Visual Perception from Thermal Image
: Dataset, Benchmark, and Challenges

Dr. Ukcheol Shin

Carnegie Mellon University
Robotics Institute

Dr. Ukcheol Shin

M.S. : Noise-aware Camera Exposure Control for Robust Robot Vision

EE, Korea Advanced Institute of Science and Technology (KAIST)
- Robotics and Computer Vision (RCV) Lab
- Advisor: Prof. In So Kweon

2017 - 2019

2019 - 2023.Aug

Ph.D. : Self-supervised 3D Geometric Perception in Adverse Real-world Environment

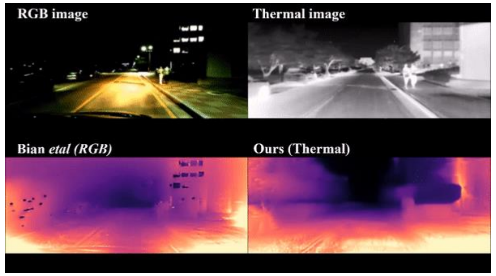
EE, Korea Advanced Institute of Science and Technology (KAIST)
- Robotics and Computer Vision (RCV) Lab
- Advisor: Prof. In So Kweon

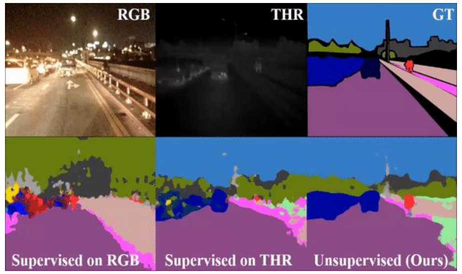
Postdoctoral Associate

RI, Carnegie Mellon University (CMU)
- roBot Intelligence Group (BIG)
- Advisor: Prof. Jean Oh

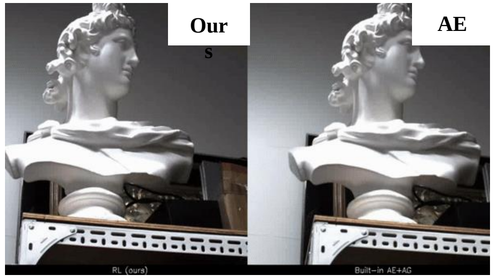
2023.Aug -

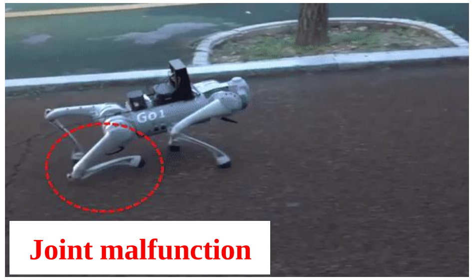
Research goal : Robust physical AI in the wild





Robust visual perception in challenging conditions





Robust sensor & actuator control

Today's agenda

Intro.	**Visual Perception in Robotics** : Limitation of visual perception from RGB camera/LiDAR
Part 1	**Spatial Perception from Thermal Image : Dataset and Benchmark** : Thermal camera is a potential rescue for robust spatial perception
Part 2	**Visual perception from Thermal Image: Challenges** : What is next?

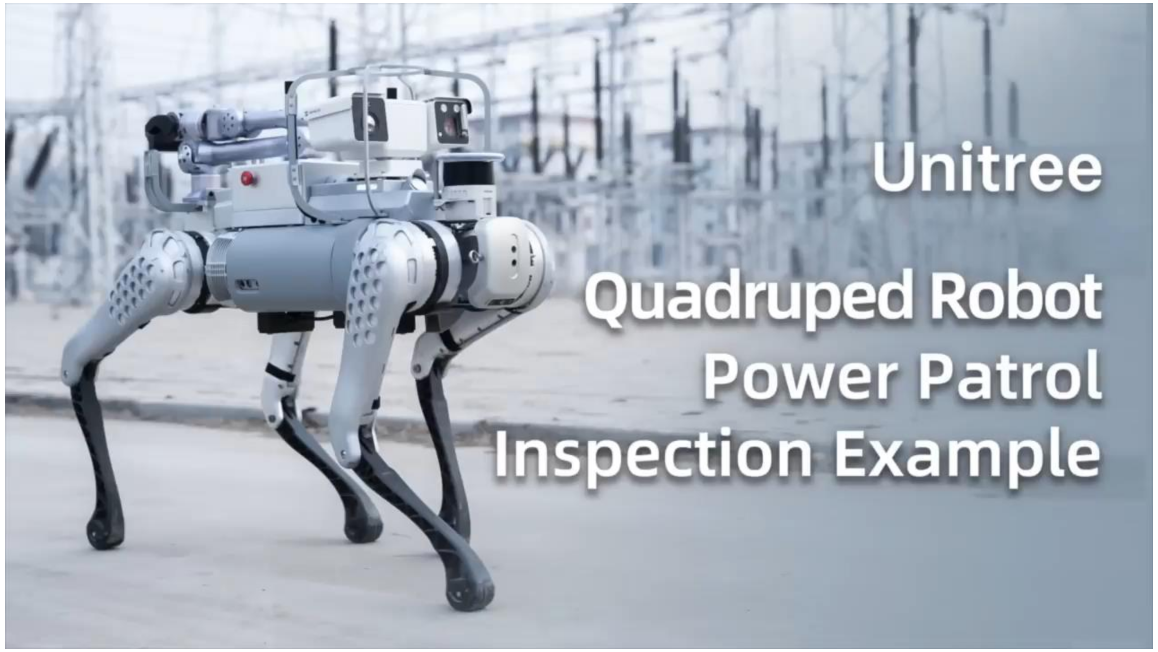
Intro.

Visual perception in Robotics

: Limitation of visual perception from RGB/LiDAR

Where are we now?

Autonomous vehicle



Quadruped robot

Credit: Tesla AI day 2022, Unitree

Where are we now?

Decision making layer

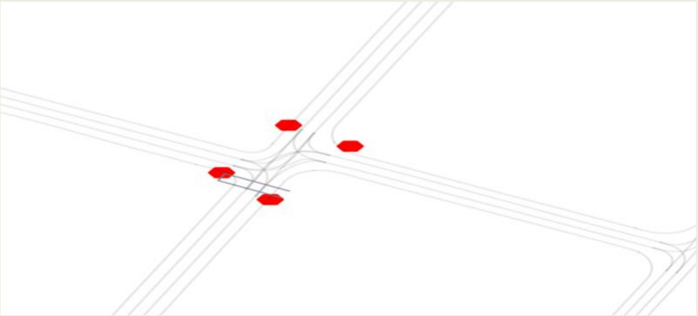
: Motion, trajectory, task planning
collision avoidance

Perception layer

Spatial perception
: Depth, occupancy, localization, mapping, tracking
Semantic perception
: Object detection, panoptic segmentation, scene graph, context reasoning

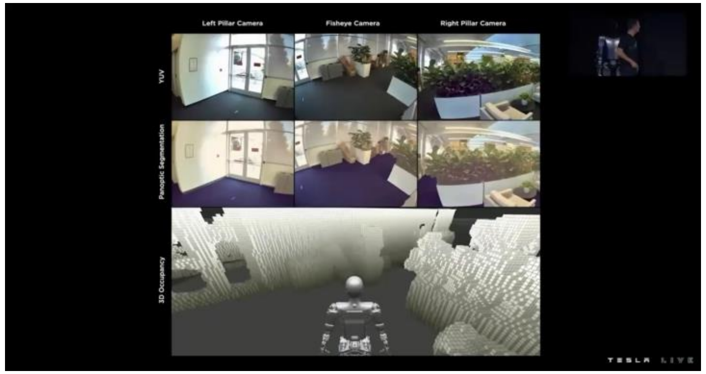
Real-time control layer

: Motor/sensor control,
Model-predictive control, RTOS

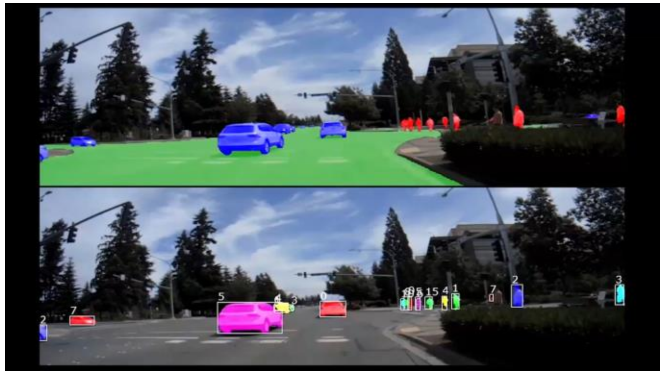


Multi-agent path planning

Visual-language navigation



Spatial perception



Semantic perception

Sensor control

Actuator control

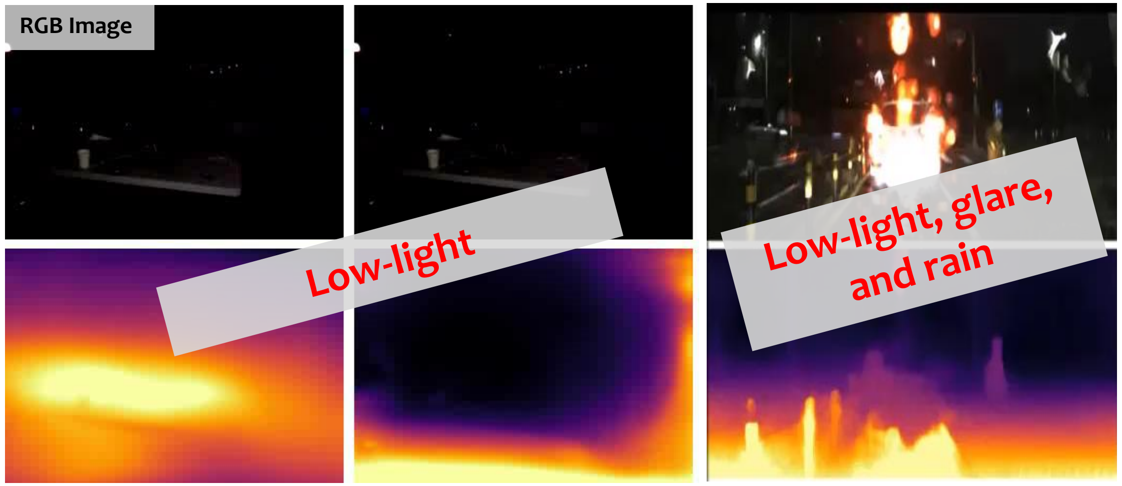
Research question

Q. Can we make AI have robust visual perception capability under challenging and hostile environments?

Limitation: visual perception from RGB camera

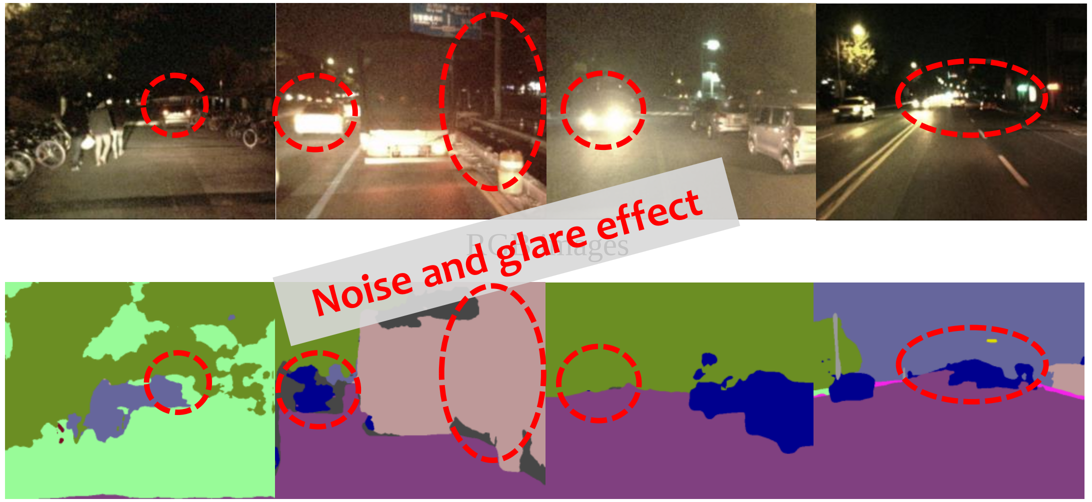
Degeneration by external factors (i.e., light & weather condition)

1. Monocular depth estimation (supervised/self-supervised)



MiDaS [1] (sup) SC-depth [2] (self-sup)

2. Semantic Segmentation (supervised)



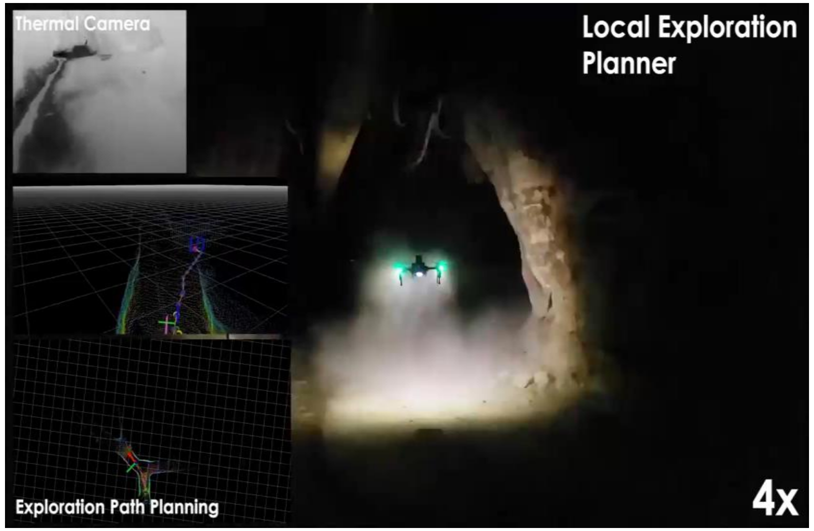
Semantic segmentation (HRNet [3])

3. RGB-Lidar depth completion (NLSPN, supervised) [4]

[1] Ranftl, René, et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer." T-PAMI 2020
[2] Bian, Jia-Wang, et al. "Unsupervised scale-consistent depth learning from video." *IJCV 2021*
[3] Wang, Jingdong, et al. "Deep high-resolution representation learning for visual recognition." *T-PAMI 2020*.
[4] Park, Jinsun, et al. "Non-local spatial propagation network for depth completion." ECCV *2020*

Limitation: visual perception from RGB camera

Q. Can RGB sensor handles such challenging conditions?



- ✓ Blinking lights
- ✓ Heavy dust

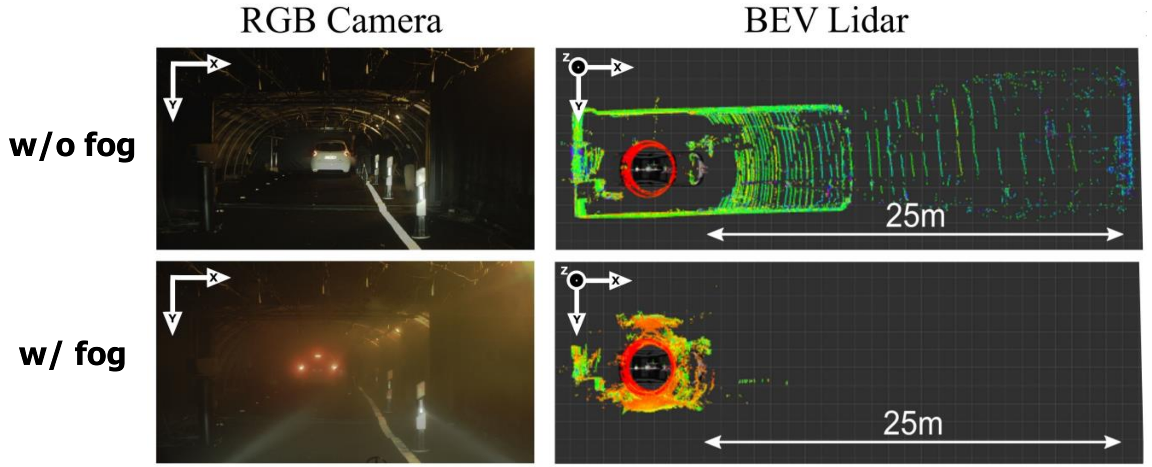
- ✓ Heavy rain
- ✓ Occlusion & blur & glare

- ✓ Heavy smoke
- ✓ Fire

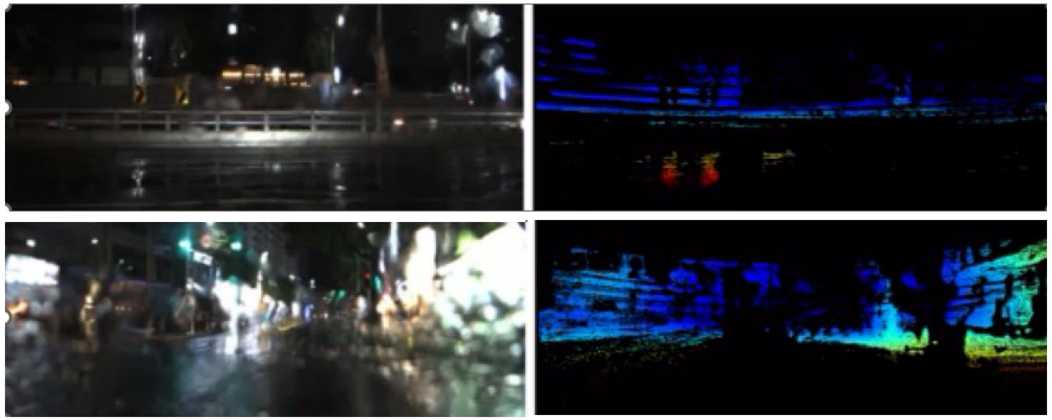
➔ RGB sensor can cause risky and unreliable predictions in adverse environments.

