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

Rohan Mehra Alexandre Riffard Mathieu Labussière

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

Contact: rohan.mehra@uca.fr, alexandre.riffard@doctorant.uca.fr, mathieu.labussiere@uca.fr

July 11, 2025

Outline

What is SWIR?

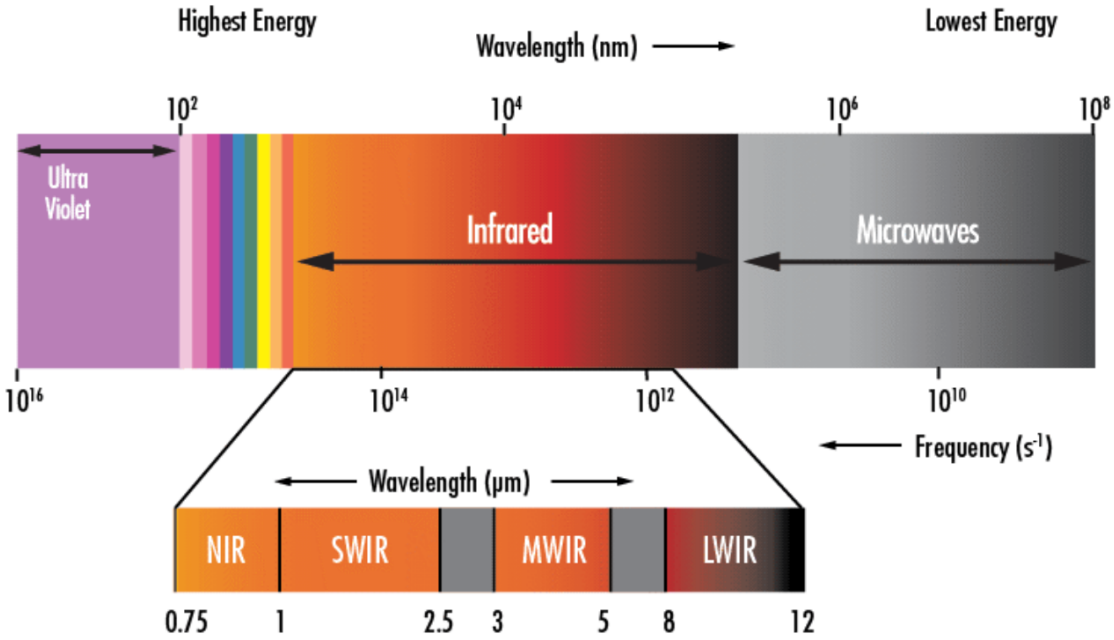
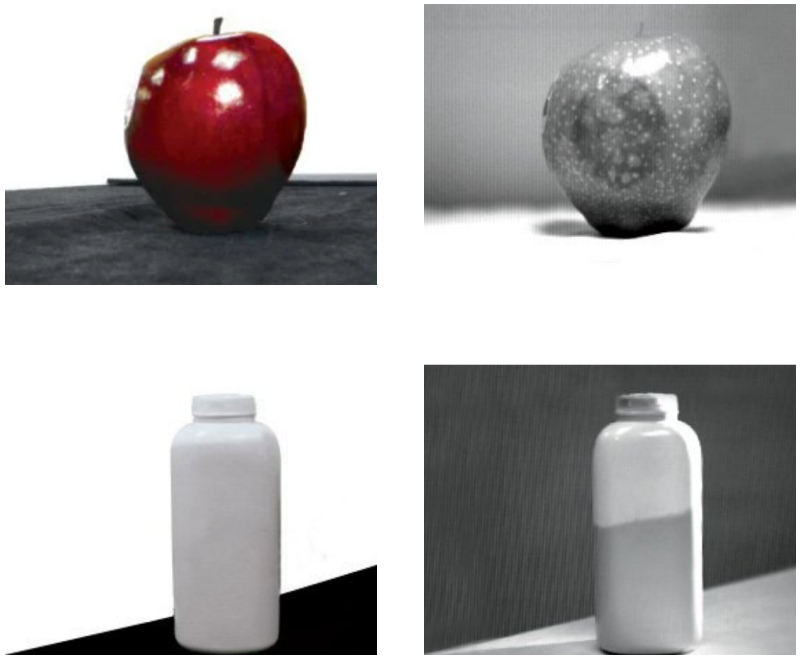


Figure: SWIR in Electromagnetic Spectrum

- Typically defined as light in the 0.9 – 1.7μm wavelength range, but can also be classified from 0.7 – 2.5μm.

