự kiện bất thường được mô tả trong Bảng 4.1. Hơn nữa, bộ dữ liệu UIT-ADrone gồm 8 lớp đối tượng: người đi bộ (pedestrian), xe đạp (bike), xe gắn máy (motorbike), xe ba gác (tricycle), xe ô tô (car), xe tải (truck), xe tải nhỏ (van) và xe buýt (bus). Bên cạnh đó, Bảng 4.2 so sánh bộ dữ liệu UIT-ADrone với các bộ dữ liệu chuẩn cho bài toán phát hiện sự kiện bất thường trong video trên không.

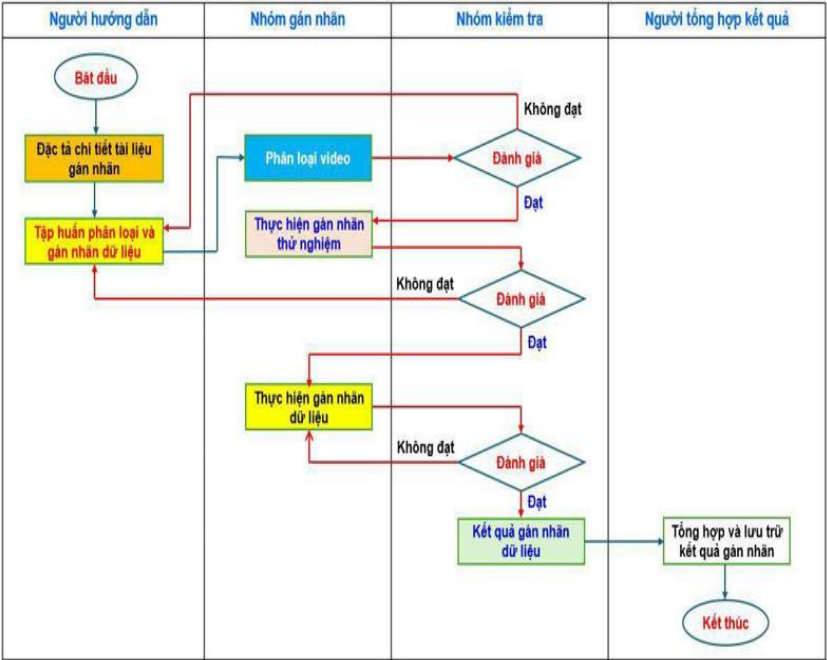
Bảng 4.2: So sánh các bộ dữ liệu bất thường phổ biến trong video trên không và bộ dữ liệu UIT-ADrone [CT.3].

STT	Bộ dữ liệu	Năm	Ngữ cảnh	Độ cao	Độ phân giải	#Khung hình	#Lớp đối tượng	#Loại sự kiện bất thường
1	Mini Drone [130]	2015	Bãi đỗ xe	10m - 20m	224 x 224	23.295	4	3
2	UTT Drone [131]	2021	Khảo viên trường đại học	10m - 20m	227 x 227	14.021	3	3
3	Brutal Running [132]	2021	Bãi đỗ xe	-	1920 x 1080	1.000	4	1
4	AU-AIR-Anomaly [133]	2021	Giao thông	10m - 30m	1920 x 1080	32.823	8	4
5	Drone-Anomaly [135]	2022	Giao thông	50m - 70m	640 x 640	87.488	7	-
6	UIT-ADrone [CT.3]	2023	Giao thông	50m - 70m	1920 x 1080	206.191	8	10