[Dark mine] CERBERUS, winner of DARPA Subterranean challenge **[Rainy road]** https://www.youtube.com/watch?v=U4qkaMSJOds&t=169s,
[Smoky fire] https://www.youtube.com/watch?v=P8zU1MjZSnE&t=178s

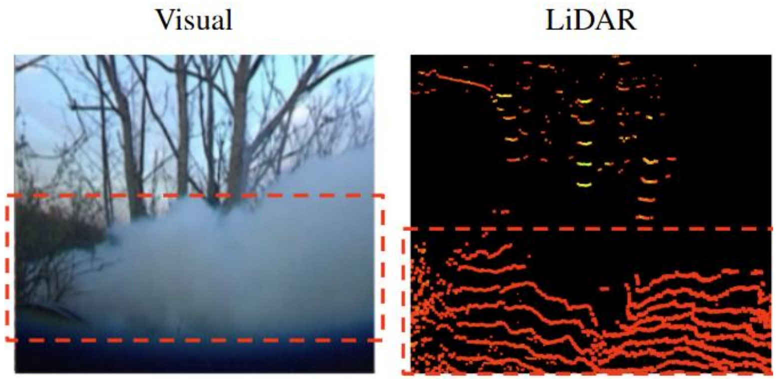
Limitation: visual perception from LiDAR

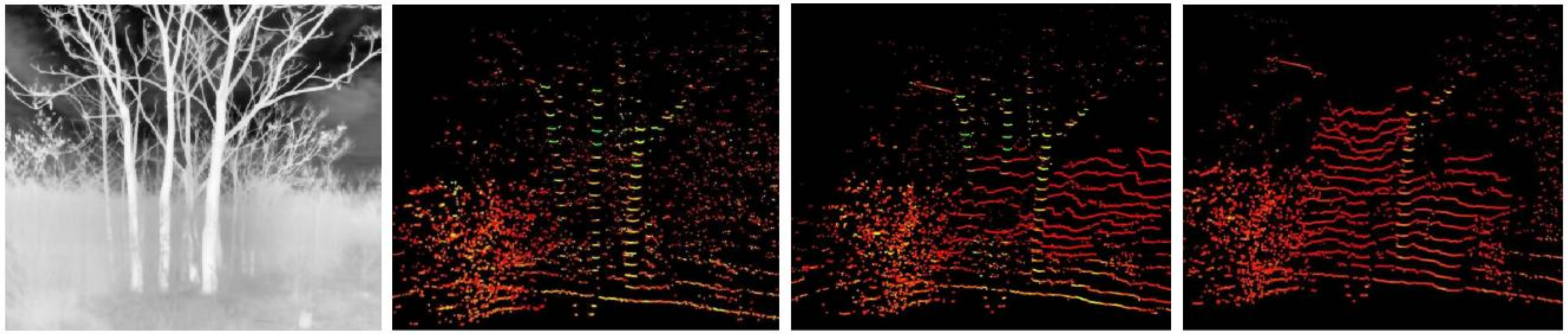


LiDAR in the fog



LiDAR in the rain





(a) Thermal image before smoke (b) No smoke (c) Light smoke (d) Heavy smoke

LiDAR in the smoke

[Fog] Bijelic, Mario, et al. "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather." *CVPR 2020*
[Rain] Ukcheol Shin, et al, "Deep Depth Estimation from Thermal Image", CVPR 2023
[Smoke] Devansh Dhrafan, et al, "FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments", Under-review (U. Shin: Co-author)

Limitation: visual perception from RGB camera/LiDAR

Q. What is the universal and robust sensor
for various vision applications and environments?

Comprehensive sensor comparison for visual perception

Ultra-violet spectrum
- UV camera

Visible spectrum
- RGB camera
- Event camera
- Spike camera
- Polarized camera
- Dual-pixel camera
- Light-field camera

Near-infrared spectrum
- NIR camera
- Gated camera
- Active stereo camera (structured light or ToF)
- Spinning LiDAR
- Solid-state LiDAR

Short~Long wave IR spectrum
- SWIR camera
- MWIR camera
- LWIR camera (Thermal camera)

mmWave spectrum
- mmWave RADAR

Audio frequency
- Sound detector
- SONAR

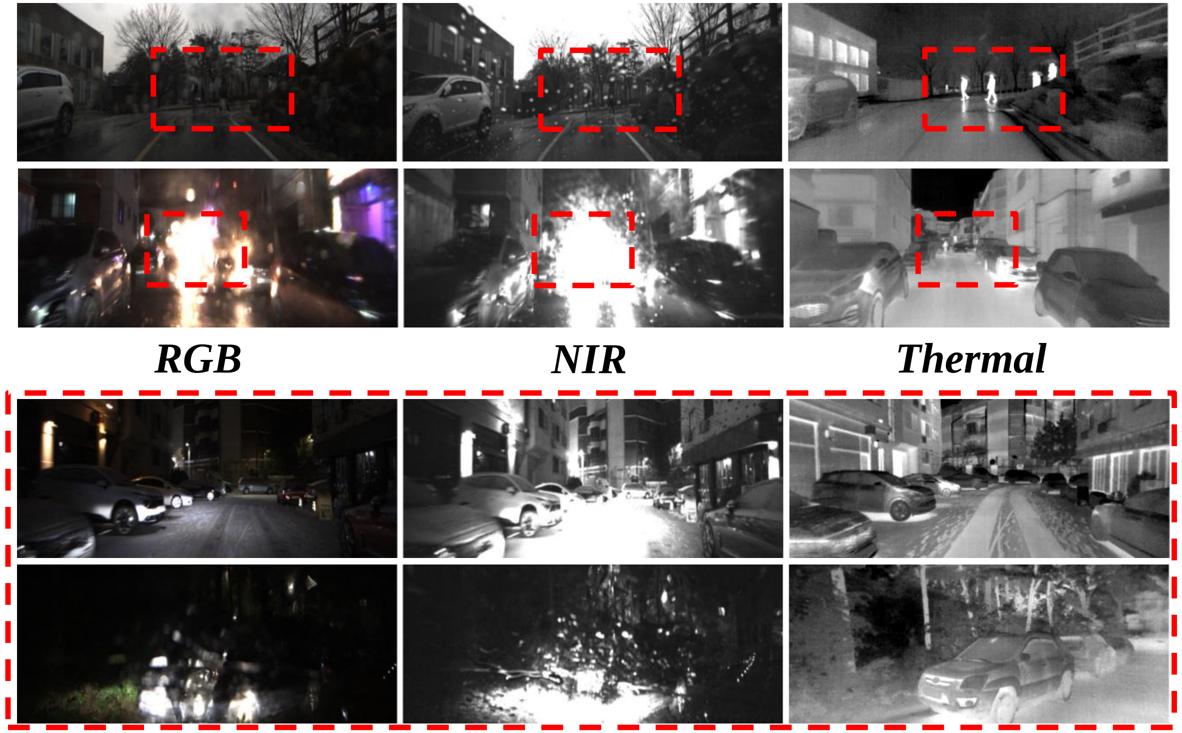
Ultra sound
- Ultrasonic sensor

Passive sensor

Active sensor

Thermal camera in challenging conditions

Thermal vision provides robustness in various challenging conditions

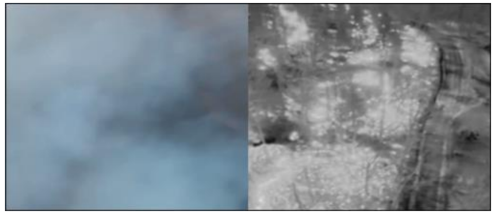


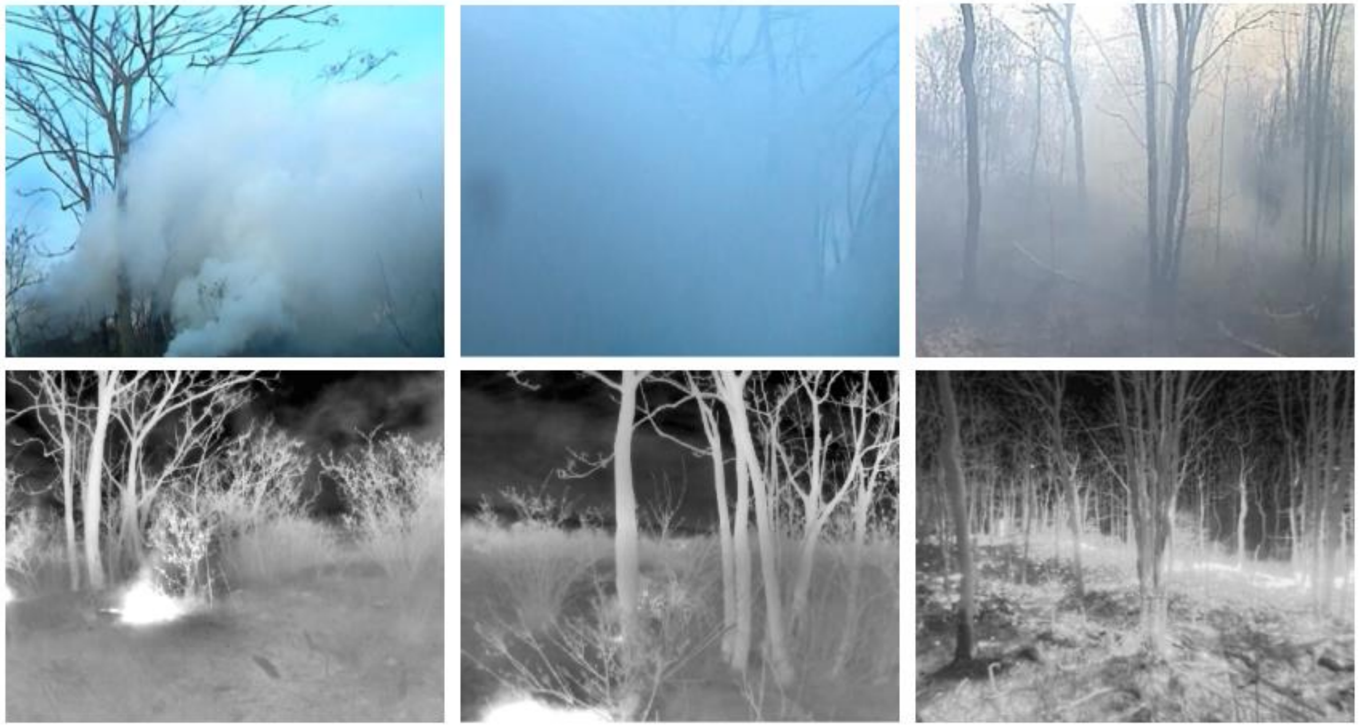
Clear visibility* against *low-light, glare, snowy, rainy, foggy, smoky conditions

[Video] FLIR Night Vision
Ukcheol Shin, et al, "Deep Depth Estimation from Thermal Image", CVPR 2023

Thermal camera in challenging conditions

Thermal vision provides robustness in various challenging conditions





Clear visibility* against *low-light, glare, snowy, rainy, foggy, smoky conditions

Devansh Dhrafan, et al, “FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments”, Under-review (**U. Shin: Co-author**)

Part 1.

Spatial Perception from Thermal Image : Dataset and Benchmark

- **Thermal camera is a potential rescue for robust spatial perception**

- [Dataset] Deep Depth Estimation from Thermal Image, CVPR 2023
- [Dataset] FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments, Under-review
- [Benchmark] Deep Depth Estimation from X: Benchmark, analysis, and challenges, TBA

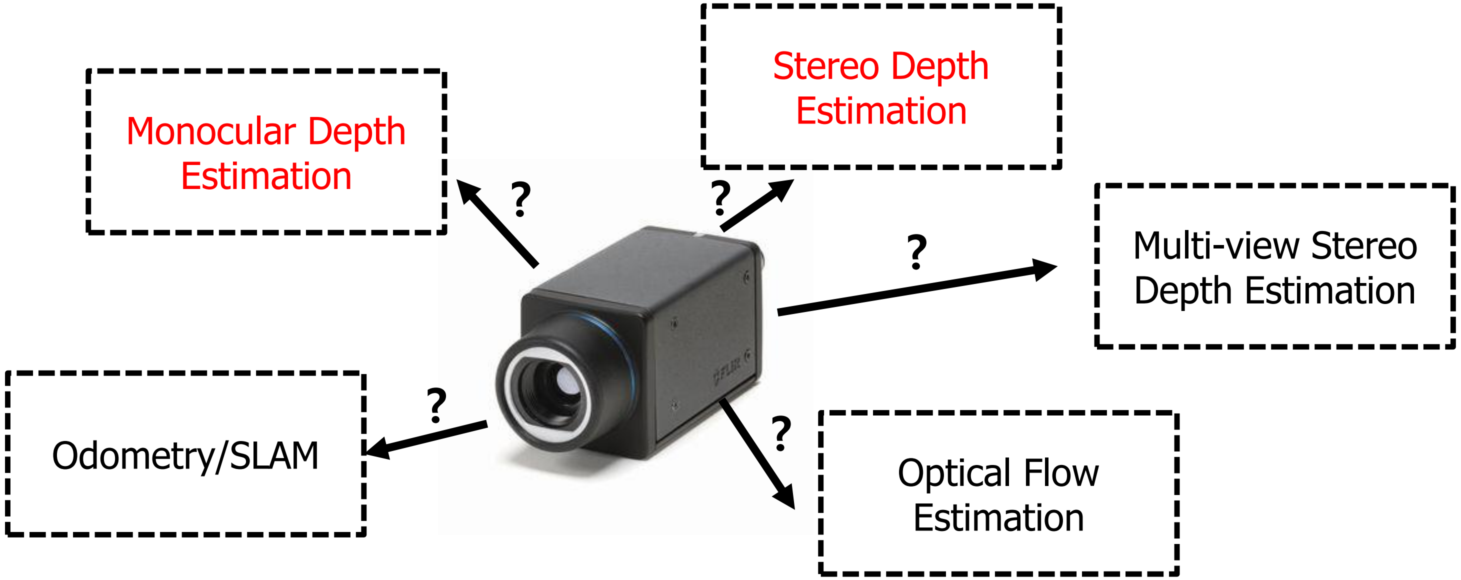
Key Challenges for spatial perception from thermal camera

[Dataset] No large-scale and open-sourced thermal 3D dataset

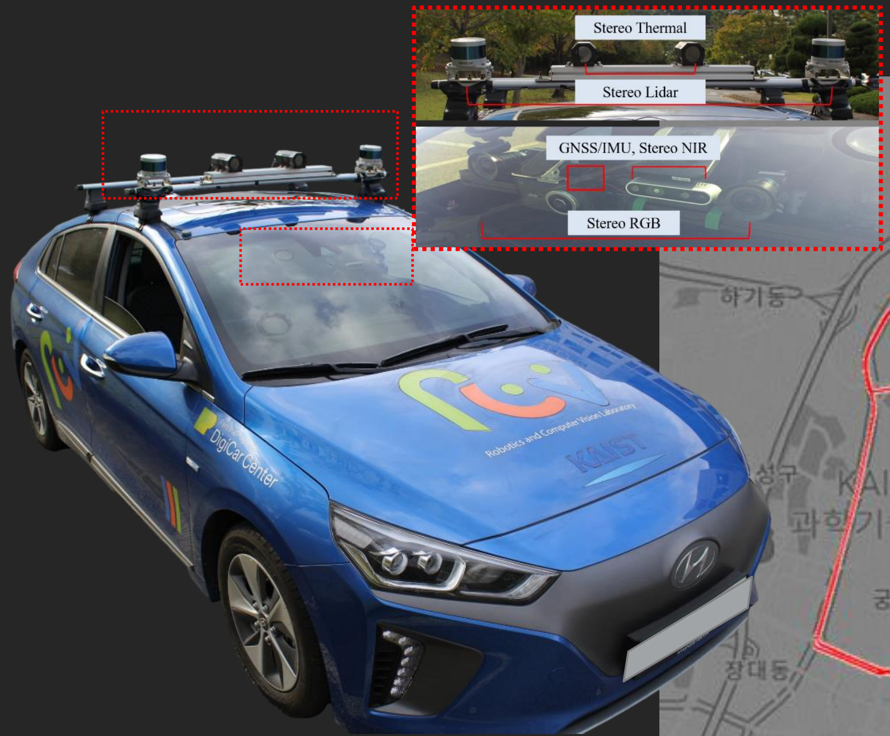
- Diverse weather, lighting, and locational conditions
- Accurate time synchronization and multi-sensor calibration

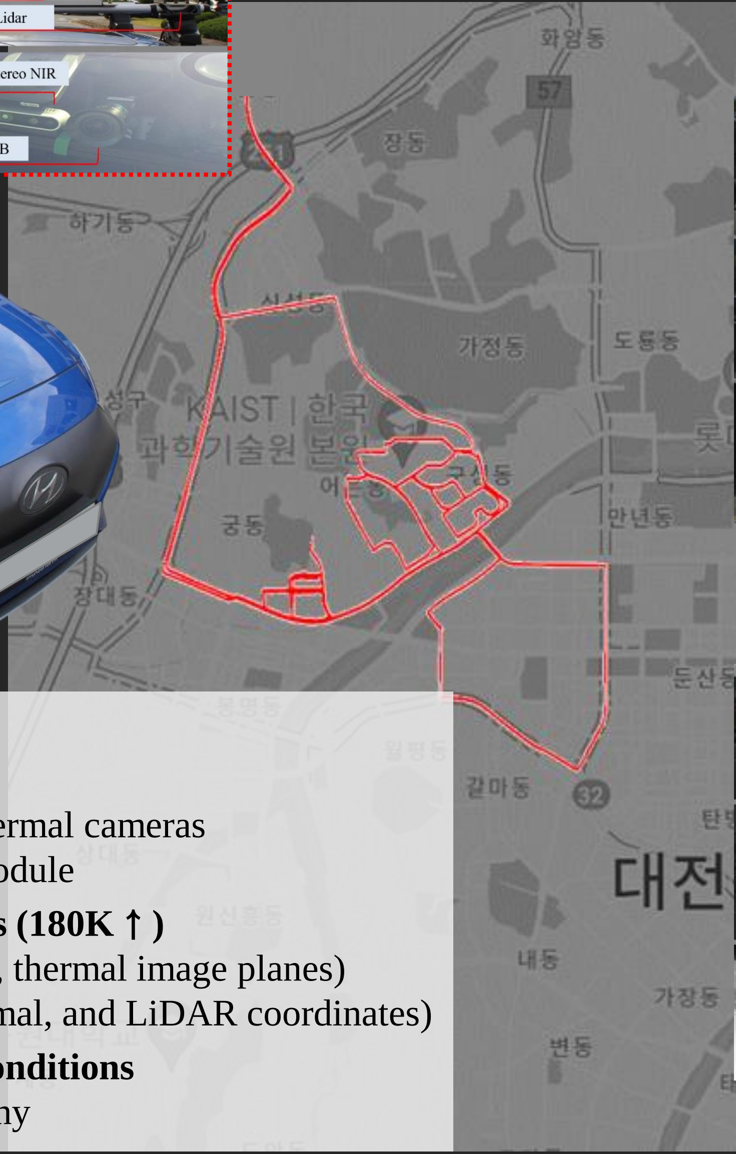
[Benchmark] It is rarely explored on thermal spectrum domain for spatial understanding.

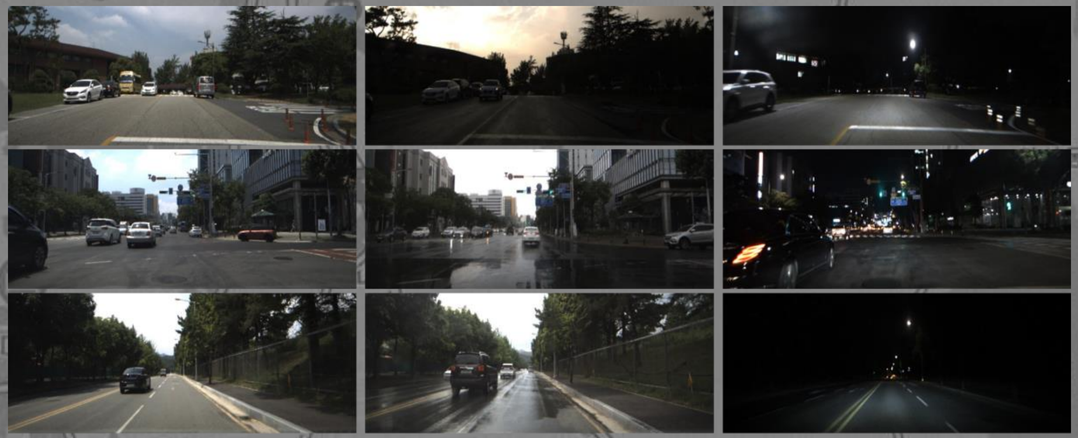
- Only a few papers on spatial perception from thermal spectrum band.
- Need to figure out advantages and disadvantages of thermal camera in various geometry tasks



