Université Clermont Auvergne
Institut d'Informatique d'Auvergne

Research Project Proposal
2nd year of Master's degree
International track in Computer Science

Submitted by

Rohan Mehra

Deep Learning on SWIR Images for Robust Perception in Harsh Weather

---

To be conducted within the ComSee Team , Institut Pascal

Supervised by:

Dr. Mathieu Labussière
Associate Professor

September 2025 – February 2026

Abstract

This proposal focuses on demostrating the benefit of using SWIR camera (short-wave infrared, 0.9–1.7 μm) in autonomous driving tasks for better perception in harsh weather, low-lighting, and glare conditions. The objective of this research is twofold. The first aim is to address the limited availability of Short-Wave Infrared (SWIR) driving datasets by generating synthetic SWIR imagery using state-of-the-art generative models, including both GAN-based and diffusion-based architectures.

The second part of this work focuses on adapting and optimizing object detection models specifically for multispectral RGB-SWIR imaging, with the goal of enhancing perception robustness in adverse weather scenarios. It is planned to explore re-training YOLOv8 and RF-DETR, leveraging both real and synthetically generated SWIR data to learn spectral cues that are absent from standard RGB imagery.

This project advances the work [15] and contributes to the advancement of multimodal perception systems for autonomous vehicles and advanced driver-assistance systems (ADAS), by exploring solutions for reliable detection in challenging conditions. The work paves the way for further integration of RGB and SWIR modalities in computer vision architectures.

Keywords: Road detection, YOLOv8, RF-DETR, multispectral imaging, GAN, Diffusion Models.

Contents

Chapter 1

Introduction

1.1 Context and Challenges

Object detection in road environments is a fundamental pillar of autonomous driving systems and advanced driver-assistance systems (ADAS). These technologies rely on robust, real-time perception of the environment, including the detection of pedestrians, vehicles, cycles, and other obstacles, even under complex conditions [5]. Historically, traditional computer vision approaches based on handcrafted features, such as HOG descriptors or Support Vector Machines (SVM), have shown their limitations when facing the variability of road scenes (lighting, weather, occlusions) [1]. The emergence of Convolutional Neural Networks (CNNs) marked a turning point, offering significant improvements in accuracy and speed—two essential aspects for real-time applications [19].

However, visible (RGB) sensors, while efficient under optimal conditions, suffer from major limitations in low-light, foggy, rainy, or snowy environments. Such adverse conditions reduce the reliability of perception systems, thus compromising the safety of autonomous vehicles [20]. To address these challenges, the integration of complementary sensing modalities—such as infrared imaging (SWIR, short-wave infrared)—could be a promising research direction [15].

1.2 Relevance of the SWIR Modality

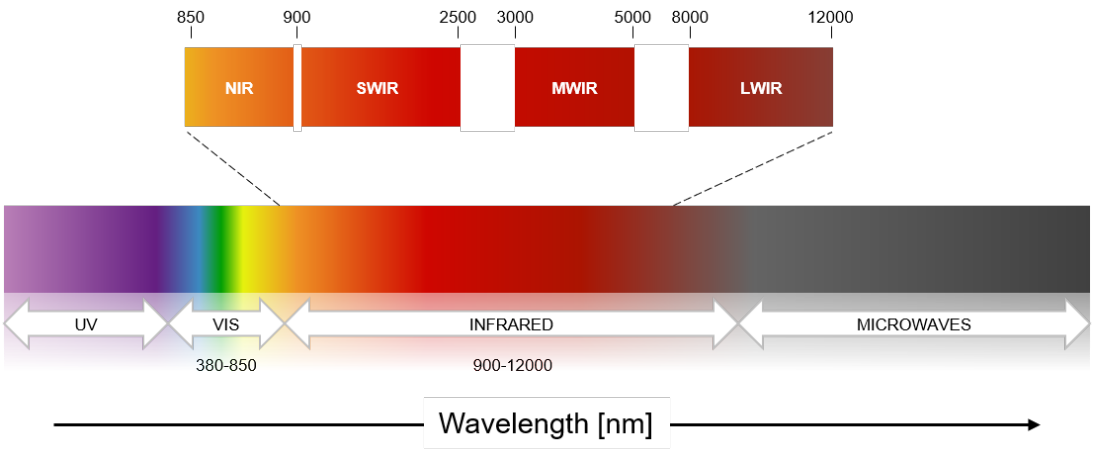


Figure 1.1: SWIR in the electromagnetic spectrum [16]

Short-Wave Infrared (SWIR) imaging—typically ranging from 0.9 to 1.7 μm—offers significant advantages for road scene perception, particularly under conditions where visible sensors perform poorly [7]. Unlike RGB images, which capture the visible spectrum (400–700 nm), SWIR sensors operate in the near-infrared range, sensitive to the material-specific properties of surfaces (reflection, absorption). This feature enables improved penetration through obstacles such as fog, rain, or dust, and enhances visibility in low-light or nighttime conditions [20, 1]. For instance, SWIR imagery reveals texture and material contrasts that are invisible in RGB, facilitating the detection of objects such as pedestrians or vehicles in complex environments [12].

Figure 1.2: Visual comparison between RGB and SWIR imaging [25]

Recent studies have shown that to deploy deep learning object detection algorithms on SWIR (or multimodal with SWIR + visible) [15] there is need for specialization: adapting the model to the SWIR domain via fine-tuning, domain adaptation, distillation etc.

1.3 Project Problem Statement

Adapting high-performing object detection models to non-conventional modalities like SWIR raises several challenges: low contrast, sensor noise, scarcity of annotated data, and domain discrepancy with RGB. While YOLOv8 achieves outstanding real-time detection performance on RGB images thanks to its optimized architecture and pre-training on COCO [23], its direct application to SWIR imagery remains limited. [15]

The objective of this research is twofold. First, we aim to address the scarcity of Short-Wave Infrared (SWIR) driving datasets by generating synthetic SWIR imagery using state-of-the-art generative models, including GAN-based and diffusion-based architectures. This synthetic data will serve as an auxiliary modality to complement limited real SWIR captures and support model training under diverse environmental conditions.

The second phase of this research focuses on the specializing of object detection models for SWIR imaging, with the goal of improving perception robustness under adverse weather conditions. We plan to investigate SWIR-adapted variants of YOLOv8 and RF-DETR, leveraging both real and synthetically generated SWIR data to learn spectral characteristics that are not present in conventional RGB imagery.

