Translation from RGB to SWIR Images for Adaptation of Detection and Semantic Segmentation Algorithms

Rohan Mehra[1] Alexandre Riffard[1] Mathieu Labussière[1]

[1] Université Clermont Auvergne, CNRS, Institut Pascal, F-63000 Clermont-Ferrand, France

Contact: rohan21@iiserb.ac.in, alexandre.riffard@doctorant.uca.fr, mathieu.labussiere@uca.fr

May 16, 2025

Outline

What is SWIR?

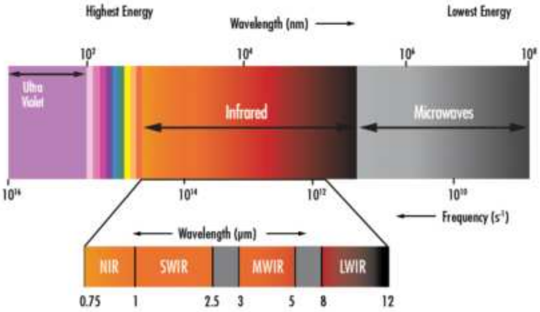


Figure: SWIR in Electromagnetic Spectrum

- Typically defined as light in the $0.9 - 1.7\mu m$ wavelength range, but can also be classified from $0.7 - 2.5\mu m$.

Why Short-Wave Infrared?

(a) Visible Images (b) SWIR Images

Figure: Comparison between visible (left) and SWIR images (right).

- Atmospheric penetration (fog, rain, etc) and vision in low light (if presence of external radiation, e.g. moon, SWIR lighting, etc.)
- Improved contrast (plastic, liquid, paint, etc.)

Cerema Data Acquisition Setup

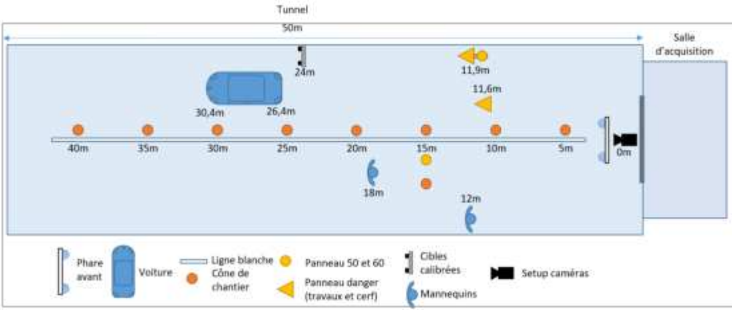


Figure: Cerema Platform setup for data acquisition [1]

- Static data acquisition was done at *Cerema, 8 Rue Bernard Palissy, Clermont-Ferrand* where rain and fog can be simulated.
- Dynamic driving data acquisition was done with Zoe.

Cameras used

Figure: L to R: SVS, Dalsa Genie & Xenics Sensor

Camera	Type	Range of Camera
Xenics Bobcat 320 [2]	InGaAs based SWIR	900 nm - 1700 nm
SVS Acuros CQD 1280 [3]	Quantum Dot based SWIR	400 nm - 1700 nm
Dalsa Genie Nano 1630 [4]	Visible Camera	380 nm - 700 nm.

Sample Images

(a) Front View

(b) Back View

(c) Side View

(d) Zoe RGB Image

(e) Cerema SWIR Image

(f) Zoe SWIR Image

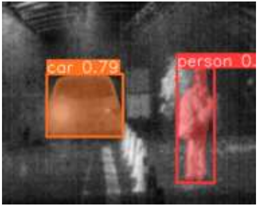
Figure: (a) to (f) Images of Cerema Platform and Captures

Previous Work: Deep Learning Semantic Seg. Models

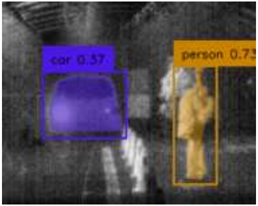
- Aim: To detect objects in a SWIR frame and segment them using RGB pre-trained models.

