Anomaly detection in computer vision for dynamic environments

S.Bolourian

Submitted for the Degree of
Master of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

August 2025

Abstract

In recent years, there has been a rise in the popularity of robotic systems used to inspect large infrastructures. Traditionally, the footage collected by such systems requires manual monitoring and analysis by professionals to detect abnormalities. However, this process can be very time-consuming and expensive. Therefore, research activities in automatic visual anomaly detection can have great practical significance in reducing the cost and difficulty of inspection and allowing for a more continuous inspection of infrastructures.

Deep learning visual anomaly detection has achieved state-of-the-art performance on various image and video anomaly detection tasks within research settings. However, applying anomaly detection models to real-world scenarios remains challenging. Real-world data is often noisy, unstructured and diverse, which can cause high false-positive rates and poor generalisation in new environments. Anomalies can also be subtle or context-dependent, which current deep learning models struggle to understand. Lastly, the black-box nature of the models and lack of interpretability and transparency can be a challenge for regulators. To overcome these challenges, this thesis introduces a novel set of anomaly detection models based on modular neural networks and graph neural networks. The results indicate that these models are promising avenues for overcoming the stated challenges while maintaining high-accuracy results.

Key words: Image anomaly detection, Video anomaly detection, Explainable AI , Graph neural networks, modular neural networks

Email: s.bolourian@surrey.ac.uk
WWW: http://www.eps.surrey.ac.uk/

Acknowledgements

I would like to thank my supervisor, Dr Simon Hadfield, for his guidance and support; without his advice and encouragement, this work would not have been possible. I will forever remain grateful for the many different ways in which he has supported me. I would also like to thank NDA and NNL for funding this research, and in particular, John-Patrick Richardson and Ed Butcher from NNL for their continuous support and patience throughout this project.

I owe a huge appreciation to my family for their unwavering love and support over the years. A special thank you goes to my girlfriend for her love and support. Her encouragement gave me strength during the most challenging times. Finally, I would like to thank my dog, Blizzard, who kept me company through countless late nights of writing, offering comfort, joy, and the reminder that life exists outside of research.

Contents

List of Figures

Chapter 1

Introduction

Visual anomaly detection (VAD) is an emerging and increasingly important area of research within computer vision and deep learning with applications in various tasks, such as product defect detection [104], infrastructure inspection [81], medical image analysis [47] and surveillance [10]. VAD models aim to identify unusual data points, patterns, and events in images and videos that deviate from the norm or expected appearance.

Recent years have seen the rapid adoption of deep learning techniques for detecting anomalies in images and videos. Compared to traditional image processing techniques, deep learning models have shown significant performance improvement due to their ability to automatically learn relevant features and capture complex non-linear relationships [8]. Models such as Convolutional Neural Networks (CNNs) [90], Autoencoders (AEs) [65], Variational Autoencoders (VAEs) [88] and Generative Adversarial Networks (GANs) [59] have been instrumental in advancing this field. Deep learning-based VAD methods can be broadly classified into supervised, semi-supervised and unsupervised approaches, depending on the type of data and the training process involved. Supervised VAD models use labelled datasets to train the model to learn discriminative features associated with anomalies [134]. In contrast, unsupervised and semi-supervised VAD models are trained using weakly labelled or unlabeled data. They are typically trained on normal data to learn the underlying patterns and feature representations of anomaly-free images or video sequences. At inference time, significant deviations

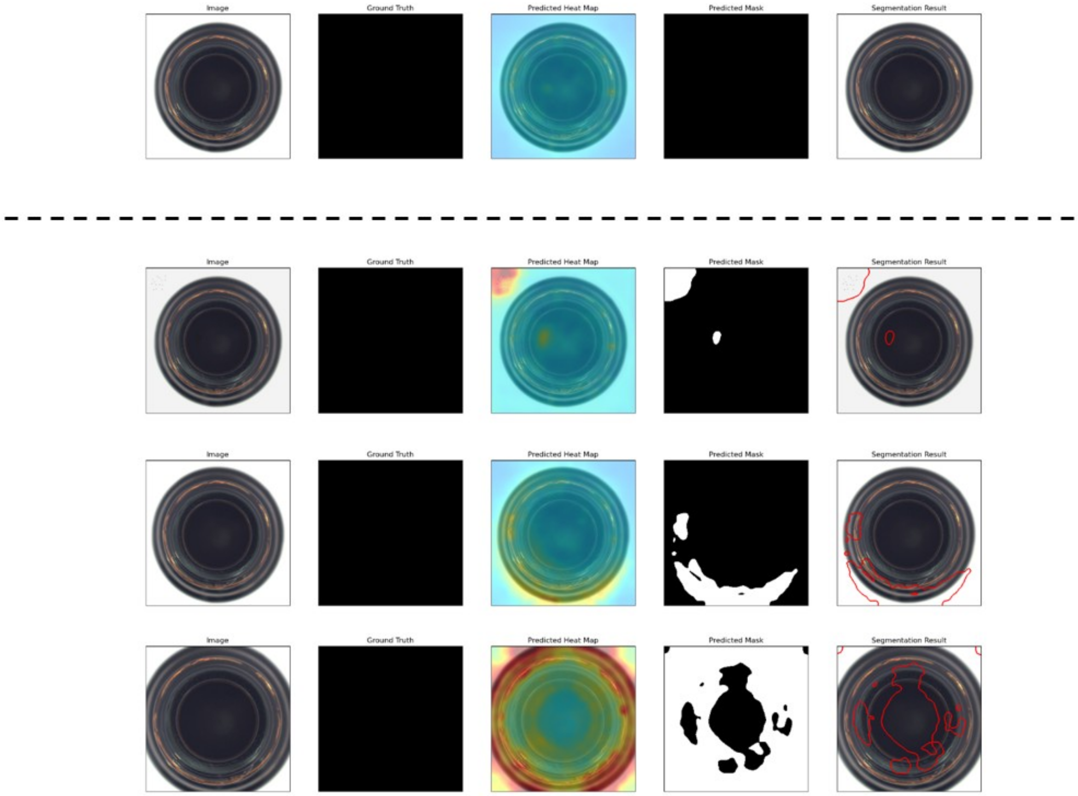


Figure 1.1: Unsupervised and semi-supervised image anomaly detection models can misinterpret normal variations in lighting, background noise, or aspect ratio as anomalies. The first row shows output from the PadIM [35] model on an anomaly-free image from MVTec dataset [18], the next rows illustrate how normal variations are incorrectly highlighted as defects

from this learned representation can be interpreted as anomalies [109] [121]. By definition, anomalies are rare and diverse, making the collection of a comprehensive labelled dataset both labour-intensive and costly [183]. As a result of these inherent challenges, existing work in the VAD domain has predominantly focused on unsupervised methods that leverage reconstruction-based or generative models [183].

While recent advances have significantly improved VAD, important challenges remain that need to be addressed to adopt and apply VAD models in real-world scenarios effectively. This is especially true for applications in dynamic environments. Most existing AD studies assume that the training and test data follow the same data distribution. As a result, current models predominantly focus on achieving high accuracy,

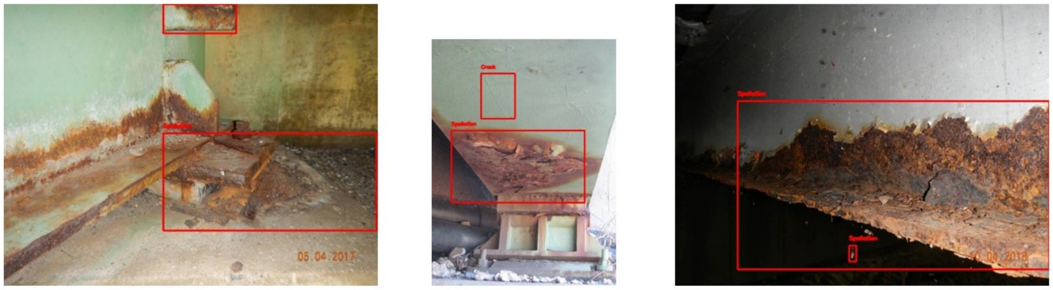


Figure 1.2: Supervised anomaly detection models can misinterpret defect types when context and material properties are not considered. In this example, the model mistakes corrosion for concrete spalling

based on evaluation on benchmark datasets, which typically contain well-defined and controlled anomalies [196]. However, such datasets do not reflect the complexities and variability found in many real-world settings. In practice, models often perform poorly in real-world dynamic environments, where natural variations such as changes in the environment, background, lighting conditions or camera configurations can cause the test data distribution to diverge significantly from the training data, where this distribution shift can have a significant negative impact on the performance of the model [196] [23].

This challenge affects both supervised and unsupervised approaches. In supervised settings, models are prone to misclassification when encountering out-of-distribution anomalies, as models often rely on low-level visual cues without contextual understanding (see Fig. 1.2) [56] [156]. At the same time, in unsupervised or semi-supervised settings, distribution shift can cause normal variations to be falsely detected as anomalies (see Fig. 1.1) [23].

Furthermore, the inherent complexity and "black-box" nature of deep learning models make their decision-making processes opaque and difficult to interpret for the end users. This lack of interpretability and explainability can be a major obstacle to the adoption and deployment of deep learning-based AD models in industrial settings [14]. In safety-critical applications, in particular, model transparency is essential not only to build trust in the system but also to meet regulatory requirements that demand a clear understanding of automated decision-making [102] [95] [123].

In this work, we aim to tackle some of the real-world challenges in visual anomaly detection applications with a particular focus towards anomaly detection tasks in nuclear power plants. In the remainder of this chapter, we provide a brief overview of inspection practices in nuclear power plants. We then outline some of the main challenges that should be addressed to develop effective visual anomaly detection methods for nuclear power plant inspections.

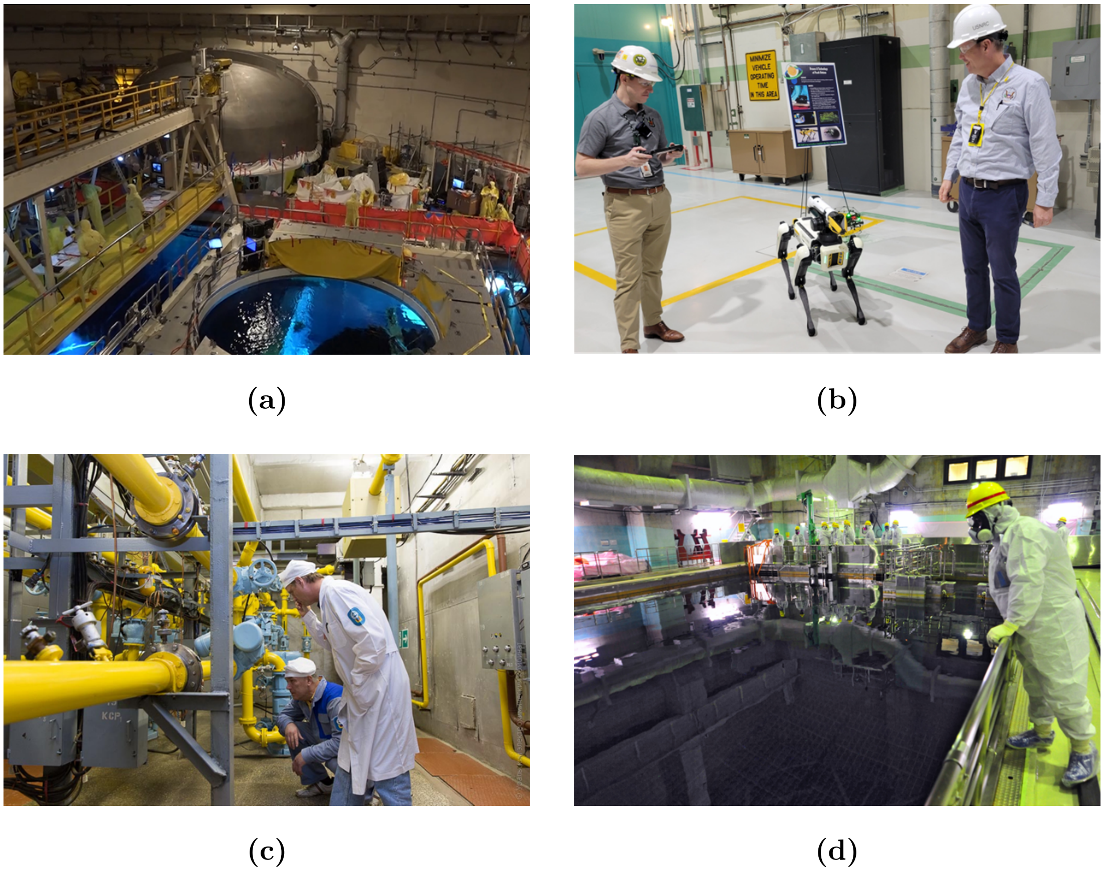
1.1 Industrial context - Nuclear power plant inspection

Structural components and equipment inside a nuclear power plant can be exposed to challenging environments of radiation, harmful gases and particulates, and high temperatures that accelerate deterioration and damage to them[202] [1] [3] [2] [7]. Therefore, periodic inspection of structural components and equipment is essential for ensuring the safety, security, and continuous operation of these sites. Failing to detect defects and damages can have significant environmental and financial costs.

Between 1952 and 2009, 99 nuclear power plant accidents [1] were reported worldwide, resulting in over 20 billion dollars in damages. Of these incidents, several were directly related to structural failures [149] of nuclear power plants, which highlights the importance of accurate and efficient inspection of nuclear facilities.

In many nuclear power plants, visual inspections serve as one of the primary means of monitoring the health and condition of the infrastructure and equipment within the facility (see Fig. 1.3). However, inspections can be a challenging process. Due to the hazardous environment, direct inspections in nuclear power plants can be difficult for human access. Therefore, many inspections are conducted manually by specialists using a variety of robotic systems such as teleoperated cameras [127] [158], legged or wheeled robots [13] [89] [143], unmanned aerial vehicles (UAV) [15] [25] [89] and autonomous underwater vehicles (AUV) [119] [127]. The areas that require inspection are also varied, ranging from general infrastructure and equipment inspection to the inspection of storage silos that can be accessed via single boreholes or underwater storage units in

[1] Accidents are defined as incidents that result in more than US$50,000 of property damage or loss of human life



(a)

(b)

(c)

(d)

Figure 1.3: a) Perry Nuclear Power Plant employees replacing fuel assemblies in the reactor core [48] (CC BY-ND 2.0), b) Inspection robot at Peach Bottom nuclear power plant [33] (CC BY 2.0), c) Inspecting pipes carrying 1.5 pct enriched uranium hexafluoride [76] (CC BY-NC-SA 2.0), d) Canadian Nuclear Safety Commission, examines the Common Spent Fuel Pool at TEPCO's Fukushima Daiichi Nuclear Power Station [74] (CC BY-SA 2.0)

large man-made pools (see Fig. 1.4). Lack of direct access to the inspection area results in poor and limited lighting conditions, which typically comes from a single light source attached to the camera. Combined with environmental radiation, this can introduce noise and reduce the quality and contrast in the captured data.

The collected images and videos are then manually reviewed by domain specialists for various types of defects and anomalies that could, if not suitably addressed and rectified, compromise the integrity and security of the infrastructure. These may include cracks, corrosion, pitting, dents, degradation or wear, structural damage, leaks, fallen debris

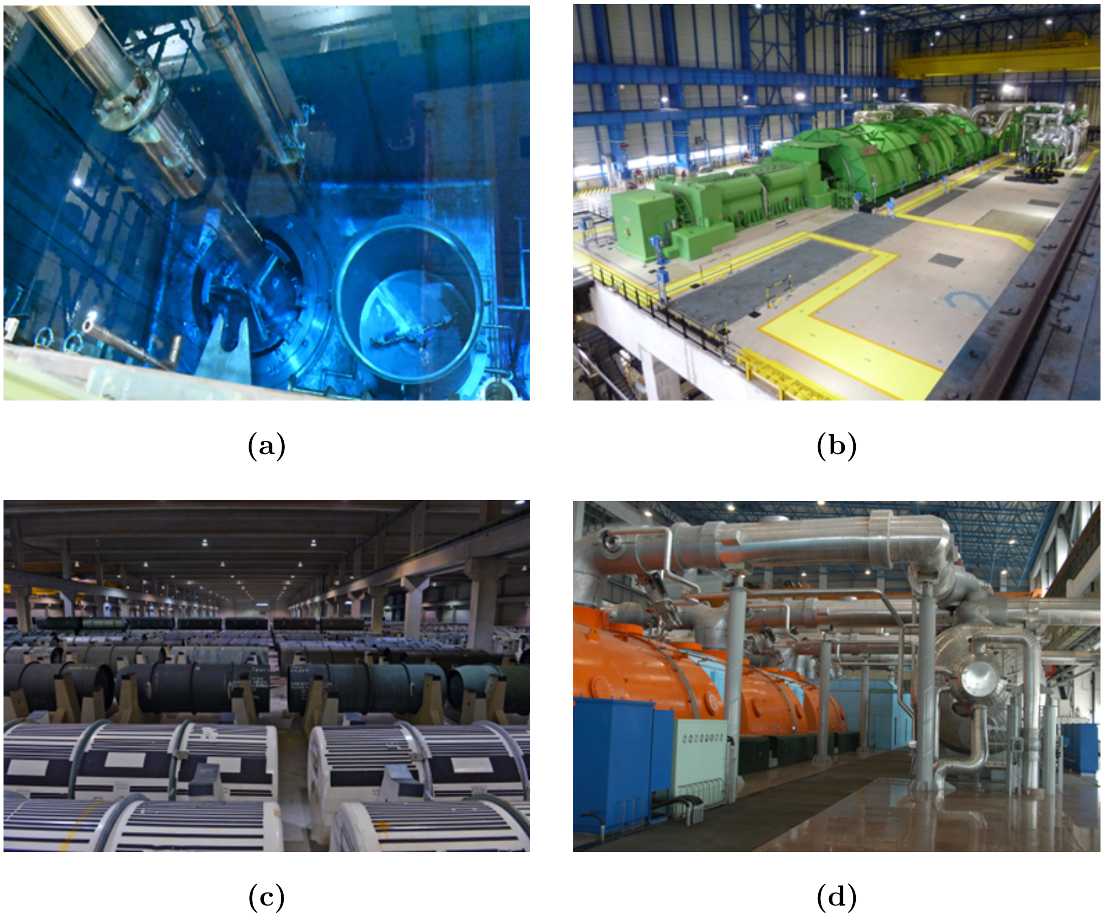


Figure 1.4: a) Central Interim Storage Facility for Spent Nuclear Fuel, Oskarshamn, Sweden [77] (CC BY 2.0), b) Civaux power plant engine room [31] (CC BY-SA 4.0), c) Safeguards Inspection at URENCO, Almelo, Netherlands [75] (CC BY-NC-ND 2.0), d) Turbine hall of Qinshan Nuclear Power Plant [73] (CC BY-SA 2.0)

and damage to equipment. However, the large number of images and videos that require analysis makes this process time-consuming, costly and susceptible to human error. Therefore, there is a real opportunity to improve efficiency and safety and reduce the cost of inspecting nuclear power plants by automating this process.

1.2 Challenges and Motivation

Deep learning-based Video Anomaly Detection (VAD) models have demonstrated strong performance and state-of-the-art accuracy on various benchmark datasets. However, most of these models have been designed and developed specifically for scenarios involving static cameras in controlled environments. In contrast, many real-world applications require models to operate in dynamic environments on images and videos captured from a moving camera [23]. Despite this, VAD in dynamic conditions, particularly with moving cameras, remains a challenging problem, with a wide range of industrial applications in security and surveillance, industrial inspection, transportation and robotics [82].