1.4 Research Project Objectives

Below is the brief summary of planned work:

- **Literature review**: Analyze the state of the art in road detection, focusing on object detection algorithms like YOLOv8 and RF-DETR etc, specialization techniques (fine-tuning, knowledge distillation), and multimodal RGB-SWIR imaging [10, 20, 14].
- **Synthetic SWIR generation**: Train state-of-the-art GAN/diffusion models for translation of RGB images to SWIR images.
- **Specialization of object detection algorithms**: Adapt the YOLOv8 and RF-DETR model to SWIR imagery through fine-tuning using the RASMD dataset, and explore with techniques involving domain adaptation and knowledge distillation.
- **Performance evaluation**: Compare performance of different generative models for translation tasks as well as draw a comparison of object detection models using various metrics.
- **Optimization and robustness**: Propose a model to enhance detection on SWIR images in adverse conditions (fog, low light). Additionalyy, propose a choice of generative model for the RGB-SWIR translation task.
- **Documentation and dissemination**: Produce a clear and structured report and present the results.

Chapter 2

Models and Methods

Road object detection faces specific constraints, particularly in difficult conditions such as night, fog, rain, or snow. Visible (RGB) sensors lose efficiency in these scenarios due to low luminosity or light diffusion by atmospheric particles [20]. These challenges necessitate the integration of complementary sensors, such as SWIR (Short-Wave Infrared, 0.9–1.7 μm) cameras, which offer better robustness due to their capacity to penetrate fog and reveal specific material contrasts [7, 12]. However, the use of unconventional modalities like SWIR poses additional problems: scarcity of annotated data, domain shift compared to RGB, and the need for geometric and temporal alignment for multimodal fusion [9]. This chapter explores a range of generative models that can be employed to produce synthetic SWIR images, as well as several object detection architectures that can be fine-tuned to enhance perception in SWIR images under harsh weather conditions. Additionally, it provides an overview of the evaluation metrics used to assess both the image generation and object detection tasks.

2.1 Generative Models for Synthetic SWIR Image

This work proposes to experiment with six GAN and diffusion-based models for translating RGB images into the SWIR domain. Table 2.1 summarizes the generative models selected for training. The set includes one diffusion-based paired translation model, two GAN-based paired translation models, and three GAN-based unpaired translation models. The following subsections provide detailed descriptions of each model, along with their architectures and loss functions etc.

Model	**Type**	**Paired/Unpaired**
CUT [17]	GAN	Unpaired
FastCUT [17]	GAN	Unpaired
CycleGAN [30]	GAN	Unpaired
Pix2Pix [8]	GAN	Paired
Pix2PixHD [26]	GAN	Paired
BBDM [13]	Diffusion	Paired

Table 2.1: List of generative models planned for RGB-to-SWIR image translation. One diffusion based paired translation, two GAN based paired translation, and 3 GAN based unpaired translation models will be used

2.1.1 CycleGAN

CycleGAN [30] enables image translation without paired training data. Its architecture contains two generators (G : RGB $\rightarrow$ SWIR and F : SWIR $\rightarrow$ RGB) and two discriminators (D_{SWIR} and D_{RGB}). The key component is the cycle-consistency loss:

$$\mathcal{L}_{\text{cyc}} = \|F(G(x)) - x\|_1 + \|G(F(y)) - y\|_1, \tag{2.1}$$

which ensures the translated image retains the geometry of the original. Additionally, CycleGAN uses adversarial losses:

$$\mathcal{L}_{\text{GAN}}(G, D_{\text{SWIR}}) = \mathbb{E}_y[\log D_{\text{SWIR}}(y)] + \mathbb{E}_x[\log(1 - D_{\text{SWIR}}(G(x)))]. \tag{2.2}$$

This makes CycleGAN suitable for RGB-to-SWIR translation when data are unaligned or collected separately. However, it may produce texture artifacts because both forward and backward mappings must be learned simultaneously.

2.1.2 CUT

Contrastive Unpaired Translation (CUT) [17] simplifies the CycleGAN architecture by removing the backward generator (F) and cycle-consistency loss. Instead, it introduces a PatchNCE contrastive objective that maximizes mutual information between corresponding spatial patches in the input and output:

$$\mathcal{L}_{\text{PatchNCE}} = -\sum_l \sum_i \log \frac{\exp(f_i^l \cdot f_i^{l,+}/\tau)}{\sum_j \exp(f_i^l \cdot f_j^{l,-}/\tau)}, \tag{2.3}$$

where $f_i^{l,+}$ is the positive feature (same location after translation), and $f_j^{l,-}$ are negatives.

CUT uses a single generator with adversarial loss and is trained with identity loss and $\lambda_{\text{NCE}} = 1$. The multi-layer PatchNCE loss ($\sum_l$ over multiple layers l) enables powerful distribution matching while effectively preserving local structures. This makes it particularly beneficial for SWIR synthesis where geometric stability is crucial.

2.1.3 FastCUT

FastCUT [17] is an optimized variant that reduces the number of contrastive layers from multiple to just one, dramatically decreasing memory and computational requirements. It employs a higher contrastive weight ($\lambda_{\text{NCE}} = 10.0$) and omits the identity loss.

FastCUT retains:

- adversarial loss for output realism,
- a single-layer PatchNCE loss for structural consistency.

While CUT performs more flexible distribution matching, FastCUT behaves more conservatively similar to CycleGAN but with significantly improved efficiency—using approximately

half the GPU memory and training twice as fast. This makes it particularly appealing for processing large unpaired RGB-SWIR datasets where computational resources are constrained.

2.1.4 Pix2Pix

Pix2Pix [8] is a paired conditional GAN model that learns a direct mapping from RGB to SWIR images. It employs a U-Net generator, which incorporates skip connections to preserve low-level spatial information, and a PatchGAN discriminator that evaluates realism at the patch level. Its objective combines:

$$\mathcal{L}_{\text{Pix2Pix}} = \mathcal{L}_{\text{GAN}}(G, D) + \lambda \|y - G(x)\|_1, \tag{2.4}$$

where the L_1 term enforces pixel-level accuracy. Because paired RGB–SWIR data are aligned, Pix2Pix can learn fine spectral transformations and produce high-fidelity outputs, making it a strong baseline when paired data are available.