Multi-Spectral Stereo (MS^2) Dataset







RGB/NIR/Thermal image in various conditions

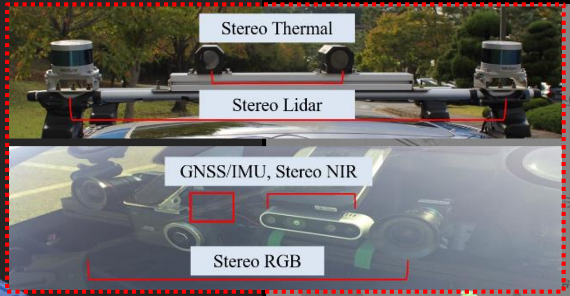
MS^2 Dataset's Features

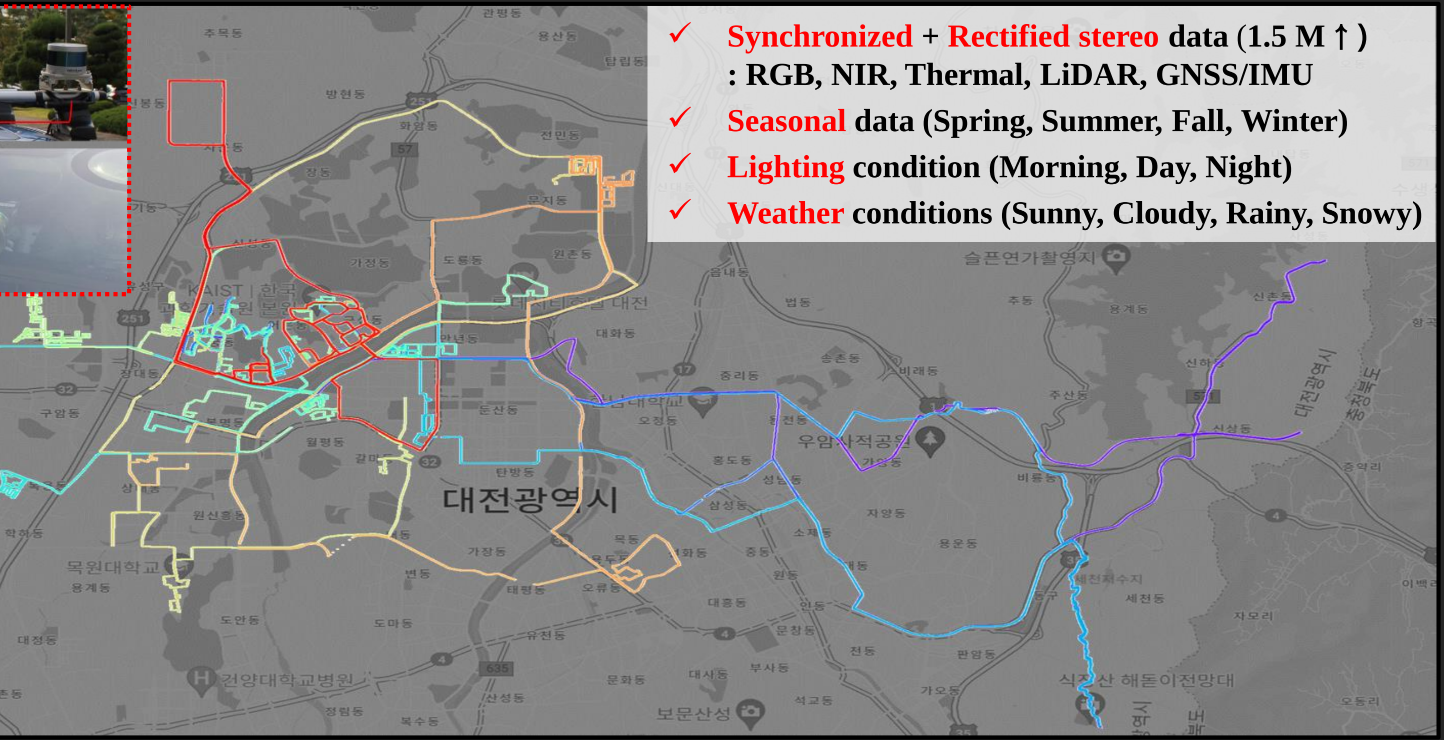
- ✓ **Multi-sensor Stereo dataset**
  - Stereo RGB, Stereo NIR, Stereo thermal cameras
  - Stereo LiDAR, single GPS/IMU module
- ✓ **Synchronized +Rectified data pairs (180K ↑)**
  - Projected depth map (in RGB, NIR, thermal image planes)
  - Odometry data (in RGB, NIR, thermal, and LiDAR coordinates)
- ✓ **A number of places with various conditions**
  - Day/Night + Clear-sky/Cloudy/Rainy

Ukcheol Shin, et al, "Deep Depth Estimation from Thermal Image", CVPR 2023

Multi-Spectral Stereo Seasonal (MS^3) Dataset

The first city-scale thermal stereo seasonal dataset





- ✓ Synchronized + Rectified stereo data (1.5 M ↑)
  : RGB, NIR, Thermal, LiDAR, GNSS/IMU
- ✓ Seasonal data (Spring, Summer, Fall, Winter)
- ✓ Lighting condition (Morning, Day, Night)
- ✓ Weather conditions (Sunny, Cloudy, Rainy, Snowy)

Seasonal changes

Rainy, snowy conditions

(from left) RGB, NIR, Thermal, Projected LiDAR

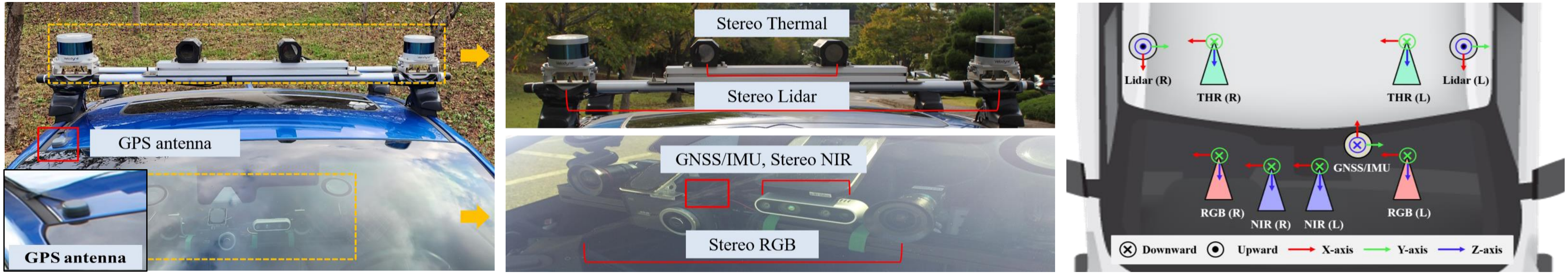
Extended version from Ukcheol Shin, et al, “Deep Depth Estimation from Thermal Image", CVPR 2023

MS^3 Dataset: Sensor System

Components of our sensor system :

- ✓ ***Stereo*** RGB cameras
- ✓ ***Stereo*** NIR cameras
- ✓ ***Stereo*** thermal cameras
- ✓ ***Stereo*** LiDAR
- ✓ Single GNSS/IMU
- ✓ ***Synchronized*** data acquisition

RCV Lab's Vehicular Sensor System

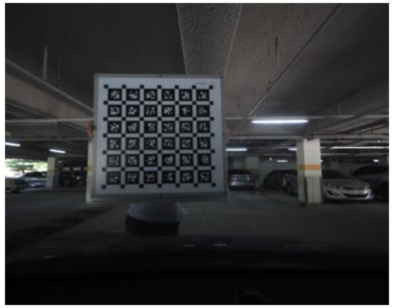


MS^3 Dataset: Calibration

Multi-sensor calibration is promising research direction!

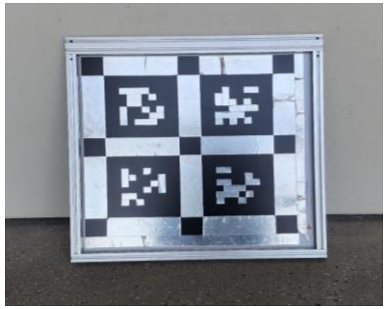
1. AprilTag (6x6)
- ✓ Stereo RGB calibration
- ✓ Stereo NIR calibration
- ✓ RGB-NIR calibration
- ✓ NIR-IMU/Lidar calibration

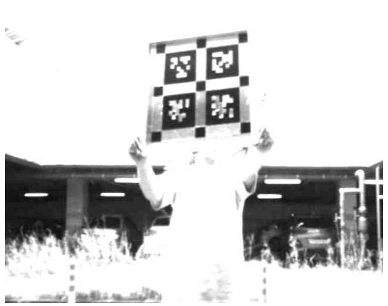

AprilTag board (6x6)


RGB image


NIR image

2. Partial metal coated AprilTag (2x2)
- ✓ NIR-Thermal calibration

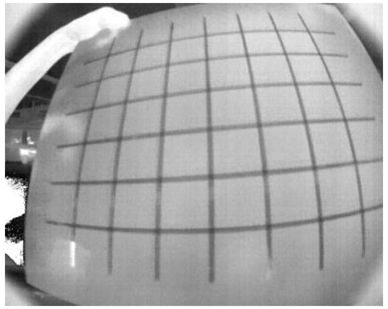

AprilTag board (2x2)

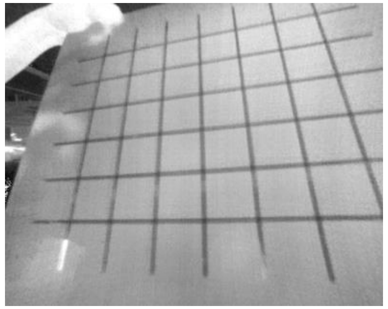

NIR image


Thermal image

3. Cooper-coated Lineboard (7x6)
- ✓ Stereo Thermal calibration


Line board (6x7)


Thermal image


After rectification

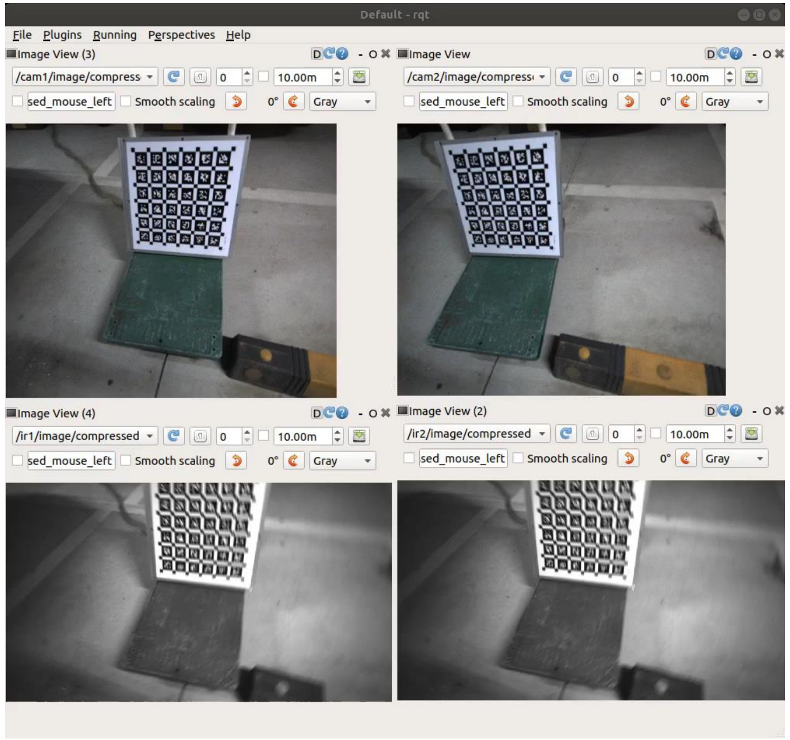
Calibration board:

[1] Olson, Edwin, "AprilTag: A robust and flexible visual fiducial system.", ICRA, 2011
[2] Choi *et al.*, "KAIST multi-spectral day/night data set for autonomous and assisted driving.", T-ITS, 2018

MS^3 Dataset: Calibration

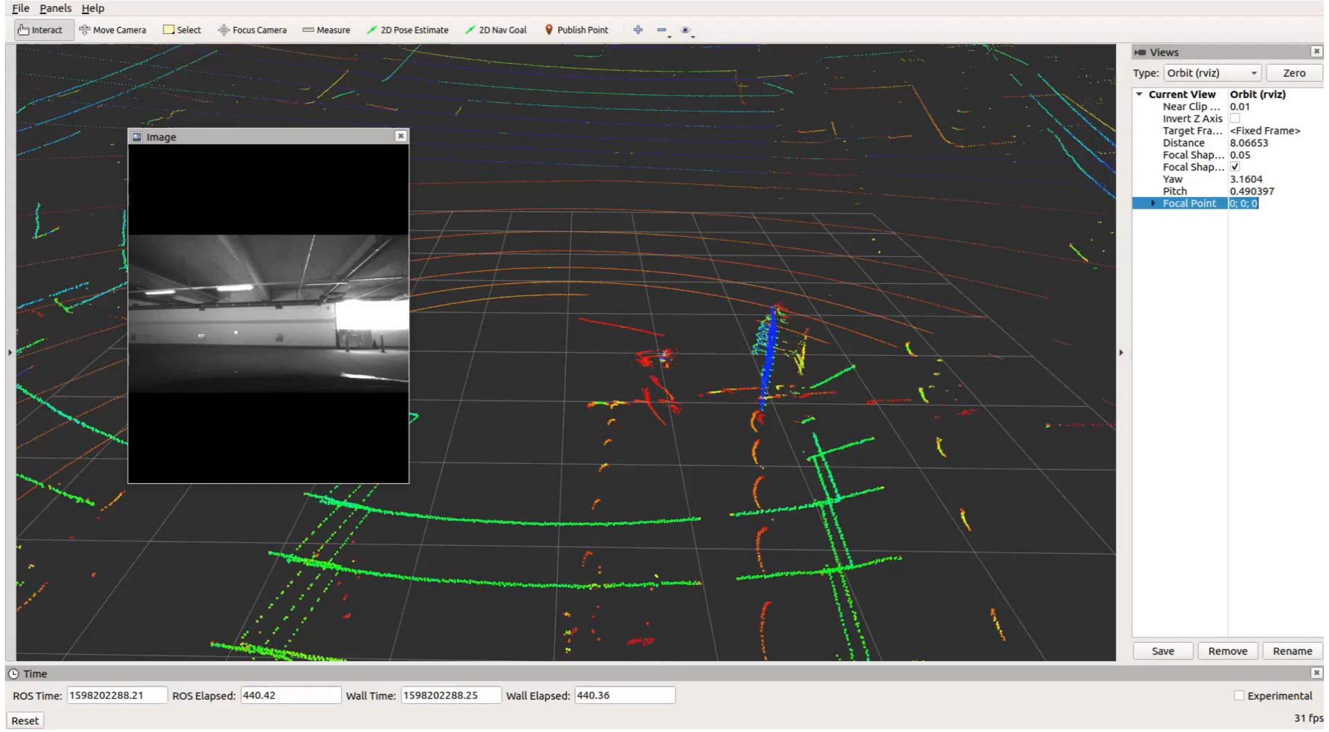
NIR-IMU calibration

- ✓ AprilTag board (6x6)
- ✓ Kalibr library



NIR-LiDAR calibration

- ✓ AprilTag board (6x6)
- ✓ Plane fitting.



[Kalibr] Rehder et al., "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes.", ICRA, 2016

MS^3 Dataset: Examples

Seasonal diversity
- ✓ Spring/summer
- ✓ Autumn/winter

Locational diversity
- ✓ City/residential
- ✓ Campus/road
- ✓ Suburban

Lighting condition
- ✓ Well-lit
- ✓ Low-light

Rainy condition
- ✓ Occlusion
- ✓ Blur
- ✓ Glare

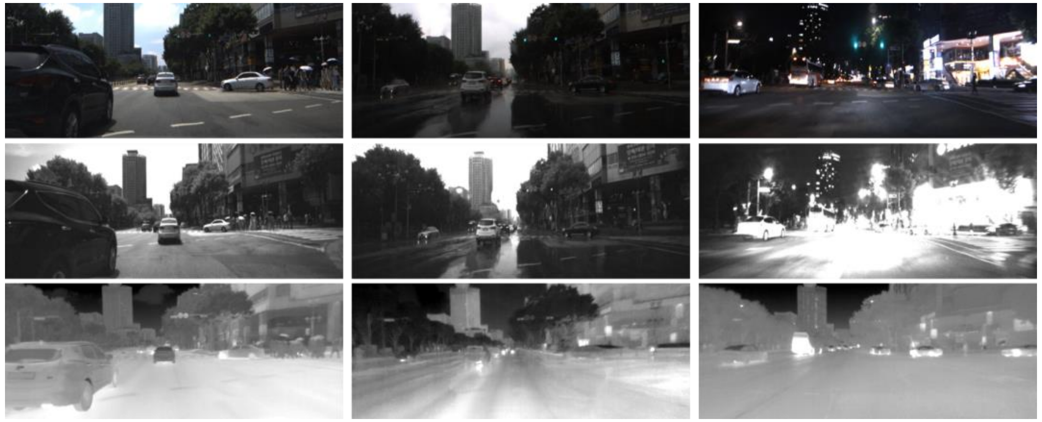
Snowy condition
- ✓ Day/Night
- ✓ Glare

MS^3 Dataset: Examples

Able to do domain analysis between
- ✓ **Modality**
- ✓ **Time**
- ✓ **Space**

Sensor diversity : RGB/NIR/Thermal images

(a) Driving scenarios – Campus (Morning, Day, Night)



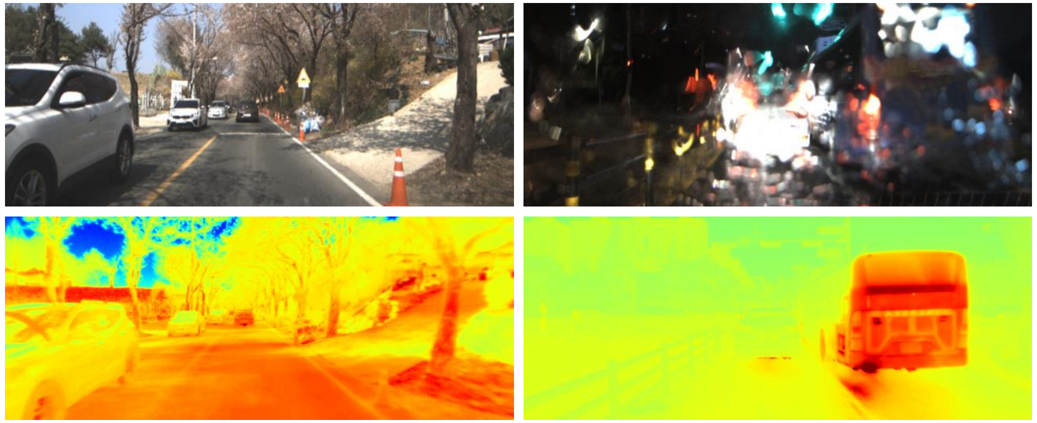
(b) Driving scenarios – City (Day, Rain, Night)

(c) Driving scenarios – Residential (Morning, Day, Night)

(d) Driving scenarios – Road2 (Day, Rain, Night)

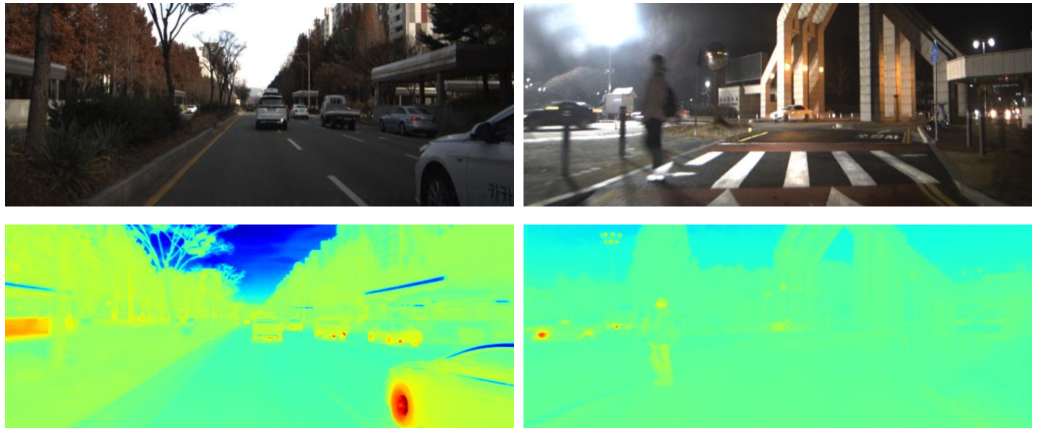
MS^3 Dataset: Examples

Temperature diversity : Seasonal, Day-Night, Rain-Snow, Clear-sky, Cloudy

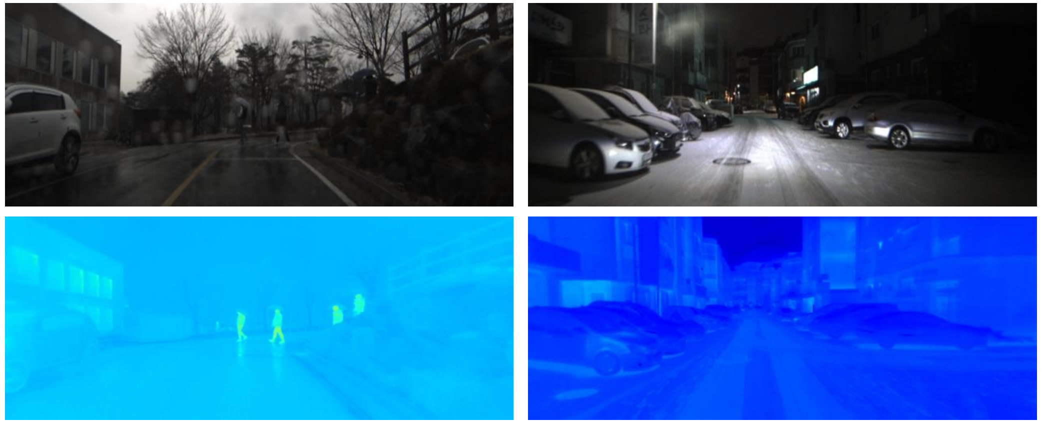


Spring (02:00 PM), Temp mean: 32.9°C, std: 6.3 °C
Spring (10:30 PM), Temp mean: 28.1°C, std: 1.4 °C

Summer (11:30 AM), Temp mean: 44.5°C, std: 6.8 °C
Summer (10:30 PM), Temp mean: 33.5°C, std: 2.6 °C



