DEPARTMENT OF ENGINEERING MATHEMATICS

Optimising Synthetic Data Use in AI for Industrial Fire Detection

Mishara Sapukotanage

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Friday 29th August, 2025

Supervisor: Dr. George Jenkinson

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Mishara Sapukotanage, Friday 29th August, 2025

Contents

List of Figures

List of Tables

Abstract

Industrial fires pose severe risks to safety and business continuity, yet collecting large, representative datasets of real fire imagery is hazardous and impractical. This thesis investigates whether synthetic imagery can reduce reliance on scarce real data while sustaining deployment-grade performance in vision-based fire detection. The study focuses on binary image classification (fire vs. no-fire) as a lightweight early-warning task and evaluates how synthetic augmentation influences accuracy, generalisation, and indoor deployability under domain shift.

A ResNet-50 backbone was applied in a four-phase experimental design, comparing real-only, synthetic-only, and mixed datasets across outdoor (D-Fire, Yunnan MSFFD) and indoor (PLOS ONE, SYN-FIRE) domains. Phase 1 established frozen outdoor baselines under varying synthetic–real ratios; Phase 2 fine-tuned the backbone; Phase 3 tested outdoor models indoors to measure domain shift; and Phase 4 trained indoor models directly to assess deployment with reduced real data. Performance was assessed using standard classification metrics, while Grad-CAM provided interpretability.

Results showed that mixed training consistently outperformed real-only or synthetic-only regimes. A fine-tuned 50:50 outdoor model achieved F1 = 0.916 on D-Fire while using one-fifth the real images of the real-only baseline. Under domain shift, the same 50:50 mix sustained higher F1 (0.878) than the real-only equivalent (0.852), demonstrating improved transferability. In the deployment setting, a 50:50 indoor model trained with 2,000 real and 2,000 synthetic images reached F1 = 0.984 and MCC = 0.972 (ROC AUC $\approx$ 0.999), slightly exceeding the indoor real-only model while halving real-data requirements. Grad-CAM confirmed attention on flames and smoke, strengthening interpretability.

The study concludes that synthetic augmentation, when balanced with real imagery, enables accurate, interpretable, and data-efficient fire classifiers. The recommended deployment model is a 50:50 indoor ResNet-50, supported by threshold tuning for site-specific safety priorities. This demonstrates that synthetic imagery is not merely a stopgap but a strategic enabler of scalable, reliable industrial fire detection.

Supporting Technologies

- **Google Colab**: GPU-accelerated execution; environment management for reproducible training and evaluation.
- **Google Drive**: Structured storage for datasets, model checkpoints, metrics JSON, and figure artefacts.
- **GitHub Repository**: Version control for dataset loaders, training/evaluation utilities, and notebooks. Publicly available at `github.com/Misharasapu/fire-detection-dissertation`.
- **PyTorch & torchvision**: ResNet-50 backbone (ImageNet-pretrained), dataloaders, training loops, and inference.
- **Python tooling**: `numpy`, `pandas`, `matplotlib`; custom `metrics.py` for scalar/curve outputs.
- **Grad-CAM utilities**: Targeting `layer4[-1].conv3` to generate image–heatmap–overlay triptychs.
- **TikZ/LaTeX**: Flowcharts and schematics (dataset pipeline, phase design, ResNet-50 blocks).

Notation and Acronyms

CNN	Convolutional Neural Network
ResNet-50	50-layer Residual Network (ImageNet-pretrained backbone)
ImageNet	Large-scale visual dataset used for pretraining deep CNNs
D-Fire	Outdoor real-world fire image dataset
PLOS ONE	Indoor real-world fire image dataset
Yunnan MSFFD	Synthetic outdoor fire dataset (Unreal Engine)
SYN-FIRE	Synthetic indoor fire dataset (Omniverse)
AUC	Area Under the Curve (ROC or PR)
ROC curve	Plot of True Positive Rate vs. False Positive Rate
PR curve	Plot of Precision vs. Recall
ROC AUC	Area under Receiver Operating Characteristic curve
PR AUC	Area under Precision–Recall curve (average precision)
TP / FP / TN / FN	True/False Positive; True/False Negative
Precision	$TP/(TP+FP)$
Recall (Sensitivity)	$TP/(TP+FN)$
Specificity (TNR)	$TN/(TN+FP)$
TNR	True Negative Rate (Specificity)
F1	Harmonic mean of precision and recall
MCC	Matthews Correlation Coefficient
Grad-CAM	Gradient-weighted Class Activation Mapping

Acknowledgements

I would like to express my sincere gratitude to SYNOPTIX Ltd. for providing this exciting and impactful project, and for their continued support throughout. In particular, I am especially thankful to George Leete, whose guidance during our weekly meetings was invaluable in shaping the direction of the work and in connecting the technical findings to industrial relevance.

I am also deeply grateful to my academic supervisor, Dr. George Jenkinson, for his consistent guidance, constructive feedback, and encouragement during both the research and the writing stages.

Finally, I would like to thank my colleagues and friends for their discussions, advice, and support, which helped me to remain motivated and focused throughout this dissertation.

Chapter 1

Introduction

1.1 Context and Motivation

Industrial fires represent one of the most serious risks to human safety and business continuity. In factories, warehouses, and power plants, even a small ignition can escalate rapidly due to combustible materials, heavy machinery, and confined layouts. Conventional detectors based on smoke or heat sensors provide reliable coverage in some situations, but they typically activate only once signals exceed a threshold [5]. By this stage, valuable time may already have been lost. Computer vision systems offer an alternative: by recognising visual indicators of fire directly in images or video feeds, they can raise earlier alerts and gain crucial seconds for containment [17]. Achieving this goal, however, depends on training models with large and diverse datasets.

Figure 1.1: Example of an indoor industrial fire scenario captured by a surveillance camera [20]. Such imagery underscores the practical need for scalable detection systems that can provide early warnings in safety-critical environments.

Acquiring real industrial fire imagery remains a major obstacle. Fire incidents are rare, ethically problematic to capture, and hazardous to stage. Available datasets are limited in scale and skewed toward outdoor or laboratory scenarios that do not reflect indoor industrial conditions [18, 19]. Indoor environments introduce additional challenges such as variable lighting, complex layouts, camera occlusion, and visual clutter from equipment. Without sufficient coverage, models risk poor generalisation, leading to dangerous missed detections or frequent false alarms that undermine confidence in automated systems.

Synthetic imagery offers a practical remedy to this bottleneck. Advances in simulation platforms and graphics engines make it possible to render realistic flames and smoke under a wide range of conditions,

complete with precise annotations. When combined with real imagery, synthetic augmentation can expand training coverage, mitigate bias, and stabilise model behaviour across domains [10, 7, 8, 1]. For industry, this represents a safe and scalable route to developing reliable fire detection systems without depending solely on rare real events.

This thesis addresses this challenge directly. It investigates whether synthetic augmentation can reduce dependence on scarce real imagery while supporting reliable indoor deployment of AI-based fire detection systems in industrial settings. The following sections define the research gap and summarise the contributions made.

1.2 Research Gap

Research on computer vision for fire detection has progressed in recent years, but several gaps remain unresolved. Many studies focus on detection or segmentation, which require bounding boxes or masks, while binary classification has received comparatively little attention despite its deployment value for lightweight early-warning systems. Evidence for synthetic augmentation is also fragmented: few studies systematically compare different synthetic–real ratios or training regimes, leaving uncertainty over when augmentation is most effective. Finally, domain shift remains a critical challenge. Models trained on outdoor imagery often fail in indoor conditions where lighting, clutter, and camera placement differ substantially. A structured evaluation of synthetic augmentation for classification, including its role in outdoor-to-indoor transfer, is therefore needed.

1.3 Aims and Contributions

The overarching aim of this thesis is to determine how synthetic imagery can reduce reliance on real data while retaining deployment-grade performance, especially under domain shift. To guide this objective, four research questions (RQ1–RQ4) are defined and addressed throughout the thesis:

RQ1 To what extent can synthetic data substitute or complement real imagery in training fire classifiers?

RQ2 Which synthetic–real composition delivers the most effective performance?

RQ3 How well do models trained with synthetic augmentation generalise across domains, particularly from outdoor to indoor environments?

RQ4 Can synthetic augmentation achieve deployment-grade accuracy with fewer real indoor images?

Addressing these questions required a structured investigation that moves beyond isolated experiments and towards a replicable framework with clear practical relevance. The thesis therefore made the following six contributions (C1–C6):

C1 Designed and implemented a phased experimental framework isolating dataset composition, training regime, and domain shift for image-level fire classification.

C2 Conducted a systematic comparison of real-only, synthetic-only, and mixed training, identifying when synthetic augmentation is most beneficial.

C3 Performed a targeted study of outdoor-to-indoor transfer, quantifying how synthetic data affects generalisation.

C4 Provided empirical evidence that balanced mixtures reduce real-data requirements while maintaining high accuracy and reliability.

C5 Delivered a deployment-ready indoor model trained on a 50/50 real–synthetic mix, showing that synthetic augmentation supports practical industrial use.

C6 Applied Grad-CAM analyses to explain both correct and failure cases, linking attention patterns to typical false positives and false negatives.

Chapter 2

Literature Review

Industrial fire detection has traditionally relied on hardware sensors such as infrared alarms or smoke detectors. While effective in some contexts, these devices typically trigger only once heat or smoke exceeds a threshold, leading to delays and false positives from benign environmental factors. In contrast, computer vision methods can recognise visual indicators of fire directly in image or video streams, offering earlier and potentially more reliable warnings. Deep learning in particular has become dominant, as it captures complex visual patterns more effectively than rule-based or sensor-driven systems.

Progress has been constrained by a persistent bottleneck: the scarcity of reliable fire imagery. Public datasets such as D-Fire and Indoor FS provide valuable resources, yet remain limited in scale, domain coverage, and annotation quality. This limitation is especially acute indoors, where real fires are rare and hazardous to capture. As a result, many models perform well in controlled outdoor or staged datasets but generalise poorly to the clutter, variable lighting, and occlusion typical of factories or warehouses.

Synthetic augmentation has emerged as a promising solution. Advances in simulation platforms and graphics engines now make it possible to render flames and smoke under diverse conditions, complete with precise annotations. Beyond fire vision, studies in domains such as multi-object detection show that dataset diversity can matter more than photorealism, suggesting synthetic imagery may provide essential coverage where real data is scarce. The central question, however, is not only whether synthetic data can substitute for real imagery, but how the two can be combined most effectively to achieve consistent and transferable performance.

This review therefore examines image-based methods for fire detection with emphasis on binary classification, while drawing on detection and segmentation research for context. Its focus is how dataset scarcity and the integration of synthetic augmentation influence model performance, cross-domain generalisation, and feasibility for industrial deployment.

2.1 Problem Landscape and Dataset Constraints

The application of deep learning to fire detection expanded rapidly in recent years, but the field continued to face persistent limitations in the availability and quality of data. Vasconcelos et al. [18] reviewed more than one hundred studies and concluded that most systems were trained on small, domain-specific datasets that restricted generalisation. This scarcity of diverse imagery, combined with the lack of consistent evaluation standards, meant that many published models performed well in controlled experiments but failed in complex real-world environments.

The absence of large-scale, representative datasets was especially visible in industrial and indoor contexts. Real fire events in such settings were rare and often unsafe to capture, leading to a concentration of data collection in outdoor surveillance or wildfire scenarios. As a result, indoor fire vision research lagged behind, despite being highly relevant for practical deployment in factories, warehouses, and other enclosed facilities. This imbalance created a significant research gap and motivated the exploration of synthetic imagery as a way to fill the void.

Several benchmark datasets illustrated both the progress and the limitations of available resources. Table 2.1 summarises the main collections. Among them, two were most relevant for this thesis: the **D-Fire** dataset [19], which contained over 21,000 labelled outdoor images, and the **Indoor Fire and Smoke (Indoor FS)** dataset [14], which included around 5,000 indoor images. These formed the real-data backbone of the study. However, both exhibited constraints. D-Fire was biased toward outdoor

surveillance imagery and showed inconsistent labelling of small flames, while Indoor FS remained modest in scale relative to modern deep learning requirements.