In response, in recent years, several deep learning-based VAD methods have been proposed, such as: semantic segmentation-based [11] [126] [173], reference frame-based [189] [96], background separation-based [57] [105], object trajectory-based [199] [135] and interaction-based approaches [166] [41] [187]. While these methods often outperform models designed for static environments, they still face major challenges that hinder their practical adoption and performance. A key limitation is that these methods are typically restricted to detecting a limited set of specific anomaly categories, constraining their versatility across a broader spectrum of anomaly types [82]. Moreover, they frequently suffer from poor generalisability, often requiring retraining when applied to new environments or scenes [160], [195], [120]. Additionally, these models are still vulnerable to external factors, such as changes in illumination, occlusions, and complex backgrounds, which frequently lead to false positives or false negatives [40].

These limitations are partly due to the fact that many of the models are derived from more traditional VAD models, which are designed for less complex environments. As a result, they inherit similar foundational limitations. While newer models introduce strategies to handle better the complexities of dynamic conditions, such as camera motion, occlusions, and changing backgrounds, these improvements are often incremental rather than transformative, leaving some of the underlying weaknesses intact.

Furthermore, reasoning and context awareness are essential aspects that remain underdeveloped in current VAD models. In the majority of cases, the model's context awareness level is often limited to spatial, semantic, or spatio-temporal information

[181] [150]. Similarly, the reasoning capabilities of these models are largely data-driven and lack the depth and flexibility characteristic of human cognition. While humans rely on abstract, context-rich understanding, prior knowledge, and causal reasoning, most models detect deviations from learned patterns without comprehending *why* something is anomalous [129] [175]. Although recent methods have begun to incorporate basic relational [150] or semantic reasoning [181], these approaches remain narrow and predefined. These limitations mean that models can struggle when the same visual pattern might signify different anomaly types or no anomalies at all, depending on the context or when the same object or action can represent different meanings given a different context.

Another crucial and underdeveloped aspect is the explainability and interpretability of the model. Existing VAD models generally lack explainability and interpretability [102], making it difficult for users to understand the rationale behind the model's decision-making. This is crucial for mass adoption in many industrial applications, especially those that are safety-critical, both from a regulatory and audit perspective and for developing trust and understanding of the model with the end user [95] [14].

In recent years, several methods have been proposed for explainable image and video anomaly detection [169]. For image anomaly detection, explainability often rests on the model's ability to show how it analysed the image via pixel or region-level explanations that highlight areas most relevant to the anomaly [169]. This is typically achieved with attention maps [91], reconstruction and prediction-error maps [172], perturbation methods [67], foundation model-based [198, 80] or post-hoc techniques such as Grad-CAM [161, 151]. Video anomaly detection similarly uses signals to highlight areas of interest during decision-making [130]. However, recent work also applies structured reasoning through scene graphs, graph neural networks, and knowledge graphs [27, 42]. Nevertheless, the explainability in these methods primarily focuses on the form of the model's output rather than on exposing the model's internal decision process, while understanding the model's reasoning and decision-making at each step remains mostly opaque.

These challenges provide the basis for the main research questions for this work:

1. How to develop a robust VAD model for dynamic environments that can generalise across diverse scenarios without the need for extensive retraining.
2. How to incorporate context-aware reasoning to identify context-dependent anomalies accurately.
3. How to improve the interpretability and explainability of VAD models to enhance transparency for end-users.

1.3 Contributions

The central contribution of this thesis is the proposition that VAD should be formulated as a context-aware compositional reasoning problem rather than as a static task addressed by a single monolithic model. Instead of applying a uniform architecture equally across all regions of an image, this thesis argues that anomaly detection should dynamically adapt its computation based on image content, allowing different regions to be analysed using specialised reasoning strategies. To this end, this thesis makes the following two primary contributions:

- A novel VAD framework that formulates anomaly detection as a compositional problem. By incorporating NMN architecture, our proposed method dynamically assembles a VAD model from a shared set of reusable modules conditioned on the context of the input image.
- A novel physics-based model for detecting anomalies in human motion from video. Our model predicts individuals' future poses by modelling human motion using a graph neural network constrained by kinematic principles.

1.4 Thesis Outline

The remainder of this thesis is organised as follows. Chapter 2 presents a comprehensive literature review on image anomaly detection. In this chapter, we present the challenges associated with each method and review prior work that addresses them.

Chapter 3 introduces a novel formulation of VAD as a compositional problem. To this end, we propose a compositional NMN framework for visual anomaly detection. We also present a comparison of our model with previously published results on three benchmark datasets.

Chapter 4 focuses on skeletal anomaly detection for human motion analysis in video. We introduce a physics-inspired graph neural network that models inter-joint dependencies and temporal dynamics. Furthermore, we evaluated our method on a benchmark dataset and compared the performance with that of existing anomaly detection techniques.

Finally, Chapter 5 concludes the thesis by summarising the motivations, the main contributions, and the performance and limitations of the proposed models. Furthermore, the possible directions for future research are discussed.

Chapter 2

Literature Review

Image and video anomaly detection is a critical task in computer vision, aiming to identify patterns that deviate from normal or expected observations. In this chapter, we review the key deep learning architectures developed for image-based anomaly detection. While video anomaly detection generally extends image anomaly detection by incorporating temporal information to capture dynamic irregularities, in the context of industrial inspection and defect detection, the temporal aspect is often not relevant. Instead, each video frame is typically analysed independently to detect defects or abnormal events, effectively treating video anomaly detection as frame-wise image anomaly detection. Therefore, here we focus primarily on the formulation and methods developed for image anomaly detection, which form the foundation for both image-based and frame-wise video-based industrial inspection systems.

2.1 Unsupervised Anomaly Detection

2.1.1 AutoEncoders

An Autoencoder (AE) is a feedforward neural network that aims to replicate its input data as the output by learning a lower-dimensional latent representation of the input. Autoencoders (AEs) typically consist of two components: an encoder f_θ that maps the input image x to a latent representation z, such that $f : X \rightarrow Z$; and a decoder g_ϕ

that reconstructs the input from the latent representation z as an output $\hat{x}$, such that $g : Z \rightarrow \hat{X}$. During training, the objective is to minimise the reconstruction loss, which is commonly measured using the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\mathrm{AE}}(\theta, \phi) = \sum_{i=1}^{N} \|\mathbf{x}_i - g_\phi(f_\theta(\mathbf{x}_i))\|^2$$

The basic idea behind using AEs for anomaly detection is to leverage the reconstruction error as an anomaly score. AEs are trained on normal data under the assumption that during inference, the model will have a higher reconstruction error for images containing anomalies. However, this assumption does not always hold, as AEs can learn to overgeneralise and reconstruct both normal and anomalous inputs with low reconstruction errors. Overgeneralisation can be especially problematic if the training data contains anomalous samples. Even in small numbers, these can lead the model to learn to minimise reconstruction loss for both normal and anomalous data, thereby reducing its discriminative ability during inference [86]. Furthermore, AEs can fail to capture high-level semantic anomalies [71, 191] while exhibiting sensitivity to noise in the data, which can lead to unstable reconstructions and inaccurate anomaly localisation [124, 180]. To address these challenges, different training strategies and architectural modifications have been explored in the literature.

Adversarial training strategies, such as introducing perturbations in the latent space during model training, have been shown to help the autoencoder extract more stable and discriminative features, thereby addressing the problem of overgeneralization [86]. Another line of research has been the use of self-supervised learning strategies. Introducing auxiliary tasks such as predicting geometric transformations or reconstructing corrupted inputs has been shown to improve model performance by allowing the model to learn richer, high-level representations [71]. An example of this is the Self-Supervision-Augmented Autoencoder (SSR-AE) that trains the model to reconstruct both the original image and the transformation applied to it. This helps the encoder to learn better discriminative features rather than memorising low-level details, making the model more sensitive to anomalies that fail to preserve transformation consistency [71]. Another example is introducing structured (non-i.i.d.) noise with spatial dependencies. Unlike traditional denoising autoencoders that use i.i.d. noise, structured noise better represents realistic anomalies. Rather than pixel-wise denoising, this forces the autoencoder to learn contextual and region-level structure. This training strategy has shown state-of-the-art results on benchmark datasets [16].

The other main path of research has been to utilise architectural modifications and hybrid architectures. For example, dual-teacher models suppress the reconstruction of abnormal features by distilling knowledge from clean reference models, allowing for a better separation of normal and abnormal data [112]. Patchwise AEs reconstruct small, localised patches instead of the whole image, which has been shown to improve anomaly localisation and detection accuracy [34, 110]. Furthermore, models such as Adversarial Dual Autoencoders (ADAE) have proposed a GAN-based method while utilising two autoencoders as the generator and the discriminator. While allowing for more stable training of a GAN-based method, ADAE has also demonstrated strong performance across various dataset types [162]. Similarly, adversarial autoencoder design methods have shown strong performance on various benchmark datasets [194]. In recent years, motivated by the ability of vision transformers to capture long-range dependencies and global context, several studies have also explored integrating transformer architectures into encoder-decoder-based anomaly detection models [98] [116]. These models have demonstrated state-of-the-art performance on benchmark datasets [186]. Finally, Memory-augmented AEs, such as MemAE and PA-MAE, have introduced the use of external memory modules, which have been shown to prevent the model from overgeneralizing to abnormal data during testing [58, 97, 157].

2.1.2 Variational AutoEncoders

A Variational Autoencoder (VAE), similar to a traditional AE, aims to replicate the input data at the output. However, unlike AEs that encode inputs into fixed latent vectors, VAEs introduce a probabilistic framework by encoding each input x as a distribution over the latent space. The encoder maps x to the parameters μ and σ of a Gaussian distribution, from which a latent vector z is sampled using the reparameterisation trick to maintain differentiability. The decoder then reconstructs the input from z. This stochastic encoding encourages the model to learn a continuous and smooth latent space, enabling both accurate reconstruction and the generation of novel, coherent data samples by drawing from the learned distribution.

The loss function is typically composed of two terms:

$$\mathcal{L}_{\mathrm{VAE}}(\theta, \phi) = \sum_{i=1}^{N} \Big(-\mathbb{E}_{q_\phi(z|x_i)} \left[\log p_\theta(x_i|z)\right] + \mathrm{KL}(q_\phi(z|x_i) \parallel p(z)) \Big)$$

where $p(z)$ is the prior distribution (commonly $\mathcal{N}(0, I)$), and $q_\phi(z|x)$ is the variational posterior. The first term encourages accurate reconstruction, while the second regularises the latent space by minimising the divergence between the approximate posterior and the prior. In terms of image anomaly detection, the reconstruction error or the reconstruction probability of VAEs can be used as an anomaly score. However, in practice, using reconstruction error or the reconstruction probability as an anomaly score can be challenging as VAEs are susceptible to overgeneralisation. The smooth and continuous latent space of the model allows it to reconstruct both normal and anomalous inputs with low reconstruction errors. This reduces the model's ability to separate normal and abnormal inputs.

To overcome this, models such as VQ-VAE discretise the latent space to prevent smooth interpolation for outliers to make it harder to accurately reconstruct abnormal inputs [114]. Furthermore, Memory-augmented VAEs prevent overgeneralisation by storing the prototypical representations of normal data during testing. This enables the model to accurately reconstruct normal images, whereas the reconstruction of abnormal images results in higher reconstruction errors. [55, 115]. Similarly, to reduce the effect of noise in training data and prevent overgeneralisation, Robust VAEs (RVAEs) replace the standard Kullback–Leibler divergence with a β-divergence formulation. This helps the model learn a more robust reconstruction boundary between normal and abnormal data [6].

VAEs also encounter difficulties in detecting subtle abnormalities and in identifying regions that do not correspond with human perception of anomalies. To address this, the Feature-Augmented VAEs (FA-VAEs) model proposed calculating reconstruction loss in both pixel space and across feature hierarchies extracted from pre-trained convolutional networks [36]. Other approaches introduce contrastive learning objectives, which, during training, encourage the encoder to cluster similar normal inputs and separate dissimilar or potentially anomalous ones. This process enables the model to learn embeddings that accurately capture high-level semantic differences. Both approaches have achieved high accuracy on benchmark datasets [111].

Lastly, VAE models can produce unreliable likelihood estimates. In practice, the model can assign higher likelihoods to out-of-distribution samples resulting from minor input variations, such as changes in lighting conditions. To address this likelihood, debiasing techniques have been proposed, which adjust the model's estimates to better separate anomalies from normal data. Additionally, ensemble-based methods, which aggregate likelihood estimates from multiple VAE models, have demonstrated improved robustness against such biases [26, 174].

2.1.3 Generative Adversarial Networks

A Generative Adversarial Network (GAN) consists of two competing networks: a generator network and a discriminator network. Generator network $G(\mathbf{z}; \theta_G)$ attempts to generate fake samples from a noise vector $\mathbf{z}$; while the discriminator $D(\mathbf{x}; \theta_D)$ aims to learn to classify if the input is real or is fake from the generator:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

GAN-based image anomaly detection models rely on the assumption that a model trained on normal images will fail to reconstruct anomalous inputs accurately. Broadly, these models can be categorised into two main types: reconstruction-based and discriminator-based approaches. In discriminator-based methods, the discriminator's ability to learn the boundary between normal and abnormal images is utilised. In reconstruction-based methods, the generator's reconstruction error is used for anomaly detection. One of the seminal works in this field was the proposed AnoGAN model [142]. During training, the GAN model is trained on normal data. During testing, the model searches the latent space for a latent vector that can best reconstruct the input image. Models such as f-AnoGAN [141] removed the need for searching the latent space by introducing an encoder. Here, the encoded image features are passed as a latent vector z to the

GAN model for reconstructing a normal image. Another seminal work in this field was the introduction of GANomaly [5]. GANomaly employed an encoder–decoder-encoder design architecture. As such, instead of relying solely on input–output reconstruction, GANomaly also utilises the discrepancy between the latent representations of the input image and the encoded representation of its reconstruction. Therefore, the anomaly score combines image-level reconstruction error and latent-space consistency error, which helps the model to distinguish between normal and abnormal inputs more effectively [5].

To improve model accuracy and performance in producing anomaly masks, later works introduced the use of self-supervised masking during training [72]. SCADN [182] introduced the use of a fixed set of strips for super-pixel segmentation as masking candidates. Similarly, AnoSeg [148] used masking and synthetic anomaly data to increase accuracy on benchmark datasets. At the same time, models such as Fence GAN [122] introduced the idea of training the generator's objective to produce samples near the normal–abnormal boundary. By doing so, the discriminator can better separate between normal and abnormal data at the boundary, improving sensitivity to smaller anomalies that standard models may miss.

Another interesting line of research has been the introduction of CycleGAN [201] methods for image anomaly detection [17] [167]. Although anomalies are rare, in some cases, we may have a few examples or be able to generate synthetic ones. CycleGANs are able to translate between normal and abnormal images, and back to normal, without the need for paired data. During training, the model learns image-to-image translation between anomalous and normal domains. During testing, the input image is translated to an anomaly-free version of that image. The anomaly score is then obtained by measuring the difference between the input image and its translated version. Furthermore, by learning bidirectional mappings, CycleGAN can enhance its generalisation capabilities. At the same time, the cycle-consistency loss helps mitigate mode collapse by acting as a regularizer. Another architectural innovation was made in Y-GAN [78]. The goal of Y-GAN is to disentangle better informative image semantics relevant to normal training data from uninformative residual information. Unlike vanilla reconstruction-based models, Y-GAN employs two encoders, which encourages the model to separate the latent space into semantic features and low-level residual features. This enables the model to generalise more effectively and better highlight subtle anomalies [78].

2.1.4 Teacher-Student

Teacher-student (T-S) architectures rely on the premise that a student model that is trained to replicate a pretrained teacher's outputs will diverge when processing anomalous data. As such, the divergence between various layers of the teacher and student model can be used to detect and localise anomalies. The seminal work in this field was introduced by Bergman et al. [19], which outperformed many other benchmark methods at the time. However, one of the main challenges with T-S architectures is feature leakage. This occurs when the student network overfits to the teacher's distribution. As such, the discrepancy between the teacher and student network for anomaly data is removed.

Reverse Distillation [39] was one of the main works that tried to address this issue. Unlike traditional T-S architectures, the Reverse Distillation model consists of a teacher encoder and a student decoder. While the teacher network outputs a low-dimensional embedding of the image, the student decoder takes the low-dimensional embedding as input and aims at generating the teacher model's representations at different levels [39]. Other methods, such as the Normal Feature Bank [168] and decoupled representation learning strategy [106], have also been proposed to mitigate feature leakage between the teacher and student networks. While the former selectively preserves representative normal features during training [168], the latter allows the model to maintain feature-space separation between normal and abnormal data [106].

Building on this, several works incorporate multi-scale feature matching by training the student's feature pyramid to replicate that of the teacher. This enables the model to capture fine-grained anomaly localisation across different resolutions. However, it requires that the student and the teacher share the same architectural design [163]. Another architectural innovation has been the introduction of dual student and dual teacher networks. Dual student models utilise two complementary student networks, which allows the model to capture richer representation and improved anomaly sensitivity [61] [184]. At the same time, multi-teacher configurations aim to increase generalisation by distilling knowledge from multiple teacher networks [132]. Furthermore, in recent years, research has also focused on incorporating Transformer-based architectures [12] to enable the model to better capture long-range dependencies and contextual information within the image.

2.1.5 Memory Banks

Memory bank architecture functions by storing a representative set of features from normal data into a dedicated memory bank. The features are typically outputs of a pre-trained model, such as variations of ResNet models. This is usually achieved by sampling normal features directly or by clustering to form representative prototypes. During testing, the extracted feature embedding of the input image is compared with the memory bank, and the similarity or distance between them is used to determine anomaly scores. In terms of memory banks, one of the main challenges is the increase in memory requirements when dealing with complex environments or multi-class anomaly detection scenarios.

Models such as SPADE [32] use a pretrained WideResNet-101 during training to build a pyramid of feature embeddings. These embeddings are stored in a memory bank of normal patch-level descriptors. During testing, anomalies are detected by measuring the distances between test patch features and the nearest normal features in the memory bank. While effective SPADE requires a high memory footprint, this can be problematic for large datasets or high-resolution images. Building on this PaDiM [35] was proposed to overcome some of SPADE's inefficiencies. Instead of storing all patch embeddings, PaDiM models the distribution of normal patch features at each spatial location using a multivariate Gaussian. During testing, each patch feature is measured against this Gaussian distribution. This formulation means that the PaDiM memory bank size is independent of the training set size and is solely dependent on the image resolution. Other methods, such as SOMAD [101] proposed using Self-Organising MAPs (SOMs) to cluster and store multi-scale CNN features while preserving the structure of the normal data.

One of the seminal works in this area is PatchCore [140], which achieved state-of-the-art results on benchmark datasets while maintaining fast and efficient inference. This was achieved by extracting patch-level features from normal images during training using a pre-trained ResNet backbone, and constructing a memory bank containing a coreset of these features. To ensure that the stored coreset of features remains representative of the full diversity of normal patches, PatchCore utilises a greedy facility location algorithm. During inference, patches are compared to the stored features using nearest neighbour search.

