Evaluation of Neural Object Detection Models for Human Detection in Infrared Images

PROJECT REPORT T1000

from the course of studies Computer Science - Artificial Intelligence

at the Cooperative State University Baden-Württemberg
Ravensburg Campus Friedrichshafen

by

Lukas Florian Richter

18.08.2025

Completion Time:	16 Wochen
Student ID, Course:	None, TIK24
Company:	Airbus Defence & Space, Taufkirchen
Supervisor in the Company:	René Loeneke

Declaration of Authorship

In accordance with clause 1.1.13 of Annex 1 to §§ 3, 4 and 5 of the Cooperative State University Baden-Württemberg's Study and Examination Regulations for Bachelor's degree programs in the field of Technology, dated 29.09.2017. I hereby declare that I have written my thesis on the topic:

Evaluation of Neural Object Detection Models for Human Detection in Infrared Images

independently and have used no other sources or aids than those specified. I further declare that all submitted versions are identical.

Taufkirchen 18.08.2025

Lukas Florian Richter

Abstract

This project report evaluates the performance of neural object detection models for detecting humans in infrared images. The study focuses on comparing different variations of the SSD (Single Shot Mutlibox Detector) model architecture, assessing their accuracy and inference speed, and identifying the most suitable model for the given task. Additionally, different preprocessing techniques are evaluated to improve the detection performance.

More specifically, the main contributions of this project are:

- Conceptualization of a simple and cost-efficient hardware setup for the purpose of on-premise human detection in infrared images
- Evaluation of different SSD model architectures
- Comparison between different preprocessing techniques
- Identification of the most suitable model for the given task
- A theoretical pipeline for the secure transmission of the detection results to a remote server

Table of Contents

List of Figures

List of Tables

Code Snippets

List of Acronyms

AI	Artificial Intelligence
AP	Average Precision
API	Application Programming Interface
AdaBoost	Adaptive Boosting
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DL	Deep Learning
FC	Fully Connected
FN	False Negative
FP	False Positive
FP16	16-bit floating point
FP32	32-bit floating point
FPS	Frames per Second
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HNM	Hard-Negative Mining
HOG	Histogram of Oriented Gradients
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
INT8	8-bit integer

IR	Infrared
IoU	Intersection over Union
ML	Machine Learning
MPS	Metal Performance Shaders
NMS	Non-Maximum Suppression
NN	Neural Network
OSI	Open Systems Interconnection
R-CNN	Region-based Convolutional Neural Network
RAM	Random Access Memory
REST	Representational State Transfer
RGB	Red, Green, Blue
RNN	Recurrent Neural Network
ROI	Region of Interest
RPN	Region Proposal Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
SGD	Stochastic Gradient Descent
SSD	Single Shot MultiBox Detector
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TN	True Negative
TP	True Positive
TPU	Tensor Processing Unit
VGG	Visual Geometry Group
VGP	Vanishing Gradient Problem

VOC	Visual Object Classes
ViT	Vision Transformer
YOLO	You Only Look Once
mAP	mean Average Precision

1 Introduction

With the increase in security threats to critical infrastructure, automated surveillance systems have become essential for ensuring the safety and security of people, infrastructure and property at scale. The ability to detect individuals on critical infrastructure premises is crucial to preventing unauthorized access and potential damage to assets. While conventional RGB-based surveillance systems remain prevalent in many application, they face inherent limitations in challenging scenarios such as low-light conditions, adverse weather, fog, smoke and complete darkness during nighttime [1].

A compelling alternative to systems operating in the visible light domain are such capturing wavelengths in the infrared spectrum and thus offering consistent detection capabilities that are fundamentally independent of ambient lighting conditions. Unlike regular RGB cameras, thermal sensors detect light with longer wavelengths that correspond to the heat signatures emitted directly by objects. This characteristic provides unique advantages for human detection, as the human body maintains a relatively constant temperature of approximately 37°C, creating distinct thermal signatures that remain visible regardless of environmental illumination [2].

