FACULTY OF ENGINEERING AND ARCHITECTURE

Transferable and Privacy-Friendly Deep Learning Techniques for Audio-Visual Urban Surveillance: From Lab to Street

Wei-Cheng Wang

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Computer Science Engineering

Supervisors

Prof. Pieter Simoens, PhD - Prof. Sam Leroux, PhD

Department of Information Technology
Faculty of Engineering and Architecture, Ghent University

GHENT UNIVERSITY

ISBN null
NUR 980, 984
Wettelijk depot: null

Members of the Examination Board

Chair

Prof. Joris Degroote, PhD, Ghent University

Other members entitled to vote

Prof. Eli De Poorter, PhD, Ghent University
Prof. Paul Devos, PhD, Ghent University
Prof. Aaron Ding, PhD, Technische Universiteit Delft, the Netherlands
Prof. Toon Goedemé, PhD, KU Leuven

Supervisors

Prof. Pieter Simoens, PhD, Ghent University
Prof. Sam Leroux, PhD, Ghent University

Table of Contents

List of Figures

List of Tables

List of Acronyms

ADR	Adversarial Disentanglement Representation
AE	AutoEncoder
AVA	Audio-Visual Active Speaker dataset
AVC	Audio-Visual Correspondence
AVS	Audio-Visual Synchronization
CCA	Canonical Correlation Analysis
CKA	Centered Kernel Alignment
CNN	Convolutional Neural Network
DCASE	Detection and Classification of Acoustic Scenes and Events
DGA	Data Governance Act
DNN	Deep Neural Network
EnsV	Ensemble Validation
EPG	Embedding-based Pair Generation
FCN	Fully Convolutional Network
FL	Federated Learning
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HITL	Human-in-the-Loop
HOTL	Human-on-the-Loop
HSIC	Hilbert-Schmidt Independence Criterion
ICT	Information and Communication Technologies
IIoT	Industrial Internet of Things
IoT	Internet of Things
IteRand	Iterative Randomization
LEAD	Logit Space Evolution for Model Selection
LogME	Log Marginal Evidence
LTH	Lottery Ticket Hypothesis
mAP	mean Average Precision
MSE	Mean-Squared Error
MSSIM	Mean Structural Similarity Index Measure

MTL	Multi-Task Learning
OCR	Optical Character Recognition
OCSVM	One-Class Support Vector Machine
PESQ	Perceptual Evaluation of Speech Quality
PSNR	Peak Signal-to-Noise Ratio
RBAC	Role-Based Access Control
RINN	Randomly Initialized Neural Network
SFUDA	Source-Free Unsupervised Domain Adaptation
SOTA	state-of-the-art
SSIM	Structural Similarity Index Measure
SSL	Self-Supervised Learning
STOI	Short-Time Objective Intelligibility
ToCaDa	Toulouse Campus Surveillance Dataset
TPG	Temporal-based Pair Generation
UD	Uncertainty Distance
UDT	Urban Digital Twin
XAI	Explainable AI

Samenvatting

Slimme bewaking is een cruciale toepassing binnen de bredere ontwikkeling van slimme steden. Naast het verbeteren van de openbare veiligheid kan het ook talrijke andere facetten van het stedelijk leven verbeteren. Zo kan het bijvoorbeeld helpen om verkeersstromen in kaart te brengen. Op deze manier dragen deze technologiëen significant bij aan een hogere levenskwaliteit en een verhoogd gevoel van veiligheid voor de bewoners. De ontwikkeling van robuuste en betrouwbare oplossingen wordt echter aanzienlijk belemmerd door een reeks praktische beperkingen. Deze variëren van de aanzienlijke kosten die gepaard gaan met data-annotatie en -opschoning, en de strikte limieten op beschikbare rekenkracht, tot complexe ethische kwesties rondom dataverzameling en het delen van gevoelige informatie. Hoewel uitgebreid wetenschappelijk onderzoek naar machine learning-modellen de theoretische prestaties van slimme bewakingstoepassingen aanzienlijk heeft verbeterd, blijft er een aanzienlijke en hardnekkige kloof bestaan tussen de prestaties van modellen die onder gecontroleerde omstandigheden zijn ontwikkeld en hun operationele betrouwbaarheid in de onvoorspelbare echte wereld.

Om deze fundamentele kloof te overbruggen, identificeert dit proefschrift in de eerste plaats de specifieke beperkingen die zich voordoen bij de toepassing van conventionele deep learning-technieken op reële, complexe bewakingsscenario's. Vervolgens ontwikkelt en valideert het een reeks innovatieve, overdraagbare, privacybewuste en labelvrije leeroplossingen. Met een gerichte focus op zowel auditieve als visuele modaliteiten, en hun synergie, pakt dit werk de uitdagingen op verschillende, met elkaar verweven vlakken aan: de onvermijdelijke prestatievermindering als gevolg van data- en taakverschuivingen (drift), de aanzienlijke en privacyrisico's, en de praktische, operationele belemmeringen die worden veroorzaakt door de beperkte toegang tot zowel brongegevens als de benodigde annotaties.

Dit proefschrift vangt aan met een diepgaand empirisch onderzoek naar de belangrijkste kenmerken van reële data—zoals dataverschuiving, signaalcomplexiteit en sensorvariabiliteit—en de operationele omstandigheden die de prestaties en betrouwbaarheid van conventionele deep learning-modellen ernstig ondermijnen. Door een grondige evaluatie van vooraf getrainde modellen voor gebeurtenislabeling (event tagging) en zelfgetrainde modellen

voor anomaliedetectie op authentieke stedelijke akoestische data, onthult en kwantificeert dit werk de ernstige impact van locatiespecifieke context, temporele verschuivingen en de beperkingen van gesloten taxonomieën op de robuustheid en betrouwbaarheid van de modellen. De studie benadrukt verder de uitdaging van de enorme datavolumes in de praktijk; een conceptueel hybride edge-cloud systeem toonde aan dat gelokaliseerde, edge-gebaseerde anomaliedetectie de hoeveelheid data die cloudanalyse of menselijke beoordeling vereist, met ongeveer 90% kan verminderen. Deze bevindingen vormen gezamenlijk een solide, empirisch onderbouwde basis die de dringende noodzaak van nieuwe, geavanceerde oplossingen bevestigt om de betrouwbare inzet van deep learning-frameworks in reële scenario's mogelijk te maken.

Als antwoord op de privacyrisico's die inherent zijn aan dataverzameling in reële stedelijke akoestische omgevingen, stelt dit proefschrift een nieuw "opt-in"' privacybeschermend raamwerk voor audiotoepassingen voor. In schril contrast met traditionele "opt-out" autorisatiemechanismen, die vaak onpraktisch zijn en niet in overeenstemming met het strikte dataminimalisatieprincipe van de GDPR, stelt het voorgestelde opt-in raamwerk de gebruiker in staat om actief en bewust te kiezen welke specifieke informatie wordt gedeeld, in plaats van te moeten specificeren welke informatie wordt achtergehouden. Dit biedt een aanzienlijk intuïtievere en meer gebruikersgerichte benadering van privacybescherming. Met het oog op een naadloze compatibiliteit met bestaande bewakingssystemen, functioneert dit raamwerk als een plug-in optie, waardoor de implementatiekosten voor integratie in bestaande systemen aanzienlijk worden verlaagd. Het vermogen om privé-informatie zoals geslacht, identiteit en emotionele toestand te beschermen wordt rigoureus geëvalueerd op verschillende publiek beschikbare spraakgebaseerde datasets. Experimentele validatie op vier spraakdatasets toont aan dat het raamwerk een zeer effectieve bescherming biedt, waarbij de herkenning van ongeautoriseerde attributen wordt gereduceerd tot bijna willekeurig niveau met slechts een minimale prestatievermindering van 2-6% in herkenningsnauwkeurigheid voor de geautoriseerde taken. Verdere experimenten op een mobiel apparaat, zoals de NVIDIA Jetson, tonen bovendien aan dat het voorgestelde raamwerk in real-time kan opereren.

Om tegemoet te komen aan de praktische noodzaak van generieke modellen die voor meerdere toepassingen kunnen functioneren zonder de noodzaak van uitgebreide, taakspecifieke labeling, stelt dit proefschrift een nieuw zelfgesuperviseerd (self-supervised) raamwerk voor. In plaats van uitsluitend audio- of videodata te gebruiken, benut het raamwerk de synergie tussen beide modaliteiten om complementaire informatie vast te leggen, wat tegelijkertijd een natuurlijke temporele aanwijzing biedt voor het proces van contrastief leren. Hoewel contrastief leren de gangbare techniek is voor zelfgesuperviseerd leren, hanteert het vaak te strikte methoden die uitsluitend vertrouwen op temporele overeenkomst om positieve (semantisch vergelijk-

bare) gebeurtenissen te definiëren. Hoewel een dergelijke techniek in de meeste domeinen goed functioneert, zijn bewakingsdata uniek omdat ze vergelijkbare gebeurtenissen op zeer verschillende, onvoorspelbare tijdstippen vastleggen. Deze eigenschap zorgt ervoor dat conventionele contrastieve leermethoden deze vergelijkbare gebeurtenissen onjuist als negatief beschouwen ("false negatives"), wat onvermijdelijk leidt tot suboptimale resultaten. In plaats van het frequent terugkeren van vergelijkbare gebeurtenissen als een foutbron te behandelen, benut dit raamwerk deze voorvallen strategisch als rijke, veelzijdige positieve paren. Met de temporele overeenkomst tussen audio- en visuele data en de similariteit van de inbeddingen (embeddings) tussen dataparen, is het centrale Embedding-based Pair Generation (EPG) mechanisme ontworpen op basis van dit principe om de kritieke problemen van "false negatives" en de informatie-bottleneck in contrastief leren te overwinnen. De doeltreffendheid en de algemeenheid van de geleerde representaties werden gevalideerd in meerdere "downstream" taken. Het raamwerk behaalde met name een prestatieverbetering van 10% ten opzichte van basismethoden in audiovisuele gebeurtenislokalisatie, en toonde ook een robuuste effectiviteit in andere taken zoals anomaliedetectie en het zoeken naar gebeurtenissen.

Tot slot introduceert dit proefschrift een praktische oplossing voor een grote uitdaging bij de ontwikkeling van locatiespecifieke modellen: het selecteren van het meest overdraagbare model onder de zware beperkingen van rekenkracht, labelkosten en ethische overwegingen. Het voorgestelde nieuwe raamwerk voor de beoordeling van overdraagbaarheid is specifiek ontworpen voor "source-free" (dat wil zeggen, zonder toegang tot de originele trainingsdata) en niet-gesuperviseerde (unsupervised) omgevingen. Het maakt gebruik van een ensemble van willekeurig geïnitialiseerde neurale netwerken (RINNs) en similariteit op het niveau van de embeddings om vooraf getrainde modellen effectief te rangschikken zonder dat brongegevens of doellabels vereist zijn, waardoor de vaak onbetrouwbare aanname van taakafstemming wordt omzeild. Geëvalueerd op reële bewakingsdata, presteert het voorgestelde raamwerk consistent beter dan de state-of-the-art basismethoden. Het behaalt hoge Kendall's correlaties met de ground-truth prestaties voor downstream taken, met waarden van 0.95, 0.94 en 0.89 voor respectievelijk object tagging, anomaliedetectie en gebeurtenisclassificatie. De kwantitatieve bevindingen benadrukken de aanzienlijke waarde van dit raamwerk in uitdagende, reële,"source-free" niet-gesuperviseerde scenario's, met name in privacygevoelige en resource-beperkte slimme stadsomgevingen. Dit raamwerk vult de fundamentele, maar vaak onderbelichte, eerste stap voor modelaanpassing bij de implementatie van locatiespecifieke modellen in slimme bewaking.

Dit proefschrift sluit af met een integratie van deze bevindingen, waarbij de beperkingen van hedendaagse slimme bewakingsframeworks op reële data worden aangetoond en de effectiviteit van de voorgestelde technieken wordt

gevalideerd. Hoewel dit werk robuuste oplossingen presenteert, erkent het ook zijn beperkingen, zoals de noodzaak om de voorgestelde methoden te valideren voor een breder scala aan taken en om de computationele haalbaarheid van grootschalige implementatie aan te pakken. Voortbouwend op deze bevindingen, komen toekomstige onderzoeksrichtingen, waaronder de ontwikkeling van nieuwe benchmarks en meer mensgerichte systeemontwerpen, naar voren als kritieke volgende stappen om de kloof tussen academische modellen en de operationele eisen van de praktijk te overbruggen. Uiteindelijk pleit dit onderzoek voor een "context-first" paradigma dat prioriteit geeft aan een diepgaand begrip van de aard en de beperkingen van data en toepassingen alvorens oplossingen te ontwikkelen. Dit biedt een pragmatisch, systeemgericht perspectief bij het ontwerpen van intelligente bewakingstoepassingen die aanpasbaar, efficiënt en ethisch verantwoord zijn.

Summary

Smart surveillance is a crucial domain in smart city development. These applications are deployed to enhance numerous facets of urban life, including public safety, scene understanding and incident response, thereby improving the living quality and safety of the residents. However, the development of robust solutions is impeded by practical constraints, including the costs of data annotation and cleaning, limitations on computational resources, and ethical issues in data collection and information sharing. Consequently, while extensive research in machine learning models has advanced the performance of smart surveillance applications, a significant gap persists between models developed under controlled conditions and their reliability in real-world deployments.

Aiming to fill the gap, this dissertation identifies limitations in applying conventional deep learning techniques to real-world smart surveillance scenarios and develops transferable, privacy-aware, and label-free learning solutions. Focusing on audio and/or visual modalities, this work tackles the challenges on different aspects: performance degradation resulting from data and task drift, significant privacy risks, and practical barriers posed by restricted access to source data and annotations.

This dissertation starts with an empirical investigation on the primary real-world data characteristics (e.g., data drift, signal complexity, sensor variability) and operational conditions that critically impair the performance and reliability of conventional deep learning models. By evaluating pretrained event tagging and self-trained anomaly detection models on real-world urban acoustic data, this work reveals and quantifies the severe impact of location-specific context, temporal drift, and the constraints of closed-set taxonomies on model robustness and reliability. The investigation further highlights the challenge of data volume in practical deployments, where a conceptual hybrid edge-cloud system showed that localized, edge-based anomaly detection could reduce the data requiring cloud analysis or human review by approximately 90%. These findings collectively establish an evidence-based foundation, confirming the critical need for novel solutions that enable the reliable deployment of deep learning frameworks in real-world scenarios.

In response to the privacy risks inherent in real-world urban acoustic data collection, this dissertation proposes a novel opt-in privacy-preserving framework for audio applications. In contrast to traditional opt-out privacy authorization mechanisms, which are often impractical and failed to align with GDPR's data minimization principle, the proposed opt-in framework allows users to actively choose what information to share rather than what to withhold, providing a more intuitive and user-centric approach to privacy protection. Aiming to be compatible with existing surveillance systems as on-edge devices, this framework functions as a plug-in option for existing surveillance systems, reducing the cost of embedding into existing systems. Despite the lack of urban acoustic datasets on sensitive attributes, the ability of protecting private information such as gender, identity, and emotional state while maintaining the performance of original tasks are evaluated on several speech-based datasets. Experimental validation on four voice datasets demonstrates that the framework provides effective protection, reducing unauthorized attribute recognition to near-random chance, with only a minimal 2-6% performance degradation in recognition accuracy for authorized tasks. Further experiments on resource-constrained device, such as the NVIDIA Jetson, show that the proposed framework can operate in real-time.

To address the practical need for general-purpose models that can function across multiple applications without extensive, task-specific labeling, this dissertation proposes a novel self-supervised framework. Instead of taking only audio or video data, the framework takes both modalities to capture complementary information, simultaneously offering a temporal cue for contrastive learning. While contrastive learning serves as the mainstream technique for self-supervised learning, it employs overly strict policies that often rely solely on temporal alignment to define positive (similar) events. While such a technique works well in most domains, surveillance data is unique in that it captures similar events across diverse time frames. This characteristic causes conventional contrastive learning methods to incorrectly consider those similar events as negatives and thus leads to suboptimal results. Instead of treating the frequent recurrence of similar events as a source of error, this framework strategically leverages these occurrences as rich, multi-aspect positive pairs. With the temporal alignment between audio and visual data and the embedding similarity between data pairs, the core Embedding-based Pair Generation (EPG) mechanism is designed based on this principle to overcome the critical issues of false negatives and the information bottleneck in contrastive learning. The efficacy and generality of the learned representations were validated across multiple downstream tasks; notably, the framework achieved a 10% performance improvement over baseline methods in audio-visual event localization, while also demonstrating robust effectiveness in other tasks such as anomaly detection and event search.

Finally, this dissertation introduces a practical solution to tackle a major challenge in the development of location-specific model: selecting the most transferable model under constraints of computational resources, labeling costs and ethical considerations. The proposed novel transferability assessment framework is designed specifically for source-free (i.e., without access to the original training data) and unsupervised settings. This framework utilizes an ensemble of Randomly Initialized Neural Networks (RINNs) and embedding-level similarity to effectively rank pretrained models without requiring source data or target labels, crucially bypassing the often-unreliable assumption of task alignment. Evaluated on real-world surveillance data, the proposed framework consistently outperforms the state-of-the-art baselines, attaining high Kendall's τ correlations with ground-truth performance across multiple downstream tasks, achieving values of 0.95, 0.94, and 0.89 for object tagging, anomaly detection, and event classification, respectively. The quantitative findings highlight the value of this framework in challenging real-world source-free unsupervised scenarios, particularly in privacy-sensitive and resource-constrained smart city environments. This framework addresses the fundamental yet often underexplored primary step for model adaptation in location-specific models deployment in smart surveillance.

This dissertation concludes by integrating these findings, demonstrating the limitations of contemporary smart surveillance frameworks on real-world data and validating the effectiveness of the proposed techniques. While this work presents robust solutions, it also acknowledges its boundaries, such as the need to validate the proposed methods across a broader scope of tasks and to address the computational feasibility of large-scale deployment. Building on these findings, future research directions, including the development of new benchmarks and more human-centric system designs, emerge as critical next steps to bridge the gap between academic models and real-world operational demands. Ultimately, this research argues for a "context-first" paradigm: one that prioritizes a deep understanding of the nature and constraints of data and application before developing solutions. It thereby offers a pragmatic, systems-aware perspective in designing intelligent surveillance applications that are adaptable, efficient, and ethically aligned.

1

Introduction

Meaning lies not in possession of results, but in the integrity of participation.
I do not strive for ownership of the outcome;
I strive because the work itself matters, and because it may allow others to go further, build higher, or see more clearly.
That, in itself, is enough.

- Unattributed reflection

As urban populations grow rapidly, cities increasingly adopt smart city infrastructure to enhance quality of life and public safety in a sustainable manner. These infrastructures rely on a wide array of interconnected sensors, including smart grids for energy distribution, video surveillance for urban monitoring, and audio sensors for capturing environmental soundscapes. Among these, video and audio data from surveillance systems play a critical role in urban situational awareness. Video sensing offers rich spatial continuity, fine-grained spatial localization, and persistent visibility of physical interactions, particularly in structured or crowded scenes where acoustic overlap may obscure key events. While visual sensing provides rich spatial information, audio offers unique advantages in operational settings. Unlike cameras, microphones are not limited by lighting or occlusion and can detect events beyond the visual field of view. This makes audio a foundational modality for detecting incidents such as shouting, crashes, or distress signals, often serving as the first indicator when visual cues are absent or ambiguous. When combined, these modalities offer a complementary perspective, enhancing detection through cross-modal cues, such as audio-visual synchrony in traffic violations, coordinated behavior in anomalous events, or reinforced identity tracking across frames and sounds. Leveraging these complementary modalities, either individually or jointly, enables a wide array of applications critical for urban management. These include real-time public safety monitoring (e.g., anomaly detection, distress signal recognition, violence detection), enhancing traffic control (e.g., detecting dangerous events, analyzing population flow), and facilitating crime investigation (e.g., tracking suspects, forensic analysis). Despite the growing deployment of these sensor networks and their diverse utility, translating such raw sensor data into actionable insights remains a major challenge.

1.1 From Smart City to Smart Surveillance

1.1.1 Smart City

Smart cities utilize advanced digital technologies, including Internet of Things (IoT) devices, Information and Communication Technologies (ICT), and big data systems, to monitor, control, and integrate diverse urban functions. These technologies are intended to increase operational efficiency and support more responsive and sustainable public service delivery [1]. While the contemporary understanding of a smart city heavily emphasizes its technological backbone, early conceptualizations were often diverse and defined through the lens of specific domains such as transportation, education, governance, and economy. This initial field-centric view has evolved, with re-

cent definitions increasingly highlighting the foundational role of ICT and IoT in integrating these diverse urban functions and enabling data-driven decision-making [2]. From environmental monitoring and traffic regulation to safety management and responsive public infrastructure, these technologies form the backbone of modern smart city systems. At their core lies a data fabric enabled by sensor networks that continuously capture the physical and social dynamics of urban life.

1.1.2 Smart Surveillance

Smart surveillance aims to develop an automatic system that monitors, analyses, and responds to environmental or behavioural signals in real time. This system encompasses sensor data collection and analytics, covering aspects such as security, safety, policy enforcement, and situational awareness [3]. Modern deployments aggregate city-scale streams of video, audio, and increasingly multimodal inputs such as thermal [4], radar [5], and LiDAR [6] data. A key distinction of modern smart surveillance from traditional approaches lies in its reliance on big data pipelines and machine learning techniques. These methods mine patterns and drive decisions at a scale and speed unattainable by manual monitoring, thereby reducing labor costs and mitigating operator bias [7].

While smart surveillance benefits multiple smart city applications across different aspects, in this dissertation, we mainly focus on the public safety. Typical smart surveillance applications comprise three functional categories: latency- or privacy-sensitive tasks executed at the edge, cloud-based semantic interpretation, and large-scale retrieval-oriented applications. Although many existing works show promising results for these applications [8, 9, 10], they are often evaluated under well-controlled conditions using large annotated datasets, computationally powerful devices, and even scripted data. Comparatively, studies explicitly designed to address the challenges related to smart surveillance deployment remain underdeveloped. The key challenges when transitioning from research settings to practical implementations arise from the available budget in computational resources, the lack of annotation, and more restricted privacy requirements. With a focus on public safety scenarios, we aim to address those obstacles by developing adaptable, privacy-aware, and label-efficient learning techniques tailored to the challenges of real-world smart-surveillance deployments.

1.2 From Sensing to Action: The Decision Pipeline

Smart surveillance systems are built upon sensor networks embedded throughout the urban fabric, such as cameras, microphones, and other devices installed in intersections, transit hubs, and public buildings that stream environmental and operational signals in real time. To balance latency, privacy, and scalability, these signals flow through a distributed edge-cloud-decision pipeline. This decision pipeline is illustrated in Figure 1.1, which demonstrates sensors collecting data, followed by processing at edge devices, further analysis in the cloud, and finally, the corresponding responses delivered by human operators or automatic response systems.

Edge infrastructure is placed near sensors. This strategic placement necessitates a dense distribution of edge devices, which consequently limits their computational power to reduce overall cost. Therefore, services deployed at the edge mainly involve lightweight tasks, such as video compression, face blurring, or audio denoising, aimed at curbing bandwidth usage and protecting personal data before transmission. Such early filtering is essential when thousands of sensors can generate petabytes per month and when privacy statutes prohibit raw imagery from traveling off-site. Also critical, low latency applications like anomaly detection or fire/smoke detection, demand immediate processing close to the source and are therefore deployed on edge devices.

Cloud infrastructure, on the other hand, offers a pool of large computational resources, ranging from data storage to processing power. These resources allow cloud services to process vast amounts of data and perform applications that are computationally expensive. Thus, in the smart surveillance pipeline, for richer analytics or multi-sensor fusion, extracted features or anonymised snippets are forwarded to cloud servers, where deeper inference supports functions such as multi-camera tracking, cross-modal event reasoning, and retrospective video search. The result of the analytics may trigger automated actuators, such as adaptive street-lighting, dynamic traffic signals, or notify human operators for interventions that still require judgment. This edge-cloud-decision loop constitutes the operational backbone of modern smart-surveillance deployments, ensuring that data-driven decisions remain possible even in large-scale, bandwidth-constrained, and privacy-sensitive urban environments.

In the remainder of this section, we present a functional overview of smart surveillance tasks relevant to public safety applications. The goal is to organize and contextualize these tasks based on the nature of the applications

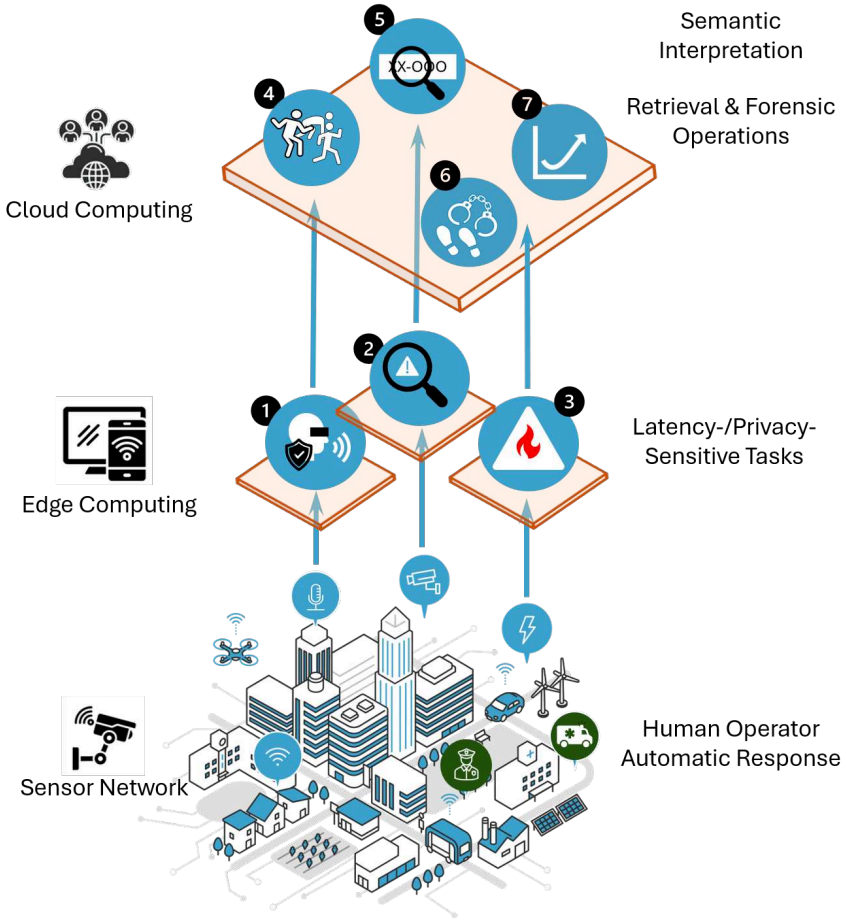


Figure 1.1: The illustration of a smart surveillance infrastructure. Data collected by sensors, such as CCTVs and microphones, are transmitted to edge devices for initial, latency- or privacy-sensitive tasks. These tasks span from ① privacy attribute filtering, ② anomaly detection, to ③ fire/smoke detection. Subsequently, these data are relayed to the cloud for more computationally intensive tasks, or post-event analysis or large-scale analysis. This includes semantic interpretation, such as ④ violent detection or ⑤ vehicle and license plate recognition, as well as retrieval, and forensic operations like ⑥ event querying or ⑦ traffic analytics. The resulting information can then be directed to human operators for further action or used to trigger automatic responses. (Incorporates images from Adobe Stock [ID: 906465341], Adobe Stock [ID: 1490526884] and Adobe Stock [ID: 594844550], modified by author).

and their corresponding placement in the edge-cloud-decision pipeline. This section further highlights how each supports incident understanding and responsive decision-making in urban environments. Rather than attempting to list every possible surveillance task, we focus on a representative subset in which audio or visual signals are critical, and where semantic reasoning or timely response is essential. This narrowed scope reflects both the practical deployment priorities of public safety systems and the focus of this dissertation.

1.2.1 Edge-level Processing

We begin with the edge-level applications, which typically require immediate response, or have strict privacy requirements. These tasks, as outlined in Table 1.1, serve as early-stage filtering, redaction, or triage before data is transmitted or stored.

The first category is filtering privacy-sensitive attributes [14, 11, 15], which includes recognizing biometric signals, such as faces and speech. The privacy-sensitive data must comply with the guidelines from the European Data Protection Supervisor [16], as well as the General Data Protection Regulation (GDPR) [17]. These regulations enforce specific rules for the collection and transmission of such data. In public spaces, GDPR clearly states that data can only be used for specified and legitimate purposes and should be limited to what is necessary (data minimization). This means that, when the scenario does not require certain privacy-sensitive information, systems should perform on-device anonymization on either visual or audio data before transmission or processing.

A second category is anomaly detection [9, 18, 19], which involves identifying events that deviate significantly from typical behavioral patterns in real-time. In such contexts, anomalies may indicate urgent or unexpected situations, such as a person collapsing, a machine malfunctioning, or an unrecognized sound spiking. These incidents require immediate action, which can be provided through a near-edge human response or, crucially, through an automated response system. Anomaly detection is crucial in environments where the range of possible threats or incidents is too broad to predefine. Consequently, it is practically impossible to create an exhaustive catalog of all potential harmful events, ruling out machine learning approaches for event classification or detection. Instead, anomaly detection systems operate by modeling what constitutes normal activity for a specific environment and by flagging deviations that warrant further inspection. Anomaly detection can be implemented using either statistical models or self-supervised approaches, leveraging patterns in audio, visual, or audio-

Stage	Core Purpose	Task Examples	Deployment Note
Early-level Processing	Low-latency, privacy-aware sensing	Privacy-Attribute Filtering, Anomaly Detection, Fire Detection	Primarily on edge devices for privacy and latency efficiency

Privacy Attribute Filtering

Anomaly Detection

Fire/Smoke Detection

Table 1.1: Three examples of smart surveillance edge-level applications. The illustrative images showcase Privacy Attribute Filtering (Source: De Coninck et al. [11]), Anomaly Detection (Source: Wang et al. [12]), and Fire/Smoke Detection (Source: Catargiu et al. [13]) as key examples.

visual signals to localize temporally or spatially abnormal segments. Given that anomaly detection is crucial for emergency incidents, this application is strategically placed at the edge for several reasons. Besides the argument of low-latency, there is a second important argument for deploying anomaly detection algorithms on the edge. Because anomalies are context-specific, a different machine learning model is required per location; ruling out the possibility to achieve scalability gains in inference cost by batching inputs from different sensors. Also, a distributed deployment helps managing the potentially immense data volumes generated by numerous sensors by processing information closer to its origin, rather than transmitting everything to a centralized cloud. This local processing also enhances operational reliability, as detection can continue even if network connectivity is unstable or temporarily lost.

A third edge-suitable application is fire or smoke detection [20, 21, 13]. In public buildings, transit hubs, or industrial zones, the ability to recognize early signs of fire is essential for reducing response time and preventing escalation. These systems detect visual or thermal cues associated with combustion, such as flame flickering, smoke plumes, or thermal hotspots, and can generate alerts before alarms are triggered through traditional means. Because emergencies can unfold rapidly and bandwidth or connectivity may be unreliable during a crisis, local execution of fire detection is both practical and necessary. This task involves processing raw video or thermal inputs and triggering low-latency alerts if indicative patterns are recognized.

1.2.2 Semantic Interpretation

When there is no demand for low-latency computation, or the tasks require integrating signals from multiple sensors, applications are deployed in the cloud. These applications, as outlined in Table 1.2, aim to produce higher-level semantic understanding of the observed environment, enabling event abstraction, cross-modal reasoning, and automated alerting.

One representative task is audio event(sound) tagging [25, 26]. Unlike event classification, which assumes a single dominant label per data sample, event tagging allows multiple events to co-occur and be labeled simultaneously. This makes it particularly suitable for surveillance contexts where overlapping activities are common; for instance, the tags *people screaming* and *siren ringing* may both occur within the same scene. Tagging operates at the short-segment level, assigning high-level semantic labels (e.g., *fire truck*, *shouting*, *siren*) to indicate the presence of predefined event types. The goal is not precise spatial localization but rather a dense, temporally-aware semantic summary that reflects the active scene components. This enables

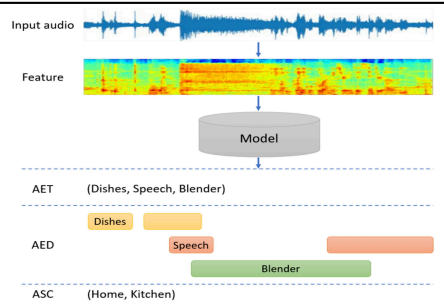
Stage	Core Purpose	Task Examples	Deployment Note
Semantic Interpretation	Real-time reasoning, category prediction, multi-source fusion	Event Tagging, Event Classification, Vehicle-Type & License-Plate Recognition	Requires richer models and broader context, typically cloud or micro-data-centers
 Audio Event Tagging(AET)		 Violence Detection	 Vehicle/License-Plate Recognition

Table 1.2: Examples of smart surveillance on semantic interpretation. The illustrative images showcase Audio Event Tagging (Source: Hou, Yuanbo [22]), Violence Detection (Source: Vijeikis et al. [23]), and Vehicle/License-Plate Recognition (Source: Al-Batat et al. [24]) as key examples.

flexible retrieval, index construction, and triage in complex environments. Due to the need for fine-grained temporal modeling, multi-label output structures, and often multi-sensor fusion (e.g., audio-visual co-occurrence), event tagging is suitable to be deployed in cloud environments where higher-capacity models can be maintained and updated. Notably, the adoption of event tagging (multi-label detection) is less of a research focus in video than in audio. Numerous works on audio tagging were spurred by academic benchmarks such as the DCASE challenge [25]. However, as these challenges concluded, this line of work is largely lessened. Conversely, visual surveillance research has consistently prioritized object detection and event localization. While both directions are important, the event tagging task is crucial yet remain underexplored in complex, real-world scenes.

Another example task is violence detection [23, 27], as violent incidents in public spaces pose significant safety risks. Detecting violence in real-world surveillance footage is particularly challenging because the same physical cues could occur in both benign and threatening contexts. For instance, people moving closely together in small groups with loud noise could reflect the celebration after a sport match. Meanwhile, the small group shouting could also an indication of a soon to be escalated conflict between different parties. These subtle scene dynamics highlight the importance of location-aware semantics and scene-level reasoning, and the importance of integrating audio as an additional cue to help identify the scene (cheering/singing for match). Technically, violent scene detection involves modeling both motion patterns (e.g., abrupt trajectories, body pose changes) and acoustic signals (e.g., raised voices, impact sounds), often using deep multi-modal architectures. The system must be robust to noisy environments and avoid false positives that could trigger unnecessary interventions. Due to its reliance on temporal fusion, multi-modal cues, and contextual understanding, this task is typically executed in the cloud, where sufficient resources and scene-wide correlation are available to support complex inference.

Lastly, vehicle-type and license plate recognition [24, 28] supports surveillance functions involving traffic safety, vehicle tracking, and law enforcement. This task combines object detection with vehicle classification (e.g., car, motorcycle, truck) and optical character recognition (OCR) to extract license plate identities from video streams. Unlike simple vehicle counting, this task must handle occlusion, varying viewpoints, and fast motion, often requiring temporal fusion across frames for stable recognition. In city-scale deployments, these systems also need to correlate observations across cameras located in different zones. For example, to perform real-time tracking on a vehicle spotted on one street, cameras in the neighboring area would

have to jointly be considered to avoid losing the target. This requires awareness of camera topology, regional timing, and possible path continuity, all of which are impractical to manage independently on edge devices. Furthermore, maintaining OCR robustness across lighting conditions and plate formats typically relies on large models and consistent updates. These demands, combined with the need to query across distributed footage and coordinate detection in real time, make this task more suitable for cloud or edge-cloud infrastructure.

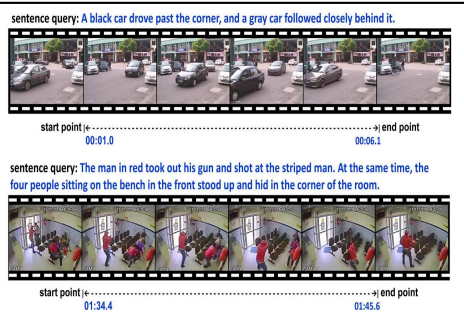
1.2.3 Retrieval and Forensic Operations

We conclude with tasks that support post-event analysis and large-scale search, typically performed over stored surveillance archives. These tasks are designed not for immediate response but for forensic inspection, situational reconstruction, and urban planning. Because they operate on historical or distributed data, they often require access to high-volume storage, semantic indexing, and multi-camera coordination, making them natural candidates for cloud-based deployment. Table 1.3 outlines this taxonomy, categorizing applications based on their core purpose and typical deployment context. The table distinguishes between objectives such as interactive Event Querying, Event Localization, and Traffic Analytics, providing illustrative examples for each.

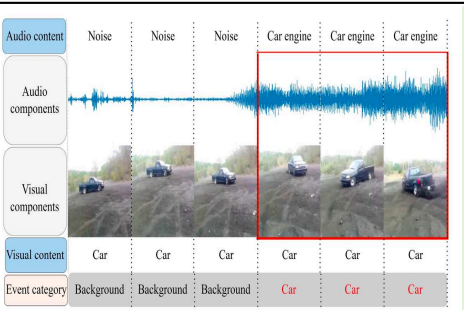
One such task is event querying [10], which addresses the challenge of navigating massive video archives where manual inspection is infeasible. Given a textual query (e.g., *a person entering through the side door*) or a sample clip (e.g., *footage of a known suspect*), the system retrieves semantically relevant segments from a large dataset. This is particularly valuable for investigative workflows, such as crime resolution, incident auditing, or compliance review. The complexity lies in representing both the query and candidate segments in a shared semantic space, often using multi-modal embeddings or cross-modal retrieval methods.

Closely related is event localization [29, 31], which focuses on identifying when a known event type occurred within a video stream, optionally including the spatial location within the video frame. Unlike querying, where the goal is to match a query to the best results, localization assumes the event class is predefined and seeks its temporal or spatial boundaries. For example, given a continuous video feed from a metro station, the system might pinpoint the exact timestamp when a fire alarm was triggered or when a person entered a restricted area. This is essential for timeline reconstruction, incident log alignment, or targeted incident review by security operators.