While the methods discussed so far use pre-trained CNNs to extract image features, another line of research has employed encoder-decoder architectures for feature extraction. For example, MemAE [58] introduced memory-augmented autoencoders in which memory items are used to reconstruct inputs; high reconstruction errors indicate potential anomalies. Another interesting line of research has been the proposed incorporation of both normal and annotated abnormal feature representations. Methods such as Dual Memory Bank Anomaly Detection (DMAD) [69] employ a CNN backbone to extract features, storing them in dedicated normal and abnormal memory banks. During testing, the distance between input image features against both memory banks is calculated. These semi-supervised approaches have shown high performance and accuracy on benchmark datasets; however, unlike previous examples discussed above, they require access to both normal and abnormal training data.

2.2 Supervised Anomaly Detection

In applications where labelled data for both normal and abnormal images is available, supervised learning techniques can be employed. For clarity, we consider supervised learning to mean any method that relies on human-labelled anomalies. If an annotated dataset is available, a straightforward approach is to use traditional image classification, object detection, and segmentation models for image anomaly detection. Architectures such as CNNs [155], YOLO [197], U-Net [152], and Mask R-CNN [176] can be trained on labelled data for image anomaly detection, including image-level classification, object detection, and semantic or instance segmentation. In recent years, the use of transformer-based models has increased due to their ability to capture long-range dependencies. Accordingly, transformer models such as ViT [147], Swin Transformer [203], and DETR [108] have also been applied to image anomaly detection. Such methods can achieve high accuracy on a variety of image anomaly detection tasks. However, their methods require a sufficient number of annotated abnormal examples and handling of class imbalance, which is common in anomaly detection tasks. In some cases, we may have access to only a few examples of labelled data or have access to labelled data at a lower granularity than is needed to train a fully supervised model. In such cases, weakly supervised, mixed-supervised, and few-shot supervised methods can be utilised.

Weakly supervised methods are trained to output pixel-level localisation of anomalies using image-level labels or coarse spatial annotations. For example, DeScarGAN [171] achieves pixel-level anomaly detection and localisation by learning image-to-image translation between normal and abnormal images using only image-level labels. A dual-branch generator and discriminator are trained with adversarial, classification, and cycle reconstruction losses, enabling the model to convert an input image into its anomaly-free counterpart. The difference between the input image and the translated output creates a pixel-level anomaly map [171]. Other proposed methods utilise Class Activation Maps (CAM) or Multiple Instance Learning (MIL). For example, CAVGA [161], an attention-based VAE model, is trained to reconstruct the input data and accurately classify images using image-level labels. Using a small percentage of examples of anomalies has shown state-of-the-art performance for this model [161].

In cases where abundant image-level labels are available, mixed supervision methods aim to combine the available data with a few strong pixel masks to close the gap to full supervision. [21] introduced a dual-head segmentation–classification network and demonstrated that adding only a handful of fully annotated samples to a large number of weakly labelled images can achieve performance comparable to that of fully supervised methods. [68] further demonstrate the effectiveness of joint weakly and fully supervised learning for surface-defect segmentation, confirming the efficiency of combining sparse masks with plentiful weak labels.

Another important line of research has also been few-shot supervised learning. Few-shot supervised learning is designed to address the core challenge of data scarcity in some applications. The main idea is to learn discriminative boundaries that generalise beyond the seen positives from only a small number of labelled anomalies. DevNet [128] introduces a deviation-based objective that enforces statistically significant separation between normal scores and the few labelled anomalies, yielding robust few-shot detectors. BGAD [185] introduces explicit boundary-guided contrastive learning to mitigate bias towards seen anomalies.

Finally, because labelled anomalies are rare, some methods synthesise pseudo-anomalies, either with generative models (e.g., CycleGAN) or via image-space augmentations. Depending on whether human-labelled anomalies are used during training, these approaches can be classified as self-supervised [99, 190] or supervised [60, 63, 170]. However, in both cases, they typically utilise supervised learning architectures and losses.

In summary, supervised image anomaly detection methods deliver strong detection, localisation, and accuracy when dense labels are available. However, due to the nature of the task, such annotation is often scarce. In these cases, research has focused on weakly supervised, mixed-supervised, and few-shot supervised methods. These methods have demonstrated high accuracy and, in many cases, outperform comparable unsupervised or semi-supervised models. Nonetheless, supervised methods remain vulnerable to domain shift and may generalise poorly to unseen environments without recalibration or adaptation.

2.3 Discussion

In this chapter, we presented a structured literature review of image anomaly detection methods, categorised according to architectural type and level of supervision. The key challenges associated with each architectural class and the strategies proposed in the literature to address them were also addressed. Overall, the review shows that substantial progress has been made in image anomaly detection approaches.

Despite this progress, the literature review also highlights several recurring challenges affecting existing models, irrespective of the supervision level or architectural choice. These challenges include sensitivity to dynamic and non-stationary environments, difficulty in distinguishing anomaly-related changes from normal environmental variations, limited access to interpretability and explainability of the model decision-making process, and a lack of explicit use of contextual information during VAD.

A key factor contributing to these challenges is that most existing methods are applied uniformly across the image under the assumption that training and test data are drawn from the same underlying distribution. A direct consequence of this is the conflation of normal with abnormal changes. Many existing approaches rely on reconstruction errors or feature-space distances that lack semantic grounding, implicitly treating any deviation from learned normality as anomalous. As a result, such models are overly sensitive to expected visual variations, including changes in illumination, background appearance, or viewpoint, while potentially overlooking subtle yet semantically meaningful anomalous events. Furthermore, these methods offer limited explainability. This lack of explainability reduces robustness and hinders trust, particularly in safety-critical applications where understanding the cause of an anomaly is essential.

To address these needs, Chapter 3 presents a compositional, object-centric visual anomaly detection model based on a neural module network architecture. Inspired by the compositional nature of human problem-solving, the proposed approach reasons explicitly about objects and regions within an image, rather than treating the image as a homogeneous input. By dynamically assembling task-specific modules conditioned on the image content, the model determines how individual objects should be analysed, or whether analysis is required at all. This compositional reasoning framework improves robustness underdynamic conditions by reducing sensitivity to irrelevant visual variations, while also providing a degree of explainability.

Chapter 4 addresses this limitation through a human-centric anomaly detection framework based on graph representations and physics-inspired graph neural networks. In this approach, the human body is modelled as a graph of joints and connections, enabling explicit representation of joint relationships and motion dynamics over time. By incorporating physics-inspired modelling principles, the proposed method distinguishes natural variability in human motion from genuinely anomalous behaviour in a structured and interpretable manner. This representation supports the separation of anomaly-relevant changes from benign variation while maintaining robustness in dynamic environments. Moreover, grounding anomaly detection in physically meaningful relational constraints enhances explainability by providing interpretable anomaly decisions rooted in motion dynamics.

Chapter 3

Compositional Anomaly Detection with Neural Module Networks

3.1 Introduction

In recent years, deep learning models have become the de facto standard architecture for both image and video anomaly detection. These models have demonstrated remarkable performance across a range of applications, largely due to their ability to learn high-level feature representations from large datasets.

However, as outlined in the previous chapter, these models operate under the key assumption that the data encountered during inference shares the same underlying statistical properties as the training data. In many real-world scenarios, however, data distribution shifts are common and can significantly undermine model performance [196] [23]. This is particularly problematic in applications that require models to operate in complex environments with dynamic and cluttered scenes or on footage captured by moving cameras. In such settings, factors such as changes in lighting, background and camera motion can introduce significant distribution shifts, leading to performance degradation and poor generalisation [196] [23].

To address the challenges introduced by dynamic environments and moving cameras, various deep-learning-based approaches have been proposed. Methods based on semantic segmentation [11] [126], the use of reference frames and background separation [189] [57], object trajectory detection from spatiotemporal information [199] [135], and interaction models leveraging spatiotemporal contextual clues [166] [41] have been extensively explored to enhance the accuracy and robustness of anomaly detection in complex real-world settings. Despite these advancements, fundamental challenges remain with respect to domain shift, environmental variability, limited contextual awareness and poor generalisations [82] [160], [195], [120] [40].

Current deep learning-based VAD models designed for dynamic environments or moving cameras are largely extensions of earlier models developed for static scenes. As such, they inherit many of the same model assumptions and design principles, which limit their ability to address the core challenges posed by environmental variability and camera motion. Rather than fundamentally rethinking the modelling approach, these methods often adopt workaround strategies such as motion compensation or background modelling, without directly addressing the root causes. Consequently, they continue to suffer from the same inherent limitations as their predecessors.

In contrast to deep learning architectures used for VAD, human visual cognition employs a far more nuanced and involved approach to VAD. Substantial evidence suggests that human visual cognition and problem-solving are inherently compositional [66], meaning that the brain represents and processes visual information in terms of structured components and their spatial, functional, and causal relationships [145]. Compositionality in visual cognition enables efficient problem-solving by allowing the decomposition of complex scenes into meaningful parts and understanding their interrelations [53] [136] [37]. This hierarchical processing also allows for rapid generalisation, as the brain can recognise new objects and understand and problem-solve in new environments by assembling familiar subcomponents rather than memorising entire scenes [20] [93] [94]. Compositionality also enhances visual search by directing attention based on expected relationships and facilitates anomaly detection by identifying deviations from expected learned structures. In dynamic environments, it enables adaptive reasoning through the flexible reuse and recombination of previously learned elements to navigate novel situations [92] [44].

Inspired by this, we propose a novel compositional VAD model based on Neural Module Networks (NMNs) architecture. NMNs have previously demonstrated strong performance across a range of computer vision tasks [45]. However, to the best of our knowledge, neural modular networks and explicit compositional reasoning have not previously been used for image or video anomaly detection. In this work, we aim to bridge this gap, demonstrating the potential of modular compositional neural networks for flexible, adaptive, and accurate anomaly detection in complex visual environments. Furthermore, we aim to move beyond task-specific solutions by proposing a unified and generalizable framework for anomaly detection. Lastly, the modular design also enhances interpretability and explainability, which are critical factors in many applications.

VAD as a Compositional Problem: Anomaly detection tasks often exhibit an underlying modular and compositional structure, regardless of the specific domain or anomaly type. A key challenge lies in modelling this structure. Scenes may contain multiple objects from different categories (e.g., pipes, beams, motors, cables), each associated with its own normal and abnormal patterns. For instance, a dent may be acceptable on one component but considered anomalous on another, depending on its material or function. Similarly, context-specific factors such as environment and location can influence whether the presence or severity of anomalous patterns is deemed acceptable. For example, a crack in a structural support beam may be critical, whereas a similar crack on an exterior ground surface may be considered acceptable. To address this complexity, we formulate VAD as a compositional problem. This enables our model to reason in terms of object-specific and context-dependent components.

Consider the inspection in Figure 3.1 . To ensure the integrity and safety of the pipe work carrying radioactive material, an inspector might prioritise inspecting pipes and the support structure to identify abnormalities while disregarding minor scratches on the ground or walls. The inspection task can be decomposed into a set of subtasks:

1. Detect pipework

2. Locate pipes → check pipes for deformation, leaks and signs of corrosion

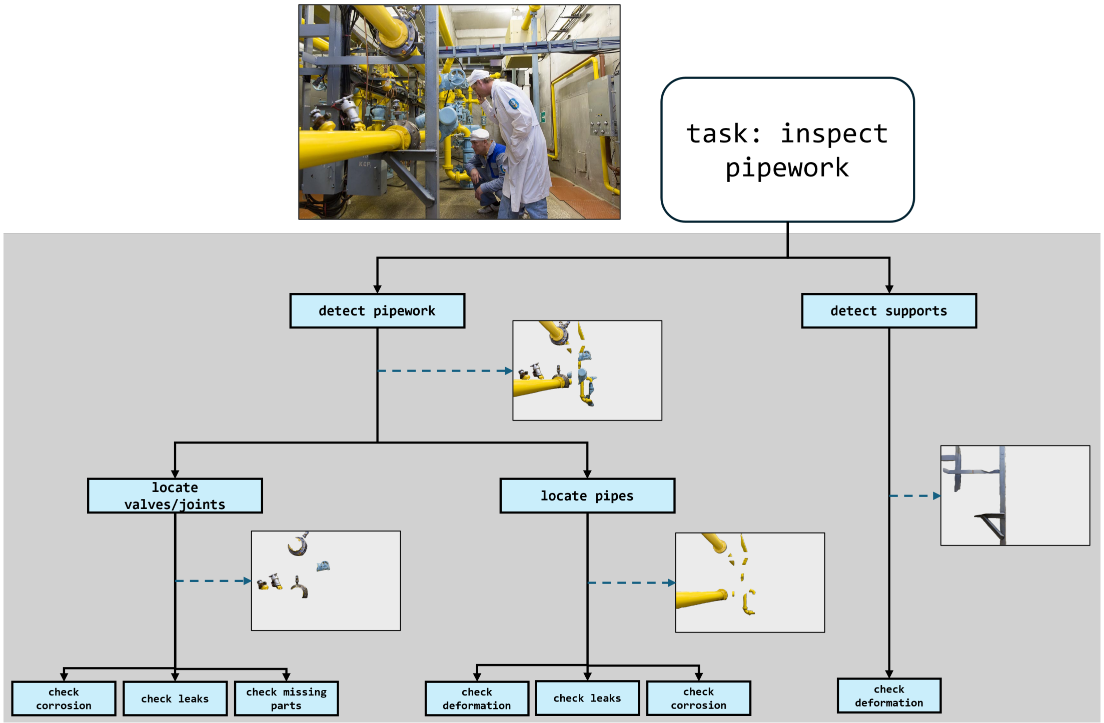


Figure 3.1: Decomposed inspection steps. Original image from [76] (CC BY-NC-SA 2.0)

3. Locate valves and joints → check them for signs of leaks, corrosion and missing parts

Here, we can consider that each subtask corresponds to a distinct computational function. The first step, detecting pipework, can be performed by a generic detection module, which maps from image features and a semantic token (e.g., ”pipework”) to a spatial distribution. The second step, locating pipes, narrows the area of interest and focuses on the pipe section, which can be considered as a form of `filter` or `reattention` module. Finally, specialised modules can check the area of interest for signs of object and material-specific deformation and abnormalities.

Therefore, to achieve this formulation, we propose an NMN model that dynamically assembles its structure from a collection of reusable neural modules, each specialised for a distinct function, based on the content of the input image. This approach reflects the compositional nature of the task and allows the model to adapt its reasoning process to the specific context of the scene. Unlike monolithic deep learning architectures that rely on fixed processing pipelines, our model enhances adaptability, interpretability, and robustness across diverse anomaly types and complex environments.

3.2 Related Work

In this section, we mainly focus on Neural Modular Networks (NMNs) and the concept of compositionality in Visual Question Answering (VQA), as it aligns most closely with the formulation of our problem. Compositionality is an important aspect of generalisation as it enables models to address unseen scenarios by recombining previously acquired knowledge [28] [153]. Furthermore, it plays a central role in human cognition, as it provides the ability to combine simpler concepts to form more complex ones [146] [133]. NMNs are designed to decompose complex tasks into simpler sub-tasks, where specialised modules complete each subtask.

The NMN, proposed by Andreas et al. [9], was an early attempt to make neural networks compositional by dynamically assembling a task-specific network at inference, rather than using a fixed, monolithic network. The proposed method leveraged the natural compositional nature of questions within the framework of VQA tasks. The question is first parsed into a set of functions, and the network is then assembled by selecting and composing reusable modules, each corresponding to one of these functions. This approach achieved state-of-the-art results on both natural and synthetic datasets [9].

However, NMN was limited by the need for hand-crafted parsers that could convert questions into a dynamically assembled network. As a result, End-to-End Module Networks (N2NMNs) were introduced, incorporating the REINFORCE algorithm alongside an LSTM-based sequence model to generate network layout directly from the input question, which eliminated the need for hand-crafted parsers [70]. However, N2NMNs still required a set of handcrafted modules, which limited the model generalisation to new tasks and scalability. As such, subsequent work aimed to eliminate the need for handcrafted modules and instead proposed a method where each module shared the same architecture, regardless of its specific task. In this way, each module starts with the same architecture and only evolves into task-specific experts through training [84].

To further improve the generalisation and scalability of previous NMNs methods, the Meta Module Network (MMN) was also proposed. MMN uses function recipes to generate instance-specific modules dynamically. This architecture enables the system to adapt to a wide range of tasks without increasing model complexity [29]. More recently, researchers have integrated Transformer architectures with modular approaches to enhance NMN performance further. For example, Transformer Module Networks (TMNs) [179] , proposed the use of Transformer blocks instead of traditional CNN-based modules. Notably, their findings also suggest that even unmodified NMNS generalise better than flat Transformers, underscoring the advantages of modular structures[179].

NMN architectures have demonstrated strong performance on various VQA tasks. Although they better reflect the compositional nature of human cognition, their adoption has remained limited due to the complexity of designing and implementing these models.

3.3 Methodology

We introduce a learnable compositional model based on a neural modular network architecture for VAD. The overall architecture of the model is shown in Figure 3.2 . The core idea behind our proposed method is to replace the fixed monolithic model architectures commonly used for VAD (as discussed in Chapter 2) with a model that dynamically assembles an input-specific network from a library of specialised neural modules. This allows the computation performed by the model to adapt to the visual content of each individual input image.

This is achieved by decomposing the anomaly detection task into a set of distinct sub-tasks, where each subtask can be completed by a dedicated learnable module. In contrast to fixed VAD pipelines, our method dynamically assembles a task-specific network per input, improving adaptability and interpretability.

The model consists of three components: a set of neural modules $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$, each with learnable parameters θ_m, which serve as fundamental building blocks that can be dynamically assembled to form task-specific networks. A layout policy generator

$p(z \mid I; \theta_p)$, which predicts a scene-specific layout z to enable the dynamic assembly of a neural network, such that z specifies the subset of modules and how they are connected. Finally, an inference model $p(a \mid z, I; \theta_i)$ that uses the predicted layout z to assemble a task-specific neural network composed of a set of neural modules that produce anomaly-related predictions.

Concretely, given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the model is trained to detect and localise anomalous regions via a dynamically assembled model from a predefined set of modules. The final output of the model is produced by a dedicated prediction module that forms the terminal node of the assembled modular network, and can output either pixel-wise anomaly segmentation mask $\hat{S} \in [0, 1]^{H \times W}$ or a set of anomaly bounding boxes $\hat{B} = \{(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{c})\}_{j=1}^{N}$ such that $\hat{c}$ denotes a confidence score.

To achieve this, first given the input image I, the learnable layout generator $G(.)$ predicts a layout configuration l. The predicted layout specifies which modules are selected from the predetermined set of modules, how these modules are connected, and which external arguments are provided to each module. Based on this layout configuration, the model is dynamically assembled by instantiating the corresponding modules and connecting them according to the predicted structure. The resulting modular network can be represented as a tree-structured computation graph, in which nodes correspond to individual modules and edges define hierarchical or parallel relationships between their outputs and inputs.

Each module m_i is a function f_θ parameterised by learnable weights θ and a set of input arguments $A = \{a_1, a_2, \ldots, a_n\}$ with $n \geq 0$. Therefore, a module can operate with zero, one, or multiple inputs. The inputs to a module can originate from three sources:

1. The output of other modules, enabling hierarchical reasoning.
2. Features extracted from the input image, providing direct visual information.
3. Embeddings generated by the layout generator, which encode structural or contextual priors for dynamic network assembly.

