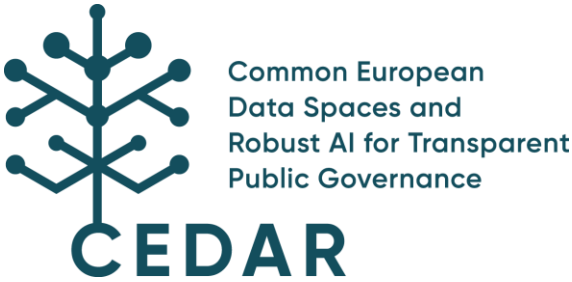

Project acronym: CEDAR

Project full title: Common European Data Spaces and Robust AI for Transparent Public Governance

Call identifier: HORIZON-CL4-2023-DATA-01

Type of action: HORIZON-RIA

Start date: 01/01/2024

End date: 31/12/2026

Grant agreement no: 101135577

[D4.1 Research Advancements in Data Analysis and Machine Learning V1]

Document description: Results from WP4 tasks, initially focusing on the SoTA analysis and refining research gaps, and then presenting results and addressing them.

Work package: WP4

Author(s): Christos Chatzikonstantinou, Stefanos Demertzis, Thodoris Semertzidis, Silvio Sorace, Marco Cipolla, Svitlana Stepanenko, Giulia Preti, Oleh Melynchuk, Jelena Sarajlic, Peter Graeff, Felix Cuadrado

Editor(s): Christos Chatzikonstanou, Silvio Sorace, Amaia Gil, Adelaide Baroncheli

Leading partner: CERTH

Participating partner: ENG, BIGS, VICOM, ART, TRE, CAU, UPM, CNT, YC

Version: 1.0

Status: Final

Deliverable type: Report

Dissemination level: PU

Official submission date: 30/09/2024

Actual submission date: 30/09/20204

The CEDAR project has received funding from the European Union's Horizon Europe project call HORIZON-CL4-2023-DATA-01 funded project Grant Agreement no. 101135577

Disclaimer

The CEDAR consortium consists of the following partners:

No.	Partner Organisation Name	Partner Organisation Short Name	Country
1	**Centre for Research and Technology Hellas**	CERTH	Greece
2	**Commissariat al Energie Atomique et aux Energies Alternatives**	CEA	France
3	**CENTAI Institute S.p.A.**	CNT	Italy
4	**Fundacion Centro de Technologias de Interaccion Visual y Comunicaciones VICOMTECH**	VICOM	Spain
5	**TREBE Language Technologies S.L.**	TRE	Spain
6	**Brandenburgisches Institut für Gesellschaft und Sicherheit GmbH**	BIGS	Germany
7	**Christian-Albrechts University Kiel**	KIEL	Germany
8	**INSIEL Informatica per il Sistema degli Enti Locali S.p.A.**	INS	Italy
9	**SNEP d.o.o**	SNEP	Slovenia
10	**YouControl LTD**	YC	Ukraine
11	**Artellence**	ART	Ukraine
12	**Institute for Corporative Security Studies, Ljubljana**	ICS	Slovenia
13	**Engineering – Ingegneria Informatica S.p.A.**	ENG	Italy
14	**Universidad Politécnica de Madrid**	UPM	Spain
15	**Ubitech LTD**	UBI	Cyprus
16	**Netcompany-Intrasoft S.A.**	NCI	Luxembourg
17	**Regione Autonoma Friuli Venezia Giulia**	FVG	Italy
18	**ANCEFVG – Associazione Nazonale Construttori Edili FVG**	ANCE	Italy
19	**Ministry of Interior of the Republic of Slovenia / Slovenian Police**	MNZ	Slovenia
20	**Ministry of Health / Office for Control, Quality, and Investments in Healthcare of the Republic of Slovenia**	MZ	Slovenia
21	**Ministry of Digital Transformation of the Republic of Slovenia**	MDP	Slovenia
22	**Celje General Hospital**	SBC	Slovenia
23	**State Agency for Reconstruction and Development of Infrastructure of Ukraine**	ARU	Ukraine
24	**Transparency International Deutschland e.V.**	TI-D	Germany
25	**Katholieke Universiteit Leuven**	KUL	Belgium
26	**Arthur's Legal B.V.**	ALBV	Netherlands
27	**DBC Diadikasia**	DBC	Greece

28	**The Lisbon Council for Economic Competitiveness and Social Renewal asbl**	LC	Belgium
29	**SK Security LLC**	SKS	Ukraine
30	**Fund for Research of War, Conflicts, Support of Society and Security Development – Safe Ukraine 2030**	SU	Ukraine
31	**ARPA Agenzia Regionale per la Protezione dell' Ambiente del Friuli Venezia Giulia**	ARPA	Italy

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.1	**11/07/2024**	**ToC released, and Partners allocation to sections**	**Christos Chatzikonstantinou (CERTH)**
0.2	**02/09/2024**	**First Consolidated Draft**	**All related partners**
0.3	**9/09/2024**	**ToC Modifications**	**Christos Chatzikonstantinou (CERTH)**
0.4	**24/09/2024**	**Second Consolidated Draft**	**All related partners**
0.5	**30/09/2024**	**Internal review**	**Boris Grivic (SNEP), Aleksander Pur (MNZ)**
1.0	**30/09/2024**	**Final version-Submitted**	**Thodoris Semertzidis (CERTH)**

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise.

Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

List of Tables

List of Figures

List of Terms and Abbreviations

AI: Artificial Intelligence

AGAN: Asymmetric learning-based Generative Adversarial Network

AMAE: Appearance-Motion United Auto-Encoder

ATLOP: Adaptive Thresholding and Localized Context Pooling

BERT: Bidirectional Encoder Representations from Transformers

BiP: Bidirectional Prediction

CBIR: Content-Based Image Retrieval

CEM: Coarsened Exact Matching

CLIP: Contrastive Language-Image Pre-Training

CNN: Convolutional Neural Network

CPI: Corruption Perceptions Index

CSO: Civil Society Organizations

DF: Deep Fake

DL: Deep Learning

DMVFN: Dynamic Multi-scale Voxel Flow Network

DTED: Deep Temporal Encoding-Decoding

ED: Entity Disambiguation

ENF: Edges based on Node Features

ESN: Echo State Network

ESS: European Social Survey

FPS: Frames Per Second

FTCN: Fully Temporal Convolution Network

GAD: Graph Anomaly Detection

GAN: Generative Adversarial Network

GHVAE: Greedy Hierarchical Variational AutoEncoder

GNN: Graph Neural Network

GPPD: Global Public Procurement Dataset

GPT: Generative Pre-trained Transformer

HDI: Human Development Index

HHF: Hashing-Guided Hinge Function

LA: Logical Access

LIVE: Learning-In-Video-strEam

LLM: Large Language Model

LMM: Large Multimodal Models

LSTM: Long short-term memory

MDS: micro-Doppler signature

MLLM: Multimodal Large Language Model

MNER: Multimodal Named Entity Recognition

MSAF: Multimodal Supervised Attentional Augmentation Fusion

NER: Named Entity Recognition

NIR: Non-Isotropy Regularization

NLP: Natural Language Processing

OCR: Optical Character Recognition

PA: Physical Access

PCA: Principal Component Analysis

PMDI: polarimetric merged-Doppler image

POI: Person-Of-Interest

PSCD: Patch-based Stride Convolutional Detector

RAG: Retrieval-Augmented Generation

RF: Radio Frequency

R-CNN: Region-based Convolutional Neural Network

RL: Reinforcement Learning

RPN: Region Proposal Networks

SA: Sentiment Analysis

SNR: Signal-to-Noise-Ratio

SoTA: State-of-The-Art

SPP: Spatial Pyramid Pooling

TED: Tenders Electronic Daily

VAE: Variational AutoEncoder

ViT-ARN: Vision Transformer Anomaly Recognition

WVS: World Values Survey

YC:YouControl

YOLO: You Only Look Once

Executive Summary

CEDAR project aims to enhance evidence-based decision-making, combat corruption, and reduce fraud in public administration. Deliverable 4.1 Research Advancements in Data Analysis and Machine Learning V1 covers the CEDAR project's initial plans and design of components, with deep analysis of the state of the art and already existing solutions. The sections describing WP4 tasks, are focusing on the SotA analysis and refining research gaps in order to identify further steps towards reaching CEDAR's goals. The deliverable includes a related work section for each task, structured on a per tool basis and the next steps that include the future directions and actions, each task should follow.

It is crucial to emphasize that the advancements in artificial intelligence (AI) and the capabilities of these technologies to process vast amounts of data and extract meaningful insights should be foundational elements in the design of Common European Data Spaces. By integrating AI-driven data analysis and machine learning techniques into the core

framework, we can ensure that these data spaces are not only efficient but also highly effective in generating actionable knowledge. This approach will facilitate widespread adoption and utilization across various sectors, enabling stakeholders to harness the full potential of data for informed decision-making, enhanced transparency, and innovation.

Each section of the deliverable, apart from Section 1 (Introduction) and Section 6 (Conclusion) covers a task of the WP. Section 2 analyzes T4.1, concerning the advancements in LLMs, NLP and NLU, to overcome the current limits of such approaches with fragmented text, dialects, jargon, and idiomatic terms that are difficult to capture automatically by general-purpose models. Named entity recognition and relationship extraction tools will perform data correlation through intra-inter-textual contents. Mention analysis tool will identify and analyze mentions of entities within a text. RAG framework and LLM-based pdf reading tools will process textual input employing LLMs and OCR models respectively.

In section 3 T4.2 is examined, a task that targets to detect corruption in multimedia content. Multimedia content can be either visual or speech content. Regarding the visual content, the video understanding techniques cover a big part of this section. More specifically, video captioning tool, object and concept/event detection tool and anomaly detection contribute in video understanding, in order to analyze and interpret video data to extract meaningful information. Disinformation detection is also crucial for corruption detection and, towards this direction, the anomaly detection tool and image and video manipulation tool will be employed, as well as retrieval and image processing tool. In respect to the audio content, speech processing in noisy environments is studied in various ways. Speech enhancement tool will improve correct audio transcription, disinformation detection in audio content can be handled through the anti-spoofing tool and keyword search and spotting tool evaluates large audio volumes simultaneously.

Section 4 focuses on T4.3, i.e., the advancements in Econometric and Graph Based Analysis. Econometric tools are employed based on SotA corruption analysis in social sciences, to discover corrupt behaviors and trends in data. Those data are compared to publicly accessible data from other countries to discover major variances that may indicate corrupt behavior. Economic literature on corruption in Public Procurement is studied, an important component of public spending, which makes it a prime target for corrupt practices. These practices result in poor outcomes and wasted public funds. A review is also performed in cryptocurrency in financial crime activities which have become a significant tool in various financial crime activities due to their decentralized and often anonymous nature. Due to those assets, they are involved in corruption, money laundering, or other illicit activities. Moreover, social science literature on corruption in Public Procurement is studied, focusing on the social culture and cultural factors that influence both the practice and the perception of illegal behavior in public procurement.

Section 5 studies data mining and correlation, introducing a number of tools to integrate, acknowledge, and correlate search outcomes in order to identify fraud and corruption operations. Correlation and graph analysis tool consolidates insights generated from previous tasks and enables both a global analysis of the entire framework and comparisons between individually analyzed components. As far as it concerns the fusion of different components, the fusion of text and visual content and fusion of audio and visual content to enhance understanding create a more comprehensive understanding of multimedia data. The financial transactions analysis tool will integrate, acknowledge, and correlate findings in order to identify fraud and corruption operations, correlating them with insights of the previous tasks. Analysis of results focuses on developing objective measures of corruption and the YouControl tool will gather and mine data from multiple sources. The footprint detector tool examine data expressing opinions from social networks.

1. Introduction

CEDAR project aims to enhance evidence-based decision-making, combat corruption, and reduce fraud in public administration. In this deliverable entitled "Research Advancements in Data Analysis and Machine Learning V1", a detailed analysis and discussion of the literature is presented for each tool. Through an extensive literature search, each tool provider attempts to refine the research gaps and then define the next steps that should be followed, to address those gaps and provide robust tools.

It is crucial to emphasize that the advancements in artificial intelligence (AI) and the capabilities of these technologies to process vast amounts of data and extract meaningful insights should be foundational elements in the design of Common European Data Spaces. By integrating AI-driven data analysis and machine learning techniques into the core framework, we can ensure that these data spaces are not only efficient but also highly effective in generating actionable knowledge. This approach will facilitate widespread adoption and utilization across various sectors, enabling stakeholders to harness the full potential of data for informed decision-making, enhanced transparency, and innovation.

Throughout the deliverable, advancements in LLMs, NLP, and NLU to overcome limitations with fragmented text, dialects, jargon, and idiomatic terms are studied, including tools. Moreover, corruption detection in multimedia content, including video understanding through captioning, object/event detection, anomaly detection, and disinformation detection using image and video manipulation tools are covered. Regarding audio content, speech enhancement, anti-spoofing, and keyword search tools are utilized. Econometric and graph-based analysis to identify corrupt behaviors and trends, comparing data with other countries, and studying economic literature on public procurement corruption and cryptocurrency in financial crimes are also analyzed. Data mining and correlation are also addressed, integrating insights from previous tasks using correlation and graph analysis tools, and enhancing understanding through the fusion of text, visual, and audio content.

Regarding the structure of the deliverable, in Section 2 the tools of the T4.1 (Advancements in LLMs, NLP, and NLU) are presented, in Section 3 the tools of the T4.2 (Advancements in Multimedia Processing), in Section 4 the ones corresponding to the T4.3 (Advancements in Econometric and Graph Based Analysis) and in Section 5 the tools of the T4.4 (Data Mining and Correlation) are presented. Finally, the Section 6 includes the conclusion of the deliverable.

2. Advancements in LLMs, NLP and NLU

2.1 Related work and research questions

2.1.1 Named Entity Recognition

When analyzing complex structured and unstructured data, especially text documents or tabular data, NER and relationship extraction are essential. Key entities, such as individuals, groups, or dates, are identified by NER, and this identification serves as the basis for understanding the connections between these entities. For example, relationship extraction models trace connections between products, whereas NER can extract items (such product names, dates, or financial figures) from tabular datasets where structured and unstructured information coexists. When it comes to databases, scientific publications, or legal documents, for example, knowing the links between entities may greatly enhance both interpretability and downstream data analytics.

However, once the entities are identified, we often encounter the problem of ambiguity. *Entity Disambiguation* (ED) helps to resolve this problem by associating the entity mentioned by the NER system with the correct one within the document's context. Without effective disambiguation, there is a risk of misinterpretations that could undermine the entire process of analysis and cross-document relationship identification.

Within a single document (*intra-document analysis*), ED allows for consistent links between mentions of entities. Without proper disambiguation, understanding the relationships between mentions would be fragmented or incorrect. ED consolidates these mentions into a coherent entity, improving the analysis of internal document relationships, such as causal, collaborative, or conflict-based relations between entities.

When analyzing multiple documents (*inter-document analysis*), ED becomes even more critical. Ambiguous mentions of entities in different documents can cause difficulties in building knowledge graphs or ontologies.

ED is crucial not only to prevent confusion but also to enable advanced text analysis applications, such as building knowledge graphs, semantic search, social network analysis, and the automatic identification of trends or patterns. Essentially, disambiguation allows for an accurate representation of relationships between entities, enabling a deeper and more precise understanding of textual content, both on a local (intra-document) and global (inter-document) scale.

A baseline implementation of NER is already available, and it uses the state-of-the-art technology, although not based on LLMs. Table 1 reports the types of entities that the NER model is able to detect.

PERSON	People, including fictional.
NORP	Nationalities, religious groups, political groups.
FAC	Facilities, buildings, airports, etc.
ORG	Institutions, agencies, etc.
GPE	Countries, states, cities.
LOC	Non GPE locations.
PRODUCT	Objects.
EVENT	Sport events, political events, named weather event (e.g., specific hurricanes), etc.
WORK_OF_ART	Title of songs, books, etc.
LAW	Named legal documents.
LANGUAGE	Named languages.
DATE	Dates and periods.
TIME	Period of time smaller than a day.
PERCENT	Percentage values, including the '%' symbol.
MONEY	Monetary values, including the unit.
QUANTITY	Measures (e.g., weight, distance, etc.).
ORDINAL	Ordinal attributes such as 'first', 'second', etc.
CARDINAL	Numeric values not falling in any of the other categories.

Table 1. Named entities.

The NER service, accessible via a RESTful API, takes as input a piece of text and return a JSON list, where each element indicates the type of entity detected and its position in the text. A sample JSON is given below:


```
[
        {
          "end": 467,
          "label": "WORK_OF_ART",
          "start": 441,
          "text": "the Nobel Prize in Physics"
        },
        {
          "end": 475,
          "label": "DATE",
          "start": 471,
          "text": "1965"
        },
        {
          "end": 505,
          "label": "PERSON",
          "start": 489,
```

```
        "text": "Julian Schwinger"
      }
]
```

2.1.2 Relationship Extraction

Relationship extraction complements NER by taking these recognized entities and determining how they relate to each other, whether through explicit associations (like "works at") or inferred ones (like "causes"). For instance, in a complex table, it could link an organization entity to a profit figure or map a drug entity to its side effects.

A baseline implementation of NER and relationship extraction is already available, and it uses the state-of-the-art technology, although not based on LLMs. The Relationship Extraction module is designed to identify relationships between entities in a sentence or document, leveraging Named Entity Recognition (NER) results to infer connections. This implementation uses the ATLOP[1] (Adaptive Thresholding and Localized Context Pooling) model for document-level relation extraction, accessible via a RESTful API. The model is pre-trained on public datasets, including those from Wikipedia, with human-readable outputs for interpretation.

In Table 2 the list of 100 relationships the model is able to detect is shown.

killed	mother	author	publisher	cast member	sister city	part of	date of birth	lyrics by	applies to jurisdiction
sexually assaulted	spouse	member of sports team	owned by	producer	legislative body	original language of work	date of death	located on terrain feature	product or material produced
injured	country of citizenship	director	located in the administrative territorial entity	award received	basin country	platform	inception	participant	unemployment rate
arrested	continent	screenwriter	genre	creator	located in or next to body of water	mouth of the watercourse	dissolved, abolished or demolished	influenced by	territory claimed by
head of government	instance of	educated at	operator	parent taxon	military branch	original network	publication date	location of formation	participant of
country	head of state	composer	religion	ethnic group	record label	member of	start time	parent organization	replaces
place of birth	capital	member of political party	contains administrative territorial entity	performer	production company	chairperson	end time	notable work	replaced by
place of death	official language	employer	follows	manufacturer	location	country of origin	point in time	separated from	capital of
father	position held	founded by	followed by	developer	subclass of	has part	conflict	narrative location	languages spoken, written or signed
killed	child	league	headquarters location	series	subsidiary	residence	characters	work location	present in work

Table 2. List of detectable relationships.

The relationship extraction service takes as input a piece of text and return a JSON list, where each element indicates the type of relationships together with the associated types of entities detected A sample JSON is given below:

[1] https://github.com/wzhouad/ATLOP

```
"relationship 1": [
{
      "entity 1": [
      "Richard Phillips Feynman",
      "PERSON"
      ]
},
{
"relationship": "conflict"
},
{
      "entity 2": [
      "World War II",
      "EVENT"
      ]
}
],
], "relationship 2": […]
```


Each detected relationship is given as a JSON list of three elements: the two entities and their type and their relationship.

As for NER, the system relies on Python and FastAPI as its web server, with deployment managed through Docker containers for ease of scalability and maintenance.

2.1.3 Retrieval-Augmented Generation (RAG) Framework

LLMs

Large Language Models (LLMs) have revolutionized natural language processing (NLP) by demonstrating remarkable capabilities in understanding and generating human-like text. Despite these advancements, LLMs encounter significant challenges, particularly in tasks that require up-to-date information or specialized domain knowledge [Kandpal et al., 2023]. Common issues include "hallucinations" [Zhang et al., 2023] — the generation of convincing yet inaccurate content — stemming from the model's reliance solely on pre-trained data, which may be outdated or incomplete. Additionally, LLMs often exhibit non-transparent reasoning processes, making it difficult to trace the origins of their outputs.

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has emerged as a promising solution to address these challenges [Lewis et al., 2020]. RAG integrates external knowledge sources with LLMs, allowing these models to retrieve relevant information from vast, dynamic repositories like databases, search engines, or knowledge graphs (**Error! Reference source not found.**). By incorporating external data during the generation process, RAG enhances the factual accuracy and credibility of the model outputs, especially in knowledge-intensive tasks where precise and current information is crucial.

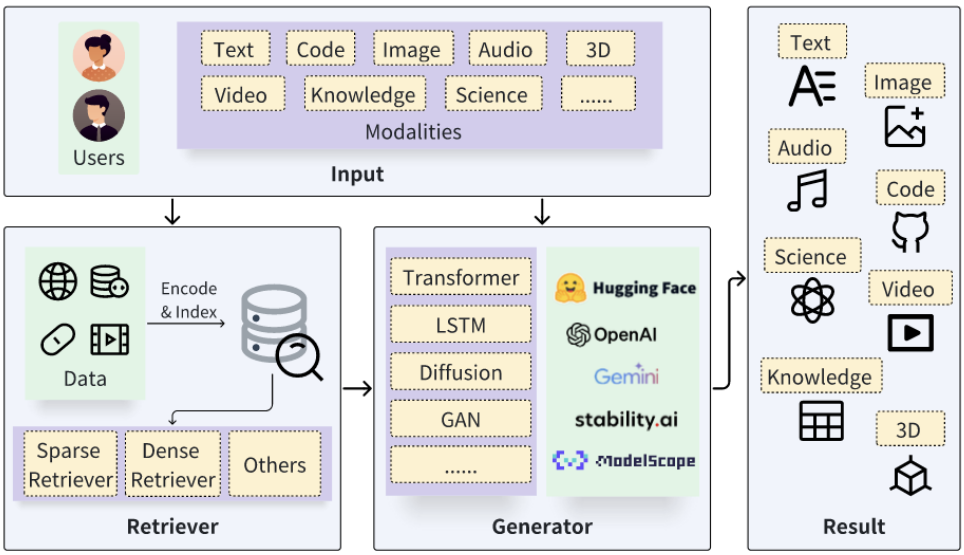

Figure 1 A generic RAG architecture. User queries in various modalities are input to both the retriever and the generator. The retriever fetches relevant information from data sources, while the generator uses this information to produce outputs across different modalities.

The core advantage of RAG lies in its ability to synergistically combine the inherent linguistic and contextual understanding of LLMs with the extensive, continuously updated information from external sources. This integration not only mitigates the problem of hallucinations by grounding the model's outputs in verifiable data but also allows for the inclusion of domain-specific knowledge that may not be present in the original training corpus of the LLM.

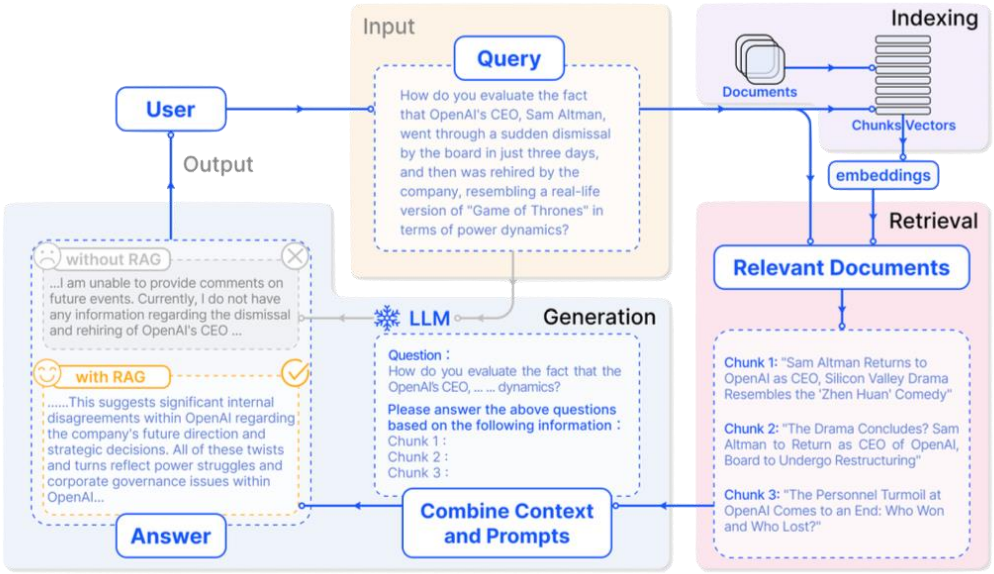

Figure 2 The basic RAG workflow

Naïve RAG Paradigm

The basic workflow of Retrieval-Augmented Generation (RAG), also referred as Naive RAG [Gao et al., 2023] involves three key phases: indexing, retrieval, and generation [Huang, 2024] (**Error! Reference source not found.**).

Indexing is the first step, where external sources are processed and organized to create an efficient system for retrieving relevant information. This involves text normalization, segmentation, and the creation of semantic vector representations to ensure fast and accurate retrieval.

Retrieval phase uses these indexed data to search for and rank documents based on their relevance to a specific query, employing both traditional methods like BM25 [Beaulieu et al., 1997] and modern approaches using pretrained language models such as BERT [Devlin, 2018] to capture semantic nuances.

Generation phase synthesizes the retrieved information with the query to produce a coherent and contextually relevant output, combining retrieved data to enhance the accuracy and informativeness of the generated content.

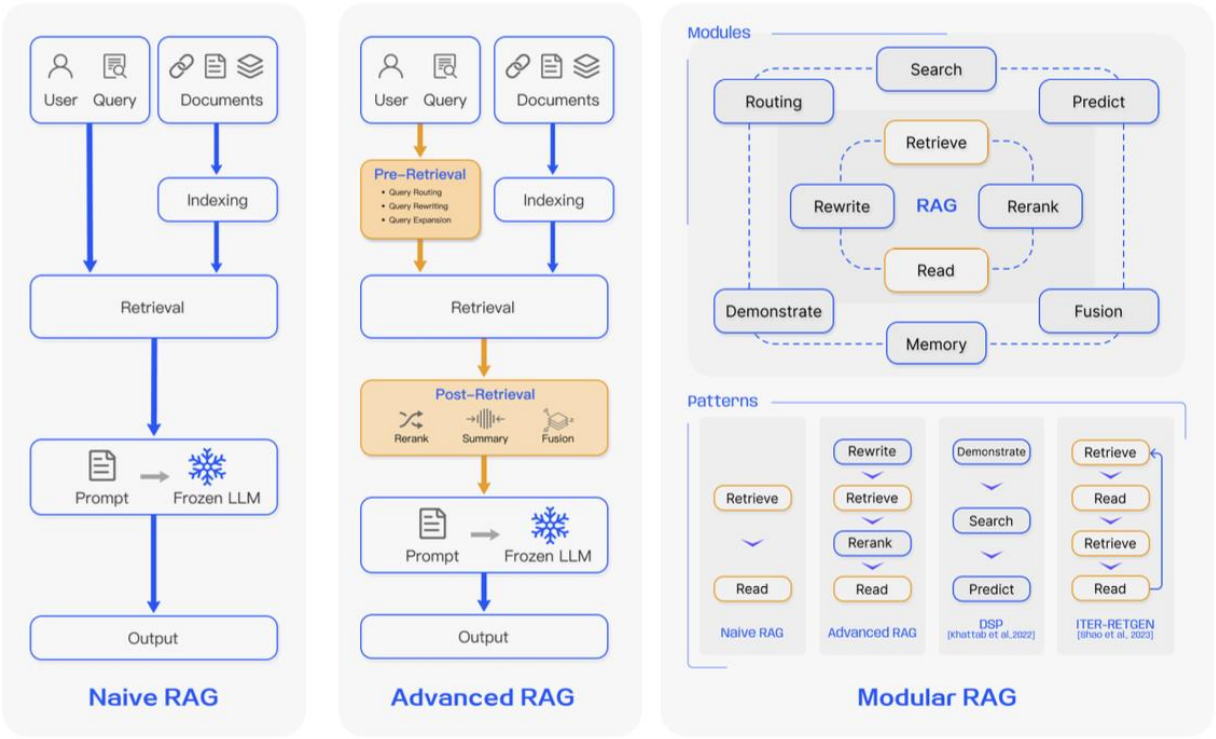

Figure 3 Comparison between the three RAG paradigms

This approach, also known as the "Retrieve-Read" framework [Ma et al., 2023], helps to ground the language model's outputs with external information. However, Naive RAG often suffers from challenges such as retrieving irrelevant or redundant information and generating responses that may still lack precision or context, limiting its effectiveness for more complex or nuanced tasks.

Unified RAG Paradigm

Advanced / Modular RAG builds on the foundational principles of Naive RAG by introducing more sophisticated techniques to enhance both the retrieval and generation processes [Ilin, 2023] (**Error! Reference source not found.**). According to [Huang, 2024] advanced RAG paradigm organizes workflow into four main phases: pre-retrieval, retrieval, post-retrieval, and generation (**Error! Reference source not found.**).

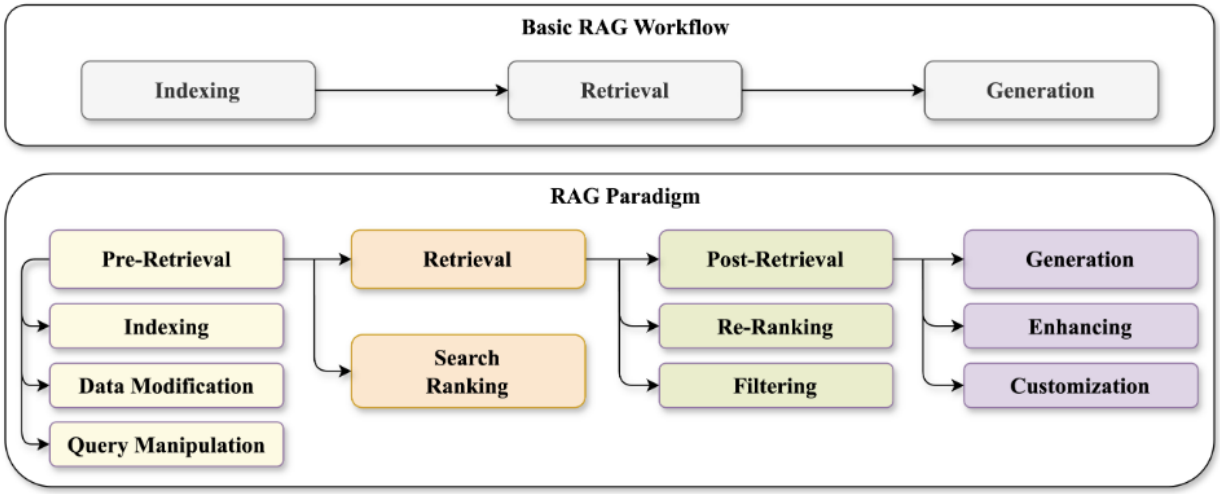

Figure 4 The RAG core concepts with basic workflow

Pre-Retrieval phase

In the **pre-retrieval** phase, foundational tasks like indexing and query manipulation prepare data for efficient access, ensuring that the retrieval process is optimized for relevance and accuracy. This phase involves several key tasks, including **data modification**, **query manipulation** and **indexing**.

Data modification focuses on enhancing the quality and relevance of the indexed data. These methods can generally be divided into two categories: *Internal Data Augmentation* and *External Data Enrichment*. Internal Data Augmentation aims to enhance the value of the information already present within documents or models. In contrast, External Data Enrichment involves incorporating additional data from external sources to address gaps, add context, or expand the scope of the existing content. The whole process involves removing redundant or irrelevant information and adding supplementary data, such as metadata, to enrich the retrieval process. [Huang, 2024]

Query manipulation, involves refining and adjusting the user's original query to better match the indexed data, using techniques such as query expansion [Izacard, 2020] or reformulation to increase the likelihood of retrieving relevant documents. [Ma et al., 2023; Zheng et al., 2023]

Indexing is the process of organizing data into a structured format, such as creating an index of external sources that can be quickly searched to find relevant information. This step often includes text normalization techniques like tokenization, stemming, and the removal of stop words to optimize the data for retrieval.

Retrieval phase

The retrieval phase in the RAG framework is key to finding information that is most relevant to the user's query. **The search and ranking process** enhances the relevance and accuracy of outputs through various specialized strategies. Atlas proposed in [Izacard et al., 2023] employs few-shot learning to optimize document relevancy, while AAR [Yu et al., 2023] adapts retrieval preferences to better suit LLMs. IRCOT [Trivedi et al., 2022] integrates retrieval with logical reasoning, and FLARE [Jiang et al., 2023] triggers retrieval based on model confidence. SURGE [Kang et al., 2023] utilizes subgraph retrieval from knowledge graphs to enhance contextual understanding, and PRCA [Yang et al., 2023] refines content for generation with a reward-driven approach. MEMWALKER [Chen et al., 2023] uniquely manages long-context queries using an internal search within a memory tree, focusing on iterative navigation rather than just initial retrieval.

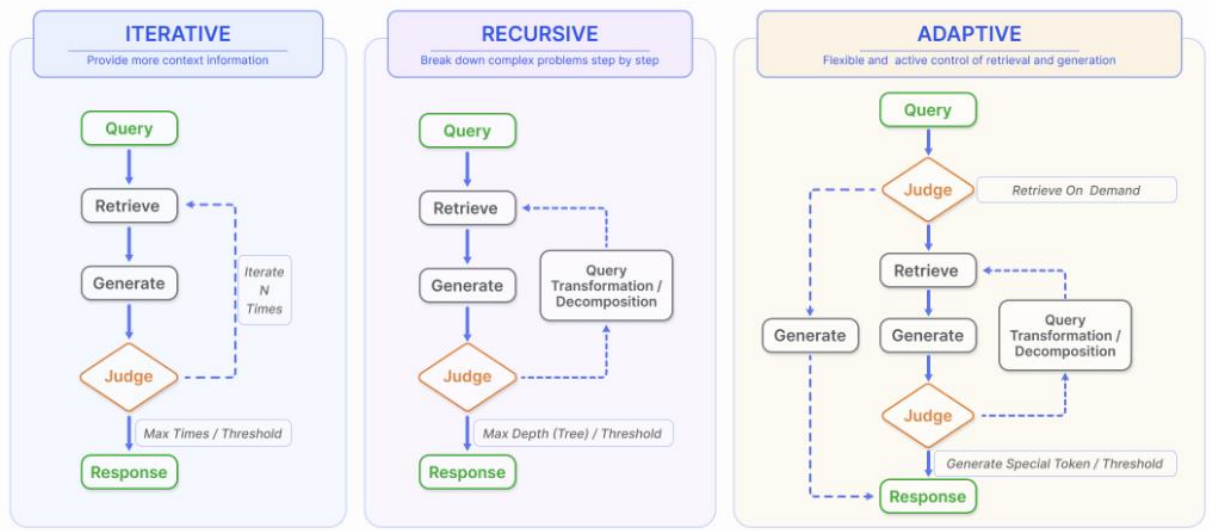

Figure 5 Three types of RAG retrieval augmentation processes.

Retrieval strategies in RAG framework customize the retrieval process to suit specific application needs, with each strategy enhancing retrieval and ranking (**Error! Reference source not found.**). **Basic RAG strategies**, like Atlas, are linear, enhancing models directly through simple retrieval phases. **Iterative strategies**, such as IRCOT, involve multi-step refinements, integrating retrieval with reasoning or continuously adjusting based on feedback, ideal for complex queries. **Recursive strategies** like those used in SURGE and MEMWALKER break down hierarchical data into simpler queries using knowledge graphs. **Adaptive strategies**, like those in AAR and FLARE, dynamically adjust retrieval based on the immediate context, optimizing relevance and precision for dynamic environments [Huang, 2024].

Post Retrieval phase

Filtering and re-ranking modules serve different purposes in the post-retrieval phase of RAG Framework. Filtering is employed to remove irrelevant or poor-quality documents from the retrieved collection, thus reducing the volume of documents and enhancing the efficiency and effectiveness of the processes that follow. On the other hand, re-ranking involves arranging the remaining documents by their relevance or usefulness for the specific task, typically giving preference to those that can significantly improve the quality of the generated responses, particularly in scenarios sensitive to the quality of the response.

Filtering modules intend to refine document sets in Retrieval-Augmented Generation (RAG) systems, aiming to improve relevance and reduce computational load. Self-RAG [Asai et al., 2023] uses a self-reflection mechanism with "reflection tokens" to evaluate and retain only the most pertinent documents, enhancing efficiency by leveraging the model's internal capabilities. Similarly, BlendFilter [Wang et al., 2024a] applies the LLM itself as a filter, assessing and removing irrelevant documents from various query augmentations. In contrast, RECOMP [Xu et al., 2024] uses selective augmentation, generating summaries from retrieved documents that are pre-appended to inputs for the language model, dynamically filtering out irrelevant information. CRAG [Yan., 2024] employs a decompose-then-recompose strategy, splitting documents into knowledge strips that are evaluated and reassembled based on their relevance. Dynamic filtering techniques are also prominent, such as FiD-TF Token Filtering [Berchansky et al., 2023], which removes less relevant tokens during decoding, and CoK self-consistency based filtering [Li et al., 2023], which processes only questions with uncertain answers, enhancing response accuracy. FILCO [Wang et al., 2023b] introduces a comprehensive approach with strategies like String Inclusion, Lexical Overlap, and Conditional Cross-Mutual Information, refining content at the sentence level and training a context filtering model to predict the most useful context during inference, thereby optimizing the relevance and accuracy of the generated output.