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.)

Previous work

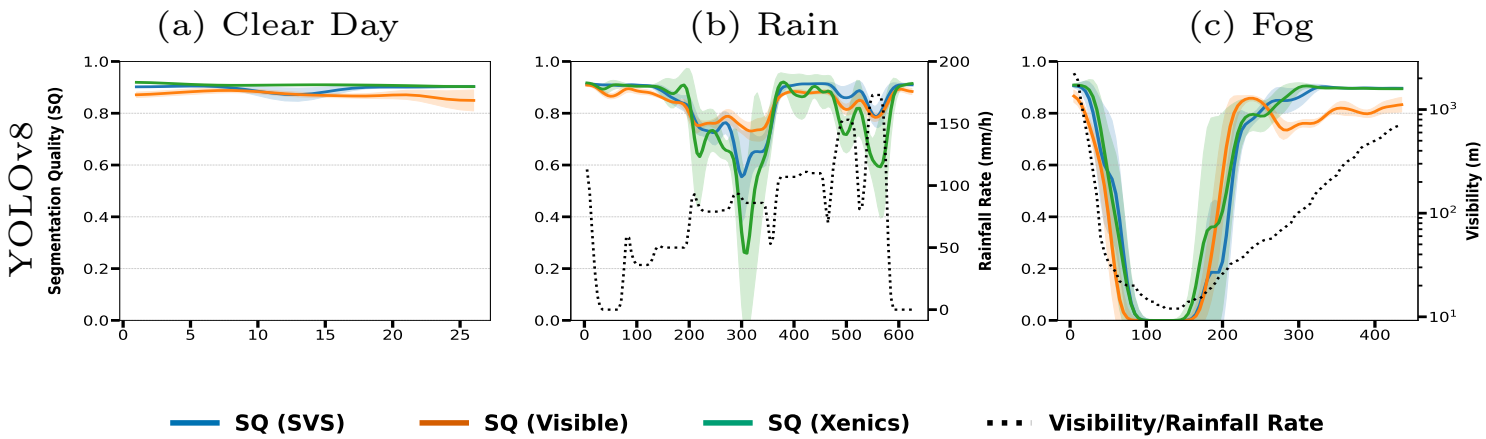


Figure: Segmentation Quality (SQ) as a function of time (with visibility/rainfall rate axis) comparing sensing modalities under different weather.

Previously: Applied RGB pre-trained object detection and segmentation models to SWIR driving images acquired in controlled weather conditions and compared results to corresponding visible-spectrum images.

Work accepted at ICCV Workshop 2025!

More at:

- rohmeh.github.io/docs/2024-SWIR-Project.pdf
- rohmeh.github.io/docs/2025-SWIR-IP-1-1.pdf

Objective and Procedure

Aim: Train models to translate RGB images into SWIR images in order to create a fake SWIR dataset fine-tune object detection models.

Procedure:

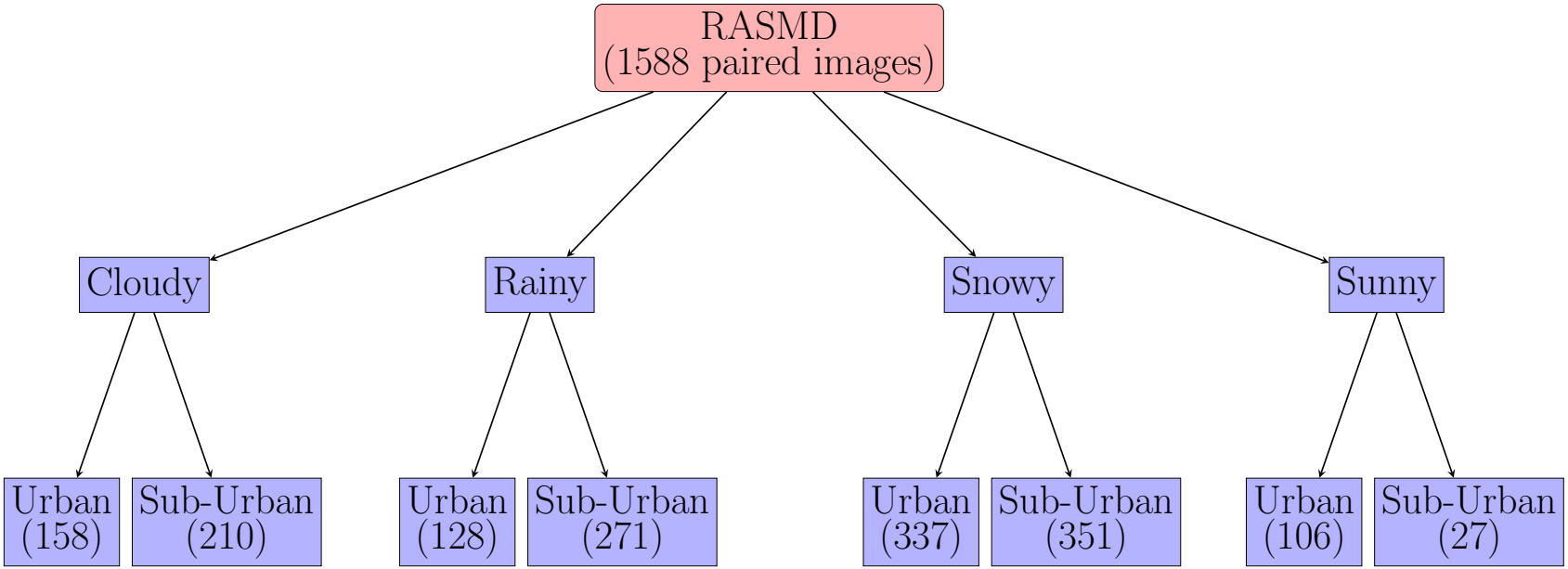
- Create dataset as part of RASMD [6] and split it into train/test/val.
- Train various models on the dataset.
- Evaluate on various traditional/learned-based metrics.