The integration of deep learning architectures with thermal imaging thus opens new possibilities for automated systems that can reliably detect humans in the aforementioned scenarios with adverse conditions for conventional RGB-based concepts. However, most state-of-the-art object detection models have been primarily developed for and trained on RGB imagery. Given that the spectral, tectural and contrast characteristics of infrared images differ substantially from visible-light imagery, both due to the properties of those wavelengths themselves and of the sensors, those existing models might need to be adapted to achieve optimal performance.

This research addresses the critical need for systematic evaluation of neural object detection models specifically tailored for thermal human detection applications. The study focuses on the Single Shot MultiBox Detector (SSD) architecture, a prominent one-stage detection framework known for its balance between accuracy and computational efficiency. By examining multiple model variants with different backbone networks (VGG16 and ResNet152), initialization strategies (pretrained versus scratch training), and thermal-specific preprocessing techniques (image inversion and edge enhancement), this work provides comprehensive insights into optimal configurations for infrared surveillance systems.

The practical significance of this research extends beyond academic interest, addressing real-world challenges faced by the security and defense industry. In partnership with Airbus Defence & Space, this project explores the development of cost-efficient, edge-deployable thermal surveillance solutions that can operate reliably in challenging environments where traditional RGB systems fail.

1.1 Research Objectives and Contributions

This thesis makes several key contributions to the field of thermal image processing and computer vision:

1. **Preprocessing Technique Analysis**: Quantitative evaluation of thermal-specific image enhancement methods, including polarity inversion and edge enhancement, and their impact on detection accuracy.

2. **Backbone Network Comparison**: Detailed comparison between VGG16 and ResNet152 architectures in the context of thermal imagery, addressing the trade-offs between model complexity and performance.

3. **Practical Implementation Guidelines**: Development of actionable recommendations for deploying thermal surveillance systems in real-world environments, considering computational constraints and accuracy requirements.

4. **Dataset Integration Framework**: Unified evaluation approach across five diverse thermal datasets (FLIR ADAS v2 [3], AAU-PD-T [4], OSU-T [5], M3FD [6], KAIST-CVPR15 [7]), enabling robust performance assessment.

1.2 Thesis Organization

The remainder of this thesis is structured to provide a comprehensive examination of thermal human detection using neural networks. Section 2 presents a thorough review of object detection fundamentals, SSD architecture principles, and thermal image processing techniques, establishing the theoretical foundation for the experimental work. Section 3 details the systematic approach employed for model evaluation, including dataset preparation, experimental design, and evaluation metrics. Section 4 presents comprehensive performance analysis across all model configurations and preprocessing techniques. Section 5 interprets the findings within the context of practical deployment scenarios and industrial requirements. Finally, Section 6 synthesizes the key contributions and outlines directions for future research in thermal surveillance technologies.

2 Literature Review and Theoretical Background

The field of object detection has undergone significant evolution from traditional computer vision techniques to sophisticated deep learning architectures. Understanding this progression is essential for contextualizing the current work's contribution to thermal image analysis. This section examines the theoretical foundations of object detection, with particular emphasis on the Single Shot MultiBox Detector (SSD) architecture and its applicability to thermal imagery processing challenges.

Key areas to develop:

- Evolution from traditional methods (HOG, SIFT) to deep learning
- Comparison of one-stage vs. two-stage detection models
- SSD architecture fundamentals and anchor box mechanisms
- Backbone network analysis (VGG vs. ResNet trade-offs)
- Thermal imaging characteristics and preprocessing challenges
- Existing work on infrared human detection
- Gap analysis: Limited research on SSD for thermal surveillance

2.1 Object Detection Fundamentals

Most object detection methods can be broadly categorized into two main approaches: traditional methods and deep learning-based methods. Traditional methods mainly rely on handcrafted features and sliding window techniques, while deep learning-based methods in this field leverage Convolutional Neural Networks (CNNs) or Vision Transformer (ViT) architectures to automatically learn features from data.

2.1.1 Traditional Object Detection Methods

Simple approaches to object detection entail applying manually constructed feature detector kernels in a sliding window fashion to images.