Autumn (1:30 PM), Temp mean: 21.4°C, std: 7.4 °C
Autumn (08:00 PM), Temp mean: 18.3°C, std: 1.7 °C

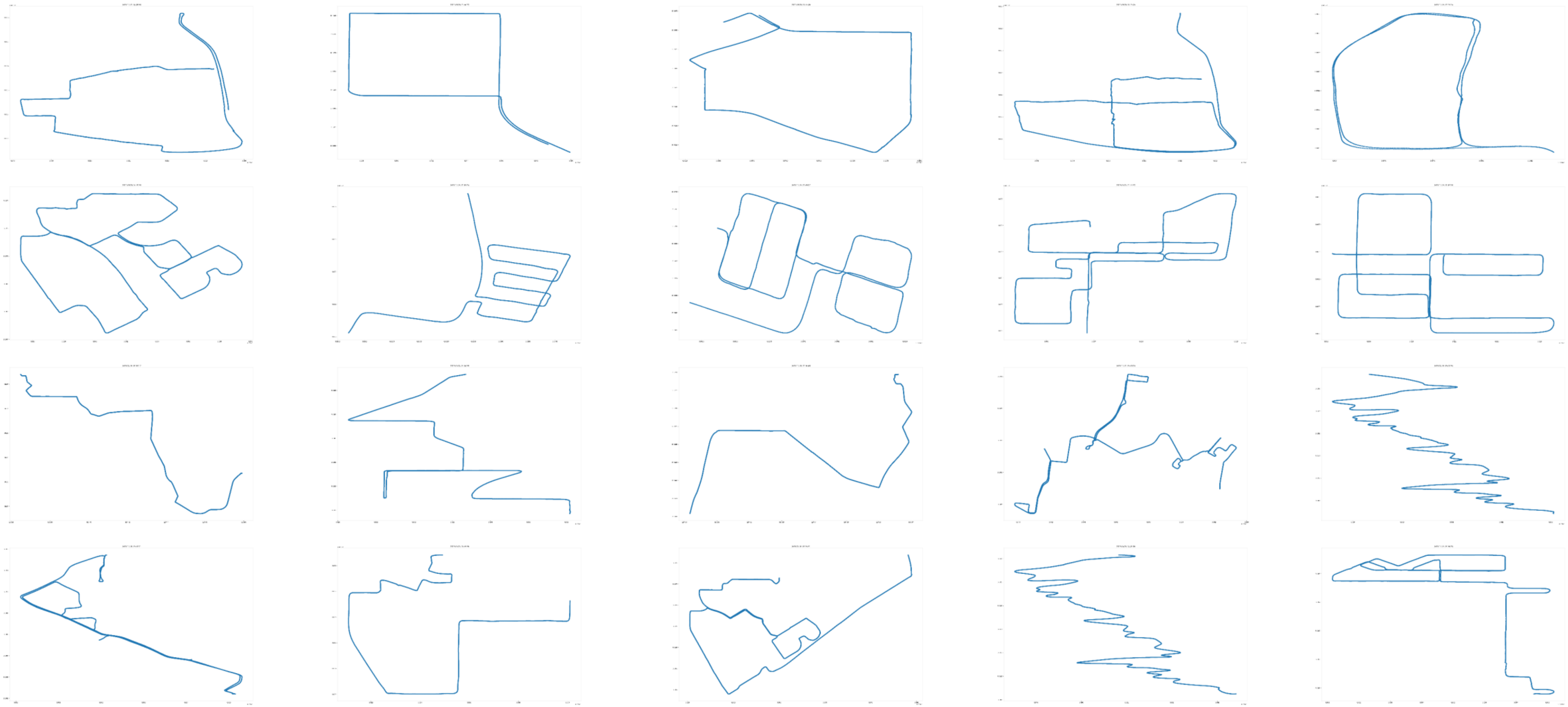


Winter (12:00 AM), Temp mean: 10.7°C, std: 0.6 °C
Winter (08:00 PM), Temp mean: 0.4°C, std: 1.7 °C

MS^3 Dataset: Examples

Possible research :
VO/SLAM/3D recon/NeRF from thermal/multi-sensor

Trajectory diversity : Open loop, (single/multiple) closed loop, forward/backward moving, frequent rotational scenario



MS^2 Dataset: Examples

Left ➔ Right : RGB, NIR, Thermal, Depth, Trajectory

- Seq: Summer, Day, Clear-sky, Campus

- Seq: Summer, Day, Rainy, Campus

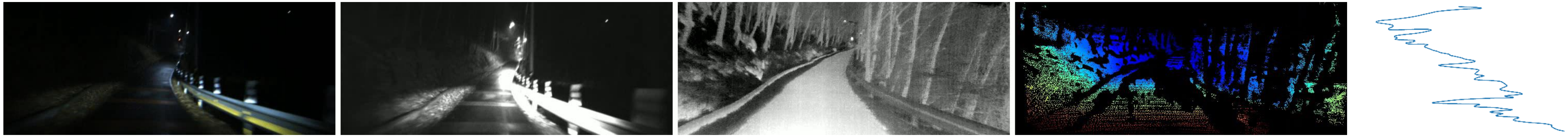
- Seq: Summer, Night, Clear-sky, Campus

- Seq: Summer, Day, Rainy, Road

MS^3 Dataset: Examples

Left ➔ Right : RGB, NIR, Thermal, Depth, Trajectory

- **Seq: Autumn, Night, After rain, Suburban**



- **Seq: Winter, Night, Snowy, Residential**

- **Seq: Spring, Night, Rainy, Road**

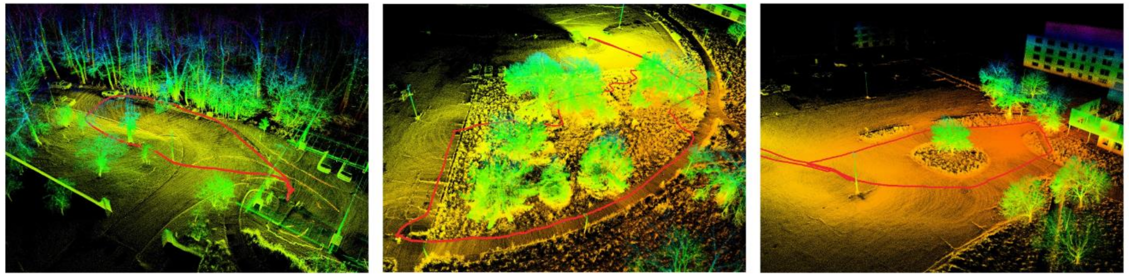
- **Seq: Spring, Day, Rainy, Residential**

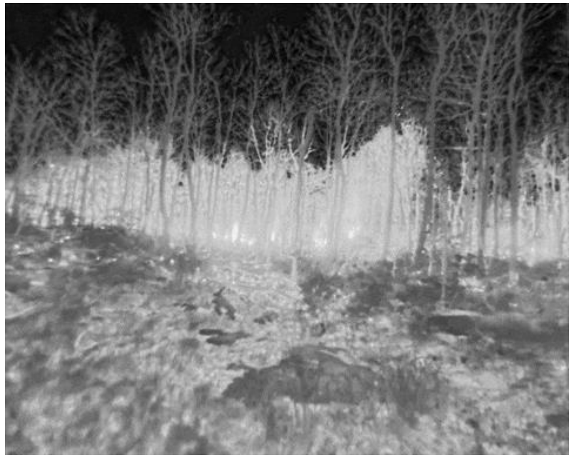
FIReStereo: Forest InfraRed Stereo Dataset

Carnegie Mellon University The Robotics Institute

AIRLAB

The first thermal-stereo dataset in forest fire & smoke





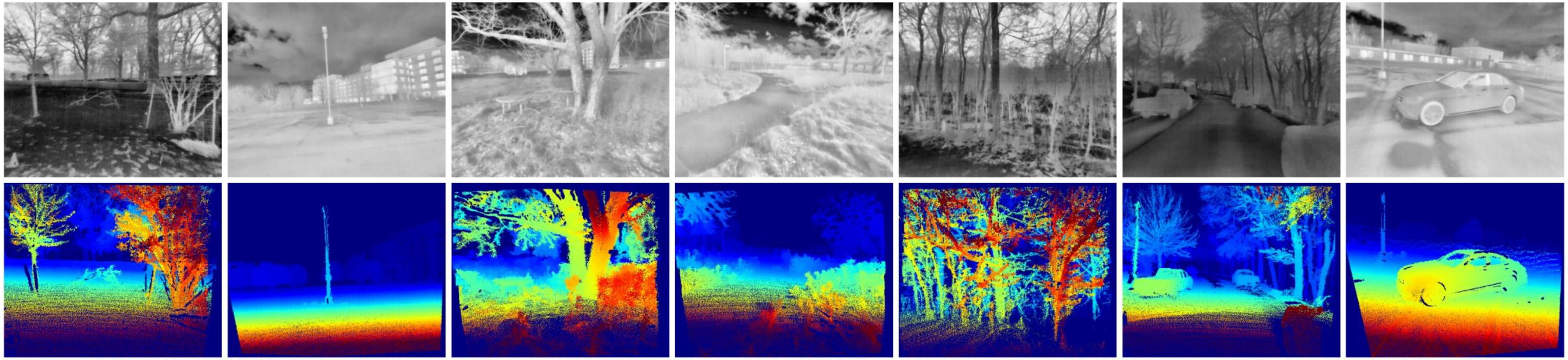
Devansh Dhrafan, et al, “FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments”, Under-review (U. Shin: Co-author)

FIReStereo: Forest InfraRed Stereo Dataset

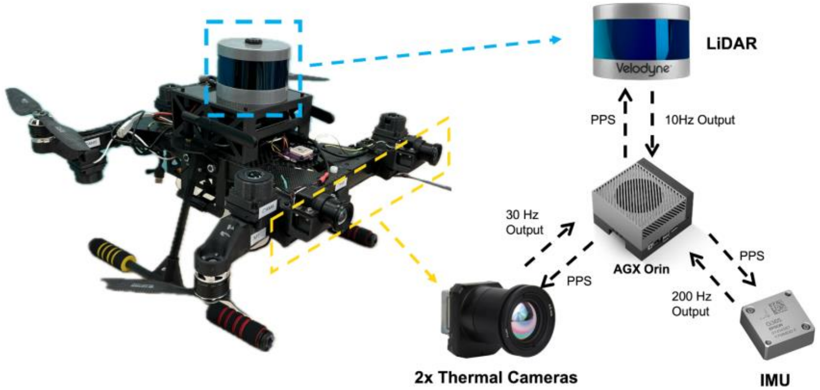
Carnegie Mellon University The Robotics Institute

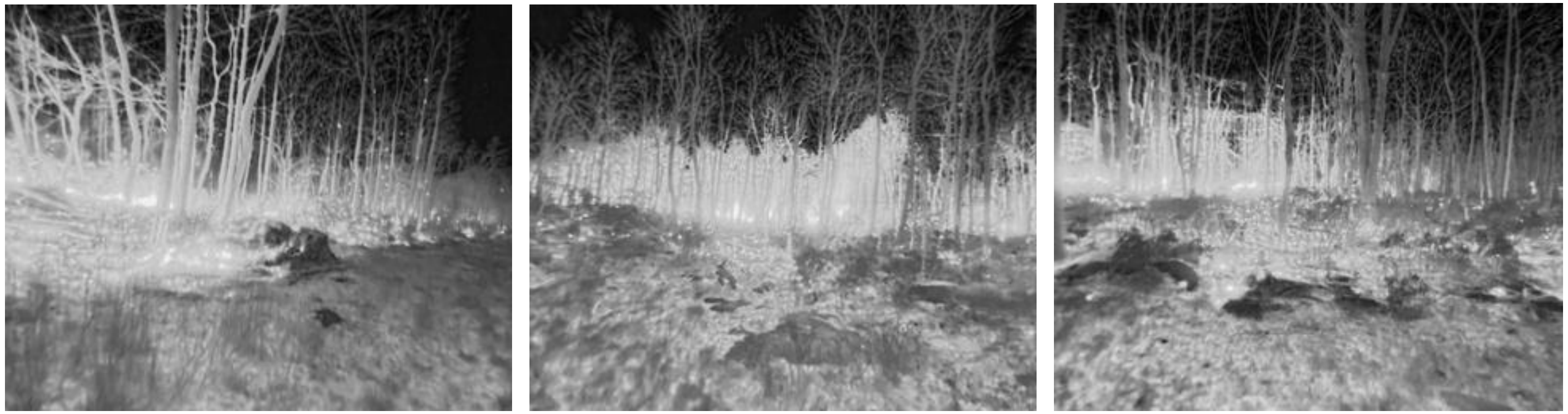
AIRLAB

The first thermal-stereo dataset in forest fire & smoke



Sparse Trees | Lamp Post | Single Tree | Park | Dense Trees | Night | Car





Devansh Dhrafan, et al, “FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments”, Under-review (U. Shin: Co-author)

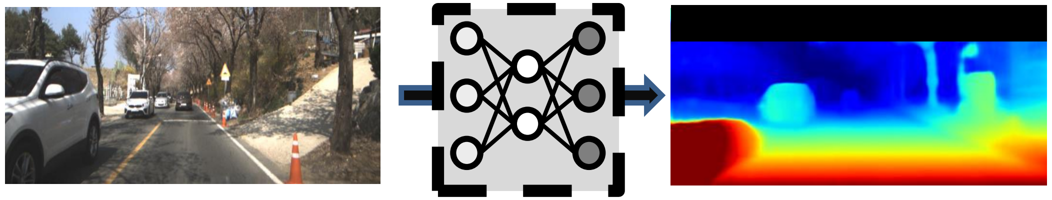
Spatial Perception from Thermal Image

Unique information & Safety

Clean visibility against *low-light, snowy, rainy conditions*

Q. Can we leverage the robustness of thermal image in spatial perception tasks?
+ is it better than spatial perception from RGB or NIR images?

Deep Depth Estimation from X

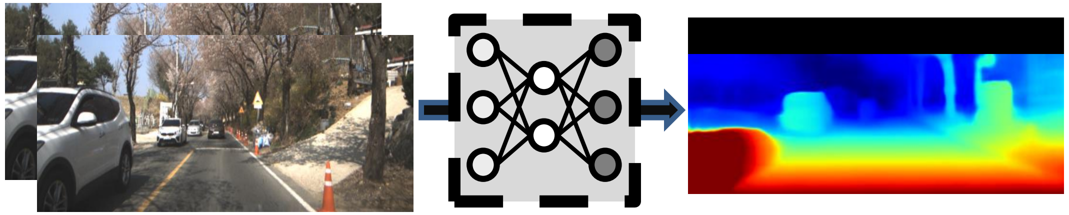


Monocular Depth Estimation

Classification based methods :
- Soft labels for ordinal regression, CVPR19
- Deep ordinal regression, CVPR 19

Regression based methods :
- Vision transformers for dense prediction., ICCV 21
- Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset Transfer, T-PAMI 22
- **Neural window fully-connected crfs for monocular depth estimation, CVPR 22**

Hybrid methods :
- Adabins: deep estimation using adaptive bins, CVPR21
- Binsformer: revisiting adaptive bins for monocular depth estimation, Arxiv preprint 22



Stereo Depth Estimation

3D cost volume based methods :
- Learning for disparity estimation through feature constancy, CVPR 18
- Real-time self-adaptive deep Stereo, CVPR 19
- **AAnet: Adaptive aggregation network for efficient stereo matching, CVPR 20**

4D cost volume based methods :
- Pyramid stereo matching network, CVPR 18
- Group-wise correlation stereo network, CVPR19
- CFnet: Cascade and fused cost volume for robust stereo matching, CVPR 21
- Attention concatenation volume for accurate and efficient stereo matching., CVPR22

Ukcheol Shin, et al, "Deep Depth Estimation from Thermal Image: Dataset, Benchmark, Challenges", To be available in Arxiv

Depth from X: Training and Evaluation Splits

Exp1. Evaluation on MS^2 dataset

Non-overlapped train/val/test subset

In-distribution test

Train set

- ✓ Season: **Summer**
- ✓ Light condition: **Day, Night**
- ✓ Weather condition: **Clear-sky, Cloudy, Rain**

Test set

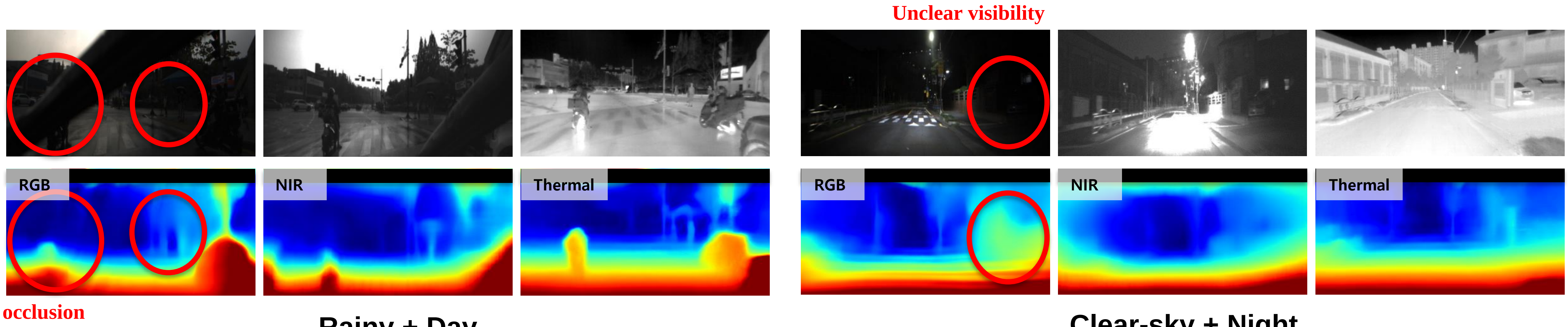
- ✓ Season: **Summer**
- ✓ Light condition: **Day, Night**
- ✓ Weather condition: **Clear-sky**, **Cloudy, Rain**

Seasonal data

Rainy, snowy

Depth from X: Benchmark and findings

Findings 1. Monocular depth from thermal image performs the best in day, night, rainy conditions



Rainy + Day

Clear-sky + Night

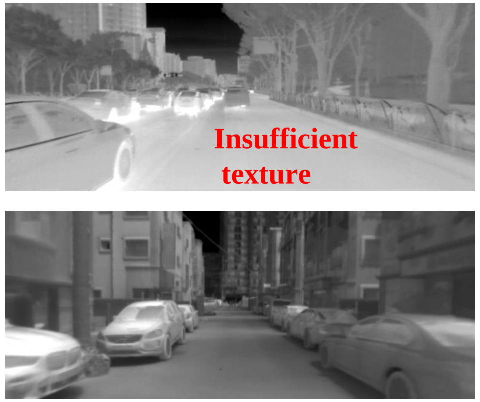
Test set : summer (clear-day, clear-night, rainy-day)

Red: best, purple: runner-up

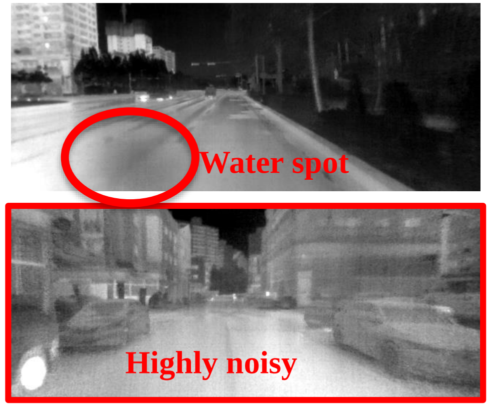
Monocular	RGB		NIR		THR	
NeWCRF	RMSE(↓)	$\delta < 1.25$(↑)	RMSE(↓)	$\delta < 1.25$(↑)	RMSE(↓)	$\delta < 1.25$(↑)
Sm_Clear_Day	3.111	**94.8**	**3.071**	93.3	**2.717**	**95.1**
Sm_Clear_Night	3.573	89.9	**3.157**	**91.2**	**2.544**	**95.2**
Sm_Rainy_Day	**4.447**	**87.0**	5.042	81.0	**3.503**	**90.9**

Depth from X: Benchmark and findings

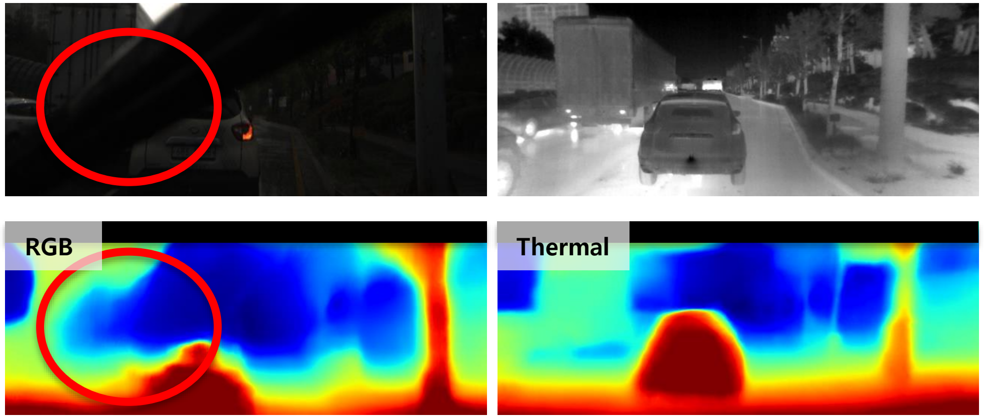
Findings 2. Thermal images have disadvantages in matching problem. But, still perform better in depth.



