UNIVERSITÀ DEGLI STUDI DI CATANIA

DIPARTIMENTO DI MATEMATICA E INFORMATICA

DOTTORATO DI RICERCA IN INFORMATICA (INTERNAZIONALE) XXXVII CICLO

Salvatore Fabio Onorato

Segmentation and Anomaly Detection in Solar Plants by UAV: A Framework Based on Deep Neural Networks and Multimodal Dataset PV30

PH.D. THESIS

Tutor: Prof. Sebastiano Battiato

Anno Accademico 2023 - 2024

Abstract

The growing global demand for sustainable energy solutions has led to the rapid expansion of photovoltaic (PV) installations across residential, commercial, and utility-scale infrastructures. As these systems scale in size and complexity, ensuring their optimal performance and operational reliability becomes increasingly critical. One of the key challenges in this context is the early detection of faults and performance degradation, which, if left unresolved, can significantly impact the energy yield and lifespan of PV systems.

Conventional inspection methods, such as manual visual assessments and handheld thermal imaging, are inherently labor-intensive, time-consuming, and impractical for large-scale deployment. These limitations have catalyzed the adoption of Unmanned Aerial Vehicles (UAVs), or drones, equipped with high-resolution RGB and thermal cameras for aerial inspections. Complemented by advances in artificial intelligence (AI) and deep learning, this technological convergence has enabled the development of automated, data-driven pipelines for PV panel monitoring.

Despite the increasing interest in applying computer vision techniques to solar plant inspection, several significant gaps remain in the research landscape. First, the majority of existing studies rely on proprietary or limited-access datasets, which often lack multimodal imaging (RGB + IR), pixel-level ground truth annotations, or sufficient diversity in operational scenarios. Second, performance evaluation of deep learning models is frequently conducted in isolated contexts, without standardized benchmarks or rigorous comparisons across architectures, which hinders reproducibility and fair

assessment.

This thesis addresses these limitations through two key contributions. Firstly, we introduce **PV30**, a novel, publicly available dataset comprising over 5,000 high-resolution RGB and thermal infrared images acquired from UAV missions over real-world PV installations. The dataset includes detailed pixel-level annotations for semantic segmentation of solar panels, as well as bounding box annotations for thermal anomaly detection. The acquisition process was meticulously designed to capture a wide range of environmental and operational conditions—including different lighting scenarios, panel orientations, soiling levels, and fault types—ensuring the dataset's representativeness and utility for both academic and industrial applications.

Secondly, we perform an extensive benchmarking of state-of-the-art deep learning architectures across the two primary tasks: semantic segmentation and thermal anomaly detection. For segmentation, we evaluate models such as U-Net++, DeepLabV3+, PSPNet, PAN, and SegFormer, comparing their performance in terms of Intersection-over-Union (IoU), precision, recall, and F1-score. For anomaly detection, we test leading object detection models including YOLOv5/YOLOv11, DINO, Grounding DINO, and GLIP, assessing their mean Average Precision (mAP), inference latency, and robustness to noise and environmental variability.

In addition, we propose a reproducible and extensible **benchmarking framework** that includes automated pipelines for model training, validation, and visualization of results. The framework is designed to support multimodal data fusion, allowing researchers to test early and late fusion strategies between RGB and thermal inputs. This toolset not only facilitates objective performance comparison but also serves as a foundation for future research in smart inspection systems.

Overall, this study aims to bridge the gap between academic research and real-world deployment of AI-driven PV inspection solutions. By providing both a robust dataset and a standardized evaluation methodology, this work contributes to accelerating the development of scalable, intelligent, and reliable monitoring systems for next-generation solar energy infrastructure.

Contents

Chapter 1

Introduction

1.1 Motivation

In recent decades, climate change and the growing need to reduce greenhouse gas emissions have accelerated the global transition towards renewable energy sources. Among these, photovoltaic (PV) solar energy has emerged as a leading solution due to its scalability, low system maintenance requirements, and suitability for integration across urban, rural, and industrial settings. As a result, PV systems have experienced rapid global deployment and now represent an increasingly significant share of total energy production.

However, the long-term reliability and efficiency of photovoltaic installations can be compromised by various factors, including material degradation, harsh environmental conditions, dust accumulation, partial shading, and malfunctioning components such as bypass diodes or damaged solar cells. These issues can lead to energy yield losses or, in severe cases, permanent system failures. To ensure optimal plant operation and maximize return on investment, the timely detection of such anomalies is crucial.

Traditionally, PV system monitoring has relied on manual inspections or the use of fixed instrumentation, such as environmental sensors and smart inverters. While effective in limited contexts, these approaches are often impractical at scale due to high labor demands, long response times, and logistical challenges. In contrast, more agile and automated solutions are gaining prominence—particularly aerial inspections using drones equipped with optical and thermal sensors. These platforms enable the rapid acquisition of high-resolution imagery across large surfaces, allowing for the detection of defects that are invisible to the naked eye and facilitating the creation of comprehensive datasets for downstream analysis.

In parallel, recent advances in artificial intelligence—particularly in the field of deep learning—have opened new frontiers for the automated processing of aerial imagery. Techniques such as semantic segmentation, classification, and object detection are increasingly employed to automate the recognition of PV modules and the identification of anomalies such as hot spots, degraded cells, or partially obstructed panels. Nonetheless, the lack of suitable datasets remains one of the main barriers to the widespread adoption of these technologies.

Specifically, most existing datasets are limited to RGB imagery and often lack accurate pixel-level annotations. This significantly impairs the ability of models to learn meaningful features for thermal diagnosis, preventing full exploitation of multispectral imaging capabilities. Furthermore, model performance is typically evaluated in a fragmented manner, with no unified benchmarking framework to enable direct and rigorous comparisons across different solutions.

1.2 Objectives and Contributions

This dissertation aims to address the aforementioned challenges by making a significant contribution to the advancement of automated monitoring technologies for photovoltaic (PV) systems. Specifically, the objective is to develop a methodological and experimental platform that enables a systematic analysis of the role of multimodal imaging and deep learning techniques in the detection of thermal anomalies in solar panels. To this end, the research is structured around three main contributions:

1. **Development of the PV30 Dataset**:

One of the central elements of this dissertation is the creation of **PV30**, a new multimodal dataset composed of nine flight missions over real photovoltaic (PV) installations, acquired using a drone equipped with high-resolution RGB and thermal sensors. The images cover a wide range of environmental conditions (sunlight, cloud cover, module tilt, presence of dirt) and include detailed annotations of the solar panels and the observed anomalies. The defect classes comprise single and multiple hot spots, bypass diode anomalies, localized shading, and progressive degradation.
The dataset was manually annotated with the support of industry experts and validated using certified thermographic technologies, making it a potential benchmark for both the scientific and industrial communities.

2. **Benchmarking of Deep Learning Models**:

Leveraging the PV30 dataset, this dissertation systematically examines and compares numerous deep learning models, selected among the most representative in the field of computer vision. These include Fully Convolutional Networks, encoder-decoder architectures such as U-Net and DeepLab, detection networks like YOLO and Faster R-CNN, as well as multimodal architectures employing both early and late fusion strategies between RGB and thermal data. The objective is to identify the most effective architectures for each task (segmentation vs. detection) and for each configuration (RGB only, thermal only, RGB+thermal).

Model performance is evaluated using standard metrics (IoU, precision, recall, F1-score), with particular attention given to robustness under varying environmental conditions and the presence of noise in the data.

3. **Proposal of an Evaluation Framework**:

To address the methodological gap in the current literature, an open and reproducible benchmarking framework is introduced, enabling objective and transparent comparison of different approaches. This framework includes automated pipelines for training, validation, and result visualization, and is designed to be easily extendable to new datasets or models. This contribution serves as a foundation for future academic research and industrial applications, fostering the development of intelligent monitoring systems that are scalable and adaptable to various geographical and technical contexts.

1.3 Thesis Outline

The structure of the dissertation reflects the evolution of the research project and progressively guides the reader from the problem definition to the experimental validation of the obtained results.

- **Chapter 1 – Introduction**: Presents the motivation, objectives, and contributions of the research, providing an overview of the scientific and applicative context.
- **Chapter 2 – State of the Art**: Reviews the existing literature on PV monitoring systems, drone-based thermal imaging, available datasets, and the main deep learning techniques used for segmentation and anomaly detection. The most relevant gaps in the field are identified, justifying the initiation of this research.
- **Chapter 3 – The PV30 Dataset**: Describes in detail the data acquisition procedures, the technical specifications of the sensors used, the annotation methodology, and the defect classes considered. A statistical analysis of data distribution and the correlations between anomalies is also provided.
- **Chapter 4 – Methodology and Models**: Introduces the selected network architectures, multimodal fusion techniques, training and optimization strategies, and the evaluation criteria used for comparative analysis. The experimental framework is presented with particular emphasis on reproducibility and scalability.
- **Chapter 5 – Experimental Results**: Provides a detailed analysis of model performance on the PV30 dataset, including qualitative result visualizations and a discussion of the most common errors. Trade-offs between accuracy,

computational complexity, and generalization capability are highlighted.

- **Chapter 6 – Conclusions and Future Work**: Summarizes the main results achieved, discusses their theoretical and practical implications, and proposes future research directions such as integration with explainable AI techniques, adaptation to new climatic conditions, and the development of self-supervised models.

1.4 Potential Applications

The work presented in this dissertation holds significant relevance not only in the academic domain but also for the solar energy industry and the broader ecosystem of predictive maintenance for photovoltaic (PV) systems. With the global increase in installed PV capacity and the shift from small residential systems to large utility-scale solar farms, the ability to effectively monitor the condition of PV modules in a scalable and automated manner has become a strategic necessity.

Currently, many companies involved in PV system operations and maintenance (O&M) rely on scheduled periodic inspections and manual analysis of thermographic images. These processes are costly, time-consuming, and prone to human error. The introduction of automated solutions based on computer vision and deep learning, as explored in this dissertation, can drastically reduce operational costs, improve the timeliness of maintenance actions, and enhance diagnostic reliability.

In particular, the PV30 dataset, accompanied by an open benchmarking framework,

represents a valuable tool for companies aiming to develop, test, or validate proprietary algorithms for visual and thermal monitoring. The high-resolution annotated images enable robust model training, while the systematic benchmarking allows for rigorous comparison of new solutions against validated approaches. This facilitates the accelerated development of intelligent commercial systems, such as:

- **Software platforms** for the automated analysis of drone inspections, integrated with defect visualization dashboards and intervention recommendations;
- **Embedded systems** onboard drones, capable of performing real-time analysis during flight;
- **Cloud-based services** for continuous multi-site monitoring, tailored to large-scale energy operators;
- **Decision-support tools** for O&M teams that estimate fault priority and severity based on their impact on energy production.