2.1.5 Pix2PixHD

Pix2PixHD [26] extends the original Pix2Pix framework for high-resolution image-to-image translation. To handle the increased complexity and instability of training at high resolutions, it introduces a multi-scale, coarse-to-fine generator and employs three distinct discriminators (D_1, D_2, D_3) that assess image authenticity at different spatial scales. A key component for stabilizing this adversarial training is a feature-matching loss, which minimizes the difference in intermediate feature representations between real and generated images as seen by each discriminator:

$$\mathcal{L}_{\text{FM}}(G, D_k) = \mathbb{E}_{(x,y)} \sum_{l=1}^{L} \frac{1}{N_l} \|D_k^{(l)}(y) - D_k^{(l)}(G(x))\|_1, \tag{2.5}$$

where $D_k^{(l)}$ denotes the l-th layer feature extractor of the k-th discriminator. By enforcing perceptual consistency across scales, Pix2PixHD effectively preserves fine textures and global structure, which is essential for translating RGB to SWIR where material reflectance properties must be accurately rendered.

2.1.6 BBDM

The Browning Bridge Diffusion Model (BBDM) [13] is a diffusion-based framework for paired image-to-image translation. It introduces a novel "Browning Bridge" which is a forward diffusion process that starts from the *target* image y and converges towards the *source* image x, rather than pure noise. This creates a structured, continuous pathway between the two domains.

The model features a bilateral architecture with two branches that denoise this bridging process:

- A **global branch** that operates on a low-resolution version of the noisy image to capture the overall structure and semantic layout.

- A **local branch** that processes high-resolution patches to recover fine-grained textures and local details.

This design efficiently separates the learning of global scene composition from local texture rendering. The training objective is based on denoising the Browning Bridge:

$$\mathcal{L}_{\text{BBDM}} = \mathbb{E}_{(x,y),t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, x, t)\|_2^2\right], \tag{2.6}$$

where z_t is the noisy image at timestep t on the bridge between y and x, and ϵ_θ is the denoising network conditioned on the source image x. By leveraging paired data and this more structured diffusion process, BBDM generates high-quality, artifact-free translations, making it highly valuable for producing precise synthetic SWIR imagery when paired datasets are available.

2.2 Deep Learning Object Detection Models

Deep learning-based object detectors fall into two main categories: two-stage and one-stage detectors. Two-stage detectors, such as Faster R-CNN, first proceed by generating region proposals (Region Proposal Network), followed by classification and bounding box refinement. These models offer high accuracy but are often too slow for real-time applications [19]. In contrast, one-stage detectors, like the YOLO (*You Only Look Once*) and Detection Transformer (*DETR*) family, perform detection and classification in a single pass, prioritizing *speed* over a slightly lower accuracy in some cases [10, 21]. This speed is essential for road object detection.

2.2.1 YOLO (You Only Look Once)

YOLO (*You Only Look Once*) is a family of one-stage object detectors that treats detection as a regression problem, performing simultaneous object localization and classification in a single forward pass [19]. Unlike two-stage approaches, YOLO divides the input image into a grid where each cell predicts bounding boxes (with coordinates, dimensions, and objectness scores) and class probabilities. This unified architecture enables significantly faster inference, making YOLO particularly suitable for real-time applications.

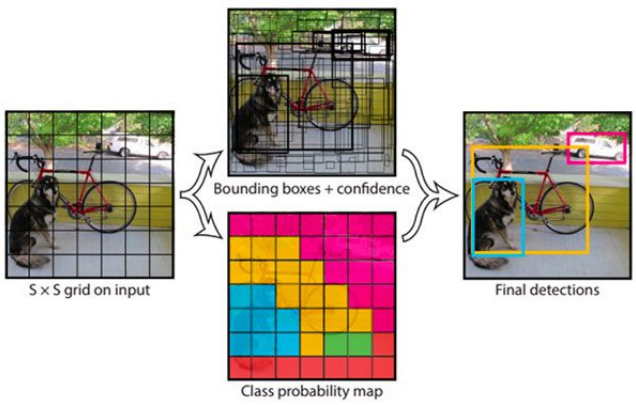


Figure 2.1: Operating principle of YOLO models: image grid division with bounding box and class probability prediction per cell [19]

The YOLO architecture comprises three main components: a **backbone** (e.g., Darknet, CSPDarknet) for feature extraction, a **neck** (e.g., PANet, FPN) for multi-scale feature aggregation, and a **head** that generates final predictions. The training typically combines localization loss (IoU, CIoU), classification loss (BCE, Focal Loss), and in recent versions, a Focal Distribution Loss (DFL) for improved accuracy [22].

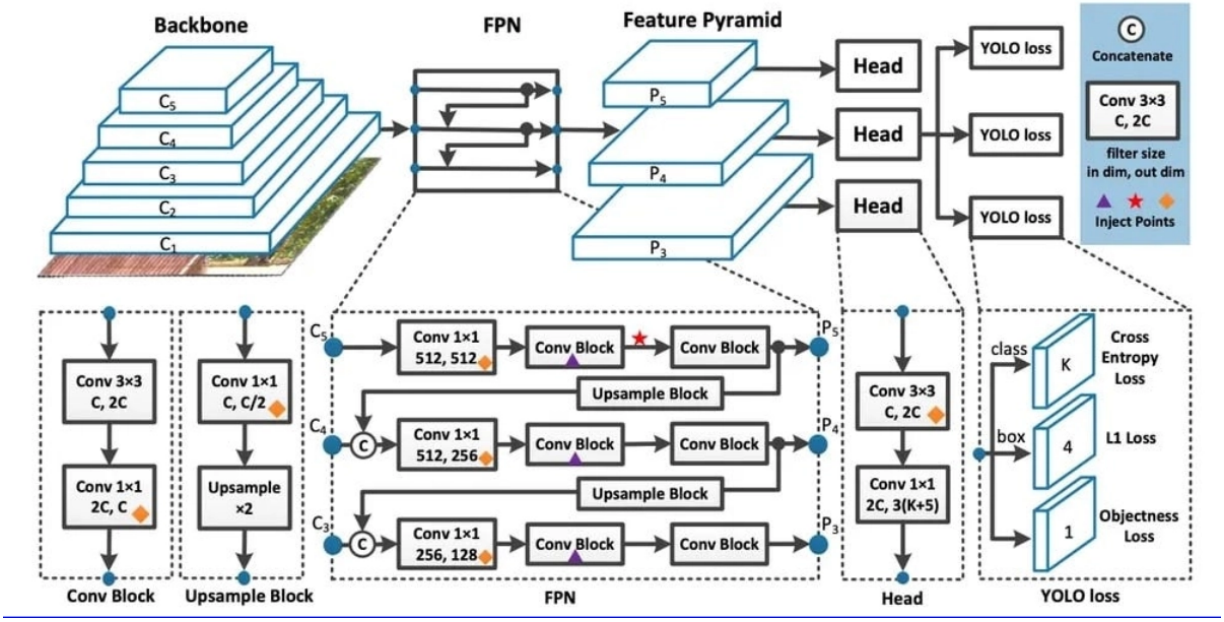


Figure 2.2: YOLOv8 Architecture showing backbone, neck, and head components [24]