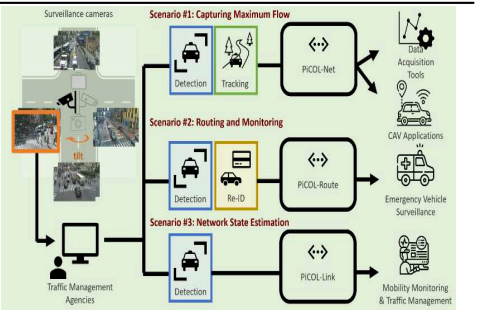
Stage	Core Purpose	Task Examples	Deployment Note
Retrieval and Forensic Operations	Archive-driven retrieval, temporal and spatial indexing	Event Localization, Event Querying, Traffic Analytics	Runs on cloud systems with access to archival data and storage indices



Event Querying



Event Localization



Traffic Analytics

Table 1.3: Examples of smart surveillance on retrieval and forensic operations. The illustrative images showcase Event Querying (Source: Yuan et al. [10]), Event Localization (Source: Ran et al. [29]), and Traffic Analytics (Source: Li et al. [30]) as key examples.

A more data-intensive example is traffic analytics [32, 30], which involves aggregating vehicle or pedestrian flow over extended periods. These analytics are not aimed at individual event detection but at extracting population-level patterns: average vehicle speed, pedestrian congestion hotspots, or route usage over time. Such metrics support urban planning, signal optimization, and emergency preparedness. Because the insights depend on long-term, multi-camera observations and often require historical correlation, traffic analytics tasks rely on archival systems and high-throughput batch processing, typically in the cloud.

1.3 Core Challenges in Urban Surveillance

Before developing applications for smart surveillance, we should first identify the challenges and limitations brought by the characteristics of surveillance data and the practical constraints. In the following sections, we will describe these characteristics, along with their challenges and limitations.

1.3.1 Real-World Data Challenges

While both video and audio in surveillance captures rich information, real-world urban environments exhibit several distinctive properties that shape how such data should be interpreted and utilized in smart surveillance systems. These properties are described from the following aspects: continuity and high volume, data drift across contextual and temporal dimensions, signal complexity, labeling challenges and supervision gaps, and privacy sensitivity.

1.3.1.1 Continuity and High Volume

Video and audio data are collected by tens of thousands of sensors 24/7, recorded without predefined semantic boundaries or event markers. These streams accumulate over extended durations and contain large volumes of mostly uneventful or repetitive content, such as idle street scenes or ambient environmental noise. Events of interest, ranging from brief auditory signals like a car horn to prolonged visual developments such as crowd buildup or vehicle collisions, are temporally sparse and embedded within this background. The lack of clear temporal segmentation complicates the design of event detection systems, which must process vast amounts of data while isolating semantically meaningful segments in real-time. This places significant demands on filtering, temporal abstraction, and scalable storage. Moreover, the prohibitive cost of annotating this high-volume of data makes supervised training of deep learning models infeasible at operational scale;

motivating the use of unsupervised or self-supervised learning strategies. Taken together, these characteristics introduce challenges along two critical axes: scalability and supervision.

To address these challenges, researchers have proposed a range of system-level and learning-level strategies. On the scalability front, recent work has focused on reducing the cost of analyzing all data by introducing mechanisms that prioritize segments likely to contain meaningful activity. In the video domain, Elmir et al. [33] proposed an intelligent video recording system that combines motion detection with object recognition to retain only activity-rich segments, such as scenes involving people or vehicles, while discarding low-activity footage. Another example, Das et al. [34] proposed to first identify video segments that potentially contain certain objects, such as guns, then perform frame-level detection to locate the target object. In the audio domain, despite being less developed than the video domain, recent works have also shown growing interest in scalable event processing. For example, Neri [18], Rodríguez et al. [19] introduced a structured pipeline to first detect anomaly segments before further classification, which could reduce the cost of performing intense computation on audio classification. These approaches illustrate a broader shift in how detection is operationalized: not as a final analytic target, but as an early-stage filter required to cope with the scale and continuity of surveillance streams.

On the supervision front, unsupervised learning has long been explored in both audio and visual domains as ways to mitigate the lack of densely labeled data [35, 36]. These approaches are especially attractive for surveillance, where manual annotation is infeasible at scale. However, defining effective pretext tasks for these methods remains non-trivial. These pretext tasks, which are artificial, auxiliary tasks, are used to train the model as an alternative to human-labeled annotation. Contrastive learning, for example, relies on pairwise similarity assumptions that often fail in continuous streams, where temporally distant segments may still be semantically related, leading to false negatives and degraded representations. On the other hand, reconstruction-based methods are computationally expensive and inefficient in high-dimensional modalities like video or audio. These limitations highlight the tension between general-purpose self-supervised frameworks and the structural characteristics of unsegmented surveillance data. As a result, both scalability and supervision remain open challenges in systems that aim to operate under real-time constraints and uncertain conditions.

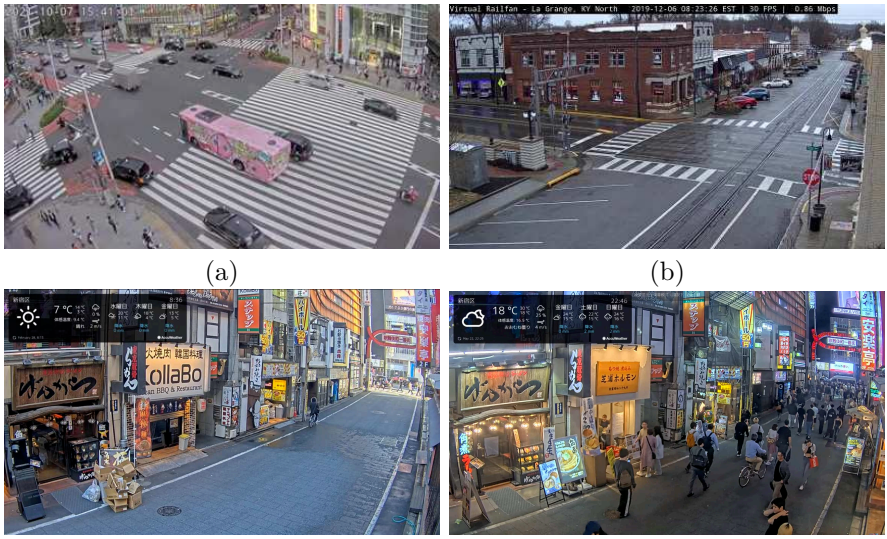
While tasks such as anomaly detection offer practical entry points for managing data overload, the lack of reliable annotation affects nearly every

application discussed in this work, from event retrieval and localization to cross-modal matching and behavior analysis. In this context, unsupervised learning is not merely an efficient alternative, but an essential tool for enabling learning-driven surveillance systems to function under realistic data and deployment constraints.

1.3.1.2 Data Drift Across Contextual and Temporal Dimensions

Urban visual and audio events are strongly shaped by their surrounding context, both in space and time. For instance, a sensor near a traffic intersection may capture frequent honking, structured vehicle flows, and dense pedestrian crossings, while one in a park records slower-paced movement, sparse acoustic activity, and unstructured group formations. Over time, contextual conditions at a single site may also evolve: visual and acoustic signatures can shift with seasonal changes, lighting conditions, infrastructural updates, and human activity cycles. Figure 5.1 further illustrates this with example frames from different cameras or at different times. These variations make surveillance data inherently non-stationary and create two major challenges. First, they alter the very definition of events and anomalies: what is considered abnormal in one location (e.g., standing in the middle of the road) may be entirely normal in another (standing in the middle of the park). Second, they pose a risk to model performance, as systems trained on data from one context may not generalize well to another. This mismatch complicates anomaly detection and weakens smart surveillance applications in general, requiring either context-specific learning or robust adaptation strategies.

A growing number of works have highlighted the limitations of one-size-fits-all models in real-world surveillance. Prior studies [37, 38, 39] point to the importance of adapting models to individual deployment sites, especially when the definition of normal state or risk varies drastically across environments. To acquire the location-specific machine learning models more efficiently, instead of struggling with training from scratch with limited, unlabeled data for the target deployment location, an alternative solution is to apply transfer learning. This technique leverages knowledge learnt from a source domain to improve a model's performance when applied to a target domain. It is particularly effective for addressing the performance degradation caused by the data drift, concept drift, or task drift between source and target domain. These transfer learning approaches have often utilized models pretrained on large, annotated datasets for general tasks like image or video classification. However, for domains such as smart surveillance where target-specific annotations are limited, leveraging models pretrained



(a) (b)

(c) (d)

Figure 1.2: Example frames to illustrate contextual and temporal drift, captured from publicly accessible YouTube live CCTV streams. (a) A high-level view from a stream in Omoide Yokocho, Shinjuku, Tokyo, Japan, showing a busy intersection with pedestrians and various vehicles. (b) An intersection with a railway and a passing train, from a stream in La Grange, Kentucky, USA. (c) A walking street with a lower viewing angle, primarily containing pedestrians, from a stream in Kabukicho, Shinjuku, Tokyo, Japan. (d) From the same YouTube live stream as (c), but captured at a different timestamp. Frames (a)-(c) are from cameras with different configurations and locations, illustrating contextual drift. The differences in lighting and crowdedness between (c) and (d) (from the same camera at a different time) demonstrate temporal drift.

via self-supervision on extensive unannotated data is particularly effective. Despite increasing attention in the broader transfer learning community, such strategies remain relatively uncommon in real-world surveillance deployments, due to scalability concerns and the maintenance overhead associated with managing context-specific models.

1.3.1.3 Signal Complexity

Audio and visual surveillance streams in urban settings are marked by high signal complexity. In both modalities, events occur in dense environments with overlapping sources and limited separation, creating significant challenges for clean interpretation. This density is not merely a matter of volume but arises from entangled content and diverse object or source types; consequently, there are only limited properly annotated datasets available that reflect these characteristics.

Visual surveillance data is often impaired by occlusion and multi-scale variation, where people, vehicles, and objects may appear at varying resolutions, be partly obstructed, or be embedded in crowded scenes. As illustrated in Figure 1.3, these conditions complicate consistent object detection or tracking, and make high-level activity interpretation brittle. This frame also shows how multiple events may unfold simultaneously. Audio data, meanwhile, is challenged by overlapping sources and reverberation, which distort temporal structure and blur signal boundaries, especially in single-channel settings.

For both modalities, a significant limitation is the misalignment between the outputs of available machine learning models and the requirements of comprehensive scene understanding, largely due to the structure of available datasets and benchmarks. The computer vision community has traditionally prioritized object detection [40] and recognition [41]. Object detection identifies and localizes discrete entities with a frame (e.g., 'car', 'person'); which provides an inventory of a scene but does not, by itself, describe the events taking place. Similarly, classification models, which map a data sample to a single category, are insufficient for scenarios with concurrent activities. However, video surveillance data contains multiple, co-occurring events. Therefore, event tagging is required to assign semantic labels to video segments that describe actions, interactions, and contextual states, such as 'vehicle turning left', 'pedestrian crossing', or 'person falling.' An event tagging model assigns multiple labels flagging all relevant categorizations above a certain confidence. While existing datasets for visual models are often developed for object detection tasks, in the audio domain, tagging-based scenarios are more common in particular due to the DCASE

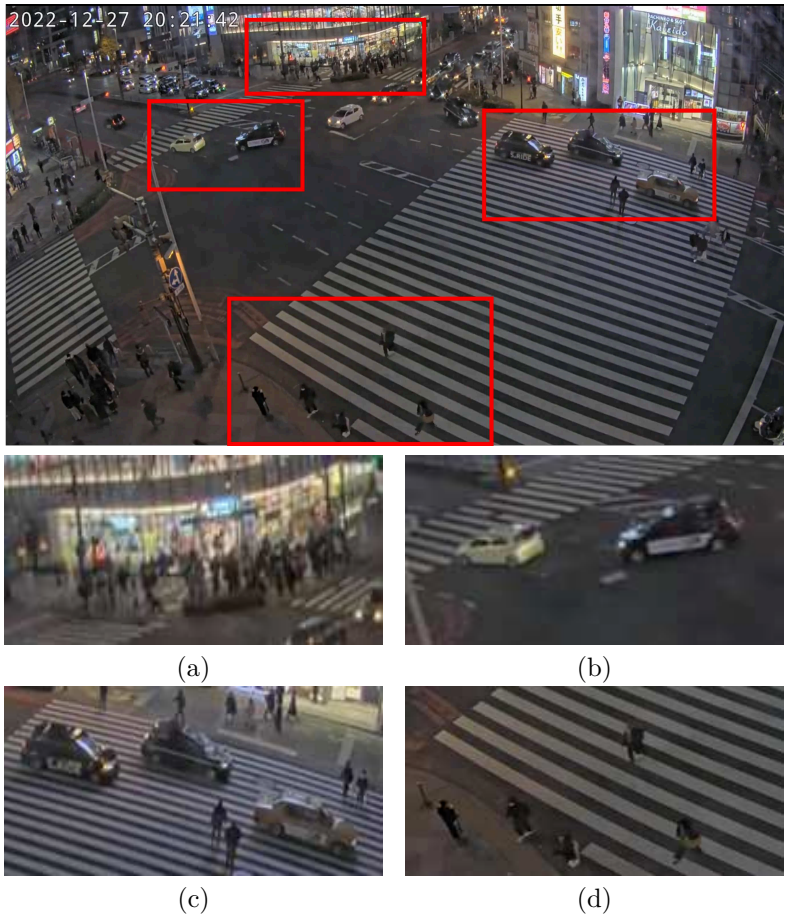


Figure 1.3: An example frame from a stream in Omoide Yokocho, Shinjuku, Tokyo, Japan demonstrates the complexity of a real-world scenario. The top image illustrates the camera view, while (a)-(d) are the four rectangular areas in the scene. From a single frame, we observe vehicles with diverse sizes and viewing angles. Furthermore, this timestamp contains more than four different events: (a) a group of people gathering for more than 10 minutes, (b) several cars making a right turn, (c) a traffic violation (taxis try to pass while pedestrians are crossing), and (d) pedestrians crossing the zebra line.

benchmark [25]. Moreover, most visual and audio models are developed for closed-set conditions, where the classes of objects or events are predefined. A surveillance context, however, is an open-set problem, encompassing a vast and unpredictable range of events and object interactions. Therefore, the ability to move beyond simple object identification to tag and understand novel or unseen events is crucial for real-world applications.

Altogether, the complementary nature of audio and visual systems, coupled with the limitations of their respective unimodal approaches, points toward the value of multi-modal fusion. While this direction has been extensively explored in the broader video understanding community [42, 43, 44], its application in urban surveillance remains comparatively underdeveloped. Leveraging the strengths of each modality may provide a pathway toward richer and more adaptable representations that overcome the blind spots of unimodal systems when facing this inherent signal complexity.

1.3.1.4 Labeling Challenges and Supervision Gaps

Surveillance systems often suffer from sparse and inconsistent annotations, which limit the feasibility of supervised learning and challenge evaluation. A major contributor is the cost of producing labeled data tailored to each individual sensor deployment. Since definitions of normality and anomaly are often context-dependent, annotations collected in one environment may not transfer well to another, requiring location-specific effort that does not scale.

Compounding this, the complex and unstructured nature of real-world urban scenes makes annotation inherently difficult. Events are often ambiguous, overlapping, or dependent on localized social norms, which challenges annotator agreement and task design. Moreover, dense and diverse scene content often demands tagging, as multiple concurrent events or co-occurring entities may need to be captured within the same frame or clip. Rare events such as aggression, crashes, or distress signals are especially difficult to capture and label consistently, resulting in limited ground truth for many high-value tasks.

To better cope with such variability, recent works have proposed transferability assessment techniques to determine which pretrained model best suits a given target sensor, aiming to reduce the need for full retraining. With an increasing number of such pretrained models available (referred to as a model zoo), the primary challenge becomes identifying the most adaptable one for each sensor in a scalable way. To address this, methods such as LogME [45], LEAD [46], and uncertainty-aware measures [47]

estimate deployment fitness using proxy signals such as feature alignment or model uncertainty. These are often positioned as a first step prior to model adaptation, enabling more efficient and targeted deployment. In parallel, domain adaptation remains widely used to bridge distributional gaps. Recent methods explore domain-invariant representation learning [48] and conceptual alignment [49], though many still rely on source data or partial supervision, limiting their use in fully unsupervised settings. Altogether, these techniques reflect incremental progress toward addressing annotation sparsity, though robust solutions remain elusive in dynamic surveillance environments.

1.3.1.5 Privacy Sensitivity

Beyond the technical challenges introduced by data drift, complexity, and annotation gaps, urban surveillance systems must also navigate stringent privacy constraints. Video data, particularly from fixed or high-resolution deployments, can expose faces, license plates, body language, and social interactions. Similarly, audio recordings may capture speech, vocal identity, or environmental cues that indicate individual presence or behavior. These risks are heightened by the passive and pervasive nature of sensing, where individuals are often unaware they are being recorded.

These privacy risks are not merely theoretical: legal frameworks such as the GDPR [17] in Europe and other regional policies impose strict regulations on the collection, processing, and storage of personally identifiable information. Even when models do not explicitly analyze sensitive content, the act of capturing such data can undermine public trust and raise ethical concerns. To mitigate these risks, many systems adopt obfuscation techniques targeting specific attributes, such as face obfuscation or de-identification [15] and license plate encryption [50]. In contrast, studies on audio privacy remain limited. Moreover, existing works are typically implemented as opt-out defaults, with the aim of identifying and removing prespecified visual or acoustic privacy-sensitive information. However, the opt-out mechanism is limited to predefined categories and thus does not comply with the data minimization principle in GDPR. From a legal and ethical standpoint, data protection should begin at the point of acquisition. An opt-in model, where individuals retain control over whether and how their data is collected, offers a more robust alignment with GDPR principles and public expectations. Achieving this would require substantial rethinking of sensing infrastructure, consent mechanisms, and governance protocols. During my PhD, my research focused primarily on privacy aspects of audio, while the complementary area of video privacy was concurrently addressed by my colleague,

(a) (b) (c)

Figure 1.4: An example frame demonstrates the possible privacy sensitivity of a real-world scenario. The top image illustrates the camera view, while (a)-(c) are the three rectangular areas in the scene. Deployed in Urban area, it is common that the recorded images contains privacy sensitive information such as (a) faces, and (b) car or (c) motor plates.

Sander De Coninck, particularly in the domain of privacy-preserving visual analysis [14, 11].

1.3.2 Operational Challenges of Model Deployment

Beyond the inherent challenges posed by the nature of urban surveillance data (as discussed in Section 1.3.1), deploying models in real-world settings introduces further operational difficulties. This section will address these practical constraints, which include restrictions on data access and usage permissions that necessitate source-free approaches, computational limitations of deployment hardware 1.3.2.2, and the challenges in robustly evaluating model performance under realistic operational conditions 1.3.2.3.

1.3.2.1 The Source-free Setting

While many recent research frameworks assume access to source data for domain adaptation, practical deployments rarely permit it. Legal and contractual restrictions, particularly in surveillance contexts involving sensitive or private footage, often prevent data storage or redistribution. Although regulatory frameworks such as the European Union's Data Governance Act (DGA) [51] encourage cross-sectoral data sharing to encourage innovation, these efforts sometimes conflict with stringent data protection laws like the GDPR or concerns over proprietary leakage. As a result, the direct exchange of raw data is often infeasible. In response, model sharing, rather than data sharing, has emerged as a practical and privacy-conscious alternative for enabling collaborative development. This shift reinforces the practical relevance of source-free adaptation techniques that operate using only pretrained models and unlabeled target data.

1.3.2.2 Computational Bottlenecks in Edge Deployment

Section 1.2 emphasized the strategic importance of edge computing, particularly for achieving low-latency responses in time-sensitive surveillance applications. Edge devices are generally designed for lightweight tasks due to factors like deployment density and cost. This subsection focuses on the inherent computational bottlenecks of these devices as a critical operational challenge. We will explore how specific limitations in processing capacity, memory, and power directly constrain the feasibility of deploying and adapting more sophisticated deep learning models at the network edge. Surveillance deployments often rely on edge devices with limited computational and memory capacity, such as embedded GPUs or low-power processing units. While exact capabilities vary, they typically constrain model size, memory access, and real-time throughput. These limitations

make it infeasible to deploy large models or resource-intensive adaptation techniques, such as gradient-based fine-tuning, domain-specific normalization, or ensemble-based uncertainty estimation, at the edge. As a result, developers face a trade-off between model expressiveness and deployability. Inference pipelines must be compact, energy-efficient, and robust to hardware variability. While edge-based inference helps satisfy real-time and privacy requirements, it also limits the scope for dynamic model selection or multi-modal reasoning, especially when working under tight memory, power, or platform constraints. This operational bottleneck places a premium on architectural efficiency, early-stage signal triage, and source-free model reuse.

1.3.2.3 Lack of evaluation datasets

While many public datasets exist for benchmarking, few reflect the full complexity, diversity, or constraints of real-world surveillance environments. Evaluation often relies on synthetic [52] or scripted [53] data, under-representing deployment noise, drift, or social sensitivity. In addition, many publicly available surveillance video datasets suffer from significant structural and annotation limitations. They often include only a small number of unique actors or scenes, lack temporal continuity, and are typically composed of short, non-continuous clips that fail to represent real-world temporal dynamics. Severe gaps in annotation, ranging from coarse labels to completely unlabeled segments, further limit their utility in evaluating detection, tracking, or adaptation strategies. Moreover, some datasets have been retracted or deprecated due to ethical concerns or violations of privacy standards, complicating reproducibility and long-term benchmarking. A notable example is the DukeMTMC dataset [54], which was taken offline following investigative reports and public scrutiny over its use by commercial surveillance vendors in China, raising serious concerns about consent, governance, and downstream misuse [55]. These limitations are particularly pronounced in the case of audio-visual surveillance datasets, which remain rare and often under-annotated due to their increased privacy sensitivity and higher collection complexity. As a result, researchers face significant difficulties in benchmarking complex tasks such as cross-modal adaptation, long-term drift handling, or privacy-preserving modeling under realistic conditions.

1.3.3 Research Questions

The goal of this dissertation is to investigate the primary real-world data characteristics (e.g., data drift, signal complexity, sensor variability) and operational conditions that critically impair the performance and reliability

of machine learning models for audio, video, or audiovisual tasks, and to identify solutions, such as system-level or methodological adaptations, that can be applied to mitigate these performance gaps in practical deployments. Based on the challenges detailed in the preceding sections, four research questions can be defined:

Research Question 1:
What are the primary challenges and limitations of deploying conventional acoustic surveillance techniques in real-world urban environments?

While the challenges in visual surveillance, such as varying lighting, occlusions, or open-set events, are well-documented and explored, the specific limitations for acoustic surveillance were less established at the start of this research. This question addresses that gap by empirically evaluating the performance of existing acoustic anomaly detection and sound tagging models on unprocessed, real-world urban audio data. The investigation focuses on quantifying the impact of location-specific context, temporal drift, and the constraints of closed-set taxonomies on model robustness and reliability.

Research Question 2:
How can we learn effective, robust, and generalizable data representations from complex, large-scale urban surveillance streams by leveraging the unique characteristics of multimodal surveillance data to overcome the limitations of conventional unsupervised representation learning methods, such as false negatives and information bottlenecks?

Given the significant challenges posed by factors like lack of annotation, this question explores how unsupervised representation learning techniques can make use of the fundamental properties and structure of surveillance data itself to enable effective analysis without relying on extensive manual labeling.

Research Question 3:
How can a principled and practical privacy-preserving framework be designed to operate on-edge, protecting sensitive attributes in acoustic data through opt-in mechanisms while maintaining compatibility with pre-existing recognition systems?

As urban surveillance inevitably captures sensitive data and existing privacy protection measures often fall short, this question drives the search for robust and practical privacy-preserving solutions, the opt-in mechanisms, suitable for real-world deployment, including on edge devices.

Research Question 4:
How can a principled assessment framework be developed to accurately predict the transferability of pre-trained models under source-free and unsupervised conditions, in order to both guide effective model deployment in the near-term and inform the future development of more adaptable solutions?

This question investigates methods for reliably evaluating the suitability of pretrained models and guiding their adaptation to new surveillance deployments, particularly when faced with common operational constraints such as the absence of source data or target annotations.

1.4 Research Contributions

To address the research questions formulated previously, this doctoral dissertation investigates how state-of-the-art deep learning models perform when deployed in real-world urban surveillance environments, where data is noisy, unlabeled, and privacy-sensitive. Motivated by the gap between data-driven development and operational constraints, this work makes four key contributions: it empirically exposes the limitations of pretrained models on real data; proposes a privacy-aware, opt-in audio framework deployable on edge devices; introduces a self-supervised pipeline for learning robust audio-visual embeddings tailored for smart surveillance applications; and develops a novel, source-free transferability assessment framework grounded in representation structure rather than pretrained biases. Together, these contributions offer a pragmatic, systems-aware perspective on designing intelligent surveillance applications that are adaptable, efficient, and ethically aligned.

Chapter 2: Real-World Acoustic Surveillance

The primary motivation of this work is to experimentally evaluate how state-of-the-art (SOTA) models perform when applied to real-world urban acoustic surveillance data, and to uncover the practical limitations that arise during deployment. Unlike many benchmark-driven studies, this chapter grounds its exploration in data collected from two European cities, offering a lens into real environmental complexity.

We first demonstrate that pretrained models, particularly those for sound tagging, are limited by their predefined taxonomies. Many common urban acoustic events, such as bells, fireworks, or seasonal noises like birdsong, fall outside these taxonomies, leading to misclassifications or complete omissions. Furthermore, it was observed that sound tagging models flagged

numerous acoustic events, the majority of which were of minor operational interest. When encountering vast data volumes, this becomes particularly problematic for efficient real-time monitoring. To address this, a conceptual hybrid edge-cloud system was utilized, wherein edge devices perform lightweight anomaly detection and only forward potentially relevant segments to the cloud for tagging and operator review. This pipeline reduced the data volume forwarded to the cloud by approximately 90% while preserving event coverage.

The chapter also evaluates the cross-context generalizability of both anomaly detection and tagging models. Empirical results across different dates and sensor placements reveal significant performance drops when models trained in one temporal or spatial context are applied to others. These findings underscore the necessity of location- and location-specific models, especially for tasks like anomaly detection where environmental norms vary sharply (e.g., traffic sounds being normal near a road but anomalous in a park).

Despite efforts such as data filtering to reduce the volume of transmitted information, the work in this chapter underscored that privacy remains a significant and unresolved challenge in real-world urban acoustic surveillance. The potential for incidental capture of sensitive information, including human voices, in segments flagged for further analysis, was observed. Such findings drive the developing truly robust and deployable privacy-preserving techniques that address the nuances of complex urban audio data, a task made more challenging by regulatory constraints and the limited availability of suitable datasets for development and evaluation, remains a critical area requiring further dedicated investigation.

The findings suggest key directions for developing tailored architectural and methodological response to improve performance, efficiency, and compliance in smart surveillance deployments. In achieving this, the work in this chapter primarily addresses Research Question 1, offering an empirical understanding of real-world performance gaps in urban acoustic surveillance, while also bringing to operational constraints and privacy considerations that inform other research questions. Ultimately, Chapter 2 establishes a reality-grounded foundation: that pretrained models fall short when faced with complex, unlabeled, and privacy-sensitive urban audio data.

Chapter 3: An Opt-in Framework for Privacy Protection in Audio

This chapter addresses the critical challenge of privacy protection in audio-based applications, aiming to tackle the privacy concern in smart surveil-

lance. However, due to the lack of properly designed datasets, this work was evaluated on a more general acoustic dataset. The motivation stems from the inadequacy of conventional *opt-out* privacy paradigms, where users must specify which attributes to protect, a strategy that cannot feasibly account for all sensitive information that may be extracted from raw audio.

To this end, this chapter presents the first opt-in privacy-preserving framework for audio applications, wherein users explicitly authorize only a target task, and all other information is actively suppressed. The approach centers around a deep neural obfuscator trained via adversarial learning and a novel privacy loss metric. Motivated by the constraints of real-world deployment, the design prioritizes compatibility with edge devices and third-party processing pipelines.

Evaluation is conducted across four speech datasets using three different sensitive attributes (gender, speaker identity, and emotion). Results show strong protection performance under both ignorant and informed attacker models, with only minor degradation in task accuracy.

Importantly, the framework's computational viability is tested on a range of devices, from server-class GPUs to embedded edge devices like the Raspberry Pi and NVIDIA Jetson TX1. While real-time execution is only feasible on GPU-accelerated hardware, these results offer a realistic view of current deployment constraints and justify future work on model compression for broader applicability. The development of this novel opt-in privacy-preserving framework, its underlying mechanisms, and its empirical validation directly address Research Question 3, which focuses on how principled and practical privacy-preserving techniques can be effectively developed and implemented for smart surveillance systems.

Chapter 4: Unsupervised Audio-Visual Representation Learning

This chapter investigates how to construct general-purpose audio-visual representations tailored for the unique characteristics of surveillance data. The central objective is to train models in a self-supervised manner on real-world, long-duration surveillance recordings, enabling them to serve as feature extractors for a variety of downstream tasks without relying on extensive annotation. A key aspect of this work is the explicit integration of audio and visual signals as complementary modalities, rather than treating them independently. By learning from their joint structure, the model captures cross-modal patterns that improve robustness and semantic understanding across varied surveillance tasks.

Two key challenges are addressed in this chapter. The first challenge is the issue of false negatives when applying contrastive learning, which can misguide the contrastive learning objective. Contrastive learning involves a pair selection process, which is typically done by considering temporal-alignment, or Temporal-based Pair Generation (TPG). As surveillance data contains repetitive, irregularly appearing events, such cue is no longer feasible as it would mistake the same event in different time as negative, complicating training convergence. The second is the information bottleneck often seen in contrastive learning. When relying solely on temporally aligned audio-video pairs, the learned features may suffice only for the pretext task but fail to capture broader semantic context. Considering this false negative problem alongside the information bottleneck, the bottleneck is further tightened as the pretext task limits the learnt representation to only mapped to the temporally-aligned pair. This bottleneck limits the generalizability of the representations to downstream tasks, which is particularly problematic as the representations should be used for multiple tasks in real-world deployment. To address both challenges, we propose Embedding-based Pair Generation (EPG), a new mechanism that samples semantically similar pairs based on mutual proximity in the learned embedding space rather than raw timestamps. This approach reduces false negatives, enhances pair quality, and allows the model to capture richer, more diverse relationships between modalities. A modified contrastive loss further supports the inclusion of multiple positives per anchor, increasing the expressiveness of the learned embeddings.

We evaluate the learned representations across multiple downstream tasks, including supervised audio-visual event localization, unsupervised anomaly detection, and event search. We first train the models on four long-form, publicly accessible audio-visual recordings from an urban traffic surveillance camera, totaling over 16 hours of footage across day and night conditions. These recordings enable the model to learn from realistic noise, occlusions, and temporal event continuity. Results show that our method performs competitively with task-specific baselines while generalizing across tasks. Additionally, we observe evidence of concept drift: models trained on scenes with similar context (such as time-of-day and traffic composition) yield better retrieval and detection results, underscoring the importance of context-awareness in real-world deployments.

The development of the EPG mechanism and the self-supervised learning pipeline presented in this chapter directly addresses Research Question 2 by exploring how to effectively leverage the unique characteristics of surveillance data for robust representation learning. Furthermore, the improved

audio-visual representations also contribute to Research Question 1 through its potential to mitigate real-world performance gaps in downstream surveillance tasks.

Chapter 5: Source-Free Model Transferability Assessment

This chapter addresses the challenge of identifying the most adaptive models for real-world surveillance systems under the real-world smart surveillance constraints: source-free and label-free. The primary goal is to develop a transferability assessment framework that requires no access to source training data or annotated target labels, and avoids unreliable assumptions such as representational similarity between pretext and downstream tasks that often undermine current methods.

We propose a new strategy based on Randomly Initialized Neural Networks (RINNs), which serve as an unbiased, task-agnostic reference for evaluating how well the structure of a model's learned representations aligns with the target data. In contrast to traditional approaches that rely on pretrained embedding spaces or task-specific uncertainty metrics, our method leverages RINNs for a direct comparison of embedding consistency that does not require model fine-tuning or pseudo-labeling.

To ensure efficiency and scalability, we incorporate minibatch Centered Kernel Alignment (CKA) as the core similarity measure. This not only preserves robustness in estimating representational structure but also significantly reduces the computational footprint, making the method suitable for large model zoos and resource-constrained surveillance deployments.

We evaluate our method on a comprehensive set of real-world surveillance datasets, including both source and target domains. Models are tested across multiple tasks: object tagging, event classification, and anomaly detection. In all settings, our embedding-level RINN assessment consistently outperforms existing baselines, demonstrating task-agnostic stability and clear alignment with ground-truth performance rankings. Additionally, evidence of concept drift is again observed: models trained on source domains that are environmentally or structurally closer to the target domain yield superior results, highlighting the importance of context-aware selection. By decoupling assessment from pretrained embedding biases and relying instead on RINN-based structural alignment, the framework offers a reliable path toward model selection in privacy-sensitive, heterogeneous, and data-limited smart surveillance systems.

The development of this novel RINN-based transferability assessment frame-

work directly confronts the challenges described in Research Question 4, by providing a practical methodology for evaluating and selecting pretrained models under source-free and label-free operational constraints. Furthermore, by enabling model selection, this work also contributes a valuable tool towards mitigating real-world performance gaps, as addressed in Research Question 1.

1.5 Publications

The research results obtained during this PhD research have been published in scientific journals and presented at an international workshop. The following list provides an overview of these publications.

1.5.1 Journal Publications

[1] **W.-C. Wang**, S. De Coninck, S. Leroux, and P. Simoens, *An opt-in framework for privacy protection in audio-based applications*. Published in Pervasive Computing(IEEE), 21(4) p.17-24, 2022.

[2] S. De Coninck, **W.-C. Wang**, S. Leroux, and P. Simoens, *Privacy-preserving visual analysis: training video obfuscation models without sensitive labels*. Published in Applied Intelligence(Springer), 54(8) p.6041-6052, 2024.

[3] **W.-C. Wang**, S. De Coninck, S. Leroux, and P. Simoens, *Embedding-based pair generation for contrastive representation learning in audio-visual surveillance data*. Published in Robotics and AI(Frontiers), 11:1490718, 2025.

[4] **W.-C. Wang**, S. Leroux, and P. Simoens, *Source-Free Model Transferability Assessment for Smart Surveillance via Randomly Initialized Networks*. Published in Sensors(MDPI), 13: 3856, 2025.

1.5.2 Conference and Workshop Publications

[1] S. De Coninck, **W.-C. Wang**, S. Leroux, and P. Simoens, *Selective manipulation of disentangled representations for privacy-aware facial image processing*. Published in 4th Workshop on Machine Learning for CyberSecurity, 2022.

1.6 References

[1] United Nations Centre for Regional Development (UNCRD). Smart cities: Training material. UNCRD Training Resource, 2023. URL `https://uncrd.un.org/sites/uncrd.un.org//files/smart-city-training-material_1_smart-cities.pdf`.

[2] Filipe Moura and João de Abreu e Silva. Smart cities: definitions, evolution of the concept, and examples of initiatives. In *Industry, innovation and infrastructure*, pages 989–997. Springer, 2021.

[3] Arun Hampapur, Lisa Brown, Jonathan Connell, Sharat Pankanti, Andrew Senior, and Yingli Tian. Smart surveillance: applications, technologies and implications. In *Fourth international conference on information, communications and signal processing, 2003 and the fourth pacific rim conference on multimedia. Proceedings of the 2003 Joint*, volume 2, pages 1133–1138. IEEE, 2003.

[4] Manoj Kumar, Susmita Ray, and Dileep Kumar Yadav. Moving human detection and tracking from thermal video through intelligent surveillance system for smart applications. *Multimedia Tools and Applications*, 82(25):39551–39570, 2023.

[5] Vera Lobanova, Dmitry Bezdetnyy, and Lesya Anishchenko. Human activity recognition based on radar and video surveillance sensor fusion. In *2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pages 025–028. IEEE, 2023.

[6] Pedro Torres, Hugo Marques, and Paulo Marques. Pedestrian detection with lidar technology in smart-city deployments–challenges and recommendations. *Computers*, 12(3):65, 2023.

[7] Himani Sharma and Navdeep Kanwal. Video surveillance in smart cities: current status, challenges & future directions. *Multimedia Tools and Applications*, pages 1–46, 2024.

[8] Kishan Bhushan Sahay, Bhuvaneswari Balachander, B Jagadeesh, G Anand Kumar, Ravi Kumar, and L Rama Parvathy. A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques. *Computers and Electrical Engineering*, 103:108319, 2022.

[9] Huu-Thanh Duong, Viet-Tuan Le, and Vinh Truong Hoang. Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, 23(11):5024, 2023.

[10] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22052–22061, 2024.

[11] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. Privacy-preserving visual analysis: training video obfuscation models without sensitive labels. *Applied Intelligence*, 54(8):6041–6052, 2024.

[12] Wei-Cheng Wang, Sander De Coninck, Sam Leroux, and Pieter Simoens. Embedding-based pair generation for contrastive representation learning in audio-visual surveillance data. *Frontiers in Robotics and AI*, 11:1490718, 2025.

[13] Constantin Catargiu, Nicolae Cleju, and Iulian B Ciocoiu. A comparative performance evaluation of yolo-type detectors on a new open fire and smoke dataset. *Sensors*, 24(17):5597, 2024.

[14] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. Selective manipulation of disentangled representations for privacy-aware facial image processing. Presented at the 4th Workshop on Machine Learning for CyberSecurity (MLCS), co-located with ECML PKDD, September . Grenoble, France.

[15] Lamyanba Laishram, Muhammad Shaheryar, Jong Taek Lee, and Soon Ki Jung. Toward a privacy-preserving face recognition system: A survey of leakages and solutions. *ACM Computing Surveys*, 57(6): 1–38, 2025.

[16] European Data Protection Supervisor. Video surveillance —reference library, 2024. URL `https://www.edps.europa.eu/data-protection/data-protection/reference-library/video-surveillance_en`. Accessed: 2025-05-22.

[17] European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2016. Accessed: 2025-05-07.

[18] Michael Neri. Anomaly detection and classification of audio signals with artificial intelligence techniques. *Science Talks*, 10:100351, 2024.