An example of this is the **Viola-Jones-Algorithm** [8]:

1. Compute the integral image of the input image, that is, the sum of pixel intensities from the top-left corner of the image to each pixel. This allows for quick computation of the sum of pixel intensities in any rectangle in the image by subtracting the value of the upper left pixel of the rectangle from that of the lower right pixel.

2. Apply a series of Haar-like features to detect potential objects. Haar-like features are computed by subtracting the sum of pixels in one rectangle from the sum of pixels in an adjacent rectangle. These features capture various simple patterns, such as edges and lines.

3. Use the Adaptive Boosting (AdaBoost) technique to build a cascaded strong classifier consisting of several weak classifiers that can detect simple patterns consisting of Haar-like features.

4. Split the image into subwindows and classify each subwindow using the cascaded classifier as either containing the object or not.

Other approaches employ Histogram of Oriented Gradients (HOG) descriptors. The HOGs are attained by dividing the image into a grid of cells, contrast-normalizing them and then computing the vertical as well as horizontal gradients of their pixels. The gradients for each cell are accumulated in a one-dimensional histogram which serves as that cell's feature vector. After labeling the cells in the training data, a Support Vector Machine (SVM) can be trained to find an optimal hyperplane separating the feature vectors corresponding to the object that should be detected from those that do not contain the object.

2.1.2 Deep Learning-Based Object Detection

However, those methods are either highly dependent on engineering the correct priors, such as the Haar-like features, or limited to binary classification scenarios, as is the case for HOG-based SVMs. Thus, newer Object Detection methods employ more complex deep-learning architectures that require less manual feature engi-

neering. The best-performing models nowadays are ViTs using Attention mechanisms [9] to learn relationships between patterns in different parts of images. However, they will not be further examined in this thesis, due to computational constraints that make them unfeasible for the edge-deployable solution sought in this work [9].

Relevant for this examination are their predecessors, CNNs. The main mechanism they use to extract information from images are convolutional layers. Those convolutional layers get passed an image in the form of a tensor and perform matrix multiplication on that input tensor and a kernel tensor in a sliding window fashion to compute subsequent feature maps. Those will be passed on as input to the next layer. [10]

At their core, these convolutional layers do not work inherently different from fully connected layers that compute several weighted sums across all components of the input tensor. More specifically, fully connected layers can be described as convolutional layers whose kernel dimensions are identical to those of the input tensor.

Resorting to smaller kernels, however, serves as a prior making use of the heuristic that in most cases, the features composing an object in an image lie closely together. Thus, it is not necessary to process the entire image to detect an object that occupies only part of it. Convolutional neural nets hence save computational resources by focusing on smaller regions. In many cases it is advantageous to use those savings to increase network depth in order to make it possible for the network to learn more complex high-level features in subsequent layers.

Object detection, as opposed to image classification, consists of two main tasks: locating where an object is and classifying which class it belongs to. In the context of machine learning, that means two concepts must be used: regression to approximate the location of an object and classification to determine its class. CNNs solving these tasks can be categorized into two main categories:

- **Two-Stage Detectors**: These detectors operate in two stages. The first stage proposes regions of interest and the second stage classifies which object they contain. In more detail, that means regressing bounding boxes and assessing the “objectness” of that region, for example by using logistic regression. If the confidence this region contains an object exceeds a given threshold, the second stage then classifies the object in that region. That requires a second pass of the extracted region through a classifier network. This two-stage approach can be computationally expensive, especially when dealing with a large number of proposals. Examples of two-stage detectors include Region-based Convolutional Neural Networks (R-CNNs) [11], Fast R-CNNs [12], and Faster R-CNNs [13].

- **Single-Stage Detectors**: These detectors perform both tasks simultaneously in a single pass through the network. That means passing the image through a network that both regresses bounding boxes and classifies objects in those boxes at the same time. Examples include You Only Look Once (YOLO) [14] and SSD [15]. This approach can be faster but may sacrifice some accuracy compared to two-stage detectors, as the feature extractor is not optimized for both tasks .