Table 2.1: Fire datasets at a glance, showing domain, size, and primary task. This thesis focuses on D-Fire and Indoor FS, while smaller legacy sets are included here for context to illustrate the scarcity and fragmentation of real fire imagery. Synthetic datasets (e.g. Yunnan, SYN-FIRE) are discussed separately in Section 2.3.

Dataset	Domain	Size	Task
D-Fire	Outdoor	21,527	Detection/Classification
Indoor FS	Indoor	5,000	Detection/Classification
BoWFire	Outdoor	2,000	Detection/Classification
OVIFIRE	Outdoor	1,400	Detection
FIRESENSE	Outdoor	3,000	Detection
VisiFire	Outdoor	1,200	Detection
MIVIA	Mixed	62,690 frames	Video-based detection

Other datasets, such as BoWFire, OVIFIRE, FIRESENSE, and VisiFire, were widely cited in the literature but limited in size and diversity, typically containing fewer than 3,000 images. Many of these sets were generated by overlaying fire elements on static backgrounds, which restricted their realism. Historical video collections such as MIVIA [3] established early benchmarks, but their annotation detail no longer matched the demands of current architectures. These smaller or outdated datasets underlined why synthetic augmentation became increasingly important. They highlighted the lack of representative, large-scale fire imagery, especially for industrial deployment.

Synthetic datasets such as Yunnan and SYN-FIRE further illustrated both the opportunities and challenges of augmentation. They provided diverse fire scenarios, but they were heavily biased toward fire-positive imagery, making them unsuitable for binary classification unless combined with real data that supplied the necessary negative cases. This limitation directly motivated the hybrid strategies examined later in this thesis.

2.2 Methods in Fire Vision: Classification vs Detection vs Segmentation

Early research on image-based fire detection relied on handcrafted features and rule-based models. Chen et al. [2] developed one of the first generic colour models, using the YCbCr colour space to distinguish flames from the background. While effective under controlled conditions, such rule-based systems were prone to false positives when exposed to distractors such as lights or sunlight. Later approaches incorporated motion and shape cues. Foggia et al. [3] proposed a multi-expert system that combined colour evaluation, shape irregularity, and motion descriptors, which improved reliability and demonstrated feasibility on embedded devices. These contributions laid the groundwork for computer vision-based fire detection but were limited in scalability and adaptability.

The introduction of deep learning architectures shifted the field toward end-to-end feature learning. Vasconcelos et al. [18] summarised how convolutional neural networks became dominant for classification, YOLO variants for detection, and U-Net derivatives for segmentation. Several architectures became recurring backbones across these tasks. ResNet-50 [6], adopted in this thesis, introduced residual connections that stabilised deep network optimisation and quickly became standard in fire vision studies. VGG [13] provided an earlier deep CNN benchmark but was computationally heavy. MobileNetV2 [11] prioritised efficiency with inverted residuals and proved suitable for embedded deployment, while EfficientDet [16] extended this trajectory with scalable detection using bidirectional feature fusion. Together, these models illustrated the evolution from sequential to residual, lightweight, and compound-scaled architectures.

Detection approaches have been the most widely adopted, particularly for localising fire and smoke in real time. Sozol et al. [14] introduced the Indoor Fire and Smoke dataset and optimised YOLOv5 for indoor scenarios, showing that it outperformed both Faster R-CNN and later YOLO variants. Their model integrated DeepSORT tracking and Grad-CAM interpretability, reinforcing the importance of temporal and visual explanations in practical deployments. Such work illustrated the maturity of detection pipelines

Table 2.2: Comparison of fire vision tasks. Classification is lightweight but relatively underexplored, whereas detection and segmentation dominate research due to their detailed outputs. This thesis focuses on classification because of its deployment value for early-warning systems.

Task	Purpose	Typical Architectures
Classification	Decide whether an image contains fire or not (lightweight early warning).	ResNet-50, VGG, MobileNetV2
Detection	Localise fire or smoke with bounding boxes for actionable alerts.	YOLOv5/7/8, EfficientDet
Segmentation	Delineate flames or smoke at the pixel level for detailed scene understanding.	U-Net family, U-Net++ (Swin)

but also highlighted how classification remained relatively underexplored, despite its value for lightweight early-warning systems.

The contrast between handcrafted systems and deep architectures therefore showed a clear trajectory. While early methods relied on colour and motion rules, modern pipelines used deep backbones trained on large or augmented datasets. Within this landscape, classification remained a comparatively neglected task but one with practical value, particularly when paired with synthetic augmentation to counter dataset scarcity. This gap motivated the systematic evaluation of classification with real and synthetic data mixtures in this thesis. A concise comparison of the three major computer vision tasks for fire detection is presented in Table 2.2, highlighting their purposes and typical architectures.

2.3 Synthetic Data for Fire Vision

A growing body of work has investigated how synthetic imagery can compensate for the scarcity of labelled fire datasets. Although these approaches varied in generation technique and task focus, a consistent trend emerged: hybrid training with both synthetic and real images outperformed either source alone.

Park et al. [10] investigated wildfire classification using Progressive GANs. While their models benefited from mixed training, synthetic-only regimes produced excessive false positives, highlighting weak generalisation. Hu et al. [7] introduced the FireFly dataset, built in Unreal Engine, which demonstrated that pretraining on synthetic embers followed by fine-tuning on a small real set improved detection mAP by up to 8.6% while reducing annotation effort.

Indoor contexts received increasing attention. Kim et al. [8] developed digital twins of buildings and rendered synthetic fire and smoke in Unity, generating large labelled datasets that enabled real-time Jetson deployment with reduced false alarms via temporal post-processing. Arlovic et al. [1] released SYNFIRE, a photorealistic indoor dataset containing only fire-positive segmentation masks. While valuable for pixel-level tasks, its lack of negative samples and absence of classification labels limit its direct use for binary classification unless paired with real no-fire data. Beyond fire-specific studies, Staniszewski et al. [15] evaluated synthetic strategies for object detection and concluded that dataset diversity mattered more than photorealism, reinforcing the case for combining varied synthetic imagery with real-world data.

The main findings of these studies are summarised in Table 2.3. Across different domains, synthetic-only training consistently struggled with precision, while mixed or pretrain–fine-tune strategies improved generalisation and reduced the domain gap. Balanced combinations proved most effective because real data grounded the model while synthetic imagery supplied diversity that limited real datasets could not provide.

Table 2.3: Limitations of key synthetic data studies. Despite varied approaches, a consistent weakness was identified: synthetic-only models lacked reliability, whereas mixed datasets proved more transferable.

Study	Task	Main Limitation
Park (2022)	Classification (GAN wildfire images)	Synthetic-only models produced excessive false positives and failed to transfer reliably.
Hu (2023)	Detection (FireFly, UE4 embers)	Coverage restricted to embers and aerial perspectives, limiting diversity.
Kim (2024)	Detection (Unity digital twin)	Required site-specific RGB-D capture for each environment.
Arlovic (2025)	Segmentation (SYN-FIRE, Omniverse)	Fire-positive only; provided segmentation masks without negatives, limiting suitability for classification.
Staniszewski (2023)	Multi-object detection	Analysis not validated on fire imagery, though it showed diversity was more important than photorealism.

2.4 Domain Shift and Generalisation

One of the most persistent challenges in vision-based fire detection is domain shift, where models trained under one set of conditions perform poorly when exposed to another. For industrial deployment, this often means that models trained on outdoor surveillance footage fail to transfer effectively to cluttered indoor environments. Changes in lighting, camera geometry, occlusion, and visual background clutter are particularly problematic, as they alter the appearance of fire cues in ways the model has not seen during training.

Vasconcelos et al. [18] highlighted this issue in their survey, noting that models validated on curated datasets often collapsed in uncontrolled environments. Earlier handcrafted systems such as Chen et al. [2] were especially vulnerable, as fire-coloured distractors frequently produced false positives. Even modern deep networks continue to suffer when the deployment distribution diverges from the training set.

Several studies suggest that synthetic augmentation can help mitigate these effects. Park et al. [10] reported that GAN-generated wildfire imagery reduced overfitting and improved classification robustness in noisy outdoor conditions. Hu et al. [7] showed that pretraining on the synthetic FireFly dataset followed by fine-tuning on a small real dataset improved transferability to real wildfire video, particularly in cluttered or low-visibility scenes. Arlovic et al. [1] extended this evidence to indoor environments, demonstrating that mixed training with SYN-FIRE and real images improved segmentation performance on previously unseen warehouse fires.

Taken together, these findings indicate that broader and more diverse training distributions improve generalisation across domains. Synthetic augmentation appears especially valuable for reducing the severity of domain shift, offering a pathway to more reliable deployment in industrial contexts where indoor data are scarce.

2.5 Interpretability and Deployment Considerations

Interpretability became a key requirement for deep learning models in fire detection, as stakeholders needed to understand why an alert was issued. Park et al. [10] applied Class Activation Mapping (CAM) to classifiers trained with synthetic augmentation and demonstrated that attention was correctly focused on flame regions. Sozol et al. [14] extended this by applying Grad-CAM to YOLOv5 detectors, showing that predicted bounding boxes aligned with visible smoke and fire. Such visual explanations strengthened user trust. Earlier rule-based systems, such as Foggia et al. [3], can also be regarded as interpretable, since their decisions were based on explicit combinations of colour, shape, and motion cues.

Deployment feasibility was equally important. Venâncio et al. [19] demonstrated that models trained on D-Fire could run on a Raspberry Pi in real time, balancing accuracy with efficiency. Kim et al. [8] advanced this further by generating synthetic training data via digital twins and deploying detectors on Jetson hardware, where temporal filtering reduced false alarms. These examples showed that both model

architecture and operational safeguards were necessary to achieve practical deployment. Lightweight backbones such as MobileNetV2 [11] and scalable detection models such as EfficientDet [16] reinforced this trend by offering efficiency without sacrificing capability.

False positive mitigation remained a consistent theme. Sozol et al. [14] integrated DeepSORT temporal tracking to suppress spurious alerts, while Kim et al. [8] applied temporal coherence filters to achieve a similar effect. These studies indicated that reliable deployment depended not only on improved training data but also on system-level policies such as temporal smoothing and threshold tuning—an approach that aligns with the industrial emphasis on stability and trustworthiness.

2.6 Summary and Sharpened Gap

This review has traced the evolution of vision-based fire detection from handcrafted approaches through deep learning and into the emerging use of synthetic data. Early systems, such as those proposed by Chen et al. [2] and Foggia et al. [3], relied on colour, motion, and shape cues. These methods were interpretable but prone to false positives and lacked robustness across diverse environments. The introduction of deep learning architectures, including VGG [13], ResNet-50 [6], MobileNetV2 [11], and EfficientDet [16], shifted the field toward end-to-end feature learning and has since dominated classification, detection, and segmentation. Nevertheless, as Vasconcelos et al. [18] emphasised, the underlying challenges of limited datasets and weak cross-domain generalisation remain unresolved.

Larger curated datasets have provided partial progress. D-Fire [19] established a widely used outdoor benchmark, while Indoor FS [14] addressed the scarcity of indoor fire imagery. Yet both remain modest in coverage, and rare or safety-critical scenarios are underrepresented. Synthetic data has therefore emerged as a strategic solution. Studies by Park et al. [10], Hu et al. [7], Kim et al. [8], and Arlovic et al. [1] consistently reported that balanced mixtures of synthetic and real data outperformed either source alone. Staniszewski et al. [15] further showed that dataset diversity and domain randomisation were often more valuable than extreme photorealism, reinforcing the role of synthetic augmentation as more than a temporary substitute.

Interpretability and deployment feasibility have evolved in parallel. Techniques such as Grad-CAM [14], WSOL [10], and modular expert systems [3] demonstrated that fire detection models can remain explainable. Deployment studies confirmed that models could run effectively on constrained hardware such as Raspberry Pi [19] and Jetson [8], provided efficiency was balanced with safeguards such as temporal filtering or threshold tuning. These findings underscored that successful fire detection required both robust training data and operational reliability.