[19] Martha Rodríguez, Diana P Tobón, and Danny Múnera. A framework for anomaly classification in industrial internet of things systems. *Internet of Things*, 29:101446, 2025.

[20] Kuldoshbay Avazov, Mukhriddin Mukhiddinov, Fazliddin Makhmudov, and Young Im Cho. Fire detection method in smart city environments using a deep-learning-based approach. *Electronics*, 11(1):73, 2021.

[21] Ashutosh Sharma, Rajeev Kumar, Isha Kansal, Renu Popli, Vikas Khullar, Jyoti Verma, and Sunil Kumar. Fire detection in urban areas using multimodal data and federated learning. *Fire*, 7(4):104, 2024.

[22] Hou, Yuanbo. *Advancing machine listening : understanding acoustic scenes and events and the emotions they evoke*. PhD thesis, Ghent University, 2024.

[23] Romas Vijeikis, Vidas Raudonis, and Gintaras Dervinis. Efficient violence detection in surveillance. *Sensors*, 22(6):2216, 2022.

[24] Reda Al-Batat, Anastassia Angelopoulou, Smera Premkumar, Jude Hemanth, and Epameinondas Kapetanios. An end-to-end automated license plate recognition system using yolo based vehicle and license plate detection with vehicle classification. *Sensors*, 22(23):9477, 2022.

[25] Linus Ng, Kenneth Ooi, and Gan Woon Seng. Urban sound tagging dcase 2019 challenge task 5. Technical report, Technical report, DCASE2019 Challenge (September 2019), 2019.

[26] Jisheng Bai, Jianfeng Chen, and Mou Wang. Multimodal urban sound tagging with spatiotemporal context. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):555–565, 2022.

[27] Pablo Negre, Ricardo S Alonso, Alfonso González-Briones, Javier Prieto, and Sara Rodríguez-González. Literature review of deep-learning-based detection of violence in video. *Sensors*, 24(12):4016, 2024.

[28] Muhammad Usama, Hafeez Anwar, and Saeed Anwar. Vehicle and license plate recognition with novel dataset for toll collection. *Pattern Analysis and Applications*, 28(2):57, 2025.

[29] Yue Ran, Hongying Tang, Baoqing Li, and Guohui Wang. Self-supervised video representation and temporally adaptive attention for audio-visual event localization. *Applied Sciences*, 12(24):12622, 2022.

[30] Tao Li, Zilin Bian, Haozhe Lei, Fan Zuo, Ya-Ting Yang, Quanyan Zhu, Zhenning Li, and Kaan Ozbay. Multi-level traffic-responsive tilt camera

surveillance through predictive correlated online learning. *Transportation Research Part C: Emerging Technologies*, 167:104804, 2024.

[31] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6420–6429, 2023.

[32] Zijian Hu, William HK Lam, SC Wong, Andy HF Chow, and Wei Ma. Turning traffic surveillance cameras into intelligent sensors for traffic density estimation. *Complex & Intelligent Systems*, 9(6):7171–7195, 2023.

[33] Youssef Elmir, Hayet Touati, and Ouassila Melizou. Intelligent video recording optimization using activity detection for surveillance systems. *arXiv preprint arXiv:2411.02632*, 2024.

[34] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Accurate and efficient two-stage gun detection in video. *arXiv preprint arXiv:2503.06317*, 2025.

[35] Tung Minh Tran, Doanh C Bui, Tam V Nguyen, and Khang Nguyen. Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[36] Zahidul Islam, Sujoy Paul, and Mrigank Rochan. Unsupervised video highlight detection by learning from audio and visual recurrence. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8702–8711. IEEE, 2025.

[37] Sam Leroux, Bo Li, and Pieter Simoens. Automated training of location-specific edge models for traffic counting. *Computers and Electrical Engineering*, 99:107763, 2022.

[38] Tetsu Matsukawa and Einoshin Suzuki. Convolutional feature transfer via camera-specific discriminative pooling for person re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8408–8415. IEEE, 2021.

[39] JunHa Hwang, SeungDong Lee, HaNeul Kim, and Young-Seob Jeong. Subset selection for domain adaptive pre-training of language model. *Scientific Reports*, 15(1):9539, 2025.

[40] Sani Abba, Ali Mohammed Bizi, Jeong-A Lee, Souley Bakouri, and Maria Liz Crespo. Real-time object detection, tracking, and monitoring framework for security surveillance systems. *Heliyon*, 10(15), 2024.

[41] Musrrat Ali, Lakshay Goyal, Chandra Mani Sharma, and Sanoj Kumar. Edge-computing-enabled abnormal activity recognition for visual surveillance. *Electronics*, 13(2):251, 2024.

[42] Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, and Wanli Ouyang. Exploiting visual context semantics for sound source localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5199–5208, 2023.

[43] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[44] Jiwei Zhang, Yi Yu, Suhua Tang, GuoJun Qi, Haiyuan Wu, and Hirotaka Hachiya. Enhancing semantic audio-visual representation learning with supervised multi-scale attention. *Pattern Analysis and Applications*, 28(2):40, 2025.

[45] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143. PMLR, 2021.

[46] Zixuan Hu, Xiaotong Li, Shixiang Tang, Jun Liu, Yichun Hu, and Ling-Yu Duan. Lead: Exploring logit space evolution for model selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28664–28673, 2024.

[47] Jiangbo Pei, Zhuqing Jiang, Aidong Men, Liang Chen, Yang Liu, and Qingchao Chen. Uncertainty-induced transferability representation for source-free unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 32:2033–2048, 2023.

[48] Ahmed Gomaa and Ahmad Abdalrazik. Novel deep learning domain adaptation approach for object detection using semi-self building dataset and modified yolov4. *World Electric Vehicle Journal*, 15(6): 255, 2024.

[49] Vinicius PM Goncalves, Lourival P Silva, Fatima LS Nunes, João E Ferreira, and Luciano V Araújo. Concept drift adaptation in video

surveillance: A systematic review. *Multimedia Tools and Applications*, 83(4):9997–10037, 2024.

[50] Chengye Zou, Yunong Liu, Yongwei Yang, Changjun Zhou, Yang Yu, and Yubao Shang. A privacy-preserving license plate encryption scheme based on an improved yolov8 image recognition algorithm. *Signal Processing*, 230:109811, 2025.

[51] European Union. Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance (data governance act). `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0868`, 2022. Accessed: 2025-05-07.

[52] Wei Lin, Junyu Gao, Qi Wang, and Xuelong Li. Learning to detect anomaly events in crowd scenes from synthetic data. *Neurocomputing*, 436:248–259, 2021.

[53] Thierry Malon, Geoffrey Roman-Jimenez, Patrice Guyot, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninou, Julien Pinquier, Florence Sèdes, and Christine Sénac. Toulouse campus surveillance dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views. In *Proceedings of the 9th ACM multimedia systems conference*, pages 393–398, 2018.

[54] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

[55] Adam Harvey. Exposing.ai, 2021. URL `https://exposing.ai`.

2

Real-World Acoustic Surveillance

In theory, theory and practice are the same.
In practice, they are not.

- Yogi Berra

Real-World Acoustic Surveillance: An Empirical Evaluation of Anomaly Detection, Sound Tagging, and System Limitations

Wei-Cheng Wang • Sam Leroux • Pieter Simoens

The previous chapter discussed numerous challenges in smart surveillance, stemming from both the characteristics of surveillance data and the operational demands of real-world deployment. While the paradigm of visual surveillance is widely-studied, the limitations and challenges in acoustic systems in real-world urban environments are less established. By focusing solely on audio data collected from unconstrained urban settings, we can isolate and quantify the effects of context drift and other variables without the confounding factor of visual data, for which domain shift has been more extensively studied. Building upon the issues raised in Chapter 1, this chapter aims to empirically evaluate the impact of these factors, particularly the high volume of data, context drift, signal complexity, and the possible challenges of a conceptual hybrid acoustic edge-cloud framework. Establishing an empirical experiment for these performance gaps is a necessary first step toward developing the adaptable and context-aware surveillance frameworks proposed later in this dissertation.

2.1 Introduction

Real-world acoustic surveillance models are significantly impacted by challenges such as context drift and the limitations of closed-set event taxonomies. To investigate these performance gaps, this chapter utilizes unannotated urban acoustic data from the imec SenseCity project, collected from microphones deployed in Ghent and Rotterdam. These audio segments were sourced from sensors deployed at different locations and were unprocessed prior to this study. With this data, this chapter empirically evaluates two crucial smart surveillance applications. First, for anomaly detection, we investigate the performance of models trained in one context when applied to another, reflecting common deployment strategies. Second, for acoustic event tagging, or sound tagging, we examine the effectiveness of deploying models pretrained on a large-scale, publicly available dataset (which was collected under different scenarios) when applied to the audio data from the diverse urban environments investigated in this study. In both scenarios, the results reveal the need for location- and context-dependent anomaly de-

tection models with real-world data collected in Ghent and Rotterdam. Furthermore, the exploration of sound tagging demonstrates the robustness of pre-trained models but highlights limitations due to predefined taxonomies.

Thus, this study considers how such identified limitations (e.g., data volume from tagging, context-dependency of anomaly detection) would impose critical constraints on, or reveal inefficiencies in, different deployment architectures. A conceptual, hybrid edge-cloud framework combining anomaly detection and sound tagging is analyzed to quantify the potential efficiencies of an integrated system and to highlight the scalability challenges that such practical deployments face. The findings collectively underscore the necessity of developing lightweight, adaptive learning techniques to enhance the reliability and efficiency of surveillance systems in dynamic urban environments.

This chapter is organized as follows. Section 5.3.3 details the urban acoustic datasets from Ghent and Rotterdam used in this study, highlighting the real-world data characteristics that motivate this chapter's investigation into deployment limitations. Subsequently, Section 5.4.2 presents a thorough evaluation of anomaly detection techniques, while Section 2.4 offers a similar empirical analysis of acoustic event tagging, both focusing on performance limitations and contextual dependencies in real-world scenarios. Building upon these findings, Section 2.5 further explores operational limitations and constraints by considering a conceptual tiered surveillance system. Finally, Section 2.6 summarizes the key limitations identified throughout the chapter and outlines directions for future work motivated by these empirical results.

2.2 Data Collection

2.2.1 Deployment and Scope

This study utilized urban audio data collected by AsaSense[1], a partner company specializing in urban acoustic sensing solutions. To capture diverse soundscapes, microphones were deployed across various settings in Ghent and Rotterdam. These sensors recorded a wide range of acoustic signals, such as human activity, vehicle noise, construction work, and animal sounds, over a period of 10 months in Ghent and 2 years in Rotterdam. The detailed parameters of the data collection across the two cities are summarized in Table 2.1.

[1] https://asasense.com/

City	Collection Period	# of Sensors	Sampling Rate (Hz)
Ghent	Dec. 2018 - Sep. 2019	1	48000
Rotterdam	Dec. 2015 - Jan. 2017	4	24000

Table 2.1: Data collection details, including the location, number of sensors and difference in data type.

2.2.2 Privacy Considerations

During the collection of urban acoustic data, the data provider incorporated a strategy aimed at mitigating privacy concerns commonly associated with continuous audio recording in public spaces. Seeking to reduce risk of recording continuous speech that potentially contains private information, the data collection protocol involved recording 60-second audio segments at 10-minute intervals. While this intermittent sampling method is a common technique intended to limit the capture of extended private conversations, it does not entirely eliminate the risk of recording acoustic biometric information.

2.2.3 Dataset Characteristics

The extended duration of the data collection is critical for this study, as it captured significant seasonal and weather-induced variations. The resulting dataset thereby exhibits temporal drift and contextual shifts critical for rigorously testing the robustness of surveillance algorithms, as discussed in Section 1.3.1.2. This dataset therefore provides a robust empirical foundation for the subsequent investigation of anomaly detection and sound tagging in this chapter.

2.3 Urban Anomaly Detection: Sensitivity to Spatiotemporal Changes

2.3.1 Problem Statement

Anomaly detection in urban acoustic surveillance aims to identify audio segments with unusual patterns, such as glass breaking or distressed shouting, which may indicate incidents requiring further response. As discussed in Chapter 1, this task is vital for public safety, yet it is particularly challenging due to the infinite variety of potential anomalies and their subjective or context-dependent nature. A sound that is normal in one environment, such as traffic noise near a busy road, may be considered anomalous in another,

like a quiet park.

Given these complexities, this section empirically investigates the performance of contemporary anomaly detection approaches when applied to real-world urban audio. To systematically investigate the limitations of existing deep learning approaches when deployed in real-world scenarios, particularly concerning their robustness to context drift, a widely adopted analytical approach is to assess their performance when models trained with data reflecting a particular environment or time frame are then applied to data from distinct environments or time frames. Therefore, we specifically evaluate model robustness by comparing detection results across these diverse temporal and spatial contexts.

2.3.2 Experimental Setup

Due to the rarity of true anomalous events and the expensive cost of manual annotation in real-world surveillance data, our experiments rely on an unsupervised approach. The anomaly detection experiments use a deep autoregressive network [1], adapted from a WaveNet [2] architecture, comprising a total of 12.22M trainable parameters. During training, this model aims to predict future samples in an audio sequence, thereby learning a conditional distribution from the normal data. In the inference phase, the mean-squared error (MSE) between the predicted and actual waveform is utilized to estimate a sample's abnormality. To determine whether a sample is anomalous, a threshold was established using the empirical rule. Specifically, the mean (μ) and standard deviation (σ) of the MSE were calculated over the respective training data for each model version. Subsequently, based on the assumption that 95% of the training samples represent normal activity, the anomaly threshold was set at $\mu + 2\sigma$.

To investigate model performance across different conditions and further probe the limitations identified in Section 1.3.1.2, distinct sets of experiments were designed to evaluate robustness to temporal and spatial context drift:

1. First, to evaluate temporal context effects, three versions of the autoregressive models were trained using data collected by a single microphone in Rotterdam. Each version was trained on data from one of the different dates: April 30th, August 30th, and December 30th. These three models were then tested on audio data collected from this same microphone across various dates, including those used for their training and other unseen periods.

2. Subsequently, attention turned to the spatial context evaluation. Specifically, we trained four separate models on data collected by different sensors. Each model was trained using data collected from one of four distinct sensor locations in Rotterdam, using all available data from December for each respective sensor. These four sensor-specific models were then applied to same audio data, which is collected on December 30th in Rotterdam.

2.3.3 Experimental Results: Temporal Drift

The anomaly scores from the models, when applied to test data from various dates, are shown in Figure 2.1. In this figure, each of the four main sections (a-d) corresponds to a different test date. Within each section, the three plots show the anomaly scores from the three models trained on data from April 30th (top), August 30th (middle), and December 30th (bottom), respectively. The red fragments in each plot indicate events manually verified as actual anomalies, while the blue fragments represent segments confirmed as normal. Peaks in the plotted scores, whether occurring over blue or red regions, indicate that the trained model has flagged that particular audio fragment as anomalous due to a high anomaly score.

In this anomaly detection framework, a *positive* means that an event is flagged by the model as an anomaly, while a *negative* indicates it is considered normal by the model. Human annotations provide the ground truth: red fragments represent actual anomalies (ground truth positive), and blue fragments represent normal events (ground truth negative). Consequently, a false positive occurs when the model incorrectly flags a normal event as an anomaly (a model peak in a blue region). Conversely, a false negative occurs when the model fails to detect a true anomalous event (no significant model peak in a red region)

Several critical insights can be concluded from these results. First, false positives, represented by peaks over blue-colored (normal) segments, are numerous and scattered throughout the dates. These segments often correspond to normal events that are either particularly loud or were rare in that specific model's training data, rather than events requiring a corresponding response. Second, the overall performance of the three models (trained on April, August, and December data respectively) is largely similar when applied to most test dates. A notable exception occurs in the results for Figure 2.1 (d)(September 24th), where the model trained on April data performs differently compared to those trained on August and December data. Specifically, the model trained on April data produces an anomaly score profile that is more aligned with human annotations for this particular day.

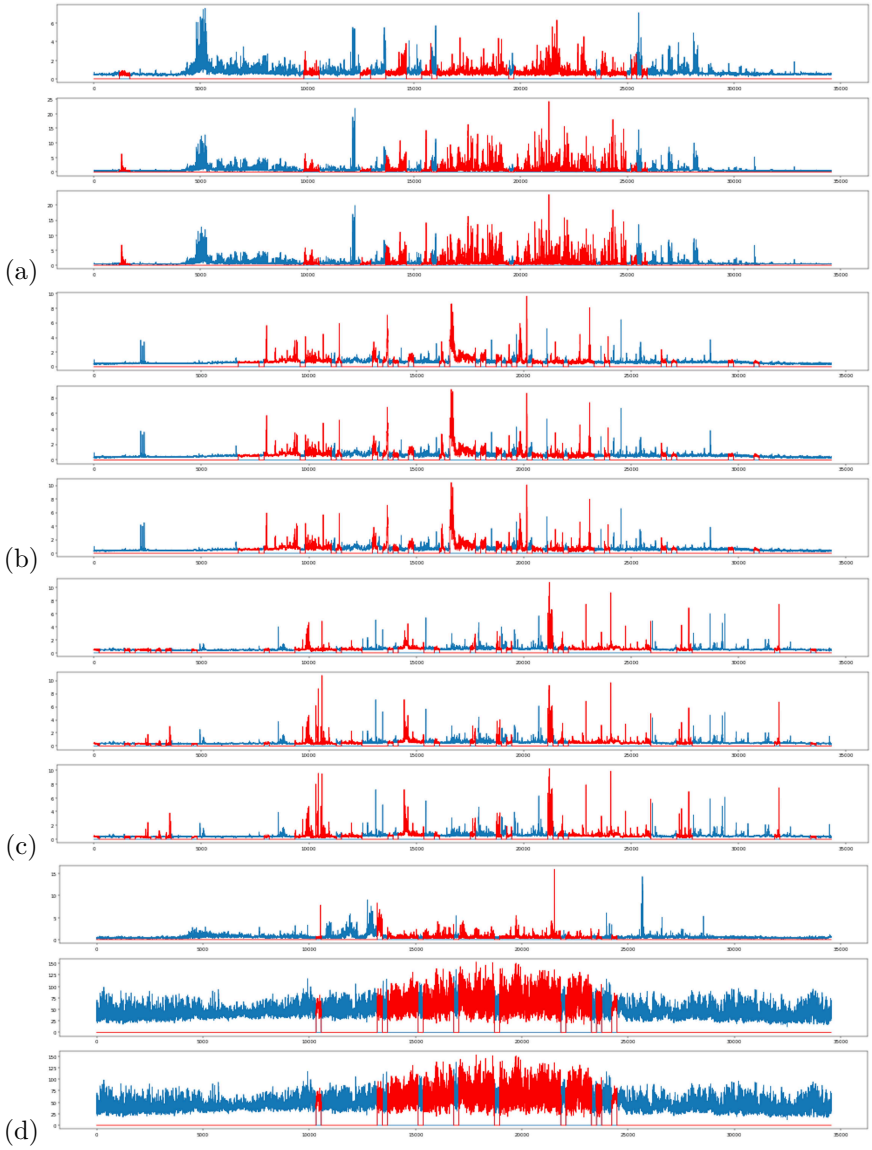


Figure 2.1: Anomaly detection performance of models trained on audio data from different dates in Rotterdam. Each row corresponds to anomaly scores for a test set of audio collected on (a) April 30, (b) August 30, (c) December 30, and (d) September 24. Models were trained on data from April 30 (top), August 30 (middle), and December 30 (bottom). Red vertical lines represent manually annotated anomalies.

Investigation into the audio data revealed that both April 30th (training data for one model) and September 24th (test data) were windy days. The similarity in these environmental conditions likely led the model trained on April to correctly classify wind sounds as normal. Meanwhile, the other two models, not having been trained on similarly windy conditions, flagged these wind sounds as anomalies. This finding highlights how temporal context shifts can significantly impact anomaly detection performance, potentially leading to an increase in false positives and the associated human resource costs for manual verification.

2.3.4 Experimental Results: Spatial Drift

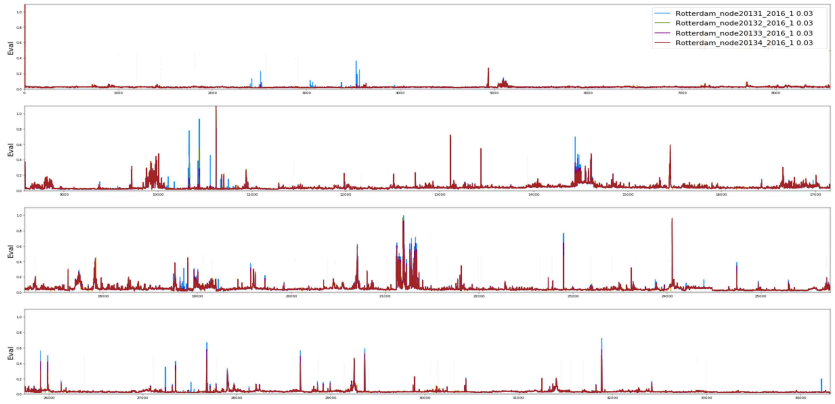


Figure 2.2: Anomaly detection performance of models trained on audio data from different sensors in Rotterdam. Each row shows the anomaly scores for audio collected on December 30, with models trained using data from sensor 1 to 4, respectively.

Further experiments were conducted to compare models trained on data collected by different sensors, aiming to assess performance across spatial contexts. Four distinct versions of the autoregressive model were trained, each using data from one of four unique sensor locations in Rotterdam (Sensor 1-4), specifically utilizing all available data from December for each respective sensor. These four sensor-specific models were then applied to the same test data: audio recordings collected on December 30th in Rotterdam.

Figure 2.2 visualizes the anomaly scores from the models trained on data collected from four different sensors. For ease of visualization, the timeline of the December 30th data is segmented and presented across four rows. Within each row, the anomaly scores from the four models trained on different sensor data are overlaid, each represented by a distinct colored line

(e.g., light blue for the model trained on Sensor 1, and so on). This allows for a direct comparison of how models trained in different spatial contexts respond to identical audio input.

As illustrated in Figure 2.2, the models exhibit varied responses to the same data. For example, the model trained on Sensor 3 (represented by the dark-red line) flagged certain events that were not detected as anomalous by the model trained on Sensor 1 (light-blue line) when analyzing the identical audio segments. Investigation suggests these differences are primarily caused by the distinct background acoustic contexts captured at each sensor's original training location. A sensor situated near a busy street, for instance, would learn to treat traffic sounds (e.g., engine noise, occasional horns) as normal, whereas a sensor in a park would build its baseline of normality based on different acoustic elements, such as bird singing or distant human speech. Consequently, when these models, each being trained to fit a specific spatial context, are deployed to analyze audio from a new or different environment, their learned definitions of *normal* cause them to flag different events as anomalous, underscoring the challenge of spatial context shift.

2.3.5 Discussion

From the experiments detailed above, we draw several key conclusions of applying anomaly detection to real-world urban audio data. A primary finding is that domain shifts significantly compromise the performance of models pretrained on data collected under different scenarios. This finding highlights the need for developing context- and location-specific models for anomaly detection, as data drift can be caused by both temporal variations and spatial differences between sensors. Blindly applying pretrained models to new target data, even if initial results appear reasonable, may lead to critical issues, particularly the oversight of false negatives. Such false negatives represent a critical failure in model generalization. For instance, a model trained at a busy intersection may learn to treat loud horns as normal background events. If this model were then deployed in a quiet residential area, it might fail to flag a horn as the urgent indicator of danger it represents in that new context, resulting in a critical false negative. Similarly, the false positives, normal events incorrectly flagged as anomalous, can overload the system by consuming resources for the investigation of common, non-critical occurrences.

To address this challenge of context dependency, one potential strategy is to train a universal model using data collected from a diverse range of sources. Although this may be seen as the most straightforward approach, it is rather impractical due to two primary factors: the complexity of the data involved

and constraints on data sharing. Data complexity implies that developing a model capable of learning representative and distinguishable features from highly varied inputs necessitates substantial model capabilities and immense data requirements. Data sharing, as indicated in Section 1.3.2.1, is often restricted due to privacy regulations or proprietary concerns.

An alternative approach is to adapt pretrained models to the specific target data, typically through finetuning. An efficient and effective adaptation process could involve two key phases: first, selecting the most suitable pretrained model from available options (e.g., a model zoo), and second, finetuning this selected model for the target context. The selection phase aims to identify the pretrained model from a collection that is estimated to perform best on the specific target data. The finetuning phase, on the other hand, involves finetuning the selected pretrained model to optimize its performance on the target data. While either model selection or finetuning can be used independently, employing both mechanisms in conjunction may be most effective; model selection can provide a better starting point, potentially mitigating severe performance degradation from context mismatch, while finetuning, although computationally intensive, allows for more precise adaptation.

2.4 Cross-Environment Sound Tagging: Robustness and Limitations

2.4.1 Problem Statement

Sound tagging aims to identify whether specific events occur within the audio segments. Unlike anomaly detection, training sound tagging models typically requires intensive data annotation, which is resource-consuming. A common strategy to mitigate this annotation burden is to train sound tagging models on large, often publicly available, annotated datasets. However, the acoustic characteristics and distribution of events in these general pretraining datasets may not fully correspond to those encountered in specific operational environments, such as the unique urban soundscapes from which our data was collected. Therefore, in this section, we aim to explore the effectiveness and limitations of applying a sound tagging model, pretrained on a general dataset, to our specifically collected urban acoustic data. This investigation will assess its robustness and identify potential challenges in such cross-environment deployment scenarios.

2.4.2 Experimental Setup

To evaluate the robustness and limitations of pretrained sound tagging models, this study employed a model originally trained on the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge Task 5 dataset [3]. This pretraining dataset originated from the SONYC project [4], which contains urban sounds recorded in New York City that were annotated with 8 coarse-level classes and 29 fine-level tags. The pretrained model selected for this evaluation was the winning entry of this challenge [5], which is a modified MobileNetV2 architecture [6] with 2.90M trainable parameters. The pretrained model takes a Log Mel-spectrogram as input, which is transformed from a 10-second audio with window length of 2560 and a hop length of 694 samples.

This pretrained model was then applied to the urban acoustic data collected for this study from Ghent and Rotterdam. Audio segments were assigned a specific event tag if the model's prediction score for that event exceeded a threshold of 0.5. Note that, each segment could be assigned multiple event tags, indicating the occurrence of more than one event. Crucially, the urban acoustic data from Ghent and Rotterdam were not annotated with ground-truth sound event labels, and an exhaustive manual re-examination of all segments was infeasible. Therefore, to explore the pretrained sound tagging model's effectiveness and limitations in these new environments, the outputs of the sound tagging model were compared against the anomaly detection results. This approach involved a qualitative in-depth analysis primarily targeting those audio segments that were either, assigned one or more event tags by the pretrained sound tagging model, or flagged as anomalous by the anomaly detection system. Such analytical approach was designed to provide deeper insights into the sound tagging model's behavior, its applicability to the new urban soundscapes, and its specific limitations when encountering out-of-distribution data. This analysis was conducted across audio data from four different dates (April, August, November, and December) to observe performance across varied temporal conditions.

2.4.3 Experimental Results

To evaluate whether the sound tagging model pretrained on New York City data is effective when applied to data collected in Rotterdam, we first compared its sound tagging outputs against the anomaly detection results. Figure 2.3 visualizes these comparisons, showing the sound tagging results (e.g., orange lines indicating segments where the model detected at least one acoustic event) alongside the anomaly scores from two different anomaly detection models (e.g., green and blue lines for models trained on April and

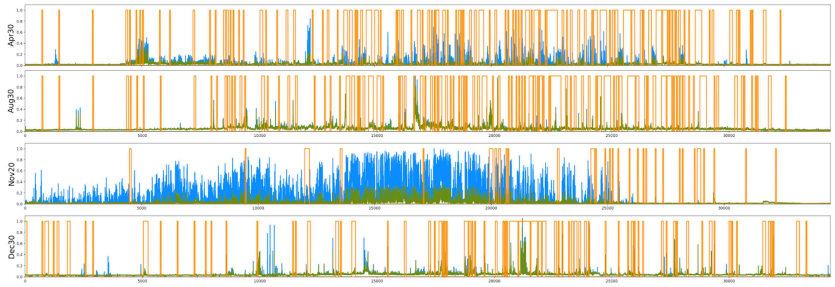


Figure 2.3: Comparison of sound tagging results and anomaly scores. Blue and green lines represent anomaly scores calculated by autoregressive models trained on April and December data, respectively. Yellow lines indicate tagged acoustic events along with their starting and ending points.

December data, respectively).

The sound tagging results on April 30, August 30, and December 30 show that, the sound tagging model tagged over 70 segments with different events daily. Further investigation has reveal several patterns. First, segments that were both tagged by the sound model and flagged as anomalous by the anomaly detector typically corresponded to distinct, predefined events from the DCASE challenge taxonomy, such as *car horn* or *shouting*. Second, vast amount of segments were tagged by the sound model but were not flagged as anomalous. These often involved common urban acoustic events like *people talking* or *engine sounds*, which the locally trained anomaly detection models considered as normal events. This suggests the pretrained sound tagging model can identify some common events correctly, even if they are not anomalous locally. Conversely, many segments flagged as anomalous by the local detector were not assigned any tags by the pretrained sound tagging model. These frequently contained acoustic events absent from the DCASE dataset's taxonomy, such as *bell ringing* or *the sound of something breaking*. In such cases, the pretrained event detector failed to tag these unfamiliar events. Finally, further investigation revealed that some untagged anomalous events (or even some tagged events) are sometimes misclassified. For instance, *bird singing* might be erroneously classified as *Alarm*, or *fireworks* as *dog barking*, highlighting limitations due to the closed-set setting of the sound tagging model.

In contrast, data from Nov. 20th showed relatively few segments tagged by the sound model, despite high anomaly scores, particularly from the anomaly detection model trained on December data. This difference was caused by the windy weather conditions on Nov. 20th, which resulted in

fewer distinct, classifiable urban events but introduced significant environmental noise (e.g., rain, wind). The anomaly model that are not trained on such stormy conditions, flagged these periods as anomalous. Concurrently, the pretrained sound tagging model failed to identify acoustic events like heavy rain, strong wind, or even the *birds chatting* because these were not part of its predefined DCASE training taxonomy.

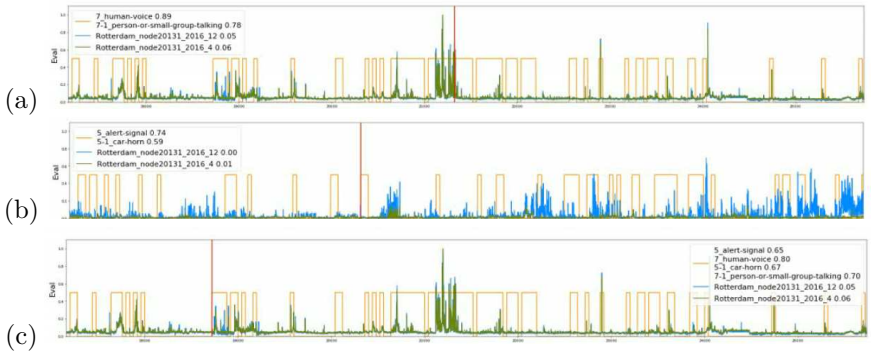


Figure 2.4: Case studies from December 30th data illustrating the performance and limitations of the pretrained sound tagging model when applied to Rotterdam urban audio, shown with corresponding anomaly score context. Subfigures show cases including: (a) an anomalous event tagged with *person-or-small-group-talking*; (b) a non-anomalous segment yet tagged with *alert* and *car-horn*; and (c) an less anomalous event that is tagged with *alert-signal* and *car-horn*.

To further investigate the results of sound tagging and anomaly detection, Figure 2.4 illustrates three cases that occurred on December 30. Figure 2.4 (a) illustrates a segment containing *people shouting*, which the pretrained model tagged with the more general label *person-or-small-group-talking*; this segment was also flagged as anomalous by the local anomaly detector, likely due to its sudden increase in volume. Meanwhile, in Figure 2.4 (b), two tags, *alert* and *car-horn*, were assigned by the sound tagging model. This segment was also not flagged by the anomaly detection, as it actually is a segment containing bird singing (distinct from people chatting). Figure 2.4 (c) is another example where the sound tagging model could fail when the acoustic event is not included in the training taxonomy. This segment described the sounds of *bicycle bell ringing* and *people shouting*, potentially an argument between biker and pedestrian. However, the sound tagging model classified this segment, which contained both non-defined *bicycle bell ringing* and pre-defined *people shouting*, with the tags *alert-signal*, *human-voice*, and *car-horn*. While the *human-voice* tag is consistent with

the *people shouting* component of the event, the non-defined *bicycle bell ringing* likely contributed to the model's misapplication of *alert-signal* and *car-horn* from its known categories. These illustrative cases present the limitations of relying solely on a pretrained sound tagging model with a fixed taxonomy in an unseen acoustic environment and highlight the complementary role that a context-specific anomaly detection model can play in identifying potentially significant events.

2.4.4 Discussion

By cross-validating the tagged events and flagged anomalies, we draw several key conclusions regarding the application of pretrained sound tagging models to real-world urban data.

First, despite the domain shift between the data used to train sound tagging model and the data that was tested, the sound tagging model in general showed some capability in identifying certain events that were part of its original taxonomy. However, this evaluation also highlighted the inherent limitations of a closed-set training approach, which can lead to the model misclassify unknown or out-of-taxonomy events (either by assigning an incorrect label or by failing to tag them altogether). Furthermore, the model failed to identify significant environmental sounds, such as heavy wind or rain, when these were not part of its original DCASE training taxonomy, even if such sounds were prominent enough to be flagged by anomaly detection systems. While this problem can be addressed by updating the detection model, such retraining is resource consuming in terms of both data collection and computation.

Second, the volume of tagged events could sometimes be overwhelming, particularly in acoustically complex urban scenarios. As the majority of these tagged events are normal in such scenarios, this could further overload the decision phase. One possible solution would be to integrate the sound tagging model with a context-specific anomaly detection model. The comparison on the sound tagging model and anomaly detection model in Figure 2.4 shows that the tagged/flagged segments are not always directly aligned, indicating the possibility of using one model to complement the other. For instance, anomaly detection can highlight unusual events missed by the sound tagging model's fixed taxonomy. Furthermore, the anomaly detection system's assessment of acoustic anomaly events offers a valuable method to filter or prioritize the sound tagging model's broad outputs, thereby mitigating information overload from numerous tags on contextually routine sounds.

2.5 A Conceptual Scalable Smart Surveillance System

To empirically investigate the system-level challenges, operational characteristics, and computational demands of acoustic surveillance applications, a conceptual edge-cloud hybrid system incorporating both edge and cloud processing stages was proposed and implemented. This system was designed to facilitate an investigation into how tasks could be distributed and how such distribution would impact data processing and resource utilization. Furthermore, such system help us to examine data flow, the interplay between different analytical modules (anomaly detection and event tagging), and associated computational loads under real-world conditions.

2.5.1 Conceptual System Architecture

In the conceptual model illustrated in Figure 3.2, edge devices were designed to perform initial, location-specific acoustic analysis. For the experiments conducted, this involved deploying location-specific anomaly detection models. These models were intended to be trained in a self-supervised manner, to identify acoustic patterns deviating from the learned normal pattern for that specific environment. The primary investigative purpose of simulating this edge layer was to assess the impact of early-stage filtering, based on the observation obtained in Section 2.4, where the anomaly detection model can effectively navigate the focus of analysis to salient events. Segments flagged as potentially anomalous by the edge module were designated for transmission to the cloud infrastructure. Meanwhile, the cloud server within this system was designed to handle more computationally intensive post-analysis of the segments forwarded from the edge. This tiered setup was intended to allow for the evaluation of sophisticated, potentially resource-heavy event detection algorithms on a pre-filtered dataset, and to study the feasibility of updating or expanding these cloud-based detection capabilities. For this investigation, a trained event detector was implemented in the cloud to further categorize the received audio segments.

2.5.2 Experimental Setup

To conduct this conceptual system for exploring its characteristics and limitations, we apply the urban acoustic data collected in Rotterdam. For the anomaly detection module, we deploy the model previously trained in Section 5.4.2. Concurrently, the event detection module in the cloud was the same as we used in Section 2.4.

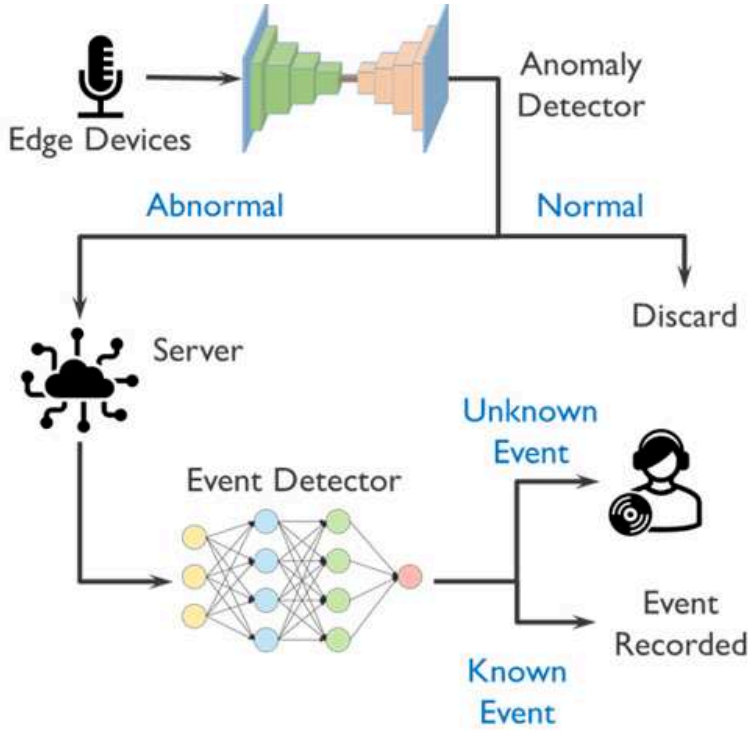


Figure 2.5: Proposed hybrid edge-cloud framework for urban acoustic surveillance systems.