Re-ranking modules in RAG systems reorder a large number of potentially relevant documents to prioritize those most likely to impact the final output effectively. These methods leverage unsupervised techniques, such as pointwise, listwise, or pairwise methods, where models like In-Context RALM [Ram et al., 2023a] use a zero-shot approach with existing language models to rank documents without supervised training, enhancing relevance based on semantic understanding. Conversely, supervised re-rankers like Re2G [M. Glass et al., 2022] and FiD-Light [Hofstätter et al., 2023] involve fine-tuning large language models (LLMs) on specific datasets, employing mechanisms such as listwise auto-regressive ranking to improve relevance determination. Additionally, data augmentation techniques are used to enhance re-ranking by generating pseudo-relevance labels or using synthetic data to train re-rankers, boosting retrieval accuracy and aligning retrieved content more closely with user queries, as seen in methods like DKS-RAC [Huang et al., 2023] and PROMPTAGATOR Framework [Dai et al., 2022] which enrich training processes and optimize re-rankers for better performance.

Generation

Enhancing module in RAG Framework aim to improve the quality and relevance of generated outputs by integrating retrieved content in various ways, and these can be categorized into three main approaches: **enhancing with queries**, **enhancing with ensemble approaches**, and **enhancing with feedback loops**. The 'Enhance with Query' approach merges the retrieved documents with the original query, allowing the generator to leverage both sources for a final output that remains aligned with the user's intent and contextually enriched, exemplified by models like RETRO [Borgeaud et al., 2022] and In-Context RALM [Ram et al., 2023b] that concatenate queries with retrieved documents into a single input sequence. The 'Enhance with Ensemble' strategy synthesizes information from multiple sources to reconcile conflicting details and blend diverse perspectives for a coherent response, as seen in FiD [Izacard, 2020] and REPLUG [Shi et al., 2023] , where multiple documents are processed independently then aggregated based on relevance. Finally, the 'Enhance with Feedback' method introduces iterative refinement into the generation process by incorporating feedback loops that evaluate and adjust initial outputs based on set criteria, with examples like PRCA [Yang et al., 2023], which distills retrieved information to optimize content, and DSP [Khattab et al., 2022], which refines queries and passages through a multi-hop process for increasingly accurate and coherent results.

Customization module adapts content to match the user's personality and specific needs, focusing on content alignment and contextual adaptation. PersonaRAG [Zerhoudi, 2024] dynamically adjusts content based on user profiles and session behavior while ERAGent [Shi et al., 2024] uses a Personalized LLM Reader to customize responses according to specific user preferences. Finally, ROPG [Salemi et al., 2024] employs a dynamic retrieval strategy that adjusts before and after generation, aligning retrieved content and responses with user-specific knowledge and preferences.

2.1.4 LLM-based PDF reading

Optical Character Recognition (OCR) will be employed for PDF-reading. OCR is the process of, electronic or mechanical, conversion of images of typed, handwritten or printed text into machine-encoded text. It can be applied to scanned documents, photos of documents, scene photos (such as the text on billboards and signs in a landscape photo, or license plates in cars), or images with subtitle text superimposed on them (like in a television broadcast). In this section, three prominent open source OCR tools are presented: Google's Tesseract, PaddlePaddle's PaddleOCR, and Jaided AI's EasyOCR.

Tesseract[2] is one of the most well-known and widely used open-source OCR engines. Developed by Hewlett-Packard and now maintained by Google, Tesseract supports over 100 languages and can handle multilingual OCR. It excels in recognizing printed text with high accuracy, especially when combined with pre-processing steps such as noise reduction and binarization. Tesseract is highly customizable, allowing users to train it on new fonts and languages, making it a versatile tool for various OCR tasks.

[2] https://github.com/tesseract-ocr/tesseract

PaddleOCR[3], developed by PaddlePaddle, is an advanced OCR tool designed to handle a wide range of text recognition tasks, including both printed and handwritten text. It utilizes deep learning models to achieve high accuracy in text detection and recognition, even in challenging scenarios such as curved text or text within complex backgrounds. Its strength lies in its ability to integrate seamlessly with other PaddlePaddle AI tools, offering robust performance for large-scale and real-time OCR applications. However, PaddleOCR can be resource-intensive, necessitating powerful hardware for optimal performance.

EasyOCR[4], developed by Jaided AI, is an open-source OCR tool that focuses on simplicity and ease of use while delivering competitive accuracy. It supports over 80 languages and can recognize a variety of scripts, including Latin, Chinese, Arabic, and Cyrillic. EasyOCR's lightweight design makes it suitable for integration into various applications, such as mobile apps and web services. Its pre-trained models provide good accuracy out-of-the-box, though customization options are more limited compared to Tesseract. EasyOCR strikes a balance between performance and ease of deployment, making it a practical choice for developers needing quick and effective text extraction capabilities.

LLaVA LLM-Vision model [Liu, 2024] is an advanced AI system designed to seamlessly integrate visual and linguistic understanding. Utilizing extensive pre-training on multimodal data, LLava excels in interpreting both images and text. Its sophisticated architecture combines vision transformers with large language models, enabling it to generate coherent and contextually relevant responses based on visual inputs.

2.1.5 Mentions analysis

The advent of large language models, particularly GPT (Generative Pre-trained Transformer), has significantly transformed the landscape of Natural Language Processing (NLP). GPT models have been leveraged across various NLP tasks due to their ability to understand and generate human-like text based on extensive training on diverse datasets. Named Entity Recognition (NER), a fundamental task in NLP involving the identification and classification of entities such as names, locations, and organizations within text, has particularly benefited from the capabilities of GPT. Unlike traditional NER systems that rely on manually labeled datasets and rule-based approaches, GPT models can perform NER with minimal examples through few-shot or zero-shot learning. This adaptability makes GPT highly effective in recognizing entities across different domains and languages, reducing the need for extensive annotated data. Recent research has demonstrated that GPT's contextual understanding and in-context learning capabilities enable it to achieve comparable or even superior performance to traditional supervised methods in certain NER tasks, paving the way for more scalable and accessible NER solutions.

[Zeghidi, 2024] evaluated Few-Shot Prompting with Large Language Models for Named Entity Recognition (NER), highlighting that while large models like GPT-4 show strong adaptability to new entity types and domains with minimal data, there remains a performance gap compared to fully supervised models. The study also explored the impact of prompt engineering, output formatting, and context length on performance, emphasizing the potential of Few-Shot Learning to reduce the dependency on large labeled datasets.

[Covas, 2023] demonstrated the effectiveness of using GPT for Named Entity Recognition (NER) in identifying comparable companies for equity valuation. The study compared GPT's performance with standard NER methods and found that GPT achieved higher precision and success rates. The research suggests that GPT-based entity extraction could automate the creation of peer groups for company valuation, outperforming traditional methods that rely on manual annotation.

[Wang S. et al., 2023] proposed a novel multimodal named entity recognition (MNER) method called Prompt ChatGPT In MNER (PGIM), which leverages ChatGPT as an implicit knowledge engine to enhance the performance of MNER tasks. The approach involves using a Multimodal Similar Example Awareness module to select suitable examples, which are

[3] https://github.com/PaddlePaddle/PaddleOCR?tab=readme-ov-file

[4] https://github.com/JaidedAI/EasyOCR

then used to generate auxiliary refined knowledge through ChatGPT. The study showed that PGIM outperformed existing state-of-the-art methods on two classic MNER datasets, addressing the limitations of previous multimodal approaches.

2.2 Next steps on LLMs, NPL and NLU

The application of a single Large Language Model (LLM) offers an advanced alternative by combining both NER and relationship extraction capabilities within a unified model. LLMs, particularly models like GPT or BERT, are fine-tuned to handle not just natural language but also structured data, including complex tables. An LLM trained or fine-tuned for specific tasks can directly parse tabular data, detect entities, and extract the relationships between them without needing separate modules for each task. This approach simplifies the architecture, offering a more cohesive pipeline for data parsing, especially when data includes mixed formats (text, tables, or even diagrams). The challenge, however, lies in the need for fine-tuning these models, which demands large amounts of annotated data to achieve high performance across diverse data sources and tasks.

Fine-tuning any relationship extraction or NER model, especially an LLM, requires vast and high-quality annotated datasets. The process involves labeling entities and their relationships across various texts or datasets, which can be resource-intensive. The success of these models depends heavily on the accuracy of these annotations, as incorrect or incomplete data will lead to poor model performance. For example, a relationship extraction model fine-tuned on legal texts will need substantial manual annotation of legal entities (such as contracts, clauses, and laws) and their connections.

2.2.1 Proposed Retrieval-Augmented Generation (RAG) architecture

The proposed RAG architecture for the CEDAR project is designed to optimize the retrieval and generation of public tender information. This system follows a modular approach, with each module serving a distinct purpose in one of the four phases: **Pre-Retrieval**, **Retrieval, Post-Retrieval** and **Generation**. These phases ensure flexibility, scalability, and efficiency in handling the complex requirements of public tenders. It is important to note that this is a preliminary architecture, and the modules may be altered or modified during the implementation phase to accommodate the evolving needs of the project. The proposed architecture can be shown on **Error! Reference source not found.**.

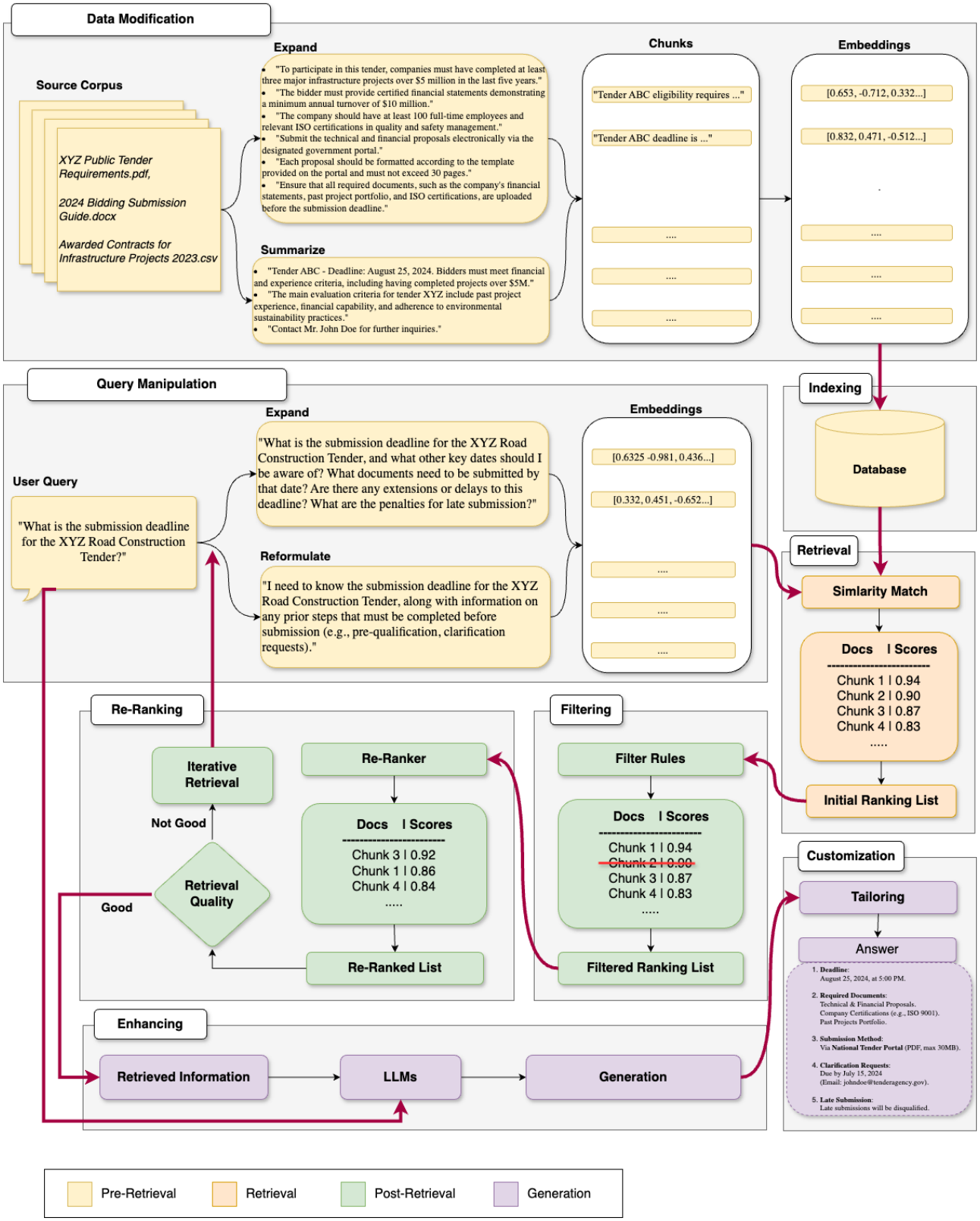

Figure 6 CEDAR RAG-LLM Framework proposed architecture

Pre-Retrieval Phase

The **Data Modification** module is responsible for preprocessing the documents in the source corpus. This module divides documents into meaningful chunks that are semantically coherent, enabling efficient downstream indexing and retrieval. Summarization and content restructuring techniques may be employed to ensure long documents are condensed while retaining key information. This preprocessing ensures that the data is ready for effective semantic search.

The **Indexing module** converts these chunks into dense vector representations (embeddings) that capture the semantic meaning of the content. These embeddings allow for efficient retrieval based on conceptual similarity rather than keyword matching. Additionally, metadata may be incorporated during indexing to support domain-specific searches.

The **Query Manipulation** module enhances the user's query before the retrieval process begins. It handles two critical tasks: **query expansion** and **query reformulation**. Query expansion adds relevant details that may not be explicitly stated by the user but are essential for retrieving comprehensive information. Query reformulation adjusts the user's query to better align with the structure of the data in the corpus, ensuring a higher quality of retrieved results. For Data Modification and Query manipulation module processes, pre trained LLMs can be used to produce extended and reformulate texts (**Error! Reference source not found.**).

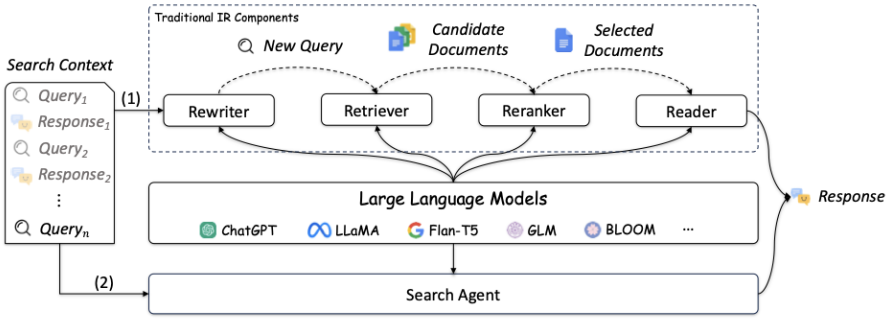

Figure 7 Overview of existing studies that apply LLMs into Information Retrieval (IR). (1) LLMs can be used to enhance traditional IR components, such as query rewriter, retriever, re-ranker, and reader. (2) LLMs can also be used as search agents to perform multiple IR tasks. [Y. Zhu et al., 2023a]

Retrieval Phase

The **Retrieval module** is responsible for matching the user's query with relevant content from the indexed corpus. By comparing the vector representations of the query to the embedded document chunks, the system identifies the most relevant chunks of text. The initial results are ranked based on their similarity scores and passed on for further refinement.

Post-Retrieval Phase

After the retrieval phase, the **Filtering module** applies a series of domain-specific rules to refine the retrieved results. It removes irrelevant, redundant or less useful chunks, ensuring that only pertinent information is retained for further processing.

The **Re-Ranking module** reorders the filtered chunks after filtering is applied. The process is based on a combination of similarity scores and domain-specific rules to ensure that the most relevant and important chunks are prioritized.

An **iterative paradigm** is employed in the retrieval phase. If the retrieved and refined results from the post-retrieval phase do not meet predefined quality standards or do not provide sufficient information to generate a satisfactory response, the system iterates back to the retrieval phase. The query may be refined, or additional information may be retrieved based on adjusted criteria.

Generation Phase

The **Enhancing module** is responsible for combining the final query (whether extended or not) with the relevant retrieved data. This module ensures that the original user query and the most relevant chunks are integrated into a coherent input, ready to be processed by the language generation model.

Finally, the **Customization module** uses the query history and contextual information from previous interactions to adapt the generated response to the specific user or session. By leveraging past queries, this module ensures that the generated answers are consistent with previous requests and tailored to the user's ongoing needs.

3. Advancements in Multimedia Processing

3.1 Related work and research questions

3.1.1 Video Captioning

In CEDAR project, the video captioning tool will be employed for corruption detection through video understanding. In this section, SoTA methods that can accomplish this goal will be presented.

Machines struggle to explain visual data at varying levels of detail like humans do. This problem becomes significantly more complicated when dealing with video data. Video captioning is the process of understanding and creating descriptive text for a video. Video captioning requires both understanding the visual content and effectively capture its semantics through human-like descriptions. Achieving this level of understanding requires the collaboration between computer vision and natural language processing research areas. The captions produced provide essential resources for video search, accessibility for visually impaired individuals, and human-robot interaction. Deep learning strategies have emerged as powerful tools in addressing the complexities of video captioning. Encoder-decoder pipelines are the most widely used architectures for deep learning video captioning. In Figure 8 the general structure of the encoder-decoder architecture is illustrated.

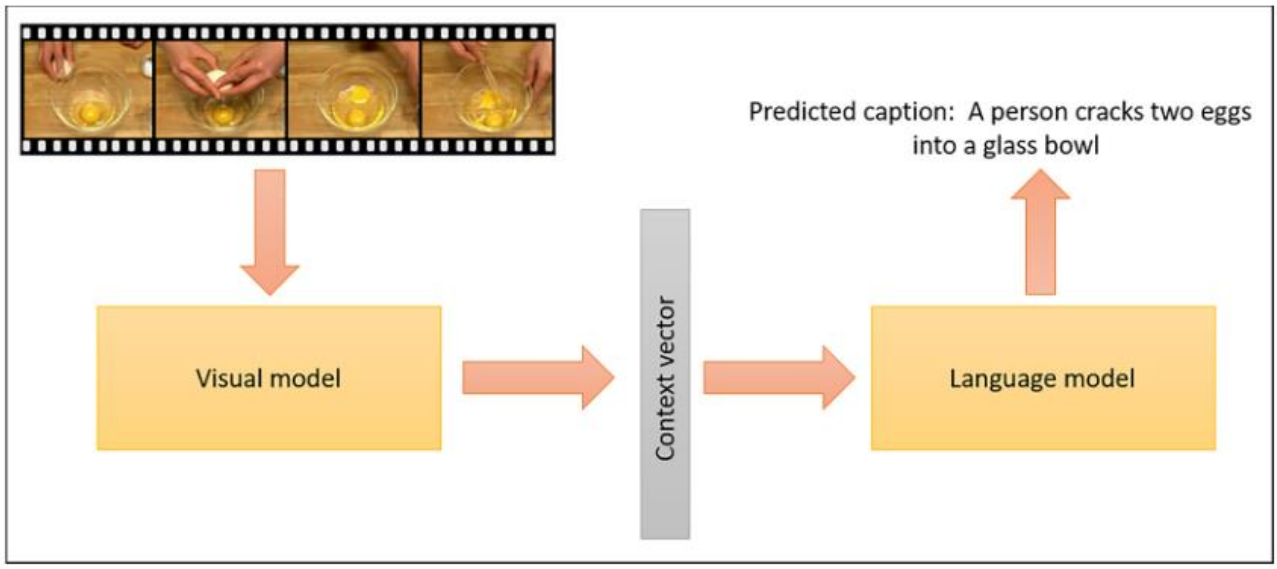

Figure 8 Overview of the basic model of video captioning task: video is first encoded into a sequence of feature vectors (context vector) using a video encoder (visual model). The context vector is then passed to the text decoder (language model) to generate a caption.

A widely used module in video captioning is the attention module, employed in machine learning algorithms to assign weights or importance to different parts of a video while creating captions. This approach allows the video caption model to focus on key frames or segments for accurate descriptions. Attention module also allows the model to choose and adaptively focus on certain sections of a video, instead of handling the entire video as a single entity. Different types of attention are employed by video captioning methods:

Spatial and temporal attention: Yan et al [Yan, 2019], as illustrated in **Error! Reference source not found.,** use global features (CNN), motion features (C3D) and features at object-level (R-CNN). Those features are fused with a two-stage attention mechanism and LSTM. In the first step, spatial attention is employed to select local features that have a higher spatial attention weight. The second stage involves selecting global and motion features using temporal attention. The attention mechanism's fused features are then sent into the LSTM decoder as input.

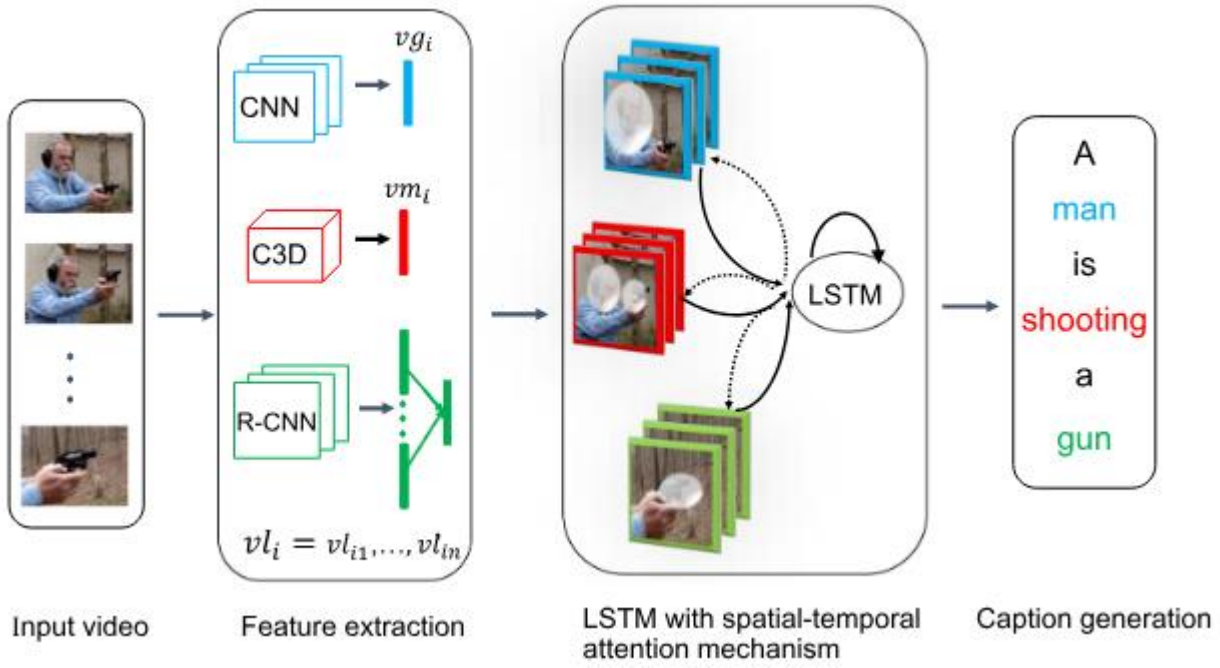

Figure 9 Illustration of the STAT [Yan,2019] video captioning framework, based on spatial-temporal attention mechanism

Hierarchical Attention: Wu et al [Wu, 2018] employ multimodal features, more specifically temporal features, motion features and semantic label features. Three hierarchical attention layers are utilized in a progressive attention manner, beginning with highly correlated modalities and moving on to the less correlated.

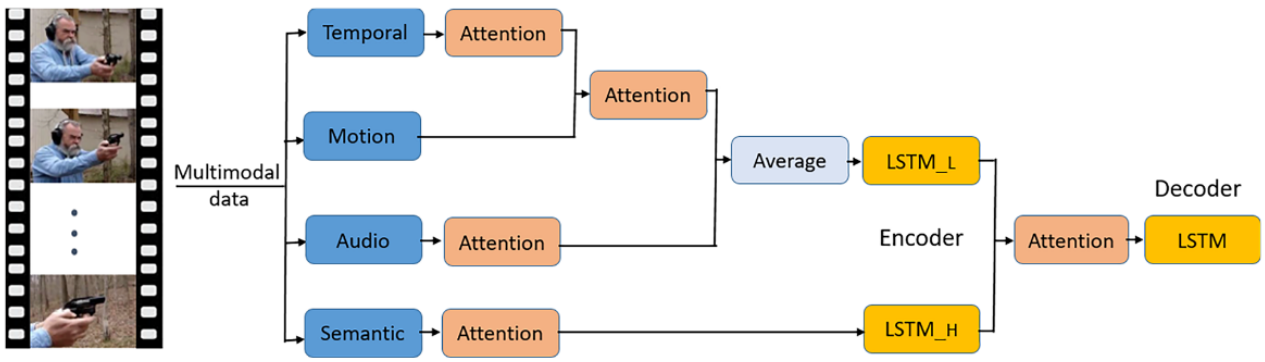

Figure 10 Video captioning with hierarchical attention used in [Wu, 2018]

Self-Attention and Multi-Head Attention: In [Choi, 2022], a parallel architecture is proposed enabling simultaneous captioning and localization. The authors try to address the challenge of information bottleneck in parallel decoding by filtering out unnecessary information using a gating network. A multi-head attention mechanism is also introduced to exploit multiple information for localization and capturing.

Video captioning methods using Reinforcement Learning (RL) have also been proposed. RL is a machine learning strategy that trains agents to make consecutive decisions to maximize cumulative rewards. Video captioning may be also formulated as a reinforcement learning problem. Deep learning networks (e.g. CNN or RNN) are the 'agents' that interact with the 'environment' of words and video features. The prediction probabilities or output words of deep learning models serve as the 'actions' that influence the internal 'state' of the models, representing weights and biases, as presented in Figure 11 [Zhao, 2021]. The 'agent' is given a 'reward' to motivate the training process. In video captioning, the 'reward' value is often derived using CIDEr or METEOR scores, which assess the similarity between produced and ground-truth captions.

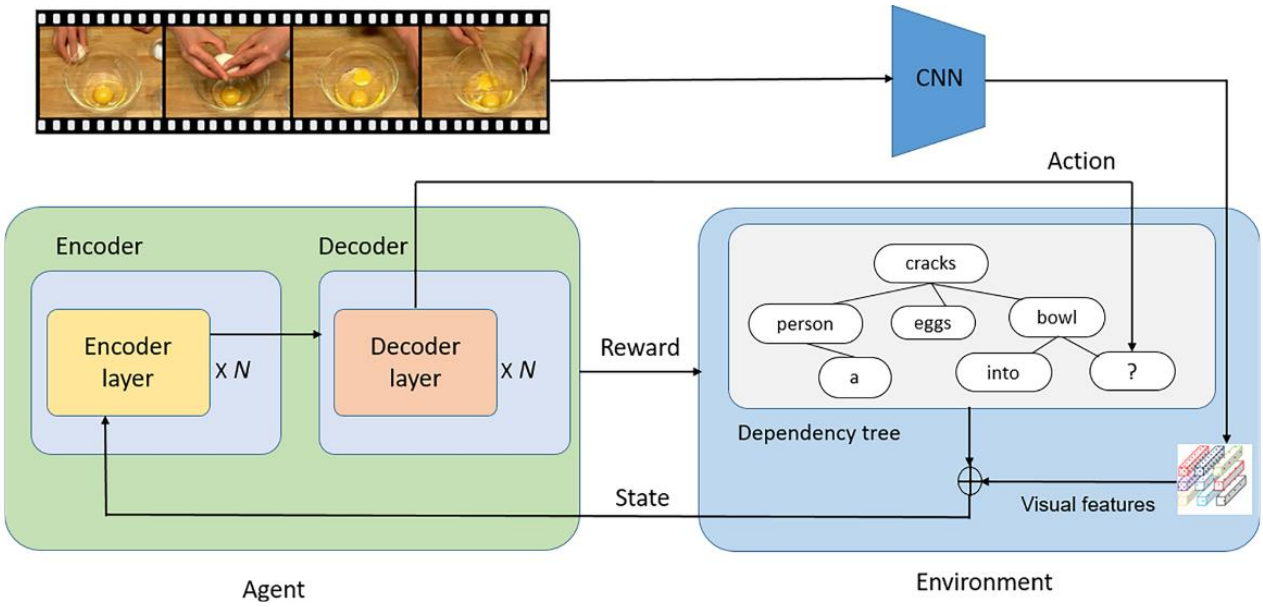

Figure 11 Video captioning with reinforcement learning [Zhao 2021]

Moreover, graph-based methods can be adapted to video captioning, exploiting their ability to capture complex dependencies and relational structures in data. Graph-based approaches are algorithms and techniques used to analyze and interpret structured data represented as graphs. A graph is made up of nodes (elements) and edges (connections) that reflect interactions or relationships between nodes. In video analysis, nodes represent objects, actions, or scenes in the frame, while edges show the interactions between them. These links might be chronological, geographical, or semantic. A graph-based method is depicted in Figure 12Error! Reference source not found. [Zhang 2020], consisted of three modules: A graph encoder exrtacting information from the source video and ground truth to build an improved graph represenation. A grounding module using region grounding and object localization to predict the relevant object words. And a sentence generator using a two-layer LSTM with a stacked design, including a selection mechanism to generate video captions.

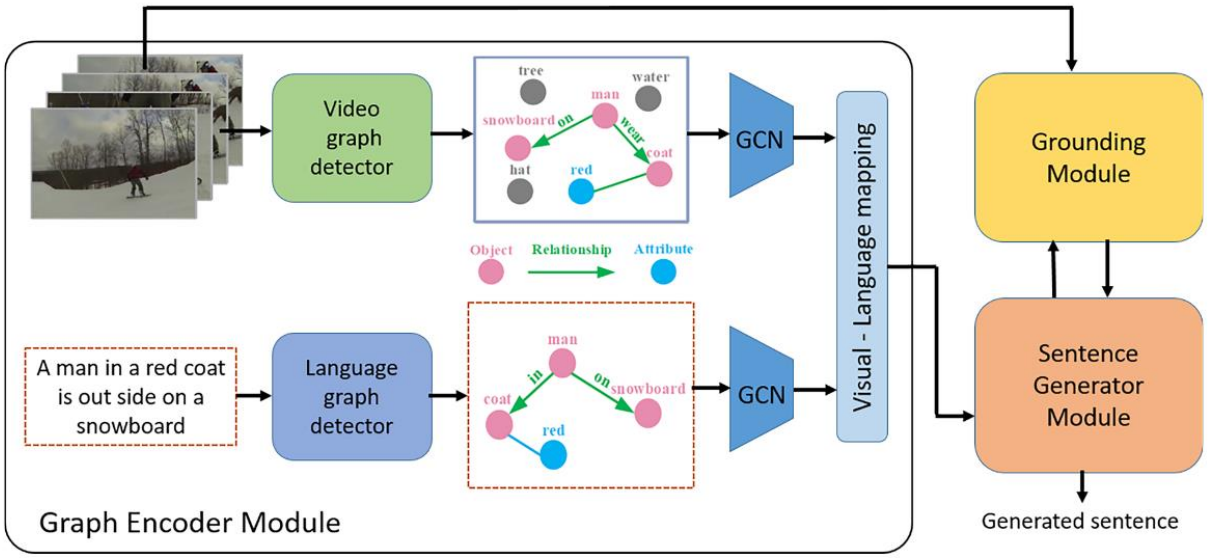

Figure 12 Graph-based video captioning [Zhang, 2020]

Powered by developments in large language models (LLMs), recent large multimodal models (LMMs) have revealed surprising capabilities. Mu et al [Mu, 2024] introduce a multi-modal embodied foundation model, consisted of a frozen vision model for visual features encoding, a frozen language model able to execute natural language for question answering, captioning and embodied planning task, an embodied former including a language mapping layer that aligns visual and embodied instructions and extracts task-relevant features and a policy network, producing low-level actions based on task-relevant features (Figure 13).

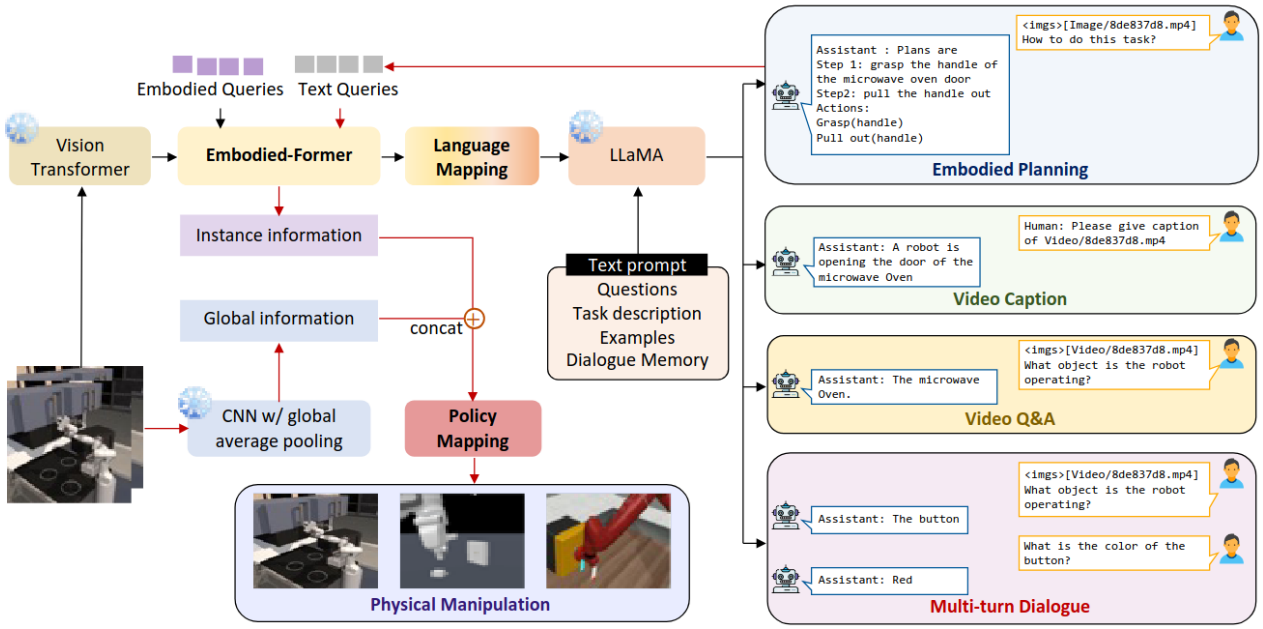

Figure 13 Framework of EmbodiedGPT [Mu, 2024]. The black arrow depicts the vision-language planning process, while the red arrow demonstrates how we use the queried language plans to improve policy learning in low-level control tasks.

The video language model Vid2Seq introduced in [Yang 2023] consists of a pretrained language model, augmented with special time tokens representing timestamps in the video. Contrary to two-stage approaches, Vid2Seq predicts event captions and the corresponding temporal boundaries at once. To deal with the lack of large-scale training data,

transcribed speech sentences are utilized as pseudo event captions and the sentence borders are reformulated as pseudo event boundaries.