Despite these advances, several gaps remain unresolved. Binary classification for industrial fire detection has been comparatively neglected, even though it offers clear value for lightweight early-warning systems. The effect of varying synthetic–real ratios has rarely been examined systematically, despite evidence that mid-ratio mixes are most effective. Domain shift, particularly between outdoor and indoor contexts, continues to undermine reliability. While synthetic augmentation can reduce this problem, it cannot fully replace diverse real-world validation. These open issues establish the rationale for this thesis: a structured evaluation of synthetic–real data mixtures for binary fire classification using ResNet-50, with explicit attention to outdoor-to-indoor transfer and the development of a deployment-ready indoor model.

Figure 2.1 summarises this trajectory. It illustrates the field's progression from handcrafted rules, through deep learning backbones, to synthetic augmentation, before highlighting the specific research gaps that motivate this work: binary classification, the optimisation of synthetic–real ratios, resilience under outdoor-to-indoor transfer, and deployability with interpretability.

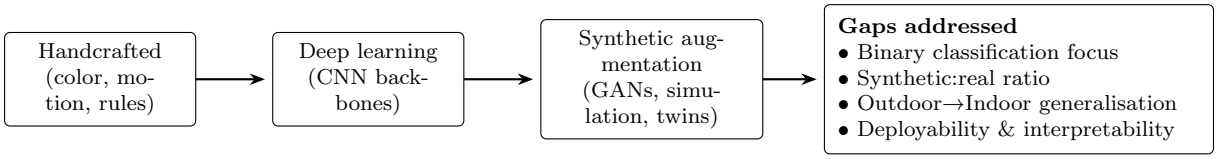

Figure 2.1: Literature trajectory and the specific gaps this dissertation targets.

Chapter 3

Methodology

This chapter sets out the methodological framework used to investigate the role of synthetic fire imagery in supplementing scarce real-world data for industrial fire detection. The overarching aim, as introduced in Chapter 1, was to determine how far synthetic augmentation can reduce reliance on real imagery while retaining deployment-grade reliability. The study was guided by the four research questions defined in Section 1.3 (RQ1–RQ4), which collectively ask whether synthetic data can substitute or complement real imagery, what balance of synthetic–real composition is most effective, how well models generalise under outdoor-to-indoor transfer, and whether deployment-grade performance can be achieved with fewer real images.

To address these questions, the experimental design was organised into four sequential phases, each isolating and then integrating different methodological factors. Phase 1 established outdoor baselines under varying synthetic–real ratios (addressing RQRQ2). Phase 2 evaluated the benefits of fine-tuning compared to frozen feature extraction (addressing RQRQ1 and RQRQ2). Phase 3 examined outdoor-to-indoor transfer to quantify the impact of domain shift (addressing RQRQ3). Finally, Phase 4 focused on indoor training for deployment, testing whether synthetic augmentation could achieve high accuracy with reduced real data (addressing RQRQ4).

The remainder of this chapter is structured as follows. Section 3.1 introduces the datasets and preprocessing pipeline, explaining how real and synthetic sources were harmonised for binary classification. Section 3.2 presents the ResNet-50 backbone and training setup, contrasting frozen and fine-tuned regimes. Section 3.3 details the four-phase experimental framework. Section 3.4 outlines the evaluation metrics and visualisation tools, Section **??** describes interpretability through Grad-CAM, and Section **??** explains the implementation strategy and reproducibility practices. Together, these methods provide a coherent and transparent foundation for the results presented in Chapter 4.

3.1 Datasets and Preprocessing

This project drew on both real and synthetic datasets to construct training, validation, and testing sets for binary fire classification. Each dataset contributed complementary strengths and exposed specific limitations, reflecting the broader challenge of balancing realism, diversity, and class balance. The integration of these datasets was motivated by two factors: the scarcity of real fire imagery—especially in indoor settings—and the ability of synthetic data to generate controlled diversity at scale. Together, the four datasets formed the backbone of the experimental design, enabling systematic tests of synthetic–real mixtures across outdoor and indoor domains.

3.1.1 D-Fire (Real, Outdoor)

The D-Fire dataset [19] contained 21,527 surveillance-style images labelled for fire, smoke, or fire+smoke. For this study, bounding box annotations were repurposed into image-level binary labels, with any image containing fire or fire+smoke marked as positive. The dataset was split into a training pool and a fixed test set, the latter reserved exclusively for evaluation in Phases 1 and 2. D-Fire provided a realistic baseline by reflecting outdoor CCTV conditions typical of industrial contexts.

3.1.2 PLOS ONE (Real, Indoor)

The PLOS ONE dataset [14] consisted of 5,000 indoor images with bounding box annotations for fire and smoke. As with D-Fire, annotations were converted into binary labels by assigning positive if any fire box was present. Unlike D-Fire, the dataset provided explicit train, validation, and test splits, which were used directly in Phase 4. It played a central role in assessing generalisation to indoor environments, representing the intended deployment domain for this project.

3.1.3 Yunnan MSFFD (Synthetic, Outdoor)

The Multi-Scenario Forest Fire Dataset (MSFFD) [7] contained 3,946 Unreal Engine 5 images of simulated outdoor fires across varied terrains, weather, and lighting. After binarisation, the dataset was heavily skewed towards fire-positive samples (approximately 83%), which encouraged overprediction of the fire class when used in isolation. It was therefore combined with D-Fire in Phases 1 and 2 to construct balanced training sets.

3.1.4 SYN-FIRE (Synthetic, Indoor)

SYN-FIRE [1] provided 2,030 photorealistic renderings of industrial indoor environments with pixel-level segmentation masks. All available images contained fire pixels, making the dataset effectively positive-only. While unsuitable for standalone binary classification, SYN-FIRE was paired with PLOS ONE negatives in Phase 4 to create a balanced indoor training set while retaining synthetic diversity.

3.1.5 Preprocessing and Label Strategy

To ensure comparability, all datasets were standardised into a common format via custom PyTorch Dataset classes. Images were resized to 224×224 pixels and converted to tensors, without applying ImageNet normalisation. This policy was adopted after early experiments showed that inconsistent preprocessing reduced performance during fine-tuning. Labels were unified under a binary scheme (Equation 3.1):

$$y = \begin{cases} 1 & \text{if fire present,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Dataset helpers encoded source-specific label rules: class ID 1 for fire in D-Fire, class ID 0 for fire in Yunnan and PLOS ONE, and any non-zero mask pixel for SYN-FIRE. To support fair comparisons, all mixed datasets were constructed at fixed sizes with deterministic sampling (`seed`=42). Outdoor mixtures (D-Fire+Yunnan) were capped at 5,260 images, while the indoor mix in Phase 4 contained exactly 2,000 real and 2,000 synthetic samples. This controlled design ensured that observed performance differences could be attributed to data composition rather than dataset size.

The four datasets are summarised in Table 3.1. While the table reports their domain, size, and role in the experiments, it should be noted that Yunnan was skewed towards fire-positive samples (83%) and SYN-FIRE was entirely positive. These imbalances were not corrected during training and are reflected upon in the Discussion as a limitation.

Table 3.1: Datasets used in this project, summarised by domain, size, and experimental role. Class imbalance details are described in the text.

Dataset	Domain	Size	Role
D-Fire	Outdoor (real)	21,527	Training + fixed test (Phases 1–2)
PLOS ONE	Indoor (real)	5,000	Indoor evaluation + Phase 4 training
Yunnan MSFFD	Outdoor (synthetic)	3,946	Mixed with D-Fire (Phases 1–2)
SYN-FIRE	Indoor (synthetic)	2,030	Mixed with PLOS ONE (Phase 4)

The preprocessing workflow is illustrated in Figure 3.1, which schematically represents the steps from raw imagery through resizing and label harmonisation to the final train/validation/test splits.

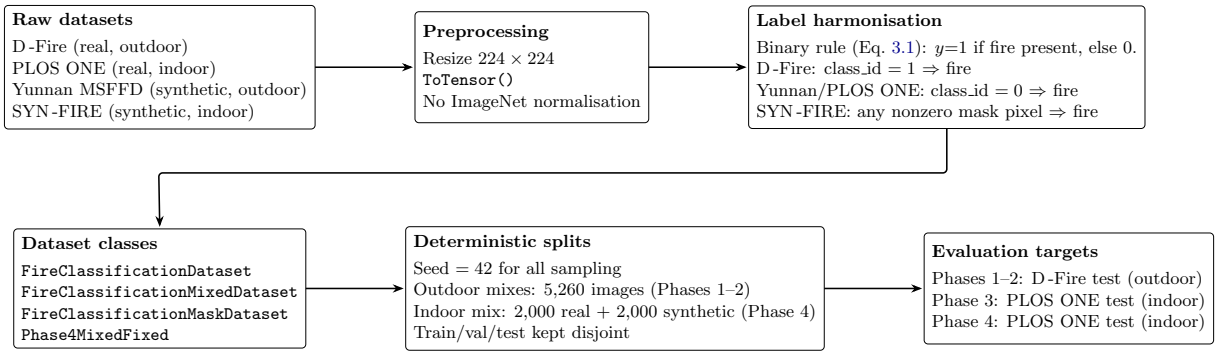

Figure 3.1: Dataset pipeline showing the flow from raw imagery through preprocessing and label harmonisation to dataset classes, deterministic splits, and phase-specific evaluation sets.

3.2 Model Architecture and Training Setup

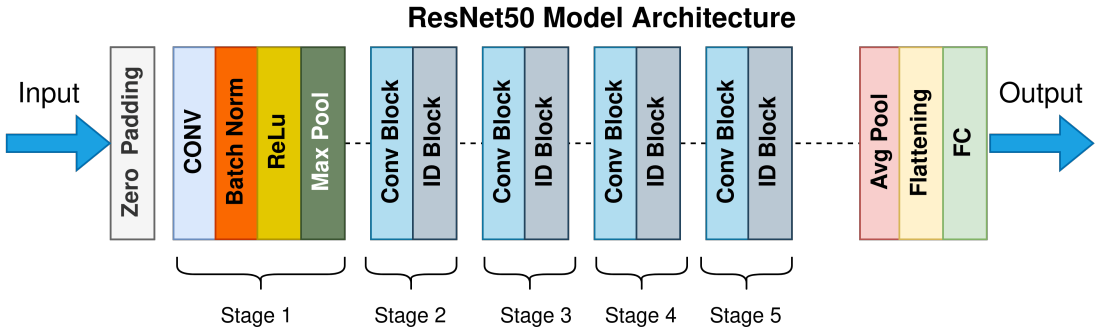

Figure 3.2: Block-level schematic of ResNet-50 used in this study, adapted from Wong's annotated diagram [21]. *Feature extraction (Phases 1 and 4):* all convolutional blocks are frozen; only the final fully connected layer is trained. *Fine-tuning (Phase 2):* `layer4` and the final fully connected layer are retrained while earlier blocks remain frozen.

The backbone for all experiments was ResNet-50 [6], a deep residual network pretrained on ImageNet. ResNet-50 was chosen because it balances accuracy and efficiency, and its residual connections mitigate vanishing gradients in deeper networks. Compared with alternatives such as MobileNetV2 [11] or EfficientDet [16], it offers stronger representational power while remaining computationally feasible. ResNet-50 has also been widely adopted in prior fire-vision studies, making it an appropriate baseline for systematic evaluation.

It is worth noting that some schematic diagrams of ResNet-50, such as Figure 3.2, show five main blocks. In PyTorch, however, the architecture is structured as four sequential residual stages (`layer1`–`layer4`), preceded by an initial convolutional stem. In this implementation, fine-tuning was performed on `layer4`—the final residual stage—and the fully connected head, while the stem and `layer1`–`layer3` remained frozen.

3.2.1 Training Regimes

Two training regimes were implemented to assess the effect of model adaptation:

- **Feature extraction (frozen base):** all convolutional layers were frozen and only the final fully connected layer was trained for binary classification. This regime was used in Phase 1 (outdoor) and Phase 4 (indoor).
- **Fine-tuning:** the final residual block (`layer4`) and the fully connected layer were unfrozen and retrained, enabling deeper layers to adapt to fire-specific features. This regime was applied in Phase 2.

The distinction between these two regimes is illustrated in Figure 3.2, which highlights the blocks that remained frozen in the feature extraction regime versus those retrained in fine-tuning.