Furthermore, the technologies developed can be transferred to related contexts, such as predictive maintenance of industrial plants, smart electrical grids, or other sectors that benefit from the integration of drones, thermography, and artificial intelligence.

Finally, due to the open and well-documented nature of the work, the results of this dissertation are fully compatible with technology transfer initiatives, academic spin-offs, public-private collaborations, and European innovation programs in the fields of energy and digitalization (e.g., Horizon Europe, the Green Deal, IPCEI – Important Projects of Common European Interest). The potential impact is therefore twofold: on one hand, to foster the adoption of AI technologies in industrial processes; on the other, to promote sustainable development and the resilience of future energy infrastructures.

Chapter 2

Background

Inspection and maintenance of photovoltaic (PV) systems have always played a crucial role in ensuring energy efficiency, operational safety, and continuity of production. Initially, these activities relied on traditional techniques such as manual visual inspection of modules, I-V curve electrical measurements, and the use of handheld ground-based thermal cameras. While effective on a small scale, these approaches are costly, time-consuming, and highly dependent on the operator's expertise—especially when applied to large-scale ground-mounted or rooftop PV installations.

In recent years, technological advancements have facilitated the introduction of remotely piloted aerial platforms, namely drones (UAVs), equipped with high-resolution optical and thermal sensors. This combination has enabled the acquisition of georeferenced, multispectral data over vast areas in extremely short timeframes, paving the way for new automated monitoring solutions. The ability to fly over remote or hard-to-reach areas, combined with the possibility of performing frequent inspections without interrupting energy production, makes drone-based inspections highly advantageous for diagnostics and predictive maintenance.

Concurrently, artificial intelligence—and deep learning in particular—has emerged as a key tool for the automated analysis of acquired imagery. Convolutional Neural Networks (CNNs), due to their ability to capture complex visual structures and spatial patterns, have been adopted for multiple applications in the PV domain: from precise segmentation of solar modules to the classification of visual defects, and automatic detection of thermal anomalies. However, traditional CNNs exhibit inherent limitations, such as difficulty in modeling long-range dependencies across different image regions and the need for large volumes of annotated data for effective training.

To overcome these barriers, the scientific community has recently turned its attention to transformer-based models, originally developed in the field of natural language processing and later adapted to computer vision. Architectures such as Vision Transformer (ViT), SegFormer, DINO, and Grounding DINO have introduced a new paradigm: instead of processing images using local convolutional kernels, they leverage self-attention mechanisms to model long-range spatial relationships. This results in enhanced abstraction and generalization capabilities, making these models particularly suitable for analyzing complex thermal data and for multimodal fusion of RGB and IR information.

In the context of semantic segmentation, several CNN-based architectures—including U-Net, DeepLabV3+, and PSPNet—have proven to be robust and accurate in localizing solar modules, even under poor lighting conditions or partial occlusions. Models such as U-Net++ and SegFormer offer further improvements through the use of dense connections and lightweight structures, making them capable of operating on low-power devices. Regarding anomaly detection, architectures such as YOLO (in its various versions), ResNet with Feature Pyramid Networks (FPN), and more recently DINO and

GLIP, have demonstrated promising performance, particularly when trained on thermal images with precise annotations.

The most directly related literature can be divided into two main areas:

1. **Semantic segmentation of PV panels on RGB and thermal images**;
2. **Anomaly detection in PV modules using thermal images**.

2.1 Semantic segmentation of PV panels on RGB and thermal images

Semantic segmentation represents a fundamental step for the precise localization of photovoltaic modules within complex scenes, facilitating subsequent operations such as defect classification and operational status analysis. Several authors have proposed solutions based on convolutional neural networks, including approaches that integrate Mask R-CNN for module segmentation and Faster R-CNN for anomaly detection, using multimodal data (RGB and thermal) acquired by UAVs.

These methodologies have demonstrated high accuracy, with true detection rates exceeding 90% and a low incidence of false positives. In particular, the use of thermal images significantly improves detection under low visibility conditions or in the presence of shading effects. Additional studies have proposed integrating 3D reconstruction with thermal imaging techniques to enhance defect localization. The use of advanced filtering for denoising has improved the readability of anomalies, effectively supporting maintenance-related decision-making processes.

2.2 Anomaly detection in PV modules using thermal images

Thermal anomaly detection is one of the most promising applications of computer vision technologies in the photovoltaic context. Deep learning models have demonstrated excellent performance in the automatic classification of defects through thermographic analysis. Among the proposed architectures, Res-CNN3 stands out as an optimized residual network for large-scale inspection, achieving a mean average precision (mAP) of 76.4%, outperforming traditional CNN-based approaches.

Subsequent developments have included the adoption of pruning techniques for lightweight models, such as ResNet-9, achieving 80.2% accuracy across 12 defect classes. Particularly noteworthy is the use of ensemble models trained on imbalanced datasets, which improved performance through data balancing techniques such as SMOTE and Focal Loss. These approaches yielded remarkable results: 94.4% accuracy in binary classification and 85.9% in multiclass classification. One of the most concrete applications was demonstrated by an autonomous UAV system used to inspect over 230,000 PV modules, efficiently detecting hotspots and other thermal anomalies. This highlights the scalability of drone-based inspection approaches for large-scale solar farms.

Additionally, a recent benchmark dataset containing 20,000 thermal images was proposed, facilitating the development of robust and generalizable models. Several studies evaluated the performance of YOLOv5, YOLOv7, and YOLOv8 for real-time inspection, with YOLOv5l achieving the best accuracy performance. Faster R-CNN has also been successfully employed in combination with segmentation for defect detection.

The literature review highlights how the integration of advanced deep learning techniques with RGB and thermal images acquired by drones has reached a high level of technical maturity. However, critical issues remain, including the lack of standardized datasets, limited generalization of models, and the interpretability of predictions. The present work aims to address these challenges through the development and analysis of the PV30 dataset, offering a novel contribution to the field.

Nevertheless, the effectiveness of these solutions is heavily influenced by the availability of precisely annotated datasets that are representative of real-world plant conditions. Most existing studies rely on RGB datasets collected in limited scenarios, which are often not publicly available or lack pixel-level annotations. This shortcoming limits the reproducibility of experiments and hinders objective comparison among different approaches. Furthermore, the poor generalization of models to unseen installations, the difficulty of effectively integrating RGB and IR data, and the lack of transparency in neural network decision processes (black-box behavior) remain open challenges in the field.

To address these issues, the construction of large-scale, carefully annotated multimodal datasets, accompanied by standardized benchmarking frameworks, becomes essential. The PV30 dataset, developed in this thesis, is presented as a concrete response to these needs, providing an innovative reference for the development and evaluation of AI models in the photovoltaic domain.

Chapter 3

Instrumentation and Data Acquisition Methodologies

3.1 Introduction

The growing integration of drones in renewable energy inspection has paved the way for automated, efficient, and safe monitoring systems for photovoltaic (PV) installations. The dataset presented in this study includes synchronized RGB and thermal images, acquired with the dual purpose of semantic segmentation and anomaly detection. This section outlines the equipment used and the systematic acquisition methodologies implemented to collect the data, thus providing a solid foundation for subsequent analytical processes.

3.2 Equipment Description

The primary UAV used for dataset collection is the DJI Mavic 3T, a platform renowned for its advanced flight performance and reliability in field operations. The drone boasts a maximum flight time of up to 46 minutes, which is crucial for covering large photovoltaic installations without the need for frequent battery changes. Additionally, it supports a transmission range of up to 15 km via its OcuSync 3+ system, allowing for remote operations over expansive areas. The DJI Mavic 3T is engineered for both stability and agility, featuring advanced obstacle detection systems and robust flight dynamics that ensure safe operations even under moderately challenging environmental conditions. Its compact, foldable design greatly facilitates transportation, and the IP43 rating provides resistance against dust and light splashes, making it well-suited for diverse outdoor settings.

3.2.1 Sensor Suit

The sensor suite integrated into the DJI Mavic 3T comprises two primary imaging systems that work in tandem to capture both visual and thermal data.

1) RGB Camera: The RGB camera is equipped with a 4/3 CMOS sensor that captures high-definition images at a resolution of 20 megapixels. This high level of detail is vital for visual analysis and the precise classification of various components within photovoltaic installations. The camera offers a variable aperture ranging from f/2.8 to f/11 and supports a broad ISO range (100–6400 for video and 100–12800 for still

images), which allows it to adapt to diverse lighting conditions. Moreover, the capability to record videos at resolutions up to 5.1K at 50 fps and 4K at 120 fps using efficient compression formats (H.264 and H.265) ensures that every nuance of the visual data is captured with clarity.

2) Thermal (Infrared) Camera: The thermal imaging system is based on an uncooled VOx microbolometer, delivering images at a resolution of 640×512 pixels. With a thermal sensitivity of ≤50 mK, this camera is capable of detecting even subtle temperature variations, which is critical for identifying hot-spots and other anomalies within photovoltaic panels. The sensor operates at a frame rate of 30 Hz, ensuring rapid thermal changes are captured effectively. The system also offers the flexibility of interchangeable lenses to tailor the field of view according to the specific dimensions of the installation. Temperature measurements are performed in two distinct modes: a high-gain mode covering a range from - 20°C to 150°C, and a low-gain mode spanning from 0°C to 500°C, thereby ensuring accurate readings under varying thermal conditions.

3.2.2 Data Acquisition Methodologies

A systematic and well-planned data acquisition strategy was essential for capturing a high-quality dataset.

1) *Flight Planning and Parameter Optimization*: Meticulous flight planning played a crucial role in ensuring the quality and comprehensiveness of the collected data. During the data collection process, the following parameters were carefully optimized:

ID	Latitudine	Longitudine	Altitude a.s.l. (m)	Flight altitude (m)	Speed (m/sec)	Overlapping	N° of RGB + IR photos	N° of IR potos	GSD (cm/pixel)
1	37.460743°	14.903484°	31	30	2,3	50%	500	250	3,96
2	37.468065°	14.824073°	49	30	2,3	50%	2.014	1.007	3,96
3	37.431377°	14.842102°	35	30	2,3	50%	840	420	3,96
4	37.456825°	14.886866°	30	30	3,3	50%	2.622	1.311	3,96
5	37.671885°	14.765921°	282	30	3,3	50%	466	233	3,96
6	37.671134°	14.763473°	278	30	3,3	50%	1.678	839	3,96
7	37.647941°	14.806170°	222	30	3,3	50%	474	237	3,96
8	37.650498°	14.801852°	226	30	3,3	50%	886	443	3,96
9	37.419450°	15.035566°	8	30	3,3	50%	522	261	3,96

- **Altitude and Speed**: The UAV was operated at altitudes between 30 and 50 meters, a range chosen to balance high image resolution with the need for extensive coverage of the photovoltaic panels. A controlled flight speed of 3 to 5 m/s was maintained to minimize motion blur, thereby ensuring that the images captured were crisp and clear.
- **Image Overlap and Coverage**: To guarantee complete coverage of the photovoltaic installations, flight paths were designed with a lateral overlap of at least 20% and a longitudinal overlap of 60%. Additionally, the camera angle was adjusted based on the orientation of the panels—using configurations such as 90° or 45°/90° for horizontal arrays and 45° for vertical elements—to capture images from optimal perspectives.