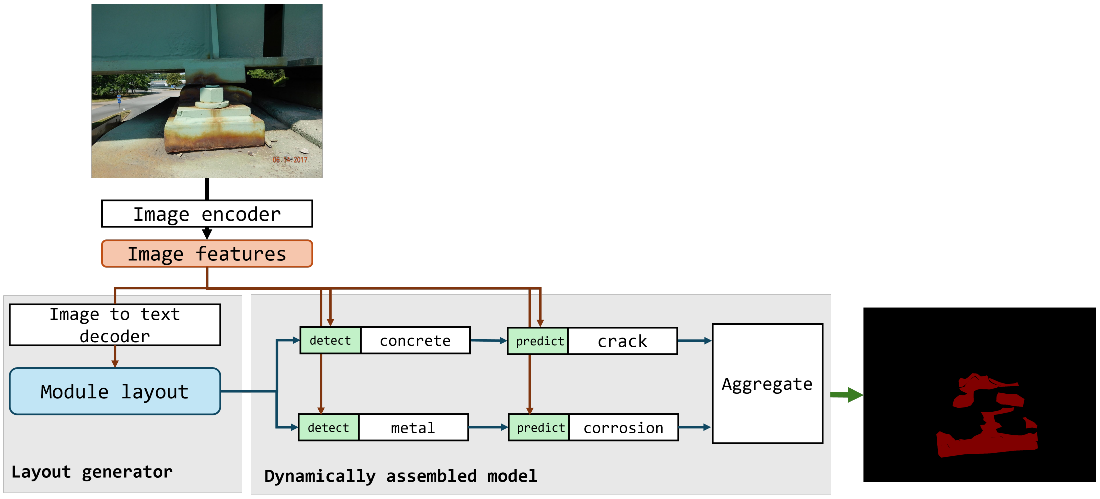


Figure 3.2: Overview of our proposed modular model for image anomaly detection

At runtime, the modules are dynamically assembled into a network based on a layout l, which defines a computational expression consisting of interconnected modules. For example, a layout can be expressed as:

$$f_{m_4}(f_{m_3}(f_{m_1}, f_{m_2})) \tag{3.1}$$

where each f_{m_i} represents a module within the network.

3.3.1 Layout Generator

To better understand the role of the layout generator model, we start by providing an intuitive example. Let us consider how a human expert analyses a scene for anomalies. For the task of inspection, an expert, based on information from observing the scene, will decide on the set of subtasks that need to be completed. For each subtask, a set of actions must be taken in the correct order. In terms of our proposed framework, a layout generator can be considered as the model that takes an image as input and, based on that image, decides on a set of subtasks, the actions that need to be taken to complete those subtasks, and the order in which those actions must be taken. Similarly, we can consider each neural module as a simple learnable action and the set of neural modules as the library of simple actions an individual has access to.

Therefore layout generator predicts a configuration L that specifies the optimal assembly of modules for analysing an input image I. The layout encodes the modules to be used, their interconnections, and any required external inputs. Previous NMN approaches in computer vision rely on an external query, such as a natural language question, to guide layout prediction. However, in VAD, such queries are not available at inference time. To address this, we treat the input image itself as the query, allowing the model to generate the layout directly from visual content, which represents a novel alternative to addressing this task. This represents a fundamental shift from prior NMN-based methods, which rely on externally provided task specifications. In contrast, our approach requires the model to infer the tasks it must perform from the image itself. Furthermore, while conventional NMNs typically predict a single layout per input, as shown in Figure 3.2, VAD may demand multiple distinct layouts for a single image, each targeting a different anomaly detection objective.

The NMN layout can be represented as a tree structure [70], where nodes correspond to modules and edges define hierarchical or parallel relationships. This tree-structured layout can be linearised using a domain-specific language that preserves structural hierarchy through delimiters, such as parentheses for subtrees and brackets for external arguments. Once linearised, layout prediction becomes a sequence-to-sequence problem, mapping an input image I to a layout sequence L.

Therefore, we model layout prediction as a sequence generation task, where an image's visual features are mapped to the corresponding layout. In our implementation, we use a pre-trained Vision Transformer (ViT) as the encoder and a decoder-only transformer for generating the layout sequence. The decoder then generates the linearised layout. Concretely, a learnable function $G()$ maps input image I to a linearised layout L such that:

$$L = G_\theta(I), L = \{\ell_1, \ell_2, \ldots, \ell_T\} \tag{3.2}$$

where each token ℓ_t represents a module identifier, structural delimiter, or external input argument. The conditional probability of the layout is defined as:

$$p(L \mid I; \theta) = \prod_{t=1}^{T} p(\ell_t \mid \ell_{<t}, F; \theta_{\text{dec}}) \tag{3.3}$$

where $F = f_{\text{enc}}(I; \theta_{\text{enc}})$ is the encoded image representation, and $\theta = \{\theta_{\text{enc}}, \theta_{\text{dec}}\}$ denotes the model parameters. As such, the layout generator produces input-dependent modular structures. In terms of VAD, this is particularly important because different images or different video frames may require different computational strategies.

However, in practice, there are some practical constraints that should be considered. Mainly, the space of all valid layouts is extremely large, and learning layout generation from the anomaly detection error signal can be difficult and unstable, particularly with limited training data size. This difficulty arises from needing to simultaneously learn the correct layout and the correct action for each module from a shared error signal. To reduce the combinatorial complexity of layout search while retaining input-dependent modularity, we introduce a simplified adaptation of the layout generator. Rather than generating an arbitrary layout sequence, we generate a sequence of a set of modules and their input arguments. Specifically, the layout is represented as:

$$L = \{(m_1, a_1), (m_2, a_2), \ldots, (m_T, a_T)\},$$

where m_t is a module type and a_t is the corresponding module argument. At each program step t, the generator produces two distributions, a distribution over module types:

$$p(m_t \mid z, h_{t-1}),$$

and a distribution over valid arguments for the selected module:

$$p(a_t \mid m_t, z, h_{t-1}).$$

where z is a global image embedding, and h_{t-1} denotes the internal controller state summarising the program prefix. Therefore, at each step, the layout generator takes as input an image embedding z, the previously selected module token, the previously selected argument token, and outputs the next module selection from a predefined set of modules, as well as module-specific argument logits from a set of defined input argument labels. Concretely, under this representation, the simplified variation of the layout generator is:

$$p(L \mid I) = \prod_{t=1}^{T} p(m_t \mid m_{<t}, a_{<t}, F)\, p(a_t \mid m_t, m_{<t}, a_{<t}, F).$$

Given the predicted module and argument tokens, the dynamic computation graph is instantiated by selecting which module is applied at each step and which argument embedding conditions its behaviour. Importantly, module parameters are shared globally across all inputs. While the module inventory is fixed, the execution path is conditional on the input image, resulting in input-dependent computation.

3.3.2 Neural Modules

The proposed NMN is composed of a fixed set of neural modules, each aimed at performing a single operation. This design is motivated by the observation that industrial inspection can be decomposed into a sequence of interpretable steps, such as localising a relevant object or surface type, refining attention to a sub-region, and finally predicting defect locations. Each module is implemented as a small trainable neural network specialised for one simple function. The module weights are learned end-to-end during the training process, such that the module weights are shared across all instances of the same module type. Variation in behaviour across different instances of the same module type is achieved by conditioning each module on a learned embedding of input argument a such that $e(a) \in \mathbb{R}^d$. For each module type, the input argument is selected from a predefined set of valid arguments specific to that module. Therefore, the argument embedding functions as a compact conditioning vector. In this work, argument embeddings are not taken from a pre-trained word embedding model but rather trained from scratch jointly with the full system. Furthermore, all modules operate on the feature representations extracted by the backbone model $C()$:

$$F = C(I) \in \mathbb{R}^{C \times H \times W}.$$

In the remainder of this section, we describe the modules used.

START Module. The start module provides a fixed initial input to initiate execution of the modular network. It is also used as the starting token for the layout generator. In practice, the module outputs a uniform attention mask that does not restrict the flow of information to the next module.

FIND Module. The FIND module $f_{find}()$ with learnable parameters θ_{find} predicts an attention mask that identifies the spatial region corresponding to an object or surface category. The module uses a conditional convolutional network architecture. The FIND module takes as input image feature F generated by the backbone model, an attention mask M_{t-1} from the previous module and an argument embedding $e(a) \in \mathbb{R}^d$ such that $a \in A_{find}$ where A_{find} denotes the predefined set of arguments for the FIND module. The argument set should contain the names of objects and/or the material types.The module outputs a soft attention mask $M_{find} \in [0,1]^{1\times H\times W}$ where:

$$F_t = F \odot M_t,$$

$$M_{find} = M_t \odot \sigma(f_{\text{find}}(F_t,\, e(a_{\text{find}})))$$

FILTER Module. The aim of the FILTER module $f_{filter}()$ is a trainable function parameterised by learnable weights θ_{filter} that refines an existing attention mask. The module uses a conditional convolutional network architecture. The FILTER module takes as input image feature F generated by the backbone model, an attention mask M_{t-1} from the previous module and an argument embedding $e(a) \in \mathbb{R}^d$ such that $a \in A_{filter}$ where A_{filter} denotes the predefined set of arguments for the filter module. When available, the argument set includes material and object attributes, which can help to further constrain the region from the previous module. The module outputs a soft attention mask $M_{filter} \in [0,1]^{1\times H\times W}$ where:

$$F_t = F \odot M_t,$$

$$M_{filter} = M_t \odot \sigma(f_{\text{filter}}(F_t,\, e(a_{\text{filter}})))$$

REATTEND Module. The REATTEND module $f_{reattend}()$ is a trainable module with learnable parameters $\theta_{reattend}$ that apply spatial transformation to the attention mask from previous steps. The module uses a conditional convolutional network architecture. The REATTEND module takes as input image feature F generated by the backbone model, an attention mask M_{t-1} from the previous module and an argument embedding $e(a) \in \mathbb{R}^d$ where $a_{\text{reattend}} \in \{\texttt{left}, \texttt{right}, \texttt{up}, \texttt{down}\}$. The module outputs a soft attention mask $M_{reattend} \in [0,1]^{1\times H\times W}$ where:

$$F_t = F \odot M_t,$$

$$M_{reattend} = M_t \odot \sigma(f_{\text{reattend}}(F_t,\, e(a_{\text{reattend}})))$$

Prediction Modules. The Prediction modules are trainable prediction heads which act as the final modules of the architecture. They are responsible for generating model outputs in the form of labels, segmentation masks, or bounding boxes. Here, we outline two modules: PREDICT_BOX and PREDICT_SEGMENT.

The PREDICT_BOX module follows a DETR-style [24] architecture. However, unlike the vanilla DETR model, we introduce the use of defect-conditioned queries based on the argument embedding $e(a)$. The PREDICT_SEGMENT module is based on a U-Net [139] style decoder architecture. The output of PREDICT_SEGMENT is also conditioned on $e(a)$.

As such, both PREDICT_BOX and PREDICT_SEGMENT are conditioned on $e(a)$, where a denotes the set of possible anomaly types. In addition to the input argument embedding, both modules take as input the backbone feature map F and the attention mask M from the previous module. To ensure that both detection and segmentation are restricted to spatial regions highlighted as important by previous modules, the backbone image features are gated using the attention mask such that

$$F_T = F \odot M_T.$$

The PREDICT_BOX module outputs K bounding box parameters $b \in [0,1]^{K\times 4}$ and corresponding objectness scores $o \in [0,1]^K$, while the PREDICT_SEGMENT module outputs a probabilistic segmentation mask $\hat{S} \in [0,1]^{1\times H\times W}$.

Concretely, the DETECT_BOX module first converts the masked feature map into a flattened sequence of spatial tokens, which serve as the input to the query-based detection mechanism:

$$X = \text{flatten}(F_T) \in \mathbb{R}^{HW\times d}.$$

The module employs a fixed set of K learned query vectors $Q \in \mathbb{R}^{K\times d}$, conditioned on the defect embedding $e(a_{\text{box}})$, which attend over the spatial tokens via transformer cross-attention:

$$S = \text{Transformer}(Q, X) \in \mathbb{R}^{K\times d}.$$

Bounding box prediction is performed using lightweight feed-forward heads applied independently to each resulting latent vector S_k:

$$b_k = \sigma(g_{\text{box}}(S_k)).$$

On the other hand, in DETECT_SEGMENT, the masked feature map is passed to a lightweight convolutional decoder g_{seg}, which is conditioned on the defect embedding:

$$\ell_{\text{seg}} = g_{\text{seg}}(F_T, e(a_{\text{seg}})), \quad \hat{S} = \sigma(\ell_{\text{seg}}),$$

where ℓ_{seg} denotes the segmentation logits. The resulting mask may be upsampled to match the input image resolution if required by the application.

3.4 Training

The training strategy depends on the type of supervision available in the dataset. When the training annotations only provide the final inspection outcome (e.g., anomaly bounding boxes, segmentation masks, and defect labels), the model must be trained end-to-end, jointly optimising both the layout generator and the neural modules using the downstream detection/localisation loss. In this regime, the layout is treated as a latent decision process, and the generator learns to select module sequences and arguments indirectly through task performance. In addition, it is often possible to incorporate weak or auxiliary supervision to stabilise learning. For example, intermediate supervision can be provided for specific modules (e.g., supervising the FIND module using object/surface labels), or partial program supervision can be provided when approximate ground-truth layouts are available. When such layout annotations exist, training can be performed in a two-stage manner, where the layout generator is first trained with direct token-level supervision and subsequently fine-tuned jointly with the modules.

In all cases, the model relies on extracting image features using a pretrained backbone. During the initial training stages, the backbone parameters are kept frozen, and only the layout generator and neural modules are optimised, which allows for stabilised learning. After the modular reasoning components reach satisfactory performance, a second training stage is performed in which selected backbone layers are unfrozen and fine-tuned jointly with the detection and reasoning modules using a reduced learning rate.

3.4.1 Joint Training

In a joint training setting, the model is trained using only supervision from the anomaly prediction task. Given an input image I and corresponding image anomaly annotation Y, the model jointly optimises the parameters of the layout generator and the modules without access to the ground-truth layouts. The layout generator produces a discrete layout sequence $P \sim p(P \mid I; \theta_L)$, which is assembled and executed to output a prediction $\hat{y} = f_{\text{NMN}}(P, I; \theta_M)$, where θ_L and θ_M are the parameters of the layout generator and modules, respectively.

Loss $\mathcal{L}_{\text{NMN}}(\hat{y}, Y)$ is computed by comparing the model output $\hat{y}$ with the corresponding ground truth image label Y. However, since the layout P is discrete, the layout generator cannot be trained via direct backpropagation. Instead, similar to previous works [84], we formulate the layout generation process as a reinforcement learning problem and optimise θ_L using the REINFORCE algorithm [84]. The reward is defined as the negative task loss:

$$R(P, I) = -\mathcal{L}_{\text{NMN}}(f_{\text{NMN}}(P, I; \theta_M), Y) \tag{3.4}$$

The objective is to maximise the expected reward:

$$J(\theta_L) = \mathbb{E}_{P \sim p(P|I;\theta_L)} \left[R(P, I)\right] \tag{3.5}$$

The gradient of this objective is estimated using:

$$\nabla_{\theta_L} J(\theta_L) \approx \frac{1}{K} \sum_{i=1}^{K} \Big(R(P^{(i)}, I) - b \Big) \nabla_{\theta_L} \log p(P^{(i)} \mid I; \theta_L) \tag{3.6}$$

where $P^{(i)}$ denotes the i-th sampled layout, K is the number of samples, and b is a baseline to reduce variance. The module parameters θ_M are optimised via gradient descent using the loss $\mathcal{L}_{\text{NMN}}$, treating the layout P as fixed during each update.

3.4.2 Alternative Training Strategies

During training, given an input image I and corresponding anomaly annotations Y, the model needs to jointly optimise the parameters of both the layout generator and the modules. While this can be formulated as a joint training problem, it is often challenging in practice, as the model must simultaneously learn to predict appropriate layouts and train individual modules to perform task-specific computations, all from supervision that reflects only the final anomaly detection output. This makes it difficult to disentangle whether poor performance is due to layout prediction or incorrect module behaviour. To help the model, we can first train the layout generator or the modules separately before training both jointly.

If examples of optimal layout for images are available, the layout generator can be trained as a decoder-only transformer:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(y_t^* \mid I, y_{<t}^*)$$

Such that $\pi^* = \{y_1^*, ..., y_T^*\}$ is the optimal sequence of tokens the transformer should learn to predict. Alternatively, if image annotation for some or all of the modules' behaviour is available, we can first pre-train some of the modules before jointly training the modules and the layout generator.

3.5 Experiment

3.5.1 Datasets

For this work, we use three publicly available benchmark datasets: DACL10K [51], CODEBRIM [118], and DTU [144].

DACL10K contains 9,920 images and 62,327 labelled objects related to bridge inspection. The labelled objects include 12 damage classes and 6 bridge components that are essential for assessing the condition of the bridge. CODEBRIM is another dataset with a focus on bridge defects, consisting of 1,590 images. The dataset covers 6 different types of concrete defects with 8,323 labelled objects. Finally, the DTU dataset is a collection of drone images of wind turbines. The dataset consists of 13,470 images from various parts of the wind turbine. Although the dataset does not distinguish between different types of damage, it does contain labels for dirt marks that can be incorrectly associated with surface-level defects. All the datasets above comprise images collected under varying environmental conditions and include samples captured from multiple viewpoints and scales.

3.5.2 Model Implementation

We implement and train the model using the joint training formulation described above. While semantic segmentation and image classification naturally lend themselves to NMN formulation, it is more challenging to formulate object detection as an NMN model. This is because object detection involves a separate mechanism that does not map cleanly onto a modular "choose-the-expert" paradigm. Therefore, below we outline how semantic segmentation and object detection can be formulated based on NMN design

Semantic Segmentation with NMN

For semantic segmentation, we implement the NMN framework using a ResNet-101 [62] backbone as the shared feature extractor. The high-level features from the backbone can then be passed through *a find* and or *re-attend* attention module that adaptively highlights defect-relevant patterns (e.g., cracks, corrosion stains) while suppressing background noise. This reweighted representation is then fed into a U-Net–style decoder, the *predict* modules, which serve as the expert head for segmentation. Within this NMN design, the attention block and the prediction head function as a modular component that can be chosen and tuned based on the context.

Object Detection with NMN

For object detection, the NMN framework is integrated into a Faster R-CNN [137] architecture. Here, the ResNet backbone first extracts multi-scale features, which are then passed through the *find* and or *re-attend* positioned after the backbone and before the Region Proposal Network (RPN). These attention modules act as foreground priors, reweighting the feature maps so that the RPN is less likely to generate proposals in uniform background areas and more likely to focus on regions with defect-like textures. After proposals are generated and refined through ROIAlign, the features are passed into modular *predict* modules acting as the prediction heads.

3.5.3 Results

To better assess our proposed method, we compare its performance against other models' published results on the DACL10K dataset to analyse pixel-level defect detection, and on the DTU and CODEBRIM datasets to evaluate bounding-box defect detection performance. Although different backbone architectures are used, the comparison remains informative as all methods are evaluated under the same training and testing protocol.