3.2.2 Loss Function

All models were trained using the binary cross-entropy objective shown in Equation 3.2:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right], \tag{3.2}$$

where $y_i \in \{0,1\}$ is the ground-truth label and $\hat{y}_i$ is the predicted probability of fire. This loss is the standard formulation for binary classification [4] and is consistent with prior fire detection studies.

3.2.3 Optimisation and Hyperparameters

Training was standardised across all phases to ensure comparability. Each model was trained for five epochs using the Adam optimiser [9] with a learning rate of 1×10^{-4} and a batch size of 32. These values balanced computational efficiency with stable convergence and ensured fair comparisons between frozen and fine-tuned regimes. The hyperparameters are summarised in Table 3.2.

Table 3.2: Training hyperparameters applied consistently across all phases.

Hyperparameter	Value	Rationale
Backbone	ResNet-50	Accuracy–efficiency balance
Loss function	Binary cross-entropy	Standard for binary logits
Optimiser	Adam	Adaptive, stable for small datasets
Learning rate	1×10^{-4}	Matches fire vision baselines
Batch size	32	Fits GPU memory consistently
Epochs	5	Efficiency and comparability

3.3 Experimental Design

The experimental design was organised into four sequential phases, each isolating and then integrating different methodological factors. This phased structure made it possible to disentangle the effects of dataset composition, training regime, and domain shift in a way that reflects both academic inquiry and industrial deployment scenarios. By progressing from outdoor baselines to indoor deployment, the framework ensured that each research question (RQ1–RQ4, Section 1.3) was addressed systematically. The overall sequence of phases is summarised in Figure 3.3, which illustrates how the study progressed from Phase 1 (frozen outdoor baselines) through Phase 2 (fine-tuned outdoor models) and Phase 3 (domain shift) to Phase 4 (indoor deployment).

3.3.1 Phase 1: Frozen Outdoor Training

The first phase established outdoor baselines using feature extraction. Five ResNet-50 models were trained with the backbone frozen and only the final fully connected layer retrained: a real-only baseline (D-Fire), a synthetic-only baseline (Yunnan MSFFD), and three mixed configurations with 25%, 50%, and 75% synthetic samples. To ensure fairness, all mixed datasets were constrained to 5,260 samples (`seed`=42), preventing differences in dataset size from influencing results. The real-only baseline, however, was trained on the full D-Fire training pool of approximately 17,000 images, which established the real-data ceiling against which all other configurations were compared. Evaluation was performed on the fixed D-Fire test set, which served as the canonical outdoor benchmark across all outdoor experiments.

Rationale: Phase 1 isolated the impact of synthetic–real ratios in a controlled setting, directly addressing RQ1 and RQ2.

3.3.2 Phase 2: Fine-Tuned Outdoor Training

In the second phase, the same outdoor configurations were revisited but with fine-tuning enabled. The final residual block (`layer4`) and the classification head were unfrozen, allowing deeper adaptation to fire-specific features while still leveraging pretrained ImageNet knowledge. Three models were trained: a real-only baseline (D-Fire), a synthetic-only baseline (Yunnan MSFFD), and a balanced 50/50 mixture. As in Phase 1, all mixed datasets were capped at 5,260 images for fairness, while the real-only model was trained on the full D-Fire training pool of approximately 17,000 images to represent the real-data ceiling. Evaluation was carried out on the D-Fire test set to enable like-for-like comparison.

Rationale: Phase 2 tested whether fine-tuning provides measurable advantages over frozen feature extraction, and whether mixed datasets derive greater benefit from deeper adaptation (RQ1, RQ2).

3.3.3 Phase 3: Domain Shift (Outdoor to Indoor)

The third phase addressed the key industrial challenge of domain shift. Outdoor-trained models from Phases 1 and 2 (real-only, synthetic-only, and the best-performing 50/50 mix) were re-evaluated on the indoor PLOS ONE test set without retraining.

Rationale: This design quantified how far synthetic augmentation improves robustness when models face outdoor-to-indoor transfer, directly addressing RQ3. Including both frozen and fine-tuned variants revealed whether the benefits of fine-tuning extend to cross-domain generalisation.

3.3.4 Phase 4: Indoor Training for Deployment

The final phase focused directly on the deployment domain: indoor industrial environments. Two models were trained from scratch using feature extraction: one real-only model using the full PLOS ONE training set, and one mixed model using a fixed 50/50 composition of PLOS ONE real images (2,000) and SYN-FIRE synthetic positives (2,000). Both models were evaluated on the PLOS ONE test set.

Rationale: Phase 4 tested whether synthetic augmentation could reduce the amount of real indoor data required while still achieving deployment-grade performance (RQ4). The fixed 2,000+2,000 design demonstrated that near-ceiling performance could be reached with half the real data, addressing both data scarcity and cost considerations.

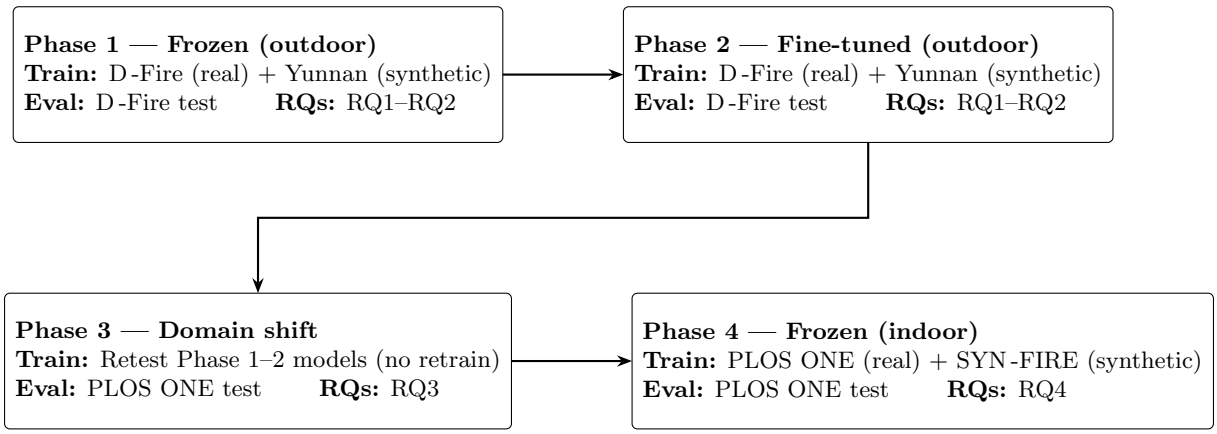

Figure 3.3: Four-phase experimental design, combining sequencing and content. Each phase lists its training sources, evaluation set, and linked research questions (RQs). Detailed mixture compositions (e.g., 25/75, 50/50, 75/25 in Phase 1; 50/50 in Phase 2; fixed 2000+2000 in Phase 4) are described in Section 3.3.

3.4 Evaluation Metrics and Visualisation

All models were evaluated with a standardised pipeline to ensure comparability across Phases 1–4. A dedicated helper module (`metrics.py`) computed lightweight per-epoch metrics during training and a full suite of extended metrics and plots during final testing. This arrangement maintained consistency and reproducibility while keeping the training notebooks uncluttered.

3.4.1 Core Metrics

Four core metrics were used throughout training and checkpointing: accuracy, precision, recall, and F1 score. Accuracy measured the overall proportion of correct predictions but was sensitive to class imbalance. Precision quantified the proportion of predicted fire images that were truly fire, which reduced false alarms. Recall (sensitivity) captured the proportion of true fire images that were correctly detected, which limited missed fires. The F1 score provided the harmonic mean of precision and recall and was adopted as the primary criterion for checkpoint selection. The precise definitions are shown in Table 3.3.

Table 3.3: Core metrics used during training and checkpointing. F1 was prioritised for model selection due to the need to balance false positives and false negatives in fire detection.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
F1 Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$

3.4.2 Extended Metrics

Final test evaluations employed a broader set of measures to characterise performance under imbalance and to provide safety-critical insight. Specificity (true negative rate) complemented recall by quantifying correct rejection of no-fire images. Error rates were expressed explicitly as the false positive rate (FPR) and false negative rate (FNR). The Matthews Correlation Coefficient (MCC) served as a balanced single-number summary that remained reliable with skewed class distributions. Two threshold-independent metrics were also reported: ROC AUC (separability of fire vs. no-fire) and PR AUC (average precision), which was more informative under imbalance. In addition, per-class precision, recall, and F1 with class supports were saved to highlight asymmetries between classes. These definitions are provided in Table 3.4.

Table 3.4: Extended metrics reported at test time. MCC provided a balanced summary under imbalance, while PR AUC reflected precision–recall behaviour relevant to safety-critical alarms.

Metric	Formula
Specificity (TNR)	$\frac{TN}{TN+FP}$
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$
Matthews Corr. Coef. (MCC)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
ROC AUC	Area under ROC curve (TPR vs. FPR)
PR AUC	Average precision across recall levels

3.4.3 Thresholding and Interpretability

Predicted probabilities $\hat{y}_i$ were binarised at a fixed decision threshold $\tau = 0.5$ across all phases, as defined in Equation 3.3. Here, $\hat{y}_i$ denotes the predicted probability of fire for the ith image, and $\widehat{y}_i$ is the corresponding binary class prediction.

$$\widehat{y}_i = \begin{cases} 1 & \text{if } \hat{y}_i \geq \tau \quad \text{(fire)}, \\ 0 & \text{if } \hat{y}_i < \tau \quad \text{(no fire)}. \end{cases} \tag{3.3}$$

This default threshold provided a balanced trade-off between precision and recall. Precision–recall curves were also produced to enable threshold tuning for site-specific deployments. Interpretability was incorporated via Grad-CAM: for selected models and representative cases, we generated triptychs (image, heatmap, overlay) to verify whether attention concentrated on flames or smoke rather than background artefacts. These visual checks supported the diagnosis of systematic false positives and false negatives, and they were referenced alongside quantitative results in Chapter 4.

3.5 Implementation and Reproducibility

All experiments were carried out in a cloud-based workflow designed to ensure consistency, transparency, and reproducibility. Google Colab provided GPU-accelerated execution, Google Drive acted as persistent storage for datasets, models, and results, and a private GitHub repository maintained version-controlled code. This combination avoided local hardware constraints while ensuring that every experiment could be reproduced independently.

3.5.1 Environment and File Structure

A structured hierarchy was established within Google Drive to organise assets systematically. Raw and processed datasets were stored under `data/raw`, model checkpoints under `models`, and evaluation outputs under `results` and `figures`. This separation enabled traceability across training, evaluation, and reporting stages, with each result directly linked to a specific model checkpoint and dataset split.

3.5.2 Version Control and Development Tools

Code and notebooks were maintained in a private GitHub repository, ensuring transparent tracking of changes and version reproducibility. Lightweight modules for dataset loading, training, and evaluation were developed locally in PyCharm before being committed, supporting efficient debugging and modular design. The repository was cloned into each Colab session to guarantee that the most recent tracked version of the codebase was used. Large artefacts such as datasets and model checkpoints were excluded via a curated `.gitignore`, ensuring the repository remained lightweight and fully reproducible.

3.5.3 Reproducibility Measures

Several measures were adopted to ensure that results could be reproduced exactly:

- Random seeds were fixed (`42`) for dataset splits and sampling.
- Mixed datasets were constrained to fixed sizes (5,260 images for outdoor experiments; 4,000 for indoor experiments).
- A consistent preprocessing policy was applied (resize to 224×224, convert to tensor, no ImageNet normalisation).
- Model checkpoints followed a systematic naming convention tied to dataset, composition, and training regime.

Together, these measures ensured that every figure and table presented in this dissertation could be traced back to a specific code commit, dataset composition, and evaluation pipeline.

3.6 Summary

This chapter described the methodological framework used to investigate how synthetic fire imagery could reduce reliance on scarce real data while maintaining deployment-grade performance. Four datasets spanning real and synthetic imagery across outdoor and indoor domains were harmonised through a consistent preprocessing pipeline. A ResNet-50 backbone was trained under two regimes—frozen feature extraction and fine-tuning—using standardised hyperparameters to ensure comparability.