Given the computational constraints imposed by the requirement for edge-deployment, single-stage detectors were chosen. Past research has shown that SSD-variants with Inception-v2 and MobileNet-v1 backbones perform notably faster than their Faster R-CNN counterparts, namely 4 to 7 times as fast [2].

Furthermore, benchmarks of SSD and YOLO on the MS COCO dataset yielded similar results favoring SSD in terms of speed when deployed on edge devices, namely the Raspberry Pi 4 both with and without a dedicated Tensor Processing Unit (TPU) [16]. YOLO did deliver higher mean Average Precision (mAP) scores, but the difference was not significant enough to justify the trade-off in speed, in particular taking into account the benefit of speed for real-time applications when multiple images are captured each second and fast enough processing allows for multiple attempts at detection. Additionally, the SSD models tested consumed less energy than their YOLO counterparts [16], making them a more suitable choice that minimizes the need for human intervention to replace the battery of the edge

device, which is a significant factor in the cost of deployment and maintenance of the system.

2.2 SGD as Optimizer in Deep Learning

Deep Learning Models are optimized by minimizing the loss function, which is a measure of the difference between the predicted output and the expected output, i.e. the ground truth. The loss function is typically a differentiable function, which means that it can be used to compute the gradient of the loss with respect to the model parameters. The gradient is then used to update the model parameters in the direction that minimizes the loss function, hence the name gradient descent.

In most cases, the training dataset is too large to compute the gradient with respect to the entire training dataset. Instead, the optimization takes place in so-called mini-batches of training data of a fixed size, under the assumption that the gradient computed with respect to a mini-batch of training data is a good approximation of the gradient that would be obtained if the calculation was performed across the entire training dataset.

This differentiation of the loss function with respect to a mini-batch of training data is called Stochastic Gradient Descent, (SGD).

As desribed before, SGD is a first-order optimization algorithm, which means that it only considers the first-order derivatives of the loss function with respect to the model parameters. This is in contrast to second-order optimization algorithms, such as Newton's method, which consider the second-order derivatives of the loss function with respect to the model parameters. However, first-order optimization algorithms are generally preferred in deep learning because they are computationally more efficient and can be easily implemented on hardware accelerators such as Graphics Processing Units (GPUs) and TPUs, which is why second-order optimization algorithms will not be further discussed in this work.

One limitation of SGD lies in the so-called Vanishing Gradient Problem (VGP) problem, which occurs due to the calculation of the partial derivates by means

of the chain rule. The chain rule for differentiation states that the derivative of a composite function is the product of the derivatives of its components:

$$f(g(x)) = f'(g(x)) \cdot g'(x) \tag{1}$$

In this equation, f and g are the functions applying the linear transformations

In the context of deep learning, this means that the derivative of the loss function with respect to a particular weight in the network

2.3 Single Shot MultiBox Detector (SSD) Architecture

The SSD architecture is a single-stage detector that uses a base network to extract features from the input image and then applies additional convolutional layers as well as Fully Connected (FC) layers to predict bounding boxes and class scores for each feature map. The following sections provide a more detailed explanation of the SSD architecture and its components.

2.3.1 Backbone Networks for Feature Extraction

Explores the role of backbone networks (VGG, ResNet) in feature extraction and their impact on SSD performance.

The SSD architecture uses a base network to extract features from the input image. The base network is typically a pre-trained CNN, such as Visual Geometry Group (VGG) or Residual Network (ResNet), which has been trained on a large dataset like ImageNet.

2.3.1.1 VGG as backbone

The VGG network is a deep CNN that consists of 16 or 19 layers, depending on the variant. The VGG network is known for its simplicity and effectiveness in image classification tasks. In its vanilla configuration, it takes 224x224 RGB images as input and outputs a 1000-dimensional vector of class probabilities. It only uses 3x3 convolutional layers and 2x2 max-pooling layers for feature extraction and the Rectified Linear Unit (ReLU) activation function for non-linearity. Eventually, it employs three FC layers for classification. The soft-max activation function is used

in the final layer to predict the class probabilities. Overall, the number of trainable parameters for the VGG−16 network is 138 million. [17]

2.3.1.2 ResNet as backbone

The ResNet network is a deep CNN that uses residual blocks to address the vanishing gradient problem. The vanishing gradients problem occurs when the gradients become too small to update the weights of the earlier layers during backpropagation. This can lead to slow convergence or even divergence of the training process. The ResNet

2.3.2 Feature Maps and Anchor Boxes

Describes the multi-scale feature maps and anchor boxes used in SSD for object detection.