4.4 Gán nhãn cho bộ dữ liệu

Luận án thực hiện gán nhãn theo quy trình được minh họa trong Hình 4.3. Để tạo dữ liệu xác thực cho mục đích đánh giá hiệu suất của các phương pháp tối tân hiện tại khác nhau cho nhiệm vụ phát hiện sự kiện bất thường trong video với ngữ cảnh giao thông, luận án sử dụng công cụ có tên là Supervisely (<https://supervise.ly/>) để thực hiện phân loại các lớp đối tượng và gán nhãn. Luận án sử dụng tệp chú thích theo định dạng MS COCO (Microsoft Common Objects in Context) [164], một định dạng dữ liệu phổ biến trong lĩnh vực phát hiện và phân đoạn đối

tượng với bộ dữ liệu gồm 80 lớp đối tượng với các hộp biên là các hộp biên ngang (horizontal bounding boxes), nghĩa là các hộp biên này không có hướng. Tiếp theo, luận án sử dụng các nhãn kiểm tra được lưu trữ dưới dạng mảng (.npy). Mỗi video được biểu diễn bằng một mảng (array). Chỉ mục cho các khung hình bắt đầu từ 0 và kết thúc với tổng số khung hình trừ 1. Mỗi phần tử trong mảng đại diện cho một khung hình được thiết lập 0, nghĩa là khung hình đó không có chứa sự kiện bất thường hoặc được thiết lập là 1, nghĩa là khung hình đó có chứa sự kiện bất thường, được minh họa ở Bảng 4.3.



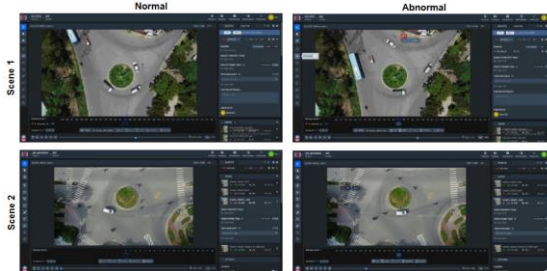
Hình 4.3: Quy trình gán nhãn dữ liệu.

Bảng 4.3: Minh họa cho chú thích nhãn cho khung hình có chứa sự kiện bất thường trong một video của bộ dữ liệu UIT-ADrone [CT.3].

Ví dụ minh họa cho chú thích nhãn bất thường trong video có tên DJI_0150.mp4														
Tên tệp: DJI_0150.npy														
0	1	2	...	27	28	29	...	152	153	154	...	2157		
0	0	1	...	1	1	1	...	1	1	1	...	0		

Giải thích: Trong mảng này, độ dài của mảng index được đánh dấu từ 0 đến số lượng khung hình (frame) trong video (2158) - 1. Mỗi index tương ứng với một khung hình trong video. Hàng dưới của mảng biểu thị hai giá trị: 0 và 1. Khi giá trị là 1, nghĩa là khung hình tương ứng có chứa sự kiện bất thường, và ngược lại, khi giá trị là 0, khung hình đó không có chứa sự kiện bất thường.

Việc gán nhãn sự kiện bất thường mức khung hình (1: khung hình có chứa sự kiện bất thường; 0: khung hình không có chứa sự kiện bất thường) trong bộ dữ liệu UIT-ADrone cho bài toán phát hiện sự kiện bất thường với ngữ cảnh giao thông có ý nghĩa lý luận và thực tiễn rất quan trọng để huấn luyện mô hình và tiềm năng ứng dụng cao trong thực tế. Việc gán nhãn theo cấu trúc này giúp nâng cao hiệu suất của các mô hình huấn luyện theo chiến lược học không giám sát, cho phép phát hiện bất thường chính xác, hiệu quả và đáng tin cậy trong video trên không với ngữ cảnh giao thông. Ngoài ra, một số ảnh chụp giao diện của công cụ Supervisely được minh họa tương ứng trong Hình 4.4 và Hình 4.5.



Hình 4.4: Một số ảnh chụp nhanh từ các cảnh giao thông của bộ dữ liệu UIT-ADrone [CT.3] từ giao diện của công cụ Supervisely. Cột bên trái hiển thị các khung hình bình thường và cột bên phải hiển thị các khung hình bất thường. Lưu ý, các hộp biên màu đỏ biểu thị các khung hình có chứa sự kiện bất thường [CT.3].



