Pedestrian Detection with Cross-Modal Fusion and a YOLO Head

Kerem Erciyes
Student ID: 910560
University of Milan

Introduction

Object detection in different lighting conditions remains a significant challenge for autonomous systems. Standard RGB cameras fail in low-light scenarios, while infrared (IR) sensors, though effective at thermal detection, often lack rich semantic detail. This project addresses the challenge of low-light pedestrian detection by combining the strengths of both RGB and IR sensors.

The core hypothesis is that by effectively combining features by fusion from these two complementary modalities, a model can achieve more robust and accurate detection performance than by using either spectrum alone. The deep learning pipeline that leverages the complementary strengths of paired RGB and IR image pairs for robust object detection. The proposed model integrates a dual-stream ResNet backbone for parallel feature extraction, a bidirectional cross-attention mechanism and axial attention mechanism for multi-modal fusion, and a lightweight YOLO-style head for efficient, single-stage object detection. The system is trained and evaluated on the LLVIP dataset.

Overview

01 **Dataset and Preprocessing**

02 **Methodology and Model Components**

03 **Loss Computation**

04 **Training and Validation**

05 **Experiments and Results**

Dataset and Preprocessing

The LLVIP Dataset

The LLVIP dataset consists of 15488 of paired visible RGB and infrared images captured in various low-light surveillance scenarios. The primary object class is 'person'.
Image Resizing: All input images (both RGB and IR) are resized to a fixed resolution of 512 × 512 pixels.
Data Augmentation: To prevent overfitting and expose the model to a wider variety of data, several dynamic data augmentation techniques are applied to the training set. These augmentations are on the fly so they are randomly applied to each batch.
– Color Jitter: The brightness, contrast, saturation, and hue of the RGB images are randomly adjusted.
– Geometric Transformations: A RandomAffine transformation is applied, introducing rotations of up to 10 degrees, translations, and scaling.
– Horizontal Flipping: With a 50% probability, both the RGB and IR images in a pair are flipped horizontally. The corresponding bounding box coordinates are adjusted accordingly.

Methodology and Model Components

- Backbone: Multi-Scale Feature Extraction
- Hybrid Feature Extraction with Progressive Fusion
- Deep Supervision Head
- YOLOv3-Style Detection Head

Backbone: Multi-Scale Feature Extraction

Dual Branch Backbone

The foundation of the model is a dual-branch backbone architecture. Two separate ResNet-18 models are employed, sourced from the timm library, to act as feature extractors: one for the RGB stream and one for the IR stream. To handle the single-channel IR input, the weights of the first convolutional layer of the IR backbone are adapted by averaging the original three-channel pretrained weights. Both backbones output feature maps from three different stages (stage2, stage3, and stage4), providing multi-scale representations of the input.

Hybrid Feature Extraction with Progressive Fusion

Bidirectional Cross-Attention Fusion

This module is the primary fusion mechanism. For a given feature scale, it performs attention in two directions: The RGB features query the the IR features, and the IR features query the the RGB features. This allows each modality to dynamically incorporate the most relevant information from the other.
A convolutional gating mechanism learns a set of four attention weight maps, allowing the model to decide on a pixel-by-pixel basis how much to trust each of the four feature streams, which are: original RGB, original IR, fused RGB, fused IR. The weights are enforced to sum to one at each pixel via a \texttt{softmax} operation across the channel dimension.

Axial Attention Block

This module can be applied after cross-modal fusion. It performs self-attention sequentially along the height and then the width axes of the feature map, efficiently providing self-attention model captures long-range spatial dependencies within the fused representation.

ViT Enhancer Block

This module is designed to refine the features from each modality before they undergo cross-modal fusion. The purpose of this module is to leverage the strength of Vision Transformers (ViT) in capturing global context, complementing the local feature extraction.

Deep Supervision Head

To provide a more stable gradient signal to the deep layers of the backbones, deep supervision strategy is applied. A small, classification-only convolutional head (SimpleConvHead) is attached directly to the highest-level feature map (layer4) of one of the backbones before fusion. This head performs dense, per-pixel classification. The resulting auxiliary cross-entropy loss ensures that the backbone is encouraged to learn, mitigating potential vanishing gradient issues.

YOLOv3-Style Detection Head

The final detection stage uses a YOLO Head, a lightweight and efficient single-stage detector head.