YOLO offers several advantages for road object detection: **high-speed inference** with millisecond-level processing, **competitive accuracy** approaching two-stage detectors in recent versions, **architectural flexibility** with variants (nano to extra-large) for different hardware constraints, and **ease of use** through implementations like Ultralytics' PyTorch framework. However, YOLO can be sensitive to domain shift and may require specialization techniques for optimal performance in specific deployment contexts [12, 10].

2.2.2 RF-DETR

RF-DETR (*Roboflow-DETR*) is a state-of-the-art object detection and segmentation models that leverage the Detection Transformer (DETR) architecture to achieve high accuracy and real-time performance [21]. Unlike YOLO that require heuristic post-processing steps such as *Non-Maximum Suppression (NMS)*, RF-DETR is an end-to-end model. It treats object detection as a direct set prediction problem, using a transformer and bipartite matching to simultaneously predict bounding boxes and class labels in a single, streamlined forward pass. This simplification enhances efficiency and predictability in the inference pipeline.

The RF-DETR architecture is engineered for speed, building upon efficient transformer variants like Deformable DETR. It maintains the core structure of detection transformers, comprising three key modules: a backbone (a robust Vision Transformer like DINOv2) for feature extraction, a Transformer Encoder-Decoder module, and a final Prediction Head. The decoder utilizes a set of learned object queries and employs computationally efficient mecha-

nisms, such as deformable cross-attention, to focus selectively on relevant feature regions and output all object predictions in parallel. Figure 2.3 demonstrated the architecture of RF-DETR in detail.

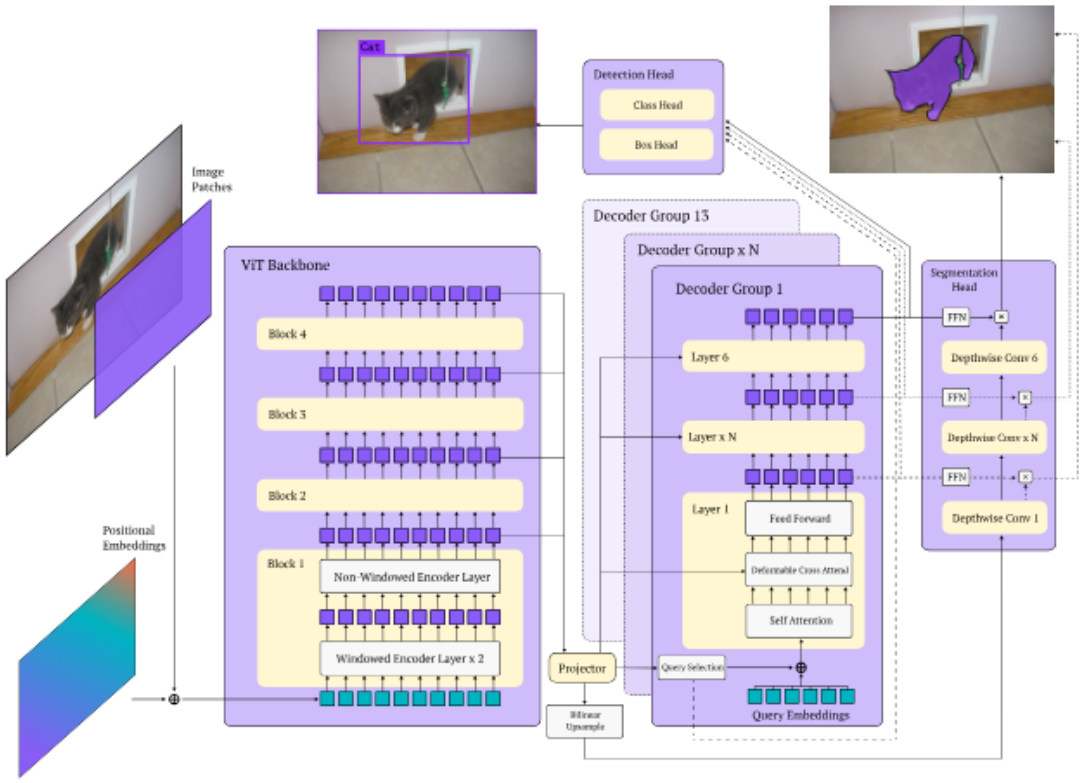


Figure 2.3: RF-DETR Architecture [21]

2.3 Evaluation Metrics

To quantitatively evaluate the performance of the different models, a set of standard metrics will be employed. The generative models for RGB-to-SWIR translation will be assessed using perceptual and pixel-wise metrics, while the object detection and segmentation models will be evaluated using standard object-level metrics.

2.3.1 Generative Model Metrics

The quality of the synthetic SWIR images generated by the models (CycleGAN, CUT, Pix2Pix, etc.) will be evaluated using the following metrics:

- **Fréchet Inception Distance (FID)** [27]: Measures the similarity between the distributions of generated and real SWIR images by comparing their statistics in a feature space from a pre-trained network. A lower FID indicates more realistic generated images.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Quantifies perceptual similarity based on deep feature distances, aligning well with human judgment. Lower scores indicate higher perceptual quality.

- **Peak Signal-to-Noise Ratio (PSNR)**: Evaluates pixel-level fidelity by measuring the ratio between the maximum possible power of an image and the power of corrupting noise. Higher values are better.
- **Structural Similarity Index Measure (SSIM)**: Assesses the perceived quality by comparing the luminance, contrast, and structure between images. Values range from -1 to 1, with 1 indicating perfect similarity.
- **Root Mean Square Error (RMSE)**: Provides a direct measure of the average magnitude of pixel-wise intensity differences. Lower RMSE values are better.