Hình 4.5: Minh họa một số sự kiện bất thường trong bộ dữ liệu UIT-ADrone [CT.3].

Hình 4.5 trình bày một số sự kiện bất thường trong các cảnh quay thuộc bộ dữ liệu UIT-ADrone. Ngoài ra, đối với 10 loại sự kiện bất thường được thống kê trong Bảng 4.1 thuộc bộ dữ liệu UIT-ADrone hiện tại chưa được gán nhãn cho từng loại sự kiện bất thường. Vì vậy, nghiên cứu tiếp theo sẽ tiến hành gán nhãn cho 10 loại sự kiện bất thường này và có thể mở rộng cho bài toán phân loại đa nhãn sự kiện bất thường (multi-label classification of abnormal events).

Chương 5. THỰC NGHIỆM VÀ KẾT QUẢ

5.1 Các bộ dữ liệu chuẩn dùng để thực nghiệm

Luận án sử dụng hai bộ dữ liệu chuẩn công khai được quay từ thiết bị máy bay không người lái cho bài toán phát hiện sự kiện bất thường trong video, có tên là UIT-ADrone [136] và Drone-Anomaly [135] được trình bày trong Bảng 5.1. Ngoài ra, Hình 5.1 minh họa một số cảnh của hai bộ dữ liệu này.

Bảng 5.1: Hai bộ dữ liệu chuẩn Drone-Anomaly [135] và UIT-ADrone [CT.3] dùng để thực nghiệm cho phương pháp đề xuất và các phương pháp tối tân hiện tại.

STT	Bộ dữ liệu	Năm	Độ phân giải	#Khung hình bất thường	#Tổng số khung hình	Tập huấn luyện	Tập kiểm tra
1	Drone-Anomaly [135]	2022	640 × 640	16.041	87,488	51,635	35,853
2	UIT-ADrone [CT.3]	2023	1920 × 1080	63.485	206,191	59,186	147,008

Hình 5.1: Minh họa các khung hình có các cảnh bình thường và bất thường trong bộ dữ liệu UIT-ADrone [CT.3] và Drone-Anomaly [135]. Hộp màu vàng minh họa một số trường hợp liên quan đến các sự kiện bất thường trong các tình huống giao thông từ hai bộ dữ liệu chuẩn này.

5.2 Độ đo đánh giá

Để tính toán hiệu suất của mô hình phát hiện sự kiện bất thường trong video giám sát, luận án sử dụng độ đo AUC-ROC (Area Under The Curve - Receiver Operating Characteristics) [100] và độ đo EER (Equal Error Rate) [13].