As shown in Table 3.1, our proposed model achieves a 43.4 mean Intersection over Union (mIoU) on the DACL10K dataset, demonstrating competitive performance on the DACL10K dataset compared to previously published work. Flotzinger et al. [52] report 42.4 mIoU using an FPN-based segmentation architecture with an EfficientNet-B4 encoder and an auxiliary loss formulation. Later, this result was improved on by the same authors by using an ImageNet-pretrained MaxViT-Base backbone within an

FPN framework [49]. Furthermore, in addition to the DACL10K dataset, the authors incorporated synthetic defect data during training, leading to improved performance. In Table 3.1, we also included the best-performing submission in the DACL Challenge competition that achieved 49.8 mIoU [50]. This was achieved by ensembling six segmentation models with different architectures, training schemes, and prediction heads. We can see that our model remains competitive against other state-of-the-art architectures and more complicated training schemes using additional synthetic defect data. However, the model showed particular difficulty in accurately detecting cracks and wet spots on concrete surfaces. Small surface defects, such as cracks, may only occupy 10-30 pixels in width and height. As such, small features are poorly represented in the downsampled representation of the image from the deeper layers of the CNN backbones. In the typical CNN backbone architecture, such as ResNet 101, a feature occupying 30 pixels in the image will correspond to only one or two output feature cells, making accurate segmentation and bounding box regression less accurate. In practical terms, due to loss of resolution in deeper layers of the backbone, the model reaches a localisation ceiling, such that despite the model receiving an output error, it can not precisely align the output with the labelled data. As such, it falls short of the ensemble-based approach, which can improve accuracy by leveraging complementary representations of spatial detail and context across architectures, rather than relying on a single resolution or feature hierarchy.

On the DTU dataset, we compare our model against the reported results of several benchmark methods, including Faster R-CNN and YOLO-based models and their variations. Our model achieves 80.2 mAP@0.5, demonstrating improved performance relative to Faster R-CNN and YOLO-based benchmarks [54]. However, our model performs less accurately than the YOLO variant, YOLO-Wind [193], which achieves 83.9 mAP@0.5 using a modified YOLOv8-based detector that integrates depthwise convolutions, MobileNet-style bottlenecks, ECA attention mechanisms, and an additional detection layer to enhance multi-scale feature representation and detection accuracy. On the other hand comapred to YOLO-Wind, our approach relies on a more standard two-stage detection pipeline and a modular design, achieving competitive performance without introducing multiple detector-specific architectural modifications, which high-

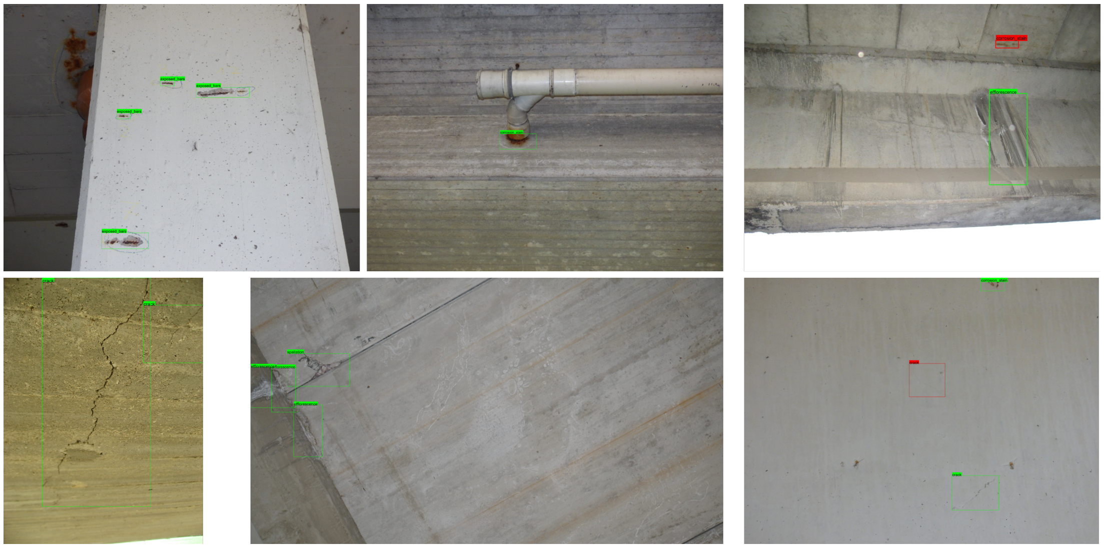


Figure 3.3: Anomaly detection results on CODEBRIM dataset. The figure shows model outputs for six images. The four images on the left demonstrate successful anomaly detection and localisation, while the two images on the right illustrate failure cases where not all anomalies are detected.

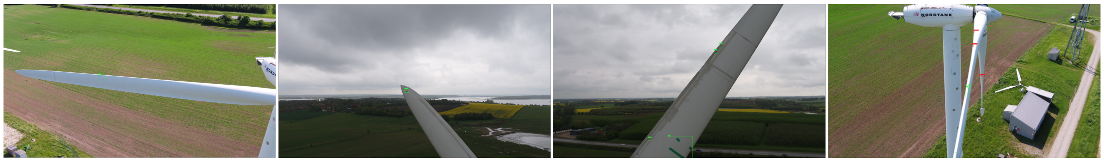


Figure 3.4: The figure shows model outputs for four images from the DTU dataset. The three images on the left demonstrate successful anomaly detection and localisation, while the image on the right illustrates failure cases when anomalies are at the edge of the object.

lights the effectiveness and generality of the proposed modular framework for aerial defect detection.

We also compare our model's performance against several other benchmarks and methods on the CODEBRIM dataset. Patel et al. report 91.2 mAP@0.5 using an improved Faster R-CNN formulation that incorporates a multi-label loss function to better handle multiple defect classes [131]. SMDD-Net reports 99.1 mAP@0.5 and introduces an attention-enhanced detection architecture that builds upon standard object detection frameworks by combining feature pyramids with attention modules to improve the localisation and classification of small-scale and low-contrast concrete surface defects, demonstrating enhanced regional feature representation compared to more basic detectors [64]. We can see that, excluding SMDD-Net, which employs a highly specialised attention-based design, our model performs better than or competitively with other baseline models and their variations. Overall, our results show that our proposed NMN model remains competitive with widely used detection frameworks.

One Factor that should be considered when comparing our results against previously published methods is the challenging aspect of training the NMN architecture. NMN are inherently challenging to train as the function of each module, and the layout generator that controls module assembly is jointly optimised from a shared error signal.

Table 3.1: DACL10K Dataset

	mIoU
Flotzinger et al., 2023 [52]	42.4
Flotzinger et al., 2024 [49]	43.37
Flotzinger et al., 2025 [50]	49.8
Ours	43.4

Table 3.2: DTU Dataset

	mAP@0.5
Faster R-CNN [54]	75.39
Foster et al., 2022 [54]	79.37
Zhanfang et al., 2025 [193]	83.9
Ours	80.2

Table 3.3: CODEBRIM Dataset

	mAP@0.5
YOLOv5 [64]	41.7
YOLOv8 [64]	59.6
Patel et al., 2021 [131]	91.2
RetinaNet [103]	88.4
SMDD-Net [64]	99.1
Ours	90.68

3.6 Conclusion

We presented the implementation of an NMN-based method for detecting image anomalies. Our proposed model demonstrates competitive performance compared to other benchmark methods. More importantly, it shows that image anomaly detection can be modelled as a compositional NMN framework. We further demonstrated that our model can select and communicate the distinct steps required for detecting various anomaly types. However, we should note that with the currently available benchmark datasets, it is challenging to test the model's capabilities thoroughly. Even when mixing multiple datasets, the images from each dataset typically contain only one type of material or object. Furthermore, current datasets lack examples where numerous types of anomalies across various materials co-occur in the same image. Finally, detecting abnormalities in the available datasets does not require complex multi-step reasoning; they can usually be resolved with only two modules. For future work, three directions can be explored:

1. Creating a more challenging dataset that reflects dynamic environments.
2. Expanding both the module architecture and the backbone to other network designs.
3. Investigating the level of granularity for each module's task to find the best balance between performance and interpretability.

In this section, we primarily focused on object-level prediction heads for surface-defect-type anomalies. However, object-level anomaly detection may also require consideration of abnormal motion, for example, in the case of individuals or vehicles that move over time. As such, in the following section, we introduce a prediction head for skeletal anomaly detection that remains aligned with the broader objectives of this thesis, particularly robustness in dynamic environments, explicit context awareness, and improved interpretability.

Chapter 4

Physics-Guided Graph Neural Networks for Skeletal Anomaly Detection

4.1 Introduction

Skeleton-based Video Anomaly Detection (SVAD) is a key task in computer vision and video surveillance, providing an interpretable and structured way to analyse human behaviour by abstracting visual data into sequences of skeletal joint positions. By focusing on the spatial and temporal evolution of these joints, SVAD systems can detect deviations from typical human motion patterns, which may signify anomalous events.

Recent work has leveraged a variety of Graph Neural Network (GNN) architectures to model the dynamics of these skeletal sequences, typically by forecasting future joint positions and flagging significant prediction errors as anomalies. Architectures such as Contextual Graph Networks (CGNs), Graph Attention Networks (GATs), and various temporal GNNs have been employed to better capture spatial relationships between joints and their temporal evolution over time. While these models have shown promise, they face notable challenges: they often suffer from noisy or missing pose estimates due to occlusions, and they tend to struggle with capturing long-range dependencies, multi-scale temporal patterns, and broader contextual cues in complex scenes.

To address these limitations, we propose a paradigm shift: instead of directly learning to predict future joint positions, we formulate SVAD as a physics-based simulation task. Specifically, we model human motion as a system governed by rigid body kinematics and dynamics, embedding these physical constraints within a learnable GNN-based architecture. Our method draws inspiration from previously proposed deep learning-based simulators that have employed Multi-Layer Perceptrons (MLPs) and GNNs to model physical interactions in domains ranging from liquid and gas flow simulations to solid object dynamics. By treating the human skeleton as a physically interacting system, our approach learns to simulate joint dynamics in a manner that respects kinematic and dynamic principles, providing physically consistent and robust motion predictions even under noisy or incomplete input data.

By treating the human skeleton as an articulated physical system, our model learns to simulate joint dynamics in a way that respects underlying physical principles. This approach enhances robustness to noisy or incomplete data and improves the generalisation and interpretability of anomaly detection. Furthermore, by grounding motion prediction in physical dynamics, the model provides richer representations of normal behaviour, allowing it to detect subtle or complex anomalies.

In this work, we aim to combine the predictive power of GNN and physics-informed modelling. By integrating GNN architectures with rigid body kinematic and dynamic modelling, we aim to introduce a novel, physically consistent framework for SVAD.

4.2 Related Work

In recent years, the use of graph-based approaches [38, 87] has gained traction in many time-series problems due to their ability to effectively capture both spatial and temporal dependencies. One important application is prediction-based skeleton video anomaly detection. Here, the human skeleton is modelled as a graph where joints are represented as nodes and edges typically represent the anatomical connections between the joints.

A seminal work introducing the concept of Spatio-Temporal Graph Convolutional Networks (ST-GCNs) for skeleton-based video anomaly detection was presented in [113]. The model used an ST-GCN encoder to learn spatial–temporal dependencies between the joints, enabling it to predict the next frame of skeleton joint positions. During inference, a high prediction error between the model prediction and the observed joint position is used as an indication of anomalies. However, as outlined in the paper, increasing the depth of ST-GCNs beyond nine layers can lead to over-smoothing and diminishing returns, causing model accuracy to drop. This limits the model's ability to learn from longer-term dependencies. At the same time, Tang et al. [154] introduced a graph-based motion prediction framework that models both spatial body structure and temporal motion dynamics. To better capture joint dependencies, the graph formulation relaxes strict adherence to anatomical connections and introduces handcrafted edges between several joints that are not anatomically connected. Together, these two works laid the foundation for prediction-based graph models in skeleton-based anomaly detection. Nonetheless, challenges such as over-smoothing and limited long-term modelling remained.

To address the limitations of earlier models, ST-GCAE-LSTM [100] combines a spatio-temporal graph convolutional autoencoder with an embedded LSTM and a dual decoder, jointly training the model to reconstruct past sequences and predict future ones. This improved the model's performance by enabling it to capture long-term dependencies better. Similarly, STEGT-AE [200] introduced a similar architecture by combining a spatio-temporal Graph-Transformer encoder and a dual-decoder autoencoder structure. To better capture long-range dependencies, multi-level skip connections are used. Furthermore, during training, the model encodes skeleton sequences using transformer-style attention over the spatio-temporal skeleton graph, which improved the model's anomaly sensitivity. While both methods improved model performance compared to basic ST-GCN-based models, they introduce higher training complexity and greater computational cost. At the same time, the deterministic modelling of future poses can lead to poor generalisation in new or dynamic scenarios [200, 100]. Furthermore, in general, using the body's predicted position as a whole can lead to missing localised anomalies if the majority of the rest of the body region is acting normally [85].

The Graph-Jigsaw Conditioned Diffusion Model (GiCiSAD) [85] tackles several of these issues. While Normal Graph predicts a single deterministic next pose, GiCiSAD predicts future skeleton poses using a Graph-Attention Forecasting module, then refines this with a diffusion model that generates multiple plausible futures instead of a single deterministic one. Furthermore, it introduced the idea of the graph-jigsaw task, in which skeleton frames are shuffled in time, and the model learns to put them back in order. This forces the model to capture temporal dependencies more effectively, improving its ability to predict future poses and detect anomalies. While GiCiSAD alleviates several limitations of earlier models, it remains computationally expensive due to diffusion sampling. GNN-based approaches have demonstrated strong performance in skeleton video anomaly detection owing to their ability to model the spatio-temporal dynamics of the human skeleton effectively. However, over-smoothing remains a challenge in capturing long-term dependencies. Several models have attempted to address this issue with more complex architectures. But doing so introduces higher computational demands, which can become problematic in more challenging environments and when training on larger datasets. Furthermore, these models do not explicitly learn the underlying dynamics and kinematics that drive human motion, limiting their ability to generalise to new scenarios or dynamic environments.

4.3 Methodology

4.3.1 Problem Formulation

We aim to formulate SVAD as a prediction problem, where, given historical joint motion features, the goal is to predict the position of each joint at the next time step. As shown in Figure 4.3.1, the model consists of three main components: (1) encoding raw joint features into a shared latent space, (2) applying a graph neural network to leverage information from other available nodes and update node states, and (3) predicting the position of the visible joints at the next time step.

Existing GNN-based approaches typically model human pose estimation as a sparsely connected graph such that joints represent graph nodes and edges correspond to anatomically accurate connections between physically adjacent joints. While this topology introduces a strong inductive bias by enforcing anatomical constraints, it can be sensitive to noise and partial observation of joints. This is because when one or more joints are missing, the sparse graph topology results in limiting the information flow and disrupting message passing between the nodes in the GNN model. This poses a significant challenge for real-world applications, where factors such as occlusion, sensor noise, motion blur, and pose estimation errors can frequently result in missing or unreliable joint observations. To overcome this limitation, we model pose estimation as a complete graph, coupled with an attention mechanism that enables adaptive, data-driven weighting of joint interactions. This allows each node to selectively incorporate information from any other available node and maintain effective information flow under partial observability instead of relying on fixed information flow based on skeletal connectivity of joints.

Motivated by this, we start by modelling the human pose estimation as a dynamic graph:

$$G_t = (V_t, E_t), \tag{4.1}$$

where

$$V_t = \{v_1, \dots, v_n\} \tag{4.2}$$

is the set of nodes observable at time t, such that each node represents a joint, and E_t is the set of edges that capture the relationships between each pair of joints. To ensure robustness under partial observability, we adopt a complete graph structure such that:

$$E_t = \{(v_i, v_j) \mid v_i, v_j \in V_t,\ i \neq j\}. \tag{4.3}$$

This allows for the information to propagate between any pair of joints, even if the intermediate joints and edges that connect them in the physical world are not observable. We also define each node $v_i \in V_t$ feature vector $x_i^{(t)}$ by the temporal feature vector associated with the corresponding joint, such that the node feature:

$$x_i^{(t)} \in \mathbb{R}^{d_i}, \tag{4.4}$$

Represents historical motion information (e.g., position, velocity, acceleration), image context, and class label for joint i at time t as defined in Eq. (4.10). When observing human poses across multiple frames, information for certain joints may be missing at some time steps. To address this, we first encode the raw input features into a shared latent space using a learnable function:

$$h_i^{(t)} = \text{Encoder}(x_i^{(t)}) \tag{4.5}$$

This encoding step ensures that all nodes are projected into a common representation space suitable for downstream graph-based computation. The encoded representations are then updated via a Graph Attention Network (GAT), which allows each node to aggregate information from all other joints in the graph, weighted by learned attention coefficients:

$$\hat{h}_i^{(t)} = \text{GAT}\big(\{h_j^{(t)}\}_{j\in\mathcal{J}_t}\big) \tag{4.6}$$

such that $\mathcal{J}_t$ denotes the set of joints at time t. To compute the joint positions at the next time step, we adopt a rigid-body kinematic formulation.

$$x^{(t+1)} = x^{(t)} + v^{(t)}\Delta t + \tfrac{1}{2}a^{(t)}\Delta t^2, \tag{4.7}$$

which requires joint position x, velocity v, and acceleration a at time t. While joint positions are observable, joint velocity can be estimated as a first-order differential approximation using backward finite differences based on past observations, and remains relatively accurate under moderate noise and partial observability. However, joint acceleration cannot be accurately estimated, as it requires position values from three time steps, $t-1$, t, and $t+1$. While acceleration can technically be estimated at time t using backward finite differences, such estimates are highly sensitive to measurement noise and amplify pose estimation errors due to second-order differentiation. As such, we consider acceleration a hidden state at time t, which the model is required to predict.

To this end, node embeddings from GAT are used to predict the change in acceleration for each joint at time t:

$$\Delta a_i^{(t)} = f_{\text{accel}}(\hat{h}_i^{(t)}). \tag{4.8}$$

Finally, the predicted acceleration is integrated into a rigid-body dynamics formulation to estimate the joint positions at the next time step. Specifically, we use the predicted change in acceleration to update the velocity and position of each joint, allowing us to

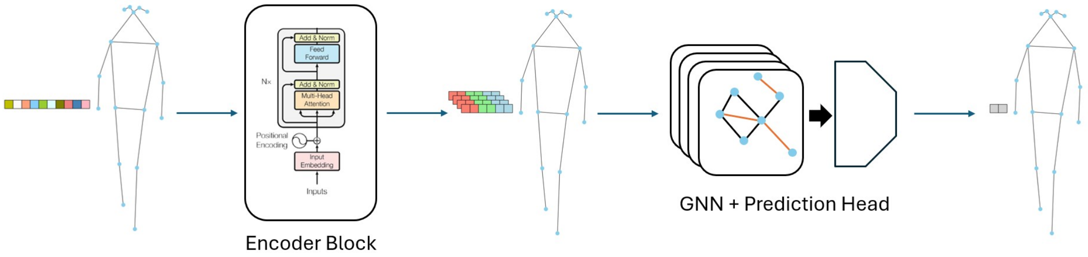


Figure 4.1: The pipeline of our proposed model. After extracting kinematic joint data from the last n consecutive frames, it is encoded using the transformer encoder block. We then model the human skeleton representation as a complete graph. The graph is passed through a multi-head GAT network. Finally, each joint representation is used to estimate the acceleration at the current time step t and the position at time step t+1

forecast the expected pose at time $t + 1$. This physically grounded formulation ensures that the predicted motion remains consistent with natural kinematic constraints.

In the remainder of this section, we provide a detailed description of different sections of the model. We begin by discussing the feature encoding process used to project raw joint observations into a shared latent space. We then describe how Graph Attention Networks are employed to model spatial relationships among joints and update node embeddings. Finally, we outline the dynamics-based prediction module used to estimate future joint positions via acceleration forecasting.