The primary evaluations focused on two aspects critical to understanding system-level limitations. First, the effectiveness of the edge filtering stage was quantified by cross-validating the segments flagged by anomaly detection against those subsequently identified by the cloud-based event tagging model. This allowed for an assessment of the percentage of data potentially filtered out before requiring more intensive cloud processing and an analysis of what types of events were prioritized or potentially missed. Second, to investigate practical hardware constraints, the computational workload of both the anomaly detection and sound tagging algorithms was measured. This involved timing their processing performance on diverse hardware platforms, from server (Tesla V100-SXM3), laptop (GeForce GTX 980), to CPU (Intel CPU@2.7/2.4 GHz). Such investigation expose potential bottlenecks and assess the feasibility of deploying such modules, particularly in resource-constrained edge environments.

2.5.3 Experimental Results

We obtained several key observations regarding data flow, model interactions, and computational demands, which helped to identify inherent system limitations. The initial experimental observations involved a qualitative comparison between detected anomalies from the edge and acoustic events tagged by the cloud, using data collected in Rotterdam. Table 2.2 categorizes the processed segments: (a) those flagged as anomalous and subsequently tagged with a known event, (b) those flagged as anomalous but not tagged, and (c) those not initially flagged as anomalous but still tagged by the event tagging model. The distribution detailed in this table indicated that utilizing a tailored anomaly detector at the edge could substantially reduce the data volume forwarded for further analysis. In this experimental setup, the reduction of the workload has reached approximately 90% for the human operator. This observed reduction was a key quantitative result, which then allowed for a more focused examination of the remaining data, bringing certain system limitations and inter-dependencies under particular review.

The analysis of specific challenging scenarios, such as data from Nov.20th processed using an anomaly detection model trained on December data, further highlights the critical need for both context-specific anomaly detection models and comprehensive sound tagging capabilities. On such days with abnormal weather, the less-adapted anomaly detection model triggered a high volume of events. Concurrently, because wind and rainfall were not listed as acoustic events within the sound tagging model's predefined category, it failed to provide semantic labels for these sounds. Consequently,

Train	Apr30				Dec30			
Test	Apr30	Aug30	Nov20	Dec30	Apr30	Aug30	Nov20	Dec30
(a)	0	2	0	12	7	3	5	14
(b)	0	8	0	2	17	10	253	7
(c)	213	209	49	140	206	208	44	138

Table 2.2: Qualitative comparison between anomalies found and events tagged. (a) Number of segments labelled as anomalies and flagged by the sound tagging. (b) Number of segments labelled as anomalies but not flagged by the sound tagging. (c) Number of segments not labelled as anomalies but flagged by the sound tagging.

	Tesla V100-SXM3	GeForce GTX 980	Intel CPU @ 2,70 GHz	2,40 GHz
Anomaly Detection	0.104	0.372	8.164	15.482
Sound Tagging	0.004	0.008	0.078	0.072

Table 2.3: Computation time (second) to process 1-sec long audio.

these numerous, unlabeled yet flagged segments would be transmitted to a human operator for interpretation, highlighting a significant system limitation: the pipeline's overall effectiveness in reducing operator burden heavily relies on the robustness and adaptability of both its anomaly detection and event tagging components to abnormal environmental conditions or non-predefined events. While one might hypothesize that if the sound tagging model could identify these environmental sounds, a complementary knowledge base might classify them as contextually normal (thus documenting them without human interference), this particular observation pointed to a clear gap in the system's ability to automatically handle such uncatalogued events. Similar dependencies and challenges in model adaptation were noted when analyzing behaviors over extended data periods, such as a full year.

Further investigation focused on the computational load of the acoustic applications to understand practical deployment constraints, with processing times detailed in Table 2.3. To simulate varying computational resources, performance was measured on server-grade hardware (Tesla V100 GPU) and consumer-grade hardware (GeForce GTX 980 GPU), representing potential cloud capabilities, alongside more resource-constrained CPUs (Intel CPU@2.7/2.4 GHz), indicative of edge device limitations. While GPUs were capable of processing both the anomaly detection and sound tagging tasks in (or near) real-time, a critical bottleneck was observed with the CPUs:

they failed to process the computationally demanding anomaly detection algorithm in real-time. This finding directly highlights a significant practical challenge for deploying such anomaly detection methods effectively on typical resource-limited edge devices.

2.5.4 Discussion

While the hybrid edge-cloud framework permitted an empirical examination of distributed acoustic analysis and data flow, its primary utility within the context of this chapter was to reveal several significant challenges and practical limitations that remain to be addressed.

First, the computational load assessment highlighted that designing a lightweight anomaly detection model capable of real-time computation on resource-constrained edge devices remains a major and crucial direction. This difficulty includes not only the neural network architecture itself but also the training strategies required to achieve both efficiency and robust performance on such hardware. While beyond the scope of the experiments conducted here, potential avenues to mitigate the computational weight of anomaly detection models include techniques such as model distillation and quantization [7].

A second critical challenge is about data privacy, as described in Section 2.2.2. Given that the audio segments transmitted from the edge to the cloud for further analysis are raw audio data, the risk of privacy-sensitive information leakage cannot be disregarded. It is important to note that audio surveillance is often subject to even stricter legislative and ethical considerations than video surveillance; as the potential for audio surveillance to capture private conversations and allow for eavesdropping means it is very strictly regulated. Addressing this privacy risk robustly is a crucial consideration for any practical deployment of such a hybrid system.

2.6 Conclusion and Future Work

In this chapter, we investigated the key aspects of deploying acoustic surveillance systems in urban environments, focusing on the challenges posed by context shifts, resource constraints, and emerging privacy considerations. State-of-the-art audio analysis techniques, specifically anomaly detection and sound tagging, were empirically evaluated using a real-world dataset collected from Ghent and Rotterdam. Our experiments highlighted that anomaly detection models are inherently location- and context-specific. Concurrently, while sound tagging models pre-trained on large datasets ex-

hibit a degree of robustness, their utility is constrained by predefined event taxonomies and their performance in a new acoustic environments.

Our findings collectively highlight several limitations of current approaches when applied to real-world settings. Anomaly detection models often remain computationally intensive, posing challenges for deployment on resource-constrained edge devices. Training or fine-tuning location-specific anomaly detectors typically requires access to local data, which can raise data governance and privacy concerns if that data needs to be moved or accessed by centralized systems. The static nature of many current models struggles with dynamic environmental changes, indicating that mechanisms for regular updates or adaptation of anomaly detectors (e.g., to cope with seasonal variations like bird singing) should also be considered. Moreover, even with effective anomaly detection filtering location-specific normal soundscapes, the overall usability of a surveillance pipeline heavily depends on the accuracy and comprehensiveness of the subsequent sound tagging model. The need for robust adaptation in anomaly detectors would, for instance, be somewhat mitigated if more advanced sound event detection algorithms become broadly available. Additionally, the lack of ground-truth annotations for true anomalous events in real-world datasets means that evaluating the false negative rates of anomaly detection systems remains an open research question.

As part of this investigation, a hybrid edge-cloud framework was utilized as an conceptual system to explore the system-level implications of integrating self-supervised anomaly detection on edge devices with cloud-based sound tagging for post-analysis. It was observed that edge-based anomaly detection could substantially reduce the volume of data transmitted for cloud processing-by approximately 90% in our experimental setup. However, this observation also highlighted the critical challenge of ensuring that such filtering does not accidentally discard crucial information. A more critical issue is to reduce the computational cost for real-time anomaly detection on edge hardware.

To address these identified challenges, several research directions for future work should be explored. First, to enhance the practicability of anomaly detection models for edge deployment, further research into techniques such as model quantization and knowledge distillation is warranted to develop more lightweight and efficient models. Second, exploring multi-modal or federated learning approaches could enable the collaborative adaptation of models across distributed sensors while better preserving data privacy. Third, for practical deployment in scenarios lacking extensive annotations, the development of robust transferability measures for model selection in

source-free, unsupervised domain adaptation settings is essential. Pursuing these research directions leads to the development of a more robust, efficient, and ethically considerate acoustic surveillance systems capable of adapting to the diverse and dynamic environments of smart cities.

2.7 References

[1] Ellen Rushe and Brian Mac Namee. Anomaly detection in raw audio using deep autoregressive networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3597–3601. IEEE, 2019.

[2] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12, 2016.

[3] Mark Cartwright et al. Dcase 2019 challenge task 5: Urban sound tagging. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge*, 2019. URL `https://dcase.community/challenge2019/task-urban-sound-tagging-results`.

[4] Juan P. Bello, Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62(2):68–77, Feb 2019. doi: 10.1145/3224204.

[5] Sainath Adapa. Urban sound tagging using convolutional neural networks, 2019. URL `https://arxiv.org/abs/1909.12699`.

[6] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[7] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018.

3

Opt-in Framework for Privacy Protection in Audio

FBI! Open up!

—*Meme*

An Opt-in Framework for Privacy Protection in Audio-based Applications

Wei-Cheng Wang • Sander De Coninck • Sam Leroux • Pieter Simoens
Published in IEEE Pervasive Computing 2022

In the previous chapter, one of the findings was that, despite the data collection protocol deployed to reduce privacy leakage, the recorded human chatting audio confirms the privacy concerns raised in Chapter 1. Although the captured voice fragments may not be sufficiently long to identify the overall context of the conversation, the voice print within a single sentence can already reveal the identity, gender and other biometrics of the speaker. To address the risk of sensitive data being collected and used without authorization, this chapter proposes a user-centric, privacy protection framework that is scalable for deployment on smart surveillance system. The framework introduces an opt-in mechanism, allowing users to authorize the exact information they share. Furthermore, it is designed for compatibility with existing, pretrained models in an existing surveillance system. By developing such privacy protection framework, we provide a practical solution that consider all aspects from performance, privacy, to deployment. Notably, in this chapter, the framework's practicability is empirically evaluated using datasets from domestic acoustic environments. This is due to the unavailability of urban acoustic datasets containing labels for human biometrics such as gender, age and ethnicity.

3.1 Introduction

Audio data is used in an increasing number of pervasive IoT applications that are deployed in our private space. Microphones in our houses and smartphones have been proposed for acoustic event detection in ambient assisted living [1], speech processing by smart speakers of voice commands [2], or cough detection in telemedicine [3].

This rich palette of applications is realized as processing algorithms on the same data stream. All applications request direct access to the microphone, but raw audio contains more data than strictly needed to perform the task. This problem of data bundling [4, 5] opens the door for the audio being used for other purposes than the one originally agreed upon. An acoustic system for fall detection of an older person might also reveal if other persons are present. From voice commands targeted to a smart speaker, many sensitive

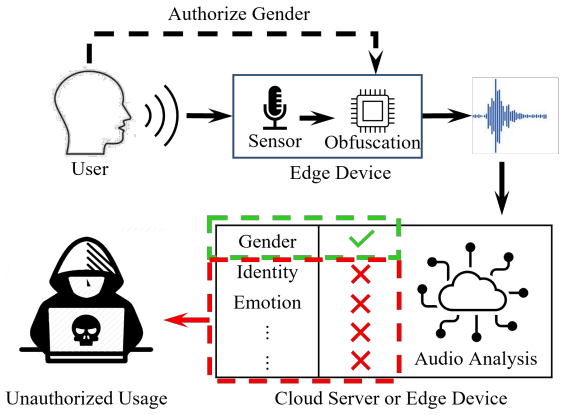


Figure 3.1: In our proposed system, users authorize one target task. Raw audio is obfuscated on the device, resulting in a transformed audio stream that is transmitted to the target task also on the edge device or in the cloud.

attributes can be derived about the user that go beyond the content of the spoken words, such as speaker identity, emotion, gender or ethnicity. Patent filings indeed show that companies consider these options as valuable sources of information for targeted advertising [6].

Information extraction from audio is typically realized by state-of-the-art deep neural networks (DNNs) with millions of parameters, such as CycleGAN-VC2 [7]. High-end edge devices, such as smartphones, have the necessary substantial computational resources to evaluate machine learning models of sound and speech applications locally. Although no raw audio data is transmitted to a cloud back-end, the risk of data misuse remains because the machine learning models are typically integrated in third-party apps. Operating systems such as Android or iOS require apps to ask permission to use the microphone, but once this permission is granted, there is no way for the user to restrict the type of information that the app can extract from the raw data. In applications like audio-based surveillance in smart cities or nursing homes, up to hundreds of edge devices need to be installed and maintained. In such cases, cloud-based audio processing reduces the installation and maintenance cost but provides data subjects with even less privacy guarantee.

To protect the privacy of the end user, we propose to obfuscate the audio on the edge device. The obfuscation is implemented as a deep neural network (DNN) with a small computational footprint that transforms the original audio in such a way that only selected sensitive attributes are retained,

such as gender, identity, etc. We use the principle of adversarial training with a newly designed privacy loss metric to train the obfuscator. The downstream analysis model (running in the cloud or on the edge device) then only has access to the obfuscated audio instead of directly to the raw microphone data. In this paper, we refer to this model as the target task, with a task consisting of the extraction of one or more permitted attributes. The principle is illustrated in Figure 3.1.

Crucially, the audio is transformed in such a way that it can still be processed by a pre-trained DNN. Our filtering approach could thus be offered as a virtual sensor to existing 3rd party applications, with the filter running in a protected hardware environment.

A second major benefit of our approach, and differentiating it from existing approaches, is that it provides an *opt-in* regime, meaning that the exposed data can only be used for authorized tasks. Alternative works are "opt-out", requiring users to enumerate the attributes they do not want to provide permission for, which is less protective of privacy.

The main contributions of this paper are threefold. First, to the best of our knowledge, this is the first work to consider the opt-in regime on audio analysis tasks. Secondly, we propose a privacy loss function that uses latent space feature representations that capture higher level attributes than the commonly used metrics that work directly with the raw audio. Finally, our solution outputs an obfuscated audio stream that is still compatible with a pre-trained DNN for the target task, making it compatible with 3rd party applications.

The remainder of this paper is structured as follows. After discussing related prior work, we describe our obfuscator framework. This framework is evaluated in an experimental set-up involving four datasets and three attributes. Finally, we discuss the limitations and scope of our work and provide pointers for future work.

3.2 Related Work

As minor clues can already reveal privacy sensitive personal information [8], there is an increasing interest in privacy-enhancing technologies for machine learning applications. Most of the existing works protect one specific attribute, such as gender [9], identity [10, 5], or emotion [11].

In the VoicePrivacy 2020 Challenge [12], the task is to protect speaker identity in automatic speech recognition (ASR) tasks. State-of-the-art in this

competition is the Distribution-Preserving X-Vector Generation approach [10]. X-vectors are fixed length embeddings of audio fragments that capture all information on the speaker identity but not on the spoken content [13]. Speech is anonymized by generating synthetic audio, replacing the original x-vector with a fake x-vector sampled from a Gaussian Mixture Model that was fitted on the principal components of the x-vectors of speakers in a large public dataset.

Other works on privacy in audio-based applications focus on acoustic event classification instead of ASR as the target task. Nelus et al. observed that feature extractors designed for event classification often produce representations containing a significant amount of speaker-dependent data [5]. They first train a feature extractor for the target classification task, which they call the trust model. Through a hyperparameter, they control during the training process the balance between classification performance and the mutual information between the original input and the extracted feature vector. Afterwards, they train a threat model that interprets the extracted feature vectors as x-vectors and aims to extract speaker information. They experimentally demonstrate the trade-off between trust and threat model performance. While they use existing architectures for both trust and threat models, the main disadvantage of their approach is that the classifier of the target task has to be retrained with the modified loss function. Our approach, on the other hand, does not require retraining the model of the target task.

Other approaches rely on disentanglement to protect certain speaker attributes. Noé et al. propose an adversarial disentangling autoencoder to conceal the gender attribute from the speaker identification task [9]. Their framework consists of a pre-trained gender classifier, an encoder, a decoder, and a gender classifier. During training, the decoder tries to generate a gender-protected x-vector from the output of the encoder and the pre-trained classifier. During the inference phase, the decoder is fed with a randomly selected gender value to generate a gender protected x-vector. The major disadvantage of this work is that it is an opt-out approach which only protects pre-specified attributes.

Whereas these approaches modify feature representations to remove a predefined sensitive attribute, we generate a transformed audio signal. This allows us to use our model in combination with an off-the-shelf recognition model without retraining. Moreover, x-vector-based approaches are limited to applications with ASR as the target task. Our approach is also applicable to other tasks. In addition, we provide an opt-in framework where all task-irrelevant information is removed.

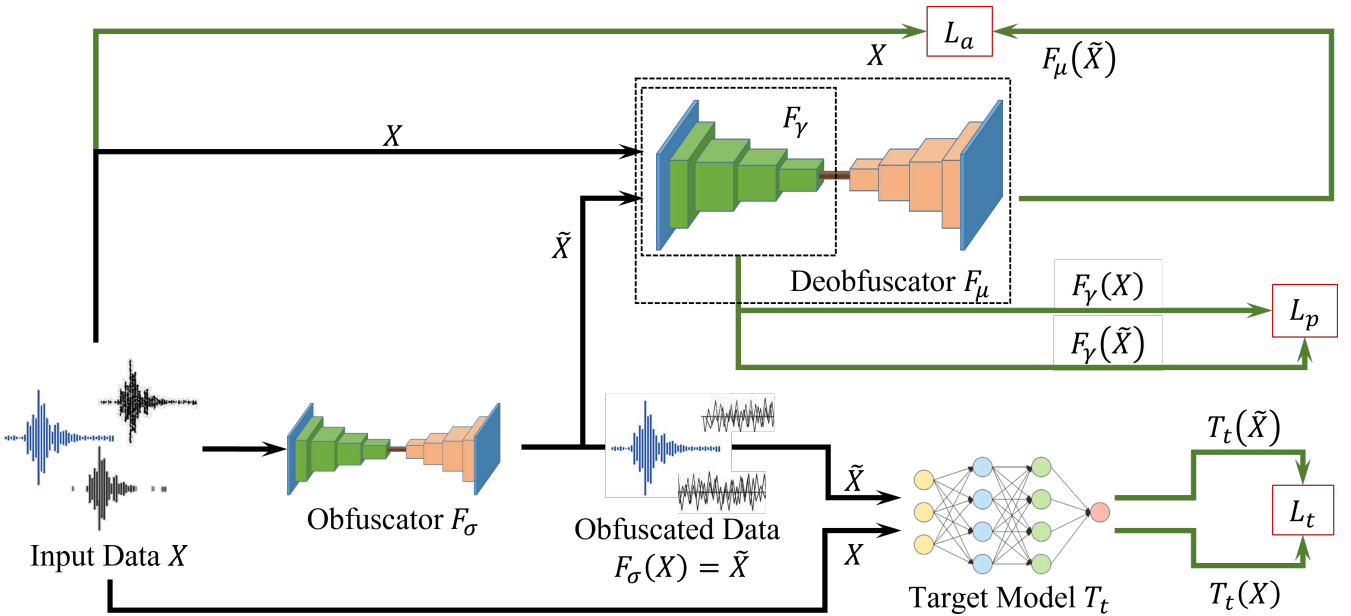


Figure 3.2: Overview of the proposed framework. During training, the raw data is fed to the obfuscator which generates a privacy-preserving version. The pre-trained target model is agnostic to the data obfuscation and can still perform the intended task. Meanwhile, the deobfuscator tries to reconstruct the original data from the privacy-preserved data. The deobfuscator is discarded for inference. L_a, L_t, and L_p are the adversarial loss, target loss, and privacy loss. The black, and green lines indicate the flow of data, and the calculation of loss, respectively.

3.3 Opt-in Privacy Protection Framework

Our framework consists of a target model, an obfuscator and a deobfuscator. The relationship between these components is visualized in Figure 3.2. Note that the deobfuscator is only used during the training phase. The target model T is a function that represents the task the user wants to opt-in for. The input signal X is the log spectrum of the raw audio, a common feature representation used as input in many audio DNN processing applications, while still allowing to decode into raw audio when needed.

Given the original signal X, the target model outputs $T(X)$, for instance, a label indicating the recognized gender, emotion or speaker. We only assume read access to the model T and never change its parameters. In modern applications, T is a pre-trained DNN.

The obfuscator F_σ is a DNN with trainable parameters σ that transforms the original signal X to $\tilde{X} = F_\sigma(X)$. To be compatible with the target model, $\tilde{X}$ has the same dimension as the input signal X. We also introduce a deobfuscator F_μ DNN with trainable parameters μ, that tries to reconstruct the original signal from the obfuscated signal. Since we aim for an opt-in approach, we cannot train the deobfuscator on the performance achieved in particular tasks. Instead, the training objective is to minimize the Mean Square error (MSE):

$$L_a = MSE(F_\mu(\tilde{X}), X). \tag{3.1}$$

The training objective of the obfuscator consists of two (weighted) loss terms L_t and L_p, reflecting the opposing goals to sustain task performance after transforming X, while removing as much information as possible in order to prevent the reconstruction of X from the modified signal. In classification target tasks, as used in this chapter, L_t is the cross-entropy loss H between the original and transformed signal:

$$L_t = H(T(\tilde{X}), T(X)). \tag{3.2}$$

Following the traditional adversarial approach with $L_p = -L_a$ did not provide satisfying results. The main reason is that the log spectrum of the audio X is a too sparse feature encoding. Instead, we use an intermediate distributed representation of dimension M that is the output after processing the first N layers of F_μ. We thus define F_γ as the sub-model of F_μ, with trainable parameters $\gamma \subset \mu$, and define the privacy loss L_p as the distance in each dimension of the latent representation:

$$L_p = \frac{1}{M} \sum_{m=1}^{M} D(F_\gamma(\tilde{X})_m, F_\gamma(X)_m). \tag{3.3}$$

Since the variance between latent dimensions might vary significantly, we compute the distance in each dimension with the following distance function D:

$$D(\alpha, \beta) = |\frac{1}{1 + e^{-(\alpha-\beta)}} - 0.5|. \tag{3.4}$$

3.4 Experimental Setup

Our evaluation focuses on sensitive attributes that can be extracted from speech. Although the principle should extend to non-speech applications as well, the choice to focus on speech was made because of the availability of public datasets with labels for multiple sensitive attributes and existing opt-out algorithms for these attributes to benchmark against.

We compare our method with the recent Adversarial Disentanglement Representation (ADR) framework[9], which is to our knowledge the work that comes closest to our approach. This opt-out framework uses x-vector as audio feature encoding. These x-vectors are converted by an encoder, which is adversarially trained against a classifier for a pre-specified protected attribute.

3.4.1 Datasets

We evaluate our proposed method on four datasets: The Emotional Voices Database (EmoV) [14], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [15], Librispeech [16], and VoxCeleb2 [17]. The EmoV and RAVDESS datasets contain audio fragments of speakers with four labeled attributes: gender, speaker identity, emotion, and utterance. The five emotions considered in EmoV are neutral, amused, angry, sleepy and disgust. The eight emotions considered in RAVDESS are neutral, calm, happy, sad, angry, fearful, surprise and disgust. The other two datasets include more speakers and utterances, but have no labels for emotion.

In real applications, it is very unlikely that the end-user of an application is one of the users whose voice was included in the labelled datasets used to train the obfuscator and target models. To mimic this setting, we split Librispeech and VoxCeleb2 into training and test sets that have different speakers. For Librispeech, we use as training set the merger of the `train100` and `train360` subsets and the `testclean` subset as our testing set. As for the VoxCeleb2 [17] dataset, we followed the experimental set-up of ADR and consider the `V2D` subset as our training set and `V2T` subset as our testing set. For EmoV and RAVDESS datasets, limited by the number of speakers in

each dataset, the training and testing datasets contain different fragments but the same speakers.

We pre-process the data by first re-sampling at 16000 Hz and compressing into mono-channel audio. After normalizing the volume, we follow the settings of deepspeech2 [18] to extract the spectrum.

3.4.2 Model Architecture

We use the DeepSpeech2 model [18] as the architecture of the target model T. This model takes raw audio as input and is widely used for speech-to-text recognition but can be easily adjusted to attribute classification tasks. This model exists in different variants, for our experiments we set the number of channels for convolution and BatchRNN to 16 and 64, respectively. To fit our target task of classifying sensitive attributes, we only adapted the dimension of the last fully-connected layer to the number of classes for the task at hand. For each segment of the input audio, the model produces a prediction and the final class is decided by vote counting over all segments. While ADR was evaluated against a different target model, the evaluation results presented in the next section indicate that performance of our target model on original audio fragments is similar to the performance of the ADR target model.

The goal of the obfuscator and deobfuscator is to transform the input data into an output with the same format. Since this task is very similar to voice conversion, we adopted for the obfuscator and deobfuscator the CycleGAN-VC2 [7] architecture, a popular state-of-the-art voice converter that maps the content and style of one speaker onto another. The obfuscator is downsized to fit on a resource-constrained edge device. We adjusted the number of residual blocks (3 instead of 6) and reduced the number of feature channels by a factor of 8. The deobfuscator was not downsized. The number of parameters of the obfuscator and deobfuscator are 0.759M and 6.847M, respectively. We train both obfuscator and deobfuscator with stochastic gradient descent with an initial learning rate of $1e-2$ and a momentum of 0.9. Obfuscator and deobfuscator were trained jointly on a Tesla V100-SXM2 model. Training converged after 8 hours on the smaller EmoV and RAVDESS datasets, and after 5 days on the larger VoxCeleb2 and Librispeech datasets.

3.4.3 Attacker model

We consider two types of attackers: an ignorant attacker and an informed attacker. The ignorant attacker is unaware of the existence of the obfusca-

tion, but is capable of training his own classification model based on the same publicly available datasets that the target task was trained on. The informed attacker on the other hand is aware of the existence of the privacy protector. He has retrieved access to the trained obfuscator and was able to generate obfuscated versions of the fragments in the public datasets. He thus possesses a dataset containing obfuscated data with ground truth labels and can train a model specifically to undo the obfuscation. For a fair comparison, all the attackers share the same architecture of the target model of our method.

3.5 Results

In the following sections, we first discuss the target task performance and privacy protection for both systems. We then show how the opt-in system can protect other attributes that were not specified beforehand. Finally, we analyze the computational cost of our approach.

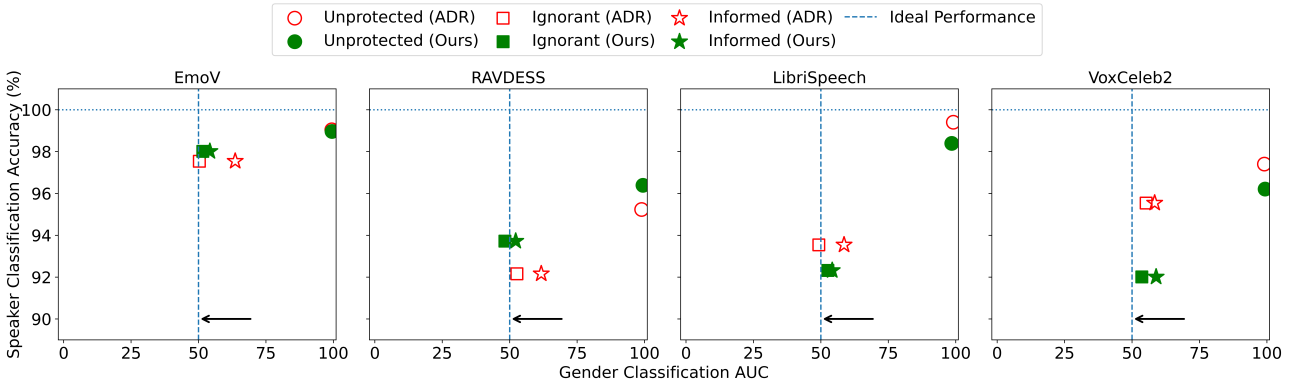
3.5.1 Privacy Protection

In the first experiment, we evaluate our opt-in system and the opt-out ADR in terms of classification accuracy on the target task and on the unauthorised attributes. We focus on the attributes of gender and speaker id, since these labels were present in all datasets.

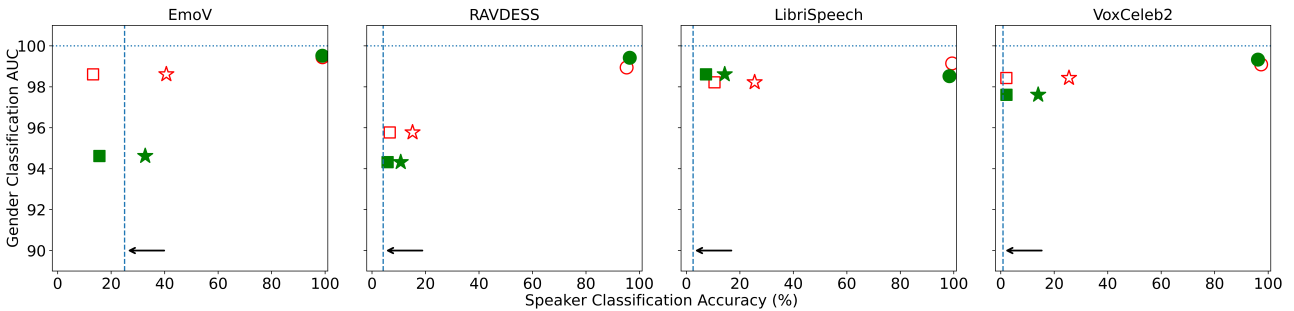
The results of the first experiment are shown in Figure 3.3. The classification performance on the target task and on the unauthorised task are shown on the Y-axis and X-axis respectively. The random classification performance on the unauthorised task corresponds with optimal protection and is shown by a vertical dashed line. Good protection on the unauthorised task and good performance on the target task correspond to the upper-left corner of the graphs.

Since we use a different audio encoding than ADR, we first confirmed if both representations contain the same amount of information about the to-be protected attributes by training classifiers on these input representations, resulting in similar classification performance as indicated by the circles in Figure 3.3. Squares and stars indicate the results obtained by the ignorant attacker and the informed attacker, respectively.

Figure 3.3 (a) illustrates the results obtained by an attacker on gender while allowing for speaker recognition. Both our framework and ADR are able to protect gender recognition up to the level of random guessing against ignorant attackers. Informed attackers who were able to train specifically



(a) Unauthorised gender retrieval on privacy-protected data for speaker identification.



(b) Unauthorised speaker retrieval on privacy-protected data for gender recognition.

Figure 3.3: Results of sensitive information retrieval on privacy-protected data for different tasks. (a) shows the cases of allowing speaker identification and defending against attacks on gender retrieval. (b) shows the cases of allowing gender recognition and defending against attacks on speaker retrieval. The vertical dash lines indicate the results of random guessing, where the arrow indicate better privacy protection. The horizontal dash lines indicate the performance on the target task, resulting the crossing of the dash lines as the best performance. Note that the axes of target task performances are scaled to 90 to 100 for better visualization.

against ADR or our obfuscator manage to retrieve more information on the protected gender attribute. Our model provides a slight but consistently better protection on all datasets.

For both frameworks, this protection comes at the cost of a degradation in classification performance of 2-6% on the target task of speaker identity recognition. Arguably, there is mutual information between gender and speaker identity as they are correlated attributes; however a more in-depth analysis, e.g. as performed in [19] is needed to determine whether this correlation fully explains the performance degradation. On the larger datasets, our model is outperformed by ADR on the task performance. This could be caused by two reasons. First, the DeepSpeech2 model takes spectrogram as input, which preserves more information but is also more complex to reconstruct. The second possible reason is the correlations between the two attributes. It is logical that some attributes, e.g. gender and speaker identity, share some information. Thus it is impossible to completely remove the information from one another.

When we switch the target task and the unauthorised tasks, similar conclusions can be drawn, see Figure 3.3 (b). Ignorant attackers are not able to perform better than randomly guessing on data protected by both frameworks, but our framework provides consistently better protection against informed attackers than ADR. This improved protection comes at the cost of a small drop in target task performance.

3.5.2 Opt-out Versus Opt-in

In the second experiment, we aim to demonstrate the differences between an opt-out and an opt-in system, and the advantages of the latter. As mentioned before, an opt-out system requires explicitly specifying which attribute has to be protected, rather than which attributes are permitted. Thus, to demonstrate the difference between an opt-in system and an opt-out system, we perform an attack on the attributes that are not specified by the opt-out system.

Besides the speaker id and gender, we include emotion as a third sensitive attribute. Since only the EmoV and RAVDESS datasets provide labels for these three attributes, this experiment is only conducted on these two datasets.

Following the description of [9], we train ADR with speaker recognition as target task. Being an opt-out system, ADR also requires us to explicitly specify which attribute to be protected. We choose gender as protected

attribute, as in the original paper. Emotion is thus the unspecified attribute that a user might inadvertently expose.

Figure 3.4: Results of sensitive information retrieval (gender and emotion) on privacy protected data for speaker identification. The horizontal and vertical dash lines indicate the results of sensitive information retrieval on hypothetically perfectly protected data, with their crossing indicating the best protection.

The results of this experiment are illustrated in Figure 3.4. The axes show the classification performance on the protected gender attribute, and the unspecified emotion attribute. Dashed lines indicate the performance of random guessing, the best possible protection level one can achieve.

Both frameworks provide protection against an uninformed attacker, as the classification on obfuscated data approaches those of a random classifier. Our framework is however much more robust against informed attacks, protecting both attributes while ADR only achieves reasonable protection against the pre-specified gender attribute. An informed attacker manages to retrieve emotion information with more than 70 % accuracy on the ADR-protected data.

This experiment shows that the opt-in regime provides better protection when the attribute of interest is not previously known. Although specifying specify multiple protected attributes could remedy this particular case, the fundamental challenges of opt-out systems remain.It is impossible to define

Table 3.1: Computational time measured on different devices.

Device	CPU	GPU	Time (ms)
Raspberry pi 2B	ARM Quad-Core Cortex-A7	Not used	1923
NVIDIA Jetson TX1	ARM Quad-Core Cortex-A57	256-core NVIDIA Maxwell	34
Server	Intel Xeon Silver 4108	NVIDIA GTX 1080 Ti	1

all the possible attributes that may be interested by all parties. Thus, opt-out systems like ADR would still leak privacy information even with multiple predefined attributes.

Summarizing the results of both experiments, we conclude that our framework provides good protection against both ignorant and informed attackers. The major advantage of our framework is the opt-in aspect, which aims to only retain information in the obfuscated signal relevant for the authorized task. However, this improved protection comes at a limited cost in classification performance on the permitted attribute.

3.5.3 Computational Cost

We further measure the computational time of the framework (without de-obfuscator) on edge devices to simulate how the proposed framework works in a real-world scenario. In Table 3.1, we show the average execution time of obfuscating one second of audio on different platforms. Only on devices with embedded GPU, the model can work in real-time.

3.6 Conclusion and Future Work

In this chapter, we introduced a novel opt-in framework to preserve privacy while using audio applications. We use adversarial training and a novel privacy-preserving loss metric to train an obfuscator that removes all but the information needed for the authorised task. Unlike existing approaches, we do not require an adaptation of the target task classification models. This allows the obfuscator to be integrated in a pipeline with existing third-party audio services.

We validated our approach on four voice datasets and compared it against one state-of-the-art approach for privacy protection. We evaluated protection against two types of attacks and show that our method can protect privacy with only a small reduction in classification accuracy on the permit-

ted task. We further showed the strength of the opt-in framework against unspecified attacks compared to the opt-out framework.

The proposed opt-in framework still has a few limitations that mandate future research before being applied in real-world scenarios. Firstly, we have evaluated our obfuscator architecture with the classification of one attribute as target task. How the current model performs on other task types such as speech-to-text recognition is yet to be investigated. In its current inception, having multiple permitted target tasks would require multiple obfuscators. Creating one obfuscator model with configurable target tasks would first require an in-depth study of the correlation between attributes. Secondly, although we do not modify the target task model, we require white-box access to backpropagate weight updates to the obfuscator. This makes our approach only compatible with third-party services with known architecture and parameter values. To overcome this limitation, one possible solution is to leverage transfer learning strategies. Finally, deployment of the obfuscator on low-end devices would require network compression techniques such as pruning and quantization [20].

Audio-based applications are very attractive, but pose significant privacy risks to the user. We hope that this paper will inspire other researchers to contribute to better protection mechanisms.

3.7 Acknowledgements

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen” programme.

3.8 References

[1] Joan Navarro, Ester Vidaña-Vila, Rosa Ma Alsina-Pagès, and Marcos Hervás. Real-time distributed architecture for remote acoustic elderly monitoring in residential-scale ambient assisted living scenarios. *Sensors*, 18(8):2492, 2018.

[2] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24, 2018.

[3] Igor DS Miranda, Andreas H Diacon, and Thomas R Niesler. A comparative study of features for acoustic cough detection using deep architectures. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2601–2605. IEEE, 2019.

[4] Andrew Trask, Emma Bluemke, Ben Garfinkel, Claudia Ghezzou Cuervas-Mons, and Allan Dafoe. Beyond privacy trade-offs with structured transparency. *arXiv preprint arXiv:2012.08347*, 2020.

[5] Alexandru Nelus and Rainer Martin. Privacy-preserving audio classification using variational information feature extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2864–2877, 2021.

[6] Huafeng Jin and Shuo Wang. Voice-based determination of physical and emotional characteristics of users, October 9 2018. US Patent 10,096,319.

[7] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6820–6824. IEEE, 2019.

[8] Aleksandra Korolova. Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops*, pages 474–482. IEEE, 2010.

[9] Paul-Gauthier Noé, Mohammad Mohammadamini, Driss Matrouf, Titouan Parcollet, Andreas Nautsch, and Jean-François Bonastre. Adversarial disentanglement of speaker representation for attribute-driven privacy preservation. *arXiv preprint arXiv:2012.04454*, 2020.

[10] Henry Turner, Giulio Lovisotto, and Ivan Martinovic. Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020. *arXiv preprint arXiv:2010.13457*, 2020.

[11] Ranya Aloufi, Hamed Haddadi, and David Boyle. Emotionless: Privacy-preserving speech analysis for voice assistants. *arXiv preprint arXiv:1908.03632*, 2019.

[12] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O' Brien, et al. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362, 2022.

[13] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing*, pages 5329–5333. IEEE, 2018.

[14] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

[15] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing*, pages 5206–5210. IEEE, 2015.

[17] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[18] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[19] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. Selective manipulation of disentangled representations for privacy-aware facial image processing. *arXiv*, 2022.

[20] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021.

4

Unsupervised Audio-Visual Representation Learning

Why choose sides when you can have it all?