Anchor-Based Detection: The head operates on the final fused feature map (e.g., a 16 x16 grid). At each grid cell, it predicts offsets for a set of predefined bounding boxes of various shapes, known as anchor boxes.

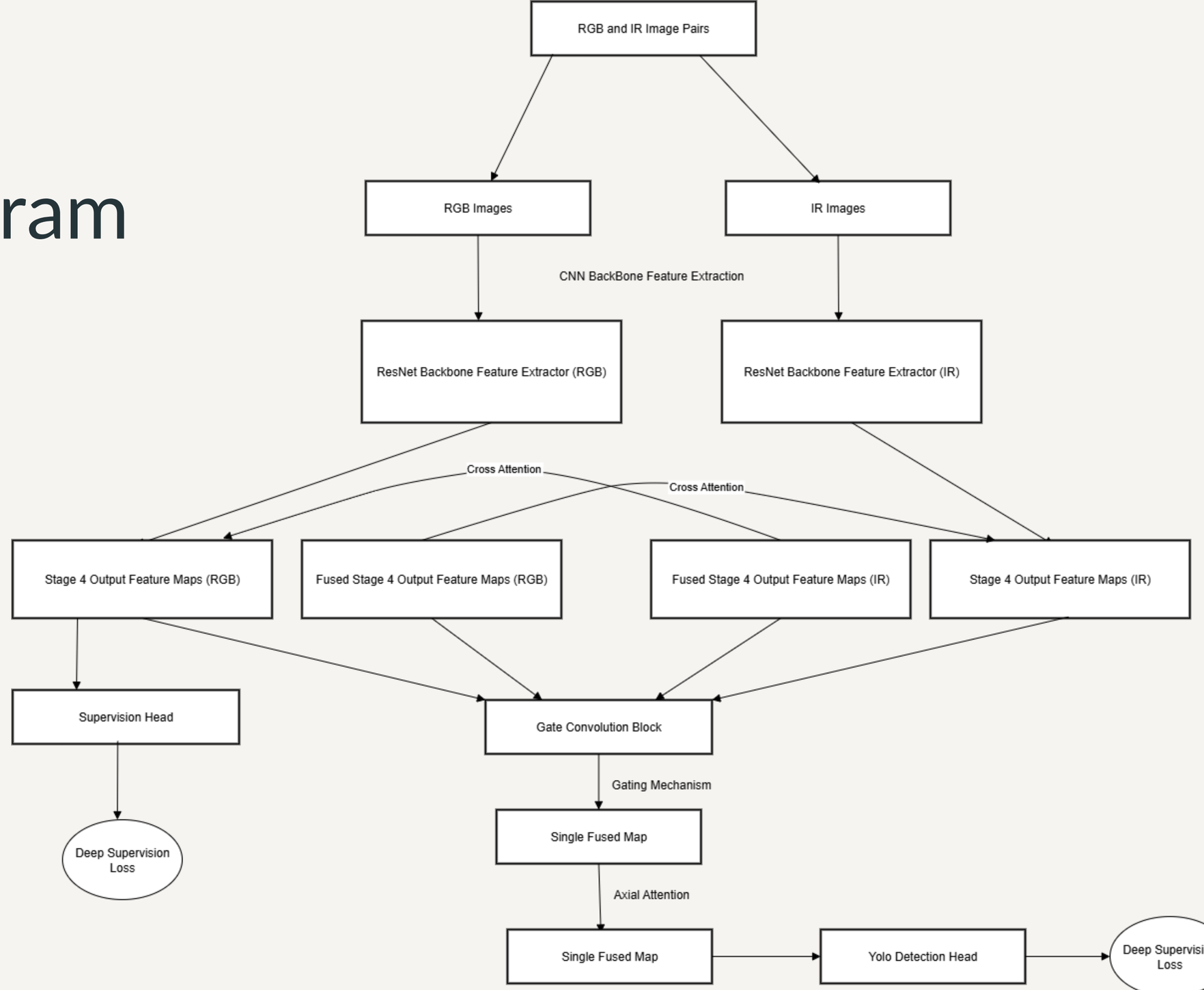
Decoupled Prediction: To improve stability, the head uses three separate 1x1 convolutional layers to predict the three distinct components of a detection:

Box Regression: Predicts the xywh coordinate adjustments for each anchor.