Normal condition (Clear-sky+Day)



Low thermal variance (rainy, night)



Rainy + Day

Stereo	RGB			THR		
AANet	RMSE(↓)	$\delta < 1.25$(↑)	>1px (↓)	RMSE(↓)	$\delta < 1.25$(↑)	>1px (↓)
Sm_Clear_Day	1.465	99.3	**2.1**	**1.203**	**99.6**	2.4
Sm_Clear_Night	1.569	99.1	**2.8**	**1.442**	**99.2**	5.4
Sm_Rainy_Day	4.114	91.4	19.0	**1.532**	**99.4**	**3.6**

Depth evaluation metrics

➔ **stereo matching metrics**

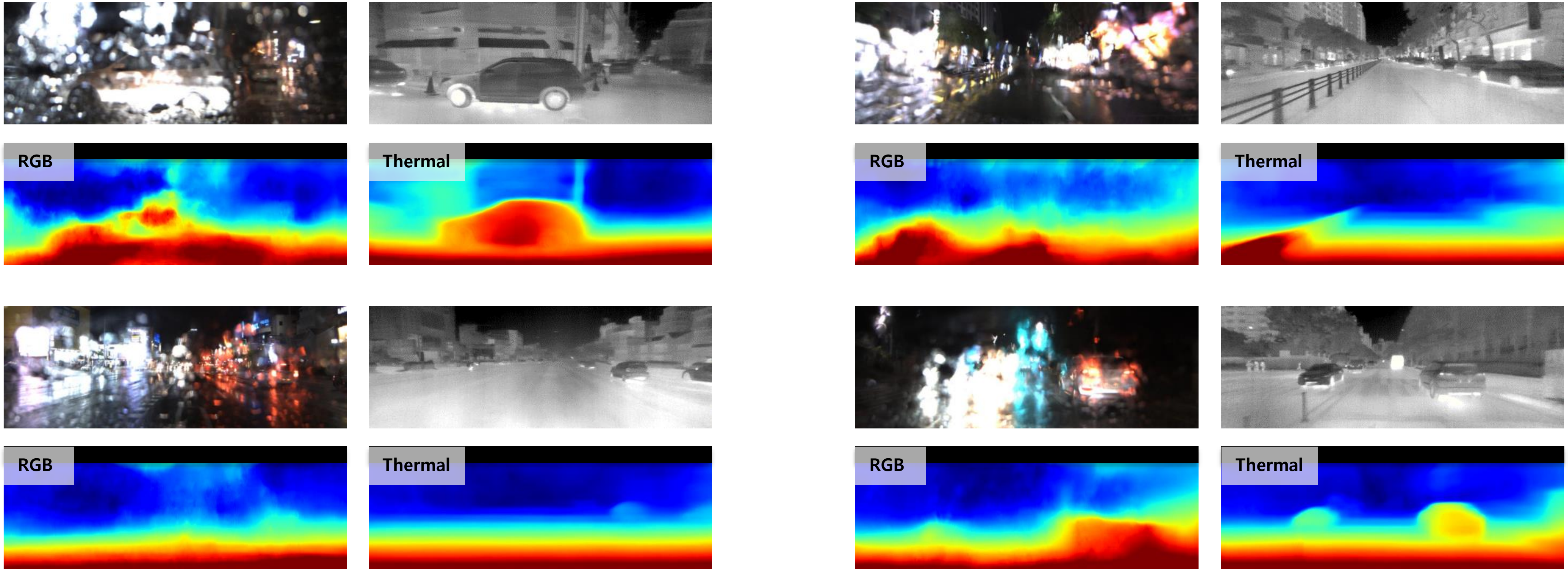
Red: best

*In stereo matching, RGB and thermal stereo has the same baseline (30cm) and resolution (640x256) / NIR stereo has a different baseline, so excluded for a fair comparison.

Depth from X: Benchmark and findings

Findings 2. Thermal images have disadvantages in matching problem. But, still perform better in depth.

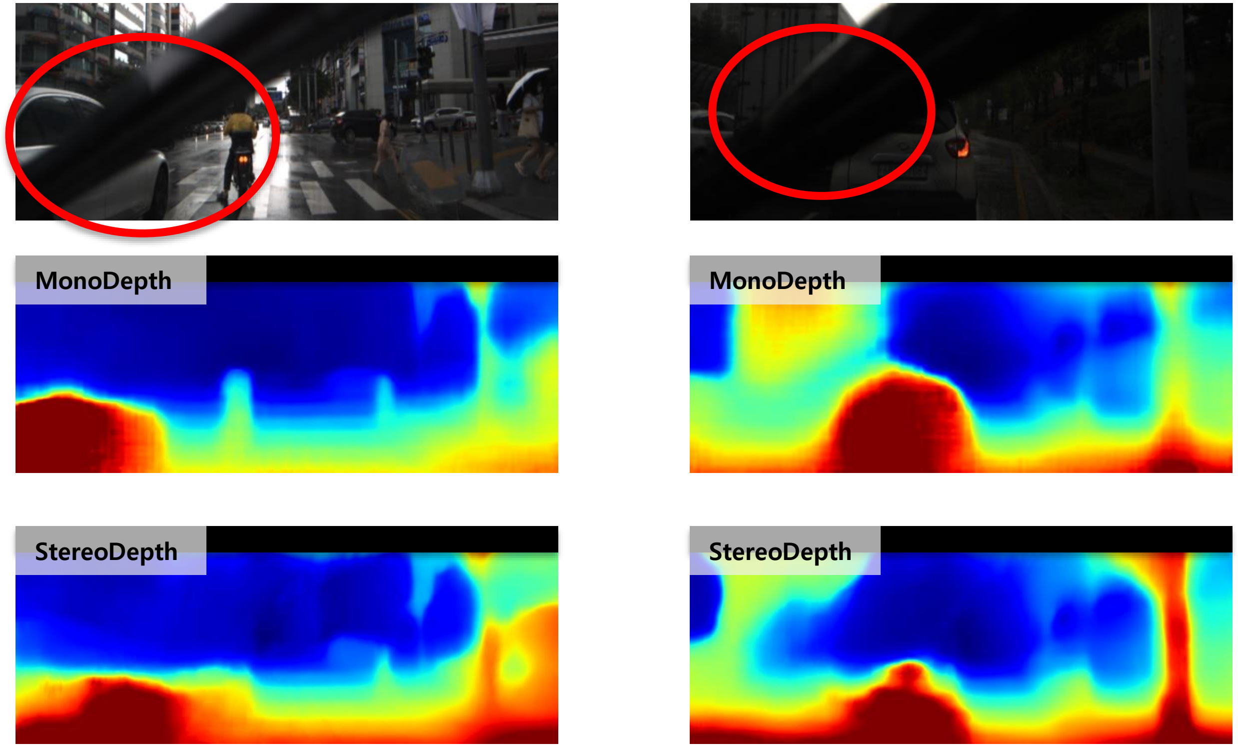
Q. Without using windshield wipers



(Spring) Rainy + Night

Depth from X: Benchmark and findings

Findings 3. In rainy conditions, monocular depth from RGB is better than stereo depth in some cases.



Test set : summer (clear-day, clear-night, rainy-day)

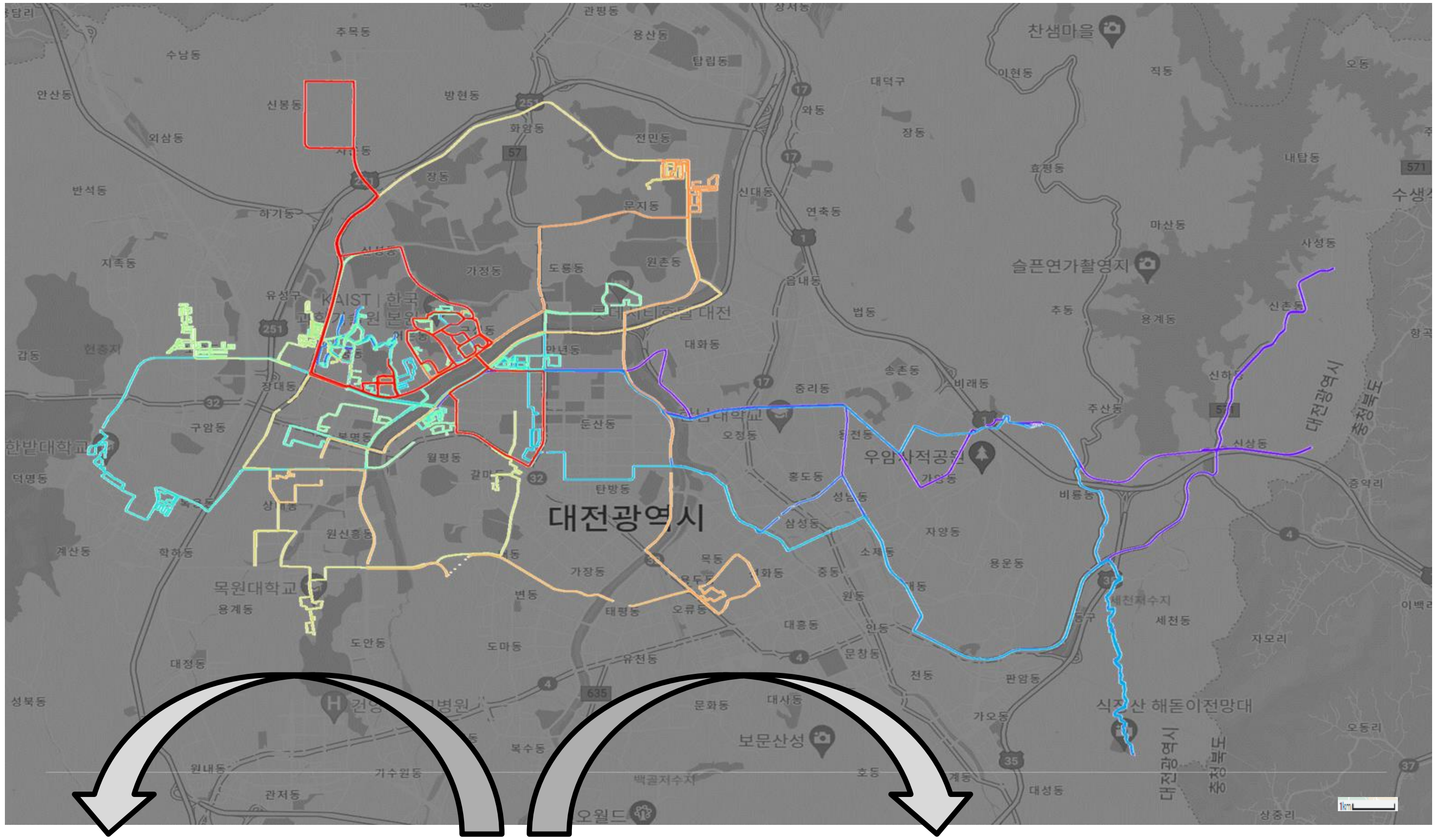
Monocular	RGB	
NeWCRF	RMSE(↓)	$\delta < 1.25$(↑)
Sm_Clear_Day	3.111	94.8
Sm_Clear_Night	3.573	89.9
Sm_Rainy_Day	**4.447**	**87.0**

Stereo	RGB	
AANet	RMSE(↓)	$\delta < 1.25$(↑)
Sm_Clear_Day	1.465	99.3
Sm_Clear_Night	1.569	99.1
Sm_Rainy_Day	**4.114**	**91.4**

Possible research:
Adaptive multi-view stereo in rainy conditions

Depth from X: Training and Evaluation Splits

Exp2. Out-of-distribution Evaluation



Non-overlapped train/val/test subset

Zero-shot generalization test

Train set

- ✓ Season: **Summer**
- ✓ Light condition: Day, Night
- ✓ Weather condition: Clear-sky, Cloudy, (Light) Rain

Test set (Remaining colored trajectory)

- ✓ Season: **Spring**, Summer, **Autumn, winter**
- ✓ Light condition: **Day, Night**
- ✓ Weather condition: **Clear-sky, (Heavy/Light) Rain/Snow**
- ✓ Various extreme conditions

Seasonal data

Rainy, snowy

Depth from X: Benchmark and findings

Findings 4. thermal images is the best domain shift robust modality

Test set (zero-shot): Spring, Fall, Winter (day/night with clear-sky/rainy/snowy)

Red: best, purple: runner-up

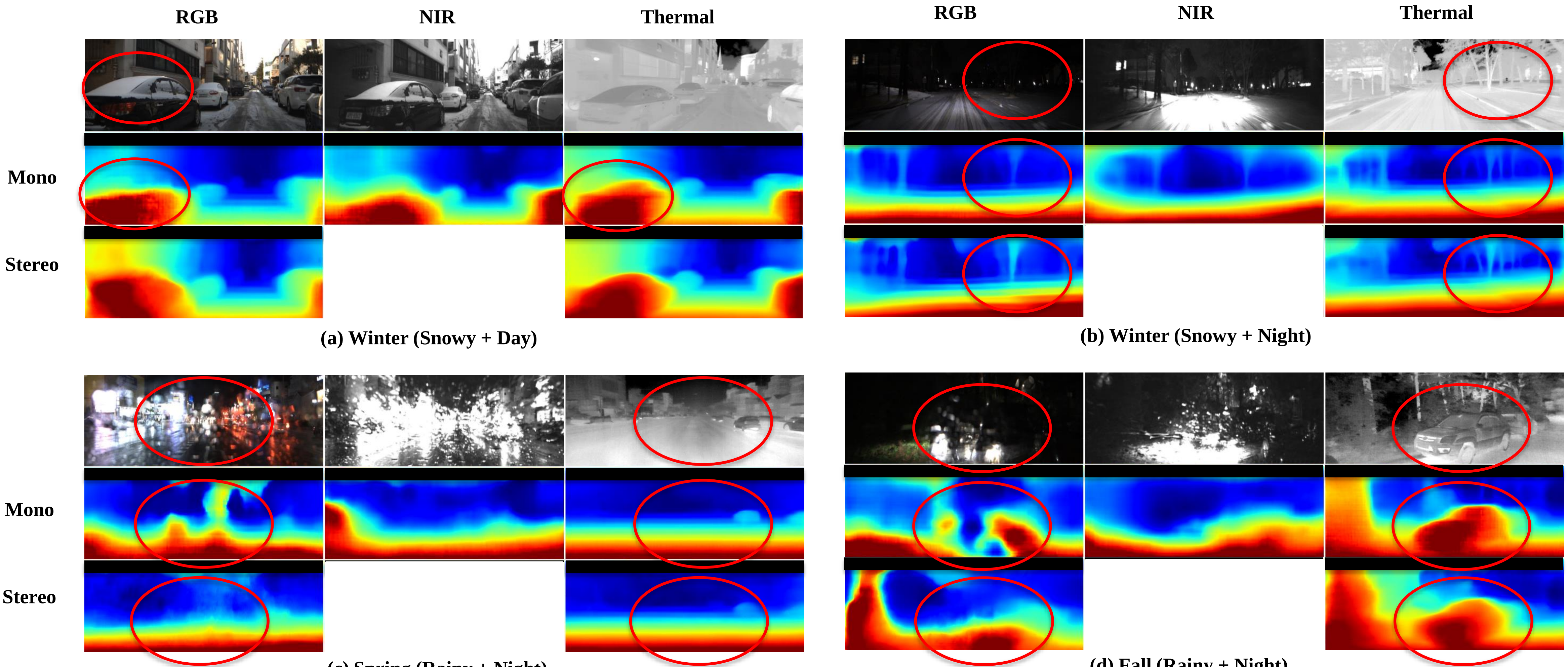
Monocular	RGB			NIR			THR		
NeWCRF*	RMSE(↓)	$\delta < 1.25$(↑)	ΔRMSE	RMSE(↓)	$\delta < 1.25$(↑)	ΔRMSE	RMSE(↓)	$\delta < 1.25$(↑)	ΔRMSE
Base(Sm_Clear_Day)	3.111	94.8	-	3.071	93.3	-	2.717	95.1	-
Spring_Clear_Day	5.473	70.0	-2.362	**4.157**	**77.4**	**-1.086**	**3.810**	**84.9**	**-1.093**
Spring_Rainy_Day	5.599	**68.9**	-2.488	**5.470**	65.8	-2.399	**3.207**	**85.5**	**-0.490**
Spring_Rainy_Night	7.282	**57.8**	-4.171	**7.207**	52.2	**-4.136**	**3.848**	**81.6**	**-1.131**
Fall_Clear_Day	5.26	80.3	-2.149	**3.814**	**89.6**	**-0.743**	**4.290**	**88.1**	**-1.573**
Fall_Rainy_Night	5.017	75.4	-1.906	**3.532**	**83.8**	**-0.461**	**3.271**	**88.1**	**-0.554**
Winter_Snowy_Day	5.092	72.9	-1.981	**4.740**	**74.5**	**-1.669**	**3.640**	**83.2**	**-0.923**
Winter_Snowy_Night	6.154	73.0	-3.043	**4.585**	**83.8**	**-1.514**	**3.362**	**91.1**	**-0.645**
Avg (Eval: zero-shot)	**5.555**	**72.5**	**-2.444**	**4.613**	**77.0**	**-1.542**	**3.567**	**84.9**	**-0.850**

➔ RMSE(each OoD scenario) - RMSE(Base, in-distribution)

***All trained models use a number of augmentations (color jitter, contrast jitter, brightness jitter, …)**

Depth from X: Benchmark and findings

Findings 4. thermal images is the best domain shift robust modality



(a) Winter (Snowy + Day)

(b) Winter (Snowy + Night)

(c) Spring (Rainy + Night)

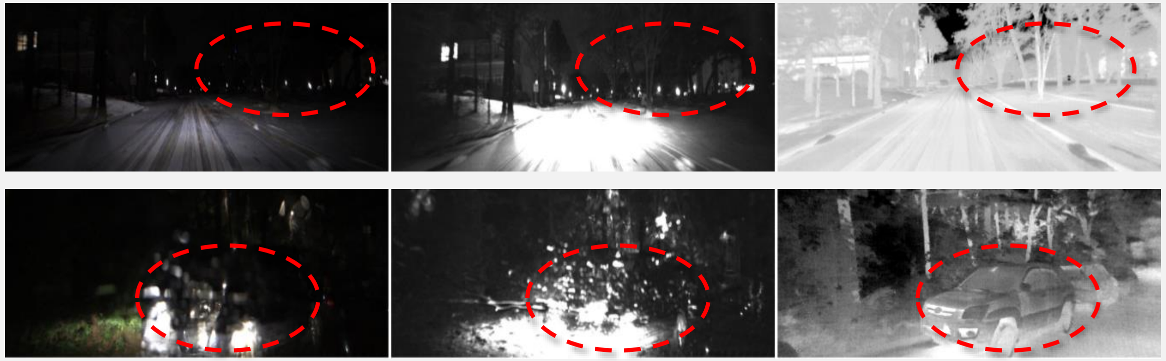
(d) Fall (Rainy + Night)

Part 1. Takeaway message

[Benchmark] Deep Depth Estimation from Thermal Image

- **Thermal camera is a potential rescue for robust spatial perception in challenging conditions**