The experimental design progressed through four structured phases. Phase 1 established outdoor baselines under varying synthetic–real ratios, Phase 2 evaluated the benefits of fine-tuning, Phase 3 assessed generalisation under outdoor-to-indoor domain shift, and Phase 4 tested indoor training with

synthetic augmentation for deployment. Evaluation combined core metrics (accuracy, precision, recall, F1) with extended measures such as specificity, MCC, and AUC (Section 3.4), while Grad-CAM provided interpretability by verifying that model attention focused on meaningful fire cues.

Finally, the implementation strategy integrated Colab for execution, Google Drive for structured storage, GitHub for version control, and PyCharm for module development, ensuring both transparency and reproducibility. These methodological choices created a coherent and replicable framework directly aligned with the central research question: to what extent synthetic data can reduce reliance on scarce real fire imagery while supporting reliable deployment in industrial environments. The next chapter presents the results obtained from this experimental programme.

Chapter 4

Results and Critical Evaluation

This chapter presents the empirical findings of the study and evaluates them in relation to the four research questions introduced in Section 1.3. Results are organised according to the phased experimental design described in Chapter 3, allowing each phase to address specific methodological factors.

- **Phase 1** established frozen outdoor baselines with varying synthetic–real ratios, directly informing RQ1 and RQ2.
- **Phase 2** examined fine-tuned outdoor models to test whether deeper adaptation improved performance, further addressing RQ1 and RQ2.
- **Phase 3** evaluated outdoor-trained models on indoor data to quantify the effect of domain shift, addressing RQ3.
- **Phase 4** trained indoor models using both real and synthetic imagery to assess deployment viability under reduced real-data availability, directly addressing RQ4.

In each phase, results are reported using the metrics and visualisations defined in Section 3.4, with ROC curves, precision–recall curves, and confusion matrices presented alongside scalar scores. Grad-CAM interpretability analyses are included to illustrate representative strengths and weaknesses of selected models. The chapter concludes with a synthesis table and critical appraisal across all phases, highlighting efficiency gains, residual limitations, and implications for industrial deployment.

4.1 Phase 1: Frozen Outdoor Models

Phase 1 established the frozen outdoor baselines and directly addressed RQRQ1 and RQRQ2. Models were trained using feature extraction on outdoor datasets, with only the final fully connected layer retrained. Results are reported for each model configuration, followed by a synthesis in Table 4.1.

4.1.1 Real-only (100% D-Fire)

The real-only baseline achieved an accuracy of 92.6% with F1 = 0.852 and MCC = 0.804 on the D-Fire test set. It demonstrated strong specificity (0.964) and reliable detection of no-fire cases, though fire recall was lower at 0.820. Figure 4.1 shows the ROC curve, PR curve, and confusion matrix for this model.

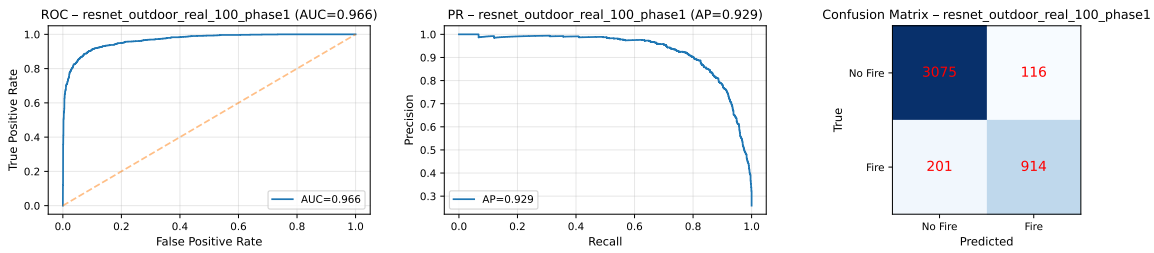

Figure 4.1: Performance of the real-only Phase 1 model on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.1.2 Synthetic-only (100% Yunnan)

The synthetic-only model transferred poorly to the D-Fire domain. It achieved extremely high recall (0.996) but very low precision (0.277), resulting in a collapse of specificity (0.090) and MCC (0.151). Nearly all no-fire cases were misclassified as fire due to dataset imbalance and the outdoor domain gap. These limitations are visible in the ROC, PR, and confusion matrix outputs in Figure 4.2.

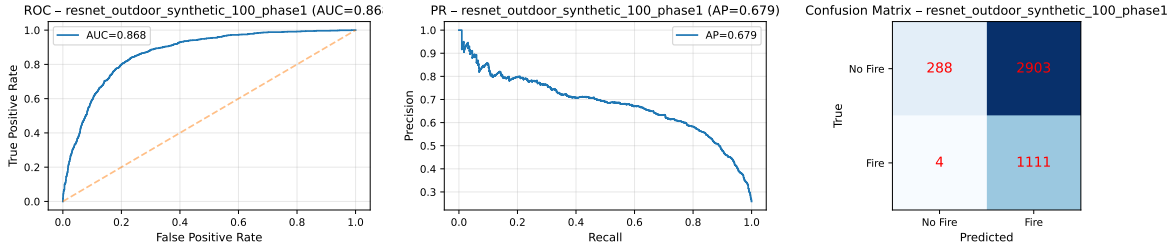

Figure 4.2: Performance of the synthetic-only Phase 1 model on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.1.3 Mixed 25% Synthetic + 75% Real

The 25/75 mixed model balanced performance effectively, achieving F1 = 0.794 and MCC = 0.732. Specificity was high (0.955) and precision strong (0.853), though recall dropped to 0.744 compared to the real-only baseline. Figure 4.3 presents the ROC curve, PR curve, and confusion matrix.

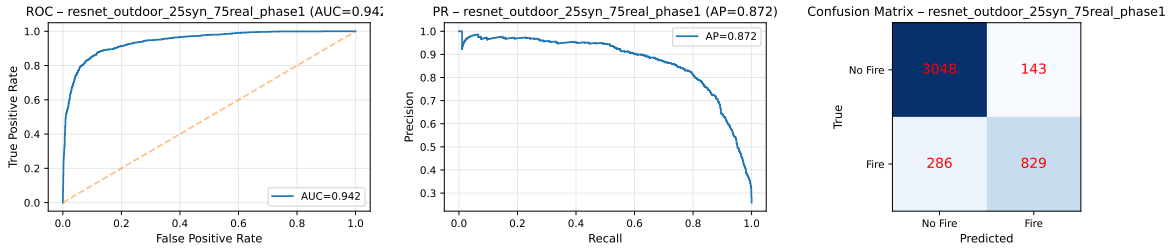

Figure 4.3: Performance of the 25/75 mixed Phase 1 model on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.1.4 Mixed 50% Synthetic + 50% Real

The 50/50 mixed model achieved F1 = 0.781 and MCC = 0.708. It maintained competitive results while using 33% less real data than the 25/75 model, highlighting the efficiency gains of synthetic augmentation. Performance outputs are shown in Figure 4.4.

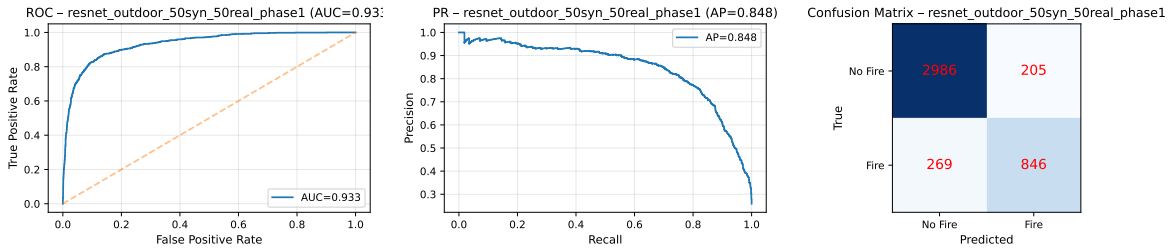

Figure 4.4: Performance of the 50/50 mixed Phase 1 model on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.1.5 Mixed 75% Synthetic + 25% Real

The 75/25 model overfit to synthetic cues. Although recall remained high (0.826), precision dropped to 0.675 and overall F1 fell to 0.743, confirming that excessive reliance on synthetic data harmed real-domain

performance. The ROC curve, PR curve, and confusion matrix are shown in Figure 4.5.

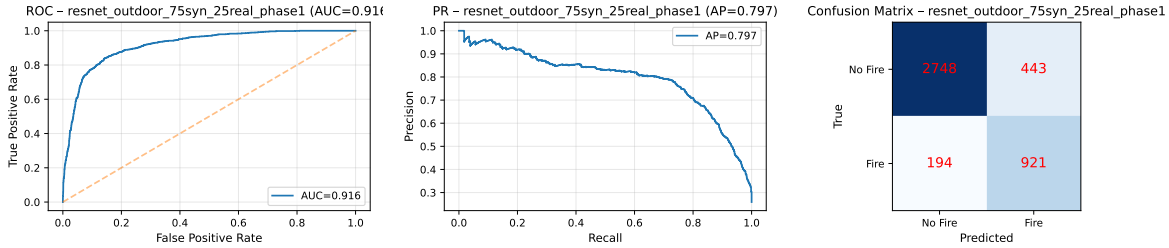

Figure 4.5: Performance of the 75/25 mixed Phase 1 model on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.1.6 Summary of Phase 1

The comparative results of all Phase 1 models are consolidated in Table 4.1. The table shows that mixed ratios stabilised performance between the extremes of real-only and synthetic-only. The 25/75 and 50/50 models offered the most balanced trade-offs, addressing RQ2 by highlighting the value of synthetic augmentation at mid ratios.

Table 4.1: Summary of Phase 1 model performance on the D-Fire test set. Mixed ratios stabilised performance between the extremes of real-only and synthetic-only, with mid-ratio models providing the most effective balance (RQ2).

Model	Accuracy	Precision	Recall	F1	ROC AUC	MCC
Real-only (100% D-Fire)	0.926	0.887	0.820	0.852	0.966	0.804
Synthetic-only (100% Yunnan)	0.325	0.278	0.996	0.433	0.868	0.151
25% Syn + 75% Real	0.900	0.853	0.744	0.794	0.942	0.732
50% Syn + 50% Real	0.889	0.805	0.759	0.781	0.933	0.708
75% Syn + 25% Real	0.852	0.675	0.826	0.743	0.916	0.647

4.2 Phase 2: Fine-Tuned Outdoor Models

In Phase 2, models were fine-tuned by unfreezing the final residual block (`layer4`) and the fully connected layer of ResNet-50, enabling deeper adaptation to the fire detection task. This phase built directly on the frozen baselines established in Phase 1 and addressed RQ1 and RQ2 by testing whether fine-tuning improved performance and whether mixed training remained effective.

4.2.1 Real-only (100% D-Fire)

The fine-tuned real-only model achieved an accuracy of 0.972, precision of 0.960, recall of 0.930, and F1 = 0.945 when evaluated on the D-Fire test set. Its specificity was very high (0.987), and MCC reached 0.926, establishing this model as the ceiling for real-only outdoor training. ROC AUC (0.992) and PR AUC (0.984) further confirmed its excellent separability (Figure 4.6).

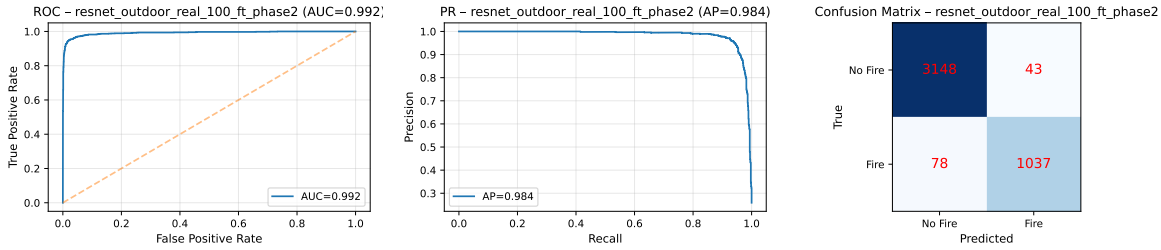

Figure 4.6: Performance of the fine-tuned real-only model (Phase 2) on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.2.2 Synthetic-only (100% Yunnan)

The fine-tuned synthetic-only model achieved substantial improvements compared to its frozen counterpart from Phase 1. It reached accuracy of 0.799, precision of 0.583, recall of 0.792, and F1 = 0.672. Although ROC AUC (0.872) and PR AUC (0.720) indicated only moderate separability, the MCC of 0.545 confirmed that the model learned transferrable cues from synthetic fire imagery. However, high false positives remained a limitation (Figure 4.7).