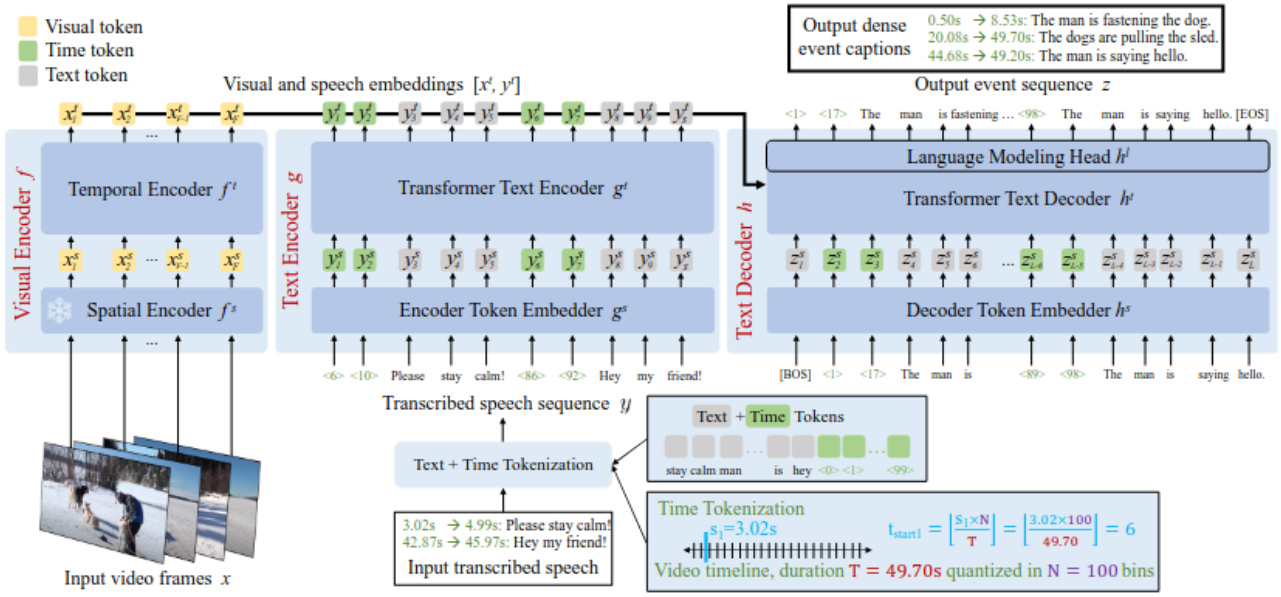

Figure 14 Vid2Seq model overview

Zhang et al [Zhang 2023] introduce an end-to-end model, computing frame representations individually using a pre-trained visual encoder, a frame embedding layer is utilized for temporal information and a video Q-Former to generate visual query tokens. A pre-trained audio encoder and an audio Q-former handle video's audio signals. The overall architecture of the Video-LLaMA model is illustrated in Figure 15**.** Song et al [Song, 2024] propose MovieChat, focusing on long video understanding. The proposed model uses a sliding window to extract and tokenize video features which are subsequently stored in short-term memory frame by frame. When the fixed-length short-term memory reaches its limit, the earliest tokens are moved to long-term memory. MovieChat incorporates two inference modes: Global mode, utilizing only long-term memory and breakpoint mode, letting the user extract a caption at a specific point of the video.

Finally, Chen et al [Chen, 2024] propose Learning-In-Video-strEam (LIVE), a model able to choose whether to produce an output or not during a video stream, allowing the model to process longer streaming videos. To augment training data, a data generation system converts offline text annotations into a streaming dialogue format. The framework is built over the CLIP vision encoder and the Llama language model and manages to achieve real-time conversation within a continuous video stream.

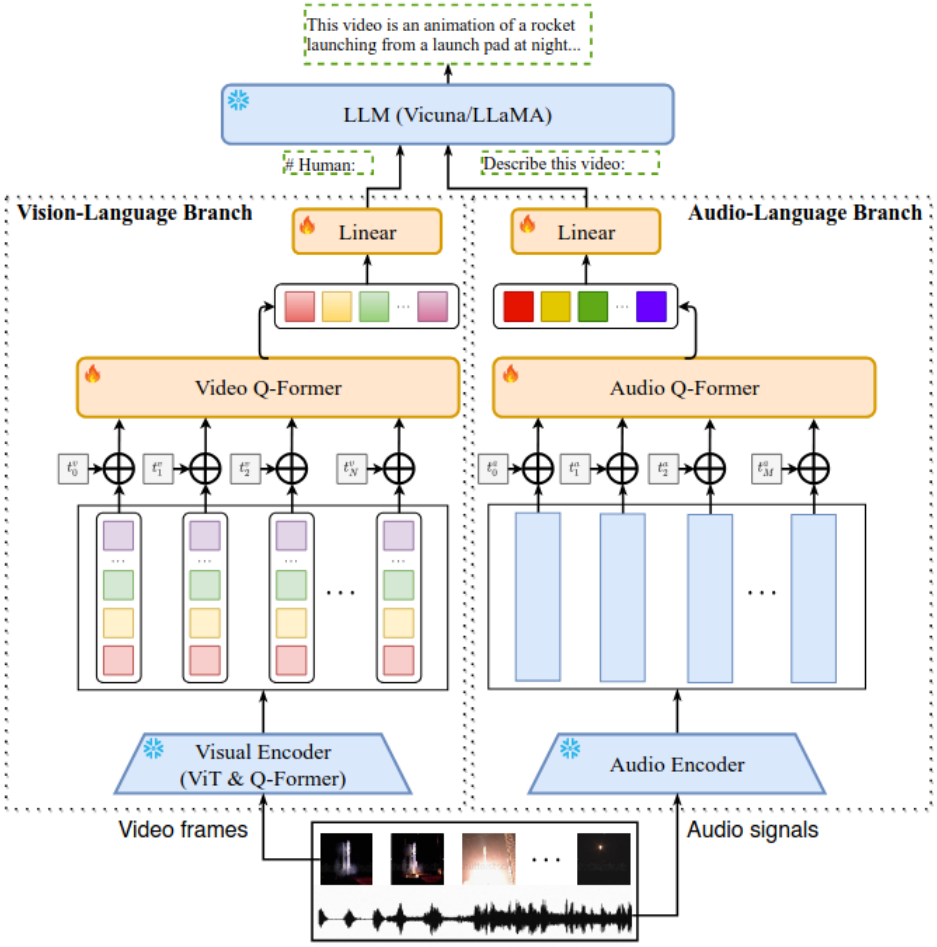

Figure 15 Overall architecture of Video-LLaMA

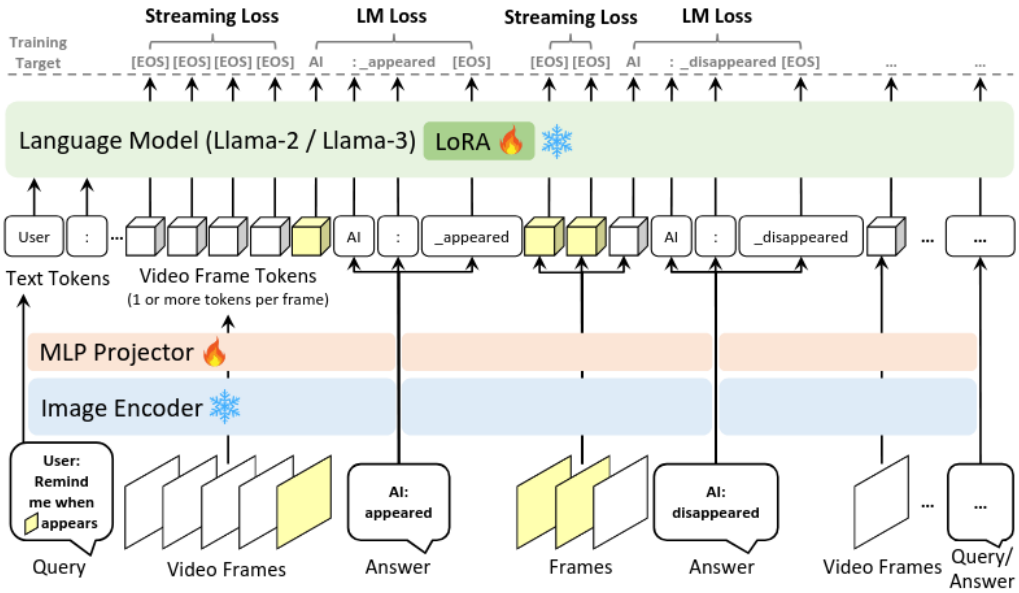

Figure 16 The training pipeline of Learning-In-Video-strEam (LIVE) [Chen, 2024]

3.1.2 Object and Concept/event detection

3.1.2.1 Concept/event detection

Concept/event detection in deep learning involves identifying and understanding specific patterns, objects, or occurrences within data, such as images or videos. This task leverages neural networks to automatically recognize and classify various elements, ranging from simple objects like cars and people to complex events like traffic accidents or sports activities. The process typically includes feature extraction, where the model learns to identify relevant characteristics from the data, and classification, where these features are used to categorize the detected concepts or events.

Dong et al in [Dong, 2020] introduced an additional motion discriminator to GAN. The dual discriminator structure improved the generator's ability to generate realistic frames with consistent motion. Yu et al in [Yu, 2021] also employed GAN to model normality. To explore the correlation, the proposed Adversarial Event Prediction (AEP) network performed adversarial learning on past and future events to explore the correlation.

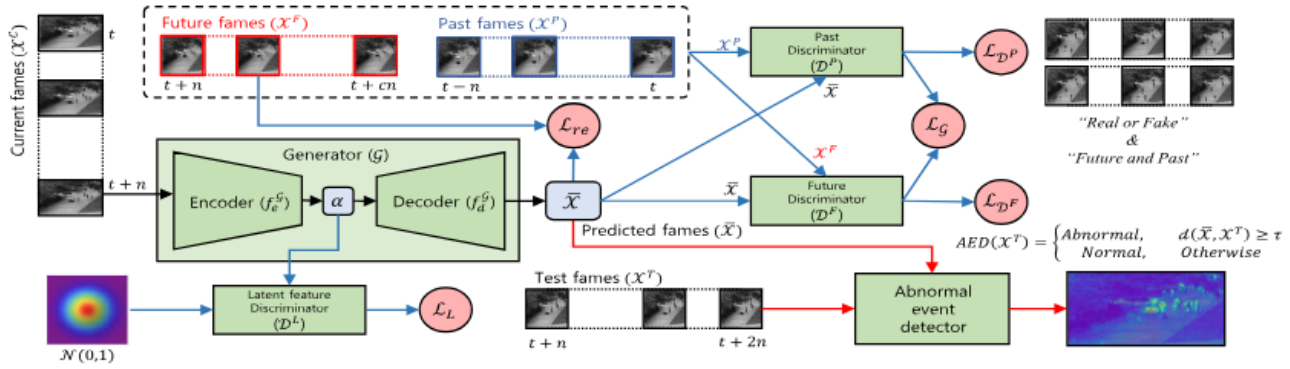

Figure 17 Model architecture as proposed in [Yu, 2021]

Liu et al [Liu, 2022a] proposed an Appearance-Motion United Auto-Encoder (AMAE) system with two distinct encoders for denoising and optical flow generation. Additionally, they employed an additional decoder to fuse spatial-temporal features and predict future frames to model spatial-temporal normality. In [Liu, 2022b] the memory was added to the dual-stream auto-encoder to store prototype appearance and motion patterns. Adversarial learning was utilized to explore the connection between spatial and temporal information of regular events.

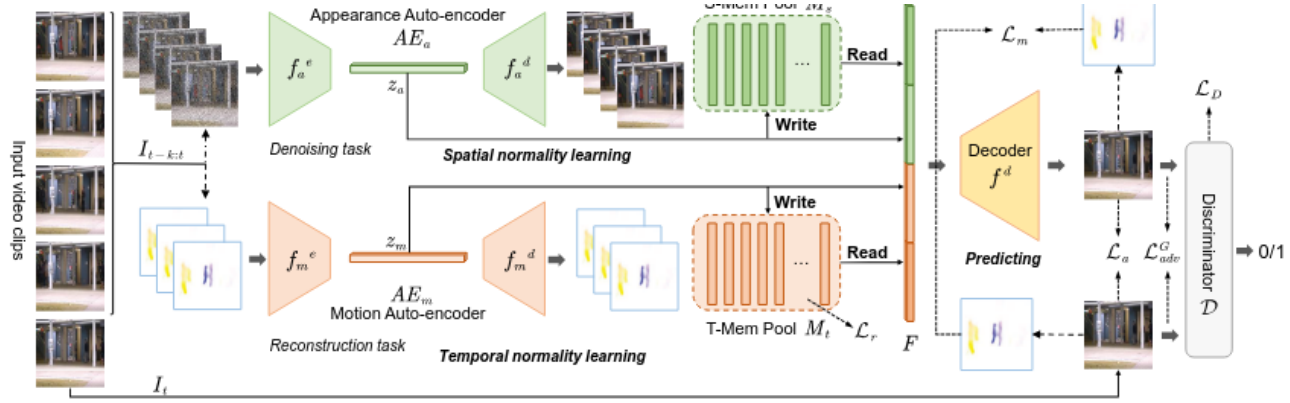

Figure 18 Overview of the framework proposed in [Liu, 2022a]

Chen et al [Chen, 2022] introduced the Bidirectional Prediction (BiP) architecture with three consistency constraints. More specifically, prediction consistency took into account the symmetry of motion and appearance in both forward and backward prediction. Association consistency considered the correlation between frames and optical flow, while

temporal consistency was employed to guarantee that BiP could create temporally consistent frames. Kamoona et al in [Kamoona, 2023] proposed Deep Temporal Encoding-Decoding (DTED) to track the temporal evolution of films across time. They treated instances of the same bag as sequential visual data rather than as independent individuals. Furthermore, DTED employs joint loss optimization to maximize the average distance between normal and abnormal videos. Wei et al [Wei, 2022] introduced a two-stage multimodal information fusion method that first refines video-level hard labels into clip-level soft labels before using an attention module to fuse multimodal information. Their extension work, [Wei, 2022a] Multimodal Supervised Attentional Augmentation Fusion (MSAF) uses attention fusion to align information, resulting in implicit alignment of multimodal data.

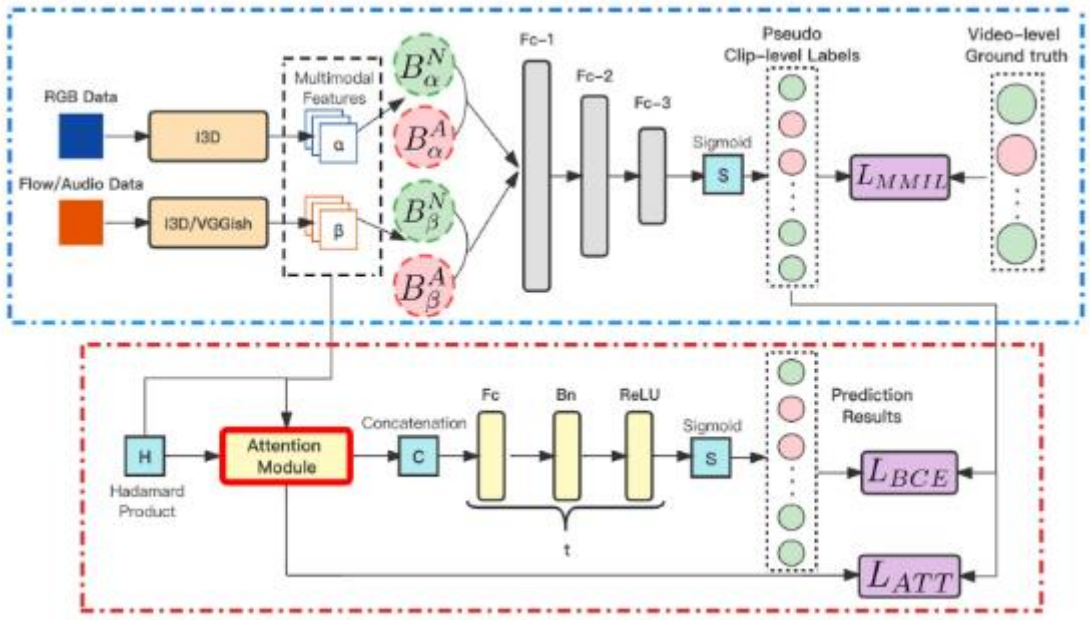

Figure 19 Illustration of the framework proposed in [Wei, 2022a]

Zaheer et al in [Zaheer, 2022] attempted to take advantage of the limited frequency of anomalous events. Their pipeline (Figure 20) consisted of a generator G and a discriminator D, which were supervised by each other cooperatively. The generator primarily generated representations for normal events. For anomaly events, the generator employed negative learning techniques to distort the anomaly representation and generated pseudo-labels to train D. The discriminator calculated the likelihood of anomalies and generated pseudo labels to improve G.

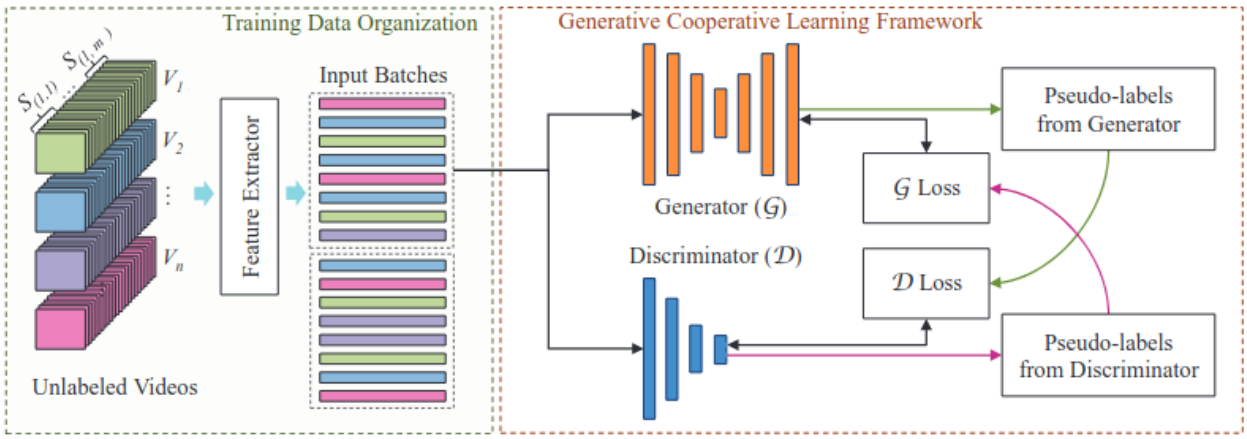

Figure 20 Proposed algorithm of [Zaheer, 2022], introducing cross-supervision for training a Generator G and a Discriminator D.

3.1.2.2 Object detection

Object detection serves, for the purposes of the CEDAR project, the goal of corruption detection in multimedia content. This will achieved through video understanding, thus in this section relevant methods are presented.

As a new type of aerial robotics that are simple to use and inexpensive, drones are more accessible to individuals and businesses. Consequently, datasets captured from drones occur more frequently to the literature. Due to the drones infrastructure, drones can perform object detection in different ways. Drone detection techniques can be divided into four groups: visual, radar, acoustics, and radio frequency-based approaches.

Radar detection-based approaches employ short duration radar pulses for detection. The use of that kind of pulses can achieve high range accuracy and range resolution. Kim et al [Kim 2020] propose a novel image structure for radar classification called polarimetric merged-Doppler image (PMDI). The proposed radar image structure highlights the periodicity of micro-Doppler signature (MDS), represents multipolarization features, and filters out unnecessary information while maintaining the same image data size. GoogleNet model is employed to minimize the amount of polarimetric data and remove extraneous information from MDS.

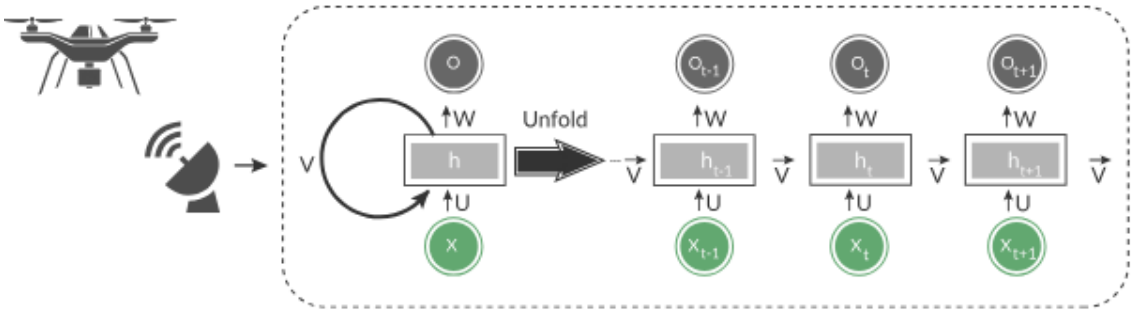

Figure 21 A drone detection architecture based on an RNN model using features capture by radar.

Acoustic detection-based techniques consist of three main modules: detection, feature extraction and classification. Detection captures the target sound in a noisy environment, while feature extraction mines human-focused characteristics to feed the classifier. Classification assigns the likelihoods of mined features to the consistent class. A general framework is illustrated in Figure 22**. T**he audio raw data is preprocessed and feature extracted before training the deep learning system. The trained model's weight is utilized to classify the type of object. In [Utebayeva, 2020] a stacked bidirectional LSTM is used to identify objects using drone audio signals. The model receives input from framewise spectral domain features, including time domain features, windowing features, perceptual features and frequency domain features.

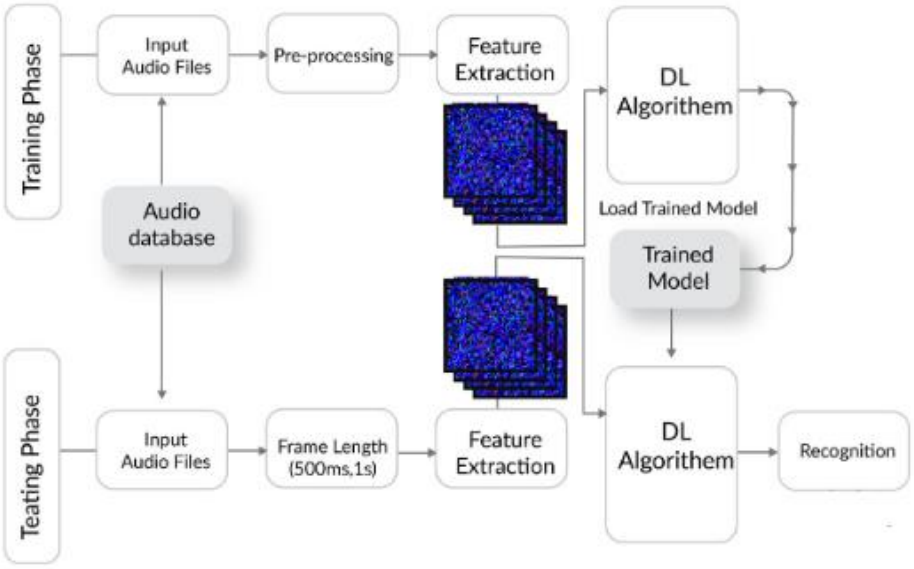

Figure 22 A typical architecture of a drone detection audio-based approach

Radio frequency detection-based approaches can be employed via the communication frequency spectrum. The RF allows the pilot to operate the drone remotely with a real-time video transmission from the camera. There are strong RF transmissions between the drone and controller. These signals can be used to detect objects in the sky. Medaiyese et al. [Medaiyese, 2022] compared and examined the performance of RF-based drone identification methods in the presence of wireless interference, including Bluetooth and WiFi. RF signals can serve as unique signatures in both steady and transient states. The authors extracted drone features by analyzing RF control signals transmitted between the drone and the controller. Wavelet transforms, such as continuous, discreet and WST, were utilized to extract RF signals from the drone in both steady and transient states to compare feature extraction methods. The authors also tested the efficiency of the models at different SNR levels.

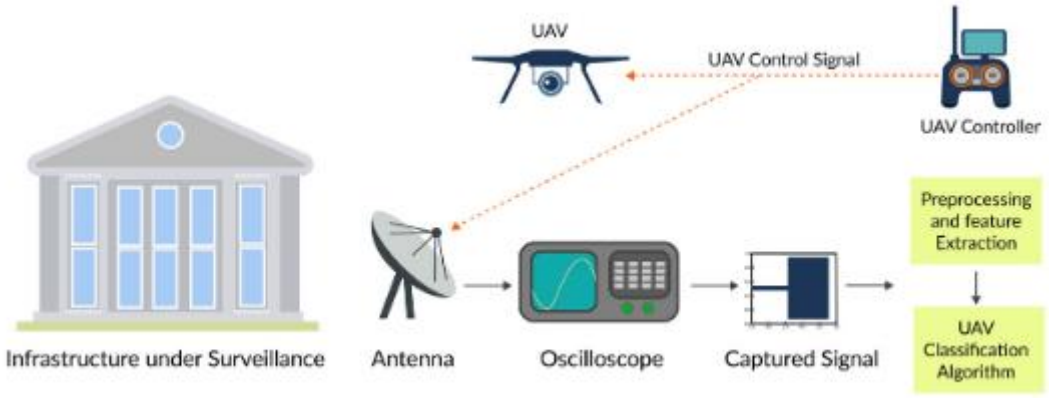

Figure 23 A general architecture of a radio-frequency-based approach

Visual detection-based methods have been shown to be successful in ensuring public safety in the areas of ship collisions, power line surveillance, border and solar farm energy inspection, and motor vehicle accidents. A baseline detection method, with visual stimuli is illustrated in Figure 24**.** In [Nalamati, 2019] the challenge of recognizig little objects by employing a popular deep learning-based approach is addressed. The authors utilized Faster-RCNN with a ResNet-101 base architecture. Pretrained models, trained models and transfer learning were used to adddress low data availability. Mahdavi et al in [Mahdavi, 2020] discussed numerous approaches for detecting flying objects using a fisheye camera. CNN classification uses a two-stage training and testing procedure. During the training phase, attributes from all photos were retrieved and utilized to train a CNN model. Next, the trained CNN categorized the input picture.

Figure 24 A baseline drone detection method using images, based on a CNN architecture

Apart from the approaches discussed above, the "traditional" detection methods can also achieve good results. Some of them have been used to the techniques described above such as the Faster-RCNN in [Nalamati, 2019]. Listed below are some indicative SoTA methods for object detection:

You Only Look Once (YOLO): In this paper, Redmon et al [Redmon, 2016] proposed a method that was computationally simpler and faster compared to the existing ones. Their method reframes the problem as a regression; the picture is divided into equal cells, and a specific number of bounding boxes are predicted with associated confidence ratings, as well as a probability score for each class (just one per cell, not per box). A basic neural network is utilized, with seven convolutional layers extracting information and two fully connected layers predicting classes and bounding boxes. However, because each cell can only predict one item, YOLO's capacity to identify tiny objects and flocks is limited.

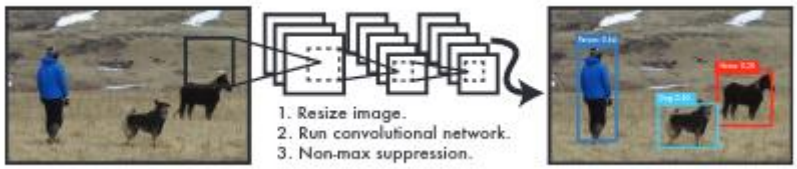

Figure 25 The YOLO detection system

YOLO-LITE: In [Huang, 2016] YOLO is modified to be executed real-time on non-GPU computers. The authors argue that dividing the input image size by 2, removing some layers and the batch normalization, which addresses covariant shift and vanishing gradient problems, does not reduce performance due to the smaller network size. A processing speed, with a CPU, above 10 FPS (Frames Per Second) is achieved with a satisfactory performance.

Faster R-CNN: In this paper, Ren et al [Ren, 2016] Region Proposal Networks (RPNs) are introduced, saving time compared to earlier approaches. The architecture is as follows: Convolutional layers are used to extract a common feature map. The sliding window predicts bounding boxes and 'objectness' scores based on pre-set scale and aspect ratio combinations, using two separate fully connected layers to process the feature map. The classes are predicted using Fast R-CNN [Girshick, 2015].

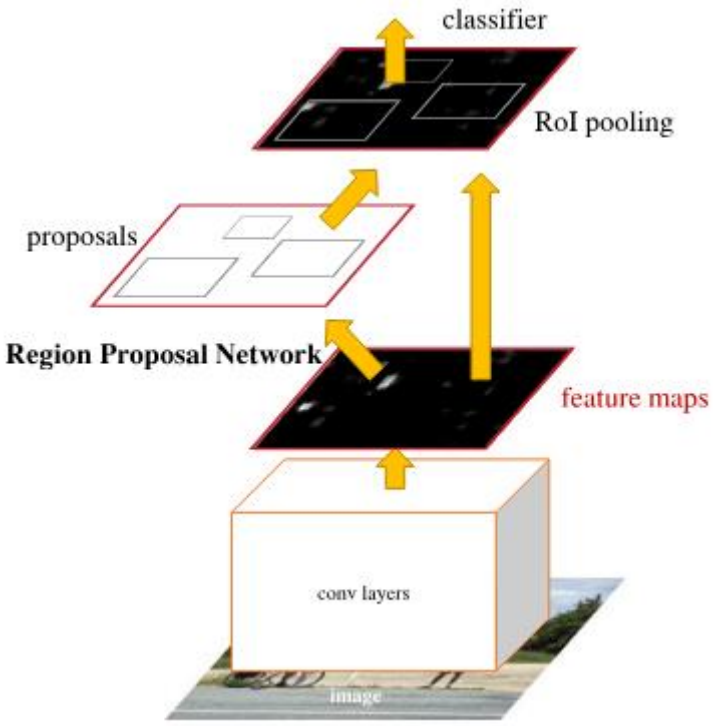

Figure 26 The pipeline of Faster R-CNN. The RPN module serves as the 'attention' of this unified network.

SSD: Single Shot MultiBox Detector: Liu et al [Liu, 2016] propose an object detection technique that improves speed and accuracy compared to previous approaches. This approach, like YOLO, saves time by not resampling pixels or features after estimating the bounding box. However, unlike YOLO, it maintains accuracy when compared to state-of-the-art methods. To predict box offsets and category scores, a similar technique to YOLO is used, but tiny convolutional filters are added to the feature maps (at different stages to predict different scales).

A specialized sub-task of object detection is the small object detection, involving detecting and localizing small objects in images and videos. The challenge of the task is the low resolution and small size of the objects. Other factors that make this task difficult are the occlusion, the background clutter and lighting conditions variations. Two different definitions of the small objects can be found in literature. Small objects can be defined based on the relative area of all instances of the same category, according to Chen et al [Chen, 2020]. Small objects can also be defined by an absolute definition. In MS-COCO dataset objects with a resolution smaller than 32*32 are considered to be small objects.

Supervised data augmentation approaches use labels from training samples to guide augmentation operations. These techniques involve adding modifications and augmentations to images or data samples while maintaining label consistency, increasing the dataset's size and diversity. In [Zeng, 2022] AutoAugmentImage and Mixup data augmentation methods are employed. AutoAugmentImage identifies appropriate data augmentation methods based on various datasets to improve the model's performance. Mixup generates new training samples by combining two samples from the training data, which enriches the dataset and improves model robustness. These data augmentation approaches introduce diverse data through various transformation techniques, effectively addressing the scarcity of small object data.

Dilated convolution methods can capture both global and local information, leveraging the connection between an object and its surroundings, improving localization and comprehension of smaller objects. Cui et al [Cui, 2020] uses parallel multi-branch dilated convolutional layers with dilatation convolutions that increase the dilatation rate for each branch. It increases the neuron's receptive field without diminishing feature map resolution, resulting in additional contextual information, which is beneficial for small object detection.

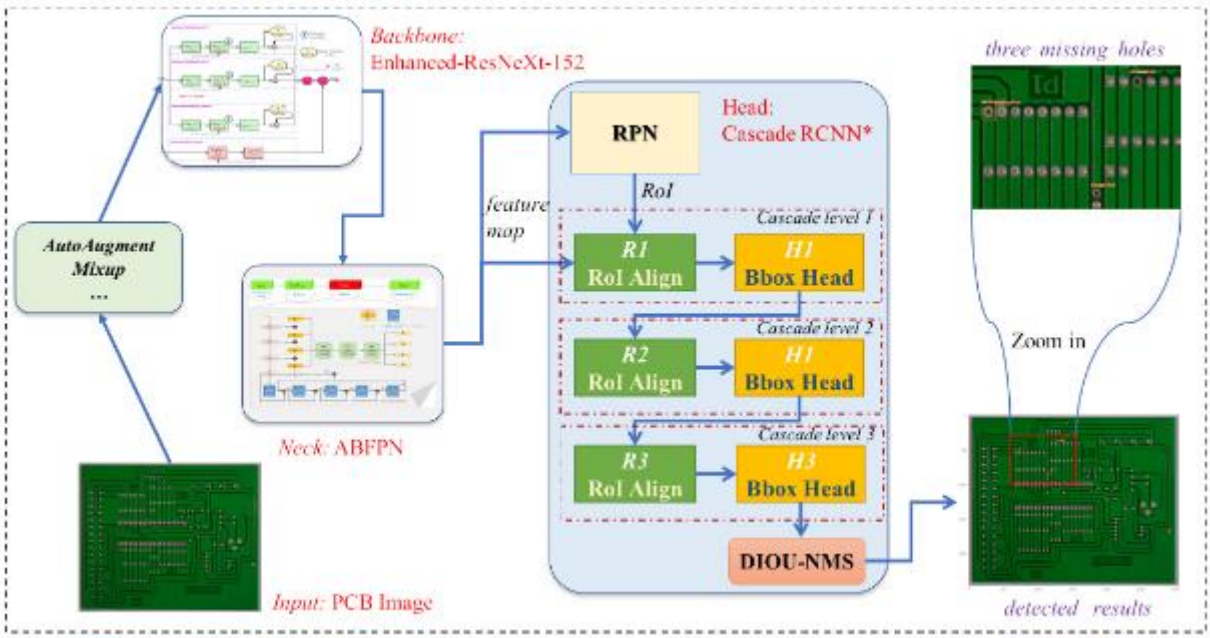

Figure 27 Framework proposed in [Zeng, 2022]

Multi-scale feature extraction methods are suitable for data with varying scales and proportions. Cao et al in [Cao, 2022] propose a YOLOv5-based small object algorithm for railway scenes. It adds a lower-down sampling detection scale, known as four-scale detection, on the baseline network. This method preserves deep semantic information while acquiring shallow feature information. It also adds an SPP (Spatial Pyramid Pooling) module into the main network.

Attention mechanism methods employ the attention mechanism, mirroring the human visual cognitive system, allowing models to focus on the most significant bits of the input data. The introduction of attention mechanisms in small object detection allows neural networks to focus on critical aspects around small objects, improving the model's perception and detection accuracy. The performance of transformers in natural language processing [Dosovitskiy et al., 2020], led to the use of their self-attention mechanism in the field of computer vision. In [Dosovitskiy et al., 2020], the self-attention system captures connections between different regions within an image. As illustrated in **Figure** the input image is divided into patches, which are then turned into vectors. These vectors are then passed into the self-attention mechanism, which calculates the correlations between each visual patch and the other patches.

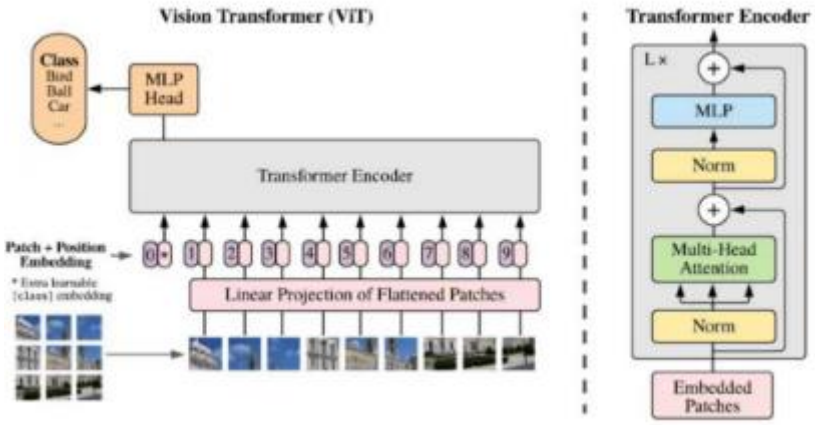

Figure 28 Vision Transformer

3.1.3 Anomaly detection

3.1.3.1 Anomaly detection, forecasting

In recent years, surveillance systems have become increasingly popular due to developments in high-definition security cameras and memory technology. Nevertheless, the widespread use of surveillance cameras in both public and private settings, necessitates the development of sophisticated algorithms to recognize and identify human activity and other relevant subjects in a recorded footage. The computer vision community has focused on abnormal human action recognition (AbHAR) for application-oriented research. This is crucial due to its numerous applications, including intelligent surveillance and security systems, human-computer interaction, robotics, sports analytics, intrusion detection, elderly healthcare, and patient monitoring. The goal of an anomaly detection systems is the detection of visual anomalies (Figure 29**).** Anomalies in video refer to irregular patterns that do not align with typical training examples, including anomalous behaviors and entities.