5.3 So sánh với các phương pháp tối tân hiện tại

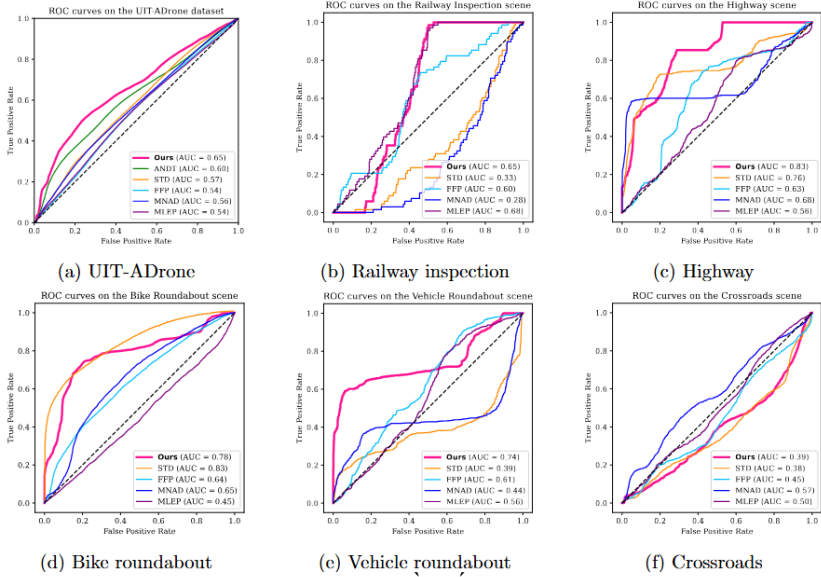
Luận án tiến hành phân tích và so sánh hiệu suất của các phương pháp tối tân hiện tại và phương pháp đề xuất của luận án trên hai bộ dữ liệu bất thường UIT-ADrone [136] và Drone-Anomaly [135], kết quả thực nghiệm lần lượt được trình bày chi tiết trong Bảng 5.2, Bảng 5.3, Bảng 5.4 và Bảng 5.5. Bên cạnh đó, luận án cũng trình bày kết quả trực quan hóa của các phương pháp khác nhau trên các bộ dữ liệu chuẩn trong Hình 5.2. Ngoài ra, ảnh chụp nhanh (snapshots) một số trường hợp phát hiện sự kiện bất thường chính xác và một số trường hợp phát hiện sự kiện bất thường không chính xác trên bộ dữ liệu này từ các thực nghiệm của luận án, được trình bày trong Hình 5.3 và Hình 5.4.

Bảng 5.2: So sánh các phương pháp tối tân hiện tại với độ đo AUC và EER ở mức khung hình (%) trên mỗi cảnh (scene) của bộ dữ liệu Drone-Anomaly [135]. Các kết quả tốt nhất được in đậm. Kết quả cao thứ hai được gạch chân [CT.1].

Phương pháp	Railway inspection		Highway		Bike roundabout		Vehicle roundabout		Crossroads	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Future Frame Prediction [64]	60.44	0.40	62.75	0.36	64.00	0.40	<u>61.41</u>	<u>0.45</u>	44.84	0.54
Spatio-Temporal Dissociation [139]	33.45	0.63	76.00	<u>0.27</u>	<u>79.50</u>	0.23	38.55	0.61	38.09	0.59
MNAD [77]	27.70	0.68	67.99	0.45	65.34	0.38	43.62	0.57	57.07	0.45
MLEP [140]	68.35	0.40	55.58	0.47	45.08	0.53	56.18	0.48	49.80	<u>0.51</u>
ANDT [135]	59.40	–	<u>68.70</u>	–	82.20	–	61.30	–	30.15 ¹	–
Phương pháp đề xuất [CT.1]	<u>64.50</u>	0.40	83.45	0.25	78.40	<u>0.25</u>	74.10	0.32	38.56	0.56

¹Kết quả dựa trên mã nguồn được cung cấp với các tham số được đề xuất từ công trình nghiên cứu [135].

Trong phần này, luận án cũng trình bày chi tiết các thực nghiệm mở rộng chuyên sâu có tinh chỉnh (fine-tuning) cho nhiệm vụ phát hiện sự kiện bất thường bằng cách hiệu chỉnh bộ dữ liệu miền chéo (cross - dataset adaptation) thông qua việc phân tích và so sánh hiệu suất dựa trên độ đo AUC với phương pháp đề xuất trên bộ dữ liệu nguồn (source dataset) và bộ dữ liệu đích (target dataset), cụ thể là UIT-ADrone [136] và các cảnh thuộc bộ dữ liệu Drone-Anomaly [135]. Trước tiên, luận án huấn luyện mô hình cho các mẫu dữ liệu huấn luyện từ bộ dữ liệu nguồn UIT-ADrone, sau đó thực hiện cài đặt học chuyển tiếp bằng cách tải các trọng số được huấn luyện trước từ tập huấn luyện trong tập dữ liệu nguồn để tiếp tục học thêm về mẫu huấn luyện từ các cảnh của bộ dữ liệu đích Drone-Anomaly. Tiếp theo, các trọng số cuối cùng sẽ được tải và tiến hành đánh giá trên tập kiểm tra của bộ dữ liệu đích.