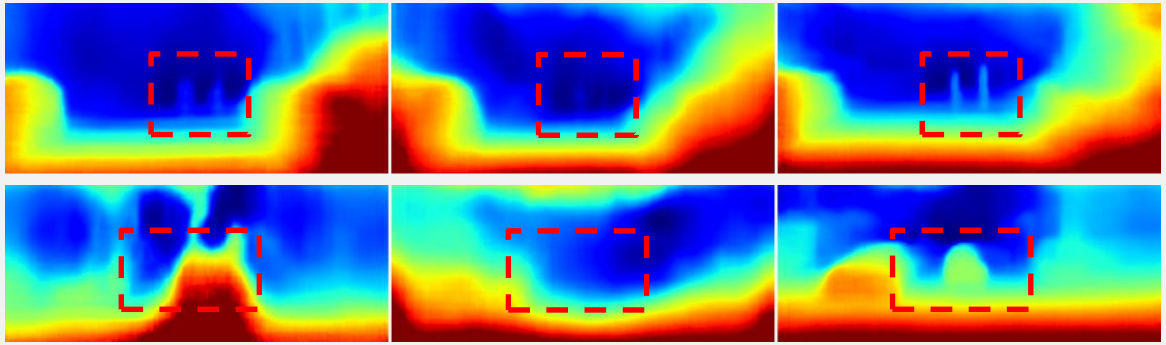
Unique information & Safety

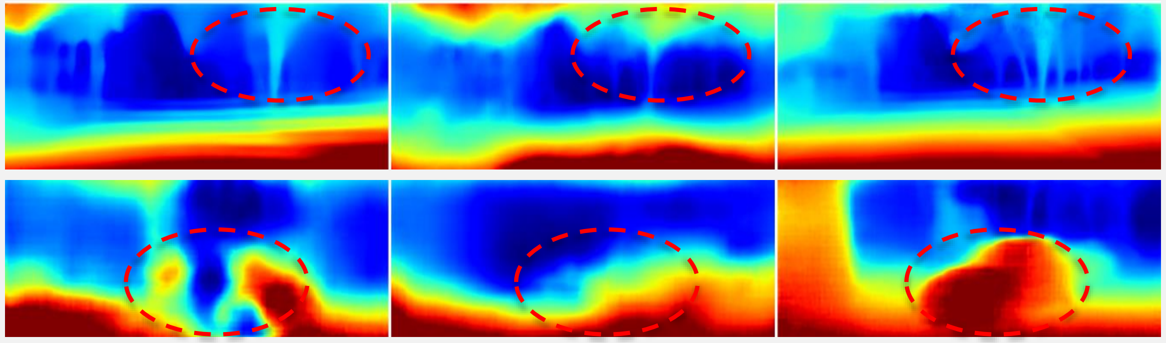


Clean visibility* against *low-light, snowy, rainy conditions

Mono

Stereo





Depth from thermal images shows the best accuracy, robustness, and generalization performance

Part 1. Takeaway message

[Take-home message]

- Thermal camera is a potential rescue for robust spatial perception in adverse weather/lighting conditions
- Thermal camera has the best domain-shift robustness against weather/lighting/seasonal changes

However,

- Suffer from low-texture, low-contrast, severe-noise
- Disadvantages in matching problem (stereo matching, optical flow, …)
- Needs extensive exploration in spatial perception tasks (odometry, SLAM, scene flow, NeRF, …)

$+ \alpha$

- RGB+NIR fusion could be a cheap and effective solution for night vision
- In rainy condition, monocular depth from RGB is better than stereo matching
- How to improve prediction results of RGB image in rainy condition?
- Why domain generalization of RGB image is worse than NIR/Thermal images?

Part 2. Visual Perception from Thermal Image : Challenges (What's next?)

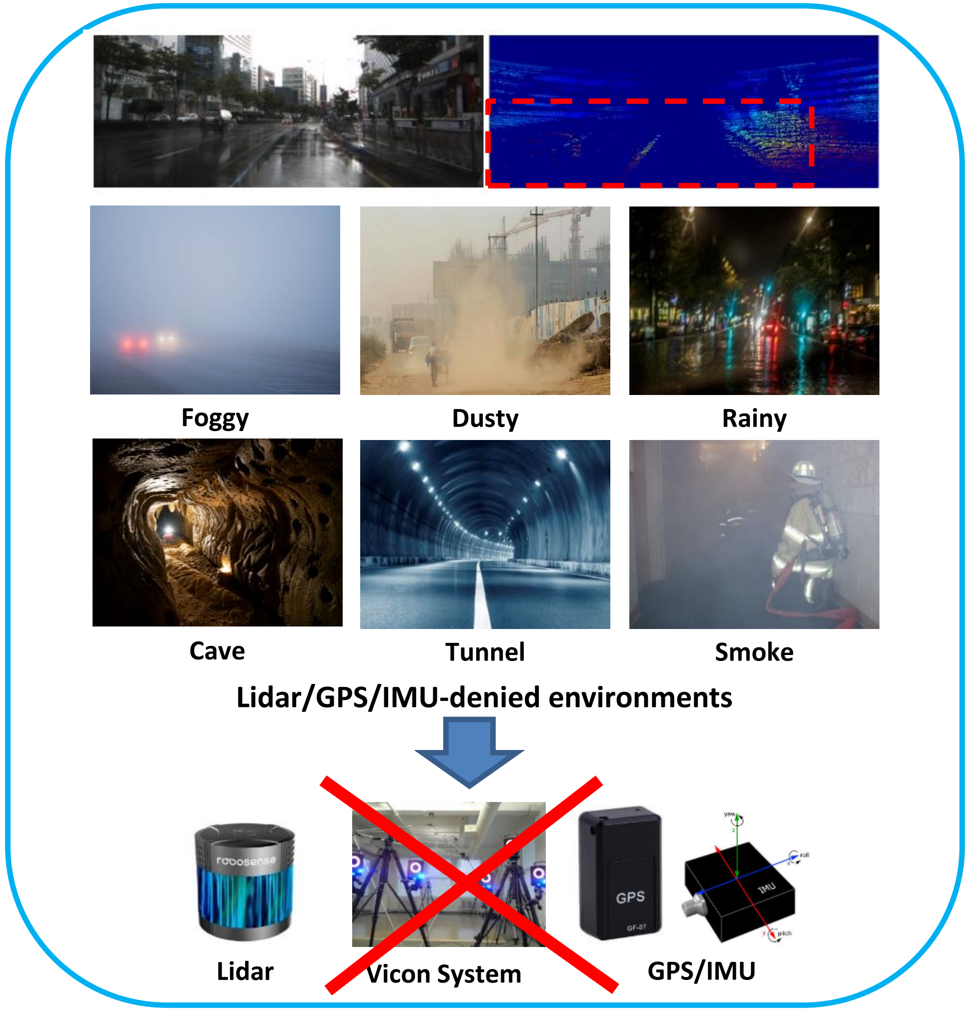
1. GT label in challenging environments
2. Thermal image enhancement
3. Traversable area detection in challenging conditions
4. Detecting transparent objects
5. Exploration on various spatial perception tasks
6. Selective sensor fusion in challenging conditions
7. Modality bias in multi-sensor fusion

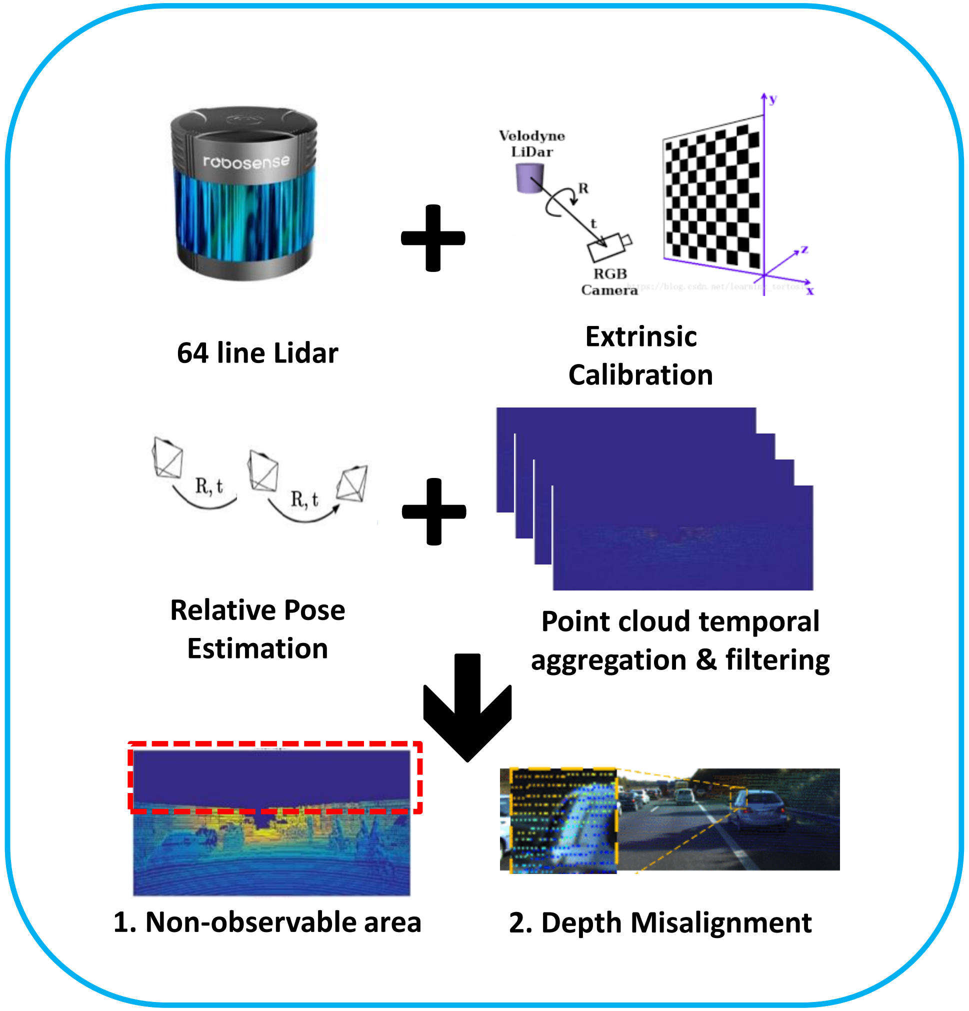
What's Next?

Q. What is the unexplored part, disadvantage, or unique property of thermal camera?

1. GT label in challenging conditions

[GT Label] infeasible to collect GT data in adverse weather and locations.





✓ Cannot collect sensor data

✓ Complicated post-process

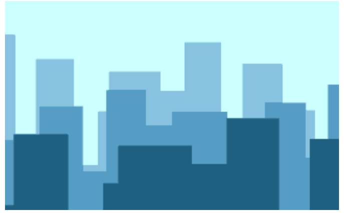
Sol: Self-supervision from thermal images

Self-supervision from images

- Monocular Depth Estimation
- Stereo Depth Estimation
- 3D reconstruction
- NeRF
- Occupancy Estimation
- Odometry/ SLAM
- Multi-view Stereo Depth Estimation
- Optical Flow Estimation
- Scene Flow Estimation
- …

Self-supervision can train various 3D geometry tasks without utilizing GT labels.

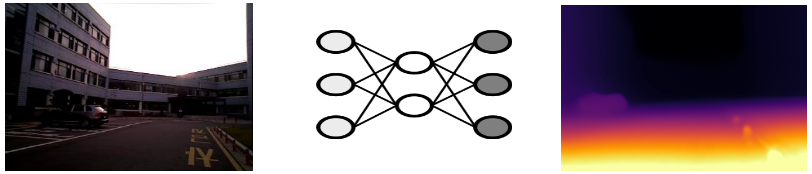
Sol: Self-supervised depth and pose estimation



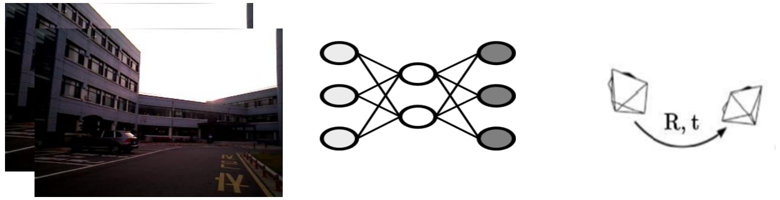
Motion parallax

Self-supervised learning of single-view depth map and multi-view pose estimation

: Networks learn depth map and relative pose that minimize motion parallax by camera in consecutive image frames.



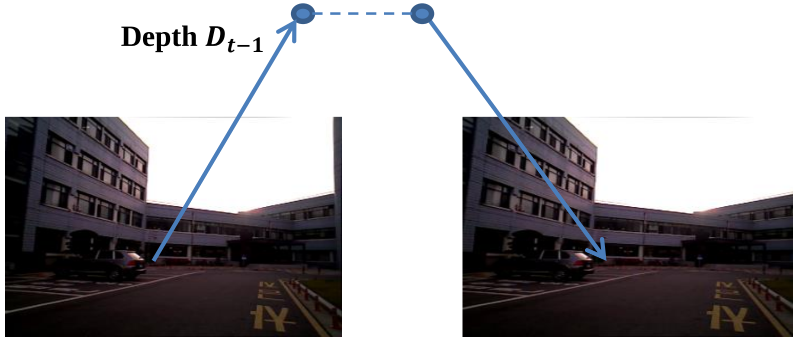
Single-view depth estimation



Two-view pose estimation

Relative camera pose P_t^{t+1}

Depth D_{t-1}



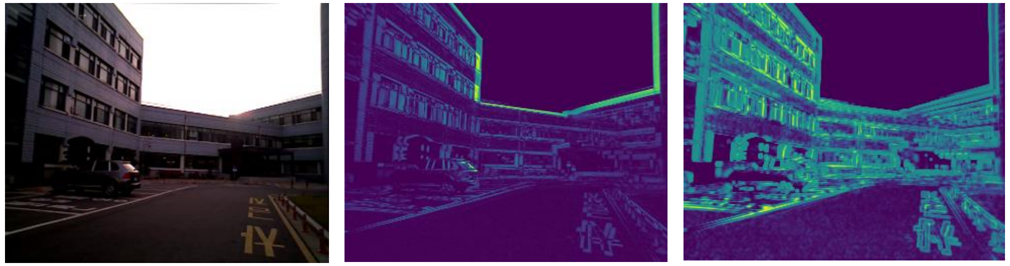
Time $t-1$ (I_{t-1})

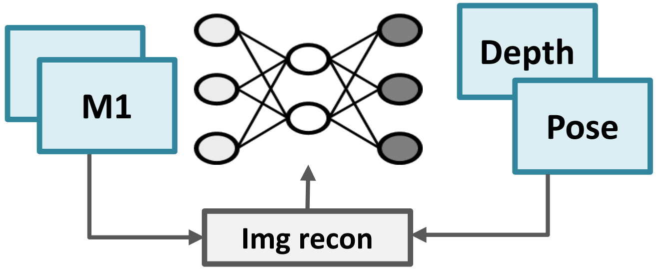
Time t (I_t)

Self-supervision from **Image reconstruction loss**

$$Loss = \lambda * L1(I_t, \tilde{I}_t) + (1-\lambda) * SSIM(I_t, \tilde{I}_t)$$

Reconstructed Img $\tilde{I}_t$ L1 SSIM



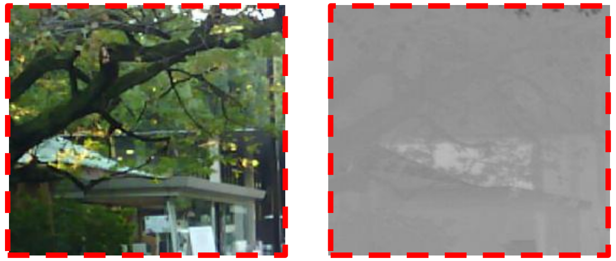


Problem: self-supervision from thermal image

Degeneration case : If images doesn't contain sufficient contents and details, supervision from image reconstruction process becomes near zero.

RGB images

Thermal images

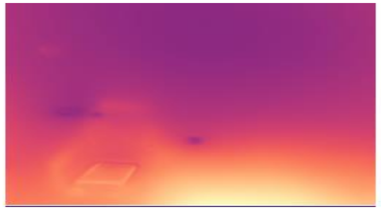


- low texture
- low contrast
- Homogenous region

RGB (time t) **RGB (time t+1)** **SSIM(I_t, I_{t+1})**

Thermal (time t) **Thermal (time t+1)** **SSIM(I_t, I_{t+1})**

insufficient self-supervision

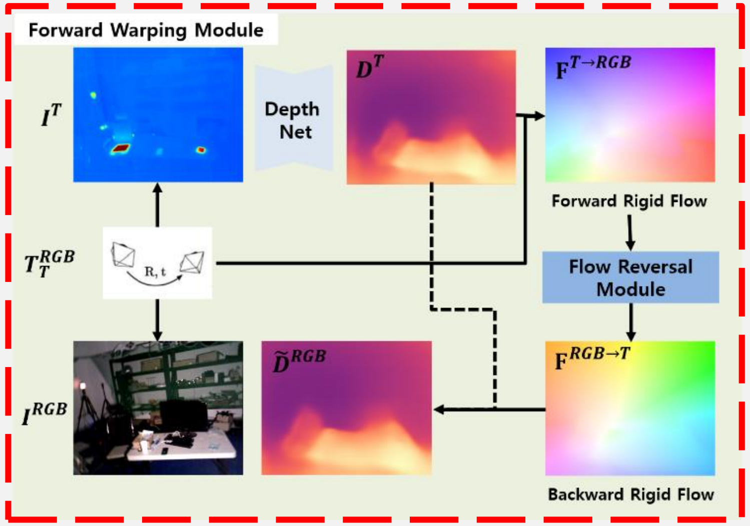


Depth from thermal image

Thermal image properties lead to weak self-supervision (image difference)

Self-supervised spatial perception from thermal image

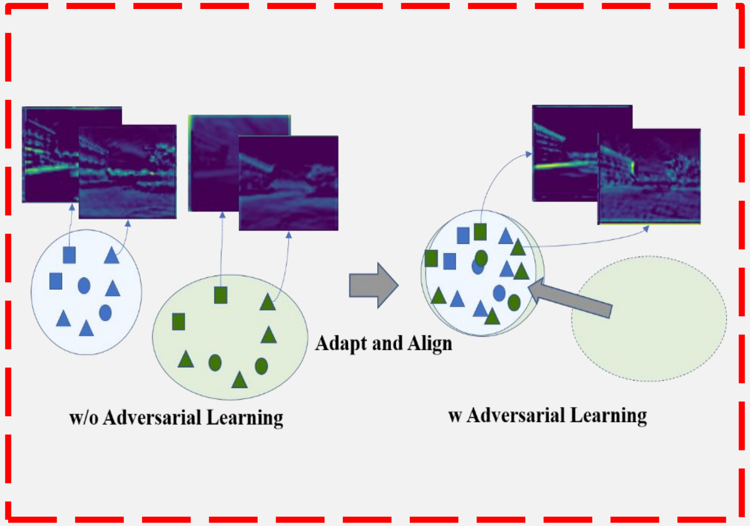
1. Self-supervision via camera geometry



Idea: Transfer self-supervision from **paired RGB images** via **camera geometric**.