Algorithms	**Tasks**	**Avg Time/ Image**
Grounding DINO [5]	Detection (open set)	10 s to 30 s
SAM [6]	Segmentation (whole image)	30 s to 3 min
Grounded SAM [7]	Detect+Segment (open set)	30 s to 2 min
YOLO[v8 & v11]-seg [8]	Detect+Segment (COCO labels)	0.3 s to 0.5 s
YOLOv8-oiv [8]	Detection (OpenImages)	1 s to 3 s
MMSegmentation [9]	Detect+Segment (Cityscrape)	5 s to 10 s

Sample Annotated Images



(a) YOLOv8



(b) Grounded SAM

(c) MMSegmentation

Figure: Sample Annotated Images with Sem. Segmentation Models [Discarding MMSegmentation]

Methodology

1. Data acquisition from all the three cameras as well as fuse images through TarDAL [10].
2. Crop and register the frames (Match the timestamp).
3. Manually annotate one frame of the sequence.
4. Preprocess the frames (Thresholding at 99% confidence interval followed by normalization.)
5. Run it through all the segmentation algorithms
6. Evaluate on IoU/F-1 score:
    - YOLOv8-seg vs Grounded SAM
    - Comparison of all three modalities

Quantitative Results - Foggy Day

Sequence	Camera	Algorithm	Preprocessed	TP	FP	FN	Precision	Recall	F1 Score
51 (Foggy Day)	SVS	GroundedSAM	Yes	122	38	54	0.762	0.694	0.726
			No	116	30	60	0.794	0.659	0.720
		YOLOv8	Yes	128	36	48	0.780	0.727	0.753
			No	125	31	51	0.801	0.710	0.753
	Visible	GroundedSAM	Yes	103	1	73	0.990	0.585	0.736
			No	116	56	60	0.674	0.659	0.667
		YOLOv8	Yes	103	6	73	0.945	0.585	0.723
			No	104	3	72	0.972	0.591	0.735
	Xenics	GroundedSAM	Yes	83	1	93	0.989	0.472	0.638
			No	68	6	108	0.919	0.387	0.544
		YOLOv8	Yes	112	3	64	0.979	0.636	0.770
			No	80	0	96	1.000	0.455	0.625

Table: Quantitative Evaluation across Foggy Day Sequence (PP validation and Algorithms Evaluation)

Quantitative Results - Rainy Day

Sequence	Camera	Algorithm	Preprocessed	TP	FP	FN	Precision	Recall	F1 Score
23 et 24 (Rainy Day)	SVS	GroundedSAM	Yes	252	78	0	0.764	1.000	0.866
			No	252	85	0	0.748	1.000	0.856
		YOLOv8	Yes	252	61	0	0.805	1.000	0.892
			No	252	51	0	0.831	1.000	0.908
	Visible	GroundedSAM	Yes	248	51	4	0.829	0.984	0.900
			No	251	28	1	0.900	0.996	0.945
		YOLOv8	Yes	251	3	1	0.988	0.996	0.992
			No	251	7	1	0.973	0.996	0.984
	Xenics	GroundedSAM	Yes	191	36	61	0.841	0.758	0.797
			No	202	53	50	0.793	0.802	0.797
		YOLOv8	Yes	193	8	59	0.960	0.766	0.852
			No	154	5	98	0.969	0.612	0.749

Table: Quantitative Evaluation across Rainy Day Sequence

Quantitative Results - Clear Day

Sequence	Camera	Algorithm	Preprocessed	TP	FP	FN	Precision	Recall	F1 Score
21 (Clear Day)	SVS	GroundedSAM	Yes	52	1	0	0.981	1.000	0.990
			No	52	4	0	0.929	1.000	0.963
		YOLOv8	Yes	52	8	0	0.867	1.000	0.929
			No	52	6	0	0.897	1.000	0.945
	Visible	GroundedSAM	Yes	47	2	5	0.959	0.904	0.931
			No	47	2	5	0.959	0.904	0.931
		YOLOv8	Yes	46	0	6	1.000	0.885	0.940
			No	52	0	0	1.000	1.000	1.000
	Xenics	GroundedSAM	Yes	52	4	0	0.929	1.000	0.963
			No	52	16	0	0.765	1.000	0.867
		YOLOv8	Yes	51	10	1	0.980	0.981	0.902
			No	50	3	2	0.943	0.962	0.952

Table: Quantitative Evaluation across Clear Day Sequences

Qualitative Evaluation

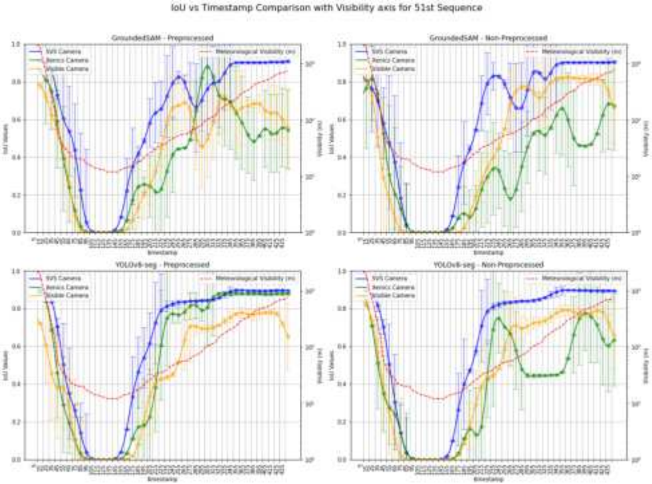


Figure: Comparison of Modalities for Foggy Day Sequence with preprocessed and non-preprocessed frames.