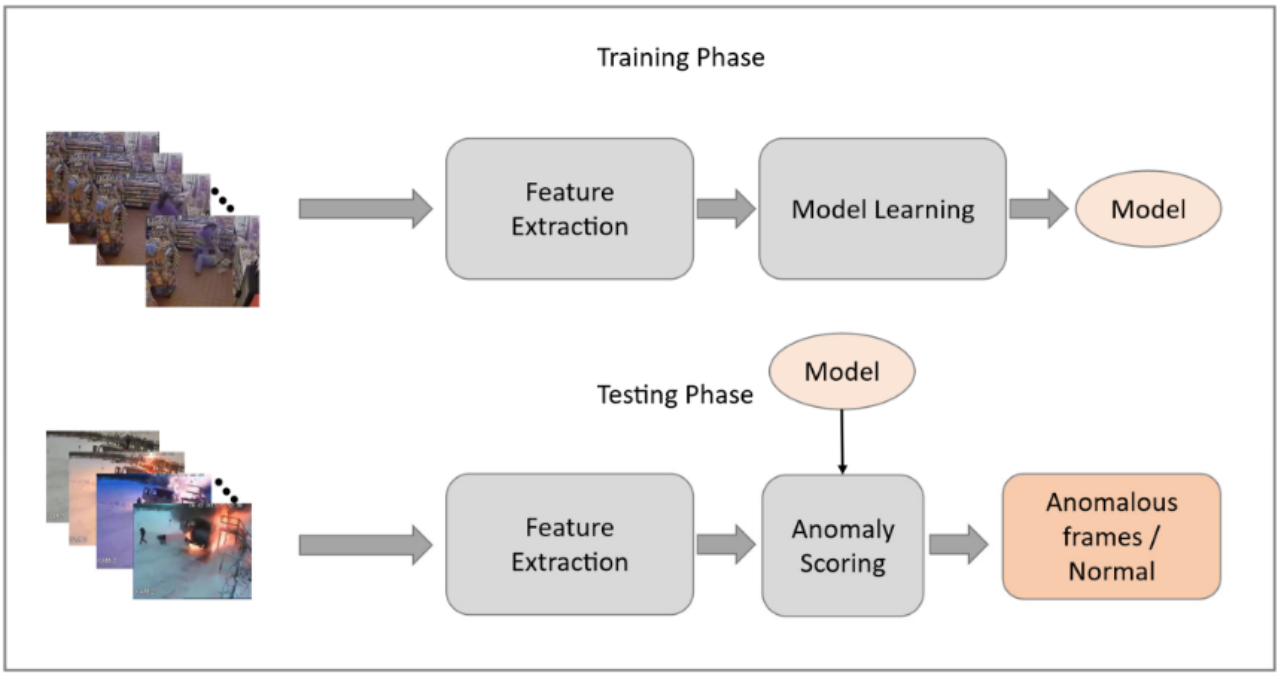

Figure 29 General pipeline of anomaly detection

Wang et al [Wang, 2023] propose an autoencoder-based Memory-Augmented-Motion Network to learn the appearance and motion features of an input frame, and a pointed patch-based stride convolutional detector (PSCD) technique, to reduce degradation effects. The PSCD method improves detection accuracy by utilizing abnormal event features at the patch-level rather than frame-level of the error map. In [Wang, 2023a] the authors used an attention mechanism-guided multi-instance learning weakly supervised video anomaly C3D network model to optimize training data for different datasets. This was motivated by the fact that the DL anomaly detection approach that utilizes visual sensors as the initial signal input suffers from the over-boundary problem caused by a lack of labeled datasets, which resulted in unsupervised training.

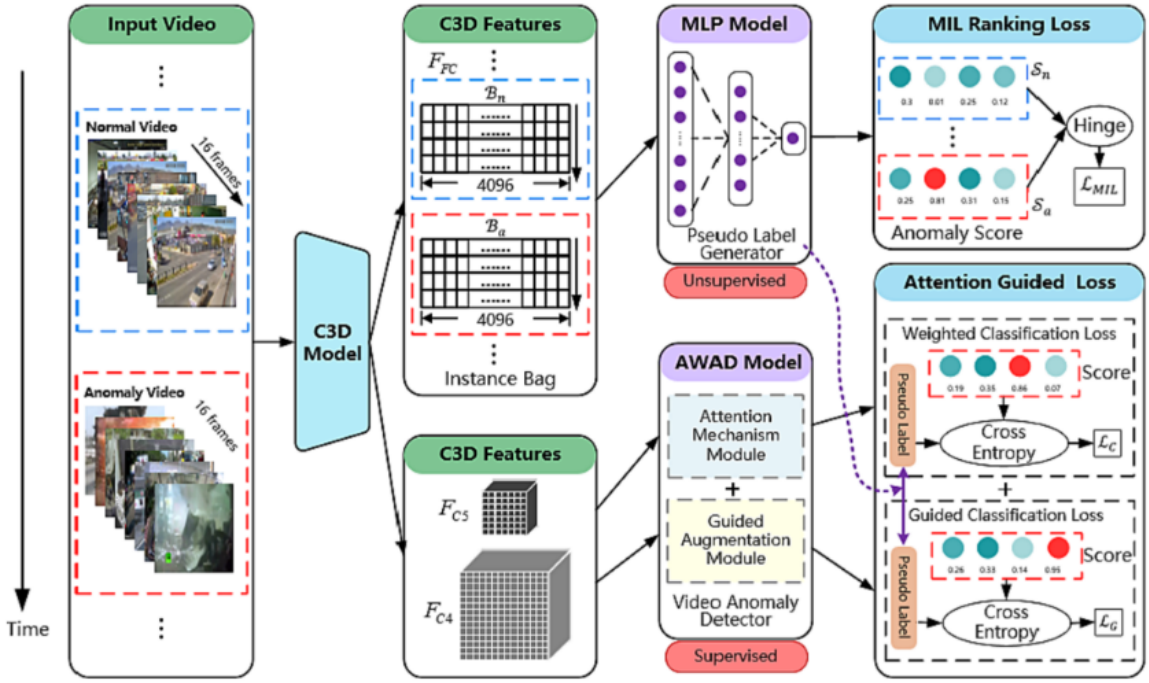

Figure 30 Flow chart of the anomaly detection model proposed in [Wang, 2023a]

Wu et al in [Wu, 2019] introduced a lightweight deep one-class model to solve one-class problems in complex scenarios. The created model learns feature representation and trains the classifier alongside CNN. Multi-frame and optical flow inputs were employed to gather spatiotemporal information. A disadvantage of this method is that the reconstructed images produced are blurry.

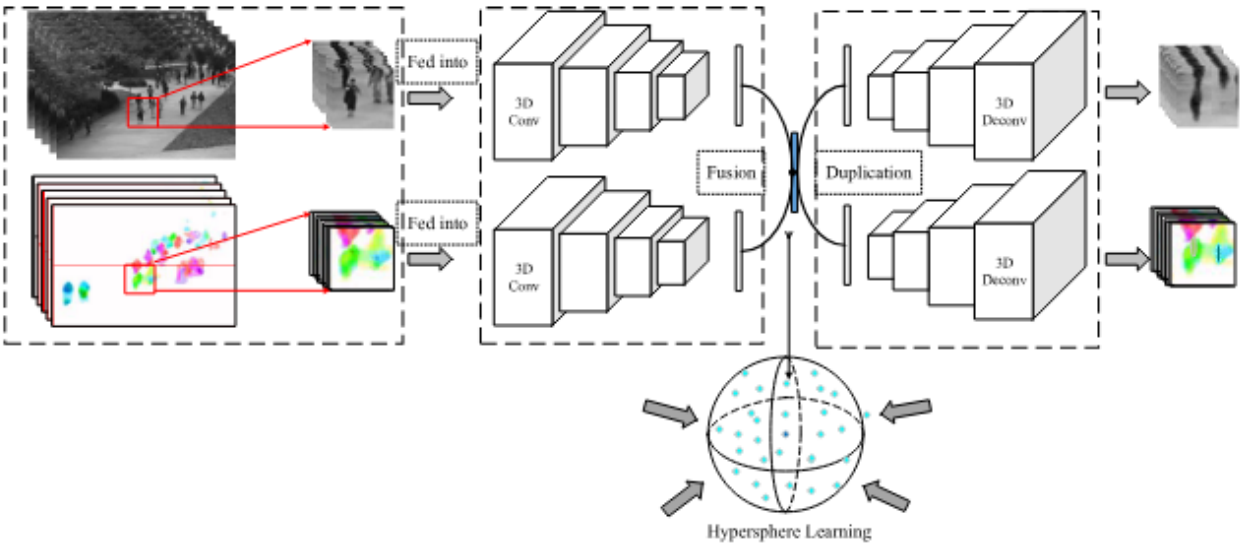

Figure 31 Model proposed in [Wu, 2019]

Islam et al [Islam, 2023] created an effective framework for spotting abnormalities in surveillance video data. 2D-CNN, autoencoder and Echo State Network (ESN) were integrated for feature extraction of input videos, sequence learning and anomaly detection. The model proposed in this work is lightweight and it may be applied on edge devices to provide safe and secure intelligent surveillance to the users.

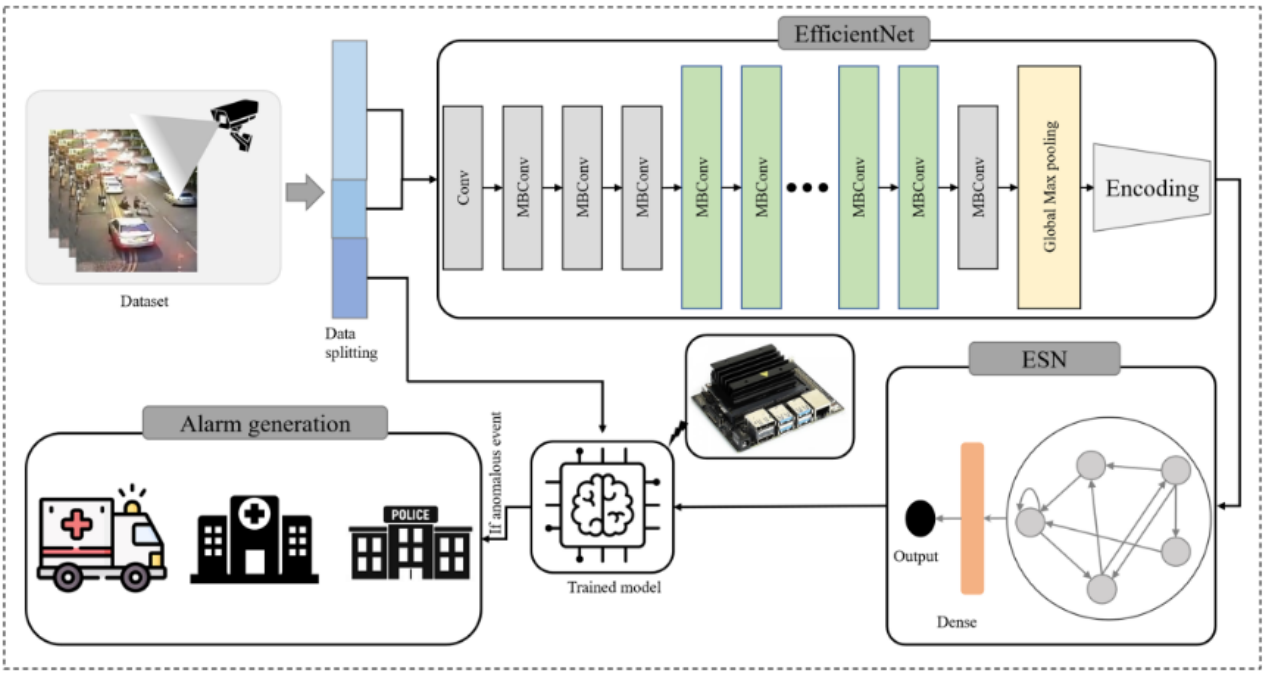

Figure 32 The model proposed for anomaly detection in [Islam, 2023]

In [Ulah, 2023] a Vision Transformer Anomaly Recognition (ViT-ARN) system for detecting and interpreting anomalies in surveillance video in smart cities is proposed. The framework consists of two stages: online anomaly detection using a customized, lightweight, one-class deep neural network in a surveillance enviroment, and classification of the detected anomalies into appropriate classes. The anomaly detection model is optimized for resource-constrained devices by employing a geometric median-based filter pruning method to reduce its size. The refined features are sent to a multi-reservoir echo state network to analyze real-world anomalies like vandalism and traffic accidents.

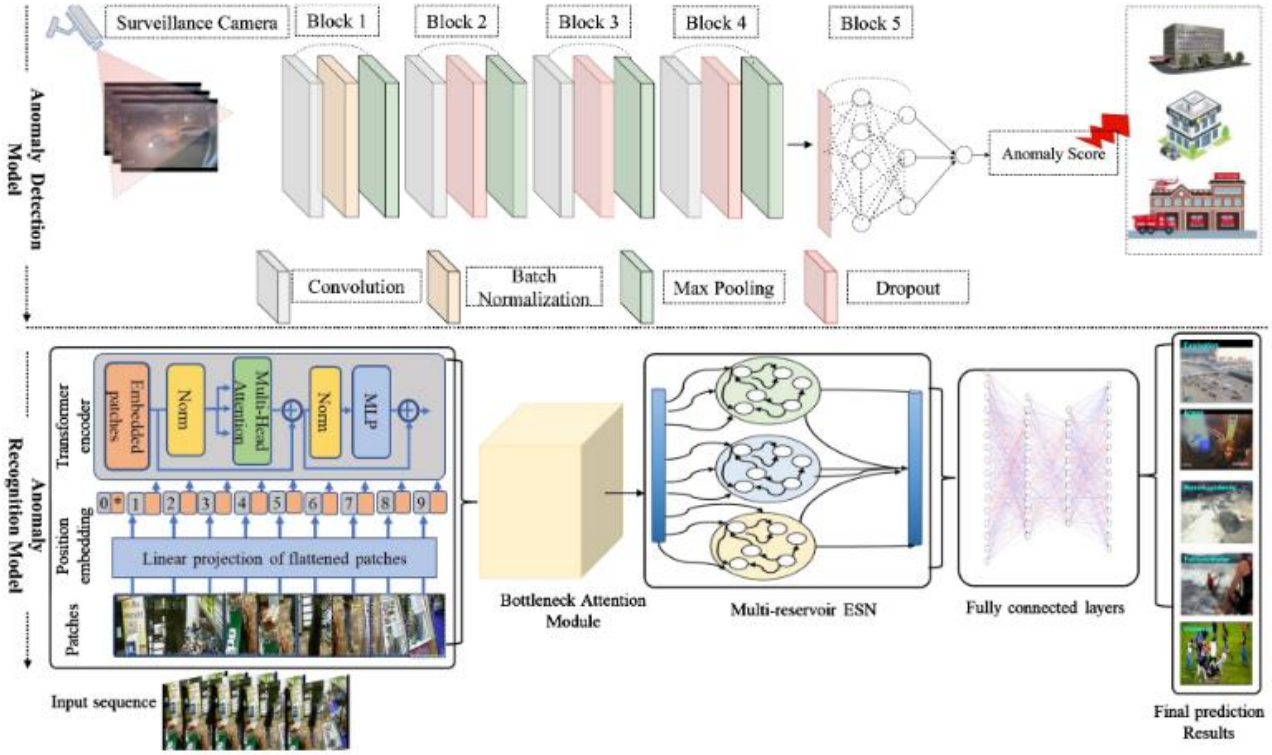

Figure 33 Structure of the ViT-ARN framework for anomaly detection and recognition [Ulah, 2023]

Apart from detecting irregular patterns in a video, another task that could be proved very useful within CEDAR needs is video prediction (forecasting), i.e., the task of predicting future frames given past video frames. Current video prediction algorithms may be broadly classified into two categories. One category involves flowing pixels from a reference frame (usually the last observed frame) to create future frames. However, this group of methods has intrinsic challenges in modelling objects' appearance and disappearance (birth and death) within the scene. These approaches can provide quite accurate forecasts in the short term, but their accuracy decreases with time. The other category includes ways for creating new frames from scratch. Although these techniques seem promising for capturing the birth-and-death processes in object dynamics, they primarily describe pixel-level distributions. As a result, they often lack an integrated understanding of the underlying real-world context, which is critical for creative prediction capabilities.

Hu et al in [Hu, 2023] propose Dynamic Multi-scale Voxel Flow Network (DMVFN) that expands the concept of dense voxel flow by incorporating a differentiable routing module. This addition improves the model's ability to capture and represent varying scales of motion. In [Wu, 2021] Greedy Hierarchical Variational Autoencoders (GHVAEs) are introduced. In contrast to standard hierarchical variational autoencoders (VAEs), a GHVAE model trains each encoder-decoder module greedily utilizing previously learnt module weights (Figure 34)**.** Greedy training avoids fitting the full model into memory, allowing bigger models to be learned on the same GPU or TPU memory. Furthermore, greedy training enhances the optimization stability of a hierarchical model by breaking the bidirectional dependencies between individual latent variables.

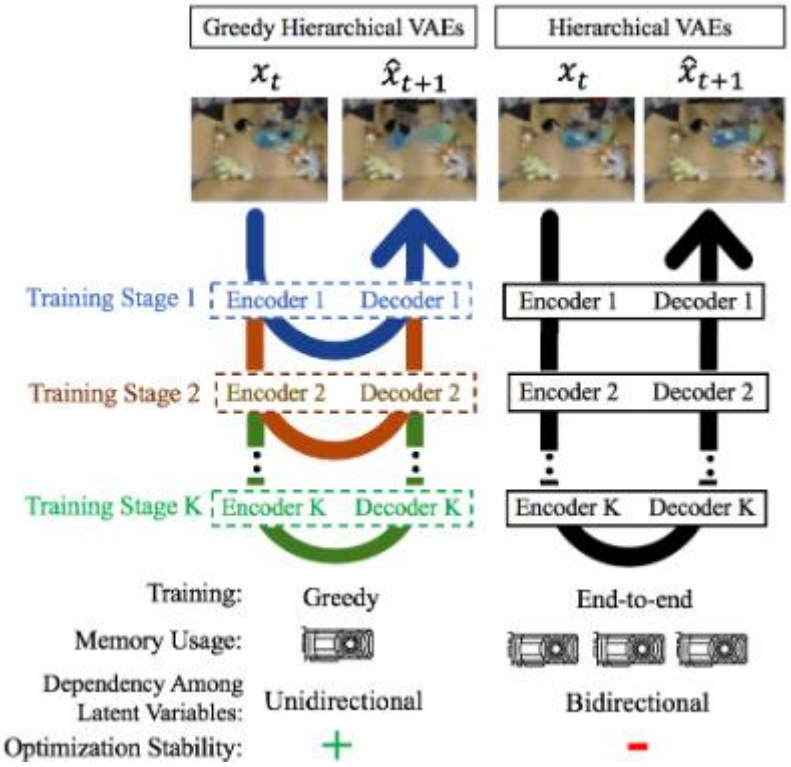

Figure 34 Overview of GHVAEs

Video LDM [Blattmann, 2023] generates videos using pre-trained image models, resulting in multi-modal, high-resolution, and long-term video predictions. LDM is initially trained on images only. Then, the image generator is turned into a video generator by adding a temporal dimension to the latent space diffusion model and fine-tuning on encoded image sequences. Seer [Gu, 2023] includes an Inflated 3D U-Net for diffusion and a Frame Sequential Text Transformer for text conditioning (Figure 35). By inflating Stable Diffusion along the temporal axis, the model can predict different outcomes using natural language instructions and reference frames.

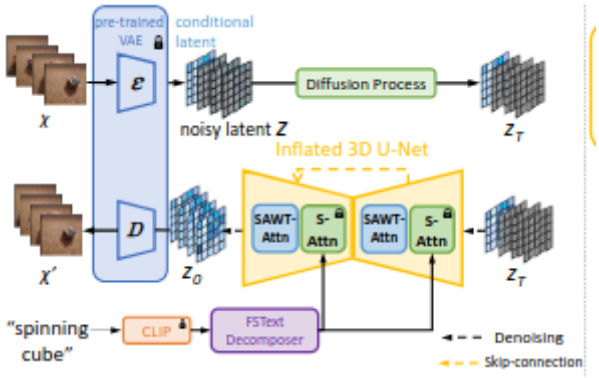

Figure 35 Seer's pipeline (Gu, 2023)

3.1.3.2 Image and video manipulation, deep fake detection

Image and video manipulation and deep fakes detection are crucial for maintaining trust and integrity in digital media, therefore, in CEDAR project, will be employed for disinformation detection. In the following paragraphs, pertinent methods from the literature will be presented.

In today's fast-changing digital world, media content manipulation has become highly advanced. One of the most notable developments is deepfake technology, which uses Artificial Intelligence (AI) and machine learning to manipulate or create video and audio content. Deepfakes offer exciting opportunities, like improving film production and virtual meetings. However, they also come with serious risks, such as spreading misinformation, identity theft, and digital fraud. Figure 36 shows typical signs of forged faces. Radar operation with adequately narrow pulse widths, can achieve restricted target classification.

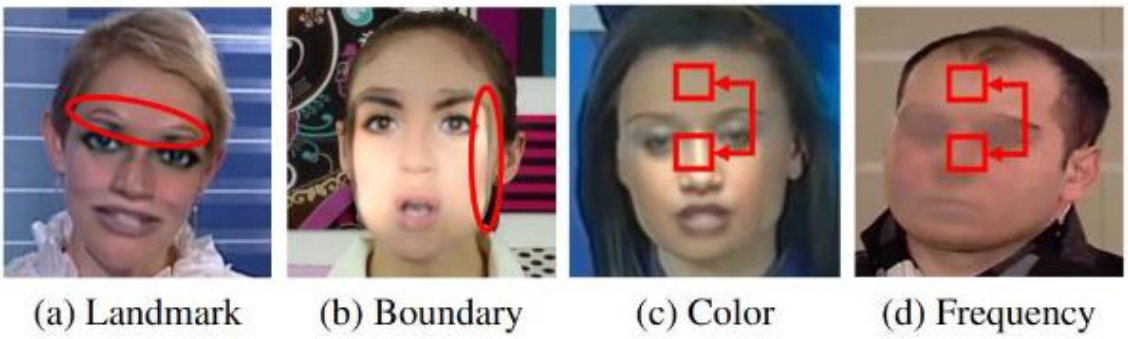

Figure 36 Typical artifacts on forged faces. a) landmark mismatch, (b) blending boundary, (c)color mismatch, and (d) frequency inconsistency [Shiohara, 2022]

The landscape of deepfake detection has evolved considerably, with researchers employing various techniques to tackle the challenges it presents. In the work by Zheng et al. [Zheng, 2021], they delve into the issue of temporal coherence in video face forgery. They introduce a two-stage framework that comprises a Fully Temporal Convolution Network (FTCN) and a Temporal Transformer network. This approach focuses on extracting meaningful temporal features, showing promise in its ability to improve deepfake video detection. In the 2022 work by Xu et al. [Xu, 2022] they propose a DeepFakes detection model that focuses on generalizability and explainability, even in open-set scenarios with unknown attacks. They utilize supervised contrastive (SupCon) loss to differentiate between authentic and DeepFake media, and further enhance the model's performance by fusing it with scores from the Xception network. While Xu et al. [Xu, 2022] focused on supervised contrastive learning for DeepFakes detection, Shiohara and Yamasaki [Shiohara, 2022] propose a novel approach using self-blended images to improve model generalization across different datasets. They use synthetic training data generated by blending pseudo source and target images. This method aims to create a more robust and generalized model by focusing on common forgery artifacts rather than overfitting to manipulation-specific cues. The approach was shown to improve model generalization across various datasets, particularly addressing the domain gap challenges that existing methods face.

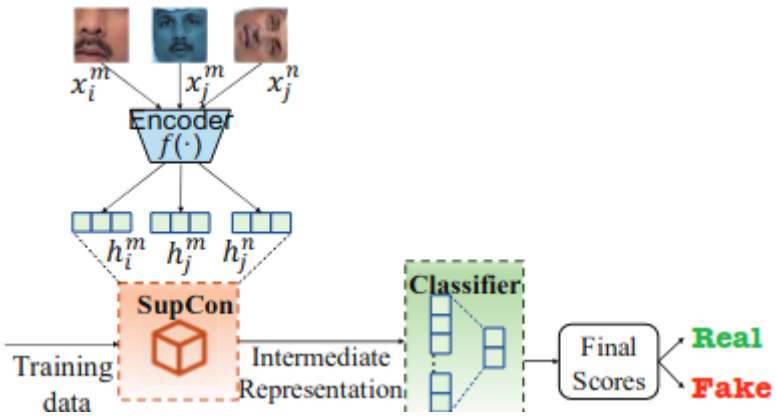

Figure 37 The DeepFake detector of [Xu, 2022]

Following the discussion on model generalizability in deepfake detection, another significant contribution comes from the work of Ganguly et al. [Ganguly, 2022]. In this work, they propose a hybrid architecture that combines Vision Transformer with Xception Network (ViXNet) to improve deepfake detection in both videos and images. The model focuses on capturing both local and global facial artifacts by employing patch-wise self-attention modules and deep convolutional neural networks. ViXNet aims to address the limitations of existing methods by offering better generalization capabilities, especially when models are trained and tested on different datasets. Advancing the field further, Kolagati et al. [Kolagati, 2022] develop a hybrid model that integrates a deep multilayer perceptron with a convolutional neural network for detecting deepfake videos. Their model utilizes facial landmarks detection to extract various facial attributes, which are then used to train both the multilayer perceptron and the convolutional neural network. The multi-input architecture aims to offer improved accuracy and generalizability in identifying manipulated videos.

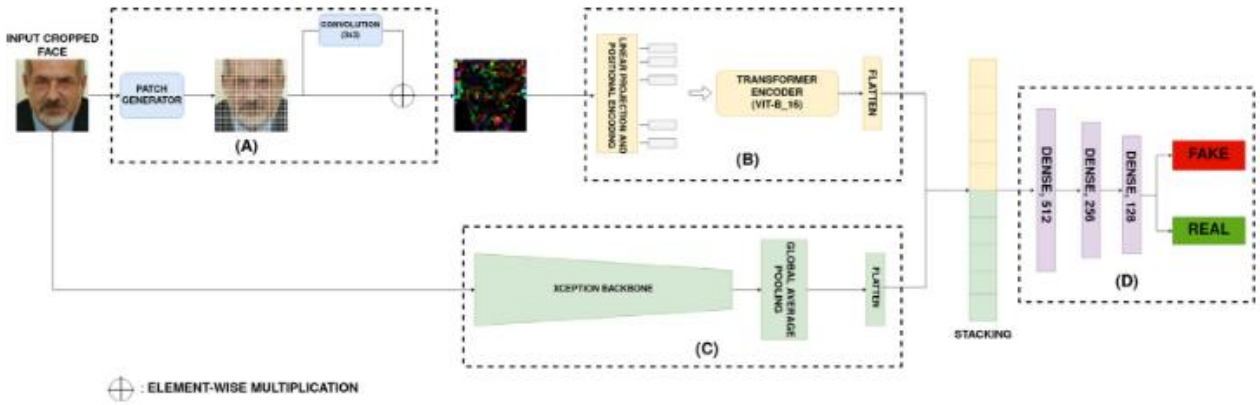

Figure 38 ViXNet model architecture

Complementing existing work on deepfake detection, Cozzolino et al. [Cozzolino, 2022] focus on creating a Person-of-Interest (POI) deepfake detector that captures unique audio-visual features of individuals. Utilizing a contrastive learning paradigm, they generate discriminative embedding for moving faces and audio segments. This enables their model to identify inconsistencies in the manipulated media, offering a more generalized and robust detection method that can handle both single and multi-modality attacks, even in low- quality or corrupted videos. Furthermore, Dong et al. [Dong, 2022] introduce the Identity Consistency Transformer, a face forgery detection method that emphasizes identity-based high-level semantics. The model employs a consistency loss mechanism to detect identity inconsistencies between inner and outer face regions. Notably, the approach shows strong generalization capabilities across different datasets and degradation types, making it particularly apt for identifying face forgeries involving celebrities. In a parallel line of inquiry, the 2023 work by Dong et al. [Dong, 2023] examines the limitations of binary classifiers in deepfake detection, specifically focusing on the phenomenon of Implicit Identity Leakage. This refers to unintended identity information being learned by the classifiers, which hampers their generalization capabilities. To mitigate this issue, the authors introduce the ID-unaware Deepfake Detection Model, which is designed to minimize the influence of identity information on detection performance. Their approach demonstrates improved results in both in-dataset and cross-dataset evaluations. In a similar effort to improve deepfake detection across various modalities, Ilyas et al. [Ilyas, 2023] introduce AVFakeNet, a unified framework employing a Dense Swin Transformer Net. This approach is notable for its ability to detect manipulations in both audio and visual streams, enhancing robustness against varied face poses and

illumination conditions, thus demonstrating strong generalizability and effectiveness across multiple datasets.

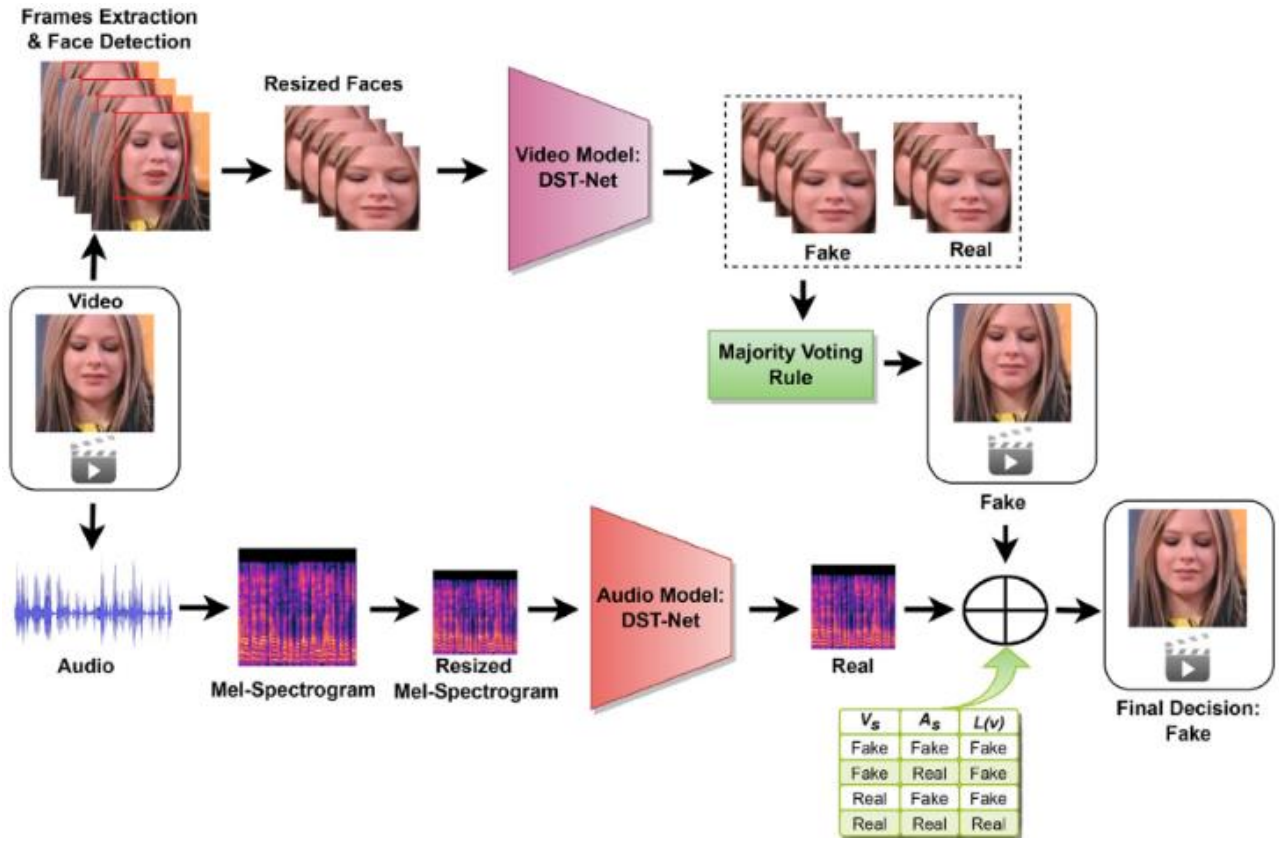

Figure 39 The workflow of AVFakeNet

3.1.4 Image and video retrieval

Image retrieval is a fundamental and long-standing computer vision task that includes searching a large database for photos that match a given query. It is commonly registered as a type of fine-grained, instance-level classfication. Regarding the video retrieval its objective is to identify the video that matches a text query from a pool of candidate videos. Typically, the videos are returned as a ranked list of candidates and scored via document retrieval metrics.

Numerous research has been conducted on similarity metric-based retrieval methods, with a particular focus on deep learning approaches. Wang et al. (2014) [Wang, 2014] were one of the early researchers to adopt a deep learning-based framework for learning an optimal similarity metric in content-based image retrieval (CBIR). This work laid the foundation for many subsequent research efforts in the area. In their 2021 survey, Kapoor et al. [Kapoor, 2021] discussed the challenging task of CBIR in the context of rapidly growing multimedia content. They highlighted the use of deep learning as a promising solution for bridging the semantic gap in CBIR and share insights from their empirical studies on the use of deep learning for CBIR tasks, pointing towards potential avenues for future research.

Image retrieval methods can be divided into two categories: **Deep Metric Learning** techniques, used to measure the similarity between data samples by learning a representation function that maps these samples into a representative embedding space. This approach is particularly useful in scenarios where traditional classification methods struggle due to high intra-class variance (differences within the same class) and low inter-class variance (similarities between different classes). **Deep Hashing** techniques involves learning hash functions through deep neural networks to convert high-dimensional data into compact binary codes. This method is highly efficient for large-scale data retrieval tasks due to its computational and storage advantages.

Deep Metric Learning: Xuan et al., (2020) [Xuan, 2020] introduced an alternative approach called "Easy Positive" mining for deep metric learning, which aims to create an embedding space where semantically similar images are close to each other and dissimilar images are far apart. Instead of pushing images from the same class as close together as possible, Easy Positive mining only requires the embedding function to map each training image to the most similar examples from the same class. This strategy results in embeddings that are more flexible and generalize better to unseen data. Sun et al., (2020) [Sun, 2020] presented Circle Loss, an adaptive optimization strategy for deep feature learning which prioritizes less-optimized similarity scores, enhancing the flexibility of the optimization process compared to conventional methods, which typically treat all similarity scores equally. This approach unifies class-level label learning and pair-wise label learning, demonstrating superior performance on various tasks, including face recognition, person re-identification, and fine-grained image retrieval. Milbich et al., (2020) [Milbich, 2020] introduced multiple complementary learning tasks for deep metric learning to improve generalization to unknown test distributions. By capturing different characteristics of the training data, such as class-discriminative, class-shared, intra-class, and sample-specific features, the proposed method jointly optimizes all tasks to achieve a diverse training signal and state-of-the-art performance on multiple DML benchmark datasets. Zheng et al., (2021) [Zhen, 2021] presented a deep compositional metric learning (DCML) framework to improve image similarity measurement without sacrificing discriminativeness. They employ an ensemble of sub-embeddings, using a set of compositors to diversely and adaptively combine them, preserving generalizable characteristics and resulting in superior performance on multiple datasets.

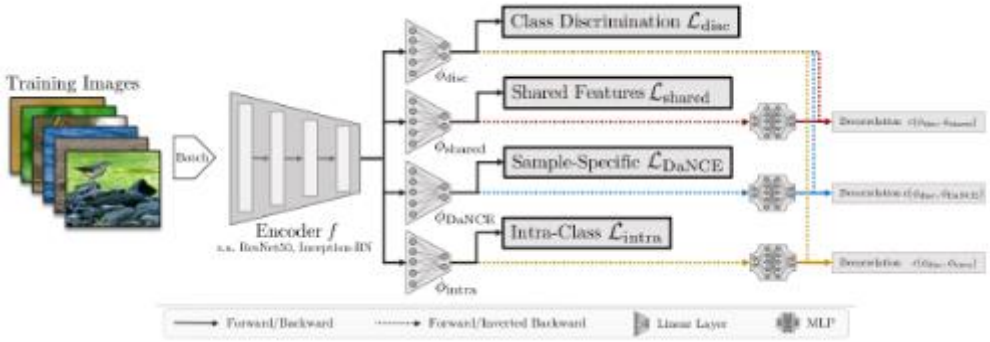

Figure 40 Architecture of the model proposed in [Milbich, 2020]

Roth et al., (2022) [Roth, 2022] proposed a non-isotropy regularization (NIR) for proxy-based DML to address the issue of locally isotropic sample distributions that can lead to loss of crucial semantic context. By employing Normalizing Flows, NIR enforces unique translatability of samples from their class proxies, allowing better learning of local structures and achieving competitive and state-of-the-art performance on standard benchmarks while retaining or improving convergence properties.