These optimized flight parameters were fundamental to acquiring data that is both high in quality and comprehensive in coverage, ensuring that the images are well-suited for subsequent processing tasks such as semantic segmentation and anomaly detection.

2) Environmental Considerations: The quality of the dataset was also dependent on external environmental factors. Data acquisition was scheduled under optimal conditions to enhance image quality. Flights were conducted during periods of low wind (with speeds below 5 m/s) to maintain UAV stability. Clear or partially cloudy skies were preferred to minimize reflective interference, thereby enhancing the thermal contrast between operational and defective cells. Moreover, most data collection activities were carried out during the central hours of the day when the thermal differential across photovoltaic panels is at its peak, further improving the quality of the thermal images.

3) Data Synchronization and Processing: A critical aspect of the data acquisition strategy was the simultaneous capture of RGB and thermal images. This synchronization ensured that the datasets were spatially and temporally aligned, which is essential for effective semantic segmentation and anomaly detection. By providing a comprehensive, multimodal perspective of the photovoltaic panels, the synchronized dataset supports robust analysis and enhances the performance of automated diagnostic algorithms.

3.2.3 Dataset Characteristics and Applications

The novel dataset was collected from eight distinct photovoltaic installations, as detailed in Table I, comprises detailed RGB and thermal imagery. The high-definition RGB images are invaluable for semantic segmentation, as they allow for precise classification and labeling of the panels and surrounding structures. In contrast, the thermal images are pivotal for detecting anomalies such as hot-spots and other temperature irregularities, which can indicate performance issues or potential failures. This multimodal dataset serves as a rich resource for developing automated diagnostic tools and advancing research in renewable energy maintenance and predictive diagnostics.

Chapter 4

DATASET AND EVALUATION METHODOLOGY

4.1 Introduction

Accurate and efficient inspection of photovoltaic (PV) panels is a critical requirement for ensuring the optimal performance, reliability, and longevity of solar energy systems. As PV installations expand globally—from residential rooftops to vast utility-scale solar farms—the demand for scalable, automated inspection methods has intensified. Central to these efforts is the availability of high-quality datasets that not only capture the visual appearance of solar modules but also reflect their thermal behavior, which is essential for identifying functional anomalies such as hotspots, bypass diode failures, and degradation patterns.

However, the current landscape of publicly available datasets is limited in several key aspects. Many datasets focus solely on RGB imagery, lack detailed and consistent annotations, or are collected under constrained environmental conditions, making them unsuitable for training and benchmarking advanced deep learning models in realistic operational scenarios. The absence of multimodal datasets and standardized evaluation protocols poses a significant challenge for both researchers and industry stakeholders aiming to develop robust, generalizable AI solutions for PV inspection.

To address these limitations, we introduce PV30, a comprehensive and novel multimodal dataset specifically designed to advance the development and evaluation of deep learning models for PV panel inspection. PV30 includes over 5,000 high-resolution RGB and infrared (IR) images, systematically collected using unmanned aerial vehicles (UAVs) across eight different photovoltaic installations under real operating conditions. These sites represent a wide range of environmental settings, module types, and defect categories, thus enhancing the diversity and representativeness of the dataset.

Each image in PV30 is meticulously annotated for both semantic segmentation and object detection tasks. The semantic annotations provide pixel-level masks of individual solar modules, enabling precise localization, while the detection annotations include bounding boxes and class labels for various types of thermal anomalies. The annotation process was carried out in collaboration with domain experts and verified using certified thermographic equipment, ensuring a high degree of accuracy and reliability.

In addition to the dataset itself, we propose a standardized evaluation framework that includes automated pipelines for training, validation, and performance visualization.

This framework supports consistent and reproducible benchmarking of state-of-the-art (SOTA) deep learning architectures, including convolutional neural networks (CNNs), transformer-based models, and multimodal fusion strategies. Performance is evaluated across multiple metrics such as Intersection over Union (IoU), precision, recall, F1-score, and mean average precision (mAP), providing a comprehensive assessment of each model's effectiveness.

By offering a richly annotated, multimodal dataset alongside an open and reproducible benchmarking framework, PV30 serves as a valuable resource for both academic research and industrial applications. It lays the groundwork for the development of next-generation AI tools for intelligent PV system monitoring and predictive maintenance, ultimately contributing to the increased efficiency, sustainability, and resilience of solar energy infrastructures.

4.2 Dataset Preparation and Preprocessing

Our framework is built upon a diverse and meticulously curated collection of aerial imagery featuring photovoltaic (PV) installations, captured using state-of-the-art UAV (Unmanned Aerial Vehicle) technology. These data acquisition efforts were carried out with a focus on achieving high spatial and temporal consistency across all collected samples. The UAV platform was equipped with a dual-sensor payload comprising a high-resolution RGB camera and a precision thermal imaging sensor. Flights were executed at a constant altitude of 30 meters above ground level, a parameter carefully selected to balance image resolution with coverage area, ensuring that the panels were captured in sufficient detail while minimizing distortion.

To maximize the diagnostic utility of the acquired images, flight missions were strategically scheduled around solar noon (approximately 12:00 PM local time). This timing minimizes the presence of shadows that can obscure defects and distort thermal signatures, while also ensuring optimal and uniform illumination across the PV arrays. By conducting all missions under comparable ambient conditions, we ensured the consistency of key imaging parameters—such as angle of capture, exposure, and radiometric calibration—thereby enhancing the overall coherence and utility of the dataset for training deep learning models.

The inclusion of both RGB and thermal imaging streams offers a multimodal perspective that significantly enriches the dataset. RGB images provide valuable information on visible surface-level anomalies, such as physical breakage, discoloration, delamination, and surface soiling. These types of defects are often early indicators of mechanical degradation and can be readily identified through visual inspection. In contrast, the thermal images reveal critical subsurface or electrical anomalies—including hotspots, bypass diode failures, and non-uniform heating patterns—that are typically invisible to the naked eye but have a direct impact on module performance and safety. This dual-sensor strategy allows for a synergistic fusion of visual and thermal data, supporting more comprehensive and accurate anomaly detection than approaches relying on a single modality.

The preprocessing pipeline adopted for this framework incorporates several best practices from the field of computer vision and deep learning. Each image was subjected to radiometric and geometric corrections to account for environmental and equipment-related variability. Standard preprocessing steps such as image normalization and resolution adjustment were applied to harmonize the input data across different sensor formats. Moreover, domain-specific data augmentation techniques—including random rotations, brightness shifts, contrast enhancements, and thermal intensity perturbations—were introduced to artificially expand the dataset and improve the generalization capabilities of the models under diverse real-world conditions.

For semantic segmentation tasks, each PV module within the dataset was annotated with precise pixel-level masks, capturing the exact boundaries of the solar panels. These detailed annotations are essential for training models to distinguish between panel surfaces and background structures, particularly in heterogeneous environments. Simultaneously, for anomaly detection tasks, bounding boxes were annotated around thermal anomalies and labeled with their corresponding defect classes (e.g., single hotspot, multi-hotspot, bypass diode anomaly). These annotations were generated with expert supervision and validated through cross-referencing with certified thermographic analyses, ensuring high fidelity and diagnostic accuracy.

To support robust model evaluation and facilitate fair performance comparisons, the dataset was stratified into three disjoint subsets: training, validation, and test sets. This partitioning strategy was designed to preserve class distributions and spatial diversity, preventing data leakage and ensuring that models are assessed on previously unseen installations and conditions. This division underpins the integrity of the benchmarking process, allowing for unbiased and reproducible evaluation of deep learning architectures under a wide array of operational scenarios.

4.3 Dataset Description

1) Overview of the Preliminary Dataset

The dataset employed in this study represents a comprehensive and well-structured collection of high-resolution imagery focused on ground-mounted polycrystalline photovoltaic (PV) panels. It was specifically designed to support a wide range of computer vision tasks, including both semantic segmentation and anomaly detection, by incorporating dual-modality imaging: RGB (visible spectrum) and thermal infrared (IR). This multimodal strategy enhances the diagnostic richness of the dataset, enabling the development and evaluation of advanced deep learning models that can exploit both visual and thermal information for more reliable fault detection.

For the task of semantic segmentation, a total of 1,200 high-definition RGB images were acquired. These images capture the spatial layout, geometry, and surface conditions of the PV panels under various lighting and environmental settings, providing essential visual cues for delineating module boundaries and detecting visible surface defects such as cracks, delamination, or soiling. Complementing these, 1,383 spatially aligned thermal images were captured using high-sensitivity infrared sensors. These IR images enable the detection of sub-surface and electrical anomalies—such as hotspots and bypass diode failures—which are often undetectable through visual inspection alone. In addition, a curated subset of 1,252 thermal images was designated specifically for object detection tasks, where the aim is to localize and classify thermal anomalies within the field of view using bounding box annotations.

The dual-modality composition of the dataset ensures that models can be trained not only to identify visible degradation but also to infer underlying electrical issues, fostering the creation of holistic diagnostic tools. By pairing RGB and thermal images for the same scenes, the dataset allows for multimodal learning and fusion strategies, which are known to enhance model performance and robustness in complex, real-world scenarios.

The dataset was methodically partitioned to support all stages of the deep learning pipeline—training, validation, and testing—using a widely adopted 70%-20%-10% split. Crucially, stratified sampling was employed to maintain a uniform distribution of anomalies and normal instances across all partitions. This ensures that each subset is representative of the full range of operational states and fault conditions present in the dataset, thus enabling models to generalize effectively and mitigating overfitting or evaluation bias. Such a design is particularly important when developing and benchmarking models intended for deployment in heterogeneous environments.

A key strength of the dataset lies in its balanced representation of both properly functioning PV modules and those exhibiting faults. This ensures that deep learning models can learn to distinguish between normal inter-panel variability and genuine anomalies, such as thermal hotspots, shading effects, delamination, and connector failures. Including healthy panels also helps to minimize false positives by providing the model with a clear reference of nominal thermal and visual characteristics.

Figure 1 presents a series of representative samples from the dataset, illustrating RGB and thermal imagery of the same PV installation. The RGB images showcase detailed structural features of the panels, including layout, alignment, frame integrity, and visible surface anomalies. In contrast, the thermal images reveal spatial temperature distributions that are crucial for identifying operational irregularities. Panels in good condition exhibit relatively uniform thermal profiles, whereas faulty panels are characterized by localized temperature elevations—commonly referred to as hotspots—or abnormal gradients, which often correspond to degraded cells or electrical faults. This visual contrast between healthy and defective modules underscores the importance of integrating both data modalities in the development of automated PV inspection systems.