4.3.2 Node Encoder

A crucial aspect of the model is encoding relevant information to accurately predict future joint positions. To this end, we aim to learn a function

$$\phi_{\text{node}} : \mathbb{R}^{d_x} \to \mathbb{R}^{d_h} \tag{4.9}$$

that maps each joint's feature vector x_i into a latent node representation. We begin by outlining how we encode the time series of motion features for a single joint at time

t. Each node $v_i \in V_t$ is associated with a time series of motion features $x_i^{(t)} \in \mathbb{R}^{n \times f}$, representing the joint's recent dynamics over a sliding window of length n. Concretely, each node's feature vector at time t is defined as:

$$x_t = [p_{t-n}, \ldots, p_t,\, v_{t-n}, \ldots, v_t,\, a_{t-n}, \ldots, a_{t-1}] \tag{4.10}$$

where p, v, and a denote the position, velocity, and acceleration vectors, respectively. However, as mentioned before, due to occlusion or model failure to localise all visible joints, some of these measurements may be unobservable. Therefore, for each time step i in the node's history, we initially define a token k as:

$$k_i = [p_i,\, v_i,\, a_i] \tag{4.11}$$

Such that:

$$x_t = [k_{t-n}, \ldots, k_t] \tag{4.12}$$

To handle missing data, we introduce a binary mask for each component:

$$m_i = [m_i^p,\, m_i^v,\, m_i^a], \quad \text{with } m_i^p, m_i^v, m_i^a \in \{0, 1\} \tag{4.13}$$

To maintain a consistent input dimensionality, missing features are imputed using learnable embeddings:

$$\tilde{v}_i = m_i^v \cdot v_i + (1 - m_i^v) \cdot e_v, \quad \tilde{a}_i = m_i^a \cdot a_i + (1 - m_i^a) \cdot e_a \tag{4.14}$$

where $e_v \in \mathbb{R}^{d_v}$ and $e_a \in \mathbb{R}^{d_a}$ are learnable vectors corresponding to velocity and acceleration, respectively, the final augmented token becomes:

$$\tilde{k}_i = [p_i,\, \tilde{v}_i,\, \tilde{a}_i,\, m_i] \tag{4.15}$$

4.3.3 Token Embedding and Positional Encoding

Here, we adopt a standard formulation for [CLS] [43] and position encoding [159]. Each augmented token $\tilde{x}_i \in \mathbb{R}^d$ is linearly projected into a latent space:

$$z_i = W\tilde{k}_i + b, \quad z_i \in \mathbb{R}^{d_e} \tag{4.16}$$

where $W \in \mathbb{R}^{d_e \times d}$ and $b \in \mathbb{R}^{d_e}$. To capture temporal information, we add positional encodings $PE(i)$ to each token:

$$\tilde{z}_i = z_i + PE(i) \tag{4.17}$$

To facilitate sequence-level representation, we prepend a special learnable [CLS] token $\tilde{z}_{\text{CLS}} \in \mathbb{R}^{d_e}$ to the input sequence. This token is intended to aggregate information from the entire sequence and does not receive a positional encoding.

The full input to the transformer becomes:

$$Z = [\tilde{z}_{\text{CLS}}, \, \tilde{z}_1, \, \tilde{z}_2, \, \dots, \, \tilde{z}_n] \tag{4.18}$$

4.3.4 Transformer Encoder with Attention Masking

The sequence Z is processed by a stack of L transformer encoder layers [159]. Each layer employs self-attention to compute contextualised representations of all tokens. For each token $z_i \in Z$, we compute:

$$q_i = W_Q z_i, \quad k_i = W_K z_i, \quad v_i = W_V z_i \tag{4.19}$$

with $W_Q, W_K, W_V \in \mathbb{R}^{d_k \times d_e}$. The attention score between tokens i and j is given by:

$$e_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}} \tag{4.20}$$

The resulting attention outputs are passed through a feed-forward network with residual connections and layer normalisation, following standard transformer design.

4.3.5 Node Features Embedding

We adopt the [CLS] token [43] strategy to obtain a summary embedding for each node. After passing the token sequence through L transformer layers, the final representation of the [CLS] token is used as a compact summary of the node's motion history:

$$h_{\text{motion}} = \tilde{z}_{\text{CLS}}^{(L)} \tag{4.21}$$

We aim to integrate visual and semantic information into the node embedding. To this end, we enrich the motion-based node embedding with both image context and joint type information. These components are concatenated to form a unified node representation:

$$h_i^{(t)} = \text{Encoder}\left(\left[h_{\text{motion}} \,\|\, x_{\text{image}} \,\|\, x_i^{\text{type}}\right]\right) \tag{4.22}$$

such that x_{image} is the [CLS] token representation from the ViT encoder of the scene and x_i^{type} is the encoding of the joint class label. Therefore, the final embedding $h_i^{(t)}$ fuses temporal dynamics, visual scene context, and anatomical semantics, enabling more informed reasoning during graph-based message passing.

4.3.6 Graph Attention Network

To model spatial dependencies and heterogeneous interactions among joints, we employ a multi-layer Graph Attention Network (GAT). At each layer l, the representation of a node is updated by aggregating information from neighbouring nodes, weighted by learned attention coefficients that reflect their contextual relevance.

Given input node features $h_i^{(l)}$, each node is first projected into a shared latent space using a learnable transformation, such that:

$$\tilde{h}_i^{(l)} = W^{(l)} h_i^{(l)} \tag{4.23}$$

Pairwise attention scores $\alpha_{ij}^{(l)}$ are computed and normalised using softmax based on the similarity of node features. Each node is then updated by taking the weighted sum of neighbour features such that:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i) \cup i} \alpha_{ij}^{(l)} \tilde{h}_j^{(l)}\right) \tag{4.24}$$

where σ is a non-linear activation function, and $\tilde{h}_j^{(l)}$ is the transformed representation of node j as shown in Eq. (4.23). To improve expressiveness and training stability, we use multi-head attention. The outputs from K independent attention heads are concatenated:

$$h_i^{(l+1)} = \|_{k=1}^{K} \sigma \left(\sum_{j \in \mathcal{N}(i) \cup i} \alpha_{ij}^{(l,k)} \tilde{h}_j^{(l,k)} \right) \tag{4.25}$$

such that:

$$\tilde{h}_j^{(l,k)} = W^{(l,k)} h_j^{(l)}. \tag{4.26}$$

This attention mechanism enables each joint to selectively attend to other joints based on contextual relevance, while remaining robust to noisy or missing observations.

4.3.7 Decoder

After L layers of GAT, we obtain the final latent representations $h_i^{(L)}$ for each joint. These embeddings are then decoded to predict the motion states for each node. We employ a separate decoder for predicting acceleration at time t and position at time $t+1$. Both decoders $\psi_{\text{dec}} : \mathbb{R}^{d_h^{(L)}} \to \mathbb{R}^{d_y}$ share the same architecture design and are implemented as a multi-layer perceptron (MLP), applied independently to each node:

$$\hat{y}_{pos} = \psi_{\text{pos}}(h_i^{(L)}) \tag{4.27}$$

$$\hat{y}_{acc} = \psi_{\text{acc}}(h_i^{(L)}) \tag{4.28}$$

where $\hat{y}_{pos}$ and $\hat{y}_{acc}$ represent the models' predicted acceleration for time t and position at time $t+1$.

4.3.8 Loss Function

In this work, the model aims to predict acceleration at the current time step and position at the next time step. Therefore, the loss function can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{acc}} + \mathcal{L}_{pos} \tag{4.29}$$

such that:

$$\mathcal{L}_{\text{acc}} = \sum_{i=1}^{N} \|\hat{a}_t^{(i)} - a_t^{(i)}\|_2^2 \tag{4.30}$$

$$\mathcal{L}_{pos} = \lambda \sum_{i=1}^{N} \|\hat{p}_{t+1}^{(i)} - p_{t+1}^{(i)}\|_2^2 \tag{4.31}$$

4.4 Experimental Setup

4.4.1 Datasets

We evaluate our proposed model on the Human-Related ShanghaiTech (HR-SHT) dataset [107]. The HR-SHT dataset is a subset of the ShanghaiTech Campus dataset [107], focusing on human skeletal and action anomaly detection. As such, six videos containing anomalies irrelevant to human motion are removed.

HR-SHT is a widely used benchmark for semi-supervised skeleton-based video anomaly detection. The HR-SHT dataset comprises of 437 videos recorded across 13 camera scenes. It is broken down into 330 anomaly-free training videos and 107 testing videos containing various types of anomalies. In total, the test set includes 130 abnormal events captured under complex lighting conditions and diverse camera viewpoints. The test set also provides frame-level and pixel-level annotations for evaluation.

In this work, we utilise the dataset under a semi-supervised learning paradigm. The model is trained using only anomaly-free videos, without access to anomaly annotations, to learn patterns of normal human motion. During testing, the model is evaluated using the test-set annotations.

4.4.2 Pose Estimation and Tracking

For our model, each joint input state vector requires a sequence of the previous n joint positions (see Eq. (4.10)). As such, we need to track each person and their associated pose estimations across the video. To achieve this, we utilise multi-object tracking to maintain consistent identities across video frames and apply pose estimation for

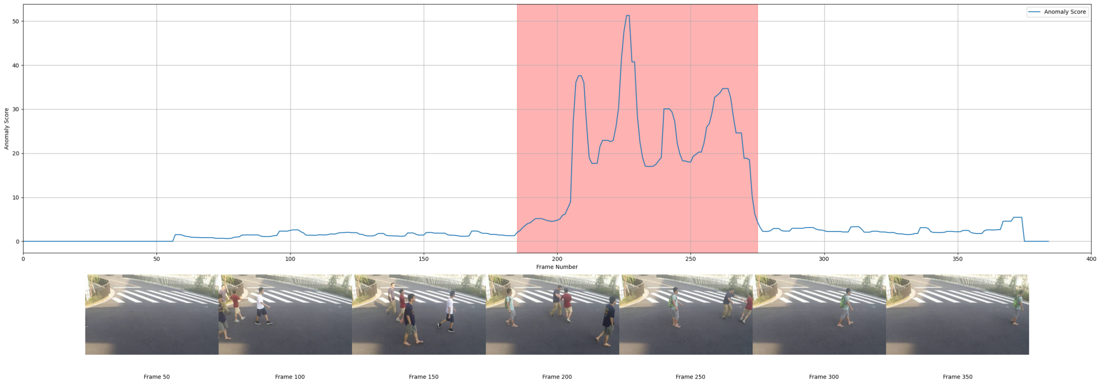


Figure 4.2: Example of anomaly detection result on ShanghaiTech dataset

each tracked person in the frame. For detecting individuals in each frame, we employ Ultralytics YOLOv8 object detector [83].

For generating pose estimation data, we utilise the PyTorch implementation [177] of the ViTPose model [178], as the ViTPose model has demonstrated strong performance in crowded scenes due to its global attention mechanism and robustness to occlusions [178]. Furthermore, for tracking, we employ the BoT-SORT algorithm [4] using the online implementation in [22].

The BoT-SORT algorithm combines appearance-based re-identification with motion modelling, enabling robust identity association under frequent occlusions and high inter-person interactions [4]. This property is well-suited for real-world crowded environments, where individuals in video footage often overlap or may be temporarily occluded before reappearing.

4.5 Results

In this section, we evaluate the performance of our proposed physics-guided GNN-based anomaly detection model. The model aims to learn the underlying kinematic and dynamic structure of human motion. Our proposed method adopts a predictive approach such that, given the joint features over a temporal window of size n, the model predicts the joint accelerations at time t and joint positions at time t+*1*. Anomalies are detected by measuring the divergence between these predicted values and the actual observed values, leveraging the assumption that anomalous motions are more complex to predict under learned physical dynamics.

Table 4.1 presents the comparative performance of our model against a range of recent state-of-the-art methods on two standard datasets: ShanghaiTech (SHT), Human-related ShanghaiTech (HR-SHT). Our model achieves competitive performance, particularly on the SHT and HR-SHT datasets, where it surpasses or matches several baseline methods such as CT-D2GAN, ROADMAP, and PoseCAVE.

Methods	**SHT**	**HR-SHT**
MTP [138]	76.03	77.04
MPED-RNN [117]	73.4	75.4
Jigsaw [164]	84.3	-
CT-D2GAN [46]	77.7	-
HSTGCNN [192]	81.80	83.40
PoseWatch-H [125]	85.75	87.23
MoPRL [188]	83.35	84.34
STGformer [165]	82.9	86.97
MSTA-GCN [30]	75.9	76.8
PoseCAVE [79]	74.9	75.7
Ours	77.29	79.18

Table 4.1: Performance comparison of different methods on SHT, HR-SHT

Our model demonstrates a strong capacity for generalisation by learning the fundamental physical dynamics of human motion. This enables effective anomaly detection without relying on dataset-specific features or action-specific patterns. Additionally, due to its physics-informed design, the model exhibits notable robustness in handling missing or occluded joint data, making it practical for real-world scenarios where imperfect pose estimations are standard.

In this section, we also compare the results from our model against various other prediction-based models, using the results reported in the original publications. It should be noted that the reported results are obtained using different image feature

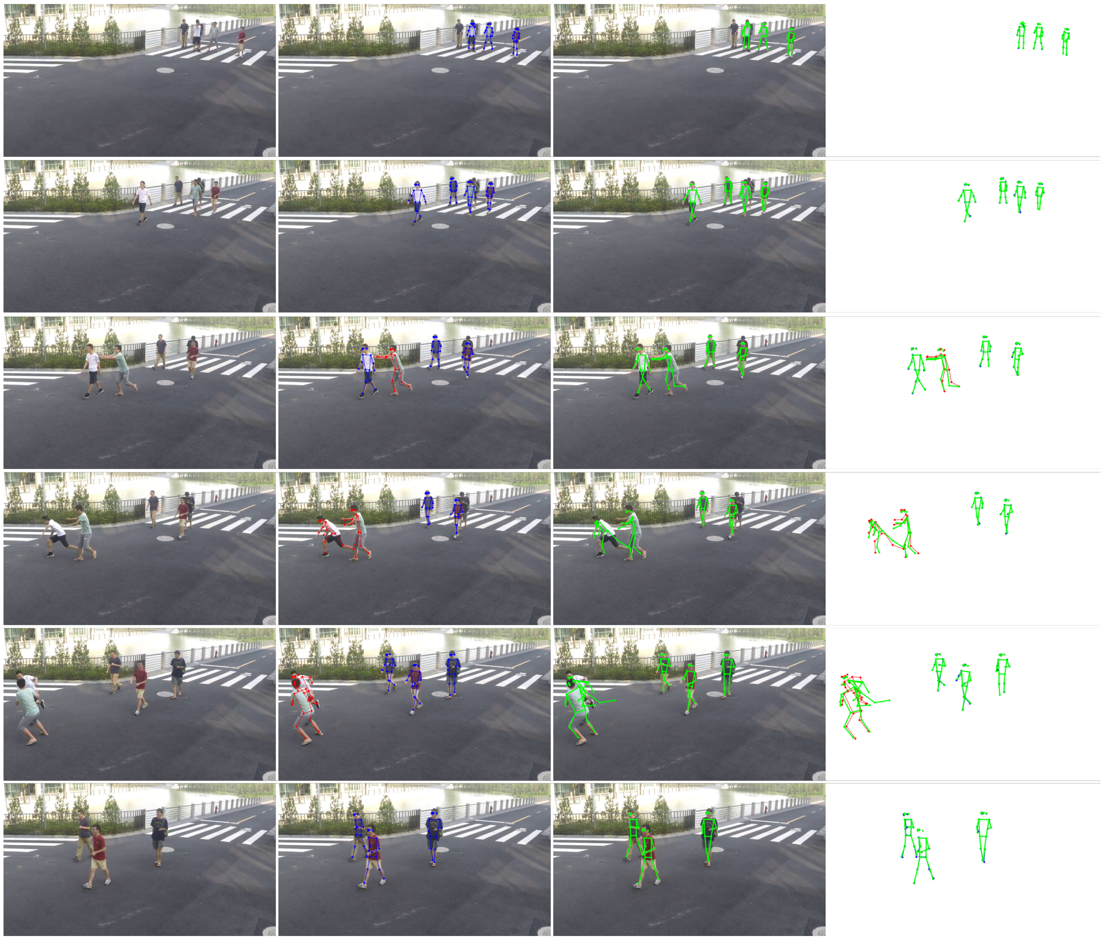


Figure 4.3: From left to right, each column shows the input image, the ground-truth pose estimation, the pose predicted by our model, and the overlap between the ground truth and the predicted pose. Results are shown for several consecutive frames from a video sequence. During anomalous motion, a clear discrepancy can be observed between the model's prediction and the ground-truth pose estimation.

extraction backbones, pose estimators, and tracking strategies. As such, these comparisons reflect end-to-end performance rather than a strictly controlled evaluation under identical preprocessing. However, the reported results provide a useful reference for comparing our model against various baseline and state-of-the-art (SOTA) prediction-based VAD models. Here, we mainly evaluate against models that operate on skeletal pose data; however, for completeness, we also include predictive methods that operate directly on RGB video data.

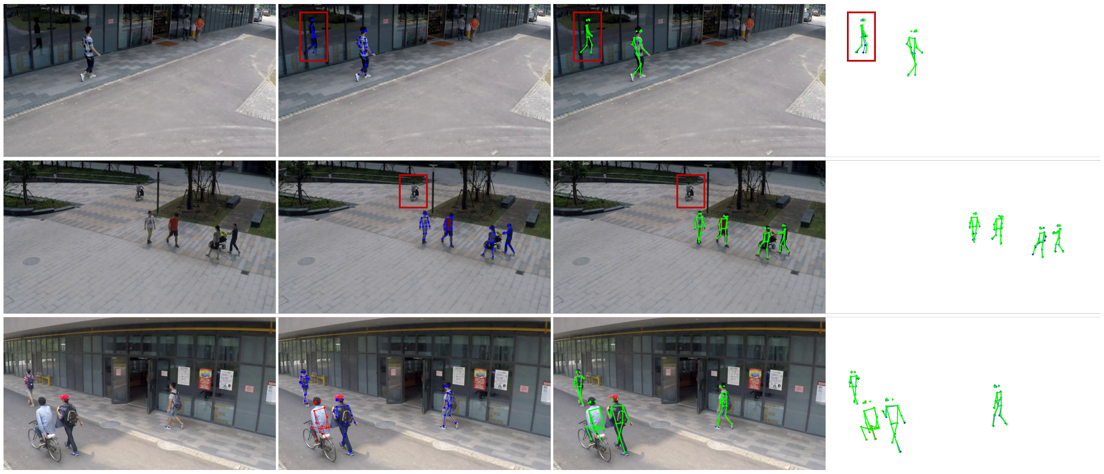


Figure 4.4: Failure and edge cases. From left to right, each column shows the image, ground-truth pose estimation, pose prediction, overlap between the ground truth and the predicted. **First row:** False detections caused by reflections on glass surfaces. **Second row:** Missed anomaly due to the failure of pose estimation and object detection to detect the person. **Third row:** Edge-motion cases where the proposed model correctly matches predictions with ground truth for previously unseen activities, particularly slow motions such as coasting on a bicycle.