2.3.2 Object Detection and Segmentation Metrics

The performance of the object detectors (YOLO, RF-DETR) will be evaluated using standard metrics based on the Intersection over Union (IoU). A detection is considered a correct True Positive (TP) if the IoU between the predicted bounding box and the ground-truth exceeds a predefined threshold (e.g., 0.5). The following metrics will be calculated using **macro** averaging, where the metric is computed independently for each class and then averaged, treating all classes equally.

Precision, Recall, and F1-Score

These metrics provide a comprehensive view of detection performance by considering the balance between correct detections and errors.

- **Precision** measures the accuracy of positive predictions, i.e., the fraction of correct detections among all predicted objects:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.7}$$

- **Recall** measures the ability to find all relevant objects, i.e., the fraction of true objects that were successfully detected:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.8}$$

- The **F1-Score** is the harmonic mean of Precision and Recall, providing a single balanced metric that is robust to class imbalance:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.9}$$

Average Precision (AP) and mean Average Precision (mAP)

Average Precision (AP) summarizes detector performance for a single class by computing the area under the Precision-Recall curve across all confidence thresholds. Following the COCO evaluation standard, we report:

- **AP** (**AP**$^{@}$[**.50** : **.95**]): Primary metric, averaged over 10 IoU thresholds from 0.50 to 0.95
- **AP**$^{@\mathbf{0.50}}$: AP at IoU=0.50 (PASCAL VOC standard)
- **AP**$^{@\mathbf{0.75}}$: AP at strict IoU=0.75

The mean Average Precision (mAP) provides the overall benchmark by averaging AP across all object classes:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2.10}$$

where N is the number of classes and AP_i is the Average Precision for class i.

2.4 Methods for Specialization with Limited Data

In order to adapt detection models for better perception through SWIR images, the following techniques are also planned to be explored:

2.4.1 Transfer Learning

Transfer learning leverages weights from models pre-trained on large, generic datasets (e.g., COCO) to initialize a model for a more specialized task [31, 3]. This approach accelerates convergence by reusing low- and mid-level representations such as edges, contours, and textures. In our work, we evaluate two fine-tuning regimes: (i) *full fine-tuning*, where all parameters are updated to capture SWIR-specific appearance statistics, and (ii) *partial fine-tuning*, where early backbone layers are frozen to retain generic visual features while only higher-level, task-adapted layers are updated. Such controlled adaptation of the backbone, neck, and detection head optimizes performance on the SWIR domain using joint classification and localization losses [22]. Nevertheless, fine-tuning without constraints can induce catastrophic forgetting of generic features [4].

2.4.2 Few-shot Learning

Few-shot learning tackles scenarios where only a small number of annotated samples are available. Methods include meta-learning, metric or prototype-based classification, and self-training [18]. In road-scene perception, these approaches enable rapid adaptation to rare weather conditions or new environments with minimal supervision. However, due to the risk of overfitting in extremely low-data regimes, few-shot strategies are often complemented with stabilization techniques such as knowledge distillation or carefully designed regularization [28].

2.4.3 Domain Adaptation (Visible $\leftrightarrow$ SWIR)

Domain adaptation mitigates the distribution shift between a labeled source domain (e.g., RGB) and an unlabeled or sparsely labeled target domain (e.g., SWIR) [29, 14]. Approaches commonly rely on adversarial feature alignment, domain-invariant representation learning, consistency regularization, or pseudo-labeling. In RGB–SWIR multimodal settings, adaptation

additionally requires handling geometric misalignment and radiometric differences between sensors, along with designing effective fusion mechanisms to exploit complementary spectral information [9].

2.4.4 Knowledge Distillation (KD)

Knowledge distillation transfers supervisory signals from a high-capacity *teacher* network to a lightweight or domain-specialized *student* model [6]. Distillation can operate at the level of logits, intermediate feature maps, or attention distributions, enabling the student to inherit the teacher's representational structure even with limited labeled data. In our context, KD is particularly beneficial when adapting RGB-pretrained models to SWIR data, serving both as a regularizer and a means of preserving generalizable features [11].

Summary

The planned work consists of using generative AI, including GANs/Diffusion-based models trained on a small set of data for the translation of RGB images to SWIR images. Further, conduction of various experiments with different training approaches, as well as different mixtures of synthetic and real SWIR data, in order to fine-tune the object detection models has been planned.

Chapter 3

Dataset and Experiments

The effectiveness of an object detection model largely depends on the quality and diversity of the data used for its training. In this work, several strategies would be considered and implemented to obtain the most suitable dataset for multimodal object detection.

3.1 Acquisition via CEREMA

Our initial objective was to collect synchronized RGB–SWIR images in a controlled environment, in collaboration with **CEREMA** (Centre for Studies and Expertise on Risks, the Environment, Mobility and Development), particularly through its site in Clermont-Ferrand [2]. CEREMA specializes in transport infrastructure, sustainable mobility, and road safety, and has a climatic testing platform capable of simulating fog and rain conditions.

The main advantage of this approach lay in the full control of acquisition conditions (illumination, calibration, spatial alignment of cameras) and the ability to produce high-quality RGB–SWIR image pairs.

The acquisition setup included the following cameras:

- a **Visible** camera (visible spectrum, RGB),
- a **SVS** camera (InGaAS-based short-wave infrared, SWIR)
- a **Xenics** camera (InGaAS-based short-wave infrared, SWIR).

However, due to time constraints and more dynamic real-world data, this initial plan could not be completed. Consequently, it was decided to rely on the only existing public dataset: **RASMD**[9].

3.2 The RASMD Dataset

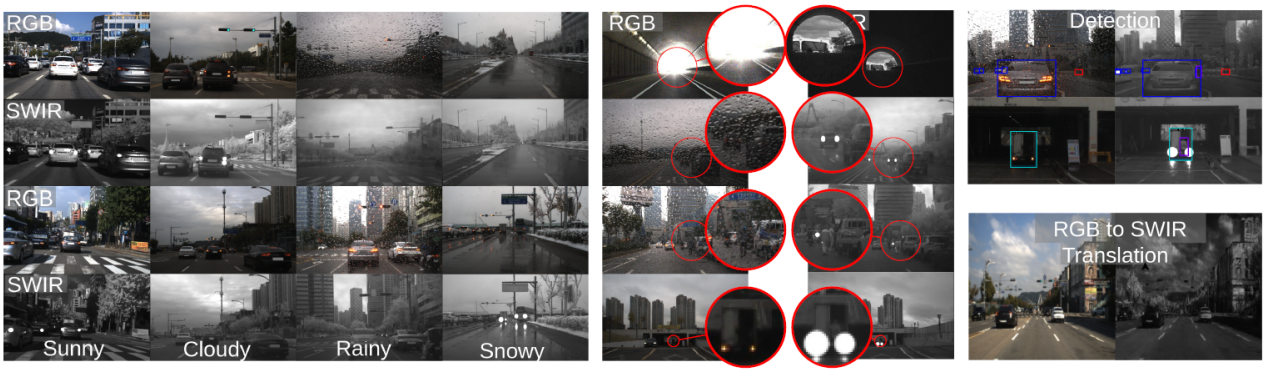
The **RASMD** (RGB and SWIR Multispectral Driving Dataset) was selected as the main data source for this project, given its strong relevance to object detection in challenging driving environments.