In summary, the PV30 preliminary dataset is a critical foundation for the research presented in this thesis, offering a rich, balanced, and meticulously annotated resource for training and evaluating state-of-the-art semantic segmentation and anomaly detection models. Its careful construction ensures relevance to real-world applications, laying the groundwork for the deployment of intelligent, UAV-based PV monitoring solutions.

2) Data Sources and Collection Methodology

The data acquisition process was meticulously designed and executed using a UAV-based aerial imaging approach, where drones were operated at a standardized altitude of 30 meters above ground level across multiple ground-mounted photovoltaic (PV) installations. This flight altitude was carefully chosen as an optimal trade-off between spatial resolution and area coverage. It allows for the capture of high-fidelity visual and thermal data, preserving the granularity necessary for detecting cell-level anomalies—such as hotspots or cracks—while ensuring that large areas of solar fields can be scanned efficiently within a single flight session.

All drone missions were systematically scheduled around solar noon (approximately 12:00 PM local time), a time window that coincides with the peak irradiance and operational temperature of the PV modules. This deliberate scheduling serves multiple purposes that are critical to the reliability of thermal and optical data acquisition:

Enhanced Thermal Contrast: Midday conditions maximize the thermal gradient between defective and functional cells, making temperature anomalies such as hotspots more visible and distinguishable in thermal imagery.

Reduced Shadowing: By capturing data when the sun is near its zenith, the influence of elongated shadows caused by module frames or surrounding objects is significantly minimized. This is essential for preserving the integrity of visual and thermal data and avoiding false positives in anomaly detection.

Consistency of Illumination: Uniform lighting conditions during midday reduce the variability in pixel intensity and color balance across RGB images, which is vital for training deep learning models that are sensitive to changes in illumination.

To further ensure data quality and consistency, all imaging operations were conducted under carefully monitored environmental conditions. Specifically, data collection was restricted to clear, sunny summer days with minimal cloud cover and low atmospheric turbulence. These ideal conditions ensure that all panels were operating under consistent irradiance and load levels, thereby standardizing the thermal response of the modules across the entire dataset. Such environmental uniformity is crucial for minimizing confounding variables and enhancing the repeatability and robustness of experimental results.

The aerial imaging protocol also included a 30% lateral overlap between consecutive frames in the flight trajectory. This overlap not only guarantees complete spatial coverage of the PV installations but also facilitates advanced post-processing techniques such as orthomosaic generation, image stitching, and 3D reconstruction if needed. Moreover, this redundancy ensures that every panel is captured in multiple images, allowing for more reliable detection and segmentation even in cases of partial occlusion or noise.

Both RGB and thermal images were captured synchronously using dual-sensor payloads mounted on the UAV platform. This simultaneous acquisition ensures perfect spatial and temporal alignment between the two modalities, allowing for a seamless fusion of optical and thermal features during downstream analysis. Such alignment is indispensable for developing multimodal models that can cross-reference texture, color, and temperature cues to improve diagnostic accuracy.

The RGB camera operated at a native resolution of 4000×3000 pixels, providing detailed visual representations of the PV modules, including physical damage, discoloration, frame alignment, and surface contaminants such as dust or bird droppings. The thermal imaging system, on the other hand, delivered infrared images at a resolution of 640×512 pixels—commonly used in commercial thermal diagnostics—sufficient for capturing subtle temperature gradients across the surface of each module. This level of detail allows for accurate detection of thermal anomalies such as microcracks, failed bypass diodes, and partial shading effects.

In summary, the UAV-based imaging strategy adopted in this study integrates high-resolution, multimodal data acquisition with rigorously controlled environmental and operational parameters. This approach not only ensures the scientific validity and reproducibility of the collected dataset but also lays a robust foundation for the development and evaluation of intelligent inspection systems in real-world photovoltaic environments.

3) Dataset Statistics and Characteristics

The PV30 dataset has been carefully curated to include a broad spectrum of thermal anomalies that are frequently encountered in operational photovoltaic (PV) systems. These include, but are not limited to, localized overheating, defective electrical components, and uneven thermal behavior due to system degradation. The dataset's anomaly composition reflects both the diversity and the statistical prevalence of real-world fault conditions, providing a solid foundation for developing robust and generalizable detection models.

The following breakdown summarizes the approximate distribution of anomaly types identified and annotated in the dataset:

- Single hotspots**:** Represent approximately 60% of the anomalous panels. These typically correspond to isolated cell-level thermal elevations caused by microcracks, soldering defects, or surface contamination. As the most prevalent form of thermal irregularity in field-deployed PV arrays, single hotspots are a critical target for early-stage fault diagnosis.
- Multi-hotspots: Account for around 38% of the annotated anomalies. These are characterized by multiple thermally elevated regions within a single panel or across adjacent modules and often suggest more systemic issues such as string-level imbalances, shading, or widespread soiling.
- Bypass diode failures: Comprise roughly 12% of the defective cases. These failures manifest as characteristic thermal patterns, often involving entire cell groups running abnormally cool or hot due to malfunctioning bypass circuits, and require sophisticated analysis to distinguish from benign thermal gradients.

The class distribution was intentionally left imbalanced to faithfully replicate the statistical occurrence of faults observed in real-world solar installations. While a uniform distribution might simplify model training, it would not accurately reflect field conditions. Instead, we address this class imbalance through tailored training strategies, such as loss reweighting and data augmentation, to ensure that models remain sensitive to minority classes while not being overwhelmed by the dominant class.

In terms of environmental representation, the dataset is heavily weighted towards panels operating under direct solar irradiance, as this represents the most diagnostically informative condition for thermal anomaly detection. Uniform sunlight ensures maximum thermal contrast between defective and functional cells, which is critical for identifying faults such as hotspots or bypass failures.

While most of the dataset was collected during cloud-free midday conditions to reduce illumination variability and avoid false detections caused by shadows, a deliberate subset of images (approximately 5–10%) includes partial shadowing. These examples are valuable for enhancing the model's robustness to edge cases, including those introduced by transient environmental conditions such as cloud cover, vegetation proximity, or structural shading.

To support reliable model development, the entire imaging process was governed by a standardized protocol that controlled for solar position, atmospheric clarity, and UAV imaging parameters. This approach minimizes inter-sample variability that could otherwise introduce noise into the learning process or lead to overfitting to specific lighting conditions. As a result, the PV30 dataset offers a well-balanced compromise between controlled acquisition conditions and realistic operational diversity, providing both high-quality training data and challenging validation scenarios for benchmarking anomaly detection algorithms in the solar energy domain.

4) Data Annotation Process

The annotation phase of the PV30 dataset was designed to ensure the creation of high-fidelity ground truth data essential for training and evaluating deep learning models in both semantic segmentation and anomaly detection tasks. To accomplish this, a rigorous and multi-stage annotation protocol was implemented, combining scalability with expert-level precision to maximize the dataset's utility across diverse research and industrial applications.

For this purpose, Label Studio (v1.x) was selected as the principal annotation platform due to its extensive support for both pixel-level segmentation masks and bounding box annotations, as well as its user-friendly interface and compatibility with industry-standard formats. Figure 2 illustrates the annotation interface used during this process. The tool's flexibility enabled seamless transitions between segmentation and detection tasks, allowing for a streamlined and uniform labeling pipeline.

The annotation workflow was structured as a two-tiered process designed to balance throughput and quality assurance:

- Tier 1 – Initial Annotation Phase: Preliminary labels were generated by a team of trained crowd workers, who underwent basic instruction in photovoltaic panel recognition, module boundary identification, and thermal anomaly interpretation. This phase allowed for rapid coverage of the large image dataset while maintaining a baseline standard of accuracy.
- Tier 2 – Expert Review and Quality Control: All annotations from the initial phase were meticulously reviewed and corrected, where necessary, by domain experts with extensive experience in solar energy system diagnostics, thermal imaging interpretation, and fault classification. These specialists ensured that annotations conformed to strict quality standards and were consistent across the dataset, particularly in edge cases such as partially occluded panels or ambiguous thermal signatures.

For the semantic segmentation task, annotators were instructed to create precise pixel-wise masks for each individual photovoltaic panel within the images. The following key guidelines were enforced during the annotation process to ensure consistency and anatomical accuracy:

- Each PV panel was to be annotated as a separate instance, regardless of adjacency or mounting configuration.
- Segmentation boundaries were to follow the exact physical outline of the panel, including visible structural components such as mounting brackets or rails.
- Panels that were partially visible at the edges of the image were to be included, thereby maintaining annotation completeness and reducing edge bias during model training.

For the thermal anomaly detection task, annotators drew bounding boxes around all detected thermal anomalies and assigned a classification label from a predefined taxonomy, which included:

- ➢ Single hotspot
- ➢ Multi-hotspot
- ➢ Bypass diode failure

Each bounding box was tightly fitted to the thermal anomaly, ensuring minimal background inclusion and accurate spatial representation of the defect.

All annotations—both segmentation masks and bounding boxes—were exported using the COCO (Common Objects in Context) JSON format, a widely adopted standard in the computer vision community. The COCO format was selected for its robust support of complex annotation structures, including multiple object instances, class hierarchies, and segmentation polygons. This decision ensures compatibility with major deep learning frameworks such as TensorFlow, PyTorch, and Detectron2, and facilitates seamless use in benchmarking pipelines that rely on standardized evaluation metrics such as mIoU (mean Intersection over Union), mAP (mean Average Precision), F1-score, and Recall.

By employing this meticulous and dual-layered annotation approach, the PV30 dataset achieves a high degree of annotation accuracy and reliability. This makes it a robust foundation for training deep learning models capable of generalizing across diverse photovoltaic systems and operational environments.

4.4 Architecture Selection Criteria

The identification and deployment of suitable deep learning architectures for semantic segmentation and anomaly detection in photovoltaic (PV) panel imagery required a rigorous and principled approach. The guiding philosophy of the framework adopted in this thesis was to evaluate models based on fundamental architectural properties rather than superficial performance indicators tied to specific datasets or tasks. This strategy was intended to produce generalizable insights that could inform future research and industrial applications beyond the scope of this study.

a) Semantic Segmentation Architecture Criteria: for the semantic segmentation task, the goal was to accurately delineate the spatial boundaries of photovoltaic modules within high-resolution RGB and thermal imagery. In selecting candidate models, we considered a variety of factors that impact both performance and applicability in real-world inspection scenarios. The key criteria were as follows:

- **Architectural Diversity:** We deliberately selected architectures representing different design paradigms–from fully convolutional networks (DeepLabV3+) to transformer-based approaches (SegFormer)–to evaluate how different architectural principles perform on the same task.