Deep hashing: Deep hashing methods have gained popularity for their ability to learn compact binary codes that can efficiently facilitate large-scale image retrieval. Notable works in this area include the simultaneous feature learning and hash coding method with deep neural networks introduced by Lai et al., in (2015) [Lai, 2015]. This method achieved significant improvements in retrieval performance over traditional hashing methods. Cao et al. (2018) [Cao, 2018] developed a novel approach called HashNet, which learned to generate binary codes using an adaptive, discrete code learning strategy. This method surpassed other deep hashing methods in retrieval performance. Zheng et al. (2020) [Yuan, 2020] presented a deep balanced discrete hashing method for large-scale multimedia retrieval, offering increased efficiency in storage and retrieval. Their method, utilizing the straight-through estimator for discrete gradient propagation, bypasses the continuous relaxation strategy commonly used in traditional supervised hash methods, thereby effectively reducing the quantization error. By directly outputting binary codes through the final layer of the Convolutional Neural Network, their approach improves retrieval performance by maintaining both label consistency and pairwise similarity. Xu et al. (2022) [Xu, 2022a] addressed the problem of semantic information loss during the hashing process in large-scale image retrieval by proposing a novel hashing-guided hinge function (HHF). They investigated the relationship between metric learning and quantization learning, revealing that the ideal metric solution cannot satisfy the optimal quantization solution for retrieval. Extensive experiments on four standard datasets

demonstrated the superiority and flexibility of HHF, which can be integrated with various off-the-shelf methods to achieve significant performance gains and state-of-the-art retrieval results. In the work of Doan et al. (2022) [Doan, 2022], an alternative approach to image hashing is proposed, which involves matching the learned continuous code distribution to a pre-defined discrete, uniform distribution in order to minimize distributional distance. By re-formulating the task, the proposed single-loss quantization objective can be integrated into existing supervised hashing methods, substantially improving their performance in achieving coding balance and low-quantization error. In their 2023 work, Hassan et al. [Hassan, 2023] proposed a novel asymmetric learning-based generative adversarial network (AGAN) to tackle the inadequacies of existing hashing algorithms in the context of content-based image retrieval. By merging feature learning with hashing into an end-to-end learning framework and introducing three distinct loss functions (encoder loss, generator loss, and discriminator loss), they improved their retrieval performance, even outperforming several state-of-the-art methods as confirmed by extensive experiments.

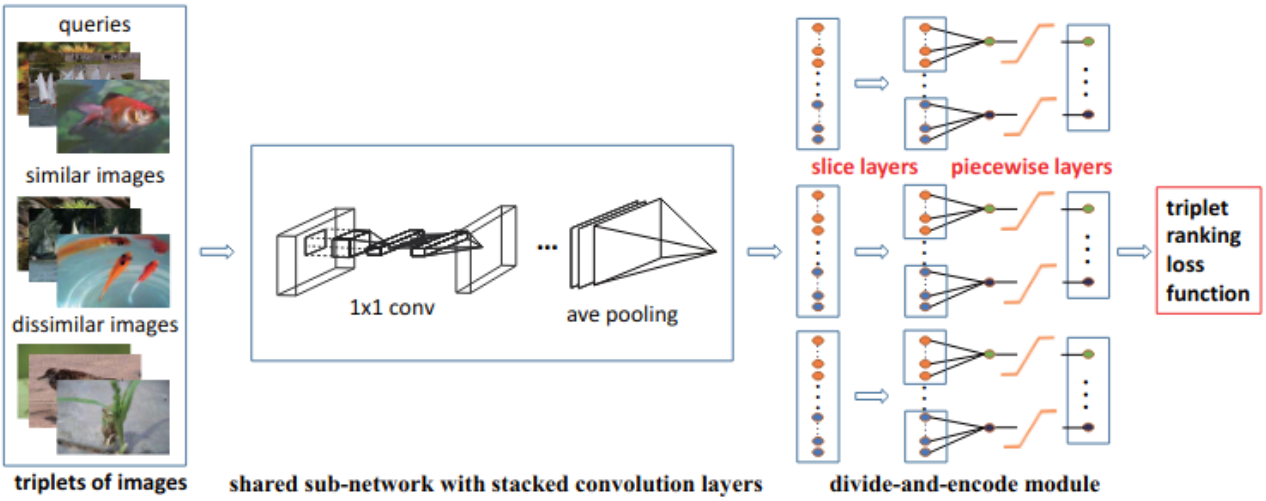

Figure 41 Overview of the hashing architecture introduced in [Lai, 2015]

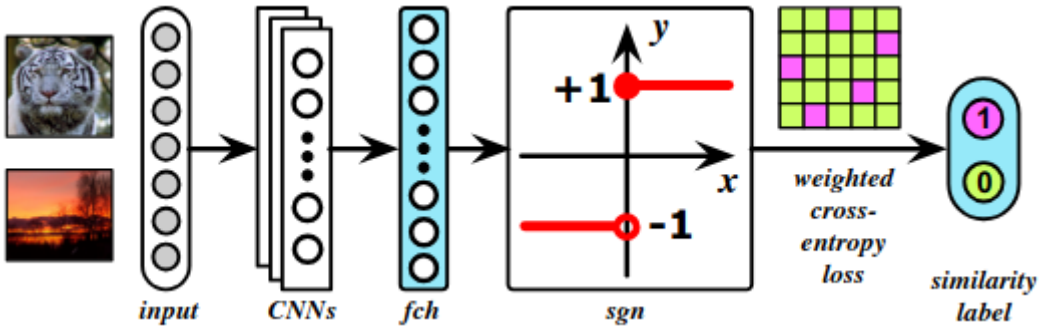

Figure 42 HashNet for deep learning, proposed in [Cao, 2018]

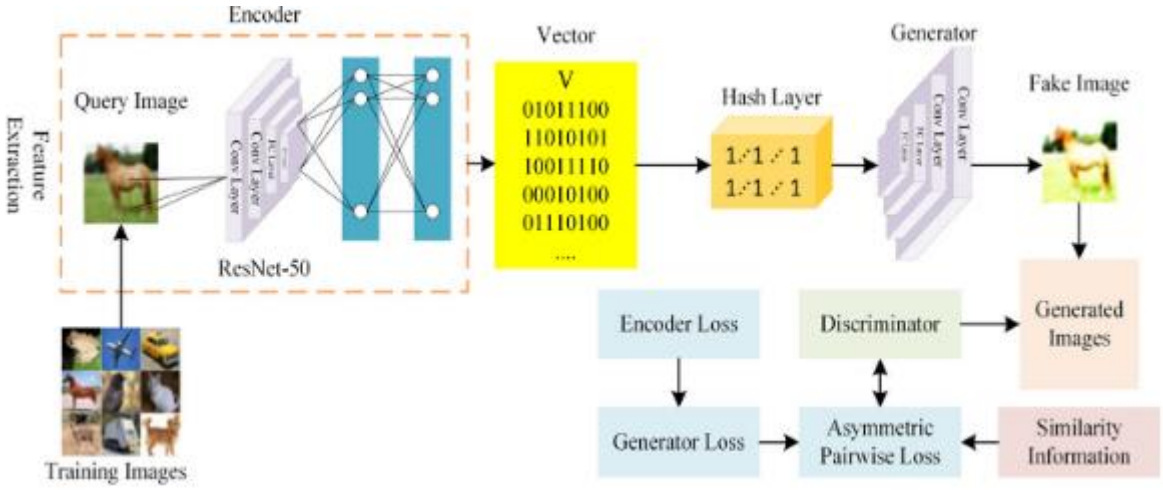

Figure 43 Pipeline of AGAN method [Hassan, 2023]

3.1.5 Speech Enhancement

Speech enhancement is a wide term that could refer to any action or operation whose goal is to improve the intelligibility of some speech recording. This can be done using several quite different techniques, such as reducing the amount of background noise in the audio (denoising), attempting to improve the quality of the speech itself (speech enhancement), and also isolating the speech from any other noise or speech (source separation).

An open source speech processing library, SpeechBrain [Ravanelli et al., 2021, Ravanelli et al., 2024], features different models and solutions pertaining to many tasks of the audio/speech processing field. This library can also serve as a useful overview of the advances in the field, as it often features both older and more recent approaches.

Speech enhancement models can be divided into those that work with time-based and time-frequency domains; and the time-frequency domain ones can further be divided onto those that learn to mask and those that learn to map the input/output [Liu et al., 2023a].

The latest approaches are often based on the transformers architecture [Vaswani et al., 2017]. Examples include MP-SENet [Lu et al., 2023], an encoder-decoder model with some transformer-based features which directly enhances magnitude and phase spectra. SepFormer [Subakan et al., 2020] is a speaker separation model that makes use of the transfomer architecture in its masking network, which is the second component of the three-part model. The encoder encodes the input, the masking network's job is to predict a mask for each of the speakers in the encoded input, and then the decoder formats the output. MossFormer2 [Zhao et al., 2023] is a hybrid model that combines transformer and recurrent neural network architectures, achieving state-of-the-art results on several speaker separation datasets.

Diffusion models [Sohl-Dickstein et al., 2015, Ho et al., 2020, Song and Ermon, 2019, as cited in [Richter et al., 2022]] have been quite successful in the image processing domain [Scheibler et al., 2022]. Some researchers have started experimenting with applying diffusion-based approaches to speech processing. [Kong et al., 2020] train a diffusion model for waveform generation, and [Lutati et al., 2023] apply it to source separation and report that the proposed upper bound (of improvement) can be surpassed. [Scheibler et al., 2022] train a diffusion-based model and demonstrate that it has potential in both source separation and speech enhancement, showing itself to be competitive with prior work in the latter task. [Richter et al., 2022] focus specifically on speech enhancement and show that their diffusion model outperforms all baselines in mismatched conditions (test data from a different corpus than training data).

[Liu et al., 2023a] aim to simplify their approach as a response to overly-complex existing models, and in doing so they create a novel neural network block and train a straightforward mask-free model called MFNet that directly maps speech or reverse noise.

A novel architecture named Mamba [Gu and Dao, 2023] was presented recently to address the shortcomings of the transformer architecture. The authors claim to achieve state-of-the-art performance in many domains, including the audio domain. [Chao et al., 2024] put Mamba to the test and apply it to the speech enhancement task - and conclude that it is a promising architecture. At the time of writing this text, their SEMamba model holds second place on the VoiceBank+DEMAND dataset [Valentini-Botinhao, 2016] leaderboard (https://paperswithcode.com/sota/speech-enhancement-on-demand).

3.1.6 Keyword Search and Spotting

Keyword search and spotting refers to the task of identifying certain words in an audio. It is different from automatic speech recognition in the sense that not every word needs to be recognized nor transcribed - we are only looking for specific words. This process is often done on edge devices and therefore should be low-power and not computationally intensive [Berg et al., 2021].

Models were usually trained with specific keywords in mind, for example, they were trained to spot ”yes”, ”no”, etc. in streaming speech. Recently a shift to language-agnostic, open-vocabulary approaches has started happening, and non-streaming circumstances are also being researched.

[Berg et al., 2021] adapt the transformer architecture [Vaswani et al., 2017] to the task of keyword spotting and find that their model, named Keyword Transformer, outperforms more complex architectures. At the time of writing this text, one of the variants of the Keyword Transformer still holds second place at the Google Speech Commands V2-12 [Warden, 2018] leaderboard (https://paperswithcode.com/sota/keyword-spotting-on-google-speech-commands). While they focus on already-recorded audios, [Wang et al., 2021] adapt the transformer to streaming audio wake-word detection, and reports a 25% improvement when compared to a baseline neural network.

[Bovbjerg and Tan, 2022] train the Keyword Transformer using self-supervised training (a teacher-student approach from [Baevski et al., 2022]) followed by task-relevant fine-tuning and note that self-supervised pretraining improves the supervised baseline by around 10%. [Mork et al., 2024] further experiment with this approach to make it more robust to noisy conditions. They test using clean and/or noisy data at every step of the training process and report that the best-performing models in noisy conditions (whose performance in clean conditions is not greatly affected) are achieved by training the student model with noisy data and the teacher model with clean data. [Li et al., 2024] note that not only noise is an obstacle to keyword spotting in real-world conditions, but also multiple simultaneous speakers. Improving and adapting a combination of older approaches [Yu et al., 2018, Yu et al., 2016, Kolbaek et al., 2017], they train a model that can separate the audio into keyword-related and keyword-unrelated channels. Moreover, the user can provide the keyword as either text, audio, or a combination of both.

[Zhu et al., 2023] present a multi-lingual open-vocabulary keyword spotting model that they trained using a variation of contrastive learning [Wu et al., 2022]. They find that their phoneme-based models are good (and better that text models) at generalizing cross-linguistically, showing this research direction as promising, especially in low-resource languages. However, the authors still note a plethora of challenges that remain present, such as the difficulty of creating and curating high quality IPA datasets.

[Ai et al., 2024] focus on user-defined keyword spotting and categorize the existing approaches into three categories: query by example, query by template, and query by cross-modal matching. They train a keyword spotting model that can be prompted using both text and audio queries, and implement a confusable keyword generation, increasing the system’s robustness. Their model showed good performance in English and Mandarin, outperforming previous approaches in both.

[Shams et al., 2024] pretrain a model based on the novel Mamba architecture [Gu and Dao, 2023] using self-supervised learning and then fine-tune it for three downstream tasks, including keyword spotting. While this architecture is much more efficient and faster than the transformer-based model that the authors compare it with, and outperforms it in most tasks, it seems to struggle the most with keyword spotting, where it does not outperform the transformer in both

keyword-spotting tasks in none of its sizes. Still, given the speed and efficiency improvements and the early stages of research, we should consider this work as a relevant research direction for KS as well, where efficiency is of high importance.

3.1.7 Anti-spoofing

Anti-spoofing is the process of detecting speech recordings created for fraudulent purposes. The bi-annual anti-spoofing challenge called ASVspoof had three tracks in 2021: LA (logical access), PA (physical access), and DF (DeepFake) [Yamagishi et al., 2021]. The LA track deals with spoofing attempts over the telephone (using speech synthesis and voice cloning), PA deals with human speech that was recorded and then played and re-recorded to be used in spoofing attempts, and DF refers to spoofing attempts that utilize artificial intelligence and access to a real person's voice to create fraudulent content. The United States of America have outlawed voice-cloning or any other AI generation of voices without consent (https://www.fcc.gov/consumers/guides/deep-fakeaudio-and-video-links-make-robocalls-and-scam-texts-harder-spot), and various news outlets have reported large financial losses caused by deep fakes (Hong Kong company - 23 million Euros[5] ; UK company - 243 000 US dollars[6]).

[Jung et al., 2021] propose a new single system solution, AASIST, which uses a novel graph-based attention layer. At the time of the publication of their results, their approach outperformed the state-of-the-art, and to this day, they are still quoted as the state-of-the-art approach in the ASVspoof-2019 [Wang et al., 2019] LA leaderboard (https://paperswithcode.com/sota/voice-antispoofing-on-asvspoof-2019-la), despite being the oldest listed approach (however, only three are listed).

[Tak et al., 2022] switch the AASIST's front-end with a version [Babu et al., 2022] of wav2vec 2.0 [Baevski et al., 2020] and also experiment with adding another aggregation layer and performing data augmentation. While the pre-training was done using only bona fide data, the model was fine-tuned on the ASVspoof2019 dataset [Wang et al., 2019]. The evaluation is done on the ASVspoof-2021 dataset [Yamagishi et al., 2021] on both the LA and DF tracks, where it was found that the model achieves state-of-the-art performance on both. The authors do note that this model was trained with extra data so the comparison is not just, but nevertheless their approach is still listed as the best one in more recent literature ([Goel et al., 2023]).

[Müller et al., 2023] divide previous work into two categories: those that work in the time domain using STFT or similar processes, and those that directly process raw audio. Stating that both groups suffer from some disadvantages (the former discards phase information, which is key for naturalness, while the latter lack transparency), their approach using complex-valued neural networks amends both issues. Their model outperforms previous methods, uses phase information in its judgment, and can be explained with explainable AI tools.

[Liu et al., 2023b] propose a model that converts audio from mono to stereo, thus exposing audio features that might indicate spoofing attempts. The authors test their model on the ASVspoof-2019 [Wang et al., 2019] PA challenge and report that it outperforms all other mono-input baselines.

[Goel et al., 2023] propose a new two-stage framework (named SSAST-CL) combining the SSAST model [Gong et al., 2022, as cited in [Goel et al., 2023]], Visual Transformer [Dosovitskiy et al., 2020, as cited in [Goel et al., 2023]], Siamese training [Koch et al., 2015, as cited in [Goel et al., 2023]], and contrastive learning [Khosla et al., 2020, as cited in [Goel et al., 2023]]. The first stage tackles representation learning, whereas in the second stage a classifier is trained. Their goal is to work in the (relatively) unexplored field of adapting the ViT to an audio field and to improve the sub-optimal performance of similar previous approaches. The authors compare their approach with some of the best performing models on the ASVspoof-2021 [Yamagishi et al., 2021] LA track and its best-performing baseline and their model ranks fourth (out of seven).

[5] https://www.euronews.com/business/2024/04/10/its-a-scam-how-deepfakes-and-voice-cloning-taps-into-your-cash
[6] https://blog.avast.com/deepfake-voice-fraud-causes-243k-scam

[Liu et al., 2023c] build upon the AASIST model [Jung et al., 2021] by making the anti-spoofing measures speaker-dependent, i.e., the process of discriminating between true and spoofed speech is done with the speaker embedding. They improve on the baseline, but do not achieve state-of-the-art results.

[Saha et al., 2024] focus on developing a Green AI solution. Green AI is a trend in the machine learning or artificial intelligence community which aims to reduce the amount of computational power, model sizes, etc. in order to lower CO2 emissions, while (ideally) still maintaining comparable performance. The authors use wav2vec 2.0 [Baevski et al., 2020] to extract the features and then experiment with several classical machine learning techniques for the classifier. Using support vector machines, they achieve a result competitive with the state-of-the-art, and at the moment of writing this text are situated as the second best result on the ASVspoof-2019 [Wang et al., 2019] LA track (https://paperswithcode.com/sota/voice-anti-spoofing-on-asvspoof-2019-la).

3.1.8 Image Processing

Person's image verification is the task of determining whether two images belong to the same person or not. In our case the goal is to make sure that a post or comment with a photo belongs to the profile being analyzed. Main approach to face verification is to use a measure of image similarity to compare faces.

[Knoche et al., 2023] addressed the need for explainability in face verification systems by proposing a method that generates confidence scores based on facial feature distances between two images and the distribution of these distances across a dataset. Additionally, they developed a novel visualization approach to highlight similar and dissimilar facial regions by systematically occluding parts of the images and analyzing the impact on verification results. This model-agnostic technique does not require access to the internal workings of deep learning models, making it applicable to a wide range of face verification systems. Their method enhances transparency in face recognition and provides meaningful explanations for verification decisions .

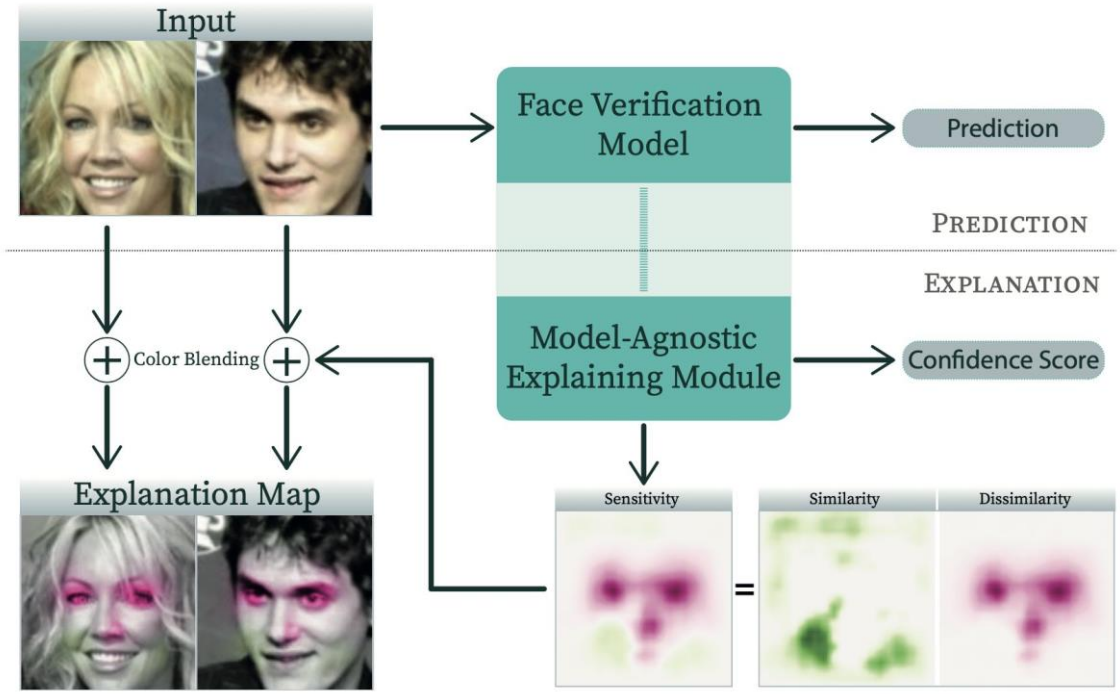

Figure 44 The proposed approach generates a similarity map and blends it with the input images into an explanation map. Besides the binary prediction of the network, we introduce a confidence score to explain the decision further.

[Huber et al., 2023] introduced the xSSAB approach for explainable face verification, which is designed to efficiently backpropagate similarity score arguments to generate visual explanation maps. These maps highlight the areas in a pair of facial images that contribute most to the system's matching or non-matching decision. The method is efficient and training-free, making it suitable for real-time applications where quick and interpretable verification is necessary. Additionally, they presented the Patch-LFW benchmark and a novel evaluation protocol, allowing the first quantitative assessment of the validity of similarity and dissimilarity maps in explainable face recognition .

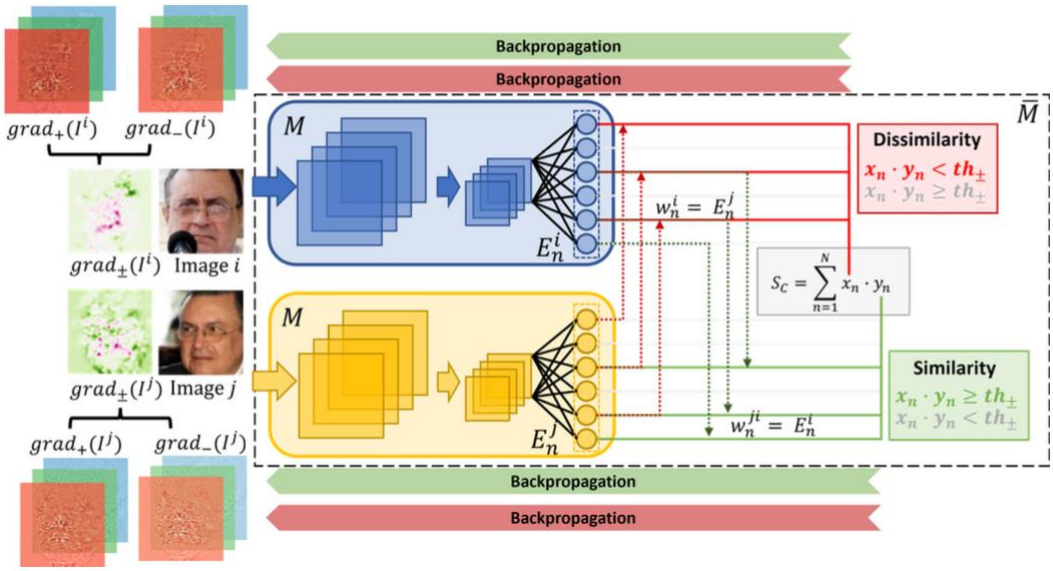

Figure 45 Overview of the proposed approach: in a siamese fashion, both face images are processed by a face recognition system M that is extended with an additional cosine similarity layer.

The xSSAB approach stands out by backpropagating the gradients that contribute to the similarity score, which are then used to generate fine-grained maps indicating similar and dissimilar regions in the images. This method operates in a Siamese network fashion, with the ability to pinpoint specific areas that influence the face verification decision. By combining these gradients, xSSAB produces a single map that visualizes both the similar and dissimilar regions, enhancing the transparency of the face verification process .

3.2 Next steps on Multimedia Processing

In this section, a literature review has been performed, presenting the SoTA of the task tools, related to corruption detection in multimedia content and speech processing in noisy enviroments. Next steps incude the identification of gaps in existing solutions and the development of appropriate tools and methods that will fullfil the project requirements.

The tools developed within this task will be utilized to the Italian, Slovenian and Ukranian pilots. In the Italian pilot multimedia processing tools will serve for issues such as emergency procedure, issue preventions and suspicious events. In the Slovenian pilot is going to deal with the inability to detect hints of fraudulent activities and in the Ukranian pilot with the lack of automation in compliance, anti-fraud and anti-corruption monitoring.

4. Advancements in Econometric and Graph Based Analysis

4.1 Related work and research questions

4.1.1 Economic literature on corruption in Public Procurement

Public procurement is the process by which governments acquire goods, services, and public works from private companies. It is a crucial component of public spending, representing about one-third of global government expenditures. In 2018, this amounted to around $11 trillion, or 12 percent of the global GDP [World Bank, 2018]. Public procurement is vital for sectors such as healthcare, defence, and infrastructure, but the large financial resources involved make it highly susceptible to corruption. Issues like bid rigging [Porter, 1993], cost overruns (Flyvbjerg, 2003), favouritism [Burgess et al., 2015], and collusion between politicians and firms [Coviello, 2017] are commonly identified in academic literature. Public procurement is often cited as one of the areas of government most prone to corruption, with every stage of the process being vulnerable to integrity concerns due to the sheer volume of transactions and the high financial stakes. These corrupt practices undermine transparency and efficiency, leading to poor outcomes and wasted public funds.

Reliable and objective measures are essential to analyse corruption dynamics in public procurement and assess its impact on the misuse of public funds. Numerous studies have suggested various methods for identifying corruption risk patterns.

One approach involves measuring discrepancies between actual and reported expenditures. For example, in a randomized controlled trial in Indonesia, [Olken, 2007], [Olken, 2009] assessed the difference between official project costs and independent engineers' estimates for public works, with unaccounted expenditures classified as corruption. Similarly, [Golden, 2005] compared the value of existing public infrastructure with the government's reported spending in Italy. They proposed a measure of corruption based on the disparity between infrastructure quantity and related public spending across Italy's 20 regions, attributing the higher expenditures to inefficiencies and corruption.

Another approach involves constructing corruption risk indices by identifying "red flags" in the procurement process. These red flags serve as indicators of potential fraud, corruption, or unethical practices, and can appear at various stages, from planning to contract execution. Red flags typically include non-transparent procedures, limited advertisement of tenders, and vague or biased evaluation criteria. A composite indicator of corruption risk can be developed by combining these red flags. For example, [Fazekas, 2020] created an index using equally weighted red flags, linking them to restricted competition. The number of bids submitted, especially when only one bid is submitted, is often used as a proxy for corruption [Klasnja, 2015].

[Lisciandra et al., 2022] combined these two approaches to develop a corruption risk indicator at the procurement level. This was achieved through a two-stage process: first, a potential waste score for each contract was estimated using data envelopment analysis, and then corruption was isolated from inefficiency by regressing waste scores on inefficiency determinants.

Corruption is a significant challenge in public procurement, with estimates suggesting that it siphons off 20% to 25% of national procurement budgets annually, totalling around USD 2 trillion globally [OECD, 2016]. Corruption severely hampers transparency and efficiency, leading to substantial public fund mismanagement and poor economic outcomes. The economic literature has thoroughly examined the detrimental effects of corruption on economic growth through various channels (Rieckmann et al. 2022; Rieckmann and Stuchtey, 2023)

First, corruption leads to the misallocation of public resources, often diverting funds to less productive sectors while simultaneously lowering the quality of public services. Corrupt officials tend to steer public spending toward sectors, such as healthcare and infrastructure, where opportunities for rent-seeking are greater. [Mauro, 1997] found a positive relationship between corruption and public investments, while [Tanzi, 1997] argued that corruption contributes to excessive investment in public infrastructure. Notably, corruption has a particularly negative impact on education

spending. [Mauro, 1998] demonstrated a significant negative correlation between corruption and public investment in education.

Second, corruption causes a decline in productivity. When business success depends more on bribing officials than on actual productive efficiency, entrepreneurs have fewer incentives to improve their productivity and competitiveness.

Third, corruption reduces the overall level of investment. It can be seen as a form of hidden taxation that decreases the expected return on investments, discouraging investors. [Mauro, 1995] found a strong negative correlation between corruption and the investment-to-GDP ratio, indicating the harmful effects of corruption on capital formation.

Finally, corruption distorts markets. [Fazekas et al., 2017] point out that corruption, especially in contexts where competition is limited, allows a small group of favoured suppliers to dominate the market, particularly in clientelist systems. This leads to an uncompetitive market environment, where a few firms secure most of the contracts. [Wittberg, 2023] expand on this, showing that companies winning single-bidder contracts tend to outperform their competitors, highlighting how corruption creates an uneven playing field.

4.1.2 Review of cryptocurrency in financial crime activities

Crypto assets have become a reliable space for the development and persistence of economic transactions stemming from illicit activities related to cybercrime, with detected activities linked to terrorist financing [Brill, 2014], human trafficking [Bartsch, 2020], ransomware attacks and subsequent payments [Connolly, 2017], darknet marketplace purchases, and money laundering [Nicholls, 2021]. Despite recognized entities involved in illicit activities being sanctioned, the advent of cryptocurrencies, stablecoins, and NFTs has provided sanctioned entities with new ways to evade sanctions and continue their illicit activities [Blanchini, 2022], [Reddy, 2018]. These digital assets, which promote decentralization and anonymity, have created an ideal environment for such activities, with nearly 40 billion dollars in illicit transactions in 2022 and over 24 billion dollars in 2023, despite representing less than 0.5% of the total on-chain transaction volume [Chainalysis, 2024].

From this perspective, several crypto assets have been associated with these kinds of activities. In their work “Cybercrime on the Ethereum Blockchain,” Hornuf et al. [Hornuf, 2023] identified, through transaction analysis, 19 different categories associated with cybercrime, with the total of these transactions exceeding 1.75 million and estimated losses surpassing 1.65 billion dollars by 2021. Other crypto assets, such as Monero, Zcash, and Dash, have also been identified as viable paths for criminals, posing an even greater challenge for law enforcement [Bele, 2021].

However, Bitcoin remains the most used cryptocurrency in cybercrime-related activities. According to Cong et al., Bitcoin has emerged as the primary means for payments in ransomware attacks, money laundering, and other forms of cybercrime due to its widespread acceptance and ease of use. Data from BitcoinAbuse.com reveals that out of the 13.6 million cybercrime-related transactions on the Bitcoin blockchain, 42.5% are linked to ransomware. This activity not only leads in the number of transactions but also dominates Bitcoin payments associated with illicit activities, accounting for 86.7% of the total received, equivalent to 156.3 million dollars as of April 2022 [Cong, 2023].

Focusing on studies aimed at detecting these types of activities, Wahrstätter et al. employed a Bitcoin user graph analysis approach to detect money laundering activities. Through unsupervised learning techniques, they identified anomalous behavioral patterns in users close to CoinJoin, utilizing a ground truth dataset from GraphSense [Wahrstätter, 2023]. Another relevant study is that of Nicholls et al., which introduces "FraudLens," an approach based on Graph Neural Networks (GNNs) for identifying illicit activities on the Bitcoin network, proposing two graph preprocessing techniques: "Edges based on Affinity (EA)" and "Edges based on Node Features (ENF)," using the Elliptic dataset [Nicholls, 2023].

4.1.3 Social science literature on corruption in Public Procurement

Public procurement, the process by which governments acquire goods, services, and public works, is not just a financial and administrative operation but a socially embedded activity where power dynamics, norms, and social networks play a critical role, particularly when it comes to deviant procedures such as fraud, embezzlement, or corruption. While economic approaches often emphasise efficiency losses and financial consequences, a social science perspective highlights the social structures and cultural contexts that shape both the practice and perception of criminal behaviour in public procurement.

In literature, crimes in public procurement are often seen as a natural extension of societal inequalities and power imbalances. Scholars argue that corrupt practices are normalised in environments where informal social networks and power relations override formal legal frameworks (van Deth & Zmerli 2010). In such contexts, procurement decisions are often based on personal connections, patronage, and kinship ties rather than transparent competition. This network-based corruption reinforces existing hierarchies, benefiting elites with greater access to social capital while marginalising groups with less influence, such as women or lower socio-economic classes (Swamy et al., 2001).

Cultural norms play a significant role in shaping attitudes toward deviant behaviour such as corruption. What may be seen as deviant behaviour in one society could be considered acceptable in another, particularly where trust in public institutions is low and informal relationships are critical for getting things done (Golden and Picci, 2005; Graeff & Svendsen 2013). Similarly, social capital, such as universal trust, or lack thereof, influences these behaviours [López et al., 2014; Putnam, 2000]. Treisman (2000) highlighted that historical legacies, such as colonialism and Protestant traditions, significantly influence corruption levels, with countries that have experienced British rule often displaying lower levels of corruption. Such insights align with social science perspectives, emphasising how deeply rooted cultural and institutional factors shape corruption patterns over time. Empirical research further shows that practices like gift-giving and favour exchanges, while violating formal procurement regulations, are often embedded in the cultural fabric of certain societies, making them resistant to reforms that focus solely on legal aspects (Søreide, 2014; Weber Abramo, 2008).

The role of social networks in perpetuating deviant behaviour, such as corruption, is another key focus in the social science literature. Such crimes often thrive in tightly knit communities where trust and reciprocity between individuals enable illicit exchanges (Graeff 2009). These informal relationships can make procurement processes opaque, as decisions are made based on loyalty or obligations within these networks rather than through transparent, merit-based competition (Larsson and Grimes, 2023). Moreover, the concentration of power among elites in such networks excludes underrepresented groups, further entrenching inequality (Blagojević and Damijan, 2013).

Corruption perception indices, such as Transparency International's Corruption Perceptions Index (CPI)(Transparency International, 2024), are often used as indicators for the amount of state-related crimes, in particular corruption. However, social scientists have critiqued these indices for relying heavily on perceptions, which may not always align with actual experiences of corruption (Donchev and Ujhelyi, 2014). In many cases, administrative crimes are hidden behind formal procedures, making it difficult for outsiders to detect. This discrepancy between perception and reality suggests the need for more nuanced, experience-based measures of corruption that capture the complex social processes at play (Olken, 2009).

Gender is another important dimension in the empirical study of corruption. Research has shown that women are generally less involved in corrupt practices due to their relative exclusion from male-dominated power structures (Swamy et al., 2001). However, when women do enter these structures, they may face pressure to conform to deviant norms in order to succeed. The exclusion of women and other marginalised groups from procurement processes reflects broader social inequalities and underscores the importance of addressing corruption as a social issue, not just a legal or economic one (Van Rijckeghem and Weder, 2001).

Finally, civil society plays a crucial role in countering such crimes. Strong civil society organizations (CSOs) can act as watchdogs, monitoring procurement processes and advocating for greater transparency (Larsson and Grimes, 2023). However, the effectiveness of these organizations depends on the broader social and political context. In societies

where democratic institutions are weak, CSOs may struggle to hold governments accountable, and anti-corruption efforts can be co-opted by powerful elites for their own benefit (Weber Abramo, 2008).

4.2 Next steps on Econometric and Graph based Analysis

4.2.1 Econometric analysis

To identify patterns of corrupt practices, the proposed methodology draws from the economic literature on public procurement and involves several key stages.

The first stage involves using publicly available and comparable data from various countries. Data sources may include the Tenders Electronic Daily (TED) covering European Union public procurement from 2006 to 2023, the Global Public Procurement Dataset (GPPD), and data from national statistical offices and international organizations like the World Bank.

The second stage focuses on identifying relevant red flags, chosen based on the literature and issues raised by key stakeholders. These red flags might include:

1. Unusually short deadlines for bid submission.

2. Vague or biased procurement specifications.

3. Limited advertisement of tenders.

4. Unjustified sole sourcing of contracts.

5. Frequent contract change orders.

6. Overly high or unusual pricing patterns.

7. Lack of transparency in the bid evaluation process.

8. Connections between procurement officials and bidders.

9. Multiple contracts awarded to the same supplier.

10. Poor contract management and monitoring.

11. Unexplained changes in project scope.

12. Unusual or unexplained payments.

13. Excessive use of consultants.

14. Absence or delay of audits.

15. Inadequate documentation.

The third stage involves creating and validating a composite index of corruption risk by combining these red flags. Following the method proposed by Fazekas and Kocsis (2020), a simple arithmetic average of red flags, scaled between 0 (lowest observed corruption risk) and 1 (highest observed corruption risk), is used to build the index. Binary logistic regression is applied to identify the most significant red flags. The index's validity is tested by correlating it with other available measures of corruption.

The fourth phase involves assessing the impact of corruption on various economic outcomes. To achieve this, a dynamic panel regression analysis is proposed. In this framework, the equation to be estimated will include our proposed measure of corruption as an independent variable, with proxies for different economic indicators as dependent

variables. Additionally, a lagged dependent variable will be included to account for temporal persistence, as economic indicators often exhibit significant inertia over time. The value of an indicator in year t is likely influenced by its value in the previous year. Furthermore, we will incorporate a set of control variables and time-invariant fixed effects to address unobserved heterogeneity.