Objectness Score: Predicts the confidence that an anchor box contains an object versus background.

Class Probability: Predicts the probability of the object belonging to a specific class (in this case, 'person').

Model Diagram



Loss Computation

First, the YOLO loss matches ground-truth boxes to anchor boxes based on their shape and location. It then computes a weighted sum of three distinct loss components for the matched pairs which come from: Box Regression and Objectness Score and Class Probability:
Box Loss: A Complete IoU (CIoU) loss is used for bounding box regression. CIoU is an advancement over standard IoU loss as it accounts for the overlap area, central point distance, and aspect ratio consistency between the predicted and ground-truth boxes, leading to faster and more accurate convergence.
Objectness Loss:}A Binary Cross-Entropy (BCE) loss is calculated across all anchors on the feature map. It trains the model to distinguish between foreground (anchors containing objects) and background.
Classification Loss): A BCE loss is calculated only for anchors that contain an object. It pushes the model to correctly classify the object within the box. Label smoothing is applied to this loss to regularize the model and prevent overconfidence.

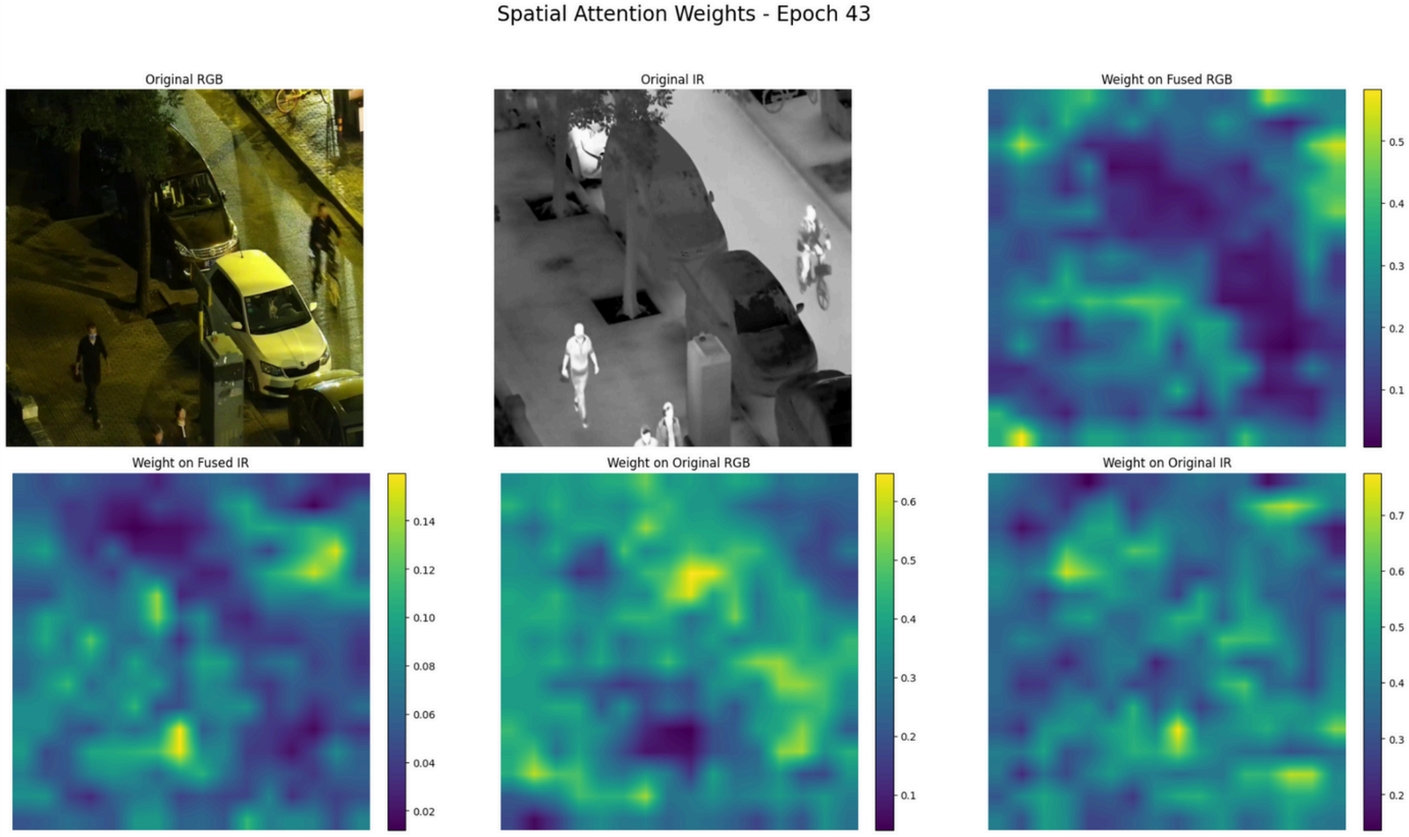
Total Loss Combination

$$\mathcal{L}_{\text{total}} = w_{\text{box}}\mathcal{L}_{\text{CIoU}} + w_{\text{obj}}\mathcal{L}_{\text{BCE\_obj}} + w_{\text{cls}}\mathcal{L}_{\text{BCE\_cls}} + \lambda_{\text{DS}}\mathcal{L}_{\text{DS}}$$