- **Feature Extraction Mechanisms:** We considered the different approaches to feature extraction, comparing the effectiveness of atrous convolutions (DeepLabV3+), nested skip connections (U-Net++), pyramid pooling (PSPNet), and hierarchical transformer encoding (SegFormer).
- **Backbone Consistency:** To enable fair comparison of architectural contributions, we standardized on ResNet- 50 as the backbone for DeepLabV3+, U-Net++, PSPNet, and PAN. This isolates the performance differences attributable to the architectural innovations rather than backbone capacity.
- **Transformer Integration:** For SegFormer, we selected the MIT-B1 variant to evaluate how a pure transformerbased approach with its own feature extraction mechanism compares to CNN-based architectures with similar parameter counts.
- **Decoder Design:** We compared different decoder strategies, from the simple design of DeepLabV3+ to the complex nested structure of U-Net++ and the all-MLP approach of SegFormer, to assess their impact on segmentation quality.

In summary, this selection strategy was designed to maximize architectural diversity, fairness of comparison, and practical relevance. By evaluating both convolutional and transformer-based models within a standardized framework, this work seeks to illuminate how different architectural innovations contribute to segmentation accuracy, robustness, and computational efficiency in the context of photovoltaic panel inspection.

b) Anomaly Detection Architecture Criteria: for anomaly detection architectures, our selection criteria emphasized:

- **Paradigm Evolution:** We selected architectures representing the evolution of object detection approaches–from anchor-based (YOLOv11) to transformer-based end-to-end detection (DINO) and language-guided approaches (Grounding DINO, GLIP).
- **Backbone Considerations:** For YOLOv11, we utilized the standard CSPDarknet backbone optimized for detection tasks. For transformer-based models (DINO), we employed ResNet-50 to maintain comparability with the segmentation architectures. GLIP and Grounding DINO employ Swin Transformer backbones to leverage their stronger feature extraction capabilities.
- **Query Formulation:** We compared different approaches to dynamically initialized queries (DINO) and languageguided query selection (Grounding DINO, GLIP)–to assess their impact on detection precision.
- **Cross-Modal Integration:** For GLIP and Grounding DINO, we evaluated how the integration of language and vision modalities affects detection capabilities, particularly in terms of semantic understanding and generalization to novel object categories.
- **Computational Scalability:** We considered architectures with different computational profiles, from the lightweight YOLOv11 to the more computationally intensive transformer-based models, to understand the performance-efficiency tradeoffs.

The ultimate selection of deep learning architectures within this research framework was carefully curated to reflect a **strategic balance** between established, widely validated approaches and more recent, **state-of-the-art innovations** that introduce novel paradigms in feature representation and information fusion. This deliberate juxtaposition was designed not only to assess raw performance but to derive **deeper insights into the architectural principles** that govern effectiveness in the context of photovoltaic panel inspection and thermal anomaly detection.

On one end of the spectrum, models such as **DeepLabV3+** and **YOLOv11** were selected due to their **proven reliability, stability, and maturity** in a wide range of computer vision applications. These architectures serve as **solid performance baselines**, offering well-understood behaviors, extensive community support, and robust training dynamics. Their inclusion ensures that the benchmarking process remains grounded in **empirically validated methodologies**, providing reference points against which more experimental models can be meaningfully compared.

Conversely, the inclusion of **cutting-edge architectures** such as **SegFormer** and **GLIP (Grounded Language-Image Pretraining)** introduces the opportunity to evaluate **novel mechanisms** for **spatial encoding, attention-based feature extraction**, and **multimodal fusion**. These models embody the frontier of vision transformer design, incorporating mechanisms such as hierarchical attention, token-based representation, and zero-shot object grounding. Their integration into the benchmarking framework reflects a commitment to **forward-looking research**, emphasizing adaptability to emerging trends in computer vision and machine learning.

Furthermore, **backbone consistency** was maintained across several models—primarily through the adoption of **ResNet-50** or similarly parameterized base networks—thereby ensuring that differences in model performance can be more confidently attributed to **architectural innovations in decoder design, feature fusion strategies, or attention mechanisms**, rather than discrepancies in the raw representational capacity of the encoder.

By structuring the evaluation framework in this way, the research moves beyond superficial metric comparisons. It instead aims to deliver **meaningful, architecture-level insights** that illuminate how different neural design principles—whether convolutional, attention-based, or hybrid—impact task-specific performance, generalization, and computational efficiency in the context of high-resolution, drone-acquired photovoltaic imagery.

This balanced approach ultimately enhances the **scientific validity and interpretability** of the benchmarking results. It provides a robust platform not only for immediate comparison but also for informing **future architectural choices**, guiding the evolution of deep learning systems aimed at renewable energy diagnostics and intelligent infrastructure monitoring.

Chapter 5

EXPERIMENTAL RESULTS

5.1 Comprehensive Evaluation of State-of-the-Art Deep Learning Models for Photovoltaic Panel Analysis

The evaluation of state-of-the-art (SOTA) deep learning models for the analysis of photovoltaic (PV) panel imagery demands a rigorous and methodologically sound benchmarking process. Such an approach must go beyond superficial accuracy metrics and consider the full spectrum of technical, operational, and practical performance indicators. This section presents a detailed assessment of the experimental results obtained using the PV30 dataset, with a particular focus on two central computer vision tasks: **semantic segmentation of photovoltaic modules** and **anomaly detection via thermal infrared imagery**.

For the semantic segmentation task, the effectiveness of each model is primarily evaluated using the **Intersection-over-Union (IoU)** metric, which quantifies the overlap between predicted segmentation masks and ground truth annotations. IoU provides a robust measure of localization accuracy, especially critical in the context of PV panel inspection where precise boundary delineation influences downstream tasks such as fault association and spatial analytics.

For the anomaly detection task, a suite of complementary metrics is employed, including **Precision**, **Recall**, **F1-score**, and **mean Average Precision (mAP)**. Precision reflects the model's ability to avoid false positives, which is crucial in high-stakes environments where unnecessary interventions carry cost implications. Recall measures the completeness of defect detection, ensuring that no critical faults go unnoticed. F1-score balances the trade-off between these two metrics, while mAP captures the model's overall detection quality across different confidence thresholds and object categories.

In addition to these accuracy-based measures, we conduct an in-depth analysis of **computational efficiency**, including **inference time, model size, and training convergence rates**. These factors are particularly important for real-world deployment scenarios, such as real-time analysis on embedded UAV platforms or large-scale batch processing in cloud-based monitoring systems.

The evaluation also addresses **model generalization capabilities** across heterogeneous environmental conditions. Given the variability inherent in solar installations—ranging from lighting fluctuations and dirt accumulation to diverse panel orientations and climatic influences—it is essential to understand how models perform outside the specific context of their training data. To this end, experiments are conducted across various subsets of the dataset that represent different operational scenarios.

Furthermore, we systematically investigate the **impact of multimodal data fusion**, comparing the performance of models trained on RGB imagery alone, thermal data alone, and combined RGB+IR inputs. This analysis aims to elucidate the contribution of each modality to the final prediction quality and to assess whether early, late, or hybrid fusion strategies yield superior results.

Overall, this evaluation framework provides a **holistic and multi-dimensional assessment** of deep learning architectures in the context of PV panel monitoring. The results obtained offer not only a benchmark of current capabilities but also actionable insights into model behavior, robustness, and applicability to real-world photovoltaic inspection workflows.

Chapter 6

IMPLEMENTED ARCHITECTURES

6.1 Semantic Segmentation Architectures

The proposed framework integrates five principal architectural paradigms, each selected for its unique structural attributes and demonstrated effectiveness in semantic segmentation tasks. These architectures have been meticulously chosen to represent a comprehensive spectrum of methodological innovations within the deep learning community, ensuring that the evaluation captures both classical and modern approaches to feature extraction, spatial reasoning, and contextual understanding. Each model has been adapted and fine-tuned specifically for the challenges inherent in photovoltaic (PV) panel segmentation from aerial drone imagery—an application domain characterized by high inter-class similarity, complex backgrounds, and variable lighting and environmental conditions.

The inclusion of fully convolutional architectures, encoder-decoder networks, pyramid pooling frameworks, and transformer-based models ensures a balanced representation of both conventional convolutional neural networks (CNNs) and emerging transformer-based methods, which are rapidly redefining the boundaries of semantic segmentation performance. These architectural choices reflect the current state-of-the-art (SOTA) in computer vision and are adapted to address the unique spatial layout and spectral variability encountered in UAV-acquired solar farm imagery.

Moreover, by benchmarking these diverse architectures under a unified evaluation pipeline, the framework provides critical insight into their relative strengths and limitations when applied to the task of photovoltaic panel identification. This includes not only segmentation accuracy and mask fidelity but also computational efficiency, robustness to noise, and scalability to large-area inspections. The outcomes of this comparative analysis are essential for both researchers aiming to push the frontier of solar panel inspection automation and industry stakeholders seeking reliable, deployable AI solutions.

Ultimately, this architectural diversity within the framework enables a comprehensive exploration of segmentation strategies, laying the groundwork for future advancements in intelligent infrastructure monitoring using multimodal aerial imagery.

1) DeepLabV3+: DeepLabv3+ represents a significant advancement in semantic segmentation architecture, combining the strengths of encoder-decoder structures with atrous convolution. The architecture extends DeepLabv3 by incorporating a simple yet effective decoder module to refine segmentation results, particularly along object boundaries.

The implementation includes specific optimizations:

- Encoder features are processed at output stride = 16
- Low-level features are extracted from the backbone network before striding
- The decoder employs a channel reduction ratio to maintain balanced feature

importance

- Final upsampling recovers full resolution with output stride = 4

This architecture demonstrates particular effectiveness in preserving fine object oundaries while capturing rich contextual information, addressing a common challenge in semantic segmentation tasks. The decoder’ s simple design proves highly effective, with just two 3×3 convolutions being sufficient for feature refinement after concatenation with low-level features.

2) U-Net++: U-Net++ introduces a novel architectural evolution of the traditional U-Net, specifically designed to reduce the semantic gap between encoder and decoder feature maps through nested and dense skip connections [15]. The architecture’s key innovation lies in its redesigned skip pathways that enable more effective feature fusion while addressing two fundamental limitations of the original U-Net: unnecessarily restrictive skip connections and disconnected decoder subnetworks.

A key advantage of U-Net++ is its flexibility during inference. Through deep supervision, the network can be deployed in two operation modes:

- Ensemble mode, where segmentation results from all branches are averaged
- Pruned mode, where only one segmentation branch is used, allowing for varying degrees of model complexity reduction and computational efficiency

The architecture exhibits particularly strong performance in domains that demand high spatial precision, such as medical image segmentation, where accurate delineation of anatomical structures is critical for both diagnostic and therapeutic applications. Its dense skip pathway design significantly enhances the flow of gradients during training, effectively mitigating the vanishing gradient problem often encountered in deep neural networks. This architectural feature also promotes feature reuse, enabling the network to learn more compact and generalizable representations without sacrificing segmentation detail.