Dataset Used

RASMD (RGB And SWIR Multispectral Driving) [6]

- Comprises of **100,000 synchronized and pixel-aligned RGB-SWIR image pairs** collected across diverse urban and suburban locations, lighting, and weather conditions (sunny, cloudy, rainy, snowy).
- Data was acquired using a vehicle-mounted platform equipped with RGB [9] and SWIR cameras [8], covering **163.3 km over 8.5 hours** of driving.
- The dataset includes annotations for object detection (6 traffic-related classes).

Training Dataset Created



- **Dataset Split: 1588 total (Paired)**
  - Training: 1111 images (70%)
  - Validation: 238 images (15%)
  - Testing: 239 images (15%)

Sample Images

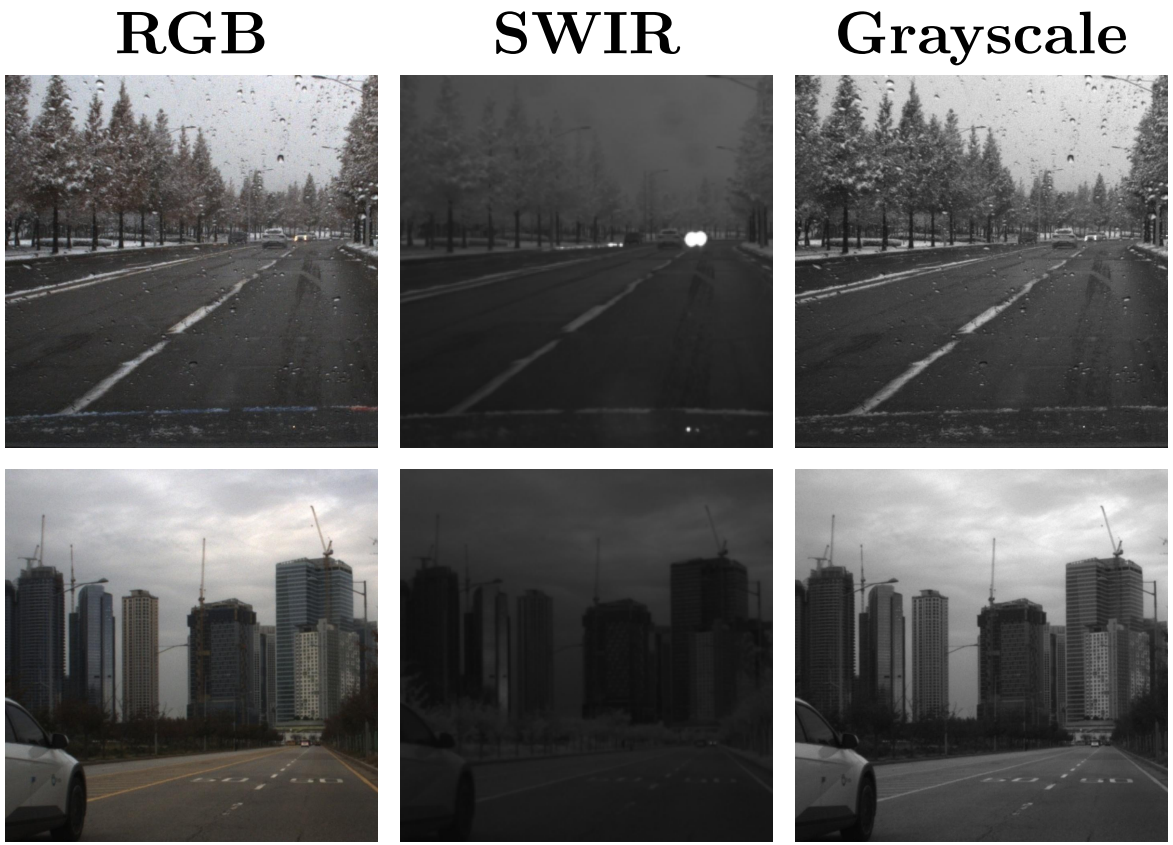


Figure: Images from RASMD [6] dataset and corresponding grayscale converted images for comparison.

Models used for translation

Model	Type	Paired/Unpaired
CUT [13]	GAN	Unpaired
FastCUT [13]	GAN	Unpaired
CycleGAN [12]	GAN	Unpaired
Pix2Pix [10]	GAN	Paired
Pix2PixHD [11]	GAN	Paired
BBDM [14]	Diffusion	Paired

- 6 Model were used - 1 diffusion based paired translation, 2 GAN based paired translation, and 3 GAN based unpaired translation models.

Pix2Pix: Conditional GANs for Paired Translation [10]

Idea:

- Designed for **paired image-to-image translation**.
- Learns a mapping from input image to output using conditional GANs.
- Uses a U-Net generator for spatial correspondence.

Loss Functions:

- **Adversarial loss:** makes output indistinguishable from real target images.
- **L1 loss:** encourages the output to be close to ground-truth in pixel space.

Pix2PixHD: High-Resolution Paired Translation [11]

Idea:

- Extension of Pix2Pix for **high-resolution** and **photo-realistic** image synthesis.
- Incorporates multi-scale discriminators and a coarse-to-fine generator.

Loss Functions:

- Adversarial loss: at multiple scales for better detail.