- Another Meme

Embedding-based Pair Generation for Contrastive Representation Learning in Audio-visual Surveillance Data

Wei-Cheng Wang • Sander De Coninck • Sam Leroux • Pieter Simoens
Published in Frontiers in Robotics and AI: Emerging Technologies in Surveillance: Novel Approaches to Video Anomaly Detection 2025

We can achieve a better understanding of the semantic content of a surveilled scene when combining audio and visual data. In this chapter, we aim to propose an audio-visual representation learning framework in an unsupervised manner, by using contrastive learning. We investigate two limitations when applying contrastive learning on audio-visual surveillance data: false negatives and the minimal sufficient information bottleneck. As explained in Chapter 1, surveillance data contains irregular, yet frequently recurring events; which can lead to a considerable number of false-negative pairs and disrupt the model's training. To tackle these challenges, this chapter proposes a novel pair generation mechanism for contrastive learning, and an modified loss to enlarge the information encoded in the audio-visual representation. Therefore, this chapter's objective is to develop and validate a framework that learns information-rich, audio-visual representations for multiple downstream tasks, including high-level tasks such as event retrieval, without expensive human annotations as supervised methods.

4.1 Introduction

Today, around 55 percent of the global population is living in an urban area or city, and this number is expected to rise to 68 percent by 2050 [1]. To support this urbanization in a sustainable way, smart cities deploy a variety of sensor, networking and data analysis technologies to improve their operational efficiency and safety measures. Cameras and microphones are two prevalent sensors in smart city applications. Cameras primarily serve surveillance functions, facilitating crime prevention and traffic monitoring, while microphones are utilized for detecting phenomena such as gunshots or glass shattering [2]. Deploying cameras and microphones in the same location enables more comprehensive situational insights. Audio and video cues provide complementary information, which enhances the robustness of event detection against challenges encountered in real-world settings, including noise, occlusions, or low-light conditions [3].

Figure 4.1: Sampling data from different timestamps may result in false negative pairs if both timestamps share a similar audio-visual context. (A) Visual frame at 00:31:52. (B) Visual frame at 01:39:40.

Deep neural networks are currently the state-of-the-art solution for audio-visual surveillance tasks such as vehicle detection [4], violent scene detection [5], and sound tagging [6]. However, training these models requires large (labelled) datasets that are expensive to collect. Furthermore, research indicates the advantages of employing location-specific models for surveillance [7], further increasing the amount of training data and associated labels that need to be collected.

The objective of this work is to design a scalable framework for learning representations of real-world audio-visual surveillance data in a self-supervised manner. The resulting representations should generalize well to a wide range of downstream surveillance tasks, meaning that the training of task-specific models that will take these representations as input requires little or no labelled data. Examples of downstream tasks for smart city surveillance include event localization [8], anomaly detection [9, 10, 11] and event search [12].

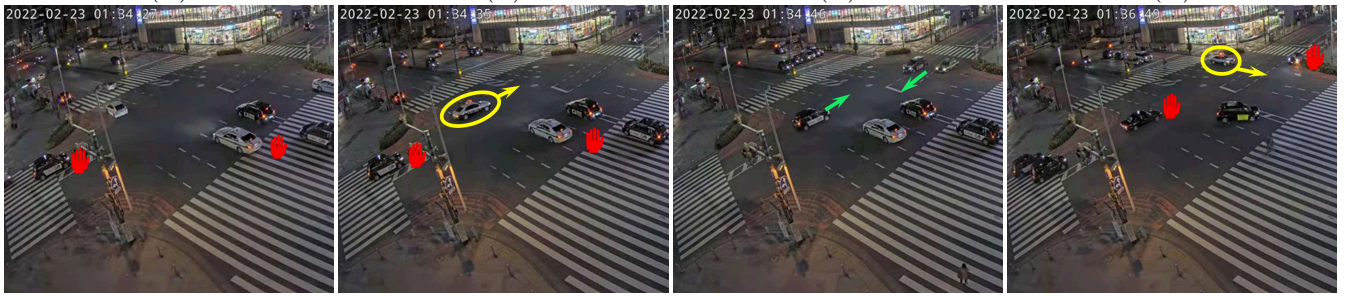
Self-supervised learning of transferable representations is typically achieved by training a feature extraction model on a pretext task. Contrastive learning, a specific type of self-supervised learning, formulates the training objective in terms of a distance metric between the representations of a pair of input samples. The goal is to minimize the distance for semantically similar instances (positive pairs) and maximize it for dissimilar instances (negative pairs). The process of generating positive and negative pairs during training is a crucial factor in obtaining transferable features. Negative pairs are often generated through random sampling from the dataset. Positive pairs can be constructed without requiring label information by pairing a sample with an augmented version of that sample. Such augmentations are straightforward in the case of static images, but much harder to design

for temporal data [13]. In the case of multi-modal data, positive pairs can be naturally formed by treating audio and video clips sampled at the same timestamp within a stream as positive pairs, a pair generation mechanisms known as Audio-Visual Synchronization (AVS) [14]. To distinguish it from our approach, we refer to it as Temporal-based Pair Generation (TPG) to highlight that typical Audio-Visual Synchronization takes temporal cues when generating data pairs.

TPG however introduces two challenges related to the semantic repetition that is observed in audio-visual surveillance data over time. First, a large temporal distance between two fragments of a recording does not guarantee a semantical difference between these fragments. Ambulances, police cars, buses, auditory beacons for visually impaired pedestrians, or vans with similar appearance are only a handful examples of scenes recurring at unpredictable and variable intervals. One example taken from a surveillance camera in Tokyo is shown in Figure 4.1. (A) and (B) shows two visually and aurally similar trucks appearing at different time frames. When sampling a data pair where the visual modality is taken from (A) and the audio modality from (B), this pair is labeled as negative based on the time stamps, despite their semantic similarity. Such mislabeled pairs, referred to as *false negatives*, compromise the training process and cause the learned embedding spaces to lose the semantic meaning [15, 16, 17].

Another limitation of relying on temporal cues to generate positive and negative pairs in contrastive learning arises from the information bottleneck in the training objective. Since all supervision information for learning a representation of one element comes from the contrasting element [18], the representations are *minimal sufficient*, meaning that they are focused on the mutual information between the samples of positive pairs. While this is effective when the downstream task is aligned with the pretext task, the minimal sufficient may not contain enough information to generalize across multiple downstream tasks [19, 20, 21]. Increasing the number of positive pairs for each sample can address this limitation, as it makes the pretext task more challenging and encourages the representations to encode richer information [17, 22, 18]. With TPG, each video clip forms a single positive pair with its corresponding audio, leading to minimal sufficient representations that only contain information on objects that are simultaneously audible and visible. However, real-world events are complex, involving multiple elements at different time intervals, as shown in Figure 4.2. Figure 4.2 (A-C) shows the progressive stages of a police car cautiously passing by a busy intersection. Due to the relative distance and velocity between the camera and the vehicles, as well as interactions between vehicles, each

(A) (B) (C) (D)



Police Cars	Audible & Invisible Stop	Audible & Visible Stop	Audible & Invisible Moving	Audible & Visible Stop
Time	34:27	34:35	34:46	36:49

Figure 4.2: Example stills illustrating two instances of a police car passing an intersection with activated siren, forcing other vehicles to stop. The police car is indicated in yellow. Stopped and moving cars are indicated with red signs and green arrows respectively. (A) Only the siren of the police car is audible. (B) The police car enters the scene. (C) The police car leaves the scene but siren is still audible. (D) On a later time, a police car enters the scene with a different view angle.

timestamp is a unique combination of visual and audio cues of the same event. Because of the information bottleneck, the representation of this scene might not contain all the visual information (police car and vehicle stopping) with its audio modality (siren). In the case of Figure 4.2 (B), when learning a minimal sufficient representation, part of the visual information could be ignored (e.g. vehicle stopping), although this information could be essential for downstream tasks. For example, during an emergency, the audio of the siren and the visual cue of other vehicles stopping are crucial for managing the traffic light before the police car enters the camera's view Meanwhile, (D) shows a similar event occuring at another time, where a police car approaches from a different road. By creating positive pairs using samples from both events, the learned representations will contain more comprehensive information, accommodating the complexities of real-world scenarios.

To reduce false negatives as well as to learn representations with suffficient information, semantically similar events should be mapped together in the embedding space. In this paper, we introduce the *Embedding-based Pair Generation* (*EPG*) mechanism as an alternative for TPG to sample positive or negative pairs. Our approach detects false negative pairs by calculating a distance between the embeddings of two instances of the *same* modality. Furthermore, we propose a new loss that considers multiple positives simultaneously to learn representations that contain more information, further improving the transferability of the learned features to a variety of downstream tasks. We train a pseudo-Siamese network to encode the audio segments and video frames to the same embedding space. After training, the model has learnt what audio typically corresponds to certain visual inputs and vice versa. The two deep neural networks can then serve as feature extractors, jointly or separately, for downstream tasks.

To summarize, our main contributions are as follows:

1. We identify the inherent flaws in applying the widely used audio-visual correspondence to smart surveillance data. An embedding-based pair generation is introduced to tackle this problem;

2. We study the limitation of minimal sufficient representation for audio-visual representation learning in surveillance. We then propose a novel loss to encode richer task-relevant information to improve the performance on downstream tasks;

3. We evaluate our approach with supervised downstream tasks and demonstrate the effectiveness of our improvement compared to state-of-the-art approaches in audio-visual representation learning. We fur-

ther qualitatively evaluate our approach on two unsupervised tasks applied on real-world surveillance data.

The remainder of this paper is structured as follows. Section 5.2 provides the related work of audio-visual representation learning, self-supervised learning, and how positives and negatives are generated for contrastive learning. In section 5.3, we describe the mechanism of embedding-based pair selection and how we incorporate multiple positives in the contrastive loss. The downstream tasks and datasets used to evaluate the learnt representation are explained in section 4.4. We then describe the implementation details and the discussion on the experimental results in section 4.5. We conclude in section 4.6 and give some directions for further research.

4.2 Related Work

Our work lies at the intersection of three domains: audio-visual representation learning, self-supervised representation learning, and pair generation for contrastive learning. In the following subsections, we provide an overview of the approaches in each of these fields that are most pertinent to our work.

4.2.1 Audio-Visual Representation Learning

The analysis of audio-visual data is gaining popularity as audio and visual information offer complementary insights on the same content. The two modalities are expected to align, either at the frame level or the instance level. Jointly considering both modalities benefits the analysis of audio-visual tasks such as active speaker detection [23], sound source localization [24], lip reading[25], or video forensics [26].

To encode audio-visual representation, most frameworks employ three components: an audio encoder, a visual encoder, and a projector. This modular design accounts for the distinct characteristics of audio and visual data, requiring different processing configurations, such as varying network architectures and learning schedules. Once high-level information is extracted by the encoders, the embeddings are connected by passing them through the projector. This setup is especially useful for tasks like anomaly event recognition [27], where labelled data is easier to obtain, allowing for straightforward end-to-end training to learn audio-visual representations.

However, in most real-world scenarios, audio-visual data is collected continuously, making it challenging to obtain annotations. To address this, self-supervised learning (SSL) is a promising approach to leverage the semantic synchronization between audio and video. The inherent correlation

between audio and visual elements serves as a natural indicator when designing the pretext task in self-supervised learning. When applying SSL to audio-visual data, the general concept involves training a model to differentiate between matching and non-matching pairs of video segments and audio excerpts. Negative pairs are constructed either by sampling audio and video from different recordings, known as audio-visual correspondence (AVC), or by sampling from different offsets in the same recording, termed audio-visual synchronization (AVS).

AVC as pretext task was first introduced by Arandjelovic and Zisserman [28], who demonstrated that the learnt representations obtained competitive results on both audio tasks, such as sound classification, and visual tasks, such as image classification and object detection. Subsequent works extended AVC to tasks such as action recognition [29], active speaker detection [23] or sound source localization [24]. More recently, Huang et al. [30] proposed a hybrid approach combining generative SSL objectives with contrastive learning. With a joint loss function, both inter-modal and intra-modal relationships can be considered by the model.

AVS, as a more nuanced pretext task, leverages the temporal synchronization between audio and video to pretrain the model. This approach has been successfully applied to tasks such as lip reading[25], video forensics [26] and active speaker detection [31].

4.2.2 Self-Supervised Representation Learning

While supervised learning has made a great achievement in many research domains, accessing reliable annotations for data is sometimes expensive or impractical. Self-supervised learning aims to learn a representative embedding by leveraging the information within the data instead of relying on the supervision of annotations. SSL introduces pretext tasks, which are auxiliary tasks designed to train the model to learn representations that can later be applied to downstream tasks. These pretext tasks may not directly relate to the target task but serve as an effective means of extracting generalizable embeddings.

According to the type of the pretext task, SSL approaches can be divided into three different categories [32]: generative, predictive and contrastive. Below, we briefly describe all three and will then focus on the contrastive approaches as these form the basis for our work.

Generative SSL employs generative models, such as AutoEncoders (AEs) or Generative Adversarial Networks (GANs), coupled with pixel-level reconstruction loss functions to learn representative features. This approach is

particularly popular in the field of computer vision [33, 9, 34]. While pixel-level reconstruction is an intuitive and effective pretext task, the generative SSL methods can be hard to train. During the training of the generative model, model tend to focus overly on background details at the expense of the foreground content. This happens particularly when the foreground content is relatively small in terms of frame ratio, known as the foreground-background imbalance. Another common challenge is the object scale imbalance, where the size of the objects varies when the camera has a more oblique view, as discussed in [35]. Both problems require additional mechanisms to focus on specific semantic information.

Predictive SSL methods utilize self-generated labels derived from predefined transformations of the input data to guide network training. Pretext tasks such as classifying rotated versions of the original image [36], or arranging image regions within a jigsaw puzzle [37] have been demonstrated to result in high-level features of images. These pretext tasks preserve the semantic meaning of the content. It is however not trivial to design good pretext tasks for temporal data, e.g. in surveillance applications, due to the added complexity of sequence dynamics.

Finally, contrastive SSL aims to overcome the challenges encountered in the aforementioned approaches. Contrastive SSL compares pairs of data samples to learn representations that maximize similarity for positive pairs (semantically similar samples) and minimize similarity for negative pairs (semantically dissimilar samples). Positive pairs typically consist of different augmentations of the same input sample [38, 39], while negative pairs are created by randomly pairing samples from the dataset. Both inputs are projected to a shared embedding space. By training the model to maximize the mutual information between embeddings of samples in positive pairs and minimize that of samples in negative pairs, the model learns to extract high-level features that can be used for downstream tasks. Contrastive SSL extends to multimodal data by forming pairs with samples from each modality. For instance, in audiovisual representation learning, video fragments paired with corresponding audio fragments represent positive pairs, while combinations of video frames and randomly selected audio snippets serve as negative pairs [16, 8]. Text is another modality commonly used in conjunction with video, in particular to learn language-video representations by pairing the video with its caption [15, 40].

While contrastive learning has been proven to be effective in many applications, there are two notable limitations, namely *false negatives* [16] and the minimal sufficient representation [18] problem. False negatives occur when semantically similar pairs are mistakenly labelled as negative due to

the design of the pretext task. Zolfaghari et al. [15] describes the impact of false negatives and proposes identifying influential samples. These samples, which are more likely to be false negative samples, have high feature similarity with other samples and should be removed. Similarly, Sun et al. [16] proposes a statistical approach to locate false negatives by considering the similarity between the same modality of different sample pairs. The information bottleneck leading to minimal sufficient representation, containing information that is sufficient for the pretext task, is rather rare in scholarly discussions. Tian et al. [18] thoroughly describes the concept with theoretical and empirical proof, stating that when the downstream tasks are not aligned with the pretext task, it might downgrade the representative of learnt features. This issue is especially pronounced when applying contrastive learning to surveillance data, where events unfold over time and involve multiple stages. A minimal sufficient representation may not be able to describe the different stages of an event. To improve the usability of learned representations in downstream tasks on such data, several works aim to incorporate multiple positives in the objective function [17, 22, 18]. By considering multiple semantically related positives in the objective function, models can better capture the diverse aspects, improving the richness and generality of the learnt features.

4.2.3 Pair Generation for Contrastive Learning

The selection of positive and negative pairs is a crucial factor in contrastive learning [41, 42]. Many works assume that randomly selected inputs lack semantic similarity and can be used as negative pairs. However, this assumption can introduce false negative pairs. This issue, a phenomenon also referred to as *sampling bias*, has been shown to hinder performance. Chuang et al. [17] empirically demonstrate significant performance gains across multiple research domains when false negatives are avoided. Other recent studies prove theoretically and empirically that the quality of the negative samples is more important than their quantity [43, 44].

Recent works have explored several strategies to refine the process of generating positive and negative data pairs. At the level of instance sampling, Kalantidis et al. [43] enhance training efficiency and the quality of the learned representations by synthesizing hard negatives. These hard negatives closely resemble positive pairs and are therefore challenging for the model to distinguish. During the training process, novel hard negatives are synthesized as feature-level linear combinations of the currently hardest examples. Tian et al. [45] focus on enhancing the diversity of the sampled positive pairs. They argue that improving the diversity in positive

pairs helps the model learn representations that are invariant to nuisance variables, since the representations are focused on the mutual information across all positive views. Zhu et al. [44] introduce a feature transformation technique that manipulates features to create both hard positives and diverse negatives. Beyond improving the pair selection, [17] propose the *debiased contrastive loss*, a novel training objective function that considers the approximated distribution of negative samples instead of relying on explicit negative samples.

4.3 Proposed Method

In the following sections, we will first explain the different components of the framework. Then, we introduce the novel embedding-based pair generation (*EPG*) mechanism designed to reduce the number of false negatives. Finally, we introduce a novel loss function and elaborate on how this loss function might address the challenge of minimal sufficient representations.

4.3.1 Architecture

As illustrated in Fig. 4.3, the audio-visual representation learning block follows a pseudo-Siamese structure with two encoders: F_v and F_a. Both encoders are deep convolutional networks, designed to process visual and audio information, respectively. Whereas in conventional Siamese architectures, the parameters are shared between the encoders, here the encoders have a different structure, hence the name *pseudo*-Siamese.

The encoders project the audio and video onto a shared embedding space $\mathcal{Z}$. Surveillance data X is first split into short clips, where each clip x comprises the sequence of frames x^v and an audio segment x^a. By selecting from the frames and segments of X, we first generate a data pair $p_{m,n} = (x^v_m, x^a_n)$, consisting of the m-th video segment and the n-th audio fragment. With F_v and F_a, $p_{m,n}$ is encoded into $\mathcal{Z}$, yielding $(F_v(x^v_m), F_a(x^a_n)) = (z^v_m, z^a_n)$. By utilizing $(F_v(x^v), F_a(x^a))$ and the pair generation mechanism, a contrastive loss $\mathcal{L}$ is calculated to train the network. Once F_v and F_a are trained, the encoders can be used as feature extractors for downstream tasks, either jointly or independently.

4.3.2 Embedding-based Pair Generation

Recognizing the limitations of relying solely on time offsets to ascertain semantic dissimilarity, we introduce an alternative solution to identify temporally non-aligned but semantically similar data pairs with the distance in

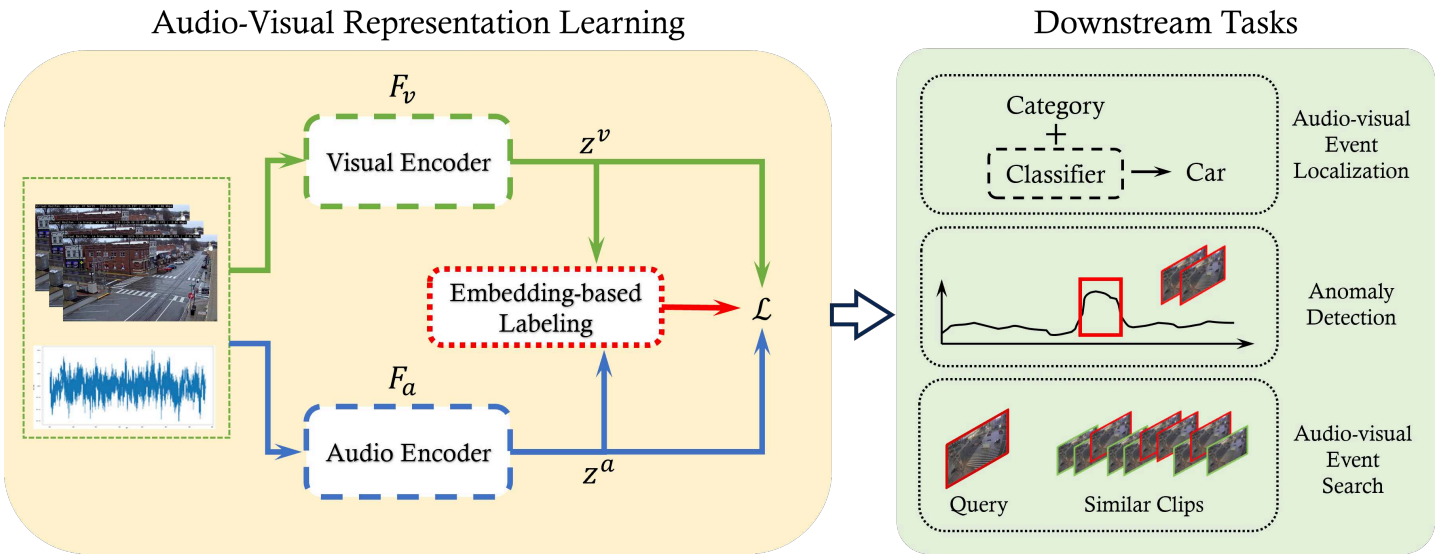


Figure 4.3: Overview of the framework. A pair of video clip x^v and audio segment x^a are served as the input of a two stream pseudo-Siamese network. The network consists of a visual encoder F_v and an audio encoder F_a to project the input into the same embedding space. A embedding-based label of the data pair is calculated to determine whether the data pair is positive or negative. These labels are later used to compute the proposed loss and to update the network. After the pseudo-Siamese network is trained, both encoders can be used as a feature extractor jointly or separately for downstream tasks.

the embedding space.

The set of possible pairs is denoted as $P = \{p_{m,n} = (x_m^v, x_n^a)\}$. Instead of solely relying on the condition $m = n$ to label the pair $p_{m,n}$ pair as positive, we propose to calculate the mutual information between x_m and x_n in the embedding space $\mathcal{Z}$. The mutual information $I(z_m, z_n)$ can be computed using a distance metric $d_{\mathcal{Z}}(\cdot)$ such as Euclidean distance or cosine similarity [46]. If the mutual information between the video or audio embeddings, $I(z_m^v, z_n^v)$ or $I(z_m^a, z_n^a)$, is higher than a threshold δ, we say x_m and x_n are semantically similar, even though they are recorded at different times.

To identify whether the two elements of the pair $p_{m,n}$ are semantically similar, we calculate the label $y^e(m, n)$ with the following equation:

$$y^e(m,n) = \begin{cases} 1, & d_{\mathcal{Z}}(x_m^v, x_n^v) \leq \delta \vee d_{\mathcal{Z}}(x_m^a, x_n^a) \leq \delta \\ 0, & d_{\mathcal{Z}}(x_m^v, x_n^v) > \delta \wedge d_{\mathcal{Z}}(x_m^a, x_n^a) > \delta \end{cases}. \tag{4.1}$$

Intuitively, this definition ascertains that when two audio fragments are semantically similar $F_a(x_m^a) \approx F_a(x_n^a)$, we assume that the corresponding video fragments in time also should be semantically similar, and vice versa. Conversely, two video fragments close in embedding space are hypothesized to have semantically close audio fragments. Hence, positive pairs can be constructed by mixing modalities of x_m and x_n. With mutual information between the video or audio embeddings, we can further identify the false negative pairs and consider them as positive. Consequently, these positive pairs can also be used to reduce the limitation of minimal sufficient representation.

4.3.3 Contrastive Loss with Multi-positive Pairs

Given that the encoders of both modalities are trained to predict similar feature representations for temporally aligned audio and video clips, they tend to focus on information present in both modalities. Contrastive loss guides the training of the encoders such that the video and audio embedding are close: $F_v(x_m^v) \approx F_a(x_n^a) \mid m = n$. This means that $F_v(x_m^v)$ learns to eliminate all information not present in x_m^a, and vice versa. As explained in the introduction, other audio segments may contain complementary information, but this will be eliminated in the representation of $F_v(x_m^v)$.

The main purpose of including multiple modalities however, is to complement each other, providing additional information when an object or person can not be observed in one of the modalities. This is particularly a problem for data pair generation that only relies on temporal alignment since it

allows only one positive pair for each time frame:

$$\forall x_n^v \in X, \exists! x_m^a \mid m = n. \tag{4.2}$$

To address this limit of minimal sufficient representation, we propose a modification to the conventional contrastive loss function that uses multiple positive pairs identified through embedding-based distance (Equation 4.1).

Different from temporal-based pair generation, the embedding-based pair generation mechanism does not solely rely on temporal information to create positive pairs with x_m^v. With the semantic similarity of the embedding space, the limitation due to Equation 4.2 may be reduced as there can be multiple positives for x_m^v:

$$|\{x_n^a \in X \mid y^e(m,n) = 1\}| \geq 1 \tag{4.3}$$

The upper bound on the mutual information between x_m^v and the union of all elements in the set in 4.3 is higher than the mutual information between x_m^v and x_m^a. As a result, including multiple positives will likely retain more information on x_m^v in the representation $F_v(x_m^v)$, which will benefit downstream task performance. Inspired by the loss function proposed in [47], we proposed a modified loss function to fit the multi-positive found by Equation 4.3.

$$\begin{aligned}\mathcal{L}^{EPG}(X) = \sum_{x_m^v \in X} \sum_{x_n^a \in X} \Big(& y^e(m,n) \cdot d_{\mathcal{Z}}(x_m^v, x_n^a)^2 + \\ & (1 - y^e(m,n) \cdot \max(\tau - d_{\mathcal{Z}}(x_m^v, x_n^a), 0)^2 \Big).\end{aligned} \tag{4.4}$$

The distance metric $d_{\mathcal{Z}}(x_m^v, x_n^a)$ measures the distance between x_m^v and x_n^a in the shared space $\mathcal{Z}$, which is bounded by a predefined constant τ in the second term of Equation 4.4. Note that, due to the symmetry between each modality of two semantically similar data pairs, Equation 4.4 is equivalent to pairing each visual modality (x_m^v) within the dataset with each audio (x_n^a). The distance function $d_{\mathcal{Z}}$ can represent any similarity metric. In this paper, we obtained the best results using a weighted combination of the Euclidean distance $\|\cdot\|$ and the cosine similarity $S_c(\cdot)$. As cosine similarity yields larger values for higher similarity; the distance function $d_{\mathcal{Z}}(\cdot)$ is defined as follows:

$$d_{\mathcal{Z}}(x_m^v, x_n^a) = \omega \cdot \|z_m^v - z_n^a\|^2 + (1 - \omega) \cdot (2 - (S_c(z_m^v, z_n^a) + 1)), \tag{4.5}$$

where ω is a value between 0 and 1 to control the ratio between the two distance functions. The use of ω is discussed in Section 5.4.4.

4.4 Experimental Setup

In this section, we describe the experiments conducted to evaluate the quality of the learned representations across various downstream tasks relevant to smart surveillance applications. The evaluation includes one supervised downstream task, namely audio-visual event detection, and two unsupervised tasks: anomaly detection and event query. For all tasks, the pseudo-Siamese network is first pretrained on the audio-visual data using the objective function of Equation 4.4 in a self-supervised manner. Subsequently, the weights of the audio and/or visual encoders were frozen and considered as fixed feature extractors while training a small network for each of the downstream tasks.

4.4.1 Implementation Details

For all experiments, we maintain consistent configurations for data preprocessing, network architecture, and training procedure. Task-specific variations are described in subsequent sections. The audiovisual dataset is first segmented into 1-second, non-overlapping recordings, each containing synchronized audio segments and video frames.

Video recordings are downsampled to 5 fps, and each frame is resized to 398×224 pixels. To align with the visual encoder's input specifications, each frame is further divided into two overlapping 224×224 crops. These two crops are treated as independent frames that map to the same audio segment, and are then jointly considered during the inference phase.

Audio segments are resampled to 44100 Hz and transformed into log-mel spectrograms, following the configuration outlined by Adapa [48]. This transformation utilizes a window size of 256, a hop length of 694, and a total of 128 bins.

The framework consists of a pseudo-Siamese network with a visual encoder F_v and an audio encoder F_a. F_v is built based on X3D-M [49], featuring one convolutional layer, four residual blocks, and a final classification layer. We take the implementation from PyTorchVideo but replace its last layer with a fully-connected layer with 512 neurons to align with the dimensions of the audio representations. F_a is implemented as a ResNet18 model, taking the log-mel coefficients as input. The parameter counts for F_v and F_a are 4.02 million and 4.16 million, respectively. Training specifics differ: F_v trains with a learning rate of $2e{-}4$ and a decay of $1e{-}5$, while F_a is trained with a learning rate of $1e{-}3$ and a decay of $1e{-}5$.

For F_v, we take the pre-trained weights of X3D-M, which is trained on

Kinetics-400, provided by PyTorchVideo. As for F_a, we train it from scratch by freezing the pre-trained F_v and exclusively training F_a. Subsequently, using a layer-wise learning approach [50], both encoders undergo iterative training.

The embedding space $\mathcal{Z}$ is designed to capture semantically meaningful information, which may not be guaranteed when training from scratch. To ensure robust initialization, we only consider a pair as positive when the two modalities are sampled from the same timestamps during the first epoch of training. In all subsequent epochs, training shifts to the embedding-based label $y^e(m, n)$, guided by the loss function $\mathcal{L}$. Throughout training, the weight parameter ω for the distance function in Equation 4.5 is kept constant at 2.5, and σ is set to 0. An ablation study of these parameters is provided in 5.4.4. As δ is the threshold to determine whether the distance in embedding space of a temporally misaligned data pair is smaller than a temporally aligned pair, we set the δ as the distance between the embeddings of the temporally aligned pair ($\delta = d_{\mathcal{Z}}(x_m^v, x_m^a)$). Thus, the threshold δ adapts dynamically during training based on the embedding space.

4.4.2 Supervised Tasks

After pretraining the audio and video encoders, they can be used as feature extractors for downstream tasks. The first task we consider is event localization, which aims to pinpoint specific predefined events within an audiovisual stream, such as the entry of a car. This task can be cast as a supervised learning problem by following the protocol outlined in [8]. Long recordings are divided into short clips, and the objective is to perform binary classification to determine whether a given clip contains the target event.

4.4.2.1 Dataset

Finding real-world surveillance datasets that contain video, audio and labels is challenging. Some established datasets, such as those used in Benfold and Reid [51], Ristani et al. [52], have been taken down due to privacy concerns. Other datasets primarily consist of very short clips gathered from diverse locations, often sourced from video-sharing platforms like YouTube [14, 53, 54]. While these datasets suffice for certain supervised tasks, such as violence detection, they fall short for our purpose of learning features in a semi-supervised manner over long audiovisual streams.

We decided to use the Toulouse Campus Surveillance Dataset (ToCaDa) [55] to validate our approach. The ToCaDa dataset encompasses two distinct scenarios, each captured by multiple cameras strategically positioned

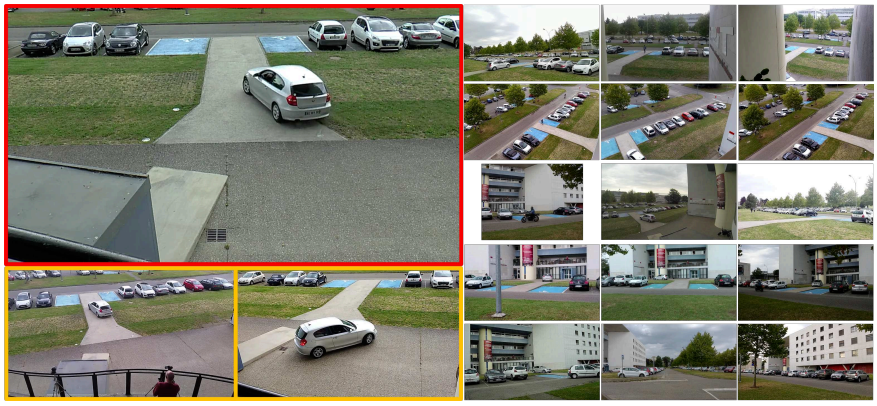


Figure 4.4: Camera setup for ToCaDa Scenario 1. The camera view used for *training* is marked in red, while the camera views that are used for *similar* are labelled as yellow. The rest of the camera views are used as *challenging*. Note that both similar sets and challenging sets are used for testing.

to record simultaneously audio and video. Some cameras have overlapping fields of view. The events in the videos are scripted to demonstrate a possible burglary involving 20 actors playing roles as pedestrians or suspects. Each video has an approximate duration of 300 seconds and comes with detailed annotations for both audio and video events. Most videos in this dataset contain only a limited number of events, making a meaningful split of each video across train and test set difficult. However, scenario one of the ToCaDa dataset contains a higher number of cameras observing the same scene, including some with slightly different viewpoints, see Figure 4.4. We therefore use the footage of Camera 2 as the training set for learning representations, and evaluate event localization on the audiovisual recordings from all other cameras.

4.4.2.2 Evaluation procedure

We evaluate the transferability of the learned representations to this supervised classification task by adopting the linear evaluation protocol from Wang et al. [20]. After pretraining, the weights of the feature extractors are frozen and a one-layer linear classifier is trained using cross-entropy loss as the objective function. After training the classifier, we evaluate performance using a segment-wise classification accuracy matrix, again following the protocol presented in [8].

4.4.2.3 Baseline Methods

We first evaluate our method by comparing it with existing audio-visual representation learning methods, as well as the *TPG* baseline. Additionally, since the data contains both audio and video, we compare the classification results of our method with visual-only and audio-only methods.

For multi-modal audio-visual representation learning, we compare our method with TACMA and MAVil. TACMA [8] is a self-supervised representation learning technique specifically designed for audio-visual event localization. TACMA employs a Barlow-Twins architecture to learn representations and includes a cross-modal attention module to enhance audio-visual information capture. However, because the cross-modal attention module is trained in a supervised manner, we exclude it and use only the AV-BT module to generate the representations. MAViL [30] is a recent method that has demonstrated strong performance in event classification tasks across general audio-visual datasets.

Both TACMA and MAViL are designed for datasets containing short videos with diverse content and scenarios. To ensure compatability with their training scheme, we segment the long ToCaDa training videos into 10-second, non-overlapping subclips. Unlike TACMA, MAViL relies on negative pairs to compute inter-modal contrastive loss. To enable this, we pair the audio and video from different subclips to create negative pairs for MAViL.

For the *TPG* baseline, we use the same training procedure as our method, with the only difference being the pair generation strategy.

For visual-only benchmarks, we employ the EfficientNet [56] and X3D [49] models. EfficientNet is pretrained on the ImageNet dataset, and we select the EfficientNet-B0 variant, which has 4.03 million parameters, making it comparable in scale to the F_v encoder in our method. For X3D, our choice is the X3D-M variant pretrained on Kinetics-400, which contains 3.76 million parameters. Both models and their pretrained weights are sourced from the PyTorch and PyTorchVideo repositories. To use these models as baselines, we remove their final classification layers and use the outputs of the remaining pretrained network as representations.

TACMA and X3D-M process input frames at a resolution of 256×256, while EfficientNet-b0 operates at 224×224. For fair comparison, we first downsize all frames to 224 *times*224 and then upsample them to 256×256 for use with TACMA and X3D-M.

For all other configurations, we follow the original preprocessing steps specified in the respective works, except for MAViL. Since the authors of MAViL

did not release their code or the pretrained model, we follow the implementation details and the pretrained model of the reproduction reported in [57].

For the audio-only benchmarks, we use the best-performing model from the DCASE19 urban sound tagging task [48]. The official implementation and pretrained weights were obtained from the author's GitHub repository [1]. Similarly to the visual-only benchmarks, we remove the classification layer from the pretrained network and use its output as feature representation for downstream evaluation.

4.4.3 Unsupervised Tasks

We also evaluate our framework in two unsupervised tasks commonly used in surveillance: anomaly detection and query-guided event search. The task of anomaly detection involves identifying inputs that deviate from normal behavior. Since the behaviors of interest are not defined beforehand, anomaly detection is a challenging task that requires high-quality input features to discern subtle deviations. The query-guided event search task is to locate events in a video similar to a given query event. For instance, if the query is a clip containing a joyriding car with distinct audio or visual characteristics, the task is to identify other timestamps in the recording where similar events occur.

4.4.3.1 Dataset

The ToCaDa dataset, while valuable for tasks like event localization, is less suited for anomaly detection and query-guided event search due to its limited video length and restricted diversity of actions occuring. Similarly, widely-used datasets with annotated anomalies, such as Avenue [58] and ShangHaiTech [59], contain only visual cues. As an alternative, we collected real-world audio-visual surveillance footage from a publicly available live stream on YouTube[2][3]. This audiovisual stream captures a main intersection in Tokyo's Shinjuku district, observed from a a high vantage point, providing a representative setting for urban surveillance. Some stills from the recordings are shown in Figure 4.1, showcasing different types of vehicles, bikes and pedestrians with their accompanying sounds. For in-depth evaluation, we recorded four 4-hour-long videos from two different dates under different lighting conditions. The videos are recorded during two time

[1] https://github.com/sainathadapa/dcase2019-task5-urban-sound-tagging
[2] https://www.youtube.com/watch?v=2gZySUir8_w$
[3] https://www.youtube.com/watch?v=xiLF6PmFZP4$

windows: 15:00 to 19:00 (daytime) and 19:00 to 23:00 (nighttime), on a Tuesday and a Thursday.

4.4.3.2 Evaluation procedure

Since the Tokyo dataset is not annotated, we conduct a qualitative evaluation between different methods of anomaly detection and query-guided event search.

The anomaly score for a clip $x_m = (x_m^v, x_m^a)$ is calculated using the distance function of Equation 4.5. A clip is flagged as anomalous if the score is higher than a threshold. For each 4-hour-long video, a separate model is trained to learn audio-visual representations. The threshold for anomaly detection is dynamically adapted for each video and set to $\mu + 2\sigma$ where μ and σ represent the mean and standard deviation of the anomaly score on the training set. This threshold considers 95.45% of the training data as normal.

For query-guided event search, separate models are trained on each video to extract features. Query events have been selected manually. Events were searched in all recordings, but events within a window of 1 minute before and after the selected event are excluded as search results. We rank the search results based on the distance between the query and the results in the embedding space.