Moreover, the model's nested encoder-decoder structure plays a pivotal role in addressing one of the core limitations of conventional encoder-decoder frameworks—namely, the semantic gap between high-resolution spatial features in the encoder and the abstract, low-resolution features in the decoder. By introducing a hierarchy of intermediate decoding paths, the network is able to gradually fuse low-level spatial information with high-level semantic features, leading to more coherent and accurate boundary predictions. This is especially beneficial in applications where object boundaries are subtle or ill-defined, as is the case with fine structures in medical imaging or complex PV module layouts in aerial inspection imagery.

The architectural innovations embedded in this design make it highly adaptable to a range of semantic segmentation tasks beyond the medical domain, including remote sensing, industrial inspection, and autonomous navigation, where precise object localization under varying imaging conditions is a key requirement. Consequently, its integration within the benchmarking framework provides a robust baseline for evaluating both general-purpose and domain-specific segmentation models.

3) PSPNet: The Pyramid Scene Parsing Network (PSPNet) introduces a sophisticated approach to capturing global context information through a pyramid pooling module.

The architecture preserves computational efficiency through a carefully designed dimensionality reduction mechanism that follows the pyramid pooling operations. Specifically, after global feature extraction via pyramid pooling, the architecture applies 1×1 convolutional layers to compress the channel dimensions at each level of the pyramid. Typically, each of these levels is reduced to 1/N of the original feature dimension, where *N* represents the total number of pyramid levels. This dimensionality compression is strategically implemented to balance the trade-off between capturing global context and preserving fine-grained local information.

By limiting the expansion of feature dimensionality, the model avoids excessive computational burden that could result from concatenating multiple high-dimensional feature maps, a common issue in context-aggregation modules. This also prevents global features from dominating the learning process and ensures that local structures critical for accurate segmentation—such as panel boundaries and defect edges—remain salient during decoding.

Importantly, this efficiency is achieved without compromising performance. Experimental benchmarks, for instance on the ADE20K dataset, demonstrate a notable improvement of 4.45% in Mean Intersection-over-Union (mIoU) compared to a baseline Fully Convolutional Network (FCN) enhanced with dilated convolutions. This gain is particularly impressive considering the modest increase in computational overhead, positioning the architecture as a highly effective and scalable solution for real-time or resource-constrained semantic segmentation tasks.

In the context of photovoltaic inspection, this balance between accuracy and efficiency is especially valuable, enabling the deployment of high-performing models on embedded systems or in edge-computing environments, such as onboard drone processors or portable diagnostic units used in field inspections.

4) SegFormer: SegFormer presents an efficient and powerful semantic segmentation framework based on Transformers, combining a hierarchical transformer encoder with a lightweight multilayer perceptron (MLP) decoder. Unlike earlier transformer-based segmentation approaches such as SETR that use a standard ViT encoder and computationally demanding decoder, SegFormer introduces a more efficient design that can generate multi-scale features without positional encoding, while significantly reducing computational complexity. This makes it more suitable for real-world segmentation applications where efficiency and accuracy must be balanced.

The SegFormer framework offers multiple model sizes (B0-B5) with the same architecture but different capacities, enabling flexible deployment based on specific application requirements. Notably, even the lightweight variants achieve competitive performance on benchmark datasets, highlighting the efficiency of the design.

5) PAN - Pyramid Attention Network: The Pyramid Attention Network (PAN) introduces an effective architecture for semantic segmentation that combines Feature Pyramid Attention (FPA) and Global Attention Upsample (GAU) modules.

Unlike previous approaches that concatenate features from different pyramid scales (as in PSPNet or ASPP), the FPA module multiplies context information with the original feature map pixel-wise. This attention mechanism allows the network to selectively enhance important features while suppressing irrelevant ones, without introducing excessive computation.

The GAU module is designed based on the insight that high-level features with abundant category information can effectively guide the selection of low-level spatial details. By using global context to weight low-level features, the module focuses on precise localization of category boundaries during upsampling, which is crucial for accurate segmentation.

Key advantages of this architecture include:

- Effective multi-scale feature extraction through pyramid attention that preserves pixel-level information

- Efficient computation by avoiding complex dilated convolutions and using multiplication-based attention rather than concatenation
- Precise boundary recovery through context-guided upsampling that selectively incorporates relevant details
- Simple yet effective global context integration at multiple levels, balancing local and global information
- Flexible design where the FPA module can be used independently as a central block between encoder and decoder, even without the GAU module, and still provide precise pixel-level prediction

6.2 Anomaly Detection Architectures

For the detection and localization of anomalies in photovoltaic panels, the framework implements several state-ofthe- art object detection architectures, each bringing unique capabilities to the task:

1) YOLOv11: YOLOv11 represents the latest evolution in the YOLO (You Only Look Once) family, introducing significant architectural improvements while maintaining the single-stage detection paradigm that has made YOLO models popular for real-time applications. This iteration builds upon previous versions with a focus on enhancing feature extraction capabilities, computational efficiency, and detection accuracy. Beyond object detection, YOLOv11 demonstrates versatility across multiple computer vision tasks, including instance segmentation, image classification, pose estimation, and oriented object detection (OBB). This multi-task capability, combined with its improved performance characteristics, positions YOLOv11 as a comprehensive solution for real-time computer vision applications across various domains, from surveillance systems to autonomous vehicles and industrial automation.

2) DINO: DINO (DETR with Improved deNoising anchor boxes) advances end-to-end object detection by introducing three key innovations that significantly improve both training efficiency and detection accuracy. The architecture builds upon previous DETR-like models while addressing their limitations in convergence speed and query initialization through a backbone network followed by a Transformer encoder-decoder structure.

3) Grounding DINO: Grounding DINO extends the DETR-based detector DINO to enable open-set object detection through tight modality fusion between vision and language.

A particularly noteworthy innovation introduced by this architecture is the implementation of sub-sentence level text representation in the attention mechanism. This enhancement addresses a persistent challenge in multimodal models that leverage natural language for object classification or detection—namely, the unintended semantic entanglement between unrelated category names during the attention computation phase.

Traditional sentence-level representations tend to aggregate contextual information across the entire input sequence, which can lead to the dilution or blurring of fine-grained semantic distinctions among individual object categories. Conversely, word-level attention mechanisms, while preserving granularity, often suffer from the overgeneration of spurious dependencies, especially when multiple category terms co-exist within a sentence. These redundant interactions not only introduce computational overhead but can also degrade classification performance by conflating semantically disjoint concepts.

To overcome these limitations, the sub-sentence level approach employs custom attention masks that are explicitly designed to suppress attention flow between semantically unrelated terms. This is achieved by selectively blocking inter-word attention pathways among conceptually distinct labels, thereby preserving local semantic integrity without introducing unwanted interference.

This technique enables the model to retain a fine-grained understanding of each category within a rich contextual framework while ensuring conceptual isolation during attention propagation. As a result, the model can generate more discriminative and disentangled feature embeddings, which significantly improves its ability to distinguish between closely related or visually similar classes—an essential capability for tasks like anomaly detection in photovoltaic systems, where subtle distinctions in thermal signatures can correspond to fundamentally different types of faults.

Moreover, this methodological refinement contributes to greater robustness and interpretability of the model, as it aligns more closely with human-like reasoning by isolating the semantic contributions of individual concepts within a structured linguistic input. This makes it particularly advantageous in safety-critical domains where transparent and reliable decision-making is paramount.

4) GLIP: GLIP (Grounded Language-Image Pre-training) introduces a unifying framework that reformulates object detection as a phrase grounding task. This conceptual shift enables a model to detect arbitrary objects described in text prompts rather than being limited to a predefined set of object categories.

A key advantage of GLIP's unified formulation is its ability to leverage both detection and grounding data during training. This expands the vocabulary of visual concepts dramatically compared to traditional detection data, which typically covers no more than 2,000 categories. For instance, the grounding datasets Flickr30K and VG Caption contain 44,518 and 110,689 unique phrases respectively, orders of magnitude larger than standard detection vocabularies.

GLIP further enables scaling up semantic-rich data through a self-training approach, where a teacher model trained on gold (human-annotated) detection and grounding data generates boxes for web-collected image-text pairs. These generated boxes, combined with noun phrases detected by an NLP parser, create a vast pool of training data that exposes the model to an unprecedented range of visual concepts.

The architecture demonstrates remarkable transfer learning capabilities on established benchmarks without task-specific fine-tuning, making it an effective foundation model for objectlevel visual understanding tasks.

Chapter 7

METHODOLOGICAL FRAMEWORK FOR TRAINING AND EVALUATING PV PANEL ANALYSIS MODELS

7.1 Training Environment

All experiments were conducted in a standardized computational environment to ensure fair comparison across architectures. The training was performed on a workstation equipped with NVIDIA RTX A100 GPUs with 80GB of VRAM. Implementation was done using PyTorch (version 1.13) with CUDA 12.4 acceleration.

The semantic segmentation models were trained using a consistent protocol with AdamW optimizer, an initial learning rate of 1e-3 with cosine annealing scheduler, and a batch size of 8 for 100 epochs. For object detection architectures, we followed model-specific training recommendations while standardizing batch sizes and optimization parameters where possible to maintain comparability.

To ensure reproducibility, all random seeds were fixed across experiments, and the same data partitioning was used for all models. Checkpointing was performed based on validation performance with early stopping patience of 10 epochs to prevent overfitting.

7.2 Evaluation Methodology

1) Performance Metrics for Semantic Segmentation: The evaluation of semantic segmentation models employed multiple complementary metrics to provide a comprehensive assessment of segmentation quality:

- **Intersection over Union (IoU):** This primary metric measures the overlap between predicted and ground truth segmentation masks, calculated as the intersection divided by the union of the prediction and ground truth areas. IoU effectively quantifies both over-segmentation and under-segmentation errors in a single measure.
- **F1-Score:** Calculated as the harmonic mean of precision and recall, the F1-score provides a balanced measure of segmentation accuracy, particularly important for datasets with class imbalance. For semantic segmentation, this metric helps evaluate the overall quality of panel boundary delineation.
- **Precision:** Measures the proportion of correctly predicted panel pixels among all pixels predicted as panels. This metric is particularly important for applications where false positives could trigger unnecessary maintenance inspections.
- **Recall:** Quantifies the proportion of actual panel pixels that were correctly identified, providing insight into the model's ability to capture the full extent

of photovoltaic installations. High recall is essential for ensuring complete panel coverage in monitoring applications.