3.2.1 Origin

Introduced by Jin *et al.* [9], RASMD is the first large-scale public dataset providing synchronized RGB and SWIR image pairs specifically designed for autonomous driving research. This dataset addresses the need for multimodal perception studies under adverse weather conditions such as fog, rain, or low illumination, where SWIR sensors offer significant advantages over conventional RGB cameras [7, 20].

3.2.2 Contents

RASMD contains a total of **100,000 synchronized and spatially aligned RGB–SWIR image pairs**, covering a wide range of locations, lighting conditions, and weather scenarios (sunny, cloudy, rainy, snowy). The object detection set contains 6 labels (person, car, truck, bus, bicycle, motorcycle).ove



(a) Weather diversity (b) Paired SWIR and RGB modalites (c) Multipurpose dataset

Figure 3.1: Visual overview of the RASMD dataset [9]

3.2.3 Acquisition

The acquisition was done with two cameras:

- **RGB**: FLIR GS3-U3-32S4C-C (2048×1536 pixels, up to 120 FPS);
- **SWIR**: CREVIS HG-A130SW (1296×1032 pixels, up to 70 FPS).

Data were collected over 163.3 km and 8.5 hours of driving, resulting in 100,000 image pairs distributed across various contexts: urban, suburban, sunny (43.2k), cloudy (33.4k), rainy (10.7k), and snowy (12.7k).

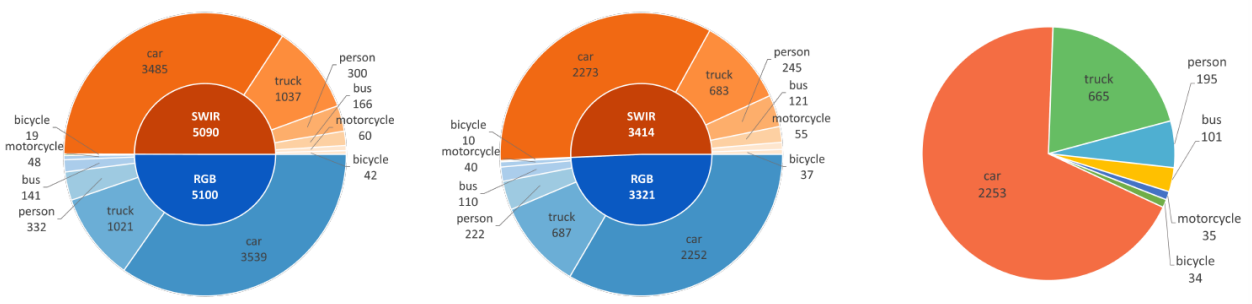
Temporal synchronization was achieved using a software trigger, and geometric alignment was ensured through calibration using a printed checkerboard visible in the SWIR spectrum.

3.2.4 Limitations of Dataset

Although RASMD provides a valuable multimodal dataset, it also presents several limitations that must be considered. These weaknesses point to directions for improving future multimodal datasets.

1. **Image quality:** some SWIR images appear very dark or lack detail, reducing interpretability and hindering object detection under challenging conditions.

2. **Annotation volume and distribution:** among the **100,000 image pairs**, only 4,776 are annotated. This limited volume restricts generalization capabilities. Moreover, the distribution between *train/test* subsets and between RGB/SWIR modalities is unbalanced, which can introduce bias. Moreover, the SWIR and RGB data has been annotated separately, limiting the pairedness of RGB-SWIR image for object detection.



(a) Train dataset (b) Test dataset (c) Merged labels test dataset

Figure 3.2: Class distribution in the RASMD dataset

3. **Annotation Classes:** annotations are limited to six classes (*person, car, motorcycle, bus, truck, bicycle*). However, real-world driving requires detecting other relevant objects such as traffic lights, road signs, crosswalks, street furniture, animals, and even birds. This limitation reduces the dataset's applicability to broader driving scenarios.

4. **Environmental diversity:** annotated data only cover daytime scenes, excluding nighttime and snowy conditions—precisely where SWIR imaging could provide benefit.

5. **Geographical diversity:** the acquisition sites lack strong geographic variation. Local features (architecture, signage, landscape) may induce bias and reduce the generalization potential of trained models.

3.2.5 Dataset Selection

Owing to the limitations presented in the above section, a small subset of paired RGB-SWIR data (around 4000 images) covering all weather conditions, as well as urban/suburban conditions, for training the generative models towards translation. Moreover, a set of around 2000 separate images covering all weather conditions were re-annotated at the lab in order to translate labels from RGB to SWIR images which could be used to evaluate the object detection models.

3.3 SWIR Image Preprocessing

To enhance the quality of SWIR images in the RASMD dataset, a preprocessing step will be applied as proposed by Mehra *et al.*[15]. The goal would be to reduce the influence of

extreme brightness values and achieve a more homogeneous grayscale distribution, facilitating the feature extraction process for detection models.

3.3.1 Methodology

The preprocessing pipeline consists of three main steps:

1. **Statistical analysis:** compute the mean μ and standard deviation σ of pixel intensities for each image.

2. **Adaptive thresholding:** define a threshold as $\mu + 2.58\sigma$, corresponding to a 99% confidence interval. Pixels exceeding this threshold are truncated to reduce the impact of outliers.

3. **Normalization:** the resulting image is linearly rescaled to the intensity range $[0, 255]$, ensuring better comparability across samples.

3.3.2 Visual Results of preprocessing

This preprocessing improves global contrast and reduces overexposed regions. The post-processed intensity histogram is more evenly distributed, making feature extraction more effective for detection. Figure 3.3 shows an example of a raw SWIR image compared with its preprocessed counterpart.

Figure 3.3: Example of preprocessing: original SWIR image (left) and after thresholding and normalization (right).