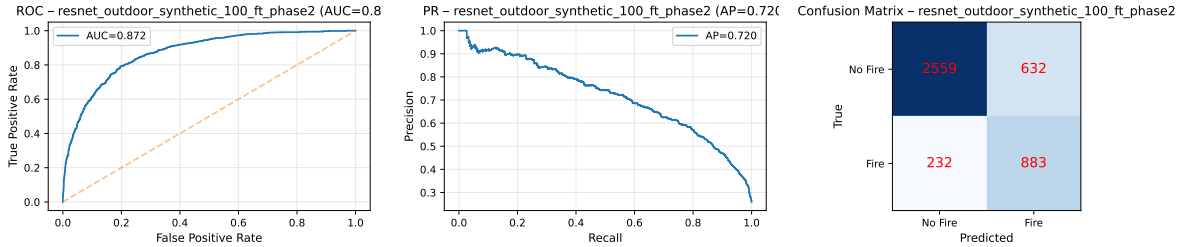

Figure 4.7: Performance of the fine-tuned synthetic-only model (Phase 2) on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.2.3 Mixed 50% Synthetic + 50% Real

The fine-tuned 50/50 mixed model achieved accuracy of 0.957, precision of 0.924, recall of 0.909, and F1 = 0.916. With specificity of 0.974 and MCC = 0.887, this model nearly matched the real-only ceiling while using only half the real training data. Its ROC AUC (0.981) and PR AUC (0.963) confirmed strong reliability (Figure 4.8).

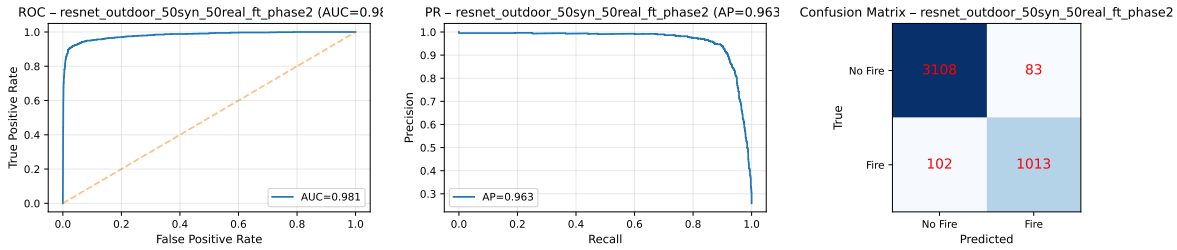

Figure 4.8: Performance of the fine-tuned 50/50 mixed model (Phase 2) on the D-Fire test set: ROC curve, PR curve, and confusion matrix.

4.2.4 Summary of Phase 2

The comparative results of all Phase 2 models are consolidated in Table 4.2. Fine-tuning improved performance across all models relative to Phase 1, with the largest gains seen for the synthetic-only baseline. The 50/50 mixed model demonstrated that synthetic augmentation maintained high effectiveness when deeper adaptation was applied, delivering strong performance with substantially reduced real data. These

findings reinforced RQ1 and RQ2, confirming that fine-tuning is beneficial and that mid-ratio mixtures remain highly effective.

Table 4.2: Summary of Phase 2 model performance on the D-Fire test set. Fine-tuning improved all models compared to Phase 1, with the 50/50 mix maintaining strong performance while halving the real-data requirement (RQ1, RQ2).

Model	Accuracy	Precision	Recall	F1	ROC AUC	MCC
Real-only (100% D-Fire)	0.972	0.960	0.930	0.945	0.992	0.926
Synthetic-only (100% Yunnan)	0.799	0.583	0.792	0.672	0.872	0.545
50% Syn + 50% Real	0.957	0.924	0.909	0.916	0.981	0.887

4.3 Phase 3: Domain Shift to Indoor

In Phase 3, the three fine-tuned outdoor models from Phase 2 (real-only, synthetic-only, and 50/50 mixed) were evaluated on the PLOS ONE indoor dataset. This tested their ability to generalise beyond the outdoor training domain and addressed RQ3 by quantifying the effect of domain shift.

4.3.1 Real-only (100% D-Fire)

The fine-tuned real-only model achieved an accuracy of 0.884 and F1 = 0.852 on the PLOS ONE test set. While precision remained very high at 0.977, recall dropped sharply to 0.756 compared with 0.930 on D-Fire, showing that many indoor fire cases were missed. This reflected strong specificity (0.986) but weaker sensitivity indoors (Figure 4.9).

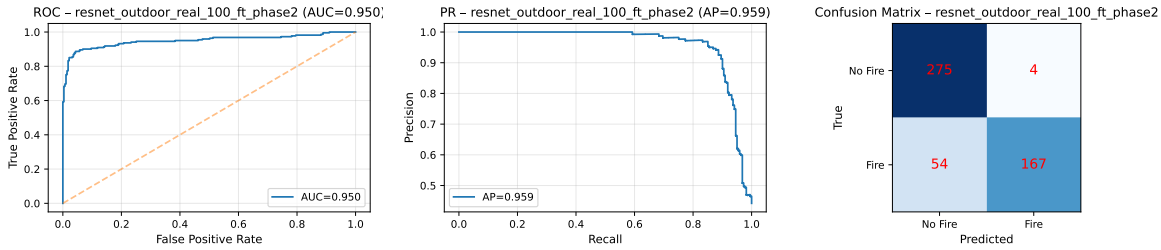

Figure 4.9: Performance of the fine-tuned real-only model (Phase 2) on the PLOS ONE indoor test set. The ROC and PR curves confirm high separability, but the confusion matrix highlights reduced recall indoors.

4.3.2 Synthetic-only (100% Yunnan)

The fine-tuned synthetic-only model performed poorly under domain shift, with accuracy of 0.624 and F1 = 0.624. Recall remained moderate at 0.706, but precision fell to 0.559 and MCC dropped to 0.265. These results confirm that while synthetic training produced transferable fire cues, it generalised weakly in indoor environments (Figure 4.10).

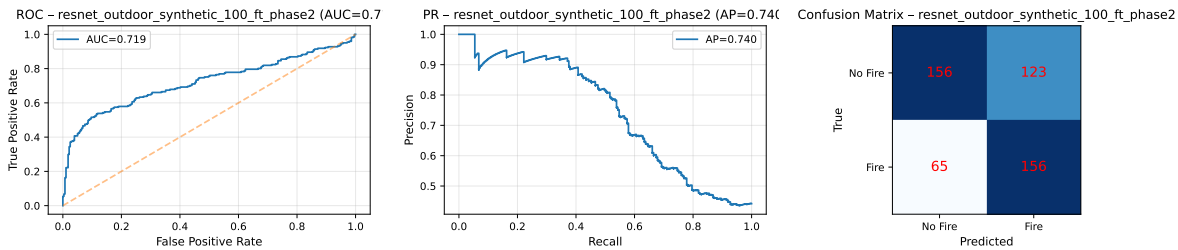

Figure 4.10: Performance of the fine-tuned synthetic-only model (Phase 2) on the PLOS ONE indoor test set. The model transferred poorly, with high false positives evident in the confusion matrix.

4.3.3 Mixed 50% Synthetic + 50% Real

The fine-tuned 50/50 mixed model achieved the best indoor generalisation. It reached accuracy of 0.902, precision of 0.978, recall of 0.796, and F1 = 0.878. Its ROC AUC (0.975) and PR AUC (0.975) were higher than those of the real-only model, confirming stronger class separation under domain shift. This level of performance was achieved while using only half the real training data (Figure 4.11).

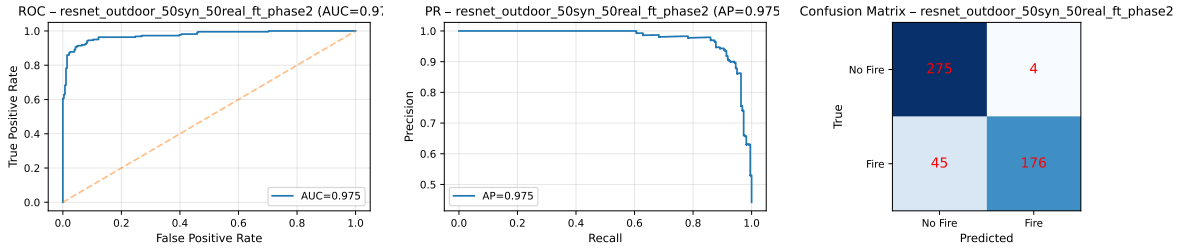

Figure 4.11: Performance of the fine-tuned 50/50 mixed model (Phase 2) on the PLOS ONE indoor test set. The model delivered the best balance between precision and recall under domain shift.

4.3.4 Summary of Phase 3

The comparative results are consolidated in Table 4.3. The real-only model provided a precision ceiling but lost recall indoors, while the synthetic-only model collapsed almost entirely under domain shift. By contrast, the 50/50 mixed model delivered the best trade-off, sustaining high precision and stronger recall, and highlighting the value of synthetic augmentation for cross-domain generalisation. These findings directly support RQ3.

Table 4.3: Summary of Phase 3 model performance on the PLOS ONE indoor test set. The 50/50 mixed model achieved the best balance under domain shift, confirming the value of synthetic augmentation for cross-domain generalisation (RQ3).

Model	Accuracy	Precision	Recall	F1	ROC AUC	MCC
Real-only (100% D-Fire FT)	0.884	0.977	0.756	0.852	0.950	0.776
Synthetic-only (100% Yunnan FT)	0.624	0.559	0.706	0.624	0.719	0.265
50% Syn + 50% Real FT	0.902	0.978	0.796	0.878	0.975	0.809

4.4 Phase 4: Indoor Training and Deployment

Phase 4 evaluated models trained directly on indoor imagery to establish a deployment-ready benchmark. Two models were considered: a real-only indoor baseline and a 50/50 mixed indoor model augmented with SYN-FIRE synthetic data. This phase addressed RQ4 by testing whether synthetic augmentation could achieve deployment-grade accuracy while reducing the number of real indoor images required.

4.4.1 Indoor Real-only (100% PLOS ONE)

The indoor real-only model achieved near-perfect performance on the PLOS ONE test set, with accuracy of 0.982, precision of 0.973, recall of 0.986, and F1 = 0.980. Its ROC AUC (0.9993) and PR AUC (0.9991) confirmed almost complete class separation. Errors were minimal, with just six false positives and three false negatives across 500 test samples (Figure 4.12).

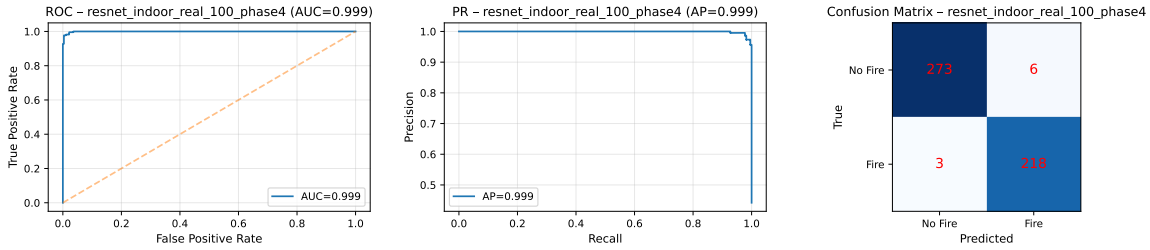

Figure 4.12: Performance of the indoor real-only model (Phase 4) on the PLOS ONE test set. Metrics confirm excellent separability, although the model still produced a small number of errors.

4.4.2 Indoor Mixed 50% Synthetic + 50% Real

The 50/50 mixed indoor model achieved accuracy of 0.986, precision of 0.991, recall of 0.977, and F1 = 0.984. Its MCC (0.972) and AUC values (ROC = 0.9990, PR = 0.9987) demonstrated deployment-grade reliability. Importantly, this performance was achieved using only 2,000 real indoor samples, surpassing the real-only baseline while requiring substantially less real data (Figure 4.13).