These metrics were computed using the default thresholds from model outputs without additional post-processing, providing a direct assessment of raw model performance. The evaluation was performed on the validation set throughout the training process to monitor convergence and on the held-out test set for final performance assessment.

2) Performance Metrics for Anomaly Detection: The anomaly detection models were evaluated using metrics specifically tailored to object detection tasks:

- **Mean Average Precision (mAP):** The primary metric for evaluating detection performance, calculated at an IoU threshold of 0.4. This threshold was selected to balance the need for localization accuracy with the inherent difficulty of precisely delineating thermal anomalies, which often have diffuse boundaries.
- **Precision:** Measures the proportion of correctly detected anomalies among all detections, helping to quantify false positive rates. This metric is particularly relevant for practical deployment scenarios where false alarms could significantly impact maintenance efficiency.
- **Class-Specific Performance:** Separate evaluation was conducted for each anomaly type (single hotspot, multi-hotspot, and bypass diode failure) to assess detection performance across different defect categories. This class-specific analysis is crucial for understanding model biases and potential weaknesses in detecting specific anomaly types.

- **False Positive Analysis:** Qualitative assessment of false positive detections was performed to identify common patterns and potential areas for improvement. This analysis revealed that temperature gradients at panel edges and reflective surfaces were the most common sources of false positives.

The detection results were evaluated without applying additional confidence thresholds beyond those produced by the models themselves, providing a realistic assessment of performance under standard operating conditions.

3) Computational Efficiency Metrics: To assess the practical deployability of the models, several computational efficiency metrics were systematically measured:

- **Inference Time:** Measured as the average time required to process images during the forward pass, reported in milliseconds per batch. For semantic segmentation, a batch size of 16 was used, while anomaly detection employed a batch size of 8, reflecting typical deployment scenarios.
- **Model Size:** The storage footprint of each model in megabytes, relevant for deployment in resourceconstrained environments such as edge devices or dronemounted systems.
- **Throughput:** Calculated as the number of images processed per second, providing a practical measure of processing capacity for large-scale installations.

These efficiency metrics were measured on the same hardware configuration used for training to ensure consistency across evaluations.

4) Testing Environment Specifications: All performance evaluations were conducted in a standardized environment to ensure reproducibility and fair comparison across architectures:

- **Hardware:** Testing was performed on NVIDIA A100 GPUs with 80GB VRAM, Intel Xeon processors, and 125GB system RAM.
- **Software:** PyTorch Lightning was used as the primary training framework, with PyTorch 12.4 as the underlying deep learning library. CUDA 11.8 provided GPU acceleration.
- **Input Processing:** Both RGB and thermal images were resized to 512×512 pixels for consistent processing across modalities. Images were normalized using appropriate mean and standard deviation values for each modality.
- **Batch Sizes:** Semantic segmentation models were evaluated with a batch size of 16, while anomaly detection models used a batch size of 8, reflecting the higher memory requirements of object detection architectures.

This standardized evaluation environment ensures that performance differences can be attributed to architectural characteristics rather than implementation or hardware variations, providing a fair basis for comparing the relative merits of different approaches.

Chapter 8

PERFORMANCE ANALYSIS

RESULTS

8.1 Semantic Segmentation Models

1) Performance Comparison Across Models (DeepLabV3+, U-Net++, PSPNet, SegFormer, PAN): After systematic optimization and tuning, our evaluation of semantic segmentation architectures (DeepLabV3+, U-Net++, PSPNet, SegFormer, and PAN) on the photovoltaic panel dataset revealed excellent performance characteristics. Table II presents the optimized performance metrics for each architecture when applied to RGB imagery.

TABLE II
OPTIMIZED PERFORMANCE METRICS OF SEMANTIC SEGMENTATION MODELS ON RGB IMAGES.

Architecture	IoU	F1-Score	Precision	Recall
DeepLabV3+	0.9441	0.9711	0.9721	0.9541
PAN	0.9437	0.9710	0.9536	0.9741
PSPNet	0.9194	0.9579	0.9397	0.9528
SegFormer	0.9150	0.9549	0.9639	0.9203
U-Net++	0.9466	0.9724	0.9340	0.9657

The optimized RGB results revealed several key observations:

- U-Net++ demonstrated the strongest performance with an IoU of 0.9466 and F1-score of 0.9724, confirming that its nested skip connection approach provides highly effective feature extraction for photovoltaic panel segmentation.
- DeepLabV3+ delivered excellent performance comparable to U-Net++, with only marginally lower IoU (0.9441), highlighting the effectiveness of its atrous convolution mechanisms for multi-scale feature capture.
- PAN achieved strong performance with an IoU of 0.9437, positioning it very close to the top-performing architectures while offering faster convergence during training.
- PSPNet and SegFormer, while exhibiting slightly lower IoU values of 0.9194 and 0.9150 respectively, provided valuable complementary capabilities in global

context modelling and efficient feature extraction.

- DeepLabV3+ achieved the highest precision (0.9721), while PAN demonstrated the strongest recall (0.9741), indicating their specialized strengths for different aspects of panel segmentation.

Our methodology employed systematic parameter tuning and advanced learning rate scheduling to achieve these results. The optimization process involved careful hyperparameter selection, architecture-specific adjustments, and custom data augmentation strategies to enhance model robustness. The convergence curves demonstrate the effectiveness of our approach, with models achieving stable performance within the training budget while avoiding overfitting.

2) Thermal Segmentation Performance: An extension of our research involved evaluating model performance on thermal imagery, which complements the RGB analysis by enabling panel segmentation based on temperature differences rather than visual characteristics, providing a completely different data perspective for more robust and comprehensive panel identification.

TABLE III
FINAL PERFORMANCE METRICS FOR THERMAL SEGMENTATION ACROSS ALL EVALUATED ARCHITECTURES.

Model	Val IoU	Precision	Recall	F1 Score	Convergence
DeepLabV3+	0.967	0.984	0.977	0.981	Medium
U-Net++	0.969	0.981	0.979	0.982	Slow
PSPNet	0.966	0.976	0.978	0.983	Slow
SegFormer	0.972	0.985	0.976	0.984	Medium
PAN	0.972	0.978	0.985	0.986	Fast

DeepLabV3+ demonstrated exceptional thermal segmentation capabilities, achieving a final validation IoU of 0.967 with particularly strong boundary delineation between panels of different temperatures. The atrous spatial pyramid pooling mechanism proved especially effective at distinguishing thermal gradients across panel boundaries.

PAN (Pyramid Attention Network) exhibited the fastest convergence in thermal segmentation tasks, reaching 90% of its peak performance within 7 epochs. Its feature pyramid attention module showed particular strength in capturing temperature variations while maintaining computational efficiency with a validation IoU of 0.972.

PSPNet leveraged its multi-scale pooling to effectively capture global thermal patterns, making it especially valuable for identifying large-scale temperature anomalies across installation arrays. While slightly less precise in boundary delineation (validation IoU of 0.966), it excelled at identifying panel-level thermal inconsistencies.

SegFormer combined efficiency with high-quality thermal segmentation (validation IoU of 0.972), with its hierarchical transformer design proving particularly adept at preserving fine thermal details even in challenging environmental conditions like partial shading.

U-Net++ achieved the highest precision in thermal boundary detection (validation IoU of 0.969), with its nested skip connections effectively preserving both semantic and spatial temperature information. This architecture proved.

3) Analysis of Architectural Strengths and Specialized Capabilities: Our comprehensive evaluation revealed distinct strengths and specialized capabilities in each optimized architecture across both RGB and thermal modalities:

DeepLabV3+:

- *RGB Capabilities:* Achieved exceptional precision (0.9721), demonstrating superior false positive rejection. The architecture's optimized encoder-decoder structure with atrous convolutions excels at multi-scale feature extraction, particularly valuable for heterogeneous panel installations.
- *Thermal Capabilities:* Excels in maintaining clear panel boundaries even with subtle temperature gradients, making it particularly valuable for identifying edge-related thermal anomalies with high precision.

U-Net++:

- *RGB Capabilities:* Delivered the highest overall IoU (0.9466), demonstrating balanced segmentation performance. The optimized nested skip connections effectively preserve spatial information while maintaining semantic context.
- *RGB Specialization:* Most effective for complex panel arrangements where detailed boundary information is critical, such as irregular installation geometries.
- *Thermal Capabilities:* Provides the most detailed thermal feature preservation with superior handling of temperature transitions, making it ideal for precise thermal isolation of individual panels in densely configured installations.

PSPNet:

- *RGB Capabilities:* Offers excellent global context integration through its pyramid pooling module, with balanced precision-recall characteristics ideal for varied installation types.
- *RGB Specialization:* Most effective for large-scale installations where understanding the global context is essential for proper panel identification.
- *Thermal Capabilities:* Demonstrates superior performance in capturing global

thermal patterns across large installations, enabling effective whole-system temperature analysis and anomaly contextualization.

SegFormer:

- *RGB Capabilities:* Achieves strong precision (0.9639) with efficient computation, leveraging its transformerbased architecture to capture long-range dependencies critical for understanding panel relationships.
- *RGB Specialization:* Combines good performance with lower computational requirements, making it suitable for edge deployment scenarios with resource constraints.
- *Thermal Capabilities:* Offers exceptional performance in challenging environmental conditions like partial shading or reflections, with its hierarchical transformer design effectively preserving fine thermal details at multiple scales.

PAN:

- *RGB Capabilities:* Delivers the highest recall (0.9741) with strong overall performance, ensuring comprehensive panel coverage even in challenging conditions through its effective feature pyramid attention mechanisms.
- *RGB Specialization:* Particularly valuable for applications where missing panels is more costly than false positives, such as maintenance inspection workflows.
- *Thermal Capabilities:* Achieves both the fastest convergence and highest overall metrics in thermal segmentation, with its pyramid attention mechanism proving exceptionally effective for thermal gradient analysis across different panel sizes.

8.2 Anomaly Detection Models

1) Performance Comparison Across Models (YOLOv11, DINO, Grounding DINO, GLIP): Following the comprehensive analysis of semantic segmentation models, a thorough evaluation was conducted on state-of-the-art anomaly detection architectures when applied to thermal imagery of photovoltaic panels. While segmentation identifies the panel boundaries, anomaly detection is crucial for localizing specific defects that affect panel performance. This analysis focused on four high-performance models: YOLOv11, GLIP, DINO, and Grounding DINO, each representing different architectural approaches to object detection.

After systematic optimization across all hyperparameters, training procedures, and model-specific configurations, exceptional performance metrics were achieved that demonstrate the viability of these approaches for thermal anomaly detection in photovoltaic installations.