To mitigate bias resulting from the dynamic nature of the model, in addition to using a traditional fixed effects estimator, we propose employing the Arellano-Bond system GMM approach. In this method, the endogenous lagged variable is instrumented using higher-order lags, which are strongly correlated with the first lag but do not bias the error term due to the presence of only one lag in the equation.

Another potential source of bias is reverse causality. To address this, we will use coarsened exact matching (CEM), which helps mitigate bias by ensuring more precise comparisons between treated and untreated groups.

4.2.2 Kriptosare: Graph-based solution for detecting illicit in cryptocurrency network

To effectively identify patterns indicative of potential fraud in cryptocurrency transactions, a comprehensive methodology will be implemented, centred on the construction of address transaction graphs. These graphs will be meticulously structured around a central node, defined by the target address, from which the graph will be generated, encompassing all associated incoming and outgoing transactions.

The methodology will unfold through several key stages:

- **Graph Construction:** The process begins with the precise identification of the target address, followed by the systematic collection of all transactions linked to this node. These data will form the foundation of the address transaction graph, capturing the entirety of transactional activity associated with the target address. This graph will serve as the central structure upon which subsequent analyses will be conducted.
- **Feature and Relationship Extraction:** Once the graph is constructed, the next stage involves a thorough analysis of its economic and structural features. Key characteristics that may reveal latent patterns or significant relationships within the transaction network will be identified and extracted. The focus will be on discovering consistent patterns that could indicate fraudulent activities, considering both economic metrics and the structural properties of the graph, and evaluating the effectiveness of various extractable features. This dual approach is crucial to enhance the reliability of pattern detection, enabling a more precise and detailed understanding of transactional dynamics.
- **Heuristic Entity Clustering:** To further refine the analysis, heuristic clustering methods will be employed. These methods will group individual entities within the graph into complex clusters that represent the aggregated characteristics of entire wallets. By adding this aggregation layer to the graph, the methodology not only simplifies the analysis but also deepens it, allowing for the detection of more sophisticated patterns that may indicate fraudulent behaviour.

For the detection of fraudulent behaviours, several advanced techniques will be employed:

- **Classification and Clustering:** Classification methods will be applied to compare new samples with previously known behaviours. Additionally, clustering techniques will be utilized to identify the proximity of new samples to already labelled data, providing an additional layer of confidence in predictions.
- **ML Techniques based on GNNs:** At the core of the advanced analysis, Graph Neural Networks (GNNs) will be employed, which are particularly well-suited for analysing complex graph structures. GNNs enable the extraction of deeper, more intricate patterns from transaction graphs, capturing relationships and dependencies that traditional methods might overlook. This approach not only enhances the accuracy of fraud detection but also ensures that the system remains adaptable to the evolving nature of fraudulent tactics, as demonstrated in recent studies.

The Kritposare tool will receive in input a cryptocurrency address, and it returns: 1) a classification of the entity behavior that control such address, 2) statistical information of the address (amount received, sent, etc.), 3) OSINT information regarding the user (if he/she has been sanctioned by OFAC o appeared in other relevant datasets), 4) if the address is related directly with potential risk services (mixers, gambling, other sanctioned entities, etc.)

- Pilot-Italy: **UC 6** (Adequacy and issue prevention)
- Pilot-Slovenia: **UC 7** (Fraudulent Activities of High-Risk Legal Entities)

4.2.3 Social Science Analysis

To explore the connections between societal factors and criminal behaviour, such as corruption, we first need to establish reliable ways to measure such crimes. Since criminal activities are often conducted covertly, absolute and reliable measures are inherently difficult to obtain. To address this challenge and allow for comparisons between countries, the Corruption Perceptions Index (CPI) published by Transparency International has frequently been used as a proxy measure for corruption-related crimes (Transparency International, 2024). However, the CPI and similar indices have limitations in that the perceived amount of crimes does not necessarily correlate well with the actual amount and is confounded by institutional variables [Bjørnskov 2006; Weber Abramo 2008, Dreher and Schneider 2010].

In recent years, more objective data on corruption and related crimes have become available, such as those provided in the Eurostat database [Eurostat, 2024]. However, this data also reflects each country's legal framework, reporting practices, and the stage at which a crime case is identified and recorded. These factors inhibit immediate cross-country comparisons.

To overcome challenges like this, our approach involves using the Tenders Electronic Daily (TED) dataset, which contains detailed information on public procurement across the European Union. By applying corruption "red flags" identified in the literature, such as those suggested by Fazekas [Fazekas and Kocsis 2017], we aim to create a more comparable measure of crimes related to procurement data that are less influenced by societal reporting biases.

Once we have developed this refined measure of corruption, the next step will be to merge this data with social science datasets such as the European Social Survey (ESS), the World Values Survey (WVS), and the Human Development Index (HDI). This integration will enable us to analyse the influence of various social factors on crimes, such as corruption, and vice versa.

Given the complexity and size of the datasets, our analysis will require an adaptive and recursive approach. We will employ methods such as Principal Component Analysis (PCA) to reduce dimensionality and identify key patterns in the data. Additionally, we will use advanced multivariate linear regression techniques, such as Elastic Net models, which allow for regularisation and variable selection in high-dimensional datasets. This flexible methodology will enable us to account for empirical results deviating from literature and previous studies, ensuring our findings are as unbiased and accurate as possible.

Ultimately, this approach allows us to respond dynamically to the data, adapting our methods to uncover the most significant social factors that either foster or prevent crimes such as corruption. This comprehensive analysis will provide valuable insights for policymakers and stakeholders seeking to address the social roots of crimes in public procurement. In particular, our results can provide a context picture that applies to the pilot states and give the opportunity to evaluate the specific pilot case data.

5. Data Mining and Correlation

5.1 Related work and research questions

5.1.1 Correlation and graph analysis

The knowledge provided by different WP4 analysis tools described in previous sections need to be integrated in a common structure. The knowledge graph can maintain the original source data structures and be enriched with the results of local level analysis. This graph structure provides a complete overview of the actual status and allows generating a global analysis of the complete structure or the comparison between separately analyzed elements.

Graph complexity has been a widely studied field. For example, the work in [Pudlák, 1988], which was published in 1988, developed a complexity theory based on graphs instead of boolean functions. In [Jukna, 2006], the complexity of a graph is defined as the minimum number of union and intersection operations needed to obtain the whole set of its edges starting from stars, and the motivation of the work is to study the complexity given that definition, with the aim of understanding the relation of it to the circuit complexity of boolean functions. Entropy has been also studied to measure graph complexity. Example of it is the work [Mowshowitz, 2012], where a taxonomy and overview of graph complexity measurements are carried on, as it is done similarly in [Dehmer, 2011] or in [Zenil, 2018], where authors survey and contrast (algorithmic) information- theoretic methods which have been used to characterize graphs and networks. Another interesting work is the one in [Chen, 2014], being the main contribution, the study of graph entropy based on a functional which considers the number of nodex with distance k to a given node.

Correlation between different nodes in a network graph is calculated by computing similarity measures considering subgraph structure of its neighbors [Newman, 2003]. Subgraph similarity can be provided by structural properties of the graph, such as shortest paths [Wang, 2020], subtrees [Shang, 2010] and random walks [Fouss,2007] [Xia, 2019], etc. As only local structure of the graph is considered, the global structure information remains regardless. Nikolentzos et al. [Nikolentzos, 2017] propose two algorithms to calculate the level of similarity of two networks applicable to labelled and unlabelled graphs working with the complete graph.

Risk assessment is a crucial process used to identify, evaluate, and prioritize potential hazards that could negatively impact an organization or individual. It involves systematically analyzing the likelihood and consequences of various risks, ranging from financial and operational threats. By assessing these risks, organizations can implement appropriate measures to mitigate or manage them, thereby minimizing potential damage. For example, in [Ganin, 2020], a framework for precisely bridging the gap between assessment and management, which is not trivial. In [Kalinin, 2021], authors propose a systemization for cybersecurity risk assessment which is applied in Smart City infrastructures. Another interesting work is the one proposed in [King, 2018]. In this work, the authors are more focused on human factor induced risk. Specifically, the work is guided in two steps. Firstly, they review human maliciousness-related literature so in a second step, they can set up an initial set of proposed assessment metrics to characterize human maliciousness.

The main characteristic of graph-based data representations is their ability to capture structural relationships, and interactions between elements of the knowledge base. There are numerous techniques within the complex networks discipline that aim at deriving insights from the underlying structure.

Graph motifs are one promising graph analysis direction for extracting common interaction patterns within a graph. Motifs are small, recurring subgraph patterns within a larger network. There are multiple types of motifs, such as triangles with 3-node interactions, or chains of links up to a certain length [Jazayeri, 2020]. These motifs capture significant topological and functional information about the overall graph [Conway, 2011].

When exploring temporal and dynamic graphs motifs can also be extended to consider explicitly the time when each of the analyzed connections took place. This way, clear boundaries on precedence, as well as validity time windows can be considered to make each identified pattern more meaningful when compared to the base structure from temporal patterns.

These characteristics have led temporal motifs to be tested for studying problems similar to CEDAR, such as financial transactions [Liu, 2023b], call networks [Kovanen, 2013], email interactions [Barnes, 2024] and patent networks.

To illustrate how temporal motifs can help we can look at what they show on transaction networks. A sample temporal motif would be as follows: A sends an amount to B, then B sends the amount to C. We can see B as the ´middleman´ in this transaction. If we see this pattern often within the temporal graph (by counting the instances of each motif appearance) we can further classify these nodes as buyers, temporal exchange venues, or currency converter. This way, motif counting within cryptocurrency networks has been used to detect mixer service addresses [Wu, 2021] (mixers are services designed to obfuscate the intended sender and recipient of a transaction) or for detecting phishing scams in Ethereum [Wu,2023b]. The results in these papers point at the usefulness of motif-based analysis to improve over base classification methods when exploring transaction networks.

5.1.2 Financial Transaction Analysis

This task addresses the research question: "How can we identify correlations and anomalous patterns in bank transactions to detect potentially fraudulent activities?" This involves detecting deviations from normal transaction behavior, recognizing suspiciously high transactions, identifying complex transaction patterns such as circular and wedge patterns, and finding correlations between financial transactions and other forms of communication, such as phone calls and emails.

State-Of-The-Art

Detecting fraudulent activities in banking transactions is a critical area of research, driven by the need to prevent financial crimes such as money laundering, terrorist financing, and various types of fraud. The state of the art in this field leverages a combination of machine learning, graph mining, and data analysis techniques to identify correlations and anomalous patterns.

Fraud detection can be effectively framed as a graph anomaly detection (GAD) task, where the goal is to identify unusual or suspicious patterns within a graph. GAD aims to detect anomalies at different levels: nodes, edges, and subgraphs, each corresponding to distinct types of fraudulent activities. For instance, anomalous nodes may represent fraudsters, anomalous edges may indicate fraudulent transactions, and anomalous subgroups may reveal patterns of money laundering.

Over recent years, traditional approaches to GAD have been increasingly supplanted by deep learning (DL) techniques, which excel at capturing the non-linear and temporal structural patterns inherent in dynamic graphs. These techniques are particularly effective in fraud detection, where patterns of behavior can evolve rapidly and in complex ways.

GAD techniques can be broadly classified into supervised, semi-supervised, and unsupervised approaches. Given the scarcity of labeled data in fraud detection, semi-supervised and unsupervised methods are often preferred. These methods do not rely heavily on large volumes of labeled anomalies.

When modeling dynamic graphs as streams of edges, DL techniques update node representations with each new edge, allowing for the continuous capture of evolving node patterns. Several neural network architectures have been proposed for this purpose:

1. JODIE [Kumar, 2019] utilizes Recurrent Neural Network (RNN) modules to derive dynamic node representations, effectively capturing the temporal dynamics.
2. DyRep [Trivedi, 2019] combines a deep temporal point process with RNNs to model the temporal evolution of node representations.
3. TGAT [Xu, 2020] integrates temporal encoding and graph attention mechanisms to incorporate time information during neighborhood aggregation.
4. DDGCL [Tian, 2021] employs a contrastive learning approach based on temporal smoothness, learning node representations by comparing the same nodes across closely spaced time intervals.
5. TGN [Rossi, 2020] incorporates a memory module updated via an RNN to store long-term patterns, enabling the modeling of both temporal and spatial characteristics.
6. SAD [Tian, 2023] enhance detection performance by combining memory banks with pseudo-label contrastive learning, further improving the robustness of fraud detection.

When dynamic graphs are modeled as a sequence of snapshots, different approaches are employed. For example:
1. NetWalk [Yu, 2018] utilizes random walks and autoencoders to create node representations, identifying anomalies based on dissimilar interactions between nodes.
2. AddGraph [Zheng, 2019] constructs node representations by combining short-term structural patterns with long-term temporal patterns, using a Graph Convolutional Network (GCN) and a Gated Recurrent Unit (GRU).
3. TADDY [Liu, 2021] leverages transformers to capture both global and local structural patterns, yielding comprehensive node representations for anomaly detection.

Generative Adversarial Networks (GANs) have also gained attention in the fraud detection domain due to their ability to model complex data distributions. For instance, the OCAN framework [Salehi, 2019] bypasses the need for labeled fraudsters by focusing solely on benign users' attributes. It employs an LSTM-based autoencoder to distinguish between benign and malicious users in feature space and uses a GAN to generate synthetic data points that help refine the discriminator's ability to identify anomalies.

Graph Neural Networks (GNNs) have become increasingly prominent in fraud detection, especially in tackling challenges such as class imbalance and camouflage. Liu et al. [Liu,2021a] propose a GNN-based approach that addresses class imbalance by using a label-balanced sampler to construct sub-graphs for training, improving detection in imbalanced datasets. CARE-GNN [Dou, 2020] counteracts feature and relation camouflage in fraud detection by optimizing neighbor selection and aggregation through reinforcement learning. Spade [Jiang, 2022] is a real-time fraud detection framework that incrementally maintains dense subgraphs in dynamic networks, enabling low-latency detection. Finally, Cheng et al. [Cheng, 2023] focus on detecting money laundering by modeling group interactions within transaction networks, uncovering complex patterns that might be missed when analyzing individual actions alone.

5.1.3 Fusion of text and visual content

Within the framework of CEDAR project the combination of different modalities is required to highlight fraud and corruption operations. In the following paragraphs, methods of the existing literature are presented, in the field of text and visual content fusion. Those methods should correlate named entities with visuals and the analysis results of task 4.2.

The fusion of multiple modalities in AI systems enhances their capabilities and effectiveness across diverse applications. By combining various data types—such as text, images, audio, and video—multimodal systems provide a more comprehensive understanding of information. This integration improves contextual interpretation and resolves ambiguities, boosts accuracy and robustness, enhances user interactions, making them more natural and intuitive.

Multimodal learning techniques can be classified into early fusion, intermediate fusion, late fusion, and hybrid fusion. In early fusion, the raw or pre-processed data from each modality are combined before being input into the model. In intermediate fusion, the features extracted from different modalities are fused and then sent to the model for decision making. In late fusion, the individual decisions from each modality are combined to form the final prediction, using methods such as majority vote, weighted average, or a meta-machine learning model that integrates the individual decisions. Finally, hybrid fusion is a combination of early, intermediate, and late fusion. [Zhao, 2024a]

The integration of multimodal fusion in large language models (LLMs) significantly enhances their functionality and applications by combining diverse data types, such as text, images, audio, and video. This approach allows LLMs to develop a more nuanced understanding of content, improving their ability to generate contextually relevant responses. Multimodal LLMs benefit from this fusion by bridging gaps between different data types. For instance, models that process both visual and textual inputs can provide more precise answers by correlating information from images with textual queries [Wu, 2023] [Jin, 2024]. Some successful multi-modal models are:

The CLIP model [Radford, 2021] is a powerful neural network designed to bridge the gap between vision and language. It achieves this by learning to associate images with their corresponding textual descriptions through contrastive learning, which involves training the model to differentiate between correct and incorrect pairs of images and captions.

CLIP leverages a large dataset of image-text pairs to create a joint embedding space where visual and textual data are aligned, enabling it to understand and generate relevant captions for images, as well as perform zero-shot classification tasks without requiring task-specific training. This approach has significantly improved the ability of AI systems to interpret and generate multimodal content, offering robust performance across a wide range of applications.

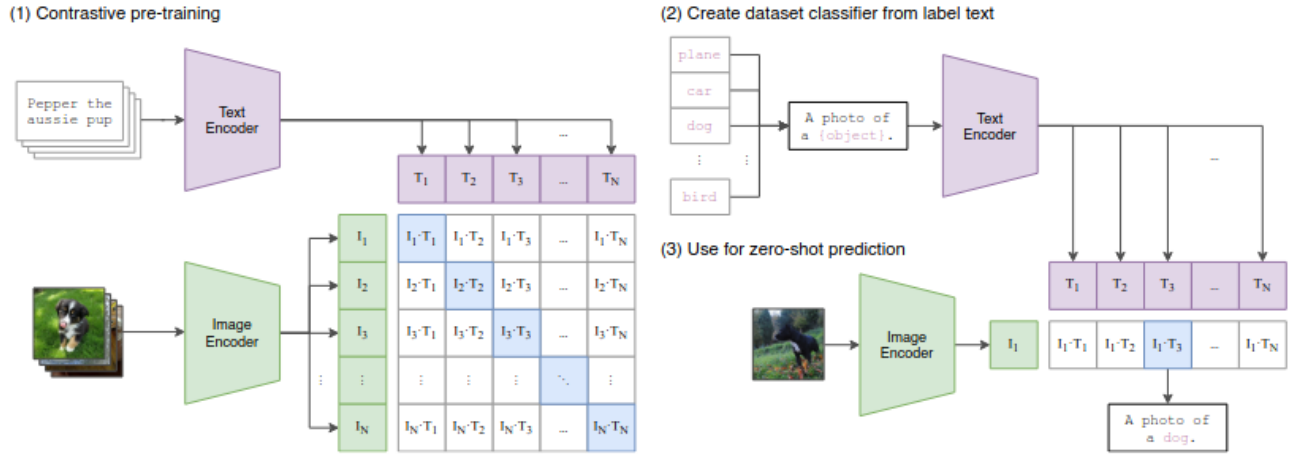

Figure 46 The CLIP architecture

BLIP-2 [Li, 2023] is an advanced vision-language pre-training method designed to leverage frozen pre-trained unimodal models. The core innovation of BLIP-2 is the Querying Transformer (Q-Former), a trainable component that acts as an intermediary between a frozen image encoder and a frozen large language model (LLM). Q-Former consists of two transformer modules with shared self-attention layers: an image transformer that interacts with the fixed image encoder to extract visual features, and a text transformer that handles both encoding and decoding tasks. This architecture uses a set of learnable query embeddings, which interact with each other through self-attention and with the frozen image features through cross-attention layers. The queries are designed to be a bottleneck, forcing them to extract the most relevant visual information for the text, thus bridging the modality gap effectively.

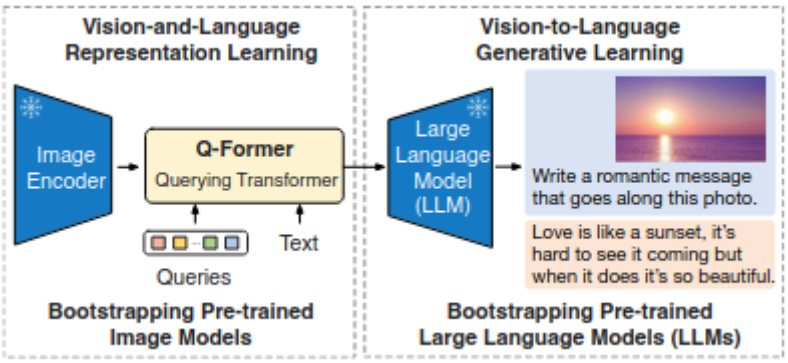

Figure 47 The BLIP-2's architecture

The LLava model [Liu, 2024] is an advanced AI system designed to integrate visual and linguistic understanding. By leveraging large-scale pre-training on multimodal data, LLava excels in tasks that require the interpretation of both images and text. This model employs a sophisticated architecture that combines vision transformers and large language models, enabling it to process and generate coherent and contextually relevant responses based on visual inputs. LLava's training involves contrastive learning and cross-modal attention mechanisms, which help it align visual features with corresponding textual descriptions effectively. This integration allows LLava to perform a variety of tasks, such as

image captioning, visual question answering, and text-based image retrieval, with high accuracy and relevance, making it a versatile tool in the realm of AI-driven multimodal applications.

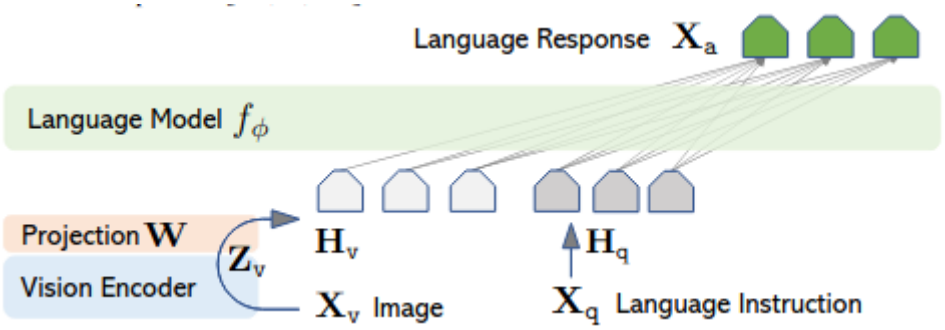

Figure 48 LLaVa network architecture

Flamingo [Alayrac, 2022], represents a significant advancement in the field of few-shot learning for multimodal tasks. Flamingo integrates large language models with powerful visual representations, leveraging pre-trained and frozen components to achieve impressive performance without the need for extensive task-specific training data. Its architecture features novel components like the Perceiver Resampler, which processes visual inputs, and cross-attention layers that enable the language model to incorporate visual information effectively. This design allows Flamingo to handle interleaved sequences of text and visual data, enabling it to perform well on tasks such as image captioning, visual dialogue, and visual question answering with only a few examples.

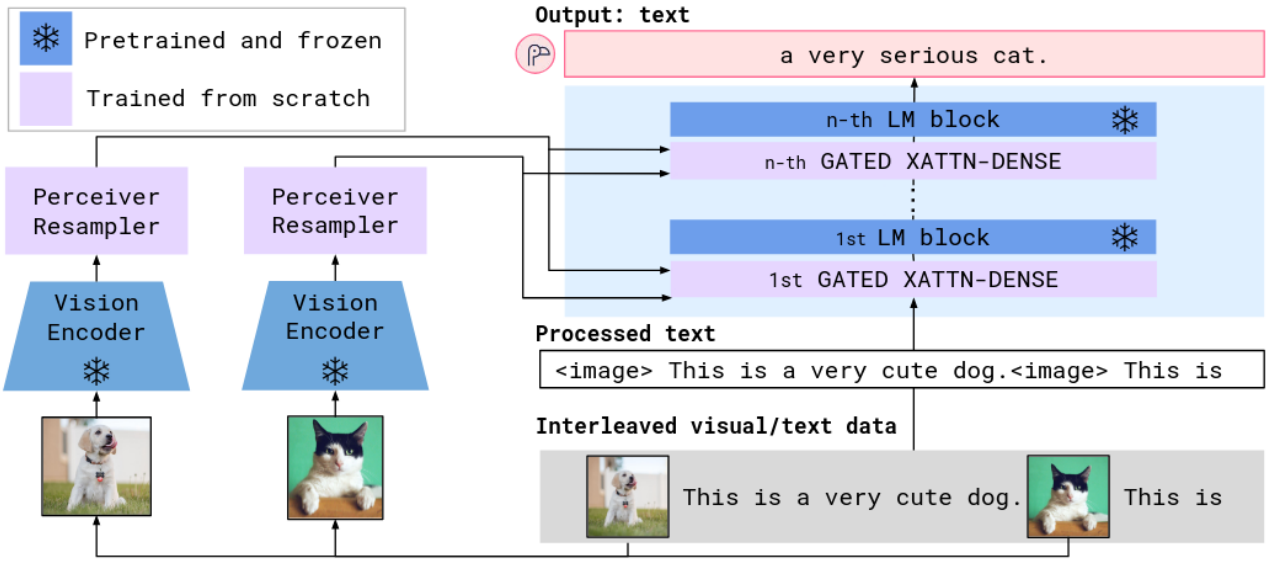

Figure 49 Flamingo architecture overview

KOSMOS-1 [Huang, 2024] is an advanced multimodal large language model (MLLM) designed to process and integrate multiple types of data inputs, such as text and images, using a Transformer-based architecture. Central to KOSMOS-1 is its Transformer decoder, which generates text in a sequential, auto-regressive manner—predicting the next token based on the previous ones. This model uniquely handles multimodal data by embedding different input types into a unified representation. For text inputs, it uses special tokens to mark the beginning and end of sequences, while images are converted into embeddings using a vision encoder and are integrated into the text sequences with additional special tokens.

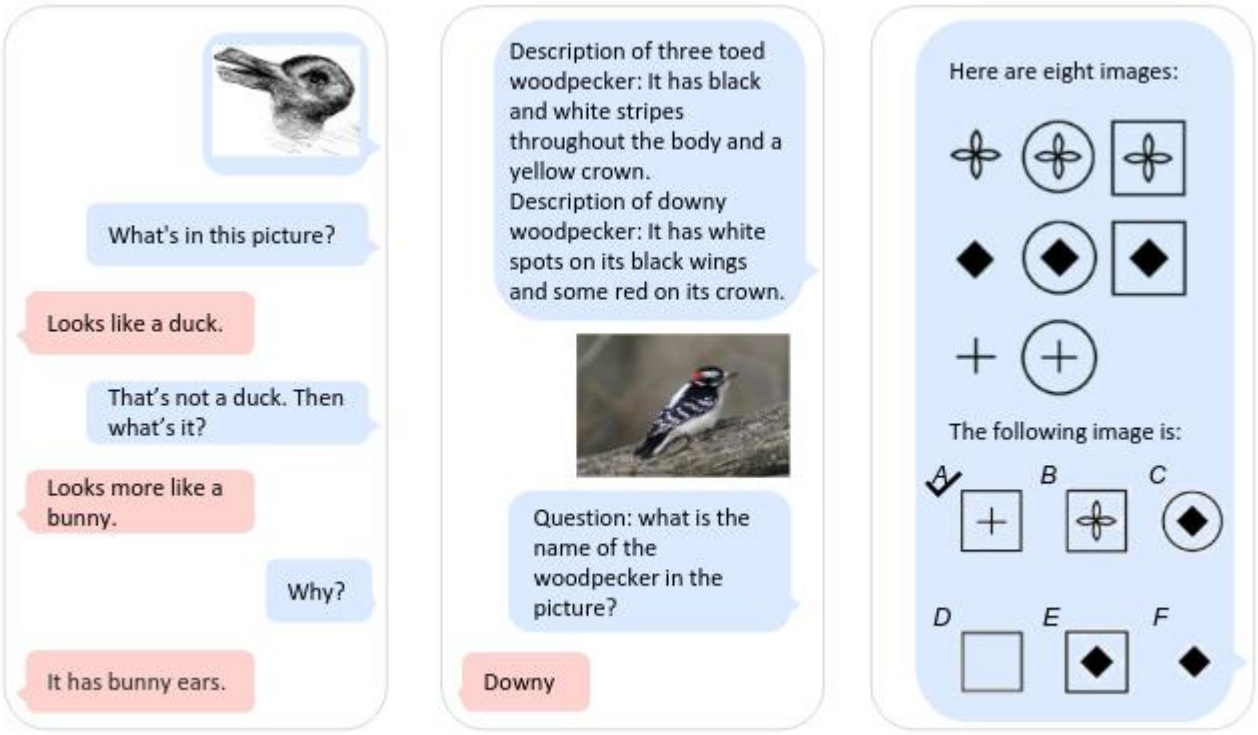

Figure 50 KOSMOS-1 is a multimodal large language model (MLLM) that is capable of perceiving multimodal input, following instructions, and performing in-context learning for not only language tasks but also multimodal tasks.

GILL [Koh, 2024] adapt a pretrained autoregressive language model to handle both text and image inputs and outputs efficiently. By keeping most of the model's weights frozen and fine-tuning a small number of parameters on image-caption data, the model learns to process interleaved image and text content. Special [IMG] tokens are introduced to the vocabulary, enabling the model to produce image outputs either through retrieval or generation. The GILLMapper, a lightweight encoder-decoder transformer, maps [IMG] token hidden states into a space suitable for image generation. For image retrieval, linear mappings translate embeddings between text and visual spaces, trained with an InfoNCE loss. Finally, a classifier, trained on human-annotated data, decides whether to retrieve or generate an image for each prompt, optimizing the model's performance on image and text tasks.

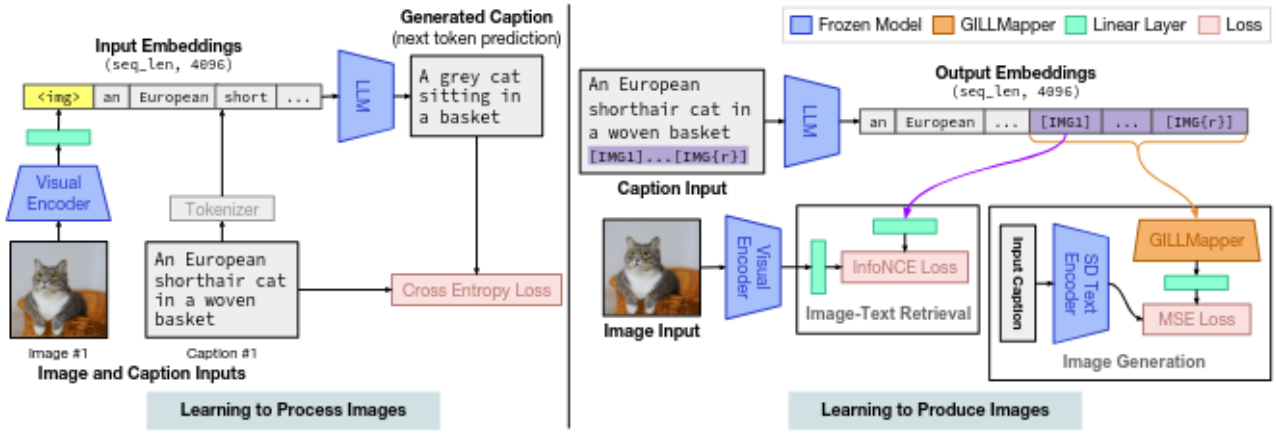

Figure 51 GILL model architecture overview

FILM [Zhao, 2024b] is designed to produce a fused image from two source images, such as infrared-visible, medical, multi-exposure, or multi-focus images. The process involves three main components. First, in the Text Feature Fusion stage, the algorithm generates textual descriptions of the images using methods like image captioning, dense captioning, and segmentation, which are then encoded into text features using BLIP2. These text features are fused into a unified representation. In the Language-Guided Vision Feature Fusion stage, the fused text features guide the extraction and refinement of visual features from the source images through a cross-attention mechanism, ensuring that the salient aspects of the images are integrated effectively. Finally, in the Vision Feature Decoding stage, these refined visual features are decoded back into a coherent fused image. This multi-stage approach enhances the fusion process by leveraging both textual and visual information, resulting in a more informative and cohesive output.

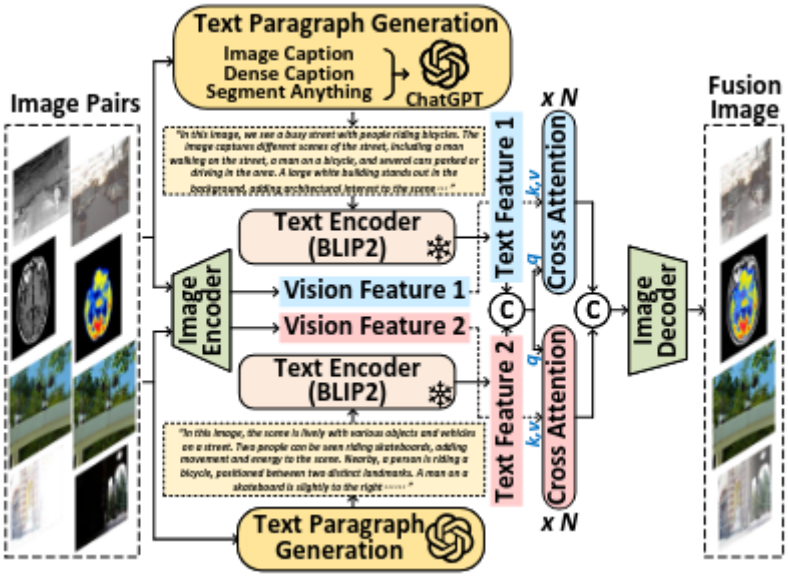

Figure 52 FILM architecture

Chameleon [Team, 2024] is an advanced multi-modal language model that excels in both understanding and generating responses involving text and images, thanks to its use of early fusion techniques. This method integrates textual and visual data at the initial stages of processing, allowing the model to create a unified embedding space where both modalities can interact seamlessly. Chameleon employs multi-modal attention mechanisms and cross-modal encoders, which enable it to learn and leverage the intricate relationships between text and images effectively. Fine-tuning on multi-modal tasks further enhances its capabilities, ensuring high performance in real-world applications that demand cohesive and contextually appropriate mixed-modal outputs. As a result, Chameleon stands out in its ability to handle diverse tasks such as designing layouts, generating visual content, and responding to complex queries that involve both descriptive and visual elements.

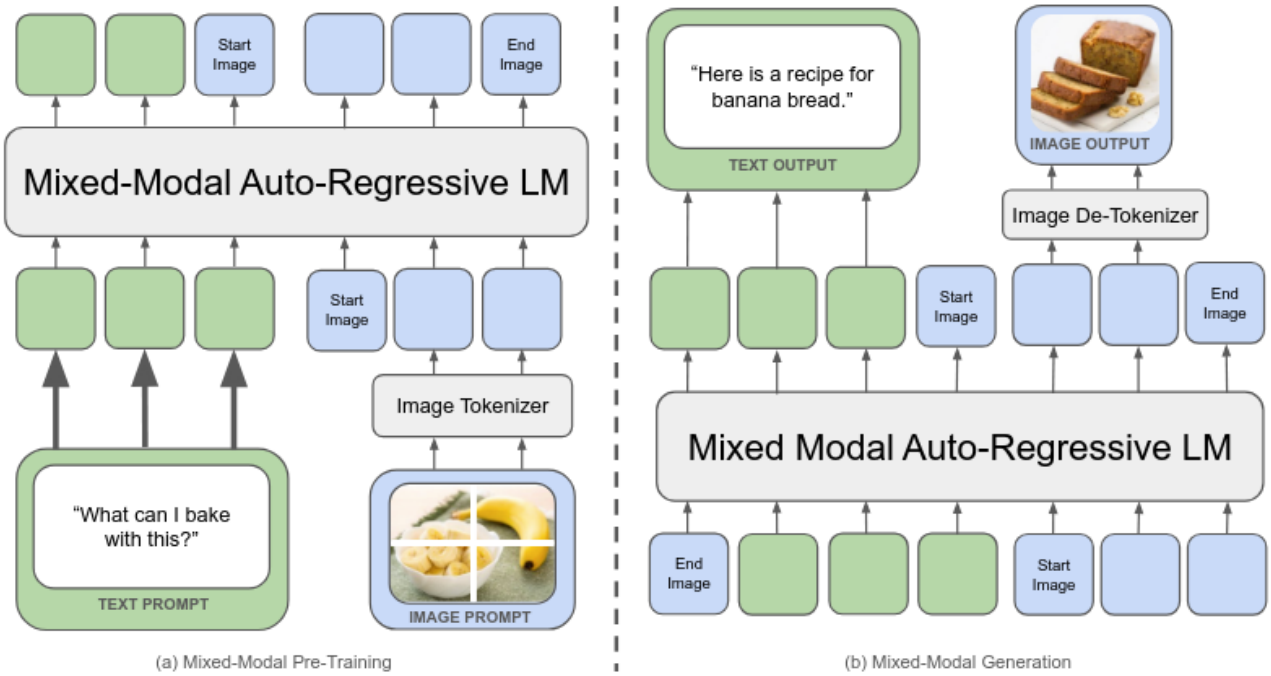

Figure 53 Chameleon network architecture

AnyRef [He, 2024] is a framework designed to process multi-modal inputs—images, audio, and text—and generate detailed textual responses and precise pixel-level perceptions. It integrates a vision encoder (ViT-L/14 from CLIP), a large language model (LLaMA-7B), multi-modal projection layers, and a mask decoder. The framework introduces a special <obj> token to facilitate instance segmentation tasks. Multi-modal prompts are converted into unified referring representations, enabling the model to handle them like text. A refocusing mechanism enhances the <obj> token embeddings with grounded text, improving mask quality. The model is trained end-to-end using a combination of text and mask losses. Training employs pre-trained encoders and fine-tunes the LLM with LoRA, optimizing for efficiency and precision. This allows AnyRef to adeptly generate grounded textual and visual outputs from diverse inputs.