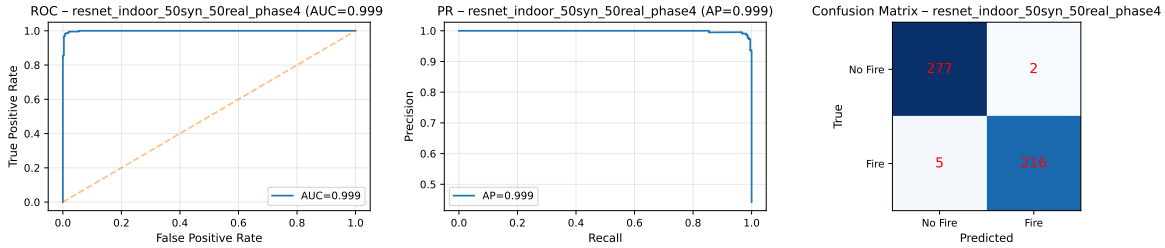

Figure 4.13: Performance of the indoor 50/50 mixed model (Phase 4) on the PLOS ONE test set. The model slightly outperformed the real-only baseline while using half the real data, confirming the value of synthetic augmentation.

4.4.3 Summary of Phase 4

The comparative results are summarised in Table 4.4. While the real-only model established a strong in-domain ceiling, the 50/50 mixed model achieved slightly higher F1 and MCC while using only half the real training data. This confirms that synthetic augmentation not only reduces data requirements but also supports deployment-ready fire detection systems, directly addressing RQ4.

Table 4.4: Summary of Phase 4 indoor model performance on the PLOS ONE test set. The 50/50 mixed model achieved deployment-grade performance while halving the real data requirement (RQ4).

Model	Accuracy	Precision	Recall	F1	ROC AUC	MCC
Indoor Real-only (100% PLOS ONE)	0.982	0.973	0.986	0.980	0.999	0.964
Indoor Mixed 50% Syn + 50% Real	0.986	0.991	0.977	0.984	0.999	0.972

4.5 Interpretability with Grad-CAM

To complement the quantitative results presented in Phases 1–4, Gradient-weighted Class Activation Mapping (Grad-CAM) [12] was applied to selected models. Grad-CAM highlights spatial regions in an image that most influenced the model's prediction, offering interpretability in safety-critical applications such as fire detection. Given page constraints, three representative panels were selected for clarity and explanatory value. Rather than exhaustively presenting every true/false case, these panels illustrate the most informative behaviours across phases: a successful detection, a domain-shift limitation, and a near-miss from the final deployment model.

Phase 1: Real-Only Baseline (Outdoor, D-Fire)

For the frozen extractor trained on D-Fire only, a true positive case was selected. The Grad-CAM heatmap (Figure 4.14) aligned strongly with visible flames, confirming that even without fine-tuning the baseline model captured relevant fire cues. This provides evidence that frozen ImageNet features were already effective for identifying clear fire regions, complementing the quantitative metrics reported in Section 4.

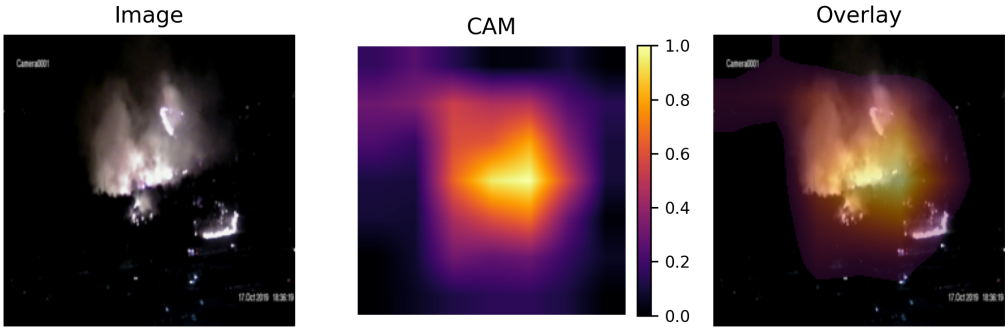

Figure 4.14: Grad-CAM visualisation for the Phase 1 real-only model on D-Fire. Heatmap concentrated on flames, showing that frozen features still captured fire cues.

Phase 2: 50/50 Mixed Fine-Tuned (Outdoor → Indoor, PLOS ONE)

For the fine-tuned 50/50 model under domain shift, a false negative case was chosen. The heatmap in Figure 4.15 showed weak activations around the flame area and stronger responses to background structures, explaining why this indoor fire was missed. This illustrates the recall limitations observed in Section 4, where even the best generalising Phase 2 model struggled with subtle indoor fires. These findings relate directly to RQ3, showing that synthetic augmentation improved transfer but did not eliminate all recall loss.

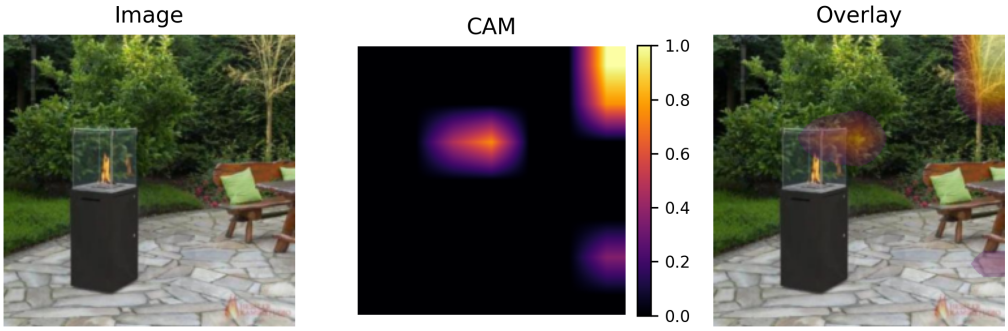

Figure 4.15: Grad-CAM for the Phase 2 50/50 mixed model on PLOS ONE. Attention diffused to background, missing subtle flames, consistent with recall loss under domain shift.

Phase 4: 50/50 Mixed Indoor (Deployment Candidate)

For the final deployment-ready model trained on 2,000 real and 2,000 synthetic indoor images, a false negative case was included to reflect realistic deployment limitations. As shown in Figure 4.16, the heatmap partially overlapped with faint smoke regions but missed the small flames, leading to misclassification. This confirms that while the Phase 4 model achieved near-perfect metrics, rare errors remained in subtle early-stage fire scenarios, which has direct implications for RQ4.

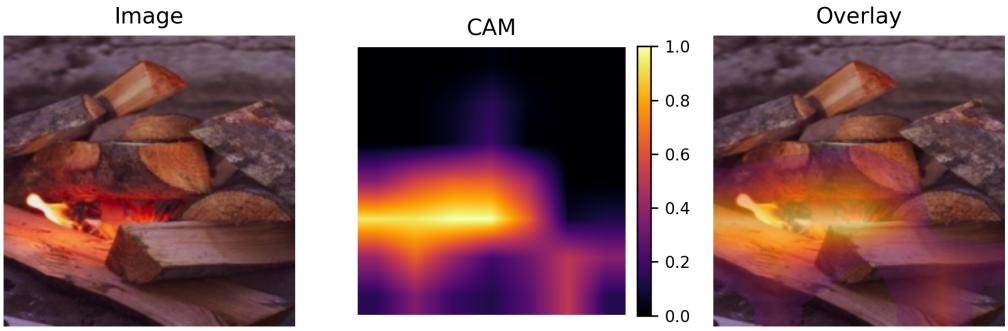

Figure 4.16: Grad-CAM for the Phase 4 indoor 50/50 model on PLOS ONE. Heatmap partially aligned with smoke but missed faint flames, showing a rare deployment limitation.

Summary

The three panels illustrate complementary interpretability insights. First, real-only frozen baselines were already capable of capturing visible flames. Second, mixed fine-tuned models generalised more effectively but remained vulnerable to domain-shift false negatives. Third, the final indoor 50/50 deployment model was highly effective but not flawless, occasionally overlooking subtle fire cues. Together, these results demonstrate that Grad-CAM not only diagnosed model weaknesses but also reinforced trust in the deployment recommendation by showing that successful predictions aligned with meaningful visual features.

4.6 Cross-Phase Synthesis and Chapter Summary

The four phases collectively addressed the research questions introduced in Section 1.3, moving from frozen outdoor baselines (Phase 1) to fine-tuned models (Phase 2), domain-shift evaluations (Phase 3), and indoor deployment candidates (Phase 4). Table 4.5 consolidates the most important results, comparing the real-only ceilings against their 50/50 mixed counterparts. This highlights both the efficiency gains from synthetic augmentation and the generalisation benefits under domain shift. Alongside F1 and MCC as balanced performance indicators, false positives (FP) and false negatives (FN) are reported to show the safety-critical trade-offs between false alarms and missed fires. Because the D-Fire test set contained over 4,000 images while the PLOS ONE indoor test set contained only 500, absolute error counts are not directly comparable across domains, but they remain useful for illustrating relative trends within each dataset.

Table 4.5: Cross-phase synthesis of key models. The table highlights data-efficiency gains, generalisation across domains, and error counts. False positives (FP) represent false alarms, while false negatives (FN) represent missed fires. Error counts are not directly comparable across datasets due to different test set sizes (D-Fire $\approx$ 4,300 images; PLOS ONE = 500 images).

Phase / Model	Real imgs (k)	Test domain	F1	MCC	FP / FN
Phase 1 Real-only	13.0	D-Fire (outdoor)	0.852	0.804	116 / 201
Phase 1 50/50 Mixed	2.6	D-Fire (outdoor)	0.781	0.708	205 / 269
Phase 2 Real-only	13.0	D-Fire (outdoor)	0.945	0.926	43 / 78
Phase 2 50/50 Mixed	2.6	D-Fire (outdoor)	0.916	0.887	83 / 102
Phase 3 Real-only	13.0	PLOS ONE (indoor)	0.852	0.776	4 / 54
Phase 3 50/50 Mixed	2.6	PLOS ONE (indoor)	0.878	0.809	4 / 45
Phase 4 Real-only	4.0	PLOS ONE (indoor)	0.980	0.964	6 / 3
Phase 4 50/50 Mixed	2.0	PLOS ONE (indoor)	**0.984**	**0.972**	2 / 5

Across phases, three consistent patterns emerged. First, fine-tuning improved all models compared to their frozen baselines, confirming RQ1. Second, balanced synthetic–real mixtures delivered strong

performance with markedly fewer real images—for example, in Phase 2 the 50/50 model required only 2.6k real images compared to 13k for the real-only ceiling—supporting RQ2. Third, under domain shift to indoor imagery, the fine-tuned 50/50 model generalised better than the fine-tuned real-only equivalent (F1 = 0.878 vs. 0.852 in Table 4.5, Phase 3 rows), addressing RQ3. Finally, Phase 4 confirmed that synthetic augmentation enabled deployment-grade indoor performance using only 2,000 real images, with the mixed model slightly outperforming the real-only indoor baseline (F1 = 0.984 vs. 0.980), directly answering RQ4.

Chapter 5

Discussion and Future Work

This chapter moves beyond the presentation of results to critically interpret their significance in relation to the project aims and research questions. While Chapter 4 reported empirical performance across four experimental phases, the present discussion synthesises these findings to evaluate how synthetic data supports industrial fire detection, particularly under domain shift from outdoor to indoor settings. The analysis highlights the advantages and limitations of mixed synthetic–real training, draws connections with the existing literature, and considers the implications for deployment in safety-critical contexts. Limitations of the study are outlined, and directions for future research are proposed. Together, these elements provide a critical perspective that frames the contributions of the thesis within both academic and industrial contexts.

5.1 Synthetic Data Effectiveness

The first research question (RQ1) examined whether synthetic imagery could substitute or complement real data when training fire classifiers. Across Phases 1, 2, and 4, the results showed that synthetic data was most effective when used in combination with real imagery rather than as a complete replacement. Synthetic-only models transferred poorly to real domains, achieving very high recall but extremely low precision due to excessive false positives. This behaviour mirrors findings in prior work [10, 7, 1], which similarly reported that purely synthetic training introduced transferable fire cues but lacked grounding in real-world negatives.