Chapter 4

Preliminary Results

This chapter presents preliminary results obtained towards training generative models for the translation of RGB images to SWIR images.

4.1 Training Dataset and Configuration

Approximately 4,000 paired images were selected, encompassing all weather conditions and diverse driving scenarios. The data set was split into (`train`, `val`, `test`) sets in (80%,10%,10%) ratio, such that every set contains images from all weather conditions containing different scenes than the images used in the training set.

4.2 Implementation Details

4.2.1 Hardware

- **GPU:** NVIDIA GeForce RTX 2080
- **VRAM:** 8 GB
- **Driver Version:** 570.124.06
- **CUDA Version:** 12.8

4.2.2 Software

- **Operating System:** Ubuntu 22.04 LTS
- **PyTorch:** Compiled with CUDA 12.5

4.3 Preliminary Results

Figure 4.1 shows visual results obtained on training six different models as listed in Table 2.1. Table 4.1 further gives us a quantitative comparison using metrics mentioned in subsection 2.3.1. Pix2PixHD and CycleGAN perform comparative and hence give us a promising direction to move ahead.

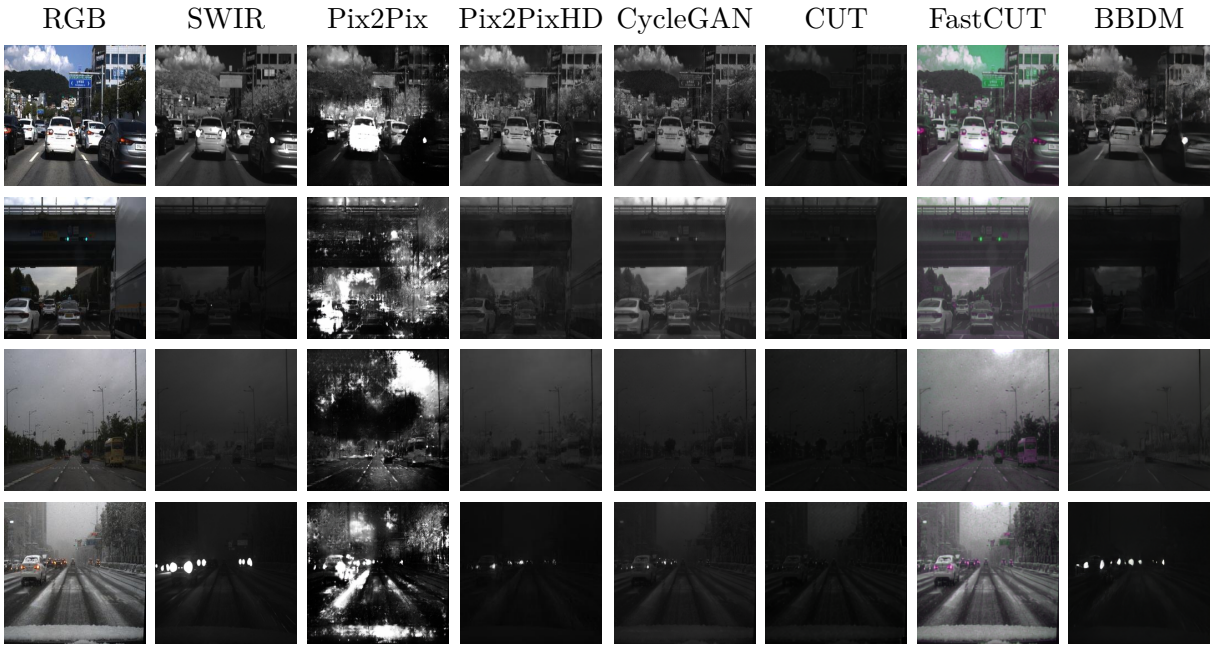


Figure 4.1: Visual Comparison of RGB to SWIR translation results across different models.

Model	**FID ↓**	**LPIPS ↓**	**RMSE ↓**	**PSNR ↑**	**SSIM ↑**
CycleGAN	56.39	0.195 ± 0.077	31.79 ± 25.83	20.44 ± 6.39	0.673 ± 0.175
Pix2PixHD	78.34	0.167 ± 0.069	23.90 ± 20.92	22.72 ± 5.88	0.724 ± 0.141
BBDM	84.39	0.218 ± 0.069	25.64 ± 17.60	21.61 ± 5.37	0.644 ± 0.149
FastCUT	99.08	0.371 ± 0.093	63.40 ± 21.13	12.73 ± 3.68	0.445 ± 0.140
CUT	112.31	0.348 ± 0.123	48.17 ± 36.86	16.82 ± 6.59	0.482 ± 0.224
Pix2Pix	145.18	0.455 ± 0.117	70.10 ± 17.63	11.44 ± 1.87	0.316 ± 0.107

Table 4.1: Quantitative comparison of different methods using image quality metrics. Higher values are better for PSNR and SSIM; lower values are better for RMSE, FID, and LPIPS.

Pix2PixHD emerges as the superior model regarding structural fidelity and pixel-level precision. It achieves the best performance in four out of five metrics: **LPIPS** (0.167), **RMSE** (23.90), **PSNR** (22.72), and **SSIM** (0.724). This dominance suggests that Pix2PixHD is particularly effective at maintaining the content and layout of the original image while performing the translation task.

While Pix2PixHD excels in structural metrics, **CycleGAN** achieves the lowest **FID score** (56.39). Since FID measures the distance between the distributions of generated images and real images in the feature space, this result indicates that CycleGAN generates images that are, distribution-wise, most similar to the target domain. This creates an interesting contrast: while Pix2PixHD is better at reconstructing specific pixel arrangements (lower RMSE), CycleGAN may produce outputs that appear more "natural" or stylistically consistent with the target class, even if the underlying structure deviates slightly more from the input.

However, Pix2Pix struggles to adapt to the SWIR domain and it maybe chosen to be discarded from further studies.

Chapter 5

Conclusion and outlook

Overall, extending [15], the planned work includes experimenting with generative models to result into the best architecture to synthesize fake SWIR images. Forward, this work tries to investigate into best model which run in *real-time*, and is capable of object detection in SWIR domain for best perception in harsh weather. Some dataset experimentation for fine-tuning YOLOv8 (CNN-based) and RF-DETR (Transfer-based) include:

- Fine-tuning on SWIR images.
- Fine-tuning on mix of RGB and SWIR images.
- Fine-tuning on a mix of synthetic generated SWIR and real SWIR.
- Fine-tuning on weather specific SWIR images.
- Fine-tuning on mix of weather specific RGB and SWIR images.