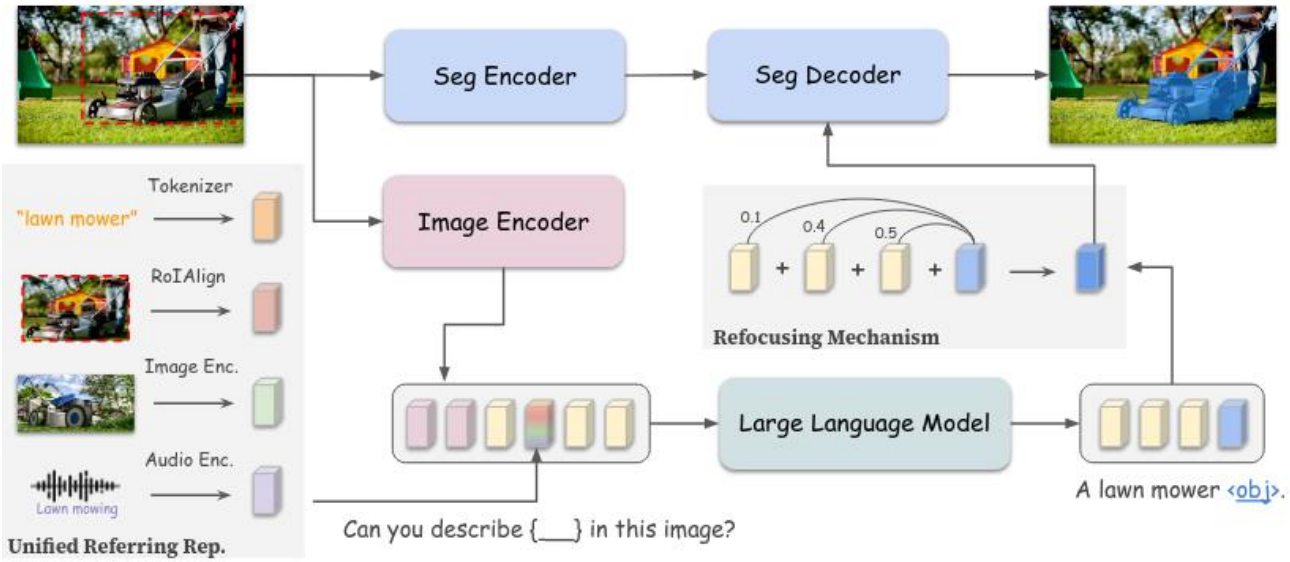

Figure 54 Overall pipeline of AnyRef

5.1.4 NoCorr: Fusion of audio and visual content – fake/fraud detection

With the advancement of technology and artificial intelligence, it has become quite simple to create convincing fake content, be it fake news, videos, audios, etc. Some of such content, namely the audios, videos, or images that have been generated using AI, is referred to as deepfakes, which is a portmanteau of deep learning and fake [Goodfellow et al., 2014, as cited in [Liu et al., 2023]]. Malicious deepfakes are usually videos or audios of people doing or saying compromising, taboo, or illegal things. The Inter-Parliamentary Union, a global organization of national parliaments, warned about the dangers of deepfakes for politicians earlier this year (https://www.ipu.org/news/news-in-brief/202402/dangers-deepfakes-parliamentarians). One of the ideas for successful detection of deepfake videos is to jointly analyze both visual and audio segments of the suspicious content and use information stemming from both when classifying something as deepfake or real content.

[Mittal et al., 2020] combine audio and visual information and search for discrepancies in affective cues of both in order to detect possible fraudulent content. In their research they combine findings from uni- and multi-modal deepfake detection, neuroscience, psychology, and other branches of science. Features are extracted using pre-trained models, and then sent to the model made of convolutional, max-pooling, fully-connected, and normalization layers. Training is done similarly to a Siamese network using two triplet losses (one between modalities and the other between affective cues). At inference time, an input is determined as fake if the difference of the distance between modalities and the distance between perceived emotion are higher than a threshold (calculated during training). The models are tested on two datasets and achieve state-of-the-art results on one and second/third place on the other.

[Zhou and Lim, 2021] propose a framework for joint audio-visual deepfake detection. They base their approach on the synchrony between lip movements and sound, aiming to detect when either or both were corrupted. They compare their main approach (two-plus-one streams) with two other experimental approaches: independently trained networks for both modalities, and late fusion, in which the features of both modalities are aggregated just before the classification head. Their two-plus-one stream, which fuses the audio and visual inputs at every layer and utilizes intra- and inter-attention, achieved the best results in all metrics in two out of the four tested categories, and competitive results in the remaining two.

[Haliassos et al., 2022] use self-supervised (teacher-student) multi-modal (audiovisual) training to learn temporally-dense visual representations, making use of the modalities' intertwinement. Those representations are then used to train a deepfake detector. Their approach, RealForensics, is tested on unseen types of video manipulation, where it achieves state-of-the-art or second-best results. When tested on unseen datasets, it achieves state-of-the-art on all.

[Xue et al., 2023] propose a voice anti-spoofing framework called GACMNet, which extracts features from both visual and audio input, performs feature fusion, and then classifies the audio inputs as fraudulent or bona fide. They claim to achieve around 10% improvements on both the LA and PA tracks of the ASVspoof-2019 dataset [Wang et al., 2019].

[Wang et al., 2024] use two transformers-based [Vaswani et al., 2017] encoders to extract audio and visual features from the input(s), and then these features are passed on to a dynamic weight fusion module (DWF), which outputs weight features for each of the modalities. The weight features are then multiplied with the audio and visual features, and once they are concatenated a decision is made. This AVT2-DWF approach achieves state-of-the-art accuracy on all of the datasets it was tested on, and SOTA AUC in two out of four.

[Sree Katamneni and Rattani, 2024] train an audio-visual framework for the task of deepfakes detection and localization called MMMS-BA (Multi-modal Multi-sequence - Bi-modal Attention). This approach is based on recurrent neural networks (RNNs) and, as its name indicates, makes use of multiple attention mechanisms. Compared with recent work on the subject(s), MMMS-BA achieves state-of-the-art performance on deepfake detection and state-of-the-art or comparable results on deepfake localization.

5.1.5 Analysis of results

The body of economic literature focused on developing and validating objective measures of corruption is expanding. A common method for assessing corruption risk involves creating indices based on "red flags" observed during the procurement process. These red flags act as indicators of possible fraud, corruption, or unethical behavior and can emerge at various stages, from the planning phase to the execution of contracts. However, while the red flag approach shows promise, it has a notable limitation: the subjectivity in selecting and combining these indicators. Furthermore, these indicators may be influenced by contextual factors such as the legal framework, institutional capacity, or the overall transparency of the procurement system, which adds another layer of complexity to their application. The issue of validating corruption risk indicators is seldom addressed in the academic literature on corruption measurement, with a few notable exceptions.

Decarolis and Giorgiantonio (2022) contribute to this field by analyzing corruption risks in public tenders using standardized machine learning tools. One of the most intriguing findings is that many widely recognized red flags either have no significant correlation with corruption or are even negatively correlated with it. This challenges conventional wisdom and underscores the need for more robust, evidence-based criteria in corruption measurement. Gnaldi and Del Sarto (2023) also evaluate the validity of red flag indicators, using a procedure based on multidimensional Item Response Theory (IRT) models. Their findings reveal a multidimensional structure of red flags, with sub-groups that assess different aspects of corruption risk. These sub-groups vary in nature, type, and severity, and are generally non-overlapping, indicating that corruption risk is far more complex and multi-faceted than previously assumed.

5.1.6 YouControl

YouControl (YC) is a Ukrainian team that creates services for business analysis. YouControl helps businesses avoid financial risks and give journalists and public activists a chance to investigate socially important issues, PEPs, Russian, and Belarusian footprints data, corruption risks and corruption history data, etc. We collect, harmonize, and update up to 200+ data sources within YouControl Ukraine (youcontrol.com.ua) and up to 60+ data sources in YouControl.World (youcontrol.world).

YouControl is an eponymous analytical system developed by the YC team that generates a full profile for every company in Ukraine based on open data, tracks changes in state registers, and reveals links between affiliates. The unique technology allows you to get relevant (at the time of the request) information about the company or the individual entrepreneur from more than 220 official sources. In addition to real-time data, the system also provides access to historical information.

Through the “People Control” module, YouControl allows users to identify risks related to individuals, such as court decisions, sanctions, invalid documents, affiliation with PEPs, missing persons, individuals on wanted lists, terrorists, debtors, corrupt officials, and individuals subject to lustration. It can also reveal whether a person is involved in media investigations. The monitoring feature provides daily updates on changes based on data from official registries.

YouControl uses a range of algorithms to harmonize and process data from various governmental registries and online publications. YouControl processes data by receiving inputs in the form of company names, tax identification numbers, or other identifiers. The system then responds with a detailed report on the searched entity that includes connections to politically exposed persons (PEPs) and risk factors. For individuals, the system offers a comprehensive risk assessment, including links to court records, sanctions, and other legal issues. Furthermore, YouControl's API integration enables businesses to automate most tasks related to verifying counterparties and incorporating open data into their IT systems.

Due to Russia’s full-scale aggression, most valuable data registries became unavailable as datasets, files, websites, etc., and are now available via paid APIs with a pay per-request business model only. As the YC team, we are ready to commit limited access to the production API and provide all schemas for data registries and documentation.

5.1.7 Digital footprint detector

Social networks generate vast amounts of data that reflect public opinions. Leveraging Natural Language Processing (NLP), sentiment analysis extracts and interprets these sentiments from user-generated content, offering insights crucial for transparency. Algorithms such as machine learning, deep learning, and sentiment analysis, which deciphers complex linguistic patterns to provide a nuanced understanding of public sentiment, can turn large amounts of social media data into actionable information, such as indications of financial irregularities or hidden connections to Russia.

[Gunasekaran, 2023] conducted a comprehensive review of sentiment analysis (SA) techniques within the field of Natural Language Processing (NLP). This research was particularly focused on how these techniques can be employed to extract and analyze sentiments from various textual data sources, such as social media posts, news articles, and financial reports.

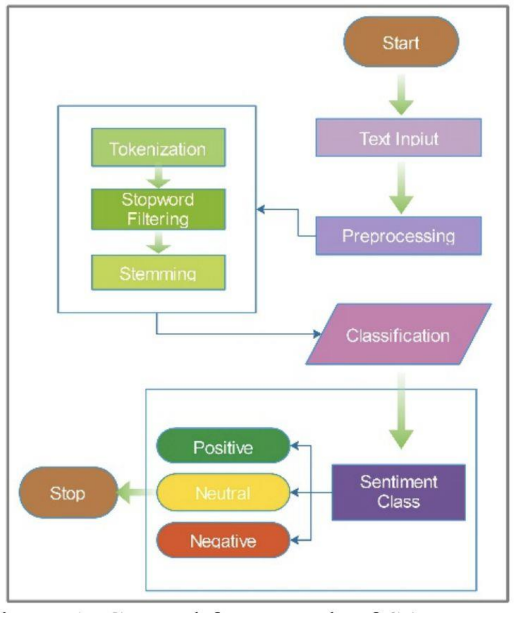

Figure 55 General framework of SA process.

Figure 1 presents the general framework of the SA process, which begins with inputting text data followed by three essential pre-processing steps: tokenization, stopword filtering, and stemming. The final step involves selecting a classification method that is crucial for understanding the overall sentiment of the text as well as the sentiment related to specific entities mentioned within it.

The purpose of SA is to comprehend both the entire sentiment of a textual information and the sentiment regarding individual features or entities referenced in the text. The terms feeling, view, opinion, and belief are used interchangeably, there are distinctions between them. Opinion is the conclusion which is subject to dispute (because of diverse expert perspectives), view is a personal opinion by any individual, belief refers to conscious acceptance and intellectual assent whereas sentiment is a view that expresses one's sentiments.

Sentiment analysis may be used on a variety of textual data sources, including social media, buyer reviews, news stories, and product descriptions.

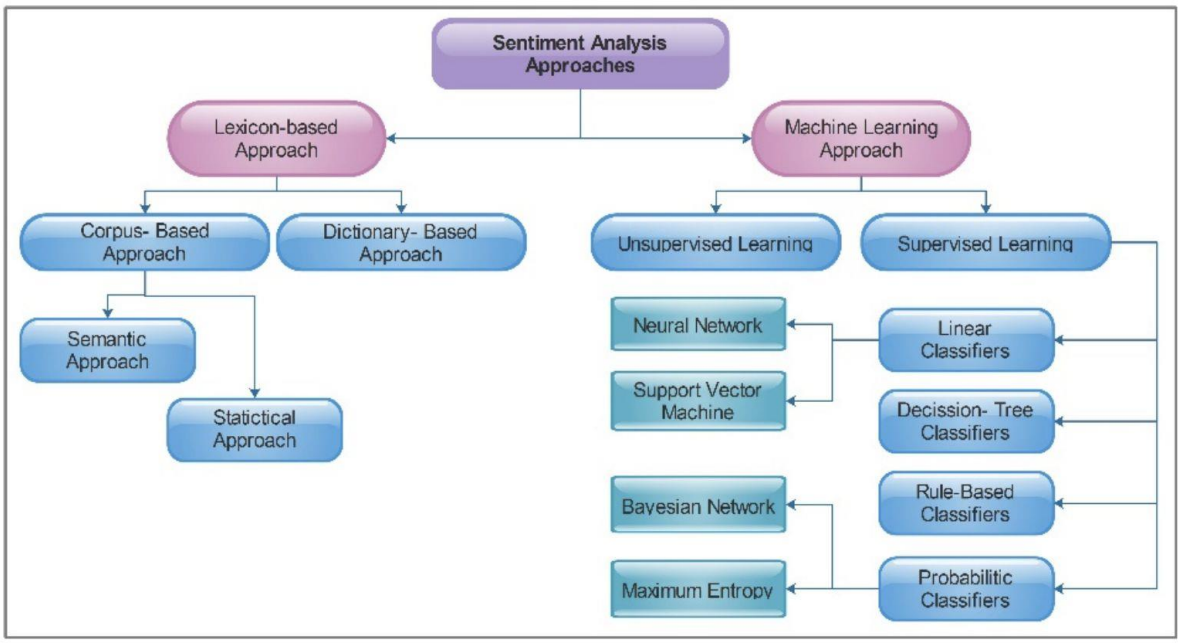

Figure 56 Widely used Sentiment analysis approaches.

Figure 2 showcases the two widely used approaches for sentiment analysis on Twitter data: the Lexicon-Based Approach and the Machine Learning Approach. The Lexicon-Based Approach involves using predefined lists of words with associated sentiment scores to analyze the sentiment of a text. This method is advantageous due to its ease of deployment and minimal training data requirements, although it may lack accuracy in complex contexts. On the other hand, the Machine Learning Approach, which includes supervised, unsupervised, and deep learning techniques, offers a more sophisticated analysis by learning from vast amounts of data. The study emphasizes that while lexicon-based methods are useful for monitoring social media and political analysis, machine learning approaches are more robust for detecting patterns.

[Tiwari et al., 2023] conducted a systematic review of sentiment analysis (SA) on social networks, with a particular focus on comparing ensemble-based techniques such as bagging and boosting. The study provided a comprehensive overview of various SA approaches, including lexicon-based, machine learning-based, graph-based, ensemble, and hybrid approaches.

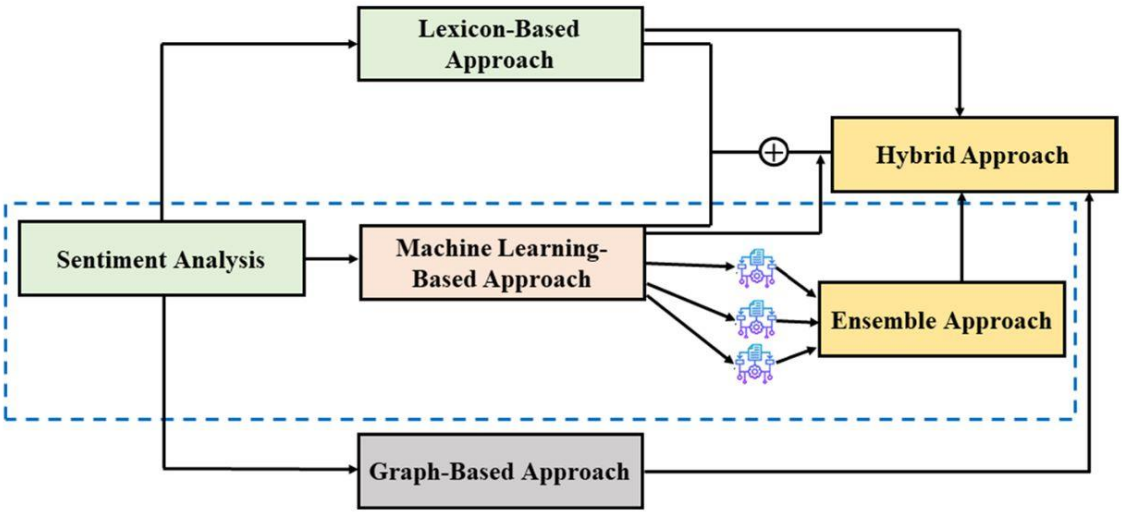

Figure 57 Classification of proposed SA approaches

In the lexicon-based approach, phrases and opinions are analyzed without prior knowledge of labels, using either a dictionary-based method, which judges sentiment based on predefined phrases, or a corpus-based method, which extracts context from the text. Machine learning-based approaches, which typically offer higher accuracy, involve using labeled data for supervised learning or discovering patterns in unlabeled data through unsupervised learning. The graph-based approach connects interrelated words in text reviews to calculate sentiment, representing these relationships as nodes and vertices in a graph. Ensemble learning combines multiple models to enhance classification accuracy and reduce errors, while the hybrid approach leverages the strengths of various techniques, such as rule-based, lexicon-based, and machine learning methods, to optimize SA outcomes. The study's comparative analysis revealed that bagging-based ensemble techniques generally outperformed boosting methods in terms of accuracy and efficiency, making them particularly effective for processing the vast and unstructured content found in social networks

5.2 Next steps on data mining and correlation

5.2.1 CorrelaX

CorrelaX will analyse AI enriched-knowledge graph providing correlation measures between different nodes considering the topological information of the network graph structure. It will be employed in all three pilots as final output aggregator

Key features of CorrelaX will include:

- **Anomaly detection:**
  CorrelaX will identify anomalously behaving nodes (companies) by learning normal patterns from historical data, such as transaction records. Using advanced data analysis techniques and machine learning algorithms, the system will develop models that represent typical company behavior, focusing on key features like transaction frequency, amounts, relationships between entities, and other financial indicators. Once trained on normal data, CorrelaX will be capable of detecting significant deviations from expected patterns. These anomalies may indicate suspicious activities, such as fraud.

- **Clustering and node correlation:**

Based on behavior patterns, CorrelaX will have the capability to correlate nodes (companies) that exhibit similar behavior, allowing it to establish clusters of homogeneous activity. These clusters of similar behavior will provide valuable insights into the relationships between companies and the underlying dynamics of their transactions. CorrelaX can highlight groups that follow typical or expected behavior, as well as those that deviate from established norms, potentially identifying networks of companies engaged in similar activities. This clustering will enhance the ability to detect broader trends, anomalies, and potential risks across a network of entities.

- **Aggregation functions for risk assessment and result aggregations:**
  Given the relationships between nodes in the graph, CorrelaX will study possible influences by analyzing their interactions and connections. The system will extract graph information, such as, for example, graph invariants, providing valuable topological information. These invariants, such as node degree, clustering coefficient, and shortest paths, can reveal fundamental insights into the structure and behaviour of the network.
  By focusing on that information, CorrelaX will identify direct relationships and influences between nodes or entities. This analysis will help uncover patterns of influence, where certain nodes may have a stronger impact on others, or where groups of entities might act in coordinated ways.

5.2.2 FraudAtor (CNT)

FraudAtor will detect fraud and suspicious activities in bank transactions through data analysis and pattern recognition techniques.

Due to constraints on accessing real transaction data during development and training phases, we will rely on synthetic or publicly available data to build and train our models. Consequently, the performance of the tool will depend on the quality and representativeness of the synthetic or public data used. During the testing phase, we will have access to real transaction data provided by MNZ, allowing us to validate the tool's performance.

The key features of FraudAtor will include:

- **Anomaly Detection**: FraudAtor will learn normal transaction behavior from the data and thus flag transactions that deviate significantly from the normality.
- **Pattern Matching and Mining**: FraudAtor will detect specific transaction patterns that are indicative of fraudulent activities. These include circular transaction patterns, where money moves in a loop through several accounts, and wedge patterns, where transactions split and merge in ways that may indicate money laundering schemes.
- **Network Analysis**: FraudAtor will incorporate network analysis techniques to identify relationships and connections between different entities involved in transactions. This will help in detecting networks of accounts that may be working together to execute fraudulent schemes.

Use in CEDAR

FraudAtor will play a crucial role in CEDAR's mission to combat crime and corruption in public procurement through two specific use cases:

- **Pilot 2 UC7** (*Fraudulent Activities of High-Risk Legal Entities*): FraudAtor can help identify connections between companies that have made bids for tenders and high-risk entities. By analyzing transaction patterns and identifying suspicious activities, it can aid in uncovering potential collusion and fraudulent behavior among bidding companies.
- **Pilot 3 UC1** (*Lack of Automation in Compliance, Anti-Fraud, and Anti-Corruption Monitoring*): FraudAtor can address the challenge of automating the monitoring of compliance and detecting fraudulent activities. Through its anomaly detection capabilities, FraudAtor can streamline the process of identifying suspicious transactions and support ongoing anti-fraud and anti-corruption efforts.

5.2.3 Fusion of text and visual content (CERTH)

In this section, a literature review has been performed, presenting the SoTA in the field of text and visual content fusion. Next steps will include a study focused more on the multimodal fusion of LLMs and the selection of the language model that can best adapt to the project's needs.

The tool will be utilized in all three pilots. It will be used in the Italian pilot to handle issues, including suspicious events, emergency procedures, and issue prevention. In the Slovenian pilot is going to deal with the inability to detect hints of fraudulent activities and in the Ukranian pilot with the lack of automation in compliance, anti-fraud and anti-corruption monitoring.

5.2.4 NoCorr: Fusion of audio and visualcontent – fake/fraud detection (TRE)

The fake/fraud detector utilizing the fusion of audio and visual content, named NoCorr, will be able to classify both visual and audio segments in each input as either fake/fraudulent (deepfakes) or real, trustworthy content.

NoCorr will be based on the state-of-the-art approaches in this domain. It will be trained using open-source, publicly available data. We also might create extra synthetic training data, if deemed necessary.

Use in CEDAR

NoCorr will be used in all three pilots for the following use cases:

- **Pilot-Italy**: UC5 (Suspicious events and results); UC6 (Adequacy and issue prevention); and UC7 (Adoption of emergency procedure in public procurement)
- **Pilot-Slovenia**: UC8 (Inability to Detect Hints of Fraudulent Activities)
- **Pilot-Ukraine**: UC1 (Lack of Automation in Compliance, anti-Fraud and Anti-corruption Monitoring)

An unofficial example of a NoCorr usage can be analyzing social media content of tender participants or candidates.

5.2.5 Analysis of results (SoA) (BIGS)

To validate the findings on corruptive practices, a range of statistical and econometric methods will be employed, carefully selected based on the nature of the data provided by collaborating partners.

- First, the results can be cross-referenced with the corruption index developed in Task 4.3 to identify significant patterns or correlations.
- Second, logistic regression will be used to evaluate the relevance of each red flag in predicting corruption.

These methods must be tailored to the specific characteristics of the data—whether categorical, continuous, or time-series—ensuring that the analysis is both appropriate and effective. Additionally, advanced techniques, such as machine learning algorithms or sensitivity analysis, could be integrated to refine the identification of key indicators and enhance the robustness of the results. By adapting the methodology to the data, we aim to draw statistically sound and practical conclusions for identifying corruption risks.

5.2.6 YouControl (SoA)(YC)

YC will collect, harmonize, and perform data mining in various data sources including data from third-party APIs like Ukrainian and Russian official registries. Those sources include PEPs, Russian, and Belarusian footprints data, corruption risks and corruption history data, negative media mentions, public procurement data, company history, beneficiary info, etc.

In the CEDAR project, YouControl plays a key role in providing its data analytics tools and methodologies to support the transparent management of foreign aid for Ukraine's reconstruction efforts. YC will ensure collection and analysis of

data from governmental registries. Our API will provide clean harmonized data about bidders and results of compliance checks versus various official sources, as well as automated detection and highlighting of predefined risk indicators, helping to identify tenders where fraud or manipulations may occur.

YC will provide an algorithm-based scoring mechanism capable of identifying and highlighting bidders' legal and financial issues, as well as inconsistencies and irregularities in their registration data as compared to the bid submission package.

Use in CEDAR

YouControl will be used in every specific use case of Pilot 3 “Transparent management of foreign aid for rebuilding Ukraine”. Specifically, in:

- **Pilot-Ukraine**: UC1 - Lack of Automation in Compliance, anti-Fraud and Anti-corruption Monitoring
- **Pilot-Ukraine**: UC2 - Legally Incompliant Tender Participants (bidders)
- **Pilot-Ukraine**: UC3 - High-Risk and Low Reputation Tender Participants
- **Pilot-Ukraine**: UC4 - Elimination of Competition Through Bid Coordination and/or Tailored Tenders
- **Pilot-Ukraine**: UC5 - Insufficient Accessibility and Transparency of Anti-fraud and Anti-corruption Monitoring Results

5.2.7 Digital footprint detector (SoA)(ART)

ART will provide a tool for digital footprint analysis using NLP techniques. The tool will analyze unstructured textual data including information about Ukrainian companies from online search results, Ukrainian job portals, and company profiles on social media. ART will utilize algorithms such as machine learning, deep learning, and sentiment analysis, to decipher complex linguistic patterns and provide an understanding of public sentiment. The tool will focus on detecting insights that could indicate suspicious activity, low reputation, possible links to Russia and signs of possible fraud or corruption.

Use in CEDAR

The tool will be tested in the following CEDAR use cases:

- Pilot-Ukraine: **UC1** - Lack of Automation in Compliance, Anti-Fraud, and Anti-Corruption Monitoring,
- Pilot-Ukraine: **UC3** - High-Risk and Low Reputation Tender Participants
- Pilot-Ukraine: **UC4** - Elimination of Competition Through Bid Coordination and/or Tailored Tenders,
- Pilot-Ukraine: **UC5** - Insufficient Accessibility and Transparency of Anti-fraud and Anti-corruption Monitoring Results.

5.2.8 RaphtoryMotifs (SoA)(UPM)

RaphtoryMotifs will develop scalable techniques that can detect temporal graph motifs over Knowledge graph temporal interactions. Over the project multiple temporal motifs will be identified for their application to the collected CEDAR knowledge graphs. The component will implement two categories of algorithms: counting motif algorithms that provide whole graph information on the most significant interaction patterns, and identification of the individual involvement of knowledge graph nodes for each temporal motif.

The component will aim to support efficient computation of 3-message temporal motifs, including triangle and bi-node patterns. Additionally, with the input of the use cases, algorithms for computing attribute-based temporal motif algorithms will be developed for specific pilots where the behavior of key entities needs to be monitored for an identified use case.

Use in CEDAR

As part of the T4.4 Tools RaphtoryMotifs will be used in all 3 Use Cases to enrich the available information regarding the behavior of the Knowledge Graph entities.

6. Conclusion

This deliverable has provided a comprehensive overview of the current state and future directions in the field, focusing on the requirements of the CEDAR project. Each tool provider focused on the part of the literature, concerning its tool, which could optimally cover the project requirements, i.e., extracting insightful information about corruption and identify patterns and abnormalities, carefully concealed beneath large amounts of data.

The tools presented in this deliverable may accept different inputs such as multilingual text, images, videos and audio, or financial and econometric data, also multimodal data. In order to exploit those data, the mining of subtle patterns is required in addition to the linking and combining of the insights into actionable intelligence.

Finally, the next steps section, following the related work section of each task serves as the linking point between the D4.1 and the D4.2, aspiring to present the first coherent results, based on the research methods studied in this deliverable.

List of References

[Ai et al., 2024] Ai, Z., Chen, Z., and Xu, S. (2024). MM-KWS: Multimodal prompts for multilingual user-defined keyword spotting. ArXiv, abs/2406.07310.

[Alayrac, 2022] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35, 23716-23736.

[Asai et al., 2023] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. arXiv Preprint arXiv:2310.11511.

[Auriol, 2016] Auriol, Emmanuelle, Thomas Flochel, and Stephane Straub. 2016. Public Procurement and Rent-Seeking: The Case of Paraguay. World Development

[Babu et al., 2022] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In Proc. Interspeech 2022, pages 2278–2282.

[Baevski et al., 2020] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460.

[Baevski et al., 2022] Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. (2022). data2vec: A general framework for self-supervised learning in speech, vision and language. ArXiv, abs/2202.03555.

[Barnes, 2024] Barnes, M.R., Karan, M., McQuistin, S., Perkins, C., Tyson, G., Purver, M., Castro, I. and Clegg, R.G., 2024, May. Temporal Network Analysis of Email Communication Patterns in a Long Standing Hierarchy. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 126-138).

[Bartsch, 2020] Robert Bartsch. 2020. The relationship of drug and human trafficking and their facilitation via Cryptomarkets and the Dark Web: A recommendation for cryptocurrency regulation. (2020).

[Beaulieu et al., 1997] Beaulieu, M., Gatford, M., Huang, X., Robertson, S., Walker, S., & Williams, P. (1997). Okapi at TREC-5. Nist Special Publication SP, 143–166.

[Bele, 2021] Bele, Julija Lapuh. "Cryptocurrencies as facilitators of cybercrime." SHS Web of Conferences. Vol. 111. EDP Sciences, 2021.

[Berchansky et al., 2023] Berchansky, M., Izsak, P., Caciularu, A., Dagan, I., & Wasserblat, M. (2023). Optimizing retrieval-augmented reader models via token elimination. arXiv Preprint arXiv:2310.13682.

[Berg et al., 2021] Berg, A., O'Connor, M., and Cruz, M. T. (2021). Keyword transformer: A self-attention model for keyword spotting. In Proc. Interspeech 2021, pages 4249–4253.

[Bjørnskov, 2006] Bjørnskov, C. (2006). The multiple facets of social capital. European Journal of Political Economy 22(1): 22–40.

[Blagojević & Damijan, 2013] Blagojević, S., & Damijan, J. P. (2013). The impact of corruption and ownership on the performance of firms in Central and Eastern Europe. Post-communist Economies, 25(2), 133-158.

[Blanchini, 2022] Blanchini, M., Cerreta, M., Di Monda, D., Fabbri, M., Raciti, M., Ahmad, H.S., Costa, G.: Supporting criminal investigations on the blockchain: A temporal logicbased approach (2022).

[Blattmann, 2023] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22563-22575).

[Borgeaud et al., 2022] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., & others. (2022). Improving language models by retrieving from trillions of tokens. International Conference on Machine Learning, 2206–2240.

[Bovbjerg and Tan, 2022] Bovbjerg, H. S. and Tan, Z. (2022). Improving label-deficient keyword spotting through self-supervised pretraining. 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pages 1–5.

[Brill, 2014] Alan Brill and Lonnie Keene. 2014. Cryptocurrencies: The Next Generation of Terrorist Financing? Defence Against Terrorism Review 6 (2014).

[Burgess, 2015] Burgess, R., Jedwab, R., Miguel, E., Morjaria, A., & Padró i Miquel, G. (2015). The value of democracy: evidence from road building in Kenya. American Economic Review, 105(6), 1817-1851.

[Cao, 2018] Cao, Z., Long, M., Wang, J., Yu, P.S.: Hashnet: Deep learning to hash by continuation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5608–5617 (2018)

[CAo, 2022] Cao, Y., Pan, H., Wang, H., Xu, X., Li, Y., Tian, Z., & Zhao, X. (2022, July). Small object detection algorithm for railway scene. In 2022 7th International Conference on Image, Vision and Computing (ICIVC) (pp. 100-105). IEEE.

[Chainalysis, 2024] Chainalysis Inc.: The 2024 Crypto Crime Report (2024).

[Chao et al., 2024] Chao, R.-Y., Cheng, W.-H., Quatra, M. L., Siniscalchi, S. M., Yang, C.-H. H., Fu, S.-W., and Tsao, Y. (2024). An investigation of incorporating Mamba for speech enhancement. ArXiv, abs/2405.06573.

[Chen, 2014] Chen, Z., Dehmer, M., & Shi, Y. (2014). A note on distance-based graph entropies. Entropy, 16(10), 5416-5427.

[Chen, 2020] Chen, Y., Zhang, P., Li, Z., Li, Y., Zhang, X., Qi, L., ... & Jia, J. (2020). Dynamic scale training for object detection. arXiv preprint arXiv:2004.12432.

[Chen, 2022] Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W., ... & Ma, L. (2022, June). Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 1, pp. 230-238).

[Chen et al., 2023] Chen, H., Pasunuru, R., Weston, J., & Celikyilmaz, A. (2023). Walking down the memory maze: Beyond context limit through interactive reading. arXiv Preprint arXiv:2310.05029.

[Chen, 2024] Chen, J., Lv, Z., Wu, S., Lin, K. Q., Song, C., Gao, D., ... & Shou, M. Z. (2024). VideoLLM-online: Online Video Large Language Model for Streaming Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18407-18418).

[Cheng, 2023] Cheng, D., Ye, Y., Xiang, S., Ma, Z., Zhang, Y. and Jiang, C. 2023. Anti-money laundering by group-aware deep graph learning. IEEE Transactions on Knowledge and Data Engineering, 35(12), pp.12444-12457.

[Choi, 2022] Choi, W., Chen, J., & Yoon, J. (2022). Parallel pathway dense video captioning with deformable transformer. IEEE Access, 10, 129899-129910.

[Cong, 2023] Cong, Lin William, et al. An anatomy of crypto-enabled cybercrimes. No. w30834. National Bureau of Economic Research, 2023.

[Connolly, 2019] Lena Y Connolly and David S Wall. 2019. The rise of crypto-ransomware in a changing cybercrime landscape: Taxonomising countermeasures. Computers & Security 87 (2019), 101568.

[Conway, 2011] Conway, D., 2011. Modeling network evolution using graph motifs. arXiv preprint arXiv:1105.0902.

[Covas, 2023] Covas E. (2023). Named Entity Recognition Using GPT for Identifying Comparable Companies. arXiv preprint arXiv:2307.07420v2.

[Coviello, 2017] Coviello, D., & Gagliarducci, S. (2017). Tenure in Office and Public Procurement. American Economic Journal: Economic Policy, 9(3), 59-105.

[Cozzolino, 2022] Davide Cozzolino, Matthias Nießner, and Luisa Verdoliva.: Audio-visual person-of-interest deepfake detection. arXiv preprint arXiv:2204.03083, 2022. 1, 2, 5

[Cui, 2020] Cui, L., Lv, P., Jiang, X., Gao, Z., Zhou, B., Zhang, L., ... & Xu, M. (2020). Context-aware block net for small object detection. IEEE Transactions on cybernetics, 52(4), 2300-2313.

[Dai et al., 2022] Dai, Z., Zhao, V. Y., Ma, J., Luan, Y., Ni, J., Lu, J., Bakalov, A., Guu, K., Hall, K. B., & Chang, M.-W. (2022). Promptagator: Few-shot dense retrieval from 8 examples. arXiv Preprint arXiv:2209.11755.

[Decarolis, 2022] Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: new evidence from Italian calls for tenders. EPJ Data Science, 11(1), 16.

[Dehmer, 2011] Dehmer, M., & Mowshowitz, A. (2011). A history of graph entropy measures. Information Sciences, 181(1), 57-78.

[Devlin, 2018] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv:1810.04805.

[Doan, 2022] Doan, K.D., Yang, P., Li, P.: One loss for quantization: Deep hashing with discrete wasserstein distributional matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9447–9457 (2022)

[Donchev & Ujhelyi, 2014] Donchev, D., & Ujhelyi, G. (2014). What do corruption indices measure? Economics and Politics, 26(2), 309-331.

[Dong, 2020] Dong, F., Zhang, Y., & Nie, X. (2020). Dual discriminator generative adversarial network for video anomaly detection. IEEE Access, 8, 88170-88176.

[Dong, 2022] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, Baining Guo: Protecting Celebrities from DeepFake with Identity Consistency Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9468-9478

[Dong, 2023] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, Zheng Ge: Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3994-4004

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929.

[Dou, 2020] Dou, Yingtong, et al. 2020. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. Proceedings of the 29th ACM international conference on information & knowledge management.

[Dreher and Schneider, 2010] Dreher, A. and Schneider, F. (2010). Corruption and the shadow economy: An empirical analysis. Public Choice 144(1–2): 215–238.