RGB-based methods such as Jigsaw [164] and CT-D2GAN [46] flag anomalies based on failures in future-frame prediction and spatio-temporal consistency. This class of models typically utilises 2D or 3D CNN-based spatiotemporal architectures or transformers to model appearance and motion cues directly from pixels. Compared to skeleton-based methods, RGB-based methods are more general and can detect anomalies beyond those related to human motion. At the same time, since they process the full image, they are often more sensitive to appearance changes and background variation and provide weaker interpretability for abnormal behaviour. We observe that our model is competitive against RGB prediction baselines.

Skeleton-based models operate on sequence or individual pose estimations for different individuals in the video. Similar to our model, pose sequences are typically extracted via a combination of tracking methods (e.g., PoseFlow, SORT/DeepSORT) and pose estimation models (e.g., OpenPose, HRNet, AlphaPose). We compare our results against three distinct skeleton-based architectural families, namely recurrent model such as (MPED-RNN), spatio-temporal GNNs (MSTA-GCN), and transformer-based approaches (MoPRL, STGformer, and PoseWatch-H). Quantitatively, our approach performs better than or is competitive against most skeleton baselines such as MPED-RNN [117], PoseCVAE [79], and MSTA-GCN [30]. However, its performance remains below transformer-based models such as MoPRL [188], STGformer [165], and PoseWatch-H [125]. Although our method underperforms, it should be noted that these approaches typically rely on higher-capacity architectures that may be more difficult to support in real-world environments. Furthermore, our model provides a more interpretable anomaly signal grounded in motion dynamics.

Overall, these results indicate that our physics-inspired pose-graph formulation is competitive with SOTA and benchmark prediction-based models, while retaining the interpretability advantages of human-centric modelling.

4.6 Conclusion

We have presented a novel skeleton-based video anomaly detection framework that explicitly models human motion as a physical system. This physics-informed representation enables the model to capture the governing kinematic and dynamic formulation of human motion, allowing it to generalise effectively to unseen scenarios. Unlike prior GNN-based SVAD methods that employ expensive spatio-temporal graph convolutions, thereby limiting model depth and expressivity, this method enables the model to utilise a longer window of historical motion information.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we set out to address some of the persistent limitations in VAD, namely poor reliability in dynamic environments, limited explainability, and a lack of context awareness.

As discussed in Chapter 2, in recent years many deep learning models have shown strong performance on standard benchmark datasets under controlled environments. VAD typically falls under two paradigms: learning to recognise anomalous features, or learning a representation of normal features and using this information to recognise abnormality. Both formulations generally struggle in dynamic environments. In the latter, the distinction between anomalous and non-anomalous change can become ambiguous and context-dependent. In the former, a lack of context awareness makes learned representations of anomalous or normal features difficult to generalise and correctly apply. Furthermore, model interpretability and explainability remain open and largely unsolved problems. While some solutions have been proposed to address these challenges individually, they are typically variations of, or extensions to, existing architectures; as such, the underlying limitations largely remain unresolved. These limitations motivated the central argument of this thesis: that VAD models should move beyond static, monolithic formulations and instead be viewed as a compositional problem.

Accordingly, a key contribution of this thesis is the reframing of anomaly detection as a compositional reasoning problem rather than a monolithic classification or reconstruction task. This reframing was motivated by considering how expert individuals analyse and inspect images or videos for anomalies. When analysing an image or video sequence, experts do not treat all regions equally. Instead, they first identify objects, surfaces, or areas of interest that require further analysis, each of which can be considered a task to be completed. Following this, different inspection and analysis strategies are applied to complete each task. This process is therefore inherently context-aware and compositional, and requires explicit reasoning. In response, this thesis introduced a novel compositional VAD model based on an NMN architecture, designed to emulate expert inspection behaviour during anomaly analysis.

To this end, Chapter 3 introduced the formulation of VAD as a compositional problem and proposed an NMN-based architecture for anomaly detection. Instead of relying on a single monolithic model, the proposed method dynamically assembles a model based on information captured from the input image. This represents a second key novelty of the approach. NMNs typically require an external input query to assemble a model; in contrast, this thesis demonstrated that the image itself can be used as the input query to guide dynamic model assembly. As such, the proposed model is able to first identify regions that require further analysis and then dynamically assemble a model to perform the anomaly detection task. Furthermore, the VAD task can be formulated as a compositional problem in which a shared set of modules is dynamically assembled into an image-dependent model. This allows the model to adapt its internal computation to image content, enabling different regions to be processed using different combinations of learned components. This was achieved via end-to-end training of a layout generator model that outputs the mapping of the dynamically assembled model. The proposed approach was evaluated against previously published results on benchmark datasets for object detection and segmentation, and was shown to be competitive with existing methods.

Finally, this compositional approach provides improved interpretability and context-aware decision-making. The dynamic assembly of models from a set of predefined expert modules, together with the selection of regions for investigation, introduces intermediate decisions that can be inspected and analysed. This provides a form of explanation that is more closely aligned with expert reasoning. While this does not fully resolve the challenge of explainability in anomaly detection, it offers a clearer decomposition of the decision-making process.

In Chapter 3, our modular neural network model was aimed at detecting object-level anomalies, such as surface defects. However, an important aspect of the anomaly detection problem is detecting anomalies in dynamic objects that move or change over time. For this reason, in Chapter 4, we focused on pose estimation anomaly detection, which can be used as an expert module within the MNM architecture introduced in Chapter 3. Consistent with the earlier chapters, this approach aimed to address the same core challenges identified in the preceding discussion.

To this end, we proposed a novel physics-based pose anomaly detection model based on a graph neural network architecture. Similar to previous work, we use a prediction-based error to detect anomalies. Our model aims to predict the position of each visible joint at the next time step by first predicting the acceleration of each joint at the current time step and then using kinematic equations to estimate the joint positions at the next time step. Furthermore, to prevent the need for the use of spatiotemporal, which can suffer from over-smoothing and limited network depth, we embed each joint's historical features and image context individually. Finally, by using an attention-weighted complete graph, we ensure that the model can leverage information from the most informative joints to improve prediction, without being constrained by predefined human skeletal connections. We demonstrated that our model's performance is better than or comparable to many alternative architectural designs, and that it is only outperformed by significantly more parameter-heavy transformer-based networks.

In summary, in this thesis, we presented several conclusions. First, VAD can benefit from being formulated as a context-dependent and structured reasoning problem rather than a single global modelling task. Second, compositionality combined with an NMN architecture provides a practical framework for VAD in dynamic environments. Furthermore, compositionality naturally allows for less opaque and more interpretable models, which are essential for many real-world applications. Finally, deep learning models with built-in physical priors can provide strong performance, improved generalisation, and allow for better interpretability.

Despite these contributions, several limitations remain. The main two limitations are related to the training of the layout generator for the NMN architecture and the choice of modules. The proposed modular framework relies on the correct selection of regions to investigate, as well as the correct selection and execution order of modules. As the number of modules increases, the training space becomes exponentially larger, making end-to-end training of the layout generator and modules difficult to achieve. Furthermore, the modular framework introduces additional design choices related to the number, type, and complexity of modules used. Finally, it should be acknowledged that currently available benchmark datasets, even when combined, cannot fully represent the diversity observed in real-world applications. Based on these limitations, the following section outlines directions for future work that could build on this thesis.

5.2 Future Works

In this section, we recommend several directions for future work.

- **Dataset limitation:** Existing datasets are largely collected under controlled environments with limited variation, which do not adequately represent the conditions encountered in real dynamic inspection settings. Furthermore, most datasets contain anomalies from a single object or material type per image or sequence. For accurate evaluation of VAD models in dynamic environments, new datasets should capture: 1) real-world variation, 2) context-aware reasoning, and 3) multiple anomaly types from different object classes in the same image.

- **Optimisation of NMN layout generators:** Layout generators enable compositional reasoning in neural modular networks; however, their optimisation remains challenging due to the discrete and often non-differentiable nature of structure learning. As the number of available modules increases, the layout search space grows rapidly, making efficient exploration difficult. In addition, propagating training signals from module-level errors to layout-generation decisions is non-trivial and can result in unstable learning dynamics. Future work should examine different deep learning architectures and training paradigms to improve layout generators' performance.

- **Modules design considerations:** The complexity of the tasks that individual modules can perform, as well as the number of module types used, can significantly influence both detection accuracy and interpretability. However, increasing module diversity also makes layout learning more challenging and reduces computational efficiency, often requiring more training data and parameters. Future research should explore the trade-offs between model expressiveness and efficiency when increasing the number and complexity of modules.

- **Multimodal fusion and graph interaction for skeletal anomaly detection:** In our proposed method in Chapter 4, the RGB appearance of individuals was only used to extract individuals' pose estimations. Future work could explore combining skeletal graph representations with other modalities, such as the RGB appearance of different sections of individuals in the image. Furthermore, our model treats each person's motion as a stand alone indivudal withouut considering interaction with other people, objects or the environment. In future work, these interactions should be investigated.

- **More expressive physics-inspired modelling:** The current model relies on relatively simple kinematic equations to anchor predictions in physically plausible motion. Future work should consider more sophisticated kinematic and dynamic motion formulation priors, as well as additional physical constraints, such as improved modelling and formulation of motion, bone length consistency, joint angle limits, and angular velocities.

Bibliography

[1] Copyright. In G. Robert Odette and Steven J. Zinkle, editors, *Structural Alloys for Nuclear Energy Applications*, page iv. Elsevier, Boston, 2019.

[2] Front matter. In G. Robert Odette and Steven J. Zinkle, editors, *Structural Alloys for Nuclear Energy Applications*, pages i–ii. Elsevier, Boston, 2019.

[3] Preface. In G. Robert Odette and Steven J. Zinkle, editors, *Structural Alloys for Nuclear Energy Applications*, pages xvii–xviii. Elsevier, Boston, 2019.

[4] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[5] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[6] Haleh Akrami, Anand A. Joshi, Jian Li, Sergül Aydöre, and Richard M. Leahy. A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, 238:107886, 2022.

[7] M. Ershadul Alam, Todd R. Allen, Jeremy T. Busby, Jia-Chao Chen, Nicholas J. Cunningham, Yann De Carlan, Colin A. English, Concetta Fazio, Anand Garde, Malcolm Griffiths, David T. Hoelzer, David E. Holcomb, Jonathan M. Hyde, Richard J. Kurtz, Gene E. Lucas, Stuart A. Maloy, Randy Nanstad, Ken Natesan, Andre A.N. Nemith, G. Robert Odette, Michael Rieth, Lance L. Snead, Philippe Spätig, Tiberiu Stan, Lizhen Tan, Hiroyasu Tanigawa, Shigeharu Ukai, Gary S. Was, Tim Williams, Brian D. Wirth, Suresh Yagnik, Pascal Yvon, and

Steven J. Zinkle. Contributors. In G. Robert Odette and Steven J. Zinkle, editors, *Structural Alloys for Nuclear Energy Applications*, pages xv–xvi. Elsevier, Boston, 2019.

[8] Ahad Alloqmani, Yoosef B Abushark, Asif Irshad Khan, and Fawaz Alsolami. Deep learning based anomaly detection in images: insights, challenges and recommendations. *International Journal of Advanced Computer Science and Applications*, 12(4), 2021.

[9] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] S. Anoopa and A. Salim. Survey on anomaly detection in surveillance videos. *Materials Today: Proceedings*, 58:162–167, 2022. International Conference on Artificial Intelligence Energy Systems.

[11] Bilal Arain, Chris McCool, Paul Rigby, Daniel Cagara, and Matthew Dunbabin. Improving underwater obstacle detection using semantic image segmentation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9271–9277, 2019.

[12] Manoj Kumar Balwant, Shivendu Mishra, and Rajiv Misra. Madvit: A vision transformer-based multilayer distillation framework for explainable anomaly detection in aerial imagery. *Expert Systems with Applications*, 296:129166, 2026.

[13] Antonios Banos, Jim Hayman, Tom Wallace-Smith, Benjamin Bird, Barry Lennox, and Thomas B Scott. An assessment of contamination pickup on ground robotic vehicles for nuclear surveying application. *Journal of Radiological Protection*, 41(2):179, 2021.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[15] Alexander Barzilov and Monia Kazemeini. Unmanned aerial system integrated sensor for remote gamma and neutron monitoring. *Sensors*, 20(19):5529, 2020.

[16] Alexander Bauer, Shinichi Nakajima, and Klaus-Robert Müller. Self-supervised autoencoders for visual anomaly detection. *Mathematics*, 12(24), 2024.

[17] Christoph Baur, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In *International conference on medical image computing and computer-assisted intervention*, pages 718–727. Springer, 2020.

[18] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

[19] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[21] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021.

[22] Mikel Broström. Boxmot: Pluggable sota multi-object tracking modules. AGPL-3.0 license; CITATION.cff included.

[23] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6511–6523, 2023.

[24] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander

Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[25] Aneesh N. Chand, Vahid Mahalleh, Tareq Aziz, Arif Rahman, Abdullah Mohammed, and Wan Zeid. Small scale localized maintenance of industrial infrastructure using autonomous uavs. In *2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 168–175, 2021.

[26] Kushal Chauhan, Barath Mohan U, Pradeep Shenoy, Manish Gupta, and Devarajan Sridharan. Robust outlier detection by de-biasing vae likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9881–9890, June 2022.

[27] NF Chen, Zhiyuan Du, and Khin Hua Ng. Scene graphs for interpretable video anomaly classification. In *Conference on Neural Information Processing Systems Workshop on Visually Grounded Interaction and Language*, 2018.

[28] Shi Chen and Qi Zhao. Divide and conquer: Answering questions with object factorization and compositional reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6736–6745, June 2023.

[29] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 655–664, January 2021.

[30] Xiaoyu Chen, Shichao Kan, Fanghui Zhang, Yigang Cen, Linna Zhang, and Damin Zhang. Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection. *Journal of visual communication and image representation*, 90:103707, 2023.

[31] Civaux-communication. Salle des machines de la centrale de civaux, jan 2011. Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

[32] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[33] Nuclear Regulatory Commission. Nrc chair christopher hanson at peach bottom nuclear power plant, dec 2022. CC BY 2.0.

[34] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. Patch-wise auto-encoder for visual anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 870–874, 2023.

[35] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pages 475–489. Springer, 2021.

[36] David Dehaene and Pierre Eline. Anomaly localization by modeling perceptual features, 2020.

[37] Stanislas Dehaene, Fosca Al Roumi, Yair Lakretz, Samuel Planton, and Mathias Sablé-Meyer. Symbols and mental programs: a hypothesis about human singularity. *Trends in Cognitive Sciences*, 26(9):751–766, 2022.

[38] Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4027–4035, 2021.

[39] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022.

[40] Thadeu L. B. Dias, Eduardo A. B. Silva, Sergio L. Netto, and Lucas A. Thomaz. Change detection in moving-camera videos using a shift-invariant dissimilarity metric. In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6, 2022.

[41] Keval Doshi and Yasin Yilmaz. Towards interpretable video anomaly detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2654–2663, 2023.

[42] Keval Doshi and Yasin Yilmaz. Towards interpretable video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2655–2664, 2023.

[43] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[44] Eric Elmoznino, Thomas Jiralerspong, Yoshua Bengio, and Guillaume Lajoie. A complexity-based theory of compositionality. *arXiv preprint arXiv:2410.14817*, 2024.

[45] Homa Fashandi. Neural module networks: A review. *Neurocomputing*, 552:126518, 2023.

[46] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5546–5554, 2021.

[47] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[48] FirstEnergy. Perry nuclear power plant returns to service following refueling and maintenance outage, april 2017. CC BY-ND 2.0.

[49] Johannes Flotzinger, Fabian Deuser, Achref Jaziri, Heiko Neumann, Norbert Oswald, Visvanathan Ramesh, and Thomas Braml. synth-dacl: Does synthetic defect data enhance segmentation accuracy and robustness for real-world bridge

inspections? In *DAGM German Conference on Pattern Recognition*, pages 387–402. Springer, 2025.

[50] Johannes Flotzinger, Philipp J Rösch, Christian Benz, Muneer Ahmad, Murat Cankaya, Helmut Mayer, Volker Rodehorst, Norbert Oswald, and Thomas Braml. Dacl-challenge: Semantic segmentation during visual bridge inspections. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 716–725, 2024.

[51] Johannes Flotzinger, Philipp J. Rösch, and Thomas Braml. dacl10k: Benchmark for semantic bridge damage segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8626–8635, January 2024.

[52] Johannes Flotzinger, Philipp J Rösch, and Thomas Braml. dacl10k: benchmark for semantic bridge damage segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8626–8635, 2024.

[53] Jerry A Fodor. *The language of thought*, volume 5. Harvard university press, 1975.

[54] Ashley Foster, Oscar Best, Mario Gianni, Asiya Khan, Keri Collins, and Sanjay Sharma. Drone footage wind turbine surface damage detection. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[55] Xiaoyu Gao, Xiaoyong Zhao, and Lei Wang. Memory augmented variational auto-encoder for anomaly detection. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 728–732, 2021.

[56] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[57] Mariana-Iuliana Georgescu, Antonio Bărbălău, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12737–12747, 2021.

[58] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[59] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[60] Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *European Conference on Computer Vision*, pages 210–226. Springer, 2024.

[61] Jutao Hao, Kai Huang, Chen Chen, and Jian Mao. Dual-student knowledge distillation for visual anomaly detection. *Complex & Intelligent Systems*, 10(4):4853–4865, 2024.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[63] Xiangjie He, Zhongqiang Luo, Quanyang Li, Hongbo Chen, and Feng Li. Dg-gan: A high quality defect image generation method for defect detection. *Sensors*, 23(13), 2023.

[64] Loucif Hebbache, Dariush Amirkhani, Mohand Saïd Allili, Nadir Hammouche, and Jean-François Lapointe. Leveraging saliency in single-stage multi-label concrete defect detection using unmanned aerial vehicle imagery. *Remote Sensing*, 15(5):1218, 2023.

[65] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[66] D.D. Hoffman and W.A. Richards. Parts of recognition. *Cognition*, 18(1):65–96, 1984.

[67] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020.

[68] Bin Hu, Xinggang Wang, and Wenyong Yu. Joint weakly and fully supervised learning for surface defect segmentation from images. *Signal Processing: Image Communication*, 107:116807, 2022.

[69] Jianlong Hu, Xu Chen, Zhenye Gan, Jinlong Peng, Shengchuan Zhang, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Liujuan Cao, and Rongrong Ji. Dmad: Dual memory bank for real-world anomaly detection. *arXiv preprint arXiv:2403.12362*, 2024.

[70] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[71] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE Transactions on Cybernetics*, 52(12):13834–13847, 2022.

[72] Chaoqin Huang, Qinwei Xu, Yanfeng Wang, Yu Wang, and Ya Zhang. Self-supervised masking for unsupervised anomaly detection and localization. *IEEE Transactions on Multimedia*, 25:4426–4438, 2023.

[73] IAEA Imagebank. 04790007, june 2004. Petr Pavlicek/IAEA; Flickr; CC BY-SA 2.0.

[74] IAEA Imagebank. Spent fuel pool (02813601), nov 2013. Photo by Greg Webb / IAEA; Flickr; Some rights reserved (Creative Commons).

[75] IAEA Imagebank. Urenco (03210632), oct 2013. CC BY-NC-ND 2.0.

[76] IAEA Imagebank. Warheads to plowshares: Checking leu lines, may 2013. CC BY-NC-SA 2.0.

[77] IAEA Imagebank. Clab, sweden (03211120), nov 2021. Photo by Dean Calma / IAEA; Flickr; CC BY 2.0.