2.3.3 MultiBox Loss Function

Explains the MultiBox loss function that combines localization loss and confidence loss for training SSD models.

Non-Maximum Suppression (NMS):

Examines the NMS technique used to filter duplicate detections and improve detection accuracy.

2.4 Thermal Image Processing

Discusses characteristics of thermal images, preprocessing techniques (inversion, edge enhancement), and challenges specific to infrared imagery.

3 Methodology

This study employs a systematic experimental approach to evaluate the effectiveness of SSD-based neural networks for human detection in thermal imagery. The methodology encompasses dataset selection and preparation, implementation of multiple model variants with different backbone architectures, application of thermal-specific preprocessing techniques, and comprehensive evaluation metrics. The experimental design ensures reproducible results while addressing the unique challenges posed by infrared image characteristics.

Key areas to develop:

- Dataset description: FLIR ADAS v2, AAU-PD-T, OSU-T, M3FD, KAIST-CVPR15
- Model configurations: SSD300-VGG16 vs. SSD300-ResNet152
- Training setup: Pretrained vs. scratch initialization strategies
- Preprocessing techniques: Image inversion and edge enhancement
- Data augmentation and split strategies (train/validation/test)
- Evaluation metrics: mAP, precision, recall, inference speed
- Hardware setup and computational requirements
- Statistical significance testing approach

3.1 Dataset Description

Details the thermal image datasets (FLIR ADAS v2, AAU-PD-T, OSU-T, M3FD, KAIST-CVPR15) and their characteristics.

3.2 Model Implementation

Explains the implementation of SSD models with different backbones and preprocessing configurations.

TODO: Incorporate switch to multi-label setup later

3.3 Experimental Design

Outlines the systematic approach to comparing model variants and the evaluation framework.

4 Results and Analysis

The experimental evaluation reveals significant performance variations across different model configurations and preprocessing approaches when applied to thermal human detection tasks. This section presents comprehensive results from training 16 distinct model variants, combining backbone architectures (VGG16 vs. ResNet152), initialization strategies (pretrained vs. scratch), and preprocessing techniques (none, inversion, edge enhancement, combined). The analysis demonstrates clear patterns in model behavior and identifies optimal configurations for thermal surveillance applications.

Key areas to develop:

- Training convergence analysis: Loss curves and stability patterns
- Detection accuracy results: mAP scores across all model variants
- Preprocessing impact: Quantitative comparison of enhancement techniques
- Backbone architecture comparison: VGG16 vs. ResNet152 performance
- Initialization strategy effects: Pretrained vs. scratch training outcomes
- Computational efficiency: Inference speed and memory requirements
- Dataset-specific performance: Results breakdown by thermal dataset
- Error analysis: Common failure cases and detection limitations

4.1 Training Performance

Reports training loss curves, converegence behavior, and computational requirements for different mdoel variants.

4.2 Detection Accuracy Analysis

Provides detailed mAP scores and detection performance metrics for each model configuraiton and preprocessing technique.

4.3 Preprocessing Impact Evaluation

Analzyes the effects of image inversion and edge enhancement on detection performance.

5 Discussion

The experimental results provide valuable insights into the practical applicability of SSD architectures for thermal human detection systems. While certain configurations demonstrate superior performance, the choice of optimal model depends on specific deployment requirements, including accuracy thresholds, computational constraints, and operational environments. This section interprets the findings within the context of real-world surveillance applications and addresses the broader implications for thermal imaging-based security systems.