[Eurostat, 2024] Eurostat (2024). Police-recorded offences by offence category - annual data [Dataset]. https://ec.europa.eu/eurostat/databrowser/view/crim_off_cat__custom_12542525/default/table

[Fazekas, 2020] Fazekas, M., & Kocsis, G. (2020). Uncovering high-level corruption: cross-national objective corruption risk indicators using public procurement data. British Journal of Political Science, 50(1), 155-164.

[Flyvbjerg, 2003] Flyvbjerg, B. (2003). Megaprojects and Risk: An Anatomy of Ambition. Cambridge University Press.

[Fouss, 2007] Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. IEEE Transactions on knowledge and data engineering, 19(3), 355-369.

[Ganguly, 2022] Ganguly S., Ganguly A., Mohiuddin S., Malakar S, Sarkar R.: ViXNet: vision transformer with Xception Network for deepfakes based video and image forgery detection. Expert Syst Appl 210:118423 (2022)

[Ganin, 2020] Ganin, A. A., Quach, P., Panwar, M., Collier, Z. A., Keisler, J. M., Marchese, D., & Linkov, I. (2020). Multicriteria decision framework for cybersecurity risk assessment and management. Risk Analysis, 40(1), 183-199.

[Gao et al., 2023] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv Preprint arXiv:2312.10997.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[Glass et al., 2022] Glass, M., Rossiello, G., Chowdhury, M. F. M., Naik, A. R., Cai, P., & Gliozzo, A. (2022). Re2G: Retrieve, rerank, generate. arXiv Preprint arXiv:2207.06300.

[Gnaldi, 2023] Gnaldi, M., & Del Sarto, S. (2023). Validating corruption risk measures: a key step to monitoring SDG progress. Social Indicators Research, 1-27.

[Goel et al., 2023] Goel, C., Koppisetti, S., Colman, B., Shahriyari, A., and Bharaj, G. (2023). Towards attention-based contrastive learning for audio spoof detection. In Proc. INTERSPEECH 2023, pages 2758–2762.

[Golden & Picci, 2005] Golden, M. A., & Picci, L. (2005). Proposal for a new measure of corruption, illustrated with Italian data. Economics and Politics, 17(1), 37-75.

[Gong et al., 2022] Gong, Y., Lai, C.-I., Chung, Y.-A., and Glass, J. (2022). SSAST: Self-supervised audio spectrogram transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10699–10709.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[Golden, 2005] Golden, M. A., & Picci, L. (2005). Proposal for a new measure of corruption, illustrated with Italian data. Economics & Politics, 17(1), 37-75.

[Graeff, 2009] Graeff, P. (2009). Social capital: The dark side. In G. T. Svendsen & G. L. Haase Svendsen (Eds.), Handbook of Social Capital: The Troika of Sociology, Political Science and Economics (pp. 143-161). Cheltenham: Edward Elgar.

[Gu and Dao, 2023] Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. ArXiv, abs/2312.00752.

[Gu, 2023a] Gu, X., Wen, C., Ye, W., Song, J., & Gao, Y. Seer: Language Instructed Video Prediction with Latent Diffusion Models. In The Twelfth International Conference on Learning Representations.

[Gunasekaran, 2023] Gunasekaran, K. P. (2023). Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review. College of Information and Computer Sciences, University of Massachusetts, Amherst, United States.

[Hajdu, 2017] Hajdu, I., & Miklós, J. (2017). Intensity of Competition, Corruption Risks and Price Distortion in the Hungarian Public Procurement–2009-2016.

[Haliassos et al., 2022] Haliassos, A., Mira, R., Petridis, S., and Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14950–14962.

[Hassan, 2022] Hassan, M.U., Niu, D., Zhang, M., Zhao, X.: Asymmetric hashing based on generative adversarial network. Multimedia Tools and Applications 82, 389–405 (2023) https://doi.org/10.1007/s11042-022-13141-2

[He, 2024] He, J., Wang, Y., Wang, L., Lu, H., He, J. Y., Lan, J. P., ... & Xie, X. (2024). Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13980-13990).

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851.

[Hofstätter et al., 2023] Hofstätter, S., Chen, J., Raman, K., & Zamani, H. (2023). Fid-light: Efficient and effective retrieval-augmented text generation. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1437–1447.

[Hornuf, 2023] Hornuf, L., Momtaz, P.P., Nam, R.J., Yuan, Y.: Cybercrime on the Ethereum blockchain (2023)

[Hu, 2023] Hu, X., Huang, Z., Huang, A., Xu, J., & Zhou, S. (2023). A dynamic multi-scale voxel flow network for video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6121-6131).

[Huang, 2016] Huang, R., Pedoeem, J., & Chen, C. (2018, December). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE international conference on big data (big data) (pp. 2503-2510). IEEE.

[Huang et al., 2023] Huang, W., Lapata, M., Vougiouklis, P., Papasarantopoulos, N., & Pan, J. (2023). Retrieval augmented generation with rich answer encoding. Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 1012–1025.

[Huang, 2024] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., ... & Wei, F. (2024). Language is not all you need: Aligning perception with language models. Advances in Neural Information Processing Systems, 36.

[Huang, Huang, 2024] Huang, Y., & Huang, J. (2024). A Survey on Retrieval-Augmented Text Generation for Large Language Models. arXiv Preprint arXiv:2404.10981.

[Huber, 2023] Huber, M., Luu, A. T., Terhörst, P., & Damer, N. (2023). Efficient Explainable Face Verification based on Similarity Score Argument Backpropagation. arXiv preprint arXiv:2304.13409v1.

[Jazayeri, 2020]]Jazayeri, A. and Yang, C.C., 2020. Motif discovery algorithms in static and temporal networks: A survey. Journal of Complex Networks, 8(4)

[Jiang, 2022] Jiang, Jiaxin, et al. 2022. Spade: A real-time fraud detection framework on evolving graphs. Proceedings of the VLDB Endowment 16.3 (2022): 461-469.

[Ilin, 2023] Ilin, I. (2023). Advanced rag techniques: an illustrated overview. https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6

[Ilyas, 2023] Hafsa Ilyas, Ali Javed, Khalid Mahmood Malik: AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing, Volume 136, 2023

[Islam, 2023] Islam, M., Dukyil, A. S., Alyahya, S., & Habib, S. (2023). An IoT enable anomaly detection system for smart city surveillance. Sensors, 23(4), 2358.

[Izacard, Grave, 2020] Izacard, G., & Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering. arXiv Preprint arXiv:2007.01282.

[Izacard et al., 2023] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1–43.

[Jiang et al., 2023] Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active retrieval augmented generation. arXiv Preprint arXiv:2305.06983.

[Jin, 2024] Jin, Y., Li, J., Liu, Y., Gu, T., Wu, K., Jiang, Z., ... & Ma, L. (2024). Efficient multimodal large language models: A survey. arXiv preprint arXiv:2405.10739.

[Jukna, 2006] Jukna, S. (2006). On graph complexity. Combinatorics, Probability and Computing, 15(6), 855-876.

[Jung et al., 2021] Jung, J., Heo, H.-S., Tak, H., Shim, H., Chung, J. S., Lee, B.-J., Yu, H., and Evans, N. W. D. (2021). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6367–6371.

[Kalinin, 2021] Kalinin, M., Krundyshev, V., & Zegzhda, P. (2021). Cybersecurity risk assessment in smart city infrastructures. Machines, 9(4), 78.

[Kamoona, 2023] Kamoona, A. M., Gostar, A. K., Bab-Hadiashar, A., & Hoseinnezhad, R. (2023). Multiple instance-based video anomaly detection using deep temporal encoding–decoding. Expert Systems with Applications, 214, 119079.

[Kandpal et al., 2023] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. International Conference on Machine Learning, 15696–15707.

[Kang et al., 2023] Kang, M., Kwak, J. M., Baek, J., & Hwang, S. J. (2023). Knowledge graph-augmented language models for knowledge-grounded dialogue generation. arXiv Preprint arXiv:2305.18846.

[Kapoor, 2021] Kapoor, R., Sharma, D., Gulati, T.: State of the art content based image retrieval techniques using deep learning: A survey. Multimedia Tools and Applications 80, 29561–29583 (2021) https://doi.org/10.1007/s11042-021-11045-1

[Khattab et al., 2022] Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. arXiv Preprint arXiv:2212.14024.

[Khattab et al., 2022] Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2022). Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. arXiv Preprint arXiv:2212.14024.

[Klašnja, 2015] Klašnja, M. (2015). Corruption and the incumbency disadvantage: Theory and evidence. The Journal of Politics, 77(4), 928-942.

[Khosla et al., 2020] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673.

[Kim, 2020] Kim, B. K., Kang, H. S., Lee, S., & Park, S. O. (2020). Improved drone classification using polarimetric merged-Doppler images. IEEE Geoscience and Remote Sensing Letters, 18(11), 1946-1950.

[King, 2018] King, Z. M., Henshel, D. S., Flora, L., Cains, M. G., Hoffman, B., & Sample, C. (2018). Characterizing and measuring maliciousness for cybersecurity risk assessment. Frontiers in psychology, 9, 39.

[Knoche, 2023] Knoche, M., Teepe, T., Hörmann, S., & Rigoll, G. (2023). Explainable Model-Agnostic Similarity and Confidence in Face Verification. Winter Conference on Applications of Computer Vision Workshop 2023, DOI: https://doi.org/10.1109/WACVW58289.2023.00078.

[Koch et al., 2015] Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In ICML deep learning workshop, volume 2, pages 1–30. Lille.

[Koh, 2024] Koh, J. Y., Fried, D., & Salakhutdinov, R. R. (2024). Generating images with multimodal language models. Advances in Neural Information Processing Systems, 36.

[Kolagati, 2022] Santosh Kolagati, Thenuga Priyadharshini, V. Mary Anita Rajam: Exposing deepfakes using a deep multilayer perceptron–convolutional neural network model. International Journal of Information Management Data Insights, 2 (1) (2022), p. 100054

[Kolbaek et al., 2017] Kolbaek, M., Yu, D., Tan, Z., and Jensen, J. H. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25:1901–1913.

[Kong et al., 2020] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2020). DiffWave: A versatile diffusion model for audio synthesis. ArXiv, abs/2009.09761.

[Kovanen, 2013] Kovanen, L., Kaski, K., Kertész, J. and Saramäki, J., 2013. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proceedings of the National Academy of Sciences, 110(45), pp.18070-18075.

[Kumar, 2019] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In KDD.

[Lai, 2015] Lai, H., Pan, Y., Liu, Y., Yan, S.: Simultaneous feature learning and hash coding with deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3270–3278 (2015)

[Larsson & Grimes, 2023] Larsson, F., & Grimes, M. (2023). Societal accountability and grand corruption: How institutions shape citizens' efforts to shape institutions. Political Studies, 71(4), 1321-1346.

[Lewis et al., 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[Li, 2023] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (pp. 19730-19742). PMLR.

[Li et al., 2023] Li, X., Zhao, R., Chia, Y. K., Ding, B., Joty, S., Poria, S., & Bing, L. (2023). Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. arXiv Preprint arXiv:2305.13269.

[Li et al., 2024] Li, H., Yang, B., Xi, Y., Yu, L., Tan, T., Li, H., and Yu, K. (2024). Text-aware speech separation for multi-talker keyword spotting.

[Lisciandra, 2022] Lisciandra, M., Milani, R., & Millemaci, E. (2022). A corruption risk indicator for public procurement. European Journal of Political Economy, 73, 102141.

[Liu, 2016] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14 (pp. 21-37). Springer International Publishing.

[Liu, 2021] Yixin Liu, Shirui Pan, Yu Guang Wang, Fei Xiong, Liang Wang, Qingfeng Chen, and Vincent CS Lee. 2021. Anomaly detection in dynamic graphs via transformer. IEEE Transactions on Knowledge and Data Engineering (2021).

[Liu, 2021a] Liu, Yang, et al. 2021. Pick and choose: a GNN-based imbalanced learning approach for fraud detection. Proceedings of the web conference 2021.

[Liu et al., 2023a] Liu, L., Guan, H., Ma, J., Dai, W., Wang, G., and Ding, S. (2023). A mask free neural network for monaural speech enhancement. In Proc. INTERSPEECH 2023, pages 2468–2472.

[Liu et al., 2023b] Liu, R., Zhang, J., Gao, G., and Li, H. (2023). Betray one-self: A novel audio deepfake detection model via mono-to-stereo conversion. In Proc. INTERSPEECH 2023, pages 3999–4003.

[Liu et al., 2023c] Liu, X., Sahidullah, M., Lee, K. A., and Kinnunen, T. (2023). Speaker-aware anti-spoofing. In Proc. INTERSPEECH 2023, pages 2498–2502.

[Liu, 2022] Liu, P., Masuda, N., Kito, T. and Sarıyüce, A.E., 2022. Temporal motifs in patent opposition and collaboration networks. Scientific reports, 12(1), p.1917.

[Liu, 2022a] Liu, Y., Liu, J., Lin, J., Zhao, M., & Song, L. (2022). Appearance-motion united auto-encoder framework for video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(5), 2498-2502.

[Liu, 2022b] Liu, Y., Liu, J., Zhao, M., Yang, D., Zhu, X., & Song, L. (2022, July). Learning appearance-motion normality for video anomaly detection. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[Liu, 2023b] Liu, P., Acharyya, R., Tillman, R.E., Kimura, S., Masuda, N. and Sarıyüce, A.E., 2023. Temporal motifs for financial networks: A study on mercari, jpmc, and venmo platforms. arXiv preprint arXiv:2301.07791.

[Liu, 2024] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. Advances in neural information processing systems, 36.

[López & Santos, 2014] López, J. A. P., & Santos, J. M. S. (2014). Does corruption have social roots? The role of culture and social capital. Journal of Business Ethics, 122, 697-708.

[Lu et al., 2023] Lu, Y.-X., Ai, Y., and Ling, Z. (2023). Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement. ArXiv, abs/2308.08926.

[Lutati et al., 2023] Lutati, S., Nachmani, E., and Wolf, L. (2023). Separate and diffuse: Using a pretrained diffusion model for improving source separation. ArXiv, abs/2301.10752.

[Ma et al., 2023] Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. arXiv Preprint arXiv:2305.14283.

[Mahdavi, 2020] Mahdavi, F., & Rajabi, R. (2020, December). Drone detection using convolutional neural networks. In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) (pp. 1-5). IEEE.

[Mauro, 1995] Mauro, P. (1995) “Corruption and growth”, Quarterly Journal of Economics, Vol. 110, No. 3, pp. 681-712.

[Mauro, 1997] Mauro, P. (1997) "The Effects of Corruption on Growth, Investment, and Government Expenditure: A Cross–Country Analysis", Corruption and the Global Economy, pp. 83-107.

[Mauro, 1998] Mauro, P. (1998) “Corruption and the composition of government expenditure”, Journal of Public Economics, Vol. 69, No. 2, pp. 263-279.

[Medaiyese, 2022] Medaiyese, O. O., Ezuma, M., Lauf, A. P., & Guvenc, I. (2022). Wavelet transform analytics for RF-based UAV detection and identification system using machine learning. Pervasive and Mobile Computing, 82, 101569.

[Milbich, 2020] Milbich, T., Roth, K., Bharadhwaj, H., Sinha, S., Bengio, Y., Ommer, B., Cohen, J.P.: Diva: Diverse visual feature aggregation for deep metric learning. CoRR (2020) https://doi.org/10.48550/arXiv.2004.13458

[Mittal et al., 2020] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emotions don’t lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM international conference on multimedia, pages 2823–2832.

[Mork et al., 2024] Mork, J., Bovbjerg, H. S., Kiss, G., and Tan, Z.-H. (2024). Noise-robust keyword spotting through self-supervised pretraining. ArXiv, abs/2403.18560.

[Mowshowitz, 2012] Mowshowitz, A., & Dehmer, M. (2012). Entropy and the complexity of graphs revisited. Entropy, 14(3), 559-570.

[Mu, 2024] Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., ... & Luo, P. (2024). Embodiedgpt: Vision-language pre-training via embodied chain of thought. Advances in Neural Information Processing Systems, 36.

[Müller et al., 2023] Müller, N. M., Sperl, P., and Böttinger, K. (2023). Complex-valued neural networks for voice anti-spoofing. In Proc. INTER-SPEECH 2023, pages 3814–3818.

[Nalamati, 2019] Nalamati, M., Kapoor, A., Saqib, M., Sharma, N., & Blumenstein, M. (2019, September). Drone detection in long-range surveillance videos. In 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE.

[Newman, 2003] Newman, M. E. (2003). Mixing patterns in networks. Physical review E, 67(2), 026126.

[Nicholls, 2021] Jack Nicholls, Aditya Kuppa, and Nhien-An Le-Khac. 2021. Financial Cybercrime: A Comprehensive Survey of Deep Learning Approaches to Tackle the Evolving Financial Crime Landscape. IEEE Access 9 (2021).

[Nicholls, 2023] Nicholls, Jack, Aditya Kuppa, and Nhien-An Le-Khac. "FraudLens: Graph Structural Learning for Bitcoin Illicit Activity Identification." Proceedings of the 39th Annual Computer Security Applications Conference. 2023.

[Nikolentzos, 2017] Nikolentzos, G., Meladianos, P., & Vazirgiannis, M. (2017, February). Matching node embeddings for graph similarity. In Proceedings of the AAAI conference on Artificial Intelligence (Vol. 31, No. 1).

[OECD, 2016] OECD (2016). Preventing corruption in public procurement.

[Olken, 2007] Olken, B. A. (2007). Monitoring corruption: evidence from a field experiment in Indonesia. Journal of political Economy, 115(2), 200-249.

[Olken, 2009] Olken, B. A. (2009). Corruption perceptions vs. corruption reality. Journal of Public Economics, 93(7-8), 950-964.

[Porter, 1993] Porter, R. H., & Zona, J. D. (1993). Detection of Bid Rigging in Procurement Auctions. Journal of Political Economy, 101(3), 518-538.

[Pudlák, 1988] Pudlák, P., Rödl, V., & Savický, P. (1988). Graph complexity. Acta Informatica, 25, 515-535,

[Putnam, 2000] Putnam, R. D. (2000). Bowling alone: The collapse and revival of American community. New York: Simon & Schuster.

[Radford, 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[Ram et al., 2023a] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023a). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11, 1316–1331.

[O. Ram et al., 2023b] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023b). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11, 1316–1331.

[Ravanelli et al., 2021] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. ArXiv, abs/2106.04624.

[Ravanelli et al., 2024] Ravanelli, M., Parcollet, T., Moumen, A., de Langen, S., Subakan, C., Plantinga, P. W. V., Wang, Y., Mousavi, P., Libera, L. D., Ploujnikov, A., Paissan, F., Borra, D., Zaiem, S., Zhao, Z., Zhang, S., Karakasidis, G., Yeh, S.-L., Champion, P., Rouhe, A., Braun, R., Mai, F., Zuluaga, J. P., Mousavi, S. M., Nautsch, A., Liu, X., Sagar, S., Duret, J., Mdhaffar, S., Laperriere, G., Mori, R. D., and Estève, Y. (2024). Open-source conversational AI with SpeechBrain 1.0. ArXiv, abs/2407.00463.

[Razafindrakoto & Roubaud, 2010] Razafindrakoto, M., & Roubaud, F. (2010). Are international databases on corruption reliable? A comparison of expert opinion surveys and household surveys in sub-Saharan Africa. World Development, 38(8), 1057-1069.

[Reddy, 2018] Reddy, E., Minnaar, A.: Cryptocurrency: A tool and target for cybercrime (12 2018).

[Redmon, 2016] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[Ren, 2016] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

[Richter et al., 2022] Richter, J., Welker, S., Lemercier, J.-M., Lay, B., and Gerkmann, T. (2022). Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31:2351–2364.

[Rieckmann, J. and Stuchtey, T. (2023). Dark Crypto. The Use of Cryptocurrency for Illegal Purposes. Friedrich Naumann Foundation for Freedom

[Rieckmann, J., Stuchtey, T., Lenglachner J., (2022). The Economic Costs of Illicit Trade. BIGS (Brandenburg Institute for Society and Security) Policy Paper

[Rossi, 2020] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal Graph Networks for Deep Learning on Dynamic Graphs. In ICML 2020 Workshop on Graph Representation Learning.

[Roth, 2022] Roth, K., Vinyals, O., Akata, Z.: Non-isotropy regularization for proxy-based deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7420–7430 (2022)

[Saha et al., 2024] Saha, S., Sahidullah, M., and Das, S. (2024). Exploring Green AI for audio deepfake detection. ArXiv, abs/2403.14290.

[Salehi, 2019] Mahsa Salehi, Christopher Leckie, James C. Bezdek, Tharshan Vaithianathan, and Xuyun Zhang. 2016. Fastmemory efficient local outlier detection in data streams. IEEE Transactions on Knowledge and Data Engineering 28, 12 (2016), 3246–3260.

[Salemi et al., 2024] Salemi, A., Kallumadi, S., & Zamani, H. (2024). Optimization methods for personalizing large language models through retrieval augmentation. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 752–762.

[Scheibler et al., 2022] Scheibler, R., Ji, Y., Chung, S.-W., Byun, J. U., Choe, S., and Choi, M.-S. (2022). Diffusion-based generative speech source separation. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

[Shams et al., 2024] Shams, S., Dindar, S. S., Jiang, X., and Mesgarani, N. (2024). SSAMBA: Self-supervised audio representation learning with Mamba state space model. ArXiv, abs/2405.11831.

[Shang, 2010] Shang, H., Lin, X., Zhang, Y., Yu, J. X., & Wang, W. (2010, June). Connected substructure similarity search. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 903-914).

[Shi et al., 2023] Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., & Yih, W. (2023). Replug: Retrieval-augmented black-box language models. arXiv Preprint arXiv:2301.12652.

[Shi et al., 2024] Shi, Y., Zi, X., Shi, Z., Zhang, H., Wu, Q., & Xu, M. (2024). ERAGent: Enhancing Retrieval-Augmented Language Models with Improved Accuracy, Efficiency, and Personalization. arXiv Preprint arXiv:2405.06683.

[Shiohara, 2022] Kaede Shiohara, Toshihiko Yamasaki: Detecting Deepfakes with Self-Blended Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18720-18729

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning, pages 2256–2265. PMLR.

[Song and Ermon, 2019] Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In Neural Information Processing Systems.

[Song, 2024] Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., ... & Wang, G. (2024). Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18221-18232).

[Søreide, 2014] Søreide, T. (2014). Drivers of corruption.

[Sree Katamneni and Rattani, 2024] Sree Katamneni, V. and Rattani, A. (2024). Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization. arXiv e-prints, page arXiv:2408.01532.

[Subakan et al., 2020] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2020). Attention is all you need in speech separation. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 21–25.

[Sun, 2020] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[Swamy et al., 2001] Swamy, A. V., Knack, S., Lee, Y., & Azfar, O. (2001). Gender and corruption. Journal of Development Economics, 64(1), 25-55.

[Tak et al., 2022] Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., and Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation. In Proc. The Speaker and Language Recognition Workshop (Odyssey 2022), pages 112–119.

[Tanzi, 1997] Tanzi, V. and Davoodi H. (1997), Corruption, Public Investment, and Growth, International Monetary Fund Working Paper, 97/139.

[Team, 2024] Team, C. (2024). Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818.

[Tian, 2021] Sheng Tian, Ruofan Wu, Leilei Shi, Liang Zhu, and Tao Xiong. 2021. Self-supervised representation learning on dynamic graphs. In CIKM.

[Tian, 2023] Sheng Tian, Jihai Dong, Jintang Li, Wenlong Zhao, Xiaolong Xu, Bowen Song, Changhua Meng, Tianyi Zhang, Liang Chen, et al. 2023. SAD: Semi-Supervised Anomaly Detection on Dynamic Graphs. In IJCAI.

[Tiwari, 2023] Tiwari, D., Nagpal, B., Bhati, B. S., Mishra, A., & Kumar, M. (2023). A systematic review of social network sentiment analysis with a comparative study of ensemble-based techniques. Artificial Intelligence Review, 56(13407–13461). https://doi.org/10.1007/s10462-023-10472-w.

[Transparency International, 2024] Transparency International. (2024). Corruption perception index 2023.

[Treisman, 2000] Treisman, D. (2000). The causes of corruption: A cross-national study. Journal of Public Economics, 76(3), 399-457.

[Trivedi, 2019] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. 2019. Dyrep: Learning representations over dynamic graphs. In ICLR.

[Trivedi et al., 2022] Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv Preprint arXiv:2212.10509.

[Ulah, 2023] Ullah, W., Hussain, T., & Baik, S. W. (2023). Vision transformer attention with multi-reservoir echo state network for anomaly recognition. Information Processing & Management, 60(3), 103289.

[Utebayeva, 2020] Utebayeva, D., Almagambetov, A., Alduraibi, M., Temirgaliyev, Y., Ilipbayeva, L., & Marxuly, S. (2020, November). Multi-label UAV sound classification using Stacked Bidirectional LSTM. In 2020 Fourth IEEE International Conference on Robotic Computing (IRC) (pp. 453-458). IEEE.

[Valentini-Botinhao, 2016] Valentini-Botinhao, C. (2016). Noisy aware database for training speech enhancement algorithms and TTS models.

[van Deth & Zmerli, 2010] van Deth, J. W., & Zmerli, S. (2010). Introduction: Civicness, equality, and democracy—a "dark side" of social capital? American Behavioral Scientist, 53(5), 631-639.

[Van Rijckeghem & Weder, 2001] Van Rijckeghem, C., & Weder, B. (2001). Bureaucratic corruption and the rate of temptation: Do wages in the civil service affect corruption, and by how much? Journal of Development Economics, 65(2), 307-331.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Neural Information Processing Systems.

[Wahrstätter, 2023] Wahrstätter, Anton, et al. "Improving cryptocurrency crime detection: Coinjoin community detection approach." IEEE Transactions on Dependable and Secure Computing 20.6 (2023): 4946-4956.

[Wang, 2014] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1386–1393 (2014)

[Wang et al., 2019] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N. W. D., Sahidullah, M., Vestman, V., Kinnunen, T. H., LEE, K. A., Juvela, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer, S. L., Becker, M., Henderson, F., Clark, R. A. J., Zhang, Y., Wang, Q., Jia, Y., Onuma, K., Mushika, K., Kaneda, T., Jiang, Y., Liu, L.-J., Wu, Y.-C., Huang, W.-C., Toda, T., Tanaka, K., Kameoka, H., Steiner, I., Matrouf, D., Bonastre, J.-F., Govender, A., Ronanki, S., Zhang, J.-X., and Ling, Z. (2019). The ASVspoof 2019 database. ArXiv, abs/1911.01601.

[Wang, 2020] Wang, Y., Wang, Z., Zhao, Z., Li, Z., Jian, X., Xin, H., ... & Zhao, M. (2020). Effective similarity search on heterogeneous networks: A meta-path free approach. IEEE Transactions on Knowledge and Data Engineering, 34(7), 3225-3240.

[Wang et al., 2021] Wang, Y., Lv, H., Povey, D., Xie, L., and Khudanpur, S. (2021). Wake word detection with streaming transformers. ICASSP 2021 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5864–5868.

[Wang, 2023] Wang, L., Tian, J., Zhou, S., Shi, H., & Hua, G. (2023). Memory-augmented appearance-motion network for video anomaly detection. Pattern Recognition, 138, 109335.

[Wang, 2023a] Wang, L., Wang, X., Liu, F., Li, M., Hao, X., & Zhao, N. (2023). Attention-guided MIL weakly supervised visual anomaly detection. Measurement, 209, 112500.

[Wang, 2023b] Wang S., Sun X., Li X., Ouyang R., Wu F., Zhang T., Li J., & Wang G. (2023). GPT-NER: Named Entity Recognition via Large Language Models. arXiv preprint arXiv:2304.10428v3.

[Wang et al., 2023] Wang, Z., Araki, J., Jiang, Z., Parvez, M. R., & Neubig, G. (2023). Learning to filter context for retrieval-augmented generation. arXiv Preprint arXiv:2311.08377.

[Wang et al., 2024] Wang, R., Dengpan, Y., Tang, L., Zhang, Y., and Deng, J. (2024). Avt2 -dwf: Improving deepfake detection with audio-visual fusion and dynamic weighting strategies. IEEE Signal Processing Letters, PP:1–5.

[Wang et al., 2024a] Wang, H., Zhao, T., & Gao, J. (2024). BlendFilter: Advancing Retrieval-Augmented Large Language Models via Query Generation Blending and Knowledge Filtering. arXiv Preprint arXiv:2402.11129.

[Warden, 2018] Warden, P. (2018). Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. ArXiv e-prints.

[Weber Abramo, 2008] Weber Abramo, C. (2008). How much do perceptions of corruption really tell us? Economics 2(3): 1–56.

[Wei, 2022] Wei, D., Liu, Y., Zhu, X., Liu, J., & Zeng, X. (2022). MSAF: Multimodal supervise-attention enhanced fusion for video anomaly detection. IEEE Signal Processing Letters, 29, 2178-2182.

[Wei, 2022a] Wei, D. L., Liu, C. G., Liu, Y., Liu, J., Zhu, X. G., & Zeng, X. H. (2022, May). Look, listen and pay more attention: Fusing multi-modal information for video violence detection. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1980-1984). IEEE.

[Wittberg, 2023] Wittberg, E., & Fazekas, M. (2023). Firm performance, imperfect competition, and corruption risks in procurement: evidence from Swedish municipalities. Public Choice, 197(1), 227-251.

[World Bank, 2018] World Bank (2018). Global Public Procurement Spending Data. World Bank.

[Wu et al., 2022] Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. (2022). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. ICASSP 2023 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

[Wu, 2018] Wu, C., Wei, Y., Chu, X., Weichen, S., Su, F., & Wang, L. (2018). Hierarchical attention-based multimodal fusion for video captioning. Neurocomputing, 315, 362-370.

[Wu, 2019] Wu, P., Liu, J., & Shen, F. (2019). A deep one-class neural network for anomalous event detection in complex scenes. IEEE transactions on neural networks and learning systems, 31(7), 2609-2622.

[Wu, 2021] Wu, J., Liu, J., Chen, W., Huang, H., Zheng, Z. and Zhang, Y., 2021. Detecting mixing services via mining bitcoin transaction network with hybrid motifs. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(4), pp.2237-2249.

[Wu, 2021a] Wu, B., Nair, S., Martin-Martin, R., Fei-Fei, L., & Finn, C. (2021). Greedy hierarchical variational autoencoders for large-scale video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2318-2328).

[Wu, 2023] Wu, J., Gan, W., Chen, Z., Wan, S., & Philip, S. Y. (2023, December). Multimodal large language models: A survey. In 2023 IEEE International Conference on Big Data (BigData) (pp. 2247-2256). IEEE.

[Wu, 2023b] Wu, J., Liu, J., Chen, W., Huang, H., Zheng, Z. and Zhang, Y., 2021. Detecting mixing services via mining bitcoin transaction network with hybrid motifs. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(4), pp.2237-2249.

[Xia, 2019] Xia, F., Liu, J., Nie, H., Fu, Y., Wan, L., & Kong, X. (2019). Random walks: A review of algorithms and applications. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(2), 95-107.

[Xu, 2020] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. In ICLR.

[Xu, 2022] Xu, Y., Raja, K., Pedersen, M.: Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, Waikoloa, HI, USA, 4–8 January 2022; pp. 379–389

[Xu, 2022a] Xu, C., Chai, Z., Xu, Z., Li, H., Zuo, Q., Yang, L., Yuan, C.: Hhf: Hashing-guided hinge function for deep hashing retrieval. IEEE Transactions on Multimedia (2022) https://doi.org/10.1109/TMM.2022.3222598

[Xu et al., 2024] Xu, F., Shi, W., & Choi, E. (2024). RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. The Twelfth International Conference on Learning Representations.

[Xuan, 2020] Xuan, H., Stylianou, A., Pless, R.: Improved embeddings with easy positive triplet mining. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)

[Xue et al., 2023] Xue, J., Zhou, H., Song, H., Wu, B., and Shi, L. (2023). Cross-modal information fusion for voice spoofing detection. Speech Communication, 147:41–50.

[Yamagishi et al., 2021] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., and Delgado, H. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, pages 47–54.

[Yan, 2019] Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., & Dai, Q. (2019). STAT: Spatial-temporal attention mechanism for video captioning. IEEE transactions on multimedia, 22(1), 229-241.

[Yang, 2023] Yang, A., Nagrani, A., Seo, P. H., Miech, A., Pont-Tuset, J., Laptev, I., ... & Schmid, C. (2023). Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10714-10726).

[Yang et al., 2023] Yang, H., Li, Z., Zhang, Y., Wang, J., Cheng, N., Li, M., & Xiao, J. (2023). Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. arXiv Preprint arXiv:2310.18347.

[Yu et al., 2016] Yu, D., Kolbæk, M., Tan, Z., and Jensen, J. H. (2016). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 241–245.

[Yu et al., 2018] Yu, M., Ji, X., Gao, Y., Chen, L., Chen, J., Zheng, J., Su, D., and Yu, D. (2018). Text-dependent speech enhancement for small-footprint robust keyword detection. In Proc. Interspeech 2018, pages 2613–2617.

[Yu, 2018] Wenchao Yu, Wei Cheng, Charu C Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang. 2018. Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks. In KDD.

[Yu, 2021] Yu, J., Lee, Y., Yow, K. C., Jeon, M., & Pedrycz, W. (2021). Abnormal event detection and localization via adversarial event prediction. IEEE transactions on neural networks and learning systems, 33(8), 3572-3586.

[Yu et al., 2023] Yu, Z., Xiong, C., Yu, S., & Liu, Z. (2023). Augmentation-adapted retriever improves generalization of language models as generic plug-in. arXiv Preprint arXiv:2305.17331.

[Yuan, 2020] Yuan, L., Wang, T., Zhang, X., Tay, F.E., Jie, Z., Liu, W., Feng, J.: Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3083–3092 (2020)

[Yan et al., 2024] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation. arXiv Preprint arXiv:2401.15884.

[Zaheer, 2022] Zaheer, M. Z., Mahmood, A., Khan, M. H., Segu, M., Yu, F., & Lee, S. I. (2022). Generative cooperative learning for unsupervised video anomaly detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14744-14754).

[Zeng, 2022] Zeng, N., Wu, P., Wang, Z., Li, H., Liu, W., & Liu, X. (2022). A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Transactions on Instrumentation and Measurement, 71, 1-14.

[Zenil, 2018] Zenil, H., Kiani, N. A., & Tegnér, J. (2018). A review of graph and network complexity from an algorithmic information perspective. Entropy, 20(8), 551.

[Zerhoudi, M. Granitzer, 2024] Zerhoudi, S., & Granitzer, M. (2024). PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents. arXiv Preprint arXiv:2407.09394.

[Zhang, 2020] Zhang, W., Wang, X. E., Tang, S., Shi, H., Shi, H., Xiao, J., ... & Wang, W. Y. (2020, October). Relational graph learning for grounded video description generation. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 3807-3828).

[Zhang, 2023] Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858.

[Zhang et al., 2023] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., & others. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv Preprint arXiv:2309.01219.

[Zhao et al., 2023] Zhao, S., Ma, Y., Ni, C., Zhang, C., Wang, H., Nguyen, T. H., Zhou, K., Yip, J. Q., Ng, D., and Ma, B. (2023). MossFormer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural

speech separation. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10356–10360.

[Zhao, 2021] Zhao, W., Wu, X., & Luo, J. (2021). Multi-modal dependency tree for video captioning. Advances in Neural Information Processing Systems, 34, 6634-6645.

[Zhao, 2024a] Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. ACM Computing Surveys, 56(9), 1-36.

[Zhao, 2024b] Zhao, Z., Deng, L., Bai, H., Cui, Y., Zhang, Z., Zhang, Y., ... & Van Gool, L. (2024). Image Fusion via Vision-Language Model. arXiv preprint arXiv:2402.02235.

[Zheng, 2019] Li Zheng, Zhenpeng Li, Jian Li, Zhao Li, and Jun Gao. 2019. AddGraph: Anomaly Detection in Dynamic Graph Using Attention-based Temporal GCN. In IJCAI.

[Zheng, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen.: Exploring temporal coherence for more general video face forgery detection. In ICCV, pages 15044–15054, 2021. 5, 6

[Zheng, 2021a] Zheng, W., Wang, C., Lu, J., Zhou, J.: Deep compositional metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9320–9329 (2021)

[Zheng et al., 2023] Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2023). Take a step back: Evoking reasoning via abstraction in large language models. arXiv Preprint arXiv:2310.06117.

[Zhou and Lim, 2021] Zhou, Y. and Lim, S.-N. (2021). Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14800–14809.

[Zhu et al., 2023] Zhu, J., Yang, C., Samir, F., and Islam, J. (2023). The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).

[Zhu et al., 2023a] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Dou, Z., & Wen, J.-R. (2023). Large language models for information retrieval: A survey. arXiv Preprint arXiv:2308.07107.

Funded by the
European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101135577