Hình 5.2: Trực quan hóa phương pháp đề xuất và các phương pháp Future Frame Prediction (FFP) [64], Spatio-Temporal Dissociation (STD) [139], MNAD [77], MLEP [140] và ANDT [135] với độ đo ROC-AUC ở mức khung hình trên bộ dữ liệu UIT-Drone [CT.3] và Drone-Anomaly [135] [CT.1].



Hình 5.3: Trực quan hóa các trường hợp true positive, true negative, false positive and false negative được tạo ra bởi mô hình đề xuất của luận án. Các sự kiện bất thường trong mỗi khung hình video được đánh dấu bằng hình chữ nhật màu vàng [CT.1].

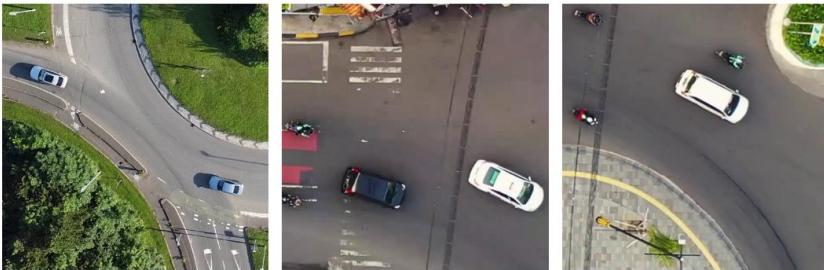
Bảng 5.3: So sánh hiệu suất của các phương pháp phát hiện sự kiện bất thường với độ đo AUC và EER trung bình ở mức khung hình (%) trên bộ dữ liệu UIT-ADrone [136] và Drone-Anomaly [135]. Các kết quả tốt nhất được đánh dấu **in đậm** [CT.1].

Phương pháp	UIT-ADrone [CT.3]		Drone-Anomaly ² [135]	
	AUC ↑	EER ↓	mAUC ↑	mEER ↓
Future Frame Prediction [64]	53.56	0.47	57.94	0.43
Spatio-Temporal Dissociation [139]	57.05	0.45	52.46	0.47
MNAD [77]	55.88	0.46	52.34	0.51
MLEP [140]	53.55	0.47	55.00	0.48
ANDT [135]	60.50	0.42	60.35	–
Phương pháp đề xuất [CT.1]	65.45	0.39	67.80	0.36

²Kết quả này được tính bằng cách tính trung bình từ kết quả của 5 cảnh trong bộ dữ liệu Drone-Anomaly: Railway inspection, Highway, Bike roundabout, Vehicle roundabout, Crossroads.

Bảng 5.4: Kết quả của các thực nghiệm thích ứng miền với độ đo AUC trung bình ở mức khung hình (%) trên bộ dữ liệu UIT-ADrone [136] và trên 5 cảnh của bộ dữ liệu Drone- Anomaly [135] có tinh chỉnh [CT.1].

Nguồn (Source) → Đích (Target)	AUC _{s→t}	AUC _{t→t}	Δ
UIT-ADrone [CT.3] → Railway inspection [135]	66.94	64.50	+2.44
UIT-ADrone [CT.3] → Highway [135]	89.14	83.45	+5.69
UIT-ADrone [CT.3] → Bike roundabout [135]	79.62	78.40	+1.22
UIT-ADrone [CT.3] → Vehicle roundabout [135]	51.81	74.10	–22.29
UIT-ADrone [CT.3] → Crossroads [135]	28.60	38.56	–9.96



Hình 5.4: Minh họa 3 cảnh phụ (sub-scenes) trong cảnh Vehicle roundabout. Do sự khác biệt đáng kể về ngữ cảnh tại vòng xoay, hiệu suất của phương pháp đề xuất giảm khi thực nghiệm tinh chỉnh trên bộ dữ liệu UIT-ADrone [136] cho cảnh Vehicle roundabout thuộc bộ dữ liệu Drone-Anomaly [135].