4.4.3.3 Baseline Methods

For the anomaly detection task, we compare our method with five other approaches: two multi-modal fusion models [10](*Fusion*) and [30](MAViL), one vision-only approach based on the X3D-M model, one audio-only baseline Adapa [48] and the multi-modal *TPG* baseline.

As for the event search task, apart from the multi-modal *TPG* and MAViL[30] baselines, we compare our approach to a baseline that involves a straightforward fusion approach in which video and audio features of separately trained encoders are concatenated. Specifically, we concatenate visual features from X3D-M [49] with audio features from Adapa [48]. We refer to this baseline as $(A+V)$.

4.5 Experimental Results

4.5.1 Audio-visual Event Localization

The experimental results for audio-visual event localization are summarized in Table 4.1, which reports the accuracy score for each method. The ToCaDa dataset consists of videos recorded from 18 different cameras. We trained on data from camera 2 and tested on all other cameras. To better assess the robustness of the learned features, we categorize the test cameras into two groups: (1) cameras with a view similar to the training camera (labeled as "similar"), and (2) cameras with distinct perspectives compared to the training camera (labeld as "challenging"). By reporting results separately, we ensue that the evaluation reflects the model's generalization ability without artificially inflating accuracy due to overlapping cameras.

	Method	# Params.(M)	Similar	Challenging
Audio-Visual	TACMA [8]	150.46	77.41	67.20
	MAViL [30]	185.66	72.69	62.34
	TPG	15.2	71.33	60.47
	EPG(Ours)	15.2	86.92	77.48
Visual	EfficientNet*	4.01	85.65	73.20
	X3D-M*	2.97	90.21	71.84
	TACMA [8]	52.87	62.15	43.60
	MAViL [30]	85.74	65.23	51.44
	TPG	4.02	60.71	44.92
	EPG(Ours)	4.02	85.42	70.02
Audio	Adapa* [48]	4.16	85.22	82.47
	TACMA [8]	97.58	76.43	65.61
	MAViL [30]	85.74	77.24	62.30
	TPG	4.16	70.66	60.43
	EPG(Ours)	4.16	81.37	79.48

Table 4.1: Audio-visual event localization results on two subsets of ToCaDa dataset. * denotes the model is pretrained on a large-scale dataset without any fine-tuning. The numbers are segment-wise classification accuracy following the protocol in TACMA [8].

The results highlight a significant performance gap between the two camera sets, with all methods achieving considerably higher accuracy on the "similar" set. These results corroborate findings in earlier research on the advantages of using location-specific methods for analyzing surveillance data [7].

When comparing our approach with the other audio-visual techniques TACMA

and MAViL, our approach demonstrates a substantial performance advantage. This difference can likely be attributed to the fact that these models are designed for more general purposes that require training on large-scale datasets with diverse content. Due to the more constrained nature of the ToCaDa dataset, these models struggle to learn sufficiently representative embeddings for event localization tasks. Examing the performance difference between TACMA and MAViL, we observe that while MAViL achieves great performance in general representation learning, its reliance on both reconstruction and contrastive loss with negative pairs defined based on temporal alignment, pose limitations in this dataset. In contrast, TACMA employs a Barlow-Twins architecture, which avoids the need for negative pairs, and only consider limited positive pairs.

Notably, when our model architecture is trained using the *TPG* pair generation strategy instead of the *EPG* strategy, a significant drop in accuracy is observed. This results highlights the advantages of accounting for semantic similarity in pair selection.

We also compare the classification results obtained using the video representations learned by our approach, TACMA and MAViL, with those from video-only based models. Our approach performs similar to the EfficientNet model, which was pretrained on the large scale ImageNet dataset. Th best performance is achieved by the pretrained X3D-M model, which is expected given its architecture's specific design for capturing motion information. In contrast, the self-supervised methods TACMA, MAViL, and *TPG* perform poorly in this task. MAViL's weaker performance can be attributed to its reliance on reconstruction loss, which is susceptible to foreground-background imbalance. This imbalance makes MAViL more sensitive to differences in the scen. These results demonstrate that the feature representations obtained from our multimodal approach also transfer effectively to purely visual tasks.

For audio-only event localization, we observe a smaller drop in accuracy between the "similar" and "challenging" locations compared to video-only localization. This underscores the robustness of audio data, which is less sensitive to the exact positioning of sensors. The best results in this setting are obtained by the pretrained Adapa model. While our model performs slightly worse than Adapa, it achieves this without the need for pretraining on extensive labeled data. Furthermore, our proposed *EPG* approach consistently outperforms the *TPG* baseline, reinforcing the effectiveness of embedding-based pair selection in enhancing the quality of learned representations.

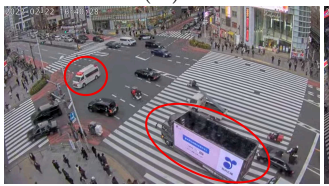
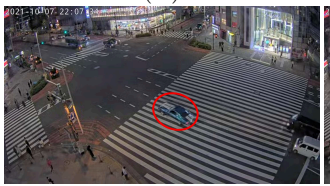
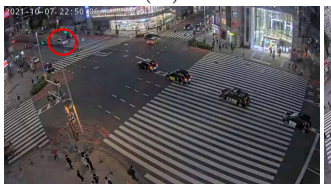
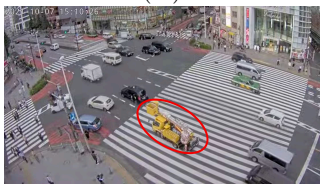
	(A)	(B)	(C)	(D)
				
Fusion	✓			
MAViL	✓	✓	✓	✓
TPG	✓	✓	✓	✓
EPG(Ours)	✓	✓	✓	✓
Audio	✓	✓	✓	
Visual	✓			✓

Table 4.2: Examples of anomalous events. Check marks indicate which models flagged this event as anomalous. (A) An advertising truck waits at the crossroad while an ambulance passes with its siren wailing. (B) A sports car speeds by with the engine roaring. (C) An ambulance enters from the upper left corner. (D) A forklift passes through without any distinct noise.

4.5.2 Anomaly Detection

Since the Tokyo dataset lacks annotations, we perform only a qualitative evaluation of anomaly detection performance. Figure 4.2 presents examples of anomalous events and indicates which methods were able to detect them. The shown events represent the four semantically distinct events with the highest anomaly score. These examples cover events with distinctive sounds, unique visual appearance, or a combination of both. This diversity demonstrates that the learned embedding space is semantically meaningful and contains information to identify various anomaly events.

In example (A), an advertising truck waits at the crossroad while an ambulance with wailing siren passes by. This anomaly is clearly identifiable through both visual and audio data. In example (B), a sports car speeds by with the engine roaring. While the sound of the sports car is highly distinctive, the car's visual appearance is not particularly notable. The Fusion baseline [10], which relies on multimodal data, fails to flag this event as an anomaly. In example (C), an ambulance arrives from the topleft corner with sirens on, then turns right and exits in the bottom left corner. In this case, the ambulance is not visually prominent but is clearly audible. Example (D) shows a yellow forklift passing through the intersection. There is no distinctive engine sound, leading audio-only methods to miss this anomaly.

These experiments show that by mapping audio and visual to the same embedding space, we can learn representations that effectively integrate information from both modalities. This enhances the ability to detect anomalies that are challenging to identify using a single modality.

4.5.3 Event Search

In our last set of experiments, we present examples of event search on the Tokyo dataset, as illustrated in Figure 4.5. The top row shows still frames of the (manually selected) query clips, while the following rows showcase the most related events identified by each method, shown in increasing order of embedding distance. We excluded a one-minute window before and after each query event from being searched. Similarly, for each method, we show only search results that are at least one minute before or after higher ranked search results.

Example (A) shows a pink bus entering the scene while the sound of a passing train can be heard in the background, though the bus itself produces no distinctive sound. All methods successfully locate a similar event where the same pink bus appears at a different time, again accompanied by the sound of a train in the background. Notably, the overlap between the bus's

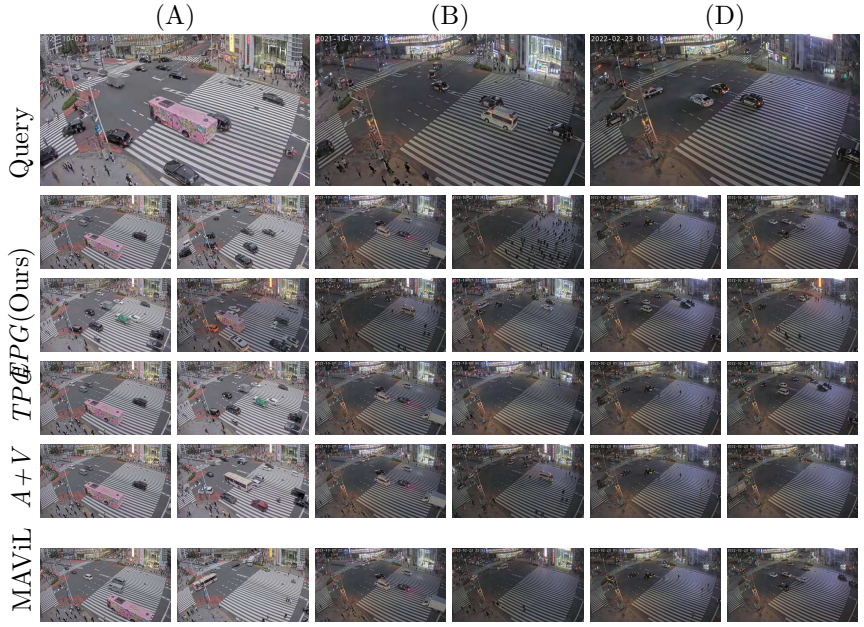


Figure 4.5: Examples of event query. The top row shows the query video, while the search results are shown in the 2-6 rows. The results of each methods are shown in a decreasing order from left to right and top to down. (A) A pink bus. (B) An ambulance with sirens on. (C) A police car with broadcast and siren on. We show the top 4 search results obtained with our method for each query events. For the other three methods, only the top two search results are reported in order not to overload the figure.

visual presence and the train's audio is brief. While the other approaches prioritize frames with visual similarity over audio similarity, MAViL selects a frame emphasizing the distinct train sound. In contrast, our method shows only a 1 % difference in preference between audio-similar and visual-similar frames. Moreover, our approach identifies an additional instance of the same bus later in the surveillance stream, this time without a train in the background. Other detections from all methods are quite diverse but typically contain either a bus, multiple black cars or the sound of a train in the background.

In example (B), an ambulance enters with its siren wailing accompanied by a broadcast announcement. Besides the passage in the query clip, the ambulance appears at least six more times during the night. All methods detect one particular instance where the ambulance enters from the bottom right corner and heads in the same direction as in the query clip. Interestingly, this event is detected earlier by *EPG* than by *TPG*, even before the ambulance was visible. This shows that even though *EPG* and *TPG* both integrate audio and video information, *EPG* is more adept at fusing both modalities, likely due to its training with embedding-based pair generation. By positively pairing segments where the ambulance is audible with segments where it is visible, *EPG* improves its ability identify such events. Furthermore, in two other appearances the ambulance follows a different trajectory, entering from the top left and turning right. Only *EPG* and *TPG* successfully detect these cases. Additionally, there is one other instance where the broadcast audio is present, detected only by *EPG* and MAViL.

In example (C), a police car enters the scene with loud sirens and flashing warning lights. Although the car itself is rather small, its visual and auditory features make it distinct from other vehicles. *EPG*, *TPG*, and MVAiL all locate another similar event. Again, *EPG* detects the event earlier, even before the car is visible. All three methods also successfully identify other occurrences where police cars drive by or sirens are aubible in the background. However, the baseline method ($A+V$) is unable to find a matching event for this example. The examples of the ambulance and the police car demonstrate that our proposed embedding based pair generation and custom loss function help to keep more information and cover more aspects of the event, which benefits the real-world application.

Although this qualitative evaluation is limited in scope, it is important to emphasize that these results are derived using the same representations applied in the anomaly detection task. This highlights the versatility and robustness of the learned embeddings across multiple tasks.

σ(sec)	0	1	10	30	60	120
Similar	86.92	86.21	85.32	87.41	87.92	86.31
Challenging	77.48	78.62	77.33	79.82	79.82	78.32

Table 4.3: Ablation study on different σ. The impact of different σ on audio-visual event localization results.

ω	0	0.25	0.5	0.75	1
Similar	DNC	85.21	86.92	85.44	DNC
Challenging	DNC	75.16	77.48	76.83	DNC

Table 4.4: Ablation study on different ω. DNC stands for *Did not converge*. The impact of different ω on audio-visual event localization results.

4.5.4 Ablation Study

To have a better understanding of our method, we analyze the results for different values of the hyperparameters used in our approach, namely δ, σ, and ω. Since δ adapts based on the similarity of positive pairs, we focus our evaluation on σ and ω, which represent the temporal constraints and the weight between the distance functions, respectively. Since the only annotated task in our experiments is audio-visual event localization, we restrict our ablation study tothis task.

4.5.4.1 Temporal Constraint σ

As shown in Table 4.3, the value of the temporal constraint σ has minimal impact on our approach. Starting from the second epoch, the embedding-based pair mechanism is introduced, which does not solely rely on the time difference but also considers the semantic similarity in selecting pairs. This dual mechanism offers flexibility, as the optimal temporal constraint σ can vary based on the content of the data. For instaces, two frames that are 1 minute apart in a scene depicting a sidewalk might contain very similar content, whereas the same time gap in a highway scene could result in significantly different content. Given this variability, the robustness of *EPG* shows another advantage in the real-world scenario.

4.5.4.2 Weight in Distance Function ω

Table 4.4 shows the effect of different values for ω, which is used in Equation 4.5. Both cosine similarity and Euclidean distance are commonly used as distance metrics in contrastive learning [47, 60]. When ω is set to 0.25, 0.5,

or 0.75, there is no significant impact on thee performance of audio-visual event localization. However, when only one of the distance metrics is used ($\omega = 0, \omega = 1$), we observe that the model sometimes fails to converge. We investigated the learning process and hypothesize why the instability arises when using only Euclidean distance or cosine similarity. On the one hand, as the pre-trained visual encoder F_v is not normalized to a unit vector, we did not normalize the output of F_a either. In the early stages of the training, using only cosine similarity occasionally leads to the model collapsing or diverging. Cosine similarity does not reflect the magnitude of the vector, which can be a crucial factor in measuring the spatial correlation between two data points in a non-unified embedding space.

In contrast, Euclidean distance provides more efficient guidance in training the F_a more efficiently. However, we found that a poor choice of the margin constant also leads to model collapse or divergence when using only Euclidean distance. Since the optimal margin for Euclidean distance can vary depending on the data, incorporating cosine similarity helps balance the loss function, leading to more stable training. Moreover, Euclidean distance complements cosine similarity by considering the absolute difference between two vectors. The combination of both metrics helps in identifying semantically similar events in the embedding space, as it captures both directional and magnitude-based relationships.

4.6 Conclusion and Future Work

In this paper, we discussed the challenges of learning audio-visual representations from multi-modal surveillance data. We addressed the limitations of relying solely on temporal alignment as pretext task, as well as the minimal sufficient representation bottleneck inherent in contrastive learning. To the best of our knowledge, this is the first study to explore these issues in the context of surveillance data.

We introduced a novel embedding-based pair generation mechanism that mitigates the problem of false negative pair generation while promoting more diversity in positive pairs. Our pseudo-Siamese network, enhanced by a new contrastive loss function that accounts for multiple positive pairs, learns more effective audio-visual representations.

We evaluated the generalization of the learned representations across various downstream tasks and compared our approach to state-of-the-art approaches using a publicly available dataset. Additionally, we demonstrated the effectiveness of our method on a more challenging dataset of real-world

surveillance data. Our results show that our approach performs similar or better than existing state-of-the-art techniques.

However, the evaluation process itself highlights a key limitation discussed in Section 1.3.2.3, Chapter 1: unsupervised learning models are paradoxically validated using annotated datasets. The difficulty of obtaining such datasets largely restricts the development of unsupervised representation learning for real-world data. This limitation motivates the work in the following chapter, which focuses on developing a model performance assessment framework for surveillance data, without the access of training data and data annotations.

In future work, we will further explore how the learned representations perform on different downstream tasks. We will also investigate techniques to make the model smaller and faster. Both audio and video data are potentially privacy sensitive and should not leave the local edge device unless absolutely necessary. To facilitate this, we aim to optimize the model for real-time operation on resource-constrained platforms, enabling scalable deployment while preserving privacy.

Funding

Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

Data Availability Statement

ToCaDa dataset can be downloaded from Zenodo. The Tokyo datasets are publicly available on Youtube, with the following link: [Archive] Tokyo-Shinjuku LiveCam Tue and [Archive] Tokyo-Shinjuku LiveCam Thu .

4.7 References

[1] UN DESA. 68% of the world population projected to live in urban areas by 2050, says UN. *United Nafions Department of Economic and Social Affairs*, 2018.

[2] Charlie Mydlarz, Justin Salamon, and Juan Pablo Bello. The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics*, 117:207–218, 2017.

[3] Dragana Bajovic, Arian Bakhtiarnia, George Bravos, Alessio Brutti, Felix Burkhardt, Daniel Cauchi, Antony Chazapis, Claire Cianco, Nicola Dall' Asen, Vlado Delic, et al. Marvel: Multimodal extreme scale data analytics for smart cities environments. In *2021 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 143–147. IEEE, 2021.

[4] Qi-Chao Mao, Hong-Mei Sun, Ling-Qun Zuo, and Rui-Sheng Jia. Finding every car: a traffic surveillance multi-scale vehicle object detection method. *Applied Intelligence*, 50:3125–3136, 2020.

[5] Fath U Min Ullah, Mohammad S Obaidat, Amin Ullah, Khan Muhammad, Mohammad Hijji, and Sung Wook Baik. A comprehensive review on vision-based violence detection in surveillance videos. *ACM Computing Surveys*, 55(10):1–44, 2023.

[6] Jisheng Bai, Jianfeng Chen, and Mou Wang. Multimodal urban sound tagging with spatiotemporal context. *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[7] Sam Leroux, Bo Li, and Pieter Simoens. Automated training of location-specific edge models for traffic counting. *Computers and Electrical Engineering*, 99:107763, 2022.

[8] Yue Ran, Hongying Tang, Baoqing Li, and Guohui Wang. Self-supervised video representation and temporally adaptive attention for audio-visual event localization. *Applied Sciences*, 12(24):12622, 2022.

[9] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022.

[10] Pratibha Kumari and Mukesh Saini. An adaptive framework for anomaly detection in time-series audio-visual data. *IEEE Access*, 10: 36188–36199, 2022.

[11] Baej Leporowski, Arian Bakhtiarnia, Nicole Bonnici, Adrian Muscat, Luca Zanella, Yiming Wang, and Alexandros Iosifidis. Audio-visual dataset and method for anomaly detection in traffic videos. 2023.

[12] Bharti Munjal, Sikandar Amin, Federico Tombari, and Fabio Galasso. Query-guided end-to-end person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 811–820, 2019.

[13] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[14] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[15] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021.

[16] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023.

[17] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[18] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[19] Y-H Tsai, Y Wu, R Salakhutdinov, and L-P Morency. Self-supervised learning from a multi-view perspective. In *Proceedings of the International Conference on Learning Representations (ICLR), 2021*, 2021.

[20] Haoqing Wang, Xun Guo, Zhi-Hong Deng, and Yan Lu. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16041–16050, 2022.

[21] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[23] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020.

[24] Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, and Wanli Ouyang. Exploiting visual context semantics for sound source localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5199–5208, 2023.

[25] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020.

[26] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[27] Junyu Gao, Hao Yang, Maoguo Gong, and Xuelong Li. Audio–visual representation learning for anomaly events detection in crowds. *Neurocomputing*, 582:127489, 2024.

[28] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017.

[29] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021.

[30] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Abudukelimu Wuerkaixi, You Zhang, Zhiyao Duan, and Changshui Zhang. Rethinking audio-visual synchronization for active speaker detection. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01–06. IEEE, 2022.

[32] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022.

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[34] Peihao Wu, Wenqian Wang, Faliang Chang, Chunsheng Liu, and Bin Wang. Dss-net: Dynamic self-supervised network for video anomaly detection. *IEEE Transactions on Multimedia*, 2023.

[35] Vignesh Sampath, Iñaki Maurtua, Juan Jose Aguilar Martin, and Aitor Gutierrez. A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of big Data*, 8:1–59, 2021.

[36] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=S1v4N2l0-`.

[37] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

[38] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[39] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[40] Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[42] Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8220–8230, 2022.

[43] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809, 2020.

[44] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10306–10315, 2021.

[45] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[46] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In *European conference on computer vision*, pages 548–564. Springer, 2020.

[47] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[48] Sainath Adapa. Urban sound tagging using convolutional neural networks. *arXiv preprint arXiv:1909.12699*, 2019.

[49] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[50] Eugene Belilovsky, Michael Eickenberg, and Edouard Oyallon. Greedy layerwise learning can scale to imagenet. In *International conference on machine learning*, pages 583–593. PMLR, 2019.

[51] Ben Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. *CVPR 2011*, pages 3457–3464, 2011.

[52] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

[53] Mauricio Perez, Alex C. Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666, 2019. doi: 10.1109/ICASSP.2019.8683676.

[54] Armin Danesh Pazho, Ghazal Alinezhad Noghre, Babak Rahimi Ardabili, Christopher Neff, and Hamed Tabkhi. Chad: Charlotte anomaly dataset. In *Image Analysis*, pages 50–66, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-31435-3.

[55] Thierry Malon, Geoffrey Roman-Jimenez, Patrice Guyot, Sylvie Chambon, Vincent Charvillat, Alain Crouzil, André Péninon, Julien Pinquier, Florence Sèdes, and Christine Sénac. Toulouse Campus Surveillance Dataset: scenarios, soundtracks, synchronized videos with overlapping and disjoint views, March 2020. URL `https://doi.org/10.5281/zenodo.3697806`.

[56] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[57] Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, et al. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6890–6894. IEEE, 2024.

[58] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[59] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[60] Donghuo Zeng, Yi Yu, and Keizo Oyama. Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–23, 2020.

5

Source-Free Model Transferability Assessment

My can opener came in a sealed plastic shell I couldn't open without a can opener.

- Common anonymous joke about modern packaging.

Source-Free Model Transferability Assessment for Smart Surveillance via Randomly Initialized Networks

Wei-Cheng Wang • Sam Leroux • Pieter Simoens
Published in MDPI Sensors, Special Issue on AI-Based Computer Vision Sensors & Systems, 2025

In the field of smart surveillance, deep learning models are often deployed without validation on annotated data from the target domain due to the scarcity of such labeled datasets. This chapter proposes a method for assessing model performance in the absence of this labeled data.

We focus on a practical use case in smart surveillance: evaluating the transferability of pretrained models for adaptation to new domains. To this end, we introduce a novel approach that leverages the embedding space generated by a collection of randomly initialized neural networks (RINNs). This method allows us to assess how well pretrained models capture the structural patterns of an unannotated dataset without requiring access to the original training data. Our approach aims to lay the groundwork for advancing unsupervised learning techniques beyond the limitations imposed by dataset availability.

Furthermore, this chapter empirically confirms the presence of data drift, as introduced in Section 1.3.1.2, Chapter 1, and presents a preliminary strategy for fine-tuning models to specific locations as a means to address this challenge.

5.1 Introduction

Visual surveillance is a foundational component of smart city infrastructure, supporting real-time applications such as traffic management and anomaly detection. These cameras operate under diverse configurations, varying in viewpoint, angle, elevation or location, as illustrated in Figure 5.1. This diversity makes it infeasible to rely on a single machine learning model that performs well across all environments. Instead, prior work has shown that camera-specific models are more effective and efficient in such settings [1, 2].

Training a unique model from scratch for each surveillance camera is often infeasible due to the high computational cost and the vast amount of devices. For example, training a single model can require eight hours on a Tesla V100 GPU, and a city such as London has over 940,000 cameras [3]. An alternative strategy is to design lightweight models tailored for a specific task

Figure 5.1: Examples of different camera configuration.

while reducing computational requirements [4]; such specialized models may not be readily available for all target tasks, particularly when the available data are unannotated. Consequently, a common strategy to obtain these location-specific models is to fine-tune pretrained models [5]. However, this approach is sensitive to domain shift: if the source and target domains differ significantly, model performance can degrade substantially [6]. The selection process itself is further complicated by the unreliability of reported performance metrics. For instance, a study on real-world systems found a discrepancy as high as 44% between a model's claimed accuracy and its measured operational performance [7]. These factors highlight the critical need for methods that can estimate a model's effectiveness on specific target data prior to deployment.

Estimating the transferability of a pretrained model to a new environment is inherently challenging as it involves quantifying the distributional shift between source and target data. In practical scenarios such as smart surveillance, access to the original training data is often restricted due to privacy concerns. While the European Union promotes increased data sharing across industrial and public sectors to drive innovation [8], the direct exchange of raw data often conflicts with GDPR [9] or risks exposing proprietary information. One solution to prevent data violation is to limit access and provide encryption on the stored data [10]. An alternative solution, the source-free setting, which shares trained models instead of data, is increasingly regarded as a privacy-preserving alternative for enabling collaboration without disclosing sensitive datasets. In addition, no labeled data are typically available for the target domain. This source-free, unsupervised setting [11] significantly complicates transferability assessment, as it precludes direct comparison between source and target domains and limits the use of traditional domain adaptation techniques.

Figure 5.2 further illustrates this concept of source-free unsupervised domain adaptation (SFUDA) in smart surveillance settings. First, a model zoo is formed of models pretrained on data collected under diverse configurations. These models could have different architectures, training data or

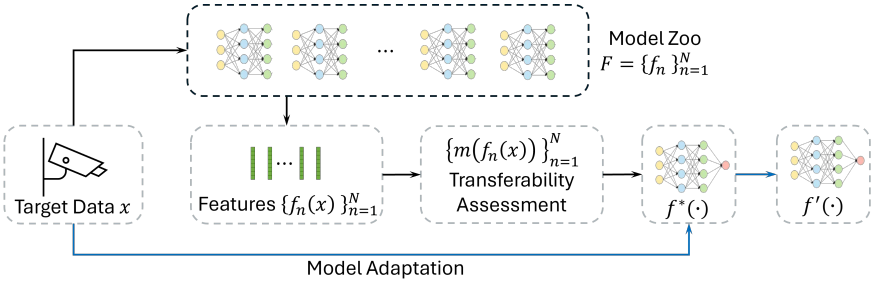


Figure 5.2: A high-level overview of source-free unsupervised domain adaptation. The black lines describe the process of transferability assessment, while the blue line describes the data flow of the model adaption. Target data is first fed into a set of pretrained models $F = \{f_n\}_{n=1}^N$. A transferability assessment is then applied to the representations $\{f_n(x)\}_{n=1}^N$ to rank the applicability of each model. Finally, the highest ranked pretrained model $f^*(\cdot)$ is adapted to target-specific mode $f'(\cdot)$ with the target data.

training procedures. When a new camera is installed in the system, a configuration procedure finds the model from the model zoo that best fits the new scene. For this, each of the models processes a small amount of camera data and the transferability assessment is used to rank them, identifying the pretrained model that is most adaptable to the target domain. Finally, the highest-ranked model is either deployed directly or after further fine-tuning on target data for a specific camera. Crucially, this procedure requires no access to the original training data of the models in the model zoo, nor does it require labeled information for the target task that will be performed on the new data (i.e., it operates in the source-free unsupervised setting).

We propose a novel approach to transferability assessment based on randomly initialized neural networks (RINNs). Unlike existing methods that rely on assumptions such as reliable model uncertainty or the generalizability of pretrained embedding spaces, assumptions which often fail in real-world surveillance scenarios, our method leverages RINNs to construct a task-agnostic, data-independent reference embedding space. Transferability is then estimated by computing the similarity between embeddings produced by each pretrained model and those from the RINN ensemble, using minibatch-Centered Kernel Alignment (CKA) for scalability and efficiency. Despite lacking training, RINNs have been shown to capture meaningful structural priors through architectural depth and compositional nonlinearity [12, 13]. This enables robust model selection without requiring access to source data, target labels, or unreliable uncertainty estimates. In the context of practical applications, our framework is designed for an of-

fline model selection phase, which would typically be executed on a server or cloud infrastructure. Following this one-time assessment, the single best-identified model could then be further fine-tuned for a specific target if needed, and subsequently deployed for efficient, real-time inference on resource-constrained edge devices at the surveillance site.

Through experimental evaluations on four unlabeled source video datasets (Tokyo Intersection [14], Tokyo Street [15], and Agdao-Market/Street [16]) and three public target sets (Street Scene [17], NWPU Campus [18], and Urban Tracker [19]), we demonstrate that this is a cost-efficient solution for source-free, unsupervised domain adaptation in smartcity environments.

To summarize, this research confronts the critical challenge of assessing model transferability under source-free and unsupervised conditions, a problem of practical importance for deploying machine learning models in real-world scenarios such as smart city surveillance. To the best of our knowledge, this specific challenge remains underdeveloped, with very few prior works addressing it directly. Our main contributions to this area are as follows:

- We introduce a novel transferability assessment framework designed for the aforementioned challenging conditions, which effectively addresses the diverse camera configurations and significant domain shifts inherent in smart surveillance applications.
- The proposed framework uniquely employs ensembles of randomly initialized neural networks (RINNs) to create a task-agnostic reference embedding space. This approach avoids biases inherent in using pre-trained models for the reference itself and enables model assessment without prior knowledge of the specific downstream task or objective.
- We developed an embedding-level score (S^E) by comparing structural similarities in data representations, reflecting intrinsic data characteristics. This approach yields more robust, task-agnostic transferability estimates, avoiding the instabilities tied to pseudo label-based methods or the reliance on task similarity, particularly under shifting downstream objectives.
- Comprehensive empirical validation on three public surveillance datasets (Street Scene, NWPU Campus, and Urban Tracker, covering 48 diverse scenes) and three downstream tasks (object tagging, anomaly detection, and event classification) demonstrates our S^E metric's efficacy. Results confirm S^E consistently identifies the most transferable models, proving its practical utility for identifying the most suitable model for adaptation within these challenging real-world surveillance

scenarios.

The remainder of the paper is structured as follows. Section 5.2 introduces past related work on transferability assessment, randomly initialized neural networks, embedding similarity and source-free unsupervised domain adaptation. In Section 5.3, we describe the proposed framework, the construction of the reference embedding space using RINNs, and the definitions of the two novel transferability scores. The evaluation process, including the details of the model zoo, datasets, downstream tasks, and implementation specifics are also explained in Section 5.3. Then, the experimental results and ablation study are presented and discussed in Section 5.4. Finally, we conclude the findings and discuss future research directions in Section 5.5.

5.2 Related Work

5.2.1 Transferability Assessment

Transferability assessment provides an estimation of how a model would perform on new data. Early works estimate transferability through partial fine-tuning on the target task. While informative, this method still incurs considerable computational cost and requires fine-tuning [20]. Moreover, the need for access to annotated target data further limits the scalability of such methods in practical settings.

More efficient methods for transferability estimation methods have been proposed, each operating under varying assumptions regarding data accessibility. One of the earliest examples, Tran et al. [21], estimates task transferability via conditional entropy. However, this method assumes access to both source data and target annotations and presumes similarity between the source and target distributions. To relax the dependence on source data, several works such as Nguyen et al. [22], You et al. [20], Ding et al. [23], and Xu and Kang [24] estimate transferability by analyzing pre-trained source representations applied to the target domain. The recent LEAD framework [25] proposes a transferability metric by modeling the evolution of a model's output logits during fine-tuning. Using a single run, it captures the initial logits and their gradients to construct an ordinary differential equation (ODE), which estimates how the logits would evolve toward their final state. This theoretical trajectory is then used to assess how well the model would adapt to the target domain, which serves as an indicator for model selection. Although these methods do not need access to the source data, offering improved privacy and reduced demands on data storage and computation, they still require annotated target data. This

limitation makes these methods hard to apply in real-world surveillance applications, where annotations are typically unavailable. Pei et al. [26] propose uncertainty distance (UD), which measures the transferability without accessing the source data and the annotation of the target data. Assuming a low model uncertainty, they calculate the distributional uncertainty with a probabilistic framework that leverages the source model's predictions on target data. In contrast, Ensemble Validation (EnsV) [27] avoids dependency on a single model and instead utilizes the joint prediction from an ensemble of models in the model zoo to form a reliable proxy for the ground truth. While this ensemble-based approach provides general stability, its performance can be suboptimal if the proxy itself is unreliable, which can occur when the majority of models in the zoo were not suitable for the target data.

While these works aim to support transferability assessment in unsupervised source-free adaptation pipelines, they still rely on pretrained models and their behavior on target data. In contrast, we propose a transferability assessment that does not depend on the embedding of pretrained models. Instead, we measure the structural consistency of target data by comparing embeddings of a given model with those from randomly initialized networks.

5.2.2 Randomly Initialized Neural Network

The potential of randomly initialized neural networks (RINNs) was first highlighted by Frankle and Carbin [28] through the Lottery Ticket Hypothesis (LTH). Their work posits that dense, untrained neural networks contain sparse subnetworks which, when trained from the same initialization, can achieve performance comparable to that of fully trained networks.

Building on this foundation, Ramanujan et al. [12] demonstrated that sufficiently over-parameterized RINNs contain subnetworks that match the performance of fully trained networks without any training, a claim now referred to as the *strong* LTH. Malach et al. [13] provide a theoretical proof of the stronger LTH, showing that, for any bounded distribution and a target network with bounded weights, a sufficiently large RINN contains a subnetwork capable of achieving comparable performance to the target network. However, their findings also indicate that the over-parameterized RINN must be sufficiently large to satisfy these conditions, thereby imposing significant memory requirements.

To address the inefficiency caused by such over-parameterization, Chijiwa et al. [29] proposed Iterative Randomization (IteRand), a framework designed to improve parameter efficiency in pruning-based settings.

In contrast to the works described above, which focus on extracting or pruning such subnetworks from RINNs for direct deployment or fine-tuning, our approach adopts a different perspective. Rather than identifying so-called "winning tickets", we assume that pruned or weakly activated neurons are either uninformative or adversely affected by domain shift. Consequently, our method concentrates on evaluating embedding consistency across models, without modifying or optimizing the RINN itself. This enables us to exploit the inherent structural properties of RINNs while avoiding the computational burden of subnetwork extraction or training.

5.2.3 Embedding Similarity as a Proxy for Transferability

Another important line of research investigates the direct comparison of representations between models, without considering the raw data or annotations. Prior studies have demonstrated that the representations learned by neural networks encode task-relevant structural information, and that the degree of similarity between these representations can provide meaningful insights into model behavior and transfer potential. Based on their computational mechanism, Klabunde et al. [30] categorize these methods into six types. To estimate the similarity between different representations in our scenario, we particularly focus on the three types that meet the following two key technical requirements: the ability to compare representations with different dimensions, and the scalability to large datasets.

Alignment-based methods aim to find an optimal transformation that minimizes the difference between one representation and the transformed version of another. For example, Williams et al. [31] apply principles from statistical shape analysis to define novel similarity metrics on embeddings. On the other hand, **CCA-based** methods find a set of weights for each dimension of both embeddings such that the weighted embeddings have the maximal correlation. While the CCA-based methods can deal with misalignment between dimensions, they are known to suffer from high computational cost in high-dimensional settings, which can be a serious issue in smart surveillance. A recent work, Tuzhilina et al. [32], proposes several strategies to reduce this computational overhead, such as utilizing structured regularization and the kernel trick. Finally, **RSM-based** methods first compute a pairwise similarity matrix within the embedding space of each representation, where the pairwise similarity matrices of the two embeddings are then compared. A well-known method in this category is Centered Kernel Alignment (CKA) [33].

For our primary similarity metric, we adopt CKA, a powerful and widely-

used benchmark shown to outperform CCA-based methods in capturing functional similarity and is flexible enough to handle representations of different dimensionalities. However, prior research also shows its limitations. Research summarized by Klabunde et al. [30] highlights that CKA can be insensitive to certain structural changes in representations and can be influenced by data manipulations that do not affect a model's function. For instance, CKA can be insensitive to the removal of functionally important principal components and can be disproportionately affected by simple manipulations of single data points that do not change the model's overall function. On the other hand, Hayne et al. [34] provide crucial insight here, demonstrating that CKA is a significantly better predictor of linear decodability (a proxy for available information) than it is of performance on a single, fixed network. While this may be a limitation when analyzing a single task, it is a distinct advantage for our goal of assessing general transferability. We seek to identify which pretrained embeddings are most versatile for a wide range of potential downstream applications. Therefore, a metric that effectively quantifies the richness of linearly available information is more appropriate than one tied to a single network's function. Furthermore, given the scale of our dataset, a full-batch computation is infeasible. The minibatch-CKA proposed by Nguyen et al. [35], which uses an unbiased estimator of HSIC [36] to compute linear CKA so that the value of CKA is independent of the batch size, enables scalable assessment across large model sets and high-dimensional embeddings.

This approach enables us to assess the embedding-level consistency between a pretrained model and the structural patterns captured by RINNs, providing an efficient, label-free estimate of transferability. The use of minibatch-CKA significantly reduces both memory footprint and computational overhead. Unlike standard CKA, which requires storing and processing the full dataset to compute Gram matrices, minibatch-CKA operates on small batches and accumulates similarity statistics incrementally. This makes it particularly well-suited for evaluating large model sets on long-form or high-resolution surveillance data, enabling scalable deployment under real-world resource constraints.

5.2.4 Source-Free Unsupervised Domain Adaptation

Source-free unsupervised domain adaptation (SFUDA) is a specialized setting of adapting pretrained models to an unannotated target domain without access to the source data or to labeled target data [37]. This setting addresses real-world constraints where source data is not universally shareable due to privacy and storage limitations, conditions commonly encoun-

tered in smart surveillance scenarios. Furthermore, such a setting alleviates computational burdens, which aligns with the need for scalable and efficient deployment in edge-based systems.

While a number of methods have been proposed under the SFUDA setting [37, 26, 38], they primarily focus on adapting a given model to the target domain. These methods can be grouped into three broad categories: (1) self-tuning via pseudo labels and information maximization [37, 39, 27], (2) feature alignment using structural cues in the target domain [26], and (3) sample generation to synthesize source-like data [38].