- Feature matching loss: stabilizes GAN training by matching intermediate discriminator features.
- Perceptual loss: improves visual quality and realism.

CycleGAN: Cycle-Consistent Adversarial Networks [12]

Idea:

- Enables **unpaired, bidirectional** image-to-image translation.
- Learns mappings between two domains using the concept of *cycle-consistency* where it ensures that translating from one domain to another and back yields the original image.

Loss Functions:

- **Adversarial loss:** ensures generated images resemble the target domain.
- **Cycle-consistency loss:** enforces that $x \to G(x) \to F(G(x)) \approx x$ and $y \to F(y) \to G(F(y)) \approx y$.
- Identity loss: preserves color composition and stabilizes training.

CUT: Contrastive Unpaired Translation [13]

Idea:

- Enables **unpaired, one-sided** image-to-image translation.
- Preserves input content by maximizing mutual information between **local patches** in input and output.
- Uses a **shared latent space** to align features from corresponding patches.

Loss Functions:

- **Adversarial loss:** ensures generated images resemble the target domain.
- **PatchNCE loss:** contrastive loss that aligns corresponding patches in input and output using negative sampling.
- **Identity loss:** regularizes training by discouraging unnecessary changes to domain B inputs.

Architecture:

- Encoder-decoder generator with multi-layer feature extraction.
- No backward generator or cycle consistency.

FastCUT [13]

Idea:

- A **lighter and faster variant** of CUT.
- Trades output quality for significantly reduced computational cost.

Loss Functions:

- Adversarial loss: ensures generated images resemble the target domain.
- PatchNCE loss: contrastive loss between input and output patches.
- *No identity loss*

Comparison to CUT:

- Removes identity loss and adjusts regularization ($\lambda_X = 10$, $\lambda_Y = 0$).