*RAL-ICRA 21

2. Self-supervision via adversarial learning

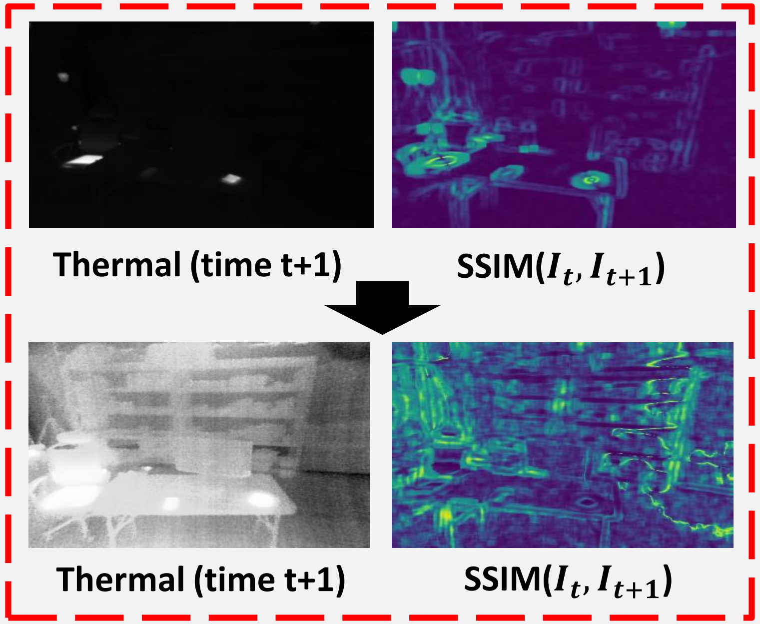


Idea: Transfer self-supervision from **unpaired RGB image** via **adversarial learning.**

ψ: discriminator

*WACV 23 (Best student paper), MVA23

3. Self-supervision via image conversion



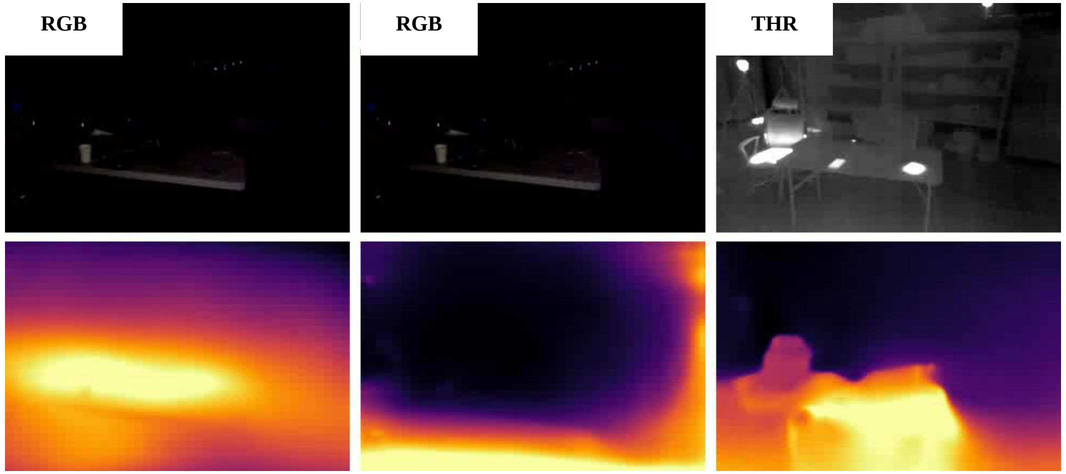
Idea: Maximize self-supervision from **thermal image** via **adaptive HE**.

*RAL-IROS 22

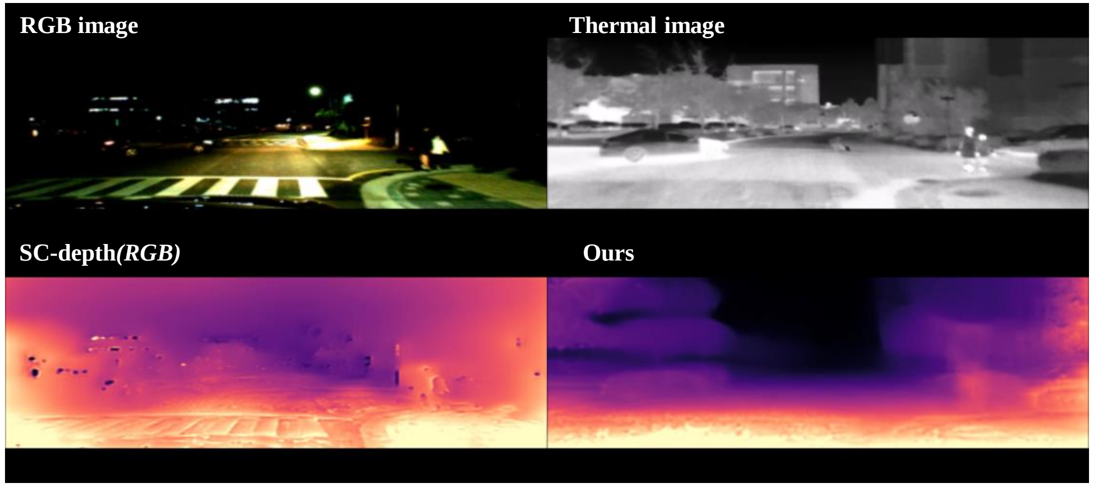
- Ukcheol Shin et al, “Self-supervised Depth and Ego-motion Estimation from Monocular Thermal Video using Multi-spectral Consistency Loss”, RA-L 2021 & ICRA 2022
- Ukcheol Shin et al, “Self-supervised Monocular Depth Estimation from Thermal Images via Adversarial Multi-spectral Adaptation”, WACV 2023 (Best Student Paper)
- Ukcheol Shin et al, “Maximizing Self-supervision from Thermal Image for Effective Self-supervised Learning of Depth and Ego-motion”, RA-L 2022 & IROS 2022

Self-supervised spatial perception from thermal image

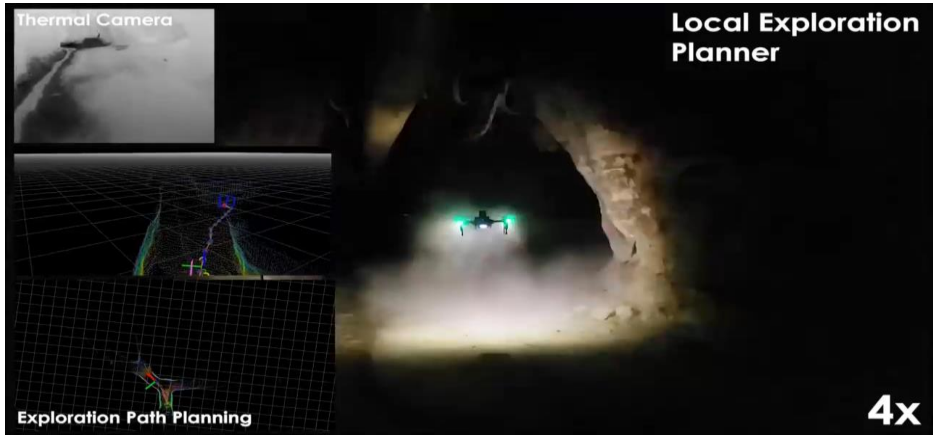
Scalable, Robust, and Self-supervised Spatial Perception in Hostile Weather, Lighting, Locational Conditions

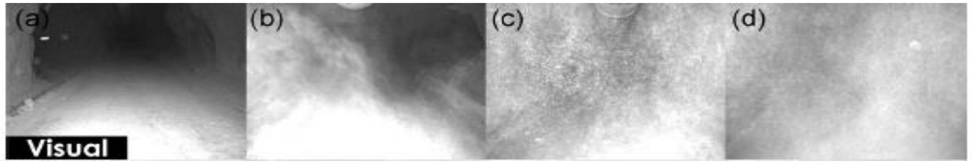


Midas v2.1 (RGB, Sup) | SC-depth (RGB, Self-sup) | Ours

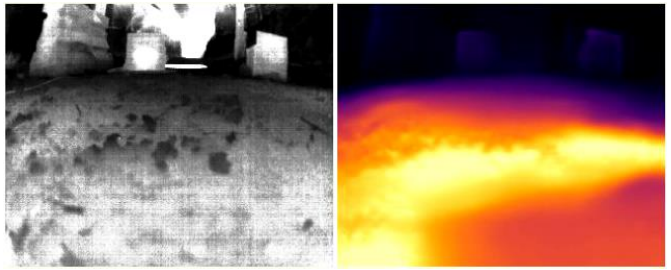


Campus (KAIST)_night [Vehicle, Outdoor]

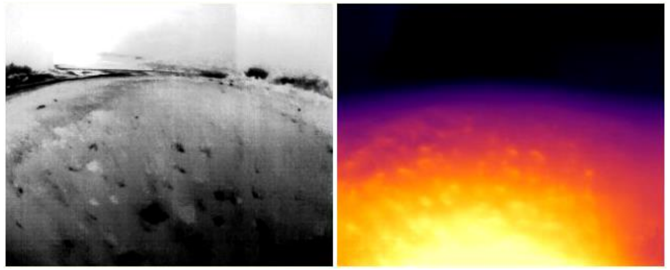




DARPA SubT challenge track
[UAV, Underground mine]



Underground mine Circuit A (Ours)



Underground mine Circuit B (Ours)

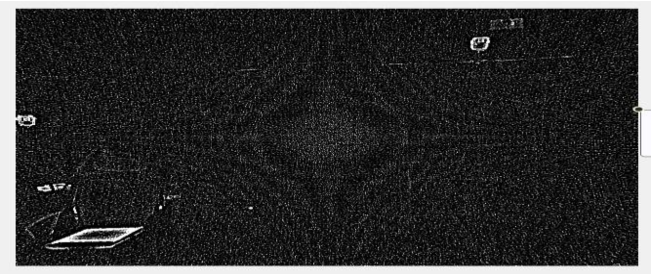
2. Thermal image enhancement

[Image quality] Disadvantages of thermal images: low-resolution, sensor noise, reflection issues
➔ They affects and degenerates (semantic/spatial) perception performance.

RGB, NIR: Higher than 2448x2048 px

Thermal: Lower than 640x512 px

Heat reflection



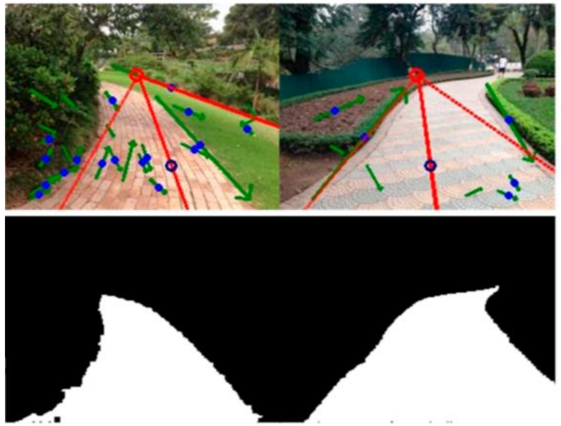
Fixed pattern noise

Potential research direction
➔ Super-resolution, denoising, colorization, contrast enhancement, RGB-thermal fusion

3. Traversable area detection in challenging conditions

[Traversable area detection] Detecting traversable area is vital for robotics and off-road vehicles

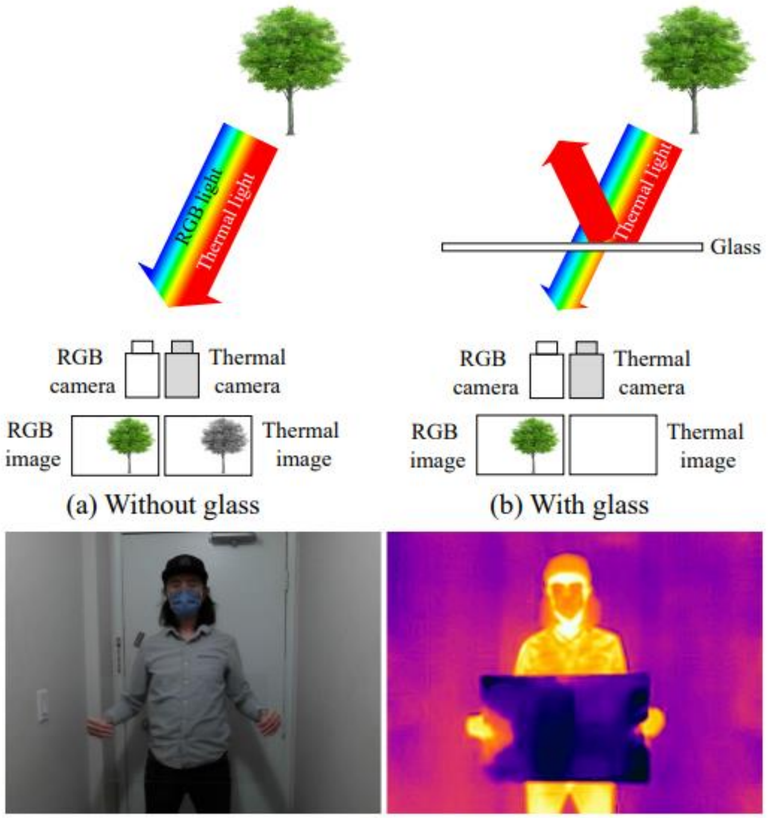
Traversable region detection with
- ✓ Geometric cue (vanishing point, ground plane detection, depth, etc)
- ✓ Semantic cue (semantic label)
- ✓ Temperature cue (black-ice detection, etc)



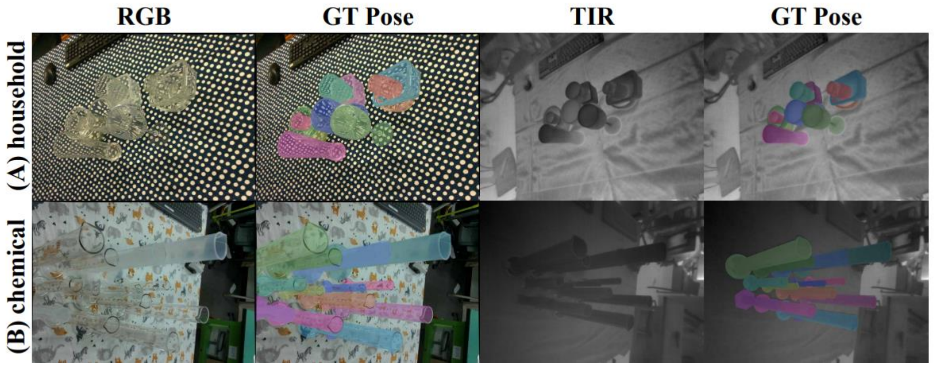
➔ Joint estimation of depth and traversable area from thermal image can bring high-level autonomy in field robotics

4. Detecting transparent objects

[Transparent object] Transparent objects (glass, window, bottles, etc) are challenging in RGB camera.



(c) A real-world pair of RGB (left) and thermal (right) images



Transparent objects cause erroneous prediction in
- ✓ semantic perception tasks
- ✓ geometric perception tasks

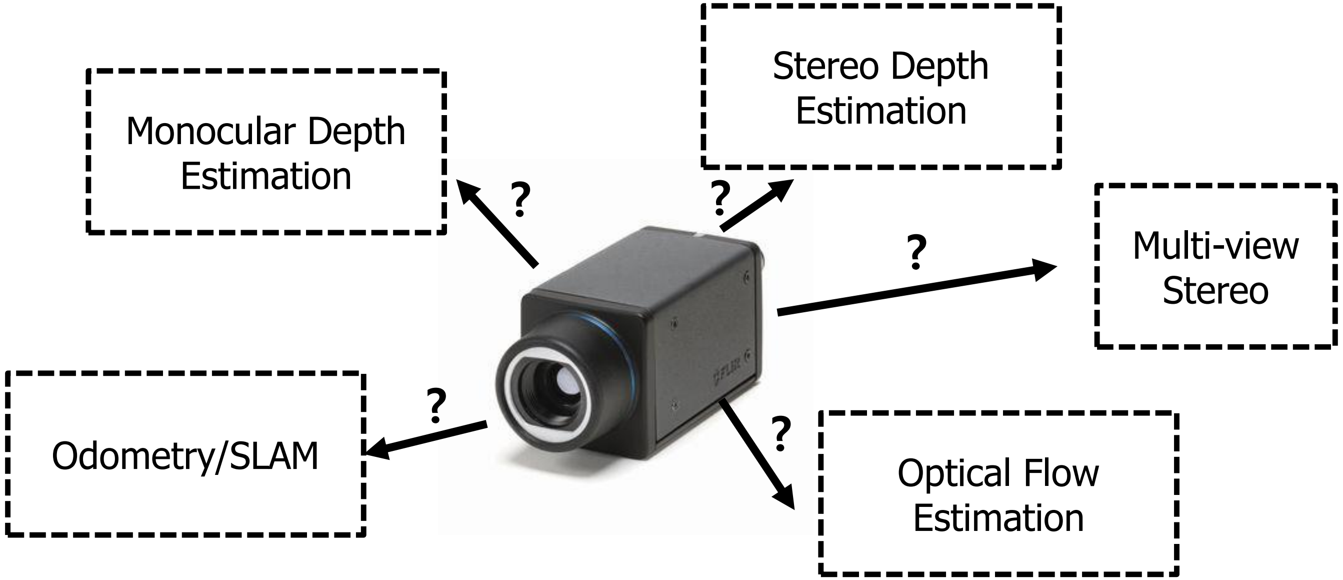
Potential research direction

➔ Transparent object grasping, 6D pose estimation, SLAM in indoor environment, detection & segmentation for transparent objects, etc.

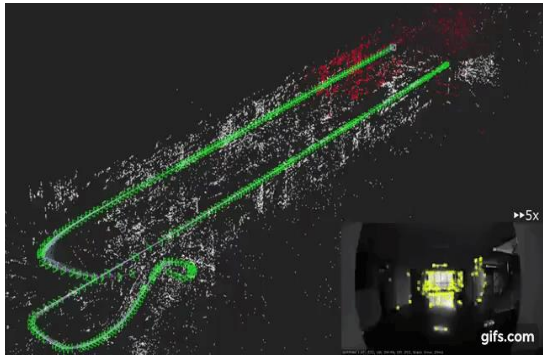
Huo, Dong, et al. "Glass segmentation with RGB-thermal image pairs." TIP 2023
Kim, Jeongyun, et al. "Transpose: Large-scale multispectral dataset for transparent object." IJRR 2024

5. Exploration on various spatial perception tasks

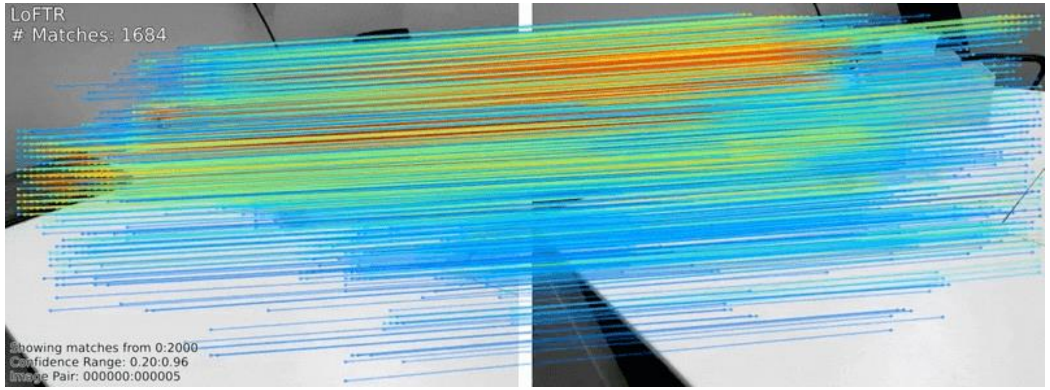
[Multi-view Geometry] Feature & descriptor, re-localization, optical flow, scene flow, visual odometry, SLAM, multi-view stereo, NeRF, etc