Key areas to develop:

- Performance trade-offs: Accuracy vs. computational efficiency analysis
- Preprocessing effectiveness: When and why certain techniques work better
- Backbone selection criteria: Situational advantages of VGG16 vs. ResNet152
- Real-world deployment implications: Edge computing considerations
- Limitations and constraints: Environmental factors affecting performance
- Comparison with existing thermal detection systems
- Cost-benefit analysis for industrial implementation
- Future optimization potential and research directions

5.1 Model Performance Comparison

Compares SSD-VGG and SSD-ResNer performance and discusses trade-offs between accuracy and computational efficiency.

5.2 Practical Deployment Considerations

Discusses real-world application scenarios and system requirements for thermal surveillance.

6 Conclusion and Future Work

This thesis has systematically evaluated the application of Single Shot MultiBox Detector architectures for human detection in thermal imagery, providing empirical evidence for optimal model configurations in surveillance applications. The comprehensive analysis of 16 model variants across multiple thermal datasets has yielded practical insights for deploying neural networks in infrared-based security systems. The findings contribute to both academic understanding and industrial implementation of thermal computer vision technologies.

Key areas to develop:

- Key findings summary: Best-performing model configurations identified
- Methodological contributions: Systematic evaluation framework for thermal detection
- Practical implications: Guidelines for industrial thermal surveillance deployment
- Technical achievements: Successful adaptation of RGB models to thermal domain
- Research limitations: Dataset constraints and environmental factors
- Future research directions: Advanced architectures and multi-modal approaches
- Industry impact: Potential applications beyond security surveillance
- Recommendations: Implementation guidelines for practitioners

7 Examples

7.1 Figures and Tables

Create figures or tables like this:

7.1.1 Figures

Figure 1 — Image Example

7.1.2 Tables

	Area	Parameters
cylinder.svg	$\pi h \frac{D^2 - d^2}{4}$ (2)	h: height D: outer radius d: inner radius
tetrahedron.svg	$\frac{\sqrt{2}}{12} a^3$ (3)	a: edge length

Table 1 — Table Example

7.2 Code Snippets

Insert code snippets like this:

```
const ReactComponent = () => {
  return (
    <div>
      <h1>Hello World</h1>
    </div>
  );
};

export default ReactComponent;
```

Listing 1 — Codeblock Example

For example this Table 1 references the table on the previous page.

Glossary

AP	Average Precision. Calculated as the area under the precision-recall curve.
AdaBoost	Adaptive Boosting. A type of ensemble learning algorithm that combines multiple weak classifiers to form a strong classifier.
Batch	A batch is a group of data processed together as a unit.
Batch Gradient Descent	Batch Gradient Descent is an optimization algorithm used to minimize the loss function in machine learning models by iteratively updating the model parameters based on their gradients with respect to the entire training dataset.
CNN	A convolutional neural network (CNN) is a type of neural network designed to process data with a grid-like topology, such as images.
CUDA	Compute Unified Device Architecture. A parallel computing platform and programming model developed by NVIDIA for general computing on GPUs.
FC Layer	A fully connected layer is a layer in a neural network where each neuron is connected to every neuron in the previous layer.
HOG	Histogram of Oriented Gradients. A feature descriptor used in computer vision and image processing for

	the purpose of object detection. It is based on the distribution of intensity gradients or edge directions in an image.
IoU	Intersection over Union. A metric used to evaluate the accuracy of object detection models. It is calculated as the ratio of the area of overlap between the predicted bounding box and the ground truth bounding box to the area of union between the two boxes.
MPS	Metal Performance Shaders. A framework for accelerating machine learning workloads on Apple Silicon devices.
MS COCO	Microsoft Common Objects in Context. A large-scale object detection, segmentation, and captioning dataset.
NMS	A technique used to eliminate redundant bounding boxes in object detection models. It works by selecting the bounding box with the highest confidence score and eliminating all other bounding boxes that have an IoU greater than a specified threshold with the selected bounding box.
PASCAL VOC	Pascal Visual Object Classes. A dataset for object detection and segmentation tasks.
ReLU	Rectified Linear Unit. A type of activation function used in neural networks. It is defined as $f(x) = \max(0, x)$.
SGD	SGD is an optimization algorithm used to minimize the loss function in machine learning models by itera-

	tively updating the model parameters based on their partial derivatives with respect to a small subset of the training data.
SSD	A single shot multibox detector (SSD) is a type of object detection model that uses a single forward pass of the network to predict the bounding boxes and class scores for all objects in the image.
SVM	Support Vector Machine. A type of supervised learning algorithm that is used for classification and regression tasks. It is based on the idea of finding a hyperplane that best separates the data into different classes.
Tensor	A tensor is a mathematical object that generalizes scalars, vectors, and matrices to higher-dimensional arrays.
VGP	Vanishing Gradient Problem. A problem that occurs in deep neural networks where the gradients of the loss function with respect to the weights become very small, making it difficult to train the network.
ViT	Vision Transformer. A type of transformer model that is designed for computer vision tasks.
YOLO	You Only Look Once. A type of object detection model that uses a single forward pass of the network to predict the bounding boxes and class scores for all objects in the image.
mAP	Mean Average Precision. Calculated as the mean of the AP values for each class.

References

[1] M. A. Farooq, P. Corcoran, C. Rotariu, and W. Shariff, “Object Detection in Thermal Spectrum for Advanced Driver-Assistance Systems (ADAS),” no. arXiv:2109.09854. arXiv, Oct. 2021. doi: 10.48550/arXiv.2109.09854.

[2] K. R. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, “Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms,” *Electronics*, vol. 11, no. 7, p. 1151, Jan. 2022, doi: 10.3390/electronics11071151.

[3] “FREE - FLIR Thermal Dataset for Algorithm Training | OEM.FLIR.Com.”

[4] N. U. Huda, B. D. Hansen, R. Gade, and T. B. Moeslund, “The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection,” *Sensors*, vol. 20, no. 7, p. 1982, Jan. 2020, doi: 10.3390/s20071982.

[5] J. W. Davis and M. A. Keck, “A Two-Stage Template Approach to Person Detection in Thermal Imagery,” in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, Jan. 2005, pp. 364–369. doi: 10.1109/ACVMOT.2005.14.

[6] J. Liu *et al.*, “Target-Aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection,” no. arXiv:2203.16220. arXiv, Mar. 2022. doi: 10.48550/arXiv.2203.16220.

[7] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, “Multispectral Pedestrian Detection: Benchmark Dataset and Baseline,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 1037–1045. doi: 10.1109/CVPR.2015.7298706.

[8] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA: IEEE Comput. Soc, 2001, p. I–511–I–518. doi: 10.1109/CVPR.2001.990517.

[9] A. Dosovitskiy *et al.*, “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” no. arXiv:2010.11929. arXiv, Jun. 2021. doi: 10.48550/arXiv.2010.11929.

[10] Y. LeCun *et al.*, “Handwritten Digit Recognition with a Back-Propagation Network,” in *Advances in Neural Information Processing Systems*, Morgan-Kaufmann, 1989.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” no. arXiv:1311.2524. arXiv, Oct. 2014. doi: 10.48550/arXiv.1311.2524.

[12] R. Girshick, “Fast R-CNN,” no. arXiv:1504.08083. arXiv, Sep. 2015. doi: 10.48550/arXiv.1504.08083.

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” no. arXiv:1506.01497. arXiv, Jan. 2016. doi: 10.48550/arXiv.1506.01497.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” no. arXiv:1506.02640. arXiv, May 2016. doi: 10.48550/arXiv.1506.02640.

[15] W. Liu *et al.*, “SSD: Single Shot MultiBox Detector,” vol. 9905. pp. 21–37, 2016. doi: 10.1007/978-3-319-46448-0_2.

[16] D. K. Alqahtani, A. Cheema, and A. N. Toosi, “Benchmarking Deep Learning Models for Object Detection on Edge Computing Devices,” no. arXiv:2409.16808. arXiv, Sep. 2024. doi: 10.48550/arXiv.2409.16808.

[17] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” no. arXiv:1409.1556. arXiv, Apr. 2015. doi: 10.48550/arXiv.1409.1556.