These methods focus primarily on how to adapt the given model rather than on whether it is the right model to begin with. However, prior work [20] has shown that the choice of a good starting point can have a substantial impact on the final performance. In contrast, our work addresses this issue, proposing a transferability assessment method that operates fully within the SFUDA constraints, yet without relying on pretrained model embeddings or source data. In doing so, our method complements domain adaptation-based approaches by guiding model selection prior to adaptation, potentially improving their overall effectiveness.

5.3 Materials and Methods

In this section, we present our novel transferability assessment framework designed for source-free unsupervised domain adaptation. We first introduce the overall structure and key components of the framework. We then introduce the two proposed RINN-based assessments, label-level score S^L, and embedding-level score S^E. Finally, we describe the benchmark datasets and models that will be used in the experimental evaluation.

This proposed framework operates on video frames obtained from the target environment. While the characteristics of this data are inevitably influenced by the camera hardware, environmental conditions including lighting, and any preliminary preprocessing, the transferability assessment method itself is designed to be independent of these specific underlying hardware components, focusing instead on the data as presented.

5.3.1 Framework

Our proposed framework is illustrated in Figure 5.3. Let $F = \{f_n\}_{n=1}^{N}$ denote a model zoo of N pretrained models and let X be the unlabelled target data. Each model f_n encodes X into a representation $f_n(X)$; these representations are then used to evaluate the transferability of each model.

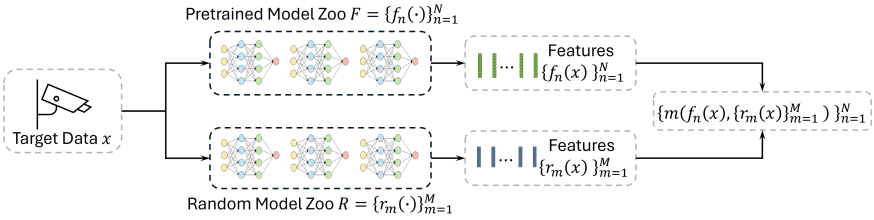


Figure 5.3: Overview of the framework. The target data is fed to both the pretrained models and the randomly initialized neural networks. We then predict the transferability of each pretrained model to the new data by estimating the mutual information between both types of representations.

In addition, we construct an RINN zoo $R = \{r_m\}_{m=1}^{M}$ based on a set of M randomly initialized neural networks. Given one pretrained representation $f_n(X)$ and the collection $R(X) = \{r_m(X)\}_{m=1}^{M}$, we define a transferability score $S(f_n(X), R(X))$, which measures how much of the information captured by the RINNs is already present in $f_n(X)$. A higher score implies that f_n is more adaptable to the target domain. We aim to rank f_n based on their adaptability to X without requiring source data or target annotations. We compare two approaches of using the RINNs for transferability estimation: label-level scoring and embedding-level scoring, as explained in the following paragraphs.

5.3.1.1 Label-Level Score

In this approach, we use the RINNs to directly generate pseudo labels $Y(X)$ for each input X: $Y(X) = c_m^r(r_m(X))$, where c_m^r is a simple, randomly initialized classifier added on top of r_m. This layer predicts a pseudo label for each input. The predicted pseudo label has no semantic meaning but instead should be seen as an assignment to a cluster grouping similar inputs. In our experiments, we set each c_m^r to predict 20 classes. We then add a similar fully-connected layer $c_n^{r_m}$ on top of each f_n and fine-tune this layer using gradient descent to predict the pseudo label. The underlying intuition for this approach is that predicting the pseudo label is only possible if the pretrained model has a rich enough feature representation that captures broad information covering the features extracted by a large set of RINNs. The transferability score is then the sum over all the r_m:

$$S^L(f_R(X), R(X)) = \sum_{M}^{m=1} \textbf{eval}(c_n^{r_m}(f_n(X)), c_m^r(r_m(X))) \tag{5.1}$$

where **eval** is a performance metric such as accuracy, F1-score or MSE.

5.3.1.2 Embedding-Level Score

A disadvantage of the label-level score is that it requires a training step which might be costly. In addition, the pseudo labels sometimes collapse where the RINN predicts the same output for each input. A large set of RINNs is needed to avoid this issue.

An alternative scoring mechanism uses internal representations of the RINNs and pretrained models. Based on knowledge distillation approaches, we can use metrics such as correlation-based similarity [40] or Centered Kernel Alignment (CKA) to directly measure the similarities between both internal representations. The embedding level score is defined as:

$$S^E(f_n(X), R(X)) = \mathbf{sim}(f_n(X), R(X)) \tag{5.2}$$

In this paper, we use minibatch-CKA, proposed in [35], to measure the structural similarity between $f_n(X)$ and $R(X)$. We adopt minibatch-CKA not only for its alignment with representation-based analysis but also for its practical benefits: it enables scalable evaluation by computing similarity over mini-batches, thereby significantly reducing memory usage compared to standard CKA. To calculate minibatch-CKA, X is first divided into batches with B samples, $X = \{X_k\}_{k=1}^{K}, |X_k| = B$. By applying pretrained model $f_n(\cdot)$ and randomly initialized model $r_m(\cdot)$ to X_k, we obtain two representations $f_n(X_k) \in \mathbb{R}^{B \times f_n^d}$ and $r_m(X_k) \in \mathbb{R}^{B \times r_m^d}$. Note that f_n^d and r_m^d represent the number of dimensions of f_n and r_m which are not necessary the same. As minibatch-CKA uses a linear kernel to calculate CKA, we further define the kernel matrices $\mathbf{K} = f_n(X_k) f_n(X_k)^\top$ and $\mathbf{L} = r_m(X_k) r_m(X_k)^\top$. The minibatch-CKA is then calculated using the following equation:

$$\mathrm{CKA}^{\mathrm{mb}}(\mathbf{K}, \mathbf{L}) = \frac{\mathrm{HSIC}_1^{mb}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathrm{HSIC}_1^{mb}(\mathbf{K}, \mathbf{K})}\sqrt{\mathrm{HSIC}_1^{mb}(\mathbf{L}, \mathbf{L})}}, \tag{5.3}$$

where

$$\mathrm{HSIC}_1^{mb}(\mathbf{K}, \mathbf{L}) = \frac{1}{k}\sum_{i=1}^{k} \mathrm{HSIC}_1(\mathbf{K}_i, \mathbf{L}_i). \tag{5.4}$$

In Equation (5.3), HSIC is the Hilbert–Schmidt Independence Criterion [41], which was used to measure the dependence between two sets of variables. To make the HSIC independent to B so that CKA can be calculated batch-by-batch, minibatch-CKA uses the unbiased HSIC estimator (HSIC_1) in [36].

For detailed derivation and proof, please refer to [33]. With the minibatch-CKA, Equation 5.2 is then:

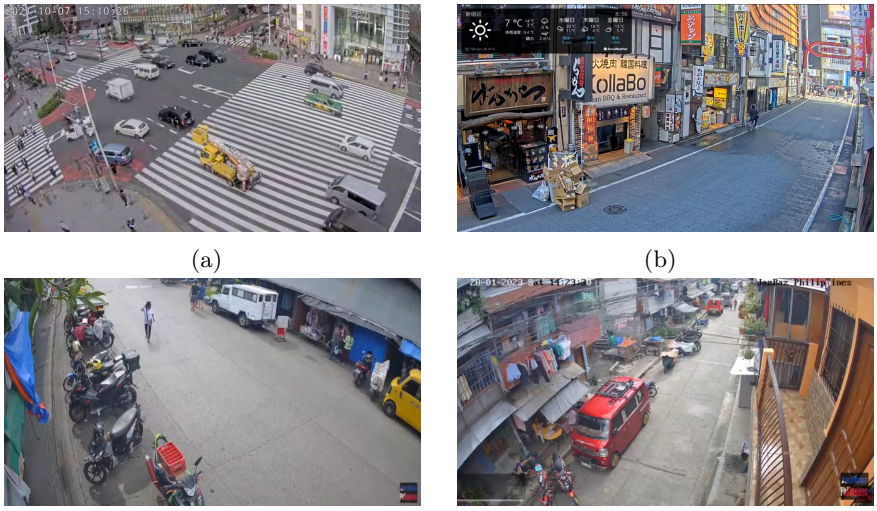
$$S^E(f_n(X), \{r_m(X)\}_M^{m=1}) = \mathrm{CKA}^{\mathrm{mb}}(f_n(X_k)f_n(X_k)^\top, r_m(X_k)r_m(X_k)^\top). \tag{5.5}$$

5.3.2 Models

We construct a model zoo comprising models with identical architectures but trained on separate source videos. This design isolates data-induced variations in learned representations while controlling for architectural differences. All models use the ASTNet backbone [42] and follow the training configuration for ShanghaiTech, as described in the original paper. In addition to the configuration-specific models, we include a universal model trained on the combined source data from all configurations. This allows us to evaluate whether a universal model can match or exceed the performance of specialized models under identical resource constraints. For evaluation, all models from the zoo were used off-the-shelf, kept frozen, and uniformly applied to each target dataset without task-specific fine-tuning. Each model encodes the target data into feature representations, which are then used to perform transferability assessments. These assessments rank the models by their adaptability to the target domain. Importantly, in alignment with real-world surveillance constraints, the source data is assumed to be inaccessible during the transferability evaluation process.

The RINN zoo is instantiated with $M = 20$ networks, a value chosen to balance computational cost and ranking stability, as confirmed by the ablation study in Section 5.4.4. Each RINN adopts the X3D-M backbone [43] from PyTorchVideo(v0.1.5), initialized with randomly sampled weights instead of pretrained checkpoints. For the label-level score S^L, we reduce X3D's output layer to 20 classes. Following [20], we train a single fully connected layer to map the pretrained model's representation $f_n(X)$ to the pseudo labels generated by the RINN, $c^{r_m}(r_m(X))$. This layer was trained using standard procedures: the learning rate was set using a learning rate range test, and an early stopping criterion on a held-out validation set determined the number of epochs. An exhaustive hyperparameter search was not conducted for this component, as the S^L score's role is to serve as a baseline demonstrating the limitations of classifier-dependent methods. This instability provides the core motivation for our main contribution, the S^E score, which is independent of such a classifier. For the embedding-level score S^E, which directly compares feature embeddings, we remove the final

classification layer of X3D. Similarity is computed using minibatch-CKA, adapted from the open-source PyTorch(v2.4.0) implementation by Ristori [44]. The resulting RINN zoo is reused across all target datasets without any fine-tuning. Consistent with real-world surveillance constraints, no target annotations are accessed during the transferability assessment.



(a) (b)

(c) (d)

Figure 5.4: Illustration of four source videos used in training the pretrained model zoo, highlighting the diversity in scene layout and viewing conditions. (a) A high-angle camera overlooking a busy intersection with dense traffic and small-scale objects. (b) A horizontal view of a pedestrian shopping street with frequent occlusions. (c) A roadside market with moderate activity. (d) A horizontal view in a domestic neighborhood with low traffic density and residential surroundings. Note that we also include the Mixed model, which is trained on all these four cameras, in the model zoo.

5.3.3 Datasets

The source data used to train the models in the pretrained model zoo consist of four real-world surveillance videos, captured by cameras positioned at distinct locations with varying viewpoints. These video streams, sourced from live surveillance feeds in Tokyo and Agdao [14, 15, 16], are illustrated in Figure 5.4. In Tokyo, one camera captures a large urban intersection (Figure 5.4(a)), while the other monitors a narrow alley lined with bars and restaurants (Figure 5.4(b)). In Agdao, the Agdao-Market camera provides a close-up view of a bustling roadside market (Figure 5.4(c)), whereas Agdao-

Street captures a quieter residential street scene (Figure 5.4(d)). Temporally, while Tokyo Street [15] covers a short period of night time, the other three datasets are recorded during the daytime. Context-wise, the urban and market scenes (Tokyo Intersection [14], Tokyo Street [15], and Agdao Market [16]) are characterized by high object density and complex layouts, leading to severe and frequent occlusions. Viewpoints range from high-angle, which reduces object scale, to eye-level, where scale can vary dynamically.

For the target data, we use three publicly available datasets: Street Scene [17], NWPU Campus [18], and Urban Tracker [19]. Street Scene and NWPU Campus are designed for anomaly detection. Street Scene contains videos captured from a fixed camera viewpoint, is composed of short, discontinuous clips, and includes 205 annotated anomaly events spanning 17 categories. NWPU Campus includes footage from 43 different cameras, covering 28 distinct anomaly classes. Based on these annotations, both datasets are used for evaluating anomaly detection and event classification tasks. For anomaly detection, we adopt frame-level AUC as the evaluation metric, while event classification is evaluated using instance-level mAP. Since the original annotations only indicate abnormal frames, we manually map each target video into 1 of the 17 or 28 predefined anomaly categories described in the respective papers. Urban Tracker, originally developed for object tracking, comprises five video sequences. We utilize four of them, specifically *Sherbrooke*, *Rouen*, *St-Marc*, and *Atrium*, and exclude *René-Lévesque* due to excessive downscaling that renders objects too small for reliable detection. As the dataset does not provide a predefined train–test split, we follow the protocol in [45], assigning the first 80% of frames for training and the remaining 20% for testing. Since Urban Tracker provides object-level descriptions but not formal semantic classes, we categorize objects based on their free-text descriptions. For object tagging, we evaluate performance using frame-level mean average precision (mAP). Evaluations across these datasets are conducted in accordance with the type of annotation available. All three target datasets were recorded during the daytime and originate from geographic regions (China, North America/Europe) distinct from the source domains. A summary of the datasets used in our experiments is presented in Table 5.1. Further details can be found in their respective original publications.

5.3.4 Baseline Models and Evaluation Protocols

Very few methods have been developed for transferability assessment under the source-free, label-free setting. Most existing approaches focus on direct domain adaptation rather than explicitly evaluating transferability. In

Dataset	Scene Characteristics			Specifications			
	Environment	Density	View	# Cams	# Frames	Resolution	fps
Tokyo Intersection [14]	Urban Intersection	Dense	High-level	1	432,000	1920x1080	30
Tokyo Street [15]	Shopping Street	Moderate	Eye-level	1	324,000	1920x1080	30
Agdao Market [16]	Roadside Market	Dense	Eye-level	1	432,000	1920x1080	30
Agdao Residential [16]	Residential Street	Sparse	Mid-level	1	432,000	1920x1080	30
Street Scene [17]	Two-lane Street	Sparse	High-level	1	202,545	1280x720	15
NWPU Campus [18]	University Campus	Varies	Mixed	43	1,466,073	Various	25
Urban Tracker [19]	Mixed	Dense	Mixed	4	600-4,540	Various	25-30

Table 5.1: A summary of the source and target datasets used in our experiments. The table details the role of each dataset, its core scene characteristics, and its technical specifications.

Task	Object Tagging	Anomaly Detection		Event Classification	
Dataset	Urban Tracker	Street	NWPU	Street	NWPU
UD	0.17 ± 0.33	1.00	0.92 ± 0.14	0.60	0.78 ± 0.27
EnsV	0.41 ± 0.20	0.89	0.85 ± 0.12	0.67	0.78 ± 0.18
S^L	0.69 ± 0.11	0.79	0.83 ± 0.09	0.48	0.85 ± 0.06
S^E	0.95 ± 0.02	0.90	0.94 ± 0.13	0.77	0.89 ± 0.09

Table 5.2: Detailed Kendall's τ scores for each transferability assessments. Results are reported as mean (±) standard deviation. No deviation is reported for Street Scene as it uses a single camera setup.

this work, we primarily compare our two proposed transferability metrics, S^L and S^E, against Uncertainty Distance (UD) [26] and Ensemble-based Validation (EnsV) [27], a state-of-the-art method in model selection. Note that EnsV [27] relies on the pretrained classifiers to generate the proxy-groundtruth for model ranking. Since our source models are feature extractors, they cannot be used with EnsV directly. Thus, to adapt EnsV as a baseline, we implement a two-step process. For each source model, we first apply a K-Means clustering algorithm to its output embeddings to generate a set of k distinct pseudo-classes. Subsequently, the label assigned to each embedding is determined by its nearest cluster centroid. To ensure a fair and controlled comparison against our own label-level baseline (S^L), we set the number of clusters to k = 20, matching the output dimensionality of the S^L classifier.

For object tagging and event classification, we attach a fully connected layer on top of each frozen pretrained model, following the protocol in [20]. This layer is fine-tuned on the training set of the target domain, and task accuracy is used as the ground truth to assess the quality of the transferability rankings. To validate whether the estimated model transferability aligns with actual model performance, we report the Kendall's τ score [46]. Kendall's τ ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating no correlation, to quantify alignment between estimated and ground-truth rankings. It is worth noting that, given only five models, Kendall's τ is highly sensitive: even a single position shift can noticeably affect the score. Similarly, for anomaly detection, we follow the evaluation protocol in [20], using Kendall's to measure the correlation between estimated rankings and actual performance.

5.4 Results and Discussion

In the following sections, we evaluate our transferability estimation approach on the tasks of anomaly detection, object tagging, and event detection. The performance of our method is compared against other baselines (UD, EnsV), with the results visualized in Figure 5.5 and the precise numerical scores reported in Table 5.2. We then perform an ablation study to investigate the trade-off between transferability prediction performance and computational cost by varying the number of RINN's in our approach.

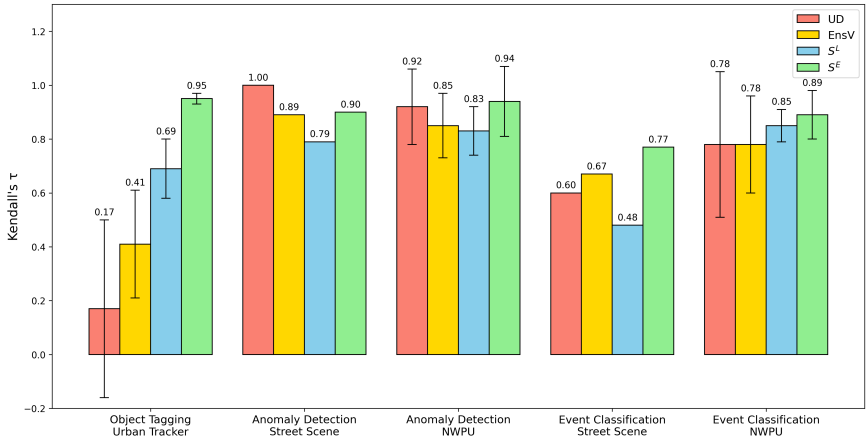


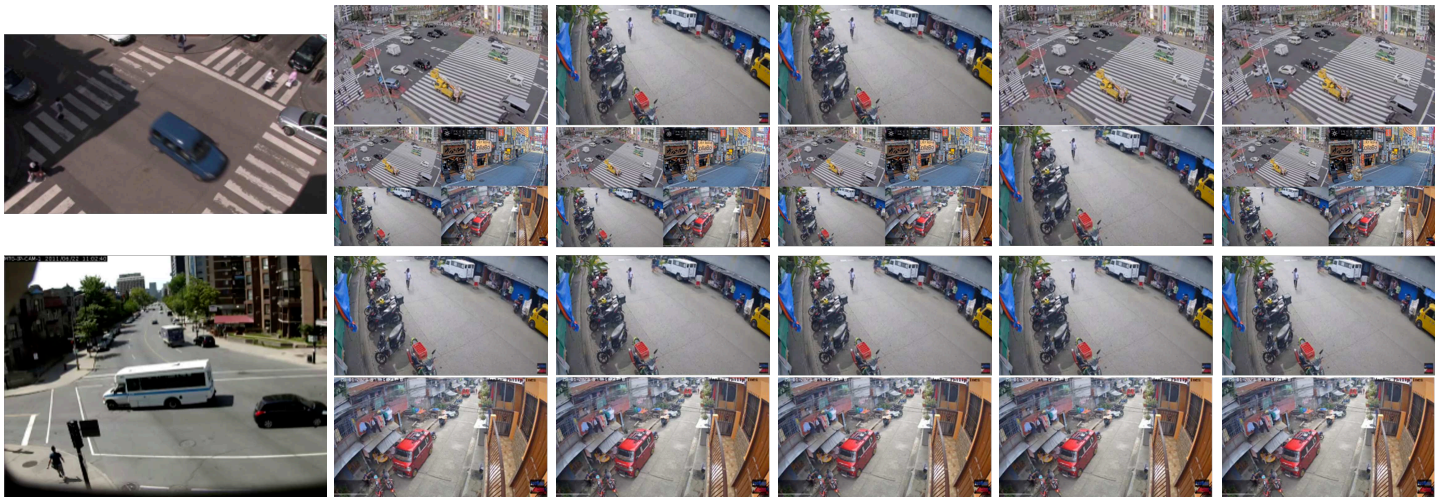
Figure 5.5: A comparison of transferability assessments. The bar chart shows the mean Kendall's τ correlation between the rankings produced by each method (UD, EnsV, S^L, and S^E) and a proxy ground-truth ranking. The evaluation is performed across five distinct downstream task–dataset combinations. Bars represent the mean performance, and the error bars indicate the standard deviation across multiple camera views where applicable. Higher τ values signify better performance. Our proposed embedding-level score, S^E, consistently achieves the highest correlation. For the exact numerical values, please refer to Table 5.2.

5.4.1 Object tagging

We evaluate performance on the object tagging task using the Urban Tracker dataset, which includes four diverse camera scenes. As shown in Table 5.2, S^E achieves the highest Kendall's τ (0.95 ± 0.02), indicating strong and consistent alignment with the ground truth ranking. In contrast, UD yields a much lower correlation (0.17 ± 0.33), suggesting poor reliability in this task. While EnsV provides a more competitive baseline (0.41 ± 0.20), its performance is still notably lower than S^E, indicating that our method's

(a) (b) (c) (d) (e) (f)

Figure 5.6: Scene-level qualitative comparison of model selection for object tagging on Urban Tracker. (a) shows example frames from the two target scenes: Atrium and St. Marc(top to bottom). (b)-(f) present the top and second model predictions (top and bottom rows per scene) selected by different methods: (b) FCN w/Label (ground truth); (c) UD; (d) EnsV; (e) S^L; and (f) S^E. Each image shown is an example frame from the original source dataset that the corresponding selected model was trained on. This allows for a visual inspection of the potential domain shift between the selected model's origin and the target scene. If the *Mixed* model (trained on all four distinct source datasets) is selected, it is represented by a grid of four distinct images, each an exemplar from one of its constituent source datasets. This figure highlights the agreement or divergence in the types of source models selected by the different methods and by the oracle. (Layout and caption adapted from Wang et al. [47].)



(a) (b) (c) (d) (e) (f)

Figure 5.7: Scene-level qualitative comparison of model selection for object tagging on Urban Tracker. (a) shows example frames from the other two target scenes: Rouen and Sherbrooke (top to bottom). (b)-(f) present the top and second model predictions (top and bottom rows per scene) selected by different methods: (b) FCN w/Label (ground truth); (c) UD; (d) EnsV; (e) S^L; and (f) S^E. Each image shown is an example frame from the original source dataset that the corresponding selected model was trained on. This allows for a visual inspection of the potential domain shift between the selected model's origin and the target scene. If the *Mixed* model (trained on all four distinct source datasets) is selected, it is represented by a grid of four distinct images, each an exemplar from one of its constituent source datasets. This figure highlights the agreement or divergence in the types of source models selected by the different methods and by the oracle.(Layout and caption adapted from Wang et al. [47].)

direct comparison of embedding structures is more effective than relying on an ensemble's prediction consensus when a significant task shift occurs. Among the four methods, S^E also demonstrates the most stable performance, likely due to its embedding-level comparison, which better preserves structural information from the target data. By contrast, S^L relies on a randomly initialized classifier to map RINN features to scalar outputs, making it more susceptible to instability from weight initialization and information compression.

UD's poor performance merits further examination. While Kendall's τ is naturally sensitive with only five models, the primary limitation lies in UD's core assumption: that low model uncertainty on a pretext task (frame reconstruction) correlates with downstream task performance. This assumption fails in object tagging, leading UD to collapse score ranges and obscure meaningful distinctions between models. For instance, in the Rouen scene, UD misidentifies the best-performing model, underscoring how pretext task uncertainty can misrepresent object-level difficulty. Interestingly, UD performs better in tasks where pretext and target objectives are more aligned (see Section 5.4.2). EnsV is also affected by this fundamental task misalignment, as its clustering is based on the same pretext task embeddings. However, its core mechanism, the reliance on a "joint agreement" from the ensemble, is a double-edged sword. On one hand, it provides a stabilizing effect that filters out noise from poorly performing models, which explains its more competitive ranking compared to UD in scenes like St. Marc. On the other hand, this same property often leads to tied or nearly identical scores among the top-competing models, making the final ranking ambiguous and potentially suboptimal.

To further validate these observations, Figure 5.6 and 5.7 provides a scene-level comparison of selected models. S^E consistently identifies the top-performing model across all scenes. Although UD occasionally agrees with S^E, even small misorderings, such as those seen in Atrium, St. Marc, and Rouen, can drive Kendall's τ to zero. These failures typically arise when UD assigns nearly identical scores to its top three models, making its ranking vulnerable to minor fluctuations. Meanwhile, S^L and S^E show strong agreement, diverging only in the second-best model for Rouen. These results emphasize the value of analyzing the score distribution itself, especially when competing models are closely matched, rather than relying solely on rank-based metrics.

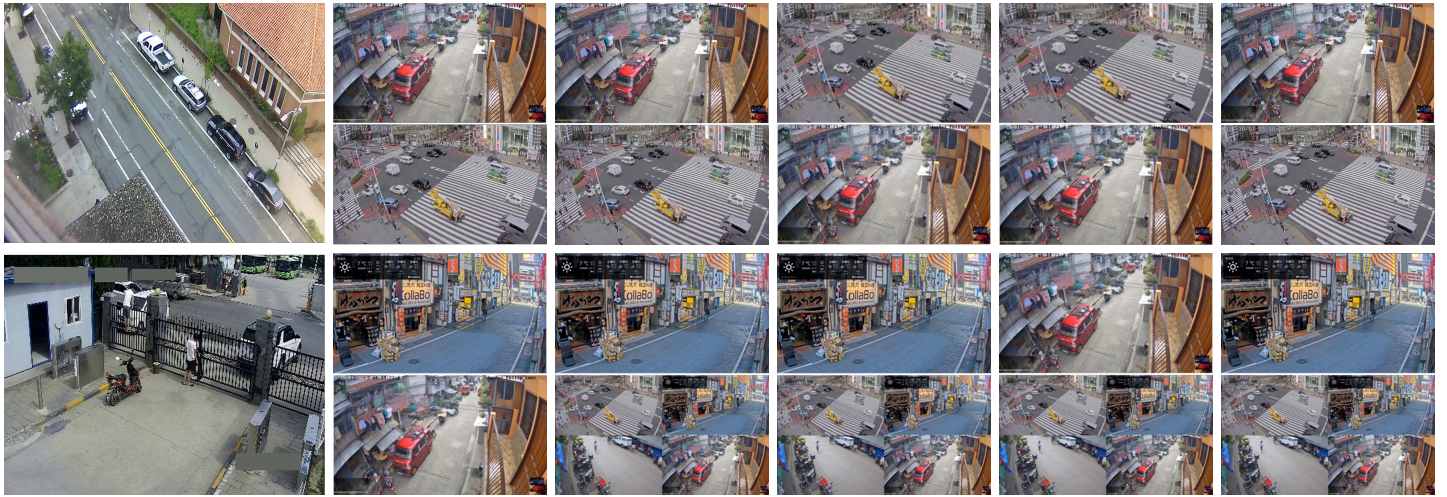
The consistent performance of S^E reinforces its independence from output uncertainty and its robustness under task misalignment. Finally, the *Mixed* model (visualized within that cell as a grid of four images), which

is trained on footage from all four sources, never secures first place in any scene, highlighting the practical benefit of maintaining a compact model zoo of cameraspecific encoders rather than relying on a single "universal" model when computational resources allow.

5.4.2 Anomaly detection

The models in the model zoo were originally trained for anomaly detection. We now evaluate how well our transferability assessment methods predict their performance on the same task across new environments. For each pretrained model, we apply a one-class SVM (OCSVM) to its extracted feature representations and compute the AUC score as the performance metric. Since S^E is derived from Equation (5.3), we directly report CKA^{mb} values, which range from 0 to 1. The resulting Kendall's τ values for Street Scene and NWPU Campus are summarized in Table 5.2. Qualitative insights into the scene-level model selections are further illustrated in Figure 5.8 and 5.9. Across most target scenes, all four methods, UD, EnSV, S^L, and S^E, achieve τ values above 0.7, indicating strong correlation with the actual model rankings despite the absence of source data or target labels. This confirms the effectiveness of our approach in identifying well-suited models for adaptation. Notably, S^E consistently outperforms S^L in ranking stability. As also observed in the object tagging task (Section 5.4.1), this difference stems from the instability of the randomly initialized classifier in S^L, which often yields imbalanced predictions. In contrast, S^E benefits from its embedding-level comparison via CKA^{mb}, offering more stable and reliable estimates. This finding is further corroborated by the ablation study in Section 5.4.4. UD also demonstrates significantly stronger performance on anomaly detection compared to object tagging. In particular, it achieves perfect alignment ($\tau = 1$) on Street Scene, outperforming both S^L and S^E. The strong performance of both baselines is expected, as they both leverage the alignment between the models' original pretext task and the anomaly detection objective. UD's success is a direct result of this, as it measures uncertainty on the relevant pretext task. EnsV's competitive ranking performance is also rooted in this alignment between the pretext task and the anomaly detection objective, as the meaningful embeddings produce effective clusters. The final ensemble further provides additional stability by aggregating these predictions. However, Street Scene involves a single fixed viewpoint and may not generalize well to more complex surveillance settings. A more comprehensive evaluation is provided by NWPU Campus, which contains 43 distinct camera perspectives.

On NWPU, both UD and EnsV maintain competitive performance, achiev-



(a) (b) (c) (d) (e) (f)

Figure 5.8: Scene-level qualitative comparison of model selection for anomaly detection on Street Scene and NWPU Campus. The top row shows results from the Street Scene dataset, while the second row shows representative examples from different scenes within the NWPU Campus dataset (D01). (a) shows example frames from the four target scenes. (b)-(f) present the top and second-best model predictions (top and bottom rows per scene) selected by different methods. (b) FCN w/Label (ground truth); (c) UD; (d) EnsV; (e) S^L; (f) S^E. Each cell shows the prediction of the model ranked best (top) and second-best (bottom) by the corresponding method, highlighting agreement or divergence with the oracle selection. For detailed quantitative Kendall's τ scores across all scenes, please refer to Table 5.2.

(a) (b) (c) (d) (e) (f)

Figure 5.9: Scene-level qualitative comparison of model selection for anomaly detection on Street Scene and NWPU Campus. The rows show representative examples from different scenes within the NWPU Campus dataset (D03, D48). (a) shows example frames from the four target scenes. (b)-(f) present the top and second-best model predictions (top and bottom rows per scene) selected by different methods. (b) FCN w/Label (ground truth); (c) UD; (d) EnsV; (e) S^L; (f) S^E. Each cell shows the prediction of the model ranked best (top) and second-best (bottom) by the corresponding method, highlighting agreement or divergence with the oracle selection. For detailed quantitative Kendall's τ scores across all scenes, please refer to Table 5.2.

ing a higher mean τ and lower variance compared to its object tagging results. It is worth mentioning that EnsV's performance is slightly more stable compared to UD, due to its characteristic of ensembling model predictions. Nonetheless, S^E still secures the highest overall mean τ, with lower sensitivity to scene variability. In contrast, S^L remains unstable, exhibiting high variance across scenes, further emphasizing the limitations of classifier-based methods in unsupervised settings and the robustness of embedding-based comparisons.

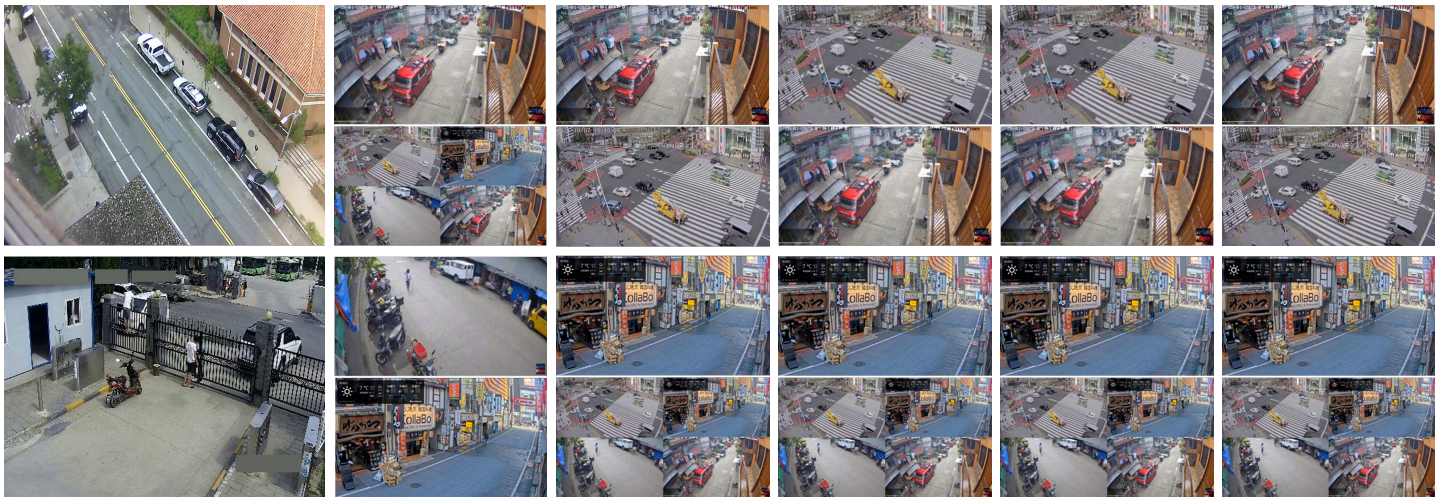
To further visualize these trends, Figure 5.8 and 5.9 presents scene-level model selections across NWPU Campus and Street Scene. Figure 5.8 and 5.9(a) shows example frames from the target scenes, while Figure 5.8 and 5.9(b)-(f) illustrate the top two models identified by each method. UD and EnsV generally align with the ground truth, deviating in NWPU-D01, where it misorders the second-best model, likely due to minimal differences in the scores. Interestingly, S^E also agrees with UD and EnsV in this case, suggesting some ambiguity in the ground-truth ranking. By contrast, S^L exhibits greater instability, with inconsistent top-three rankings, underscoring its sensitivity to classifier initialization and output imbalance.

5.4.3 Event Classification

We conclude our per-task evaluation with event classification, where each anomaly instance is assigned a semantic label, as defined in [18]. As shown in Table 5.2, the performance of both UD and EnsV drop considerably compared to their results on anomaly detection, particularly in the Street Scene dataset. In contrast, both S^L and S^E maintain stable performance across tasks, with S^E consistently achieving the highest Kendall's τ overall.

This discrepancy is especially notable because all four transferability metrics, namely UD, EnsV, S^L, and S^E, are computed without access to target labels and are therefore invariant to task changes. The only component that changes across tasks is the oracle baseline (FCN), which is retrained using event-level annotations instead of anomaly labels. Consequently, the ground-truth ranking of model performance shifts, even though the underlying target data remains the same.

To explore the implications of this task shift, Figure 5.10 and 5.11 highlights three representative scenes where the top-ranked model by FCN changes between anomaly detection and event classification. Notably, both S^L and S^E maintain consistent selections, showing resilience to the shift in downstream objectives. In contrast, both UD and EnsV are more sensitive to these shifts. For example, in the first row (Street), UD ranks the second-best model third,



(a) (b) (c) (d) (e) (f)

Figure 5.10: Scene-level qualitative comparison of model selection for event classification on Street Scene and NWPU Campus. The top row shows results from the Street Scene dataset, while the subsequent row shows representative examples from scene in NWPU Campus dataset (D01). (a) shows example frames from the four target scenes. (b)-(f) present the top and second-best model predictions (top and bottom rows per scene) selected by different methods. (b) FCN w/Label (proxy ground truth); (c) UD; (d) EnsV; (e) S^L; (f) S^E. Each cell shows the prediction of the model ranked best (top) and second-best (bottom) by the corresponding method, highlighting agreement or divergence with the oracle selection. For detailed quantitative Kendall's τ scores across all scenes, please refer to Table 5.2.

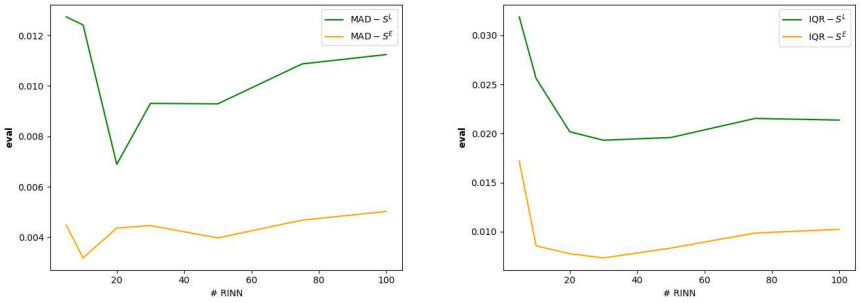
(a) (b) (c) (d) (e) (f)

Figure 5.11: Scene-level qualitative comparison of model selection for event classification on Street Scene and NWPU Campus. It shows results from the NWPU Campus dataset (D48). (a) shows example frames from the four target scenes. (b)-(f) present the top and second-best model predictions (top and bottom rows per scene) selected by different methods. (b) FCN w/Label (proxy ground truth); (c) UD; (d) EnsV; (e) S^L; (f) S^E. Each cell shows the prediction of the model ranked best (top) and second-best (bottom) by the corresponding method, highlighting agreement or divergence with the oracle selection. For detailed quantitative Kendall's τ scores across all scenes, please refer to Table 5.2.

which is similar to S^L and S^E and and is due to only minor differences in transferability scores. However, in the second (D01) and third (D48) rows, UD misranks the second-best model as fifth, indicating a sharper deviation from the ground truth. EnsV on the other hand, similar to object tagging, generates a tied rank when the task drifts. These cases underscore the vulnerability of UD and EnsV when their pretext assumption, uncertainty as a proxy for downstream performance, breaks under task misalignment, while embedding-based methods like S^E remain more robust.

Together, these results illustrate that task changes, even without altering the input data, can lead to significant divergence in model rankings. Methods like S^E that rely on structural representation comparison rather than output uncertainty or classifier initialization offer more consistent transferability assessments under such shifts, making them particularly suitable for task-agnostic deployment in smart surveillance systems.