TABLE IV
PERFORMANCE COMPARISON OF ANOMALY DETECTION MODELS ON THERMAL IMAGERY

Architecture	mAP@0.5	Precision	Recall	F1-Score	Inf. Time
YOLOv11	0.127	0.503	0.541	0.321	23 ms
DINO	0.631	0.492	0.608	0.629	68 ms
Grounding DINO	0.605	0.638	0.592	0.614	74 ms
GLIP	0.1	0.2	0.3	0.1	82 ms

The performance evolution of the anomaly detection models over the training process is visualized in Figure 5, which illustrates the gradual improvement in mAP scores as the models learned to identify increasingly subtle thermal patterns associated with different types of panel anomalies.

2) Analysis of Strengths and Weaknesses: The initial evaluation revealed inherent strengths and weaknesses in each anomaly detection architecture before optimization:

YOLOv11:

- *Strengths:* Demonstrated the fastest inference time (23 ms) among all tested architectures, making it suitable for real-time applications. Exhibited recall (0.541), indicating effective detection of various anomaly instances across the dataset.
- *Weaknesses:* Showed lower precision (0.503) compared to DINO, suggesting a tendency toward false positives, particularly for subtle thermal variations. Initial anchor configurations were suboptimal for the specific size distribution of thermal anomalies.

GLIP:

- *Strengths:* Demonstrated remarkable zero-shot capabilities through language-guided detection, allowing identification of anomaly types not well-represented in training data. The language-vision integration framework providesa promising direction for future research, particularlyfor evolving defect patterns, despite current performance limitations.
- *Weaknesses:* Significantly lower performance metrics compared to other architectures, with mAP (0.1), precision (0.2), and F1-score (0.1) far below other

models, while still requiring the longest inference time (82 ms). The complex cross-modal architecture displayed substantial domain adaptation challenges when applied to thermal imagery of photovoltaic panels, indicating the need for extensive domain-specific fine-tuning and more sophisticated prompt engineering strategies.

DINO:

- *Strengths:* Achieved the highest overall performance metrics with superior mAP (0.631) and precision (0.652), demonstrating effective denoising capabilities and feature discrimination for thermal anomalies.
- *Weaknesses:* Significantly higher computational demands with inference time (68 ms) nearly triple that of YOLOv11. The query-based detection approach showed suboptimal convergence characteristics, requiring extended training periods.

Grounding DINO:

- *Strengths:* Demonstrated remarkable zero-shot capabilities through language-guided detection, allowing identification of anomaly types not well-represented in training data. Strong precision (0.638) indicated effective crossmodal feature integration.
- *Weaknesses:* The slowest inference time (74 ms) amongall architectures, limiting real-time application potential. Language prompting strategies required refinement for optimal performance in the thermal domain.

Chapter 9

Conclusion

This doctoral dissertation presented the design, implementation, and validation of a comprehensive framework for the automated inspection of photovoltaic (PV) installations, aimed at overcoming the limitations of traditional monitoring approaches based on manual inspections and fixed instrumentation. The proposed approach is built upon three core components: multimodal data acquisition using UAVs, the creation of a rich and annotated dataset (PV30), and the establishment of a rigorous benchmarking system for evaluating state-of-the-art deep learning models.

At the heart of the research is the development of the PV30 dataset, comprising over 5,000 high-resolution RGB and thermal images collected from multiple drone missions over real-world solar plants under diverse environmental conditions. The dataset includes pixel-level annotations for semantic segmentation and bounding box labels for anomaly detection, created and validated with the assistance of domain experts. PV30 represents a significant contribution to the research community and provides a solid foundation for future advancements in solar inspection automation.

The benchmarking process included a wide range of architectures such as U-Net++, DeepLabV3+, SegFormer, YOLOv11, DINO, and Grounding DINO. The results revealed that transformer-based networks offer significant advantages in capturing global spatial relationships and in fusing RGB and IR modalities. However, optimized convolutional networks, especially lightweight models, remain highly competitive when low-latency inference is a key requirement.

Another important innovation is the open and modular benchmarking framework, designed to enable objective and reproducible comparison of models across multiple performance metrics (IoU, F1-score, mAP, inference time). This framework is intended to serve as a shared evaluation standard for academic researchers, industry developers, and O&M operators seeking to deploy automated solutions at scale.

From an industrial perspective, the outcomes of this thesis provide concrete benefits for solar energy operations. The proposed technologies enable faster, more frequent, and cost-effective inspections, improving the reliability and energy yield of PV systems. They also lay the groundwork for intelligent maintenance systems capable of early anomaly detection and efficient resource allocation.

Looking ahead, several future research directions emerge: (1) integrating explainable AI models to improve interpretability and trust in the system's decisions; (2) extending the

PV30 dataset to include different geographic and climatic scenarios, possibly enriched with multispectral or LiDAR data; (3) exploring self-supervised learning strategies to reduce dependency on annotated data; and (4) implementing embedded onboard systems for real-time processing directly on the drone.

Finally, the open-access and well-documented nature of this work encourages technology transfer and public-private collaboration, making it highly compatible with European innovation programs such as Horizon Europe and the Green Deal. In doing so, the research not only advances the scientific state of the art in AI for renewable energy but also contributes to a more sustainable and resilient global energy infrastructure.

Bibliography

[1] J. Balzategui, L. Eciolaza, and D. Maestro-Watson, “Anomaly detection and automatic labeling for solar cell quality inspection based on generative adversarial network,” Sensors, vol. 21, no. 13, p. 4361, 2021.

[2] A. Korovin, A. Vasilev, F. Egorov, D. Saykin, E. Terukov, I. Shakhray, L. Zhukov, and S. Budennyy, “Anomaly detection in electroluminescence images of heterojunction solar cells,” Solar Energy, vol. 259, pp. 130–136, 2023.

[3] F. A. Almalki, A. A. Albraikan, B. O. Soufiene, and O. Ali, “Utilizing artificial intelligence and lotus effect in an emerging intelligent drone for persevering solar panel efficiency,” Wireless Communications and Mobile Computing, vol. 2022, no. 1, p. 7741535, 2022.

[4] M. A. Napoli Spatafora, M. O. Spata, L. Guarnera, A. Ortis, and S. Bat- tiato. “Benchmarking of Blind Video Deblurring Methods on Long Exposure and Resource Poor Settings”. In: *Image Analysis and Processing – ICIAP 2023*. Ed. by G. L. Foresti, A. Fusiello, and E. Hancock. Springer Nature Switzerland, 2023, pp. 376–386.

[5] M. A. Napoli Spatafora, A. Ortis, and S. Battiato. "GLR: Gradient-Based Learning Rate Scheduler". In: *Image Analysis and Processing – ICIAP 2023*. Ed. by G. L. Foresti, A. Fusiello, and E. Hancock. Springer Nature Switzerland, 2023, pp. 269–281.

[6] M. Grisanti, M. A. N. Spatafora, A. Ortis, and S. Battiato. "E-ELPV: Extended ELPV Dataset for Accurate Solar Cells Defect Classification". In: *Intelligent Systems and Applications*. Ed. by K. Arai. Springer Nature Switzerland, 2024, pp. 837–848.

[7] F. Bougourzi, C. Distante, F. Dornaika, A. Taleb-Ahmed, A. Hadid, S. Chaudhary, W. Yang, Y. Qiang, T. Anwar, M. E. Breaban, C.-C. Hsu, S.-C. Tai, S.-N. Chen, D. Tricarico, H. A. H. Chaudhry, A. Fiandrotti, M. Grangetto, M. A. N. Spatafora, A. Ortis, and S. Battiato. "COVID-19 Infection Percent- age Estimation from Computed Tomography Scans: Results and Insights from the International Per-COVID-19 Challenge". In: *Sensors* 24.5 (2024). DOI: 10.3390/s24051557. URL: https://www.mdpi.com/1424-8220/24/5/1557.

[8] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[9] T. M. Mitchell. "Does machine learning really work?" In: *AI magazine* 18.3 (1997), pp. 11–11.

[10] K. Das and R. N. Behera. "A survey on machine learning: concept, algorithms and applications". In: *International Journal of Innovative Research in Computer and Communication Engineering* 5.2 (2017), pp. 1301–1309.

[11] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

[12] L. Breiman. *Classification and regression trees*. Routledge, 2017.

[13] N. S. Altman. “An introduction to kernel and nearest-neighbor nonparametric regression”. In: *The American Statistician* 46.3 (1992), pp. 175–185.

[14] D. Maulud and A. M. Abdulazeez. “A review on linear regression comprehensive in machine learning”. In: *Journal of Applied Science and Technology Trends* 1.4 (2020), pp. 140–147.

[15] Y. Ren, L. Zhang, and P. N. Suganthan. “Ensemble classification and regression-recent developments, applications and future directions”. In: *IEEE Computational intelligence magazine* 11.1 (2016), pp. 41–53.

[16] T. Jiang, J. L. Gradus, and A. J. Rosellini. “Supervised machine learning: a brief primer”. In: *Behavior Therapy* 51.5 (2020), pp. 675–687.

[17] M. E. Celebi and K. Aydin. *Unsupervised learning algorithms*. Vol. 9. Springer, 2016.

[18] R. S. Sutton, A. G. Barto, et al. “Reinforcement learning”. In: *Journal of Cognitive Neuroscience* 11.1 (1999), pp. 126–134.

[19] C. Bishop. “Pattern recognition and machine learning”. In: *Springer google schola* 2 (2006), pp. 5–43.

[20] Z. C. Lipton. “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” In: *Queue* 16.3 (2018), pp. 31–57.

[21] Y. LeCun, Y. Bengio, and G. Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444.

[22] F. Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[23] M. Minsky and S. Papert. “An introduction to computational geometry”. In: *Cambridge tiass., HIT* 479.480 (1969), p. 104.

[24] P. Werbos. “Beyond regression: New tools for prediction and analysis in the behavioral sciences”. In: *PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA* (1974).

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536.

[26] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “WaveNet: A Generative Model for Raw Audio”. In: *Arxiv*. 2016.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2019).

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. “Playing atari with deep reinforcement learning”. In: *Proceedings of NIPS Deep Learning Workshop* (2013).

[29] S. Chetlur, C. Woolley, P. Vandermersch, J. M. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer. “cuDNN: Efficient Primitives for Deep Learning”. In: *ArXiv* abs/1410.0759 (2014).

[30] D. Ciresan, U. Meier, and J. Schmidhuber. “Multi-column deep neural networks for image classification”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 3642–3649.

[31] V. Nair and G. E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 2010, 807–814.

[32] T. Chen. “From low-level features to high-level semantics: are we bridging the gap?” In: *Seventh IEEE International Symposium on Multimedia (ISM’05)*. 2005. DOI: 10.1109/ISM.2005.62.

[33] A. Sherstinsky. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network”. In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306. DOI: https://doi.org/10.1016/j.physd.2019.132306.

[34] J. S. S. Hochreiter. “Long Short-Term Memory”. In: *Neural Comput* 9.8 (1997), 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735.