Mixed training strategies produced a more balanced outcome. In Phase 2, the fine-tuned 50/50 model achieved F1 = 0.916 on the D-Fire test set while requiring only one-fifth of the real samples used by the real-only ceiling (Table 4.5). In Phase 4, the indoor 50/50 model not only matched but slightly exceeded the real-only counterpart (F1 = 0.984 vs. 0.980), despite relying on half as many real images. These findings confirm that synthetic data did not simply serve as filler but contributed meaningful diversity that improved generalisation and reduced the volume of real data required.

Taken together, the evidence demonstrates that synthetic imagery is best understood as a complementary resource. While it cannot fully replace real-world datasets, carefully balanced mixtures consistently achieved high performance at substantially lower real-data costs. This confirms the practical value of synthetic augmentation in industrial contexts, where collecting diverse fire imagery is difficult, hazardous, and expensive.

5.2 Domain Shift and Generalisation

The third research question (RQ3) examined how well models trained on outdoor datasets would generalise to indoor imagery, where deployment is most critical. Results from Phase 3 confirmed that domain shift remained a substantial challenge. The fine-tuned real-only model achieved high precision (0.977) but its recall dropped sharply to 0.756 when evaluated on the PLOS ONE indoor test set, indicating that many true fire cases were missed. In contrast, the fine-tuned 50/50 model delivered superior balance, with F1 = 0.878 and higher ROC and PR AUC values than the real-only counterpart (Table 4.5). This demonstrates that synthetic augmentation mitigated the decline in recall and preserved reliability under domain transfer.

These findings are consistent with wider evidence in the literature. Vasconcelos et al. [18] reported that models trained on controlled datasets often collapsed in uncontrolled conditions, while Arlovic et al. [1] showed that mixed synthetic–real training improved segmentation robustness in unseen warehouse environments. Similarly, Park et al. [10] observed that GAN-based wildfire imagery enhanced classification stability in noisy outdoor contexts. The consistent theme across these studies, and confirmed here, is that broader training distributions reduce sensitivity to dataset bias and enable better adaptation to novel domains.

Grad-CAM analysis further illustrated the mechanics of domain shift. In one false negative case, the fine-tuned 50/50 model directed attention towards background structures rather than the subtle flame region, explaining the observed recall loss. This suggests that while synthetic diversity reduces the extent of domain collapse, it does not fully eliminate the risk of overlooking visually ambiguous or small-scale fires in unfamiliar indoor settings.

Overall, the results indicate that synthetic augmentation was effective for improving generalisation, particularly by supporting recall in new domains. However, complete resolution of domain shift is likely to require both greater diversity in training distributions and extensions such as temporal modelling, which could exploit motion cues to separate subtle fire signals from background clutter.

5.3 Deployment Implications

The final research question (RQ4) asked whether synthetic augmentation could enable deployment-grade performance while reducing reliance on scarce real indoor imagery. Phase 4 provided clear evidence that this was achieved. The indoor real-only model performed strongly with F1 = 0.980, but the mixed 50/50 model reached an even higher F1 = 0.984 while using only half as many real samples (2k vs. 4k). Its Matthews Correlation Coefficient (0.972) and near-perfect ROC and PR AUC values confirmed stable and reliable classification (Table 4.5). Error counts were minimal, with only two false positives and five false negatives across 500 test images. These results confirm that synthetic augmentation was not merely a stopgap but an effective means of reaching deployment standards with reduced real-data requirements.

The implications for industrial adoption are significant. Collecting thousands of diverse indoor fire images is both impractical and unsafe, whereas generating synthetic imagery can be scaled cost-effectively. By halving the amount of real imagery required, the mixed indoor model demonstrated a pathway to more efficient development of fire detection systems that still meet the reliability demanded in safety-critical contexts. This outcome directly addresses the objectives set out in the SYNOPTIX brief, which emphasised scalability and deployability.

Interpretability analysis reinforced these findings. Grad-CAM visualisations showed that the mixed indoor model concentrated attention on flames and smoke in successful cases, supporting stakeholder trust in its predictions. At the same time, a small number of false negatives highlighted the residual risk of missing subtle or early-stage fires. In practice, such limitations could be mitigated through system-level safeguards, including threshold calibration, temporal smoothing, or integration with complementary sensor modalities such as smoke or heat detectors.

Taken together, the deployment experiments demonstrated that synthetic augmentation enabled practical fire detection systems to be trained with limited real data while sustaining the accuracy required for operational use. This represents a major step toward scalable industrial deployment and provides clear guidance on how synthetic data can be integrated into future fire detection pipelines.

5.4 Limitations and Future Directions

While the results confirmed that synthetic augmentation reduced reliance on scarce real data and enabled deployment-ready models, several limitations constrain interpretation and highlight opportunities for further research.

First, the available synthetic datasets imposed restrictions on class balance. The SYN-FIRE collection contained only fire-positive images, meaning that all negative context came from the real PLOS ONE dataset. Similarly, the Yunnan dataset was skewed toward positives, with over 80% of samples containing flames. These imbalances contributed to the tendency of synthetic-only models to overpredict fire, producing high recall but poor precision. Future work should expand synthetic datasets to include both fire and no-fire scenes, or apply rebalancing strategies such as resampling or SMOTE to improve discrimination.

Second, the preprocessing and training design were deliberately simplified to ensure comparability across phases, but this may have limited peak performance. ImageNet-specific normalisation was omitted in the final experiments to avoid instability, yet reinstating it could provide incremental gains in fine-tuned models. Training was also capped at five epochs with a single random seed, prioritising efficiency and reproducibility over full convergence. Future studies should explore longer training schedules, hyperparameter optimisation, and repeated runs under different seeds to assess stability.

Third, the scope of this project was restricted to binary image classification. This was sufficient to evaluate the role of synthetic augmentation but excluded richer tasks such as detection and segmentation. Prior work has shown that synthetic imagery is equally valuable in these domains, and extending the methodology to multi-class settings (e.g. fire, smoke, flare, reflection) or to temporal models that exploit motion cues would strengthen industrial applicability.

Finally, the evaluation pipeline assumed a fixed decision threshold of 0.5 across all contexts. While this ensured comparability, Grad-CAM analysis revealed that some subtle or early-stage fires were still missed. In practice, thresholds should be calibrated per deployment site, potentially guided by precision–recall curves to prioritise sensitivity in safety-critical contexts. Complementary safeguards such as temporal smoothing, uncertainty estimation, or integration with non-visual sensors could further reduce false negatives. Active learning also represents a promising avenue, allowing models to adapt dynamically as new synthetic and real data become available.

In summary, the limitations of this study are closely tied to opportunities for future enhancement. Addressing dataset imbalance, refining training design, extending beyond binary classification, and calibrating thresholds would all strengthen the reliability and generalisability of synthetic-augmented fire detection systems. These extensions would not only consolidate the findings of this thesis but also accelerate progress toward scalable, trustworthy solutions for industrial safety.

Chapter 6

Conclusion

This thesis investigated how synthetic fire imagery can reduce reliance on scarce real-world data while sustaining deployment-grade performance in industrial fire detection. The central aim was to evaluate whether carefully balanced synthetic–real mixtures could support reliable classification, particularly under domain shift from outdoor to indoor environments.

Four research questions guided the study. First, to what extent can synthetic data substitute or complement real imagery? Results from Phases 1 and 2 showed that synthetic-only models provided transferable cues but suffered from high false positives, while mixed datasets consistently outperformed either source alone. Synthetic data was therefore most effective as a complement rather than a replacement.

Second, what balance of synthetic and real data delivers the most effective performance? Evidence pointed clearly to mid-ratio mixtures. In Phase 2, the fine-tuned 50/50 outdoor model reached F1 = 0.916 on the D-Fire test set while requiring only one-fifth of the real data used by the real-only baseline. In Phase 4, the indoor 50/50 model slightly surpassed the real-only counterpart (F1 = 0.984 vs. 0.980) while halving the real data requirement. These results demonstrate that balanced mixtures provide both efficiency and accuracy.

Third, how well do models trained with synthetic augmentation cope with domain shift? Phase 3 confirmed that synthetic augmentation improved generalisation: the fine-tuned 50/50 model achieved F1 = 0.878 on the PLOS ONE indoor test set, outperforming the real-only equivalent (0.852) and retaining higher recall. Although domain shift remained a source of errors, particularly for subtle indoor fires, synthetic diversity reduced performance loss.

Fourth, can synthetic augmentation enable deployment-grade accuracy while using fewer real images? Phase 4 provided a definitive answer. The indoor 50/50 model achieved near-ceiling performance (F1 = 0.984, MCC = 0.972) with only 2,000 real indoor images, confirming that synthetic augmentation enabled scalable, cost-effective training pipelines. This model represents a deployment-ready solution that directly addresses the industrial requirements of SYNOPTIX.

Taken together, the project makes three main contributions. It provides a systematic evaluation of synthetic–real data mixtures in fire classification, showing that balanced combinations achieve high accuracy with substantially less real data. It demonstrates that synthetic augmentation mitigates domain shift and improves generalisation across environments. Finally, it delivers a deployment-ready model trained on a 50/50 indoor mixture, offering practical guidance for industry partners.

The findings also point to future opportunities. Expanding synthetic datasets to include negative scenes, reintroducing normalisation strategies, and exploring longer training horizons could yield further improvements. Extending beyond binary classification to segmentation, detection, and temporal models would provide richer information for industrial monitoring. Threshold calibration, uncertainty quantification, and integration with complementary sensors offer additional safeguards for deployment in safety-critical contexts.

In conclusion, this thesis has shown that synthetic data is not simply a stopgap for limited real imagery but a strategic enabler of scalable, effective, and deployable fire detection systems. By combining real and synthetic data in balanced proportions, it is possible to deliver models that are accurate, efficient, and generalisable, advancing the state of the art in AI for industrial safety and providing a foundation for future research and implementation.

Bibliography

[1] M. Arlovic and others. Syn-fire: Synthetic dataset and segmentation for indoor fire and smoke detection. *Fire Technology*, 2025.

[2] T. Chen, W. Yin, Y. Huang, and Y. Ye. Fire detection in video sequences using a generic color model. *Fire Safety Journal*, 39(5):435–445, 2004.

[3] P. Foggia, A. Saggese, and M. Vento. Real-time fire detection for video surveillance applications using a combination of experts based on color, shape, and motion. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(9):1545–1556, 2015.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[5] M. Hashem, B. Pradhan, K. Maulud, and A. Alamri. Early fire detection: A review of conventional and ai-based approaches. *Environmental Hazards*, 19(3):245–272, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] W. Hu and others. Firefly: A synthetic dataset for ember detection in wildfire. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2050–2060. IEEE, 2023.

[8] J. Kim and others. Early fire detection system by using automatic synthetic dataset generation model based on digital twins. *Applied Sciences*, 14(3):1045, 2024.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[10] H. Park and others. Advanced wildfire detection using generative adversarial network-based augmented datasets and weakly supervised object localization. *IEEE Access*, 10:12756–12769, 2022.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[14] S. Sozol, R. Mondal, and A. Thamrin. Indoor fire and smoke detection based on optimized yolov5. *PLOS ONE*, 2025.

[15] K. Staniszewski and others. Searching for the ideal recipe for preparing synthetic data in the multi-object detection problem. *Applied Sciences*, 13(6):3798, 2023.

[16] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790, 2020.

[17] S. Torabi, Y. Yilmaz, V. Atalay, and A. Cetin. Fire detection using statistical color model in video sequences. *Journal of Visual Communication and Image Representation*, 24(7):857–863, 2013.

[18] D. Vasconcelos. A survey on deep learning methods for fire detection in images and videos. *Fire*, 7(3):126, 2024.

[19] V. Venâncio, C. Silla, and A. Koerich. D-fire: An automatic fire detection system based on deep convolutional neural networks for low-power, resource-constrained devices. *Sensors*, 22(13):4729, 2022.

[20] VisionPlatform.ai. Fire and smoke detection camera. https://visionplatform.ai/fire-smoke-detection-camera, 2025. Accessed: August 2025.

[21] Terence Wong. The annotated resnet-50. https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758, 2020. Accessed: August 2025.