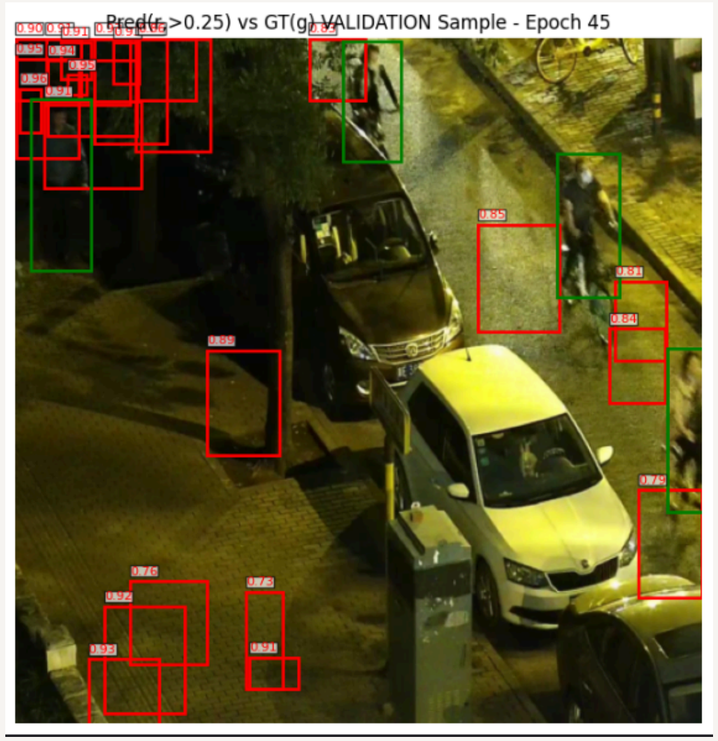
Training and Validation

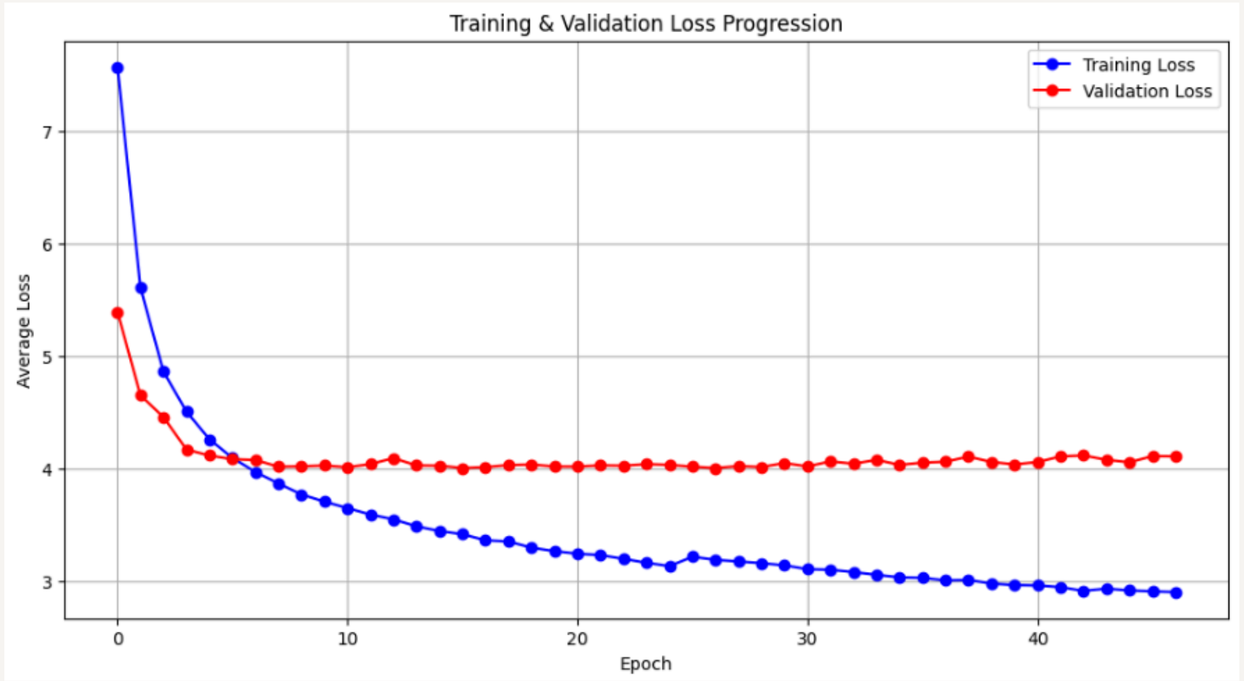
The parameters in the training loop:
Optimizer and Scheduler: A Stochastic Gradient Descent (SGD) optimizer with Nesterov momentum was used, which was found to provide better generalization than Adam-based optimizer. It uses differential learning rates (a lower rate for the backbone, a higher rate for the other layers). This is paired with a Cosine Annealing learning rate scheduler with a linear warmup phase.
Progressive Unfreezing: The backbones are initially frozen to allow the new fusion and detection layers to stabilize. After a set number of epochs, the backbone is progressively unfrozen layer by layer.
Mixed Precision: Automatic Mixed Precision is used to speed up training and reduce memory usage.
Gradient Accumulation: Gradients are accumulated over several batches before an optimizer step is taken, effectively simulating a larger batch size.
Gradient Clipping: To prevent the gradient explosion observed upon unfreezing the backbone, a differential clipping strategy is employed. Gradients for the sensitive backbone are clipped to a small norm (e.g., 1.0), while gradients for the more robust fusion and detection heads are allowed a larger norm (e.g., 10.0). A temporary, much lower learning rate is also applied for a few epochs immediately after unfreezing to ensure a smooth transition.

Experiments and Results



Experiments and Results





Conclusion