5.4 Nghiên cứu cắt lát (Ablation Studies)

Luận án tiến hành các nghiên cứu cắt lát để phân tích tầm quan trọng các thành phần khác nhau trong mô hình đề xuất được thể hiện đầy đủ trong Bảng 5.5. Các kịch bản nghiên cứu gồm: (1) STE + TTE được xem là mô hình cơ sở (baseline model) của phương pháp đề xuất, sử dụng bộ mã hóa Transformer không gian (Spatial Transformer Encoder) và bộ mã hóa Transformer thời gian (Temporal Transformer Encoder) nhưng không áp dụng cơ chế tự chú ý chéo; và (2) STE + TTE + CSA (phương pháp đề xuất của luận án).

Bảng 5.5: Một số kết quả nghiên cứu cắt lát của phương pháp được đề xuất với độ đo AUC và EER ở mức khung hình (%) trên bộ dữ liệu UIT-ADrone [CT.3]. Các kết quả tốt nhất được in đậm [CT.1].

Phương pháp	UIT-ADrone [CT.3]	
	AUC ↑	EER ↓
ANDT [135]	60.50	0.42
STE + TTE (Our baseline)	61.34	0.39
STE + TTE + CSA ([CT.1])	65.45	0.39

Chương 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết quả đạt được

(1) Đề xuất một phương pháp phát hiện sự kiện bất thường ở mức khung hình trong video trên không với ngữ cảnh giao thông dựa trên kiến trúc mạng Vision Transformer với chiến lược học không giám sát bằng cách khai thác hiệu quả các đặc trưng không gian - thời gian giữa các khung hình trong video và phát hiện khung hình có chứa sự kiện bất thường dựa trên dự đoán khung hình tương lai [CT.1].

(2) Phân loại các bài toán liên quan đến bất thường trong phân tích video, hệ thống hóa và phân tích các ưu điểm và hạn chế của các phương pháp phát hiện sự kiện bất thường. Bên cạnh đó, khảo sát các bộ dữ liệu chuẩn, thực nghiệm và thảo luận chuyên sâu các phương pháp giải quyết bài toán phát hiện sự kiện bất thường trong video [CT.2].

(3) Công bố một bộ dữ liệu có quy mô lớn, có tên là UIT-ADrone cho bài toán phát hiện sự kiện bất thường trong video trên không với ngữ cảnh giao thông tại TP.HCM, Việt Nam [CT.3].

6.2 Các đóng góp chính của luận án

(1) Luận án đã nghiên cứu và đề xuất phương pháp phát hiện sự

kiến bất thường trong video trên không với ngữ cảnh giao thông dựa trên kiến trúc mạng Vision Transformer với chiến lược học không giám sát bằng cách khai thác các đặc trưng không gian - thời gian giữa các khung hình trong video và phát hiện khung hình có chứa sự kiện bất thường dựa trên khung hình tương lai [CT.1].

(2) Luận án đóng góp một công trình nghiên cứu tổng quan cho bài toán phát hiện sự kiện bất thường trong video [CT.2].

(3) Luận án đóng góp một dữ liệu trong video trên không có quy mô lớn nhất tính đến thời điểm hiện tại, có tên là UIT-ADrone cho bài toán phát hiện sự kiện bất thường trong video được quay bằng thiết bị bay không người lái [CT.3]. Bộ dữ liệu được công bố công khai dành cho mục đích nghiên cứu tại địa chỉ <https://uit-together.github.io/datasets/UIT-ADrone/>.