Qualitative Evaluation

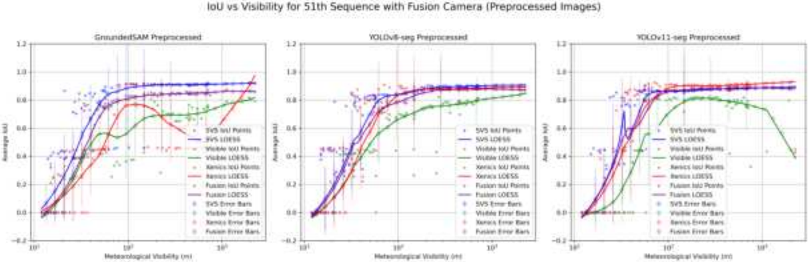


Figure: Comparison of Modalities for Foggy Day Sequence (Y axis: Visibility).

- Similarly, a comparison of modalities and algorithms had been made.
- Detailed results available at: rohmeh.github.io/docs/2024-SWIR-Project.pdf

Conclusions (May - July 2024)

1. Validated the pre-processing (Thresholding at 99% and normalization)
2. COCO Pre-trained YOLOv8-seg gives better detections compared to GroundedSAM for SWIR Images. YOLOv8 behaves more stable than YOLOv11 for SWIR Images
3. Proof of concept developed to demonstrate the effectiveness of using SWIR images with RGB pre-trained deep learning models for improved object detection and segmentation in harsh weather conditions.

Phase Two - Translation into SWIR Images

- Objective: Translating RGB images to SWIR images using GAN based models.
- Models used: CUT [11] and CycleGAN [12].
- Dataset prepared: Mixture of Zoe and Cerema acquisitions.

Dataset Information

- **Zoe Acquisitions:**
  - RGB: 6794 images
  - SWIR: 6781 images
- **Cerema Acquisitions:**
  - RGB: 262 images
  - SWIR: 402 images
- **Dataset Splits:**
  - SWIR Images: 7183 total (Unpaired)
    - Training: 5021 images (70%)
    - Validation: 1081 images (15%)
    - Testing: 1081 images (15%)
  - Visible (RGB) Images: 7056 total (Unpaired)
    - Training: 4940 images (70%)
    - Validation: 1058 images (15%)
    - Testing: 1058 images (15%)

CUT Architecture [11]

- **Key Idea:**
  - Uses contrastive learning to map unpaired images.
  - Employs a generator and a discriminator.
  - Introduces a PatchNCE loss to maximize mutual information between corresponding patches.
- **Loss Functions:**
  - **Adversarial Loss:** Ensures the generated images are indistinguishable from real images.
  - **PatchNCE Loss:** Maximizes mutual information between corresponding patches in the input and output images.
  - **Identity Loss:** Preserves color composition between input and output.

CycleGAN Architecture [12]

- **Key Idea:**
  - Learns mappings between two domains without paired data.
  - Uses cycle-consistency loss to ensure meaningful translations.
- **Loss Functions:**
  - **Adversarial Loss:** Ensures the generated images are indistinguishable from real images in each domain.
  - **Cycle-Consistency Loss:** Ensures that translating an image from domain A to B and back to A results in the original image.
  - **Identity Loss:** Preserves color composition between input and output.

Implementation Details: CUT Model

- Model: CUT (Contrastive Unpaired Translation)
- Training epochs: 56
- Batch size: 6
- Generator: `resnet_9blocks`
- Discriminator: `patch`
- Learning rate: 0.0002, with beta1 = 0.5, beta2 = 0.999
- Image size: 256x256

Implementation Details: CycleGAN Model

- Model: CycleGAN
- Training epochs: 16
- Batch size: 7
- Generator: `resnet_9blocks`
- Discriminator: `patch`
- Learning rate: 0.0002, with beta1 = 0.5
- Image size: 256x256

Environmental Setup

- **GPU:** NVIDIA RTX A4500 (20GB VRAM)
- **CUDA Version:** 12.8
- **Driver Version:** 570.124.06
- **OS:** Ubuntu 22.04
- **Software:**
  - PyTorch with CUDA 12.5
  - torchvision
  - Tensorboard, Visdom

Visual Results

(a) Original Visible Image

(b) CUT Generated SWIR

(c) CycleGAN Generated SWIR

Figure: Comparison of RGB input and generated SWIR images using CUT and CycleGAN.