Moreover, it is planned to also explore other approaches such as Few-shot learning, Knowledge Distillation etc in order to specilize the models. Here are the targeted contribution of the study:

- To develop a model capable of translating any RGB dataset to SWIR dataset.
- To identify use cases where SWIR images are suitable in autonomous driving.
- To develop a specialized object detection model capable of better perception in harsh weather using SWIR and RGB images.

This would additionally open a future domain of integrating SWIR camera in autonomous vehicles for perception in harsh weather.

References

[1] Mario Bijelic et al. “Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Bijelic_Seeing_Through_Fog_Without_Seeing_Fog_Deep_Multimodal_Sensor_Fusion_CVPR_2020_paper.pdf. 2020.

[2] CEREMA. *CEREMA Centre-Est – Site de Clermont-Ferrand*. https://www.cerema.fr/fr/cerema/directions/cerema-centre-est/site-clermont-ferrand. 2025. (Visited on 08/21/2025).

[3] Wikipedia contributors. *Transfer learning*. https://en.wikipedia.org/wiki/Transfer_learning. 2025.

[4] Vishal Gandhi and Sagar Gandhi. “Fine-Tuning Without Forgetting: Adaptation of YOLO v8 Preserves COCO Performance”. In: *arXiv preprint arXiv:2505.01016* (2025). https://arxiv.org/pdf/2505.01016.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://www.cvlibs.net/publications/Geiger2012CVPR.pdf. 2012.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: *arXiv preprint arXiv:1503.02531* (2015). https://arxiv.org/pdf/1503.02531.

[7] Edmund Optics Inc. *What is SWIR?* https://www.edmundoptics.com/knowledge-center/application-notes/imaging/what-is-swir/?srsltid=AfmBOorgYEVjuJit5FdDiL4hGEJTR4-D. 2025.

[8] Phillip Isola et al. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *CVPR* (2017).

[9] Youngwan Jin et al. “RASMD: RGB And SWIR Multispectral Driving Dataset for Robust Perception in Adverse Conditions”. In: *arXiv preprint arXiv:2405.12944* (2024). https://www.researchgate.net/publication/390671366_RASMD_RGB_And_SWIR_Multispectral_Driving_Dataset_for_Robust_Perception_in_Adverse_Conditions.

[10] Glenn Jocher, Ayush Chaurasia, Jing Qiu, et al. “YOLOv8”. In: https://github.com/ultralytics/ultralytics. Ultralytics, 2023.

[11] Dmitry Kozlov. *Awesome Knowledge Distillation*. https://github.com/dkozlov/awesome-knowledge-distillation?tab=readme-ov-file. 2025.

[12] Chih-Wei Kuan et al. “Using Yolo Technology and SWIR Images to Identify and Analyze Human and Vehicles in Smog”. In: *International Journal of Latest Engineering Research and Applications (IJLERA)* (2023). http://ijlera.com/papers/v8-i4/4.202304470.pdf.

[13] Bo Li et al. “BBDM: Image-to-image translation with Brownian bridge diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 1952–1961.

[14] Chen Ling et al. “Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey”. In: *arXiv preprint arXiv:2305.18703* (2024). https://arxiv.org/pdf/2305.18703.

[15] Rohan Mehra et al. “Would SWIR modality help for detection and segmentation in harsh weather conditions? An experimental study.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2025, pp. 2211–2219.

[16] Optics Concept. *SWIR Infrarouge à Ondes Courtes, Technologie d'Imagerie en Pleine Expansion*. Optics Concept. July 15, 2025. url: https://www.optics-concept.fr/actualites/le-swir-infrarouge-a-ondes-courtes-une-technologie-d-imagerie-en-pleine-expansion-15-07-2025-e08-5-001225.html.

[17] Taesung Park et al. “Contrastive Learning for Unpaired Image-to-Image Translation”. In: *European Conference on Computer Vision*. 2020.

[18] Archit Parnami and Minwoo Lee. “Learning from Few Examples: A Summary of Approaches to Few-Shot Learning”. In: *arXiv preprint arXiv:2203.04291* (2022). https://arxiv.org/pdf/2203.04291.

[19] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/pdf/1506.02640. 2016.

[20] Alexandre Riffard et al. *Exploitation d'un capteur proche infrarouge (SWIR) pour la perception des robots mobiles en conditions météorologiques difficiles*. https://hal.science/hal-04619118v1/document. 2024.

[21] Isaac Robinson et al. *RF-DETR: Neural Architecture Search for Real-Time Detection Transformers*. 2025. arXiv: 2511.09554 [cs.CV]. url: https://arxiv.org/abs/2511.09554.

[22] Ultralytics. *Ultralytics utils.loss Reference*. https://docs.ultralytics.com/reference/utils/loss/?utm_source=chatgpt.com. 2025.

[23] Ultralytics. *Ultralytics YOLOv8 Documentation*. https://docs.ultralytics.com. 2024. (Visited on 04/05/2025).

[24] Ultralytics / YOLOv8 Website. *What is YOLOv8*. 2025.

[25] ViewSheen Technology Co., Ltd. *What is SWIR Good for?* July 2022.

[26] Ting-Chun Wang et al. “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[27] Yu Yu, Weibin Zhang, and Yun Deng. *Frechet Inception Distance (FID) for Evaluating GANs*. Sept. 2021.

[28] Shijie Zhang, Self-Training with Noise, and Pixel-Wise Knowledge Distillation. “A Few-Shot Learning Framework for Depth Completion Based on Self-Training with Noise and Pixel-Wise Knowledge Distillation”. In: *Applied Sciences* 15.9 (2025). `https://www.mdpi.com/2076-3417/15/9/4740`, p. 4740. doi: `10.3390/app15094740`.

[29] Shizhao Zhang. “Domain Adaptive YOLO for One-Stage Cross-Domain Detection”. In: *arXiv preprint arXiv:2106.13939* (2021). `https://arxiv.org/abs/2106.13939`.

[30] Jun-Yan Zhu et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.

[31] Fuzhen Zhuang et al. “A survey of transfer learning”. In: *Journal of Big Data* 3.1 (2016). `https://link.springer.com/article/10.1186/S40537-016-0043-6`, pp. 1–40. doi: `10.1186/s40537-016-0043-6`.