6.3 Các công trình nghiên cứu đã công bố

6.3.1 Các công bố chính của luận án

[CT.1] **Tung Minh Tran**, Doanh C. Bui, Tam V. Nguyen, Khang Nguyen, “Transformer- based Spatio-Temporal Unsupervised Traffic Anomaly Detection in Aerial Videos”, IEEE Transactions on Circuits and Systems for Video Technology, IEEE, vol. 34, pp. 8292-8309, 2024. DOI: 10.1109/TCSVT.2024.3376399. ISI, Q1, Impact Factor: 8.3.

[CT.2] **Tung Minh Tran**, Tu N. Vu, Nguyen D. Vo, Tam V. Nguyen, Nguyen Khang, “Anomaly Analysis in Images and Videos: A Comprehensive Review”, ACM Computing Surveys, vol. 55, 2022. DOI: 10.1145/3544014. ISI, Q1 (Computer Science Theory and Models), Impact Factor: 16.6.

[CT.3] **Tung Minh Tran**, Tu N. Vu, Tam V. Nguyen, Khang Nguyen, “UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE, vol. 16, pp. 5590-5601, 2023. DOI: 10.1109/JSTARS.2023.3285905. ISI, Q1 (Computers in Earth Sciences), Impact Factor: 4.7.

6.3.2 Các công bố khác có liên quan đến luận án

Hội thảo quốc tế:

[CT.4] Dung T.T. Vo, **Tung Minh Tran**, Nguyen D. Vo, Khang Nguyen, “UIT-Anomaly: A Modern Vietnamese Video Dataset for Anomaly Detection”, 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), DOI: 10.1109/nics54270.2021.9701556

[CT.5] Tu N. Vu, Toan T. Dinh, Nguyen D. Vo, **Tung Minh Tran**,

Khang Nguyen, “VNAAnomaly: A Novel Vietnam Surveillance Video Dataset for Anomaly Detection”, 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), DOI: 10.1109/nics54270.2021.9701540.

Hội thảo trong nước:

[CT.6] **Trần Minh Tùng**, Nguyễn Cường Phát, Nguyễn Tấn Trần Minh Khang, “Phân tích ảnh hưởng của thời tiết bất lợi đối với phát hiện bất thường trong giao thông từ video trên không”, Hội nghị Khoa học Quốc gia lần thứ XVII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR’2024).

[CT.7] Vũ Ngọc Tú, Đinh Thanh Toàn, **Trần Minh Tùng**, Võ Duy Nguyên, Nguyễn Tấn Trần Minh Khang, “Một hàm lỗi cho phát hiện bất thường trong video giám sát”, Hội nghị Quốc gia lần thứ XIV về Điện tử, Truyền thông và Công nghệ thông tin (REV-ECIT’2021).

6.3.3 Các đề tài nghiên cứu khoa học đã tham gia

[ĐT.1] Đề tài: “Phát hiện đối tượng trong không ảnh sử dụng mạng học sâu”. Mã số DS2021-26-01. Cơ quan quản lý: Đại học Quốc gia TP.HCM.

[ĐT.2] Đề tài: “Phát hiện phương tiện và nhận biết bất thường giao thông trong không ảnh”. Mã số B2023-26-01. Cơ quan quản lý: Đại học Quốc gia TP.HCM.

6.4 Hướng phát triển

Nghiên cứu tiếp theo của luận án sẽ xem xét việc mở rộng số lượng ngữ cảnh giao thông trong thực tế như ngã ba, ngã tư trong giao thông đô thị, đường cao tốc, đường giao thông tại các khu vực nông thôn để tăng tính đa dạng của các sự kiện bất thường và mở rộng gắn nhãn cho từng loại sự kiện bất thường trong video trên không vì một khung hình có thể có nhiều hơn một sự kiện bất thường và là cơ sở mở rộng nghiên cứu cho bài toán phân loại đa nhãn các sự kiện bất thường. Bên cạnh đó, luận án có thể mở rộng nghiên cứu cho bài toán định vị sự kiện bất thường đối trong video trên không dựa trên các khung hình chứa sự kiện bất thường đã được phát hiện trong phương pháp đề xuất.