[78] Marija Ivanovska and Vitomir Štruc. Y-gan: Learning dual data representations for anomaly detection in images. *Expert Systems with Applications*, 248:123410, 2024.

[79] Yashswi Jain, Ashvini Kumar Sharma, Rajbabu Velmurugan, and Biplab Banerjee. Posecvae: Anomalous human activity detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2927–2934. IEEE, 2021.

[80] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[81] Shashi Bhushan Jha and Radu F. Babiceanu. Deep cnn-based visual defect detection: Survey of current literature. *Computers in Industry*, 148:103911, 2023.

[82] Runyu Jiao, Yi Wan, Fabio Poiesi, and Yiming Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artificial Intelligence Review*, 56(Suppl 3):3515–3570, 2023.

[83] Glenn Jocher, Ayush Chaurasia, Jing Qiu, and Ultralytics Contributors. Ultralytics YOLO: You Only Look Once Object Detection Suite, 2023. Accessed: 30 June 2024.

[84] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing pro-

grams for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[85] Ali Karami, Thi Kieu Khanh Ho, and Narges Armanfard. Graph-jigsaw conditioned diffusion model for skeleton-based video anomaly detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4237–4247, 2025.

[86] Vahid Reza Khazaie, Anthony Wong, John Taylor Jewell, and Yalda Mohsenzadeh. Anomaly detection with adversarially learned perturbations of latent space. In *2022 19th Conference on Robots and Vision (CRV)*, pages 183–189, 2022.

[87] Hwan Kim, Byung Suk Lee, Won-Yong Shin, and Sungsu Lim. Graph anomaly detection with graph neural networks: Current status and challenges. *IEEE Access*, 10:111820–111829, 2022.

[88] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[89] Shotaro Kojima, Tomoya Takahashi, Ranulfo Bezerra, Takaaki Nara, Masaki Takahashi, Naoto Saiki, Kenta Gunji, Pongsakorn Songsuroj, Ryota Suzuki, Kotaro Sato, Zitong Han, Kagetora Takahashi, Yoshito Okada, Masahiro Watanabe, Kenjiro Tadakuma, Kazunori Ohno, and Satoshi Tadokoro and. Heterogeneous robots coordination for industrial plant inspection and evaluation at world robot summit 2020. *Advanced Robotics*, 36(21):1102–1119, 2022.

[90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[91] Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection. In *European conference on computer vision*, pages 206–226. Springer, 2020.

[92] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR, 10–15 Jul 2018.

[93] Brenden M Lake. *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn.* PhD thesis, Massachusetts Institute of Technology, 2014.

[94] Brenden M Lake, Russ R Salakhutdinov, and Josh Tenenbaum. One-shot learning by inverting a compositional causal process. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[95] Rocco Langone, Alfredo Cuzzocrea, and Nikolaos Skantzos. Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *Data Knowledge Engineering*, 130:101850, 2020.

[96] Wallace Lawson, Esube Bekele, and Keith Sullivan. Finding anomalies with generative adversarial networks for a patrolbot. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 484–485, 2017.

[97] Anh Le, Quang Uy Nguyen, Ngoc Tran Nguyen, Hai-Hong Phan, and Thi Huong Chu. An integration of pseudo anomalies and memory augmented autoencoder for video anomaly detection. In *Proceedings of the 11th International Symposium on Information and Communication Technology*, SoICT '22, page 262–269, New York, NY, USA, 2022. Association for Computing Machinery.

[98] Yunseung Lee and Pilsung Kang. Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724, 2022.

[99] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9664–9674, June 2021.

[100] Nanjun Li, Faliang Chang, and Chunsheng Liu. Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. *Neurocomputing*, 490:482–494, 2022.

[101] Ning Li, Kaitao Jiang, Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Anomaly detection via self-organizing map. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 974–978, 2021.

[102] Zhong Li, Yuxuan Zhu, and Matthijs Van Leeuwen. A survey on explainable anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–54, 2023.

[103] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[104] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024.

[105] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, page 1490–1499, New York, NY, USA, 2019. Association for Computing Machinery.

[106] Mingxuan Liu, Yunrui Jiao, Jingqiao Lu, and Hong Chen. Anomaly detection for medical images using teacher–student model with skip connections and multiscale anomaly consistency. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.

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

[108] Xin Liu, Xudong Yang, Lianhe Shao, Xihan Wang, Quanli Gao, and Hongbo Shi. Gm-detr: Research on a defect detection method based on improved detr. *Sensors*, 24(11), 2024.

[109] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, Jun Liu, Azzedine Boukerche, Peng Sun, and Liang Song. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *ACM Comput. Surv.*, 56(7), April 2024.

[110] Shuai Lu, Weihang Zhang, Jia Guo, Hanruo Liu, Huiqi Li, and Ningli Wang. Patchcl-ae: Anomaly detection for medical images using patch-wise contrastive learning-based auto-encoder. *Computerized Medical Imaging and Graphics*, 114:102366, 2024.

[111] Shuai Lu, Weihang Zhang, Jia Guo, Hanruo Liu, Huiqi Li, and Ningli Wang. Patchcl-ae: Anomaly detection for medical images using patch-wise contrastive learning-based auto-encoder. *Computerized Medical Imaging and Graphics*, 114:102366, 2024.

[112] Jiaxiang Luo and Jianzhao Zhang. A method for image anomaly detection based on distillation and reconstruction. *Sensors*, 23(22), 2023.

[113] Weixin Luo, Wen Liu, and Shenghua Gao. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing*, 444:332–337, 2021.

[114] Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767, 2021.

[115] Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767, 2021.

[116] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06, 2021.

[117] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11996–12004, 2019.

[118] Martin Mundt, Sagnik Majumder, Sreenivas Murali, Panagiotis Panetsos, and Visvanathan Ramesh. Meta-learning convolutional neural architectures for multi-target concrete defect classification with the concrete defect bridge image dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[119] Franka Nauert and Peter Kampmann. Inspection and maintenance of industrial infrastructure with autonomous underwater robots. *Frontiers in Robotics and AI*, 10:1240276, 2023.

[120] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics*, 16(1):393–402, 2020.

[121] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021.

[122] Phuc Cuong Ngo, Amadeus Aristo Winarto, Connie Khor Li Kou, Sojeong Park, Farhan Akram, and Hwee Kuan Lee. Fence gan: Towards better anomaly de-

tection. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 141–148, 2019.

[123] Nikos Th Nikolinakos. *EU policy and legal framework for Artificial intelligence, Robotics and related Technologies-the AI Act.* Springer, 2023.

[124] Phong Phu Ninh and HyungWon Kim. Contour-aware anomaly detection system for autoencoder neural network. In *2023 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4, 2023.

[125] Ghazal Alinezhad Noghre, Armin Danesh Pazho, and Hamed Tabkhi. Human-centric video anomaly detection through spatio-temporal pose tokenization and transformer. *arXiv preprint arXiv:2408.15185*, 2024.

[126] Toshiaki Ohgushi, Kenji Horiguchi, and Masao Yamanaka. Road obstacle detection method based on an autoencoder with semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[127] Naoaki Okuzumi, Kenji Matsuzaki, and Satoshi Okada. Development and application of robotics for decommissioning of fukushima daiichi nps by irid. *Journal of Robotics and Mechatronics*, 36(1):9–20, 2024.

[128] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021.

[129] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), March 2021.

[130] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020.

[131] Ravi A Patel, Lucas Steinmann, Jonas Fehrenbach, David Fehrenbach, and Frank Dehn. Convolution neural network-based machine learning approach for visual

inspection of concrete structures. In *International Conference of the European Association on Quality Control of Bridges and Structures*, pages 704–712. Springer, 2021.

[132] Mingjing Pei and Ningzhong Liu. A simplified student network with multi-teacher feature fusion for industrial defect detection. In Huimin Lu, Michael Blumenstein, Sung-Bae Cho, Cheng-Lin Liu, Yasushi Yagi, and Tohru Kamiya, editors, *Pattern Recognition*, pages 245–258, Cham, 2023. Springer Nature Switzerland.

[133] Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023. Survey Certification.

[134] Michela Prunella, Roberto Maria Scardigno, Domenico Buongiorno, Antonio Brunetti, Nicola Longo, Raffaele Carli, Mariagrazia Dotoli, and Vitoantonio Bevilacqua. Deep learning for automatic vision-based recognition of industrial surface defects: A survey. *IEEE Access*, 11:43370–43423, 2023.

[135] Jianing Qiu, Lipeng Chen, Xiao Gu, Frank P.-W. Lo, Ya-Yen Tsai, Jiankai Sun, Jiaqi Liu, and Benny Lo. Egocentric human trajectory forecasting with a wearable camera and multi-modal fusion. *IEEE Robotics and Automation Letters*, 7(4):8799–8806, 2022.

[136] Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261, 2023.

[137] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[138] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2626–2634, 2020.

[139] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[140] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14318–14328, June 2022.

[141] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[142] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging (IPMI 2017), Lecture Notes in Computer Science*, volume 10265, pages 146–157. Springer, Cham, 2017.

[143] Ujwal Sharma, Uma Shankar Medasetti, Taher Deemyad, Mustafa Mashal, and Vaibhav Yadav. Mobile robot for security applications in remotely operated advanced reactors. *Applied Sciences*, 14(6):2552, 2024.

[144] ASM Shihavuddin, Xiao Chen, Vladimir Fedorov, Anders Nymark Christensen, Nicolai Andre Brogaard Riis, Kim Branner, Anders Bjorholm Dahl, and Rasmus Reinhold Paulsen. Wind turbine surface damage detection by deep learning aided drone inspection analysis. *Energies*, 12(4):676, 2019.

[145] Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. A survey on compositional learning of AI models: Theoretical and experimental practices. *Transactions on Machine Learning Research*, 2024. Survey Certification.

[146] Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. A survey on compositional learning of ai models: Theoretical and experimetnal practices. *arXiv preprint arXiv:2406.08787*, 2024.

[147] Antony Douglas Smith, Shengzhi Du, and Anish Kurien. Vision transformers for anomaly detection and localisation in leather surface defect classification based on low-resolution images and a small dataset. *Applied Sciences*, 13(15), 2023.

[148] Jouwon Song, Kyeongbo Kong, Ye-In Park, Seong-Gyun Kim, and Suk-Ju Kang. Anoseg: Anomaly segmentation network using self-supervised learning. *arXiv preprint arXiv:2110.03396*, 2021.

[149] Benjamin K. Sovacool. A critical evaluation of nuclear power and renewable electricity in asia. *Journal of Contemporary Asia*, 40(3):369–400, 2010.

[150] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 184–192, New York, NY, USA, 2020. Association for Computing Machinery.

[151] Stanislaw Szymanowicz, James Charles, and Roberto Cipolla. Discrete neural representations for explainable anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 148–156, 2022.

[152] Domen Tabernik, Matic Šuc, and Danijel Skočaj. Automated detection and segmentation of cracks in concrete surfaces using joined segmentation and classification deep neural network. *Construction and Building Materials*, 408:133582, 2023.

[153] Ruixue Tang and Chao Ma. Interpretable neural computation for real-world compositional visual question answering. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 89–101. Springer, 2020.

[154] Yao Tang, Lin Zhao, Zhaoliang Yao, Chen Gong, and Jian Yang. Graph-based motion prediction for abnormal action detection. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, pages 1–7, 2021.

[155] Xian Tao, Dapeng Zhang, Wenzhi Ma, Xilong Liu, and De Xu. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 8(9), 2018.

[156] Johannes Theodoridis, Jessica Hofmann, Johannes Maucher, and Andreas Schilling. Trapped in texture bias? a large scale comparison of deep instance segmentation. In *European Conference on Computer Vision*, pages 609–627. Springer, 2022.

[157] Yu Tian, Guansong Pang, Yuyuan Liu, Chong Wang, Yuanhong Chen, Fengbei Liu, Rajvinder Singh, Johan W. Verjans, Mengyu Wang, and Gustavo Carneiro. Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. In Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang, editors, *Machine Learning in Medical Imaging*, pages 11–21, Cham, 2024. Springer Nature Switzerland.

[158] Himanshu Vairagade, Sungmin Kim, Hoojon Son, and Fan Zhang. A nuclear power plant digital twin for developing robot navigation and interaction. *Frontiers in Energy Research*, 12:1356624, 2024.

[159] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[160] Sai Babu Veesam, Aravapalli Rama Satish, Sreenivasulu Tupakula, Yuvaraju Chinnam, Krishna Prakash, Shonak Bansal, and Mohammad Rashed Iqbal Faruque. Design of an integrated model with temporal graph attention and transformer-augmented rnns for enhanced anomaly detection. *Scientific Reports*, 15(1):2692, 2025.

[161] Shashanka Venkataramanan, Kuan-Chuan Peng, Rajat Vikram Singh, and Abhijit Mahalanobis. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer, 2020.

[162] Ha Son Vu, Daisuke Ueta, Kiyoshi Hashimoto, Kazuki Maeno, Sugiri Pranata, and Sheng Mei Shen. Anomaly detection with adversarial dual autoencoders. *arXiv preprint arXiv:1902.06924*, 2019.

[163] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature

pyramid matching for anomaly detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[164] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, pages 494–511. Springer, 2022.

[165] H Wang, J Chen, T Pan, Z Dong, L Zhang, R Jiang, and X Song. Stgformer: Efficient spatiotemporal graph transformer for traffic forecasting. arxiv 2024. *arXiv preprint arXiv:2410.00385*, 2024.

[166] Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. Gsc: A graph and spatio-temporal continuity based framework for accident anticipation. *IEEE Transactions on Intelligent Vehicles*, 9(1):2249–2261, 2024.

[167] Wenxin Wang, Zhuo-Xu Cui, Guanxun Cheng, Chentao Cao, Xi Xu, Ziwei Liu, Haifeng Wang, Yulong Qi, Dong Liang, and Yanjie Zhu. A two-stage generative model with cyclegan and joint diffusion for mri-based brain tumor detection. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3534–3544, 2024.

[168] Xiaodong Wang, Jiangtao Fan, Fei Yan, Hongmin Hu, Zhiqiang Zeng, Pengtao Wu, Haiyan Huang, and Hangqi Zhang. Unsupervised anomaly detection via normal feature-enhanced reverse teacher–student distillation. *Electronics*, 13(20), 2024.

[169] Yizhou Wang, Dongliang Guo, Sheng Li, Octavia Camps, and Yun Fu. Explainable anomaly detection in images and videos: A survey. *arXiv preprint arXiv:2302.06670*, 2023.

[170] Jing Wei, Fei Shen, Chengkan Lv, Zhengtao Zhang, Feng Zhang, and Huabin Yang. Diversified and multi-class controllable industrial defect synthesis for data augmentation and transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4445–4453, June 2023.

[171] Julia Wolleb, Robin Sandkühler, and Philippe C Cattin. Descargan: Disease-specific anomaly detection with weak supervision. In *International conference on medical image computing and computer-assisted intervention*, pages 14–24. Springer, 2020.

[172] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 650–656, 2022.

[173] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L. Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 145–161, Cham, 2020. Springer International Publishing.

[174] Tianming Xie, Qifa Xu, and Cuixia Jiang. Anomaly detection for multivariate times series through the multi-scale convolutional recurrent variational autoencoder. *Expert Systems with Applications*, 231:120725, 2023.

[175] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 20370–20382, June 2025.

[176] Yingying Xu, Dawei Li, Qian Xie, Qiaoyun Wu, and Jun Wang. Automatic defect detection and segmentation of tunnel surface using modified mask r-cnn. *Measurement*, 178:109316, 2021.

[177] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. No release version; cite commit or branch if needed.

[178] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

[179] Moyuru Yamada, Vanessa D'Amario, Kentaro Takemoto, Xavier Boix, and Tomotake Sasaki. Transformer module networks for systematic generalization in visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10096–10105, 2024.

[180] Shen Yan, Haidong Shao, Yiming Xiao, Bin Liu, and Jiafu Wan. Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises. *Robotics and Computer-Integrated Manufacturing*, 79:102441, 2023.

[181] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3110–3118, May 2021.

[182] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3110–3118, 2021.

[183] Jie Yang, Ruijie Xu, Zhiquan Qi, and Yong Shi. Visual anomaly detection for images: A systematic survey. *Procedia Computer Science*, 199:471–478, 2022. The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19.

[184] Liyi Yao and Shaobing Gao. Dual-student knowledge distillation networks for unsupervised anomaly detection. *arXiv preprint arXiv:2402.00448*, 2024.

[185] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24490–24499, June 2023.

[186] Zhiyuan You, Kai Yang, Wenhan Luo, Lei Cui, Yu Zheng, and Xinyi Le. Adtr: Anomaly detection transformer with feature reconstruction. In Mohammad Tan-

veer, Sonali Agarwal, Seiichi Ozawa, Asif Ekbal, and Adam Jatowt, editors, *Neural Information Processing*, pages 298–310, Cham, 2023. Springer International Publishing.

[187] Shih-Yuan Yu, Arnav Vaibhav Malawade, Deepan Muthirayan, Pramod P. Khargonekar, and Mohammad Abdullah Al Faruque. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):7941–7951, 2022.

[188] Shoubin Yu, Zhongyin Zhao, Haoshu Fang, Andong Deng, Haisheng Su, Dongliang Wang, Weihao Gan, Cewu Lu, and Wei Wu. Regularity learning via explicit distribution modeling for skeletal video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[189] Muhammad Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Marcella Astrid, and Seung-Ik Lee. An anomaly detection system via moving surveillance robots with human collaboration. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2595–2601, 2021.

[190] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, October 2021.

[191] Deyi Zeng. Anomaly detection by unsupervised adversarial generative self-labelling autoencoder. In *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 1012–1021, 2022.

[192] Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu. A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):200–212, 2021.

[193] Zhao Zhanfang and Li Tuo. Enhancing wind turbine blade damage detection with yolo-wind. *Scientific Reports*, 15(1):18667, 2025.

[194] Haibo Zhang, Wenping Guo, Shiqing Zhang, Hongsheng Lu, and Xiaoming Zhao. Unsupervised deep anomaly detection for medical images using an improved adversarial autoencoder. *Journal of Digital Imaging*, 35(2):153–161, 2022.

[195] Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, and Jianxin Liao. Multi-scale video anomaly detection by multi-grained spatio-temporal representation learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17385–17394, 2024.

[196] Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 151:103990, 2023.

[197] Huan Zhao, Fang Wan, Guangbo Lei, Ying Xiong, Li Xu, Chengzhi Xu, and Wen Zhou. Lsd-yolov5: A steel strip surface defect detection algorithm based on lightweight network and enhanced feature fusion mode. *Sensors*, 23(14), 2023.

[198] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

[199] Zhili Zhou, Xiaohua Dong, Zhetao Li, Keping Yu, Chun Ding, and Yimin Yang. Spatio-temporal feature encoding for traffic accident detection in vanet environment. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19772–19781, 2022.

[200] Honglei Zhu, Pengjuan Wei, and Zhigang Xu. A spatio-temporal enhanced graph-transformer autoencoder embedded pose for anomaly detection. *IET Computer Vision*, 18(3):405–419, 2024.

[201] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[202] S.J. Zinkle and G.S. Was. Materials challenges in nuclear energy. *Acta Materialia*, 61(3):735–758, 2013. The Diamond Jubilee Issue.

[203] Hüseyin Üzen, Muammer Türkoğlu, Berrin Yanikoglu, and Davut Hanbay. Swinmfinet: Swin transformer based multi-feature integration network for detection of pixel-level surface defects. *Expert Systems with Applications*, 209:118269, 2022.