Optical flow



Visual odometry

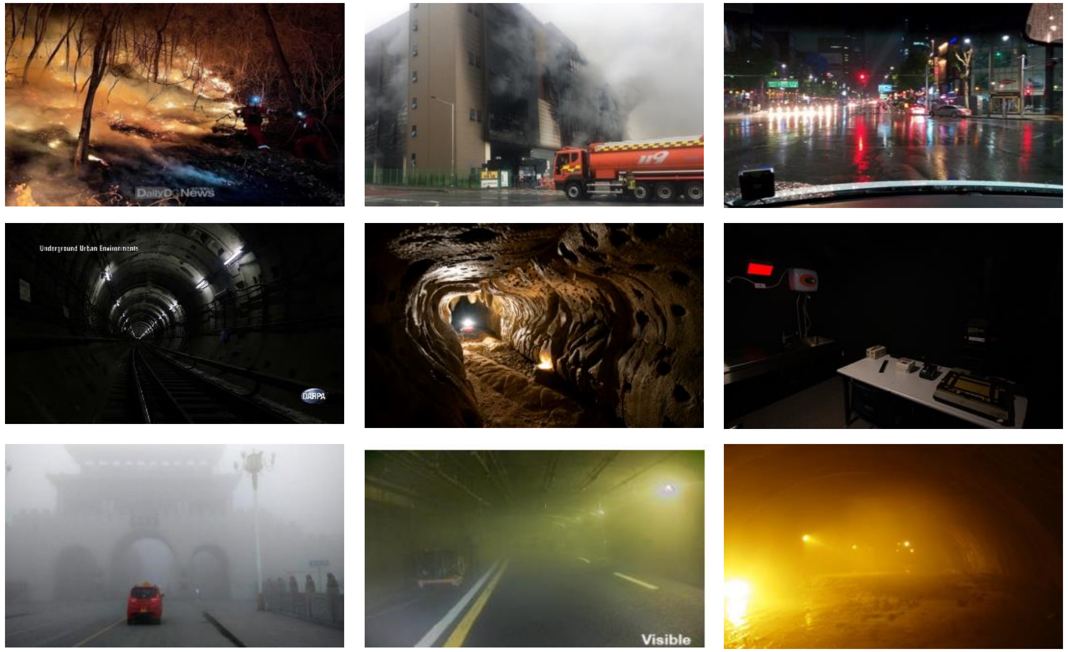


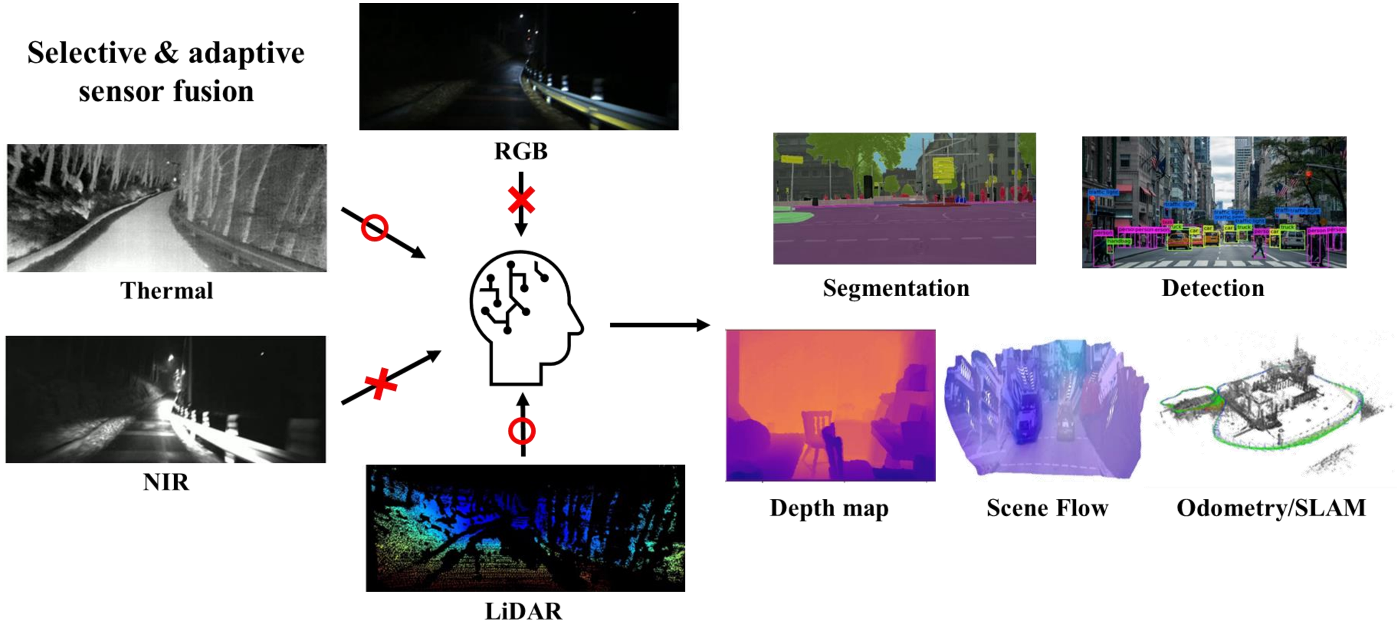
Feature matching

6. Selective sensor fusion in challenging conditions

[Sensor fusion] Thermal camera is not the one-fit-to-all solution.
➔ thermal camera is also degenerated by sensor noise, thermal homogeneity cases, reflective surface, malfunction, and non-uniformity correction (NUC), etc.

Various challenging scenarios (e.g., fire, fog, smoke)

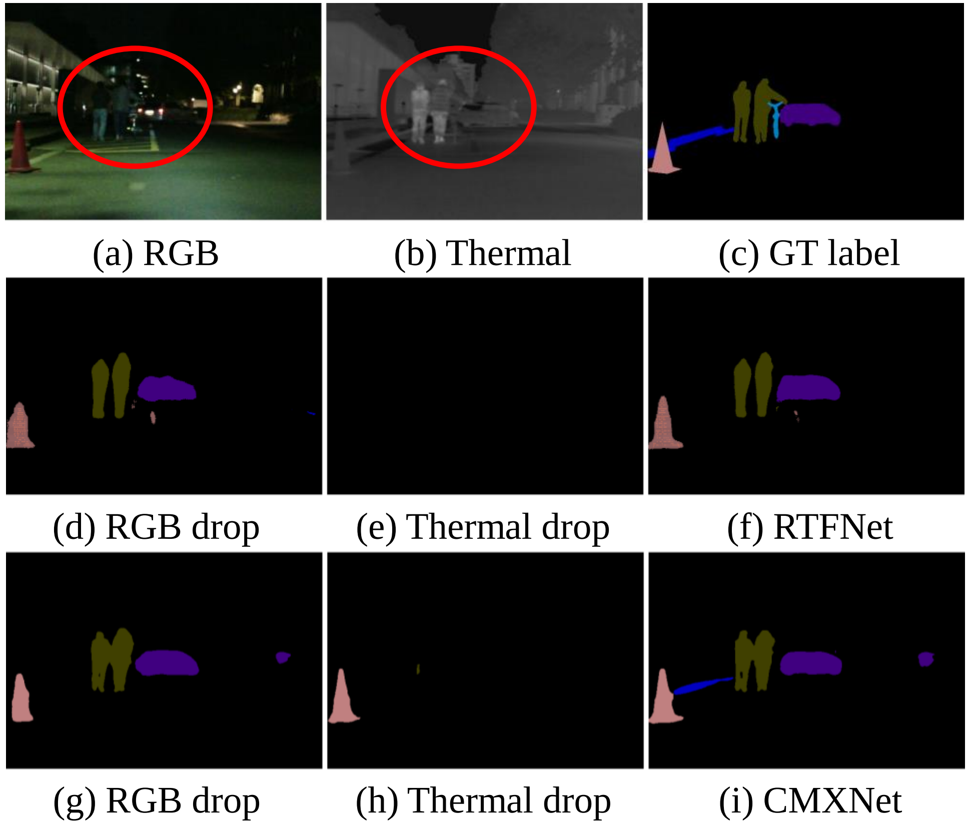




Potential research direction
➔ Shared representation learning between multiple sensors, selective sensor fusion, calibration, spatial alignment between sensors, etc.

7. Modality bias in multi-sensor fusion

[Modality bias problem] Naïve sensor fusion network bias toward one of modality.



Methods	RGB-T	RGB drop		THR drop	
	mIoU ↑	mIoU ↑	Diff ↓	mIoU ↑	Diff ↓
RTFNet [26]	53.2	45.6	-7.6	10.5	-42.7
CMXNet [16]	58.0	44.7	-13.3	39.2	-18.8
Ours	**61.2**	**53.1**	**-8.1**	**52.7**	**-8.5**

When one of modality is unavailable, the performance severely decrease.

➔ **Crucial issue for safety and robustness!**

Potential research direction

➔ Measuring uncertainty for each modality, attention mechanism, modality dropout, modality-balanced learning, etc.

Ukcheol Shin, et al, "Complementary Random Masking for RGB-Thermal Semantic Segmentation", ICRA 2024

Part 2. Takeaway message

[Take-home message]

- Self-supervision from thermal image enable scalable and label-free spatial perception in adverse weather/lighting conditions.

However, we have lots of unexplored part, disadvantages, and unique property of thermal cameras:

- [GT label] Investigate better form of supervision generation.
- [Image quality] resolve disadvantage of thermal images: low-resolution, noise, thermal homogeneity, …
- [Traversable area detection] Able to see traversable area in challenging environments.
- [Detecting transparent objects] Thermal image is effective for transparent objects.
- [Exploration] Needs extensive exploration in spatial perception tasks (odometry, scene flow, NeRF, …)
- [Selective sensor fusion] Thermal camera is not the one-fit-to-all solution.
- [Modality bias problem] Naïve sensor fusion network bias toward one of modality.

Conclusion

Intro. Visual perception in Robotics

- RGB camera/LiDAR are not best options in challenging conditions

Part 1. Spatial Perception from Thermal Image : Dataset and Benchmark

- Thermal camera is a potential rescue for robust spatial perception

Part 2. Visual perception from Thermal Image: Challenges

- Scalable and label-free geometric perception in adverse conditions
- What is next?

Research question

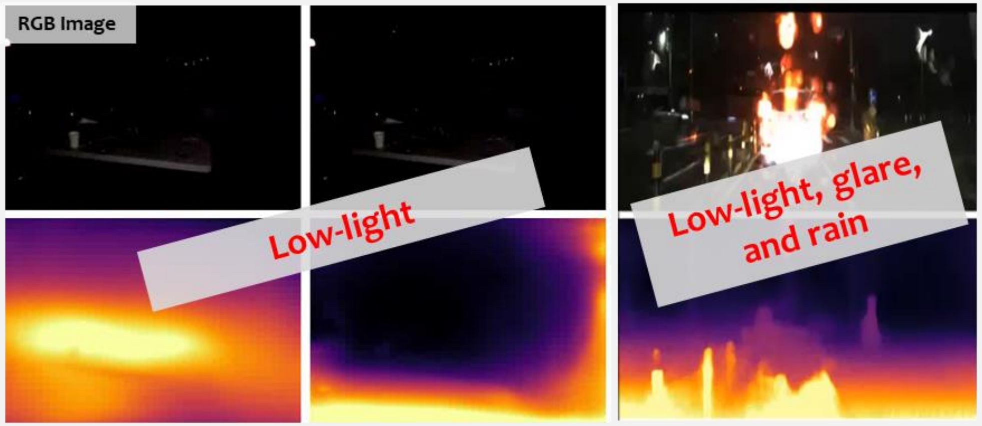
Q. Can we make AI have robust visual perception capability under challenging and hostile environments?

Intro. Takeaway message

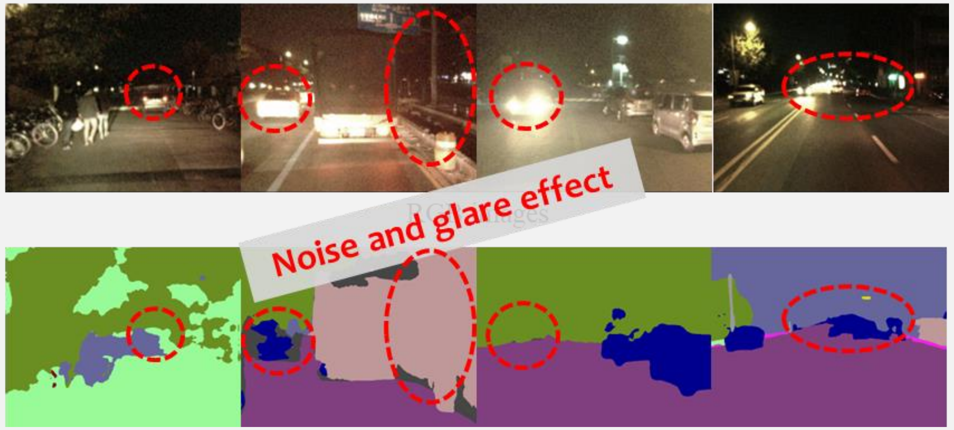
[Introduction] Limitation of visual perception from RGB/LiDAR

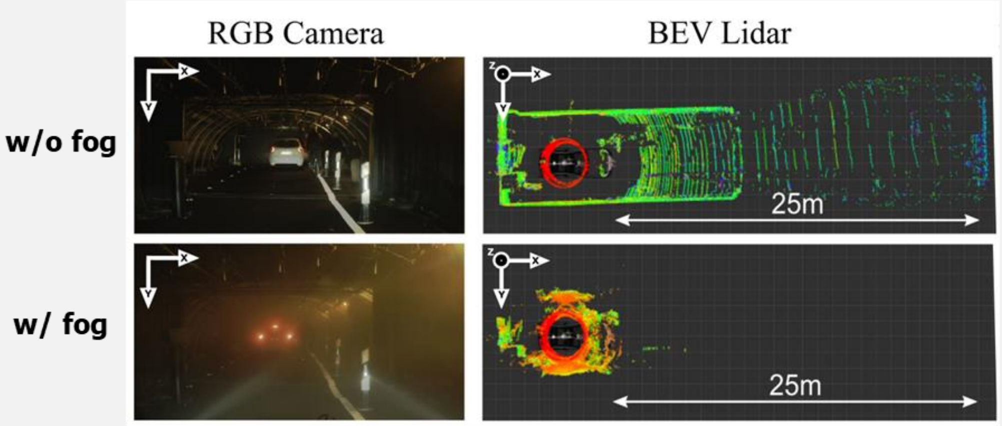
- **RGB camera/LiDAR are not best options in challenging conditions**

1. Monocular depth estimation (supervised/self-supervised)

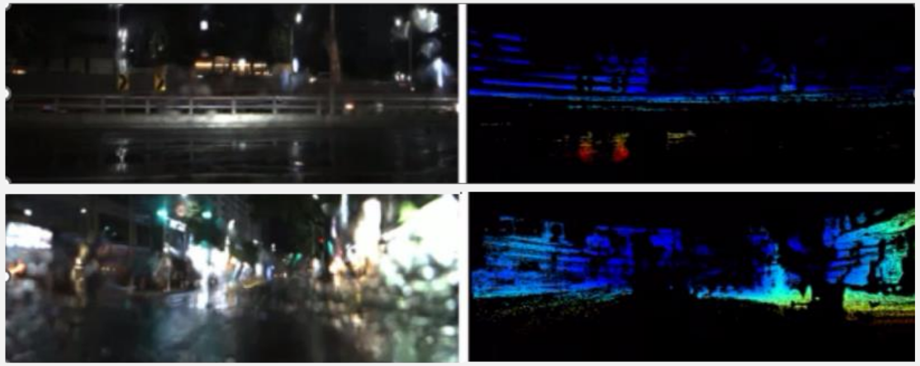


2. Semantic Segmentation (supervised)





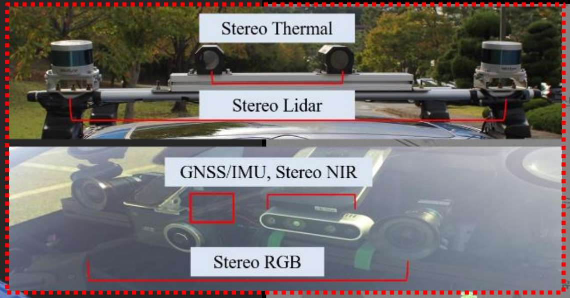
LiDAR in the fog

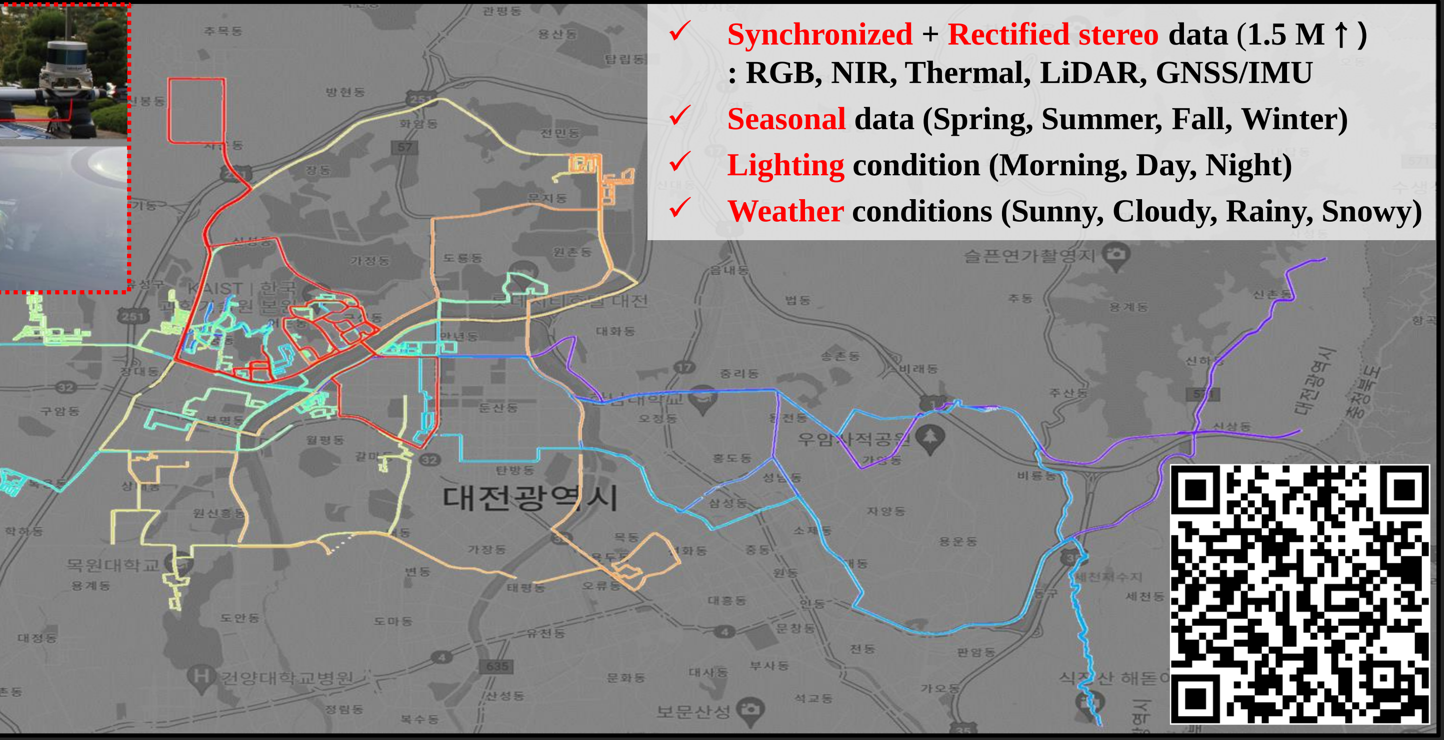


LiDAR in the rain

Multi-Spectral Stereo Seasonal (MS^3) Dataset

The first city-scale thermal stereo seasonal dataset





- ✓ Synchronized + Rectified stereo data (1.5 M ↑)
  : RGB, NIR, Thermal, LiDAR, GNSS/IMU
- ✓ Seasonal data (Spring, Summer, Fall, Winter)
- ✓ Lighting condition (Morning, Day, Night)
- ✓ Weather conditions (Sunny, Cloudy, Rainy, Snowy)

Seasonal changes

Rainy, snowy conditions

(from left) RGB, NIR, Thermal, Projected LiDAR

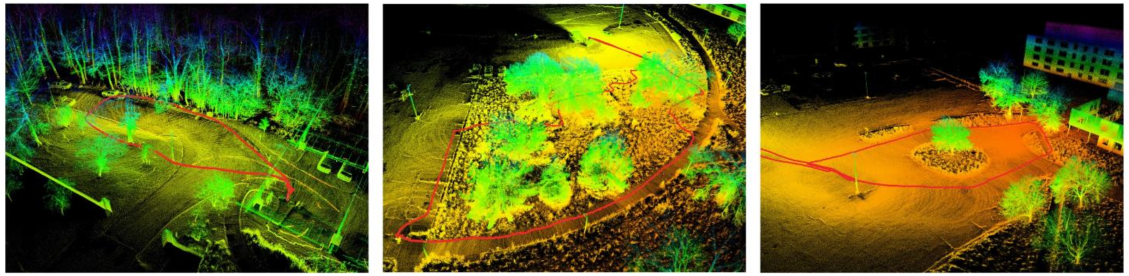
Extended version from Ukcheol Shin, et al, “Deep Depth Estimation from Thermal Image", CVPR 2023

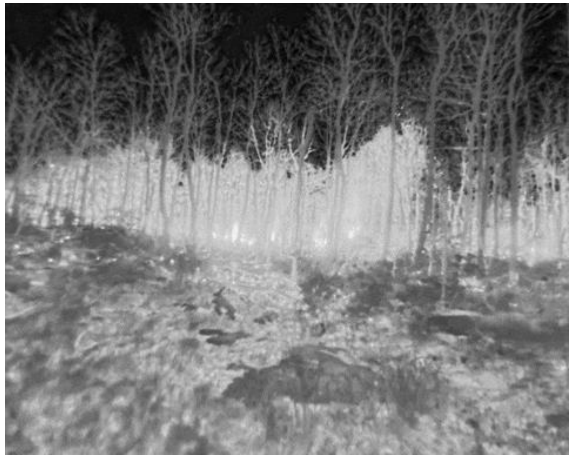
FIReStereo: Forest InfraRed Stereo Dataset

Carnegie Mellon University
The Robotics Institute

AIRLAB

The first thermal-stereo dataset in forest fire & smoke