5.4.4 Ablation Study



(a) (b)

Figure 5.12: Impact of the number of RINNs used. Shown with (a) MAD and (b) IQR, the lower the more stable.

While increasing the number of RINNs can improve the reliability of transferability assessments, it also raises computational costs. Therefore, it is essential to strike a balance between assessment stability and computational efficiency. To analyze this trade-off, we randomly initialized 100 RINNs with identical architectures and applied them to the Urban Tracker dataset to compute both S^L and S^E. We then evaluated subsets of the top $\{5, 10, 20, 30, 50, 70, 100\}$ RINNs and measured the stability of their scores using two statistical metrics: Interquartile Range (IQR) and Median Absolute Deviation (MAD). IQR quantifies the spread between the first and third quartiles, while MAD measures the median of the absolute deviations

from the median. Both are less sensitive to outliers than standard deviation, making them suitable for assessing stability in the presence of noisy or highly variable scores. Lower values in either metric indicate greater stability across RINNs. The results, shown in Figure 5.12, illustrate how stability varies with the number of RINNs, helping to identify an optimal configuration that balances reliability and efficiency.

For both S^L and S^E, we observe that beyond a certain number of RINNs, increasing the ensemble size yields diminishing returns in performance stability. The results indicate that using 20 RINNs provides a good trade-off between reliability and computational efficiency. Additionally, the stability curve for S^E demonstrates that the embedding-level assessment is consistently more robust than the label-level counterpart, reinforcing its suitability for resource-constrained deployment.

5.5 Conclusions

We address the challenge of identifying the most adaptable pretrained model for source-free, label-free domain adaptation in smart surveillance, where neither the original training data nor target annotations are accessible. Universal models often fall short of the performance achieved by camera-specific models, and only a limited number of transferability assessment methods exist. Among them, approaches like uncertainty distance rely on the questionable assumption that a model's uncertainty on one task reliably predicts its performance on another. In the absence of labeled target data, effective transferability assessment requires a task-agnostic reference embedding space; this motivates our use of ensembles of randomly initialized neural networks (RINNs) to avoid bias introduced by pretrained representations. We propose a novel and effective framework for transferability estimation in source-free unsupervised settings, specifically tailored to the unique demands of smart surveillance systems, which are characterized by heterogeneous camera configurations and pronounced domain shifts. By leveraging RINNs as unbiased feature extractors, our approach mitigates both structural and task-specific biases inherent in conventional pretrained models. This enables task-agnostic, model-independent transferability assessment. Additionally, our embedding-level metric reduces the computational overhead associated with pseudo label-based approaches, making the method scalable for large surveillance deployments. Empirical evaluations on real-world surveillance datasets demonstrate the practical utility of our method. Specifically, our embedding-level score S^E achieved strong Kendall's τ correlations with ground-truth model rankings across multiple downstream tasks,

generally performing comparably to or outperforming other source-free assessment methods. For instance, S^E obtained values of 0.95 ± 0.02 for object tagging on Urban Tracker, 0.94 ± 0.13 for anomaly detection on NWPU Campus, and 0.89 ± 0.09 for event classification on Street Scene. This consistent performance indicates that our framework can reliably identify the most adaptable pretrained models, even when the specific downstream task is unknown and both source data and labeled target data are unavailable. These quantitative findings highlight the value of our approach in challenging real-world source-free unsupervised scenarios, particularly in privacy-sensitive and resource-constrained smart city environments.

5.5.1 Limitations

Despite the positive evaluation from the conducted experiments, this work has several limitations that needed to be discussed.

First, a primary limitation arises from the symmetry of minibatch-CKA. We have proposed that a large ensemble of RINNs approximates a comprehensive, information-rich embedding space for the target data. The ideal pretrained model, therefore, is one whose representations are a large subset of this embedding space. Our goal is to measure the extent of this informational overlap. However, minibatch-CKA is symmetric which indicates the degree of overall alignment between the two embedding spaces rather than measuring this one-way, subset relationship. This is compounded by potential redundancy within the RINN ensemble. Since our method does not explicitly decorrelate the features from different RINNs, it risks overestimating certain information and impairing the model ranking.

Second, our experimental validation is limited in scope. The evaluated tasks are classification-related tasks and its efficacy for other surveillance applications, such as anomaly event localization, semantic segmentation, and crowd counting, remains unverified. Furthermore, the experiments were conducted on static datasets, and the analysis does not account for temporal dynamics or concept drift, where data distributions evolve over time.

Third, the study is bounded by the limitations and biases of the datasets used. The source datasets are mostly recorded during the daytime. Geographically, these datasets are biased to specific urban and residential environments, which presents a realistic and challenging transferability scenario when assessing models on target data from different global contexts. The target datasets also possess unique constraints: Street Scene [17] is limited by a single, fixed viewpoint and short, discontinuous clips; NWPU Campus [18] has a known data imbalance and limited object interaction; and Urban

Tracker [19] contains non-standard annotations and variable technical quality. Despite these limitations, by intentionally evaluating our method across this diverse and complementary set of target domains, we have subjected our approach to a more holistic and stringent test of its robustness. While each individual experiment is bounded by the nature of its data, the overall experiments provide broader evidence of our method's applicability across varied conditions.

Finally, practical and computational factors present further considerations. While the method avoids the cost of full fine-tuning, it incurs its own practical costs. The need to generate and store RINN weights creates a non-trivial memory footprint and computational overhead. This paper does not include a formal cost–benefit analysis comparing this assessment overhead to other lightweight baseline strategies.

Acknowledging these boundaries is crucial for contextualizing our findings and provides a clear roadmap for future research.

5.5.2 Future Work

In future work, our research direction will focus on several key aspects, focusing on methodological refinement, expanded empirical validation, and deeper theoretical analysis.

First, we will refine our core assessment method. This involves exploring asymmetric embedding similarity metrics to better estimate directional information transfer between representations. We will also investigate the impact of different RINN architectures and initialization strategies to more effectively construct the ensemble. Although RINNs require no training, they possess a non-trivial memory footprint. While the current assessment was conducted on resource-rich cloud infrastructure, developing a more memory-efficient implementation is crucial for large-scale deployment. Furthermore, our current approach is a one-step process in assessing the model transferability. To further adapt the selected model to target data, a more efficient model adaptation strategy must be investigated.

Second, we will expand the scope of our experimental validation. This includes applying the method to non-classification tasks (e.g., semantic segmentation), extending the analysis to streaming data to assess performance against temporal drift (e.g., changes in lighting and weather), and evaluating against target datasets from more diverse geographical and temporal contexts to ensure robustness. This expansion also includes a plan to expand the diversity of the model zoo. The current study utilized source models from the video surveillance domain; a comprehensive investigation

using models from disparate domains (e.g., general image or video datasets such as ImageNet or Kinetics) will be necessary to rigorously test the framework's generalizability.

Third, we will conduct a comparative analysis of the use of pretrained embeddings and RINN embedding. Our method avoids reliance on pretrained embeddings; this design is motivated by observations in prior work and our own empirical findings suggesting that such embeddings do not generalize reliably when the downstream task differs from the pretext task. This issue is particularly relevant in smart surveillance scenarios, where source models are often trained for general purposes, while the downstream task (e.g., anomaly detection, event classification, or object tagging) is not fixed at deployment. While our approach demonstrates greater consistency, it should be understood as a robust alternative to, rather than a direct solution for, the poor generalization of pretrained embedding-based methods under task drift. Future work could conduct a theoretical analysis of how pretrained embedding-based methods behave under task uncertainty, and whether RINN embedding-based assessment can adapt such reliance when the task alignment is known and further provides more stable guidance.

5.6 References

[1] Sam Leroux, Bo Li, and Pieter Simoens. Automated training of location-specific edge models for traffic counting. *Computers and Electrical Engineering*, 99:107763, 2022.

[2] Tetsu Matsukawa and Einoshin Suzuki. Convolutional feature transfer via camera-specific discriminative pooling for person re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8408–8415. IEEE, 2021.

[3] Clarion UK. How many cctv cameras are in london? [2024 statistics]. `https://clarionuk.com/resources/how-many-cctv-cameras-are-in-london/`, January 2024. Accessed on 2025-06-11.

[4] Young-Chan Lee, So-Yeon Lee, Byeongchang Kim, and Dae-Young Kim. Glbrf: Group-based lightweight human behavior recognition framework in video camera. *Applied Sciences*, 14(6):2424, 2024.

[5] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? *Advances in Neural Information Processing Systems*, 35:35710–35723, 2022.

[6] JunHa Hwang, SeungDong Lee, HaNeul Kim, and Young-Seob Jeong. Subset selection for domain adaptive pre-training of language model. *Scientific Reports*, 15(1):9539, 2025.

[7] Cristina Pronello and Ximena Rocio Garzón Ruiz. Evaluating the performance of video-based automated passenger counting systems in real-world conditions: A comparative study. *Sensors*, 23(18):7719, 2023.

[8] European Union. Regulation (eu) 2022/868 of the european parliament and of the council of 30 may 2022 on european data governance (data governance act). `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0868`, 2022. Accessed: 2025-05-07.

[9] European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, 2016. Accessed: 2025-05-07.

[10] Lu Jiang, Jielu Yan, Weizhi Xian, Xuekai Wei, and Xiaofeng Liao. Efficient access control for video anomaly detection using abe-based user-level revocation with ciphertext and index updates. *Applied Sciences*, 15(9):5128, 2025.

[11] Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. *Neural Networks*, page 106230, 2024.

[12] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11893–11902, 2020.

[13] Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR, 2020.

[14] 【LIVE】新宿大ガード交差点 Tokyo Shinjuku Live Ch. 新宿大ガード交差点ライブカメラ, 2021. URL `https://www.youtube.com/watch?v=xiLF6PmFZP4`. Accessed: 2025-04-09.

[15] 歌舞伎町ライブちゃんねる『Kabukicho Live Channel』. 新宿歌舞伎町ライブカメラ, 2023. URL `https://www.youtube.com/watch?v=bq7jWW7dfws`. Accessed: 2025-04-09.

[16] JazBaz Philippines. Philippines street view quad camera, agdao, davao city, 2023. URL `https://www.youtube.com/watch?v=NxG2Hor92DE`. Accessed: 2025-04-09.

[17] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2569–2578, 2020.

[18] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20392–20401, 2023.

[19] Jean-Philippe Jodoin, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. Urban tracker: Multiple object tracking in urban mixed traffic. In *IEEE Winter Conference on Applications of Computer Vision*, pages 885–892. IEEE, 2014.

[20] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143. PMLR, 2021.

[21] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019.

[22] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR, 2020.

[23] Nan Ding, Xi Chen, Tomer Levinboim, Soravit Changpinyo, and Radu Soricut. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *European Conference on Computer Vision*, pages 252–268. Springer, 2022.

[24] Huiwen Xu and U Kang. Fast and accurate transferability measurement by evaluating intra-class feature variance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11474–11482, 2023.

[25] Zixuan Hu, Xiaotong Li, Shixiang Tang, Jun Liu, Yichun Hu, and Ling-Yu Duan. Lead: Exploring logit space evolution for model selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28664–28673, 2024.

[26] Jiangbo Pei, Zhuqing Jiang, Aidong Men, Liang Chen, Yang Liu, and Qingchao Chen. Uncertainty-induced transferability representation for source-free unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 32:2033–2048, 2023.

[27] Dapeng Hu, Romy Luo, Jian Liang, and Chuan Sheng Foo. Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. *Advances in Neural Information Processing Systems*, 37:135883–135903, 2024.

[28] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[29] Daiki Chijiwa, Shin'ya Yamaguchi, Yasutoshi Ida, Kenji Umakoshi, and Tomohiro Inoue. Pruning randomly initialized neural networks with iterative randomization. *Advances in neural information processing systems*, 34:4503–4513, 2021.

[30] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

[31] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

[32] Elena Tuzhilina, Leonardo Tozzi, and Trevor Hastie. Canonical correlation analysis in high dimensions with structured regularization. *Statistical modelling*, 23(3):203–227, 2023.

[33] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[34] Lucas Hayne, Heejung Jung, and R Carter. Does representation similarity capture function similarity? *Transactions on Machine Learning Research*, 2024.

[35] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[36] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13:1393–1434, 2012.

[37] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020.

[38] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[39] Xi Chen, Haosen Yang, Huicong Zhang, Hongxun Yao, and Xiatian Zhu. Uncertainty-aware pseudo-label filtering for source-free unsupervised domain adaptation. *Neurocomputing*, 575:127190, 2024.

[40] Sam Leroux, Bert Vankeirsbilck, Tim Verbelen, Pieter Simoens, and Bart Dhoedt. Training binary neural networks with knowledge transfer. *Neurocomputing*, 396:534–541, 2020.

[41] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[42] Viet-Tuan Le and Yong-Guk Kim. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence*, 53(3):3240–3254, 2023.

[43] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[44] Alessandro Ristori. Pytorch implementation of centered kernel alignment. `https://github.com/RistoAle97/centered-kernel-alignment`, 2024. GitHub repository, commit `<hash>`, accessed 6 May 2025.

[45] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Fady Alnajjar, Ganzorig Batnasan, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th ai city challenge, 2024. URL `https://arxiv.org/abs/2404.09432`.

[46] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3): 261–272, 2020.

[47] Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. Source-free model transferability assessment for smart surveillance via randomly initialized networks. *Sensors*, 25(13):3856, 2025.

6

Conclusions and Future Research

"Alle beetjes helpen."

- an old Dutch saying

6.1 Conclusions

In this dissertation, we propose several solutions to reduce the implementation gap in deep learning-based smart surveillance applications that rely on audio, video, or a fusion of both modalities. These solutions encompass mechanisms, frameworks, and evaluation strategies, all aimed at developing systems that operate under real-world smart surveillance constraints.

6.1.1 Summary and Contributions

We first revisit the research questions described in Section 1.3.1 and Section 1.3.2. For each question, we proposed tailored solutions leveraging the characteristics of the sensor data, video, audio, or audio-visual.

Research Question 1:
What are the primary challenges and limitations of deploying conventional acoustic surveillance techniques in real-world urban environments?

The first research question raised in this dissertation stems from a key observation: contemporary deep learning models for urban acoustic surveillance often exhibit significant performance degradation when transitioned from controlled settings to dynamic, real-world operational environments.

To gain a deeper understanding of the missing pieces when applying existing machine learning models in real-world smart surveillance scenarios, an empirical investigation was carried out in chapter 2, with acoustic data collected from Ghent and Rotterdam for two years. We investigated the limitations of existing algorithms and the need for sensor- and context-specific models. Throughout the investigation, our findings highlighted the deployment-relevant issues such as location-specific variability, temporal context drift, and closed-set setting limitations. It also showed the model performance degradation when applied to data from different locations at different times, motivating the need for location- or context-specific models, as well as the need for temporally adaptive models. To further investigate system-level challenges, operational characteristics, and computational demands, a conceptual hybrid edge-cloud system was utilized in Chapter 2. This framework, which integrated stages for edge-based anomaly detection and cloud-based event tagging, served to empirically examine data flow dynamics and the interplay between these components, thereby highlighting further system-level challenges and dependencies.

Research Question 2:
How can we learn effective, robust, and generalizable data representations from complex, large-scale urban surveillance streams by leveraging the unique characteristics of multimodal surveillance data to overcome the limitations of conventional unsupervised representation learning methods, such as false negatives and information bottlenecks?

We addressed this question mainly in Chapter 4, in which we proposed to use audio and visual cues to learn representations in an unsupervised manner, thereby exploiting the complementarity between the two modalities. Recognizing the limited availability of annotated data, temporal alignment between audio and video initially serves as a natural cue for contrastive learning on the audio-visual data pairs. Temporal-based contrastive learning, however, introduces the problem of false negatives, which results in the degradation of representation learning, especially in continuous surveillance footage. We thus proposed an Embedding-based Pair Generation (EPG) mechanism, which incorporates semantic information, resulting in a more effective way of learning representation without suffering from the misleading effects of these false samples. Another mechanism introduced is a modified loss function, which tackles the information bottleneck by jointly considering multiple positives throughout the training, thereby improving the quality and diversity of learned representations. The effectiveness of this self-supervised framework in generating robust representations was validated on real-world surveillance data through multiple downstream tasks.

Furthermore, we also conducted reconstruction-based self-supervised anomaly detection in Chapter 2. A key empirical finding of this chapter is the issue of reconstruction-based representation learning with surveillance data. This further leads to the critical need for a more efficient and effective way of unsupervised representation learning to effectively handle the unconstrained nature of real-world surveillance events.

Research Question 3:
How can a principled and practical privacy-preserving framework be designed to operate on-edge, protecting sensitive attributes in acoustic data through opt-in mechanisms while maintaining compatibility with pre-existing recognition systems?

In Chapter 3, we address this issue by proposing an opt-in privacy protection framework. Existing approaches mostly follow the opt-out mechanism, which is likely due to the lack of intensively annotated datasets and thereby often not fulfilling the requirements brought by GDPR. Conversely, as likely the first work proposing an opt-in mechanism this context,

the proposed framework allows users to authorize specific inferences rather than declaring what to hide. Furthermore, the framework involved training an obfuscator designed to work with off-the-shelf pretrained models. This obfuscator is deployed on the edge device and does not require any modification to the target model, which could be deployed in the cloud. This further addressed the practical deployment constraints, offering privacy protection without the costs of redeployment when updating the model in the cloud. Due to the limitation of applicable datasets on urban sound where multiple attributes are annotated, we evaluate the proposed framework on four speech datasets. The evaluation is conducted with informed attackers aiming to retrieve identity, gender, and emotion, attributes which were not explicitly targeted during the training of obfuscator, revealing robustness under various inference scenarios. Finally, a further study on applicability for edge computing is performed with an evaluation of inference time and latency on Raspberry Pi, Jetson TX1, and server platforms, demonstrating feasibility on embedded edge devices.

Research Question 4:
How can a principled assessment framework be developed to accurately predict the transferability of pre-trained models under source-free and unsupervised conditions, in order to both guide effective model deployment in the near-term and inform the future development of more adaptable solutions?

The insights we obtained from Chapter 2, Chapter 3, and Chapter 4 lead to one crucial study: to identify the most adaptable model as the base from which to train location-specific model, and to develop methods to evaluate the model performance when no annotated data is available. In chapter 5, a transferability assessment methodology was developed that operates fully under source-free, annotation-free constraints, enabling model selection in practical scenarios where training data and annotations are unavailable. The few existing works on this topic rely mostly on assumptions of low model uncertainty or pseudo-labeling which relies on pretrained embedding space. In this chapter, we empirically demonstrated that the former assumption is risky and could lead to performance degradation in real-world scenario. We also raised the concern of confirmation bias, which limits the effectiveness of such assessments to task-aligned conditions. Finally, we introduced a structure-based embedding similarity score for the assessment, further improving generalization under task shift. Similar to the validation approach in Chapter 4, we demonstrated its improved ranking stability and robustness to representation misalignment compared to uncertainty-based and pseudo-labeling methods with real-world long-term surveillance video for training and multiple downstream tasks on annotated data. The

proposed assessment shows strong correlation with actual downstream performance across object tagging, anomaly detection, and event classification tasks using Kendall's τ.

With such an assessment, we were able to find the most suitable pre-trained model and finetune it to become a location- and context-specific model, the importance of which was previously addressed in Chapter 2 and Chapter 4.

To conclude:

In addressing these interconnected research questions, this dissertation underscores that there is no one-size-fits-all solution to the multifaceted challenges of real-world urban surveillance. Instead, it demonstrates the necessity of solutions tailored to specific data modalities and operational conditions, as explored for acoustic data in Chapter 2 and for audio-visual contexts in Chapter 4. While acknowledging the persistent gap between current deep learning capabilities and the stringent demands of robust, reliable, and ethical surveillance, a central finding of this work is that the path to effective solutions lies in a deep, data-centric examination of both the unique characteristics of surveillance data and the practical constraints of accessible resources. Indeed, this dissertation showed that these data characteristics, often sources of performance degradation, can be strategically leveraged in system design not only to overcome inherent challenges but also to unlock additional benefits, fostering more intelligent and context-aware applications.

Collectively, the contributions presented here offer practical methodologies, conceptual insights, and deployment-oriented perspectives critical for bridging the gap between research and real-world implementation in smart surveillance. This work highlights the importance of a paradigm shift in how such systems are designed and evaluated, placing greater emphasis on deployment feasibility, proactive privacy alignment, and a nuanced understanding of real-world data. Ultimately, these contributions provide a robust foundation that can inform both future academic inquiry and the applied development of more context-aware, privacy-preserving, and adaptable surveillance systems.

6.1.2 Limitations

While this dissertation addresses several core challenges in smart surveillance, certain structural and operational constraints remain unresolved. The following sections organize these limitations into four key themes reflecting gaps in deployment feasibility, evaluation scope, model behavior, and data complexity.

Deployment Constraints and Edge Readiness

For several applications developed in this work, including our privacy framework and embedding model, real-time deployment on resource-constrained edge devices presents an ongoing core challenge. Although our system designs are, in principle, edge-aware, key components such as the audio-visual embedding model and the audio signal obfuscator have not yet been specifically optimized (e.g., through distillation or compression) for efficient execution on hardware devices with modest requirements. Currently, both frameworks demand substantial compute and memory resources, which limits their immediate practical viability for certain early-stage sensing applications or multi-task operations directly at the edge.

Evaluation Coverage and Dataset Gaps

While this dissertation identifies structural shortcomings in existing surveillance datasets (such as short clip lengths, lack of temporal continuity, and annotation sparsity), it does not contribute new benchmarks or data collection protocols to directly tackle these particular issues. Furthermore, the evaluation of key components, like our opt-in privacy framework, was conducted primarily on paralinguistic speech tasks. Its comprehensive validation on more complex, real-world urban acoustic scenes is an area for future work.

Representation Stability and Transferability

Our proposed embedding and transferability frameworks offer valuable alternatives to label-intensive evaluation methods; however, they currently exhibit limitations concerning stability (e.g., sensitivity to initialization), direct interpretability, and task-specific sensitivity. Specifically, the RINN-based similarity metrics, while useful, function as indirect proxies and show known variability stemming from random initialization and potential biases when applied across diverse model architectures. Moreover, this transferability approach requires further validation with asymmetric or domain-specific similarity metrics. Its current memory footprint could also present scalability challenges for very large-scale model selection in some deployment contexts.

Real-World Data Complexity

Although this dissertation introduces methods to address specific aspects of real-world data, such as stream continuity, semantic sparsity, and signal entanglement, fundamental challenges tied to inherent data complexity persist. For instance, the current work does not incorporate explicit modeling for highly complex scene dynamics, such as robust multi-source acoustic sepa-

ration in noisy urban environments or the detailed analysis of visually dense crowd activities. Additionally, the reliance of some developed approaches on predefined event taxonomies still constrains their ability to effectively detect and tag out-of-distribution or truly ambiguous events not anticipated during model training. Such unresolved challenges persist across surveillance tasks and affect the performance, particularly under genuinely open-world conditions.

6.1.3 Broader Implications

Beyond the specific methodologies developed, this dissertation underscores two broader implications for the field: first, the critical importance of integrating privacy as a core design principle, particularly emphasizing the adoption of opt-in over opt-out models; and second, the pressing need for evaluation frameworks that realistically reflect the multi-task, resource-constrained realities of edge deployment in smart surveillance. More broadly, this work supports a **context-first** approach: that before new systems or frameworks are proposed, greater attention must be given to the properties of the data and the operational scenarios in which models will be applied, necessitating a detailed understanding of these factors. Embracing such a context-first methodology, rooted in a deep understanding of the specific use case, more effectively guides critical decisions regarding model selection, system architecture, and task definition. This, in turn, is crucial for developing solutions that are not only scalable and robust but also more ethically grounded.

6.2 Future Work

Building on the findings, identified limitations, and broader implications of this dissertation, this dissertation outlines a multi-dimensional vision for future research, organized into four primary domains. These visions, span from benchmarks, system design, to human centric and trustworthy AI to privacy and security aspects, emerge to bridge the gap between conventional deep learning techniques and the complex, practical demands of real-world deployment. Each domain encompasses several key research directions, which will be explored in the following sections.

6.2.1 Foundational Models and Data

Realistic Benchmarks

A critical direction for future work is to create standardized benchmark

datasets. While this dissertation leveraged existing or publicly available data, progress in the field is often constrained by the lack of datasets that fully capture the complexity of real-world surveillance, such as long-form, unsegmented streams with dense, multi-label events. Although unsupervised learning remains a primary research focus, high-quality annotated data is still essential for robust model development and evaluation. The domain of sound tagging serves as a cautionary example, where the lack of new and challenging datasets has led to reduced research interests and slower progress. Therefore, the development of new benchmarks with comprehensive, multi-label annotations and robust, standardized evaluation protocols would provide valuable support for the community and accelerate research.

A promising avenue to develop such robust benchmarks is the generation of synthetic datasets, which is closely related to Urban Digital Twins (UDTs) [1]. UDTs aim to create a virtual model that mirrors a city's physical systems and operational processes to improve urban planning, management, and sustainability through analysis and simulation. However, current UDTs rely on agent-based models, where these agents operate on pre-programmed and oversimplified rules, thus often fail to capture the complex, stochastic nature of objects' moving pattern and interaction. To overcome this, a more robust methodology is to script agent behavior with detailed observations such as trajectories or attributes extracted from real-world surveillance data. For instance, instead of setting an agent to move from A to B, setting the coordinate of the agent in each time frame allows us to mimic the authentic micro-behaviors, such as a person hesitating at a crosswalk or making a sudden turn. This also allow us to represent the realistic human responses when encountering incidents like fighting or car crash, which are essential for training and evaluating robust surveillance systems. This approach enables the precise control and replacement of objects (e.g., substituting pedestrians with specific characteristics), as well as the control over environmental factors such as weather and lighting. Crucially, this approach also provides a potential solution to privacy concerns by translating real-world events into anonymized trajectories to be reenacted by synthetic agents. Such approach thus avoids the intensive annotation cost to generate dense, multi-label data and preserves the complex, realistic patterns of movement and interaction found in real-world scenarios while simultaneously allowing for the generation of exhaustive ground-truth labels. The primary challenge of this research direction would lie in the significant engineering effort required to build an automated pipeline that can reliably extract behavioral data from real-world footage and integrate it as executable scripts within the simulation environment.

Comprehensive, Multi-purpose Systems

A key future research paradigm involves developing multi-purpose system rather than solving individual sub-tasks. While decomposing a complex problem into simpler components is a common and often necessary research strategy, it often leads to divergent research tracks that result in highly specialized models, which are rarely considered jointly. As a result, these highly specialized models are difficult to integrate into a cohesive system, while such integration is crucial for real-world deployment. To tackle this, a promising research avenue is the multi-task learning (MTL). As demonstrated by recent work [2], training multiple correlated tasks simultaneously within a unified network can enhance generalization and parameter efficiency. However, most MTL approaches require annotations for all tasks. Given that surveillance data is often unlabeled, a more suitable alternative is to develop unsupervised representation learning algorithms that learn informative representation for possible parallel downstream tasks. This is more suitable for surveillance domain as the collected data are often unannotated. Instead of creating numerous single-purpose models, future work should focus on architectures capable of learning compact, shared representative representations that are suitable for several downstream tasks. Such unified frameworks would offer a more scalable and practical solution in building robust, multi-purposed surveillance systems.

Long-term Adaptation

The third critical research direction is to develop self-adaptive techniques for neural network models, allowing models to sustain performance even under data shift or expended categories for classification or detection. Conventional deep learning models are static, having their weights fixed after training. Their performance degrades when environment changes over time. This is a significant issue in urban surveillance, where environments keep changing, whether due to seasonal effects, new urban construction, or evolving traffic patterns. Furthermore, as more data is collected, the task taxonomy itself may change, requiring the model to identify new, or previously unseen categories. Consequently, the development of efficient and robust model update strategy is a crucial area of study. To address this issue, several research avenues in continual learning [3], including task/class-incremental learning and domain-incremental learning, are considered simultaneously. Task/class-incremental learning allows the model to adapt to new task or new classes, particularly suitable when the needs of the application changes or different annotation is provided. Domain-incremental learning on the other hand, makes the model being adapted to weather changes, seasonal changes, or new buildings in the urban area. Despite their differences, a cen-

tral challenge for all continual learning approaches is mitigating catastrophic forgetting. Developing strategies that can address both of these scenarios while preserving existing knowledge is essential for creating systems that evolves throughout their operational lifetime.

6.2.2 System Efficiency and Decentralization

Dynamic Management of Performance-Efficiency Trade-offs

The first research direction within this research domain is the dynamic management of trade-offs between performance and efficiency during the inference phase. Whether for edge or cloud computing, the goal is to create frameworks that adaptively adjust their computational strategy by modifying the network's active components, scheduling tasks across different processing units, or switching between high-precision and highly-efficient computation. This involves simultaneously combining several emerging research avenues, including conditional computation [4], approximate computing [4], and heterogeneous parallelism [5]. Conditional computation allows a model to activate only certain parts of its architecture. For instance, running a lightweight anomaly detector continuously and only activating a complex event recognition module when an anomaly is found. Approximate computing modulates the precision of computations to balance accuracy against energy consumption in real time. For periods of low activity, the system could switch to a low-precision, power-saving mode. Finally, heterogeneous parallelism enables different types of processing units (e.g., CPUs, GPUs, NPUs) to work together simultaneously, can maximize hardware utilization to avoid waste from idle processors. Collectively, a system that adaptively manages its precision and efficiency during inference would significantly improve the practical feasibility of deploying computationally intensive applications.

Decentralized and Privacy-Aware Learning

Another promising research direction is the development of efficient and privacy-preserving methods for training and updating models across distributed edge and cloud systems. For instance, a model deployed on an edge device needs to adapt to contextual changes such as new building construction, altered traffic patterns, or seasonal variations. Given that surveillance data contains sensitive information, this adaptation must be done in a manner that preserves privacy. Federated Learning (FL) is a promising avenue for this, enabling decentralized training on local data without sharing the raw information itself. While FL has shown success in other fields, its application to large-scale, heterogeneous surveillance networks presents unique

and largely underexplored challenges. Existing FL frameworks often struggle with the extreme statistical heterogeneity across camera sites and the demands of real-time, multi-modal data streams. Although recent survey [6] on federated learning methods in heterogeneous scenarios address statistical, system, and model heterogeneity, the specific complexities of urban surveillance, compounded by the source-free and strict privacy-preserving constraints discussed throughout this dissertation, are not fully addressed. Therefore, future work should focus on creating FL frameworks specifically designed for the surveillance domain, capable of handling its unique data heterogeneity and operational demands.

6.2.3 Human-Centric and Trustworthy AI

Explainability and Interpretability

A primary research direction for achieving human-centric and trustworthy AI is the development of Explainable AI (XAI) techniques tailored for surveillance. For surveillance systems to be adopted and relied on, particularly in high-stakes environments, understanding their decision-making mechanisms is critical. This explainability is essential for enabling effective Human-on-the-Loop (HOTL) systems, where a human operator supervises the system and interfere if the action is malfunctioned. This supervisory role distinguishes HOTL from Human-in-the-Loop (HITL) [7], where where the human is more directly involved in the decision-making process. The XAI could be pursued through different research avenues. For instance, by visually highlight anomalous regions in a video frame and contrasting them with the expected normal state, human operator can easily understand the reason of system's decision. Further provide feedback or manually trigger fine-tuning or model adaptation mechanism. This explainability is essential not only for developing and refining the models themselves, but for empowering human operators to make correct, well-informed decisions based on the information provided by the model. Ultimately, the goal is to move beyond *black box* predictors and toward a more trustworthy system where human judgment and accountability are integrated.

Operator-Centric Usability and Human Factors

Another research paradigm involves shifting focus toward Usability and Human Factors. In any application where a human operator makes decisions based on the information delivered by the system, the system's practical utility is significantly impacted by the operator's interaction with and trust in its outputs. Future work should therefore investigate human factors such as *alert fatigue*, caused by an excessive number of false positives, and the loss

of trust that can result from numerous false negatives. Since the acceptable balance between these error types is highly dependent on the operational context, a key research direction is the development of systems that grant operators direct control over this trade-off. This would allow users to intuitively adjust the system's sensitivity. For instance, by prioritizing the detection of all potential threats in a high-security application, even at the cost of accepting more false alarms.

Role-Based Access Control

The third research paradigm focuses on developing dynamic mechanisms that manage and restrict data access and associated permissions with a high level of detail and precision, moving beyond static, all-or-nothing protection. The objective is to develop systems that can release minimal, task-specific sensitive information to authorized stakeholders for legitimate purposes, while ensuring robust privacy protection by default. To illustrate, data recorded by surveillance cameras serves multiple purposes, such as traffic control and crime investigation. These applications have distinct authorized privacy requirements: traffic control, for instance, typically does not require facial information, whereas crime investigation often does. Consequently, the encoded representations used for these two applications should ideally differ. While the conventional approach is to train different models for different access levels, it is inefficient in practical deployment. Instead, we propose to isolate representations containing sensitive information. For applications that do not require such privacy information, a neutral replacement for these representations will be provided. This aligns with foundational principles of privacy-aware role-based access control (RBAC) [8], which seeks to integrate privacy policies directly into access control mechanisms.

To achieve this, a promising avenue is representation disentanglement [9], which aims to explicitly isolate sensitive attributes from task-relevant features within a model's embedding space. Meanwhile, a complementary track is controlled data synthesis and feature manipulation [10]. This involves creating privacy-preserving data where sensitive attributes are controllably altered or replaced. For instance, a system could replace a specific sensitive feature (e.g., a speaker's vocal identity) with a neutral or average representation, thereby obfuscating the attribute while preserving utility for an authorized task. The successful development of these techniques would enable new system architectures for role-based access control, where different parties (e.g., law enforcement versus urban planners) are granted access only to the specific data streams necessary for their function, fundamentally enhancing both utility and data governance.

Adversarial Robustness

As intelligent surveillance systems are deployed for critical public safety applications, their robustness against malicious attacks becomes a primary concern. The integrity of the system relies on its ability to resist adversarial inputs designed to manipulate its output. Future work must therefore investigate the system's vulnerability to such attacks. The related research avenues include research into defending against attacks intended to either fake events [11], such as triggering fake fire alarms that could lead to the wasteful dispatch of emergency resources, or to hide real events, such as using adversarial patches [12] to render vehicles or individuals invisible to detection models [13]. Developing technical defenses to ensure the reliability and integrity of these systems against such manipulations is a critical and necessary direction for future research.

6.3 Final Remarks

Scientific progress is the collective achievement of countless researchers, each passing the torch to the next. In that spirit, this dissertation does not seek to provide definitive solutions, but rather to offer a solid foundation of practical and ethically grounded approaches for deploying intelligent systems in the complexity of real-world environments. This work contributes several frameworks that emphasize adaptability, privacy, and contextual awareness, with the hope that they may serve as concrete tools for future researchers. These contributions, from the opt-in privacy model to the source-free transferability assessment, are intended as starting points for continued investigation.

The impact of this research, therefore, lies not only in what was built, but in what it may enable. This work aims to contribute to the field not only by informing the design of intelligent systems, but by inspiring continued progress toward making smart surveillance truly serve its intended role: improving the safety, equity, and quality of urban life. Furthermore, there is a profound need to extend these technologies beyond urban convenience and into the domain of humanitarian aid. The same frameworks developed for analyzing complex urban scenes could be adapted to assist first responders in disaster relief scenarios, such as in the aftermath of a major earthquake, helping to categorize needs during mass casualty events when resources are critically limited. They could also provide tools for humanitarian organizations to document events in other crisis zones in a secure and verifiable manner.

Ultimately, it is our collective responsibility as researchers to not only build powerful tools, but to actively envision and advocate for their use in service of human dignity and safety. As the introductory reflection reminds us, meaning lies not in the possession of results, but in the integrity of participation. In doing so, may deep learning prove valuable in addressing the practical needs of our most vulnerable populations in their most difficult moments.

6.4 References

[1] Silvia Mazzetto. A review of urban digital twins integration, challenges, and future directions in smart city development. *Sustainability*, 16(19): 8337, 2024.

[2] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16076–16084, 2024.

[3] Li Yang, Zhipeng Luo, Shiming Zhang, Fei Teng, and Tianrui Li. Continual learning for smart city: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[4] Bartosz Wójcik, Alessio Devoto, Karol Pustelnik, Pasquale Minervini, and Simone Scardapane. Adaptive computation modules: Granular conditional computation for efficient inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21510–21518, 2025.

[5] Kuan-Chieh Hsu and Hung-Wei Tseng. Shmt: Exploiting simultaneous and heterogeneous parallelism in accelerator-rich architectures. *IEEE Micro*, 2024.

[6] Jiaming Pei, Wenxuan Liu, Jinhai Li, Lukun Wang, and Chao Liu. A review of federated learning methods in heterogeneous scenarios. *IEEE Transactions on Consumer Electronics*, 2024.

[7] Mangolika Bhattacharya, Mihai Penica, Eoin O' Connell, Mark Southern, and Martin Hayes. Human-in-loop: A review of smart manufacturing deployments. *Systems*, 11(1):35, 2023.

[8] Rajendra Muppalaneni, Anil Chowdary Inaganti, Nischal Ravichandran, Sai Rama Krishna Nersu, et al. Ai-powered role-based access control (rbac): Automating policy enforcement in enterprise environments. *Journal of Advanced Computing Systems*, 5(2):1–12, 2025.

[9] Zhenzhong Kuang, Jianan Lu, Chenhui Hong, Haobin Huang, Suguo Zhu, Xiaowei Zhao, Jun Yu, and Jianping Fan. Latent representation reorganization for face privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6646–6655, 2024.

[10] Pablo A Osorio-Marulanda, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Nicolas Moreno Reyes, and Andoni Beristain Iraola. Privacy

mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access*, 2024.

[11] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 19:1535–1546, 2023.

[12] Lihua Jing, Rui Wang, Wenqi Ren, Xin Dong, and Cong Zou. Pad: Patch-agnostic defense against adversarial patch attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24472–24481, 2024.

[13] Guixu Lin, Muyao Niu, Qingtian Zhu, Zhengwei Yin, Zhuoxiao Li, Shengfeng He, and Yinqiang Zheng. Adversarial attacks on event-based pedestrian detectors: A physical approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5227–5235, 2025.