Evaluation Metrics: FID

Fréchet Inception Distance (FID) [13] - Visible pre-trained

- Measures similarity between real and generated images.
- Computed using the mean and covariance of InceptionV3 features.
- Lower FID indicates better performance.

Typical FID Scores:

- **Excellent:** < 10 (Very high-quality)
- **Good:** 10 - 30 (Visually high quality with minimal artifacts)
- **Fair:** 30 - 70 (Some noticeable artifacts, but acceptable for many tasks)
- **Poor:** 70 - 150 (Significant quality issues, lacks realism)
- **Very Poor:** > 150 (Severe artifacts, unrealistic outputs)

Evaluation Metrics: Inception Score

Inception Score (IS) - Visible pre-trained

- Measures diversity and realism of generated images.
- Computed using KL divergence of InceptionV3 class probabilities.
- Higher IS indicates better performance.

Typical IS Scores:

- **Excellent:** > 8.0 (High diversity and realism, close to natural images)
- **Good:** 5.0 - 8.0 (Visually high quality with diverse samples)
- **Fair:** 2.5 - 5.0 (Some diversity, but limited realism)
- **Poor:** 1.5 - 2.5 (Low diversity, repetitive outputs)
- **Very Poor:** < 1.5 (Severe mode collapse, unrealistic images)

Results

Model	FID ↓	IS ↑
CUT	**48.0503**	**3.0735 ± 0.1804**
CycleGAN	**42.3482**	**2.9226 ± 0.0941**

Table: Comparison of CUT and CycleGAN on RGB-to-SWIR Translation.

Planned Work Ahead

- Train and compare GAN and diffusion based models:
  1. Pix2Pix [GAN based paired translation]
  2. CycleGAN [GAN based unpaired translation]
  3. CUT [GAN based unpaired translation]
  4. BBDM [Diffusion-based]
- Leverage the RASMD dataset [14] recently released.
- Translate a full-fledged RGB dataset to SWIR.
- Re-train the Detection and Semantic Segmentation Algorithms on the dataset.
- Compare the performance with results of visible pre-trained algorithms.

Questions and Reviews

Any Questions?
Reviews Please!

Merci! Thank You! धन्यवाद شكريه

References I

[1] S. Liandrat, P. Duthon, F. Bernardin, B. Daoued, and J.-L. Bicard, "A review of cerema pavin fog rain platform : from past and back to the future," 2022.

[2] Exosens, "BOBCAT/BOBCAT+," accessed: 2024-11-06. [Online]. Available: https://www.exosens.com/products/bobcat

[3] SWIR Vision Systems, "Acuros SWIR Cameras — SWIR Vision Systems," accessed: 2024-11-06. [Online]. Available: https://www.swirvisionsystems.com/acuros-swir-camera/

[4] https://www.teledynevisionsolutions.com/products/genie-nano-1-gige/?model=G3-GC11-C1630vertical=tvs-dalsa-oemsegment=tvs

[5] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L. (2024). *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. arXiv:2303.05499. Retrieved from `https://arxiv.org/abs/2303.05499`

[6] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R. (2023). *Segment Anything*. arXiv:2304.02643. Retrieved from `https://arxiv.org/abs/2304.02643`

[7] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," 2024.

[8] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online].

[9] https://mmsegmentation.readthedocs.io/en/main/

[10] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5792–5801.

References II

[11] Park, T., Efros, A. A., Zhang, R., Zhu, J.-Y. (2020). Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*.

[12] Zhu, J.-Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*.

[13] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv preprint arXiv:1706.08500*. Available at: `https://arxiv.org/abs/1706.08500`

[14] Jin, Y., Kovac, M., Nalcakan, Y., Ju, H., Song, H., Yeo, S., Kim, S. (2025). RASMD: RGB And SWIR Multispectral Driving Dataset for Robust Perception in Adverse Conditions. *arXiv preprint arXiv:2504.07603*. Available at: `https://arxiv.org/abs/2504.07603`

[15] Pavlović, M., Banjac, Z., Kovačević, B. (2023). Object Tracking in SWIR Imaging Based on Both Correlation and Robust Kalman Filters. *IEEE Access*, 11, 63834–63851. 10.1109/ACCESS.2023.3288694