BBDM: Brownian Bridge Diffusion Model [14]

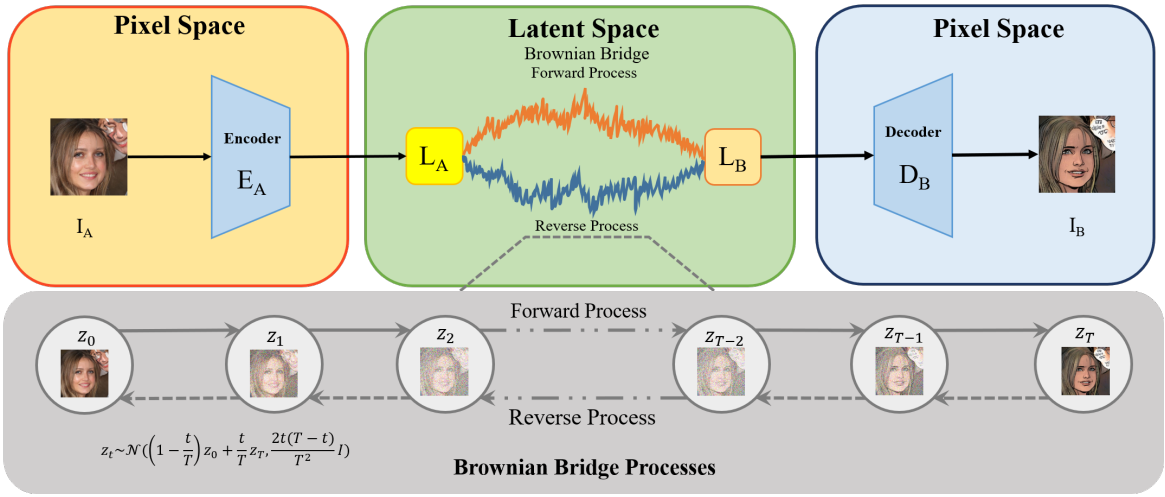


Figure: BBDM Architecture

Idea:

- Uses **diffusion models** for paired image-to-image translation.
- Models the translation as a conditional diffusion process guided by a Brownian bridge.

Loss Functions:

- Diffusion loss: score-matching loss to denoise at each step.
- Conditioning on the input ensures the diffusion path leads to the desired output.

Evaluation Metrics

Paired Data Metrics:

- **PSNR** ↑: Pixel-level fidelity (logarithmic, dB scale)
  - $\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right)$
- **SSIM** ↑: Structural similarity (0 to 1)
  - Combines luminance, contrast, and structure comparisons
- **RMSE** ↓: Root Mean Square Error
  - $\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$

Learned Metrics:

- **FID** ↓: Fréchet Inception Distance
  - Compares feature distributions in Inception-v3 space
- **LPIPS** ↓: Learned Perceptual Image Patch Similarity
  - Uses deep network features to match human perception

Implementation Details

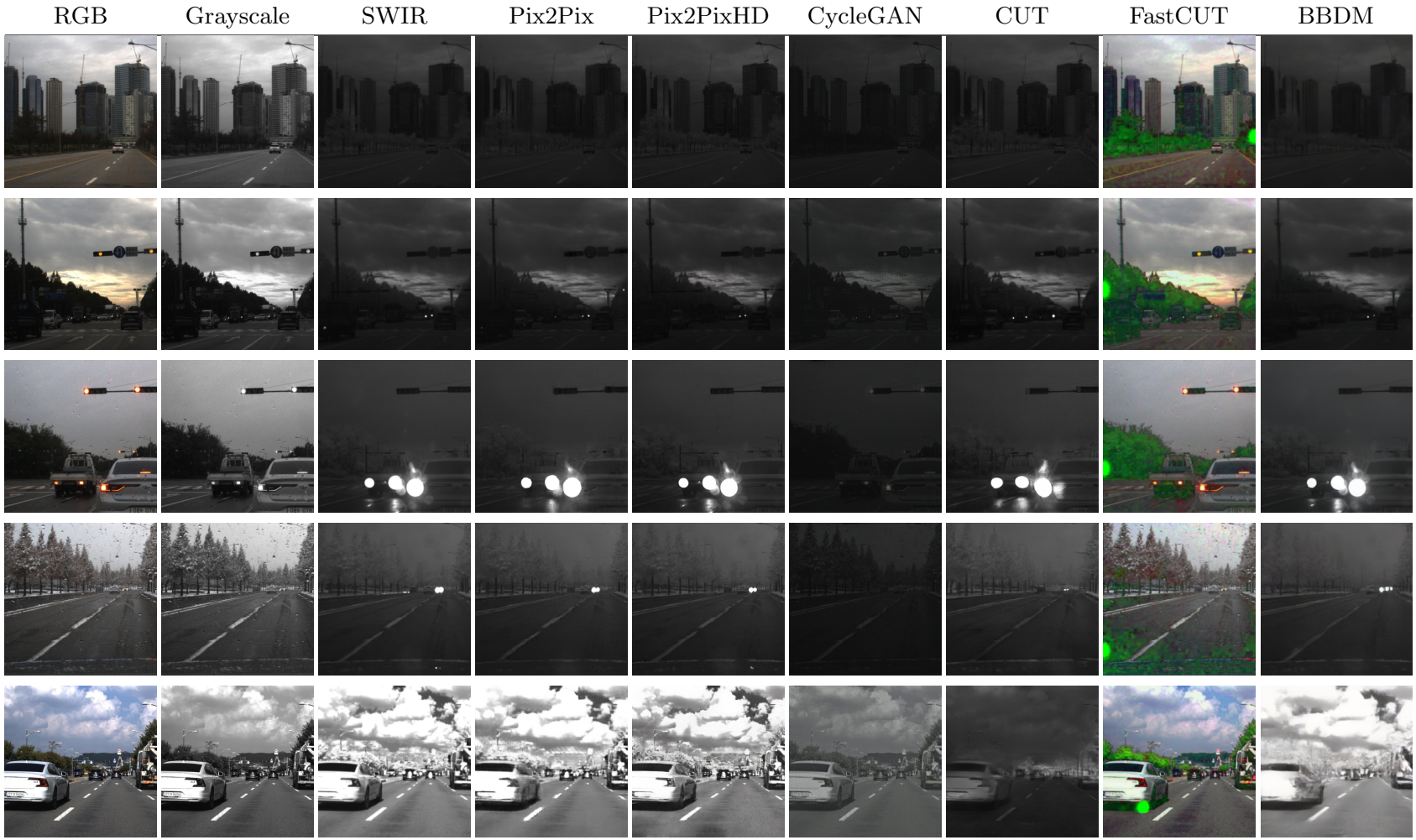
Hardware:

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

Software:

- OS: Ubuntu 22.04 LTS
- PyTorch (CUDA 12.5)
- torchvision

Visual Results



Quantitative Results

Method	**Type**	**PSNR ↑**	**SSIM ↑**	**RMSE ↓**	**FID ↓**	**LPIPS ↓**
Pix2Pix [10]	GAN	<u>33.2341</u>	<u>0.9190</u>	<u>7.3704</u>	67.4492	<u>0.0679</u>
Pix2PixHD [11]	GAN	**34.4085**	**0.9345**	**6.6343**	**38.6485**	**0.0413**
CycleGAN [12]	GAN	21.0338	0.7457	28.7373	94.8574	0.1803
CUT [13]	GAN	26.3492	0.8049	26.3492	<u>56.3365</u>	0.1348
FastCUT [13]	GAN	10.4322	0.3802	79.5787	157.0999	0.4743
BBDM [14]	Diffusion	32.1020	0.9040	8.2781	66.1329	0.0804

Table: 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.

Conclusion

- Pix2PixHD gives the best performance overall, where Pix2Pix and BBDM also give comparative scores. All of these are paired image translation methods.
- CycleGAN and CUT achieve similar scores, lower compared to unpaired translation models.
- FastCUT struggles to adapt to the SWIR domain, and produces identifiable images.

Relative Order:

Pix2Pix > Pix2Pix > BBDM >> CUT > CycleGAN >> FastCUT

Limitations and Future Works

Limitations:

- The dataset used for training has driving sequence-based images and may overfit the models.
- The images are not homographically aligned well as provided in [6] which may affect performce as well as training of models.

Future Works:

- Translate a full-fledged RGB dataset to SWIR.
- Train metric to evaluate quality of SWIR images.
- Re-train the Detection and Semantic Segmentation Algorithms on the dataset.
- Compare the performance with the 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] 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.

[3] 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)*.

[4] 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)*.

[5] 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`

[6] 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`

[7] 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

[8] CREVIS Co., Ltd. HG-A130SW SWIR Camera Product Page. Available online: `https://www.crevis.co.kr/Product/productView?idx=808` (accessed July 9, 2025).

[9] Edmund Optics. FLIR Grasshopper3 GS3-U3-32S4C-C USB 3.0 Color Camera Product Page. Available online: `https://www.edmundoptics.com/p/gs3-u3-32s4c-c-1-18-inch-grasshopper-usb-30-color-camera/33122/` (accessed July 9, 2025).

References II

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-Image Translation with Conditional Adversarial Networks*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, *High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[13] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, *Contrastive Learning for Unpaired Image-to-Image Translation*, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] B. Li, K. Xue, B. Liu, and Y.-K. Lai, *BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1952–1961, 2023.