In conclusion, in the project, by leveraging a dual-stream ResNet backbone and a progressive fusion strategy featuring Bidirectional Cross-Attention, the model creates a unified feature representation. This representation is then processed by a YOLO-style detection head. The project's findings are unique architecture and the training methodology. Mostly overcoming initial overfitting and gradient explosion required a systematic approach, with a robust training process that includes a warmup-decay scheduler, progressive unfreezing with adaptive learning rates, and differential gradient clipping. The final result is a satisfying and effective pipeline that combines the complementary strengths of RGB and IR imagery for the object detection task.

Discussion and Future Work

For future work, current model's anchor-based YOLO head may struggle with objects of unusual aspect ratios not well-represented by the predefined anchors. Firstly, focusing on maximizing the existing architecture would be better. Activating the optional TimmViTEnhancer to enrich features pre-fusion and implementing a Feature Pyramid Network (FPN) to leverage the currently unused layer2 and layer3 fused maps for improved multi-scale detection. More extensive architectural explorations could involve changing the ResNet-18 backbones for more powerful models like EfficientNetV2, or replacing the detection head with anchor-free alternatives like DETR or two-stage detectors like Faster R-CNN, though this would require significantly more training time and computational power. Finally, for real-world deployment where efficiency is key, the finalized model could be optimized through techniques such as model pruning, quantization, and particularly Knowledge Distillation, which would allow a smaller "student" model to learn from this larger, more complex one, making it suitable for edge devices.

Thank You

17 July 2025

Kerem Erciyes
Student ID: 910560

Supervisor: Professor Pasquale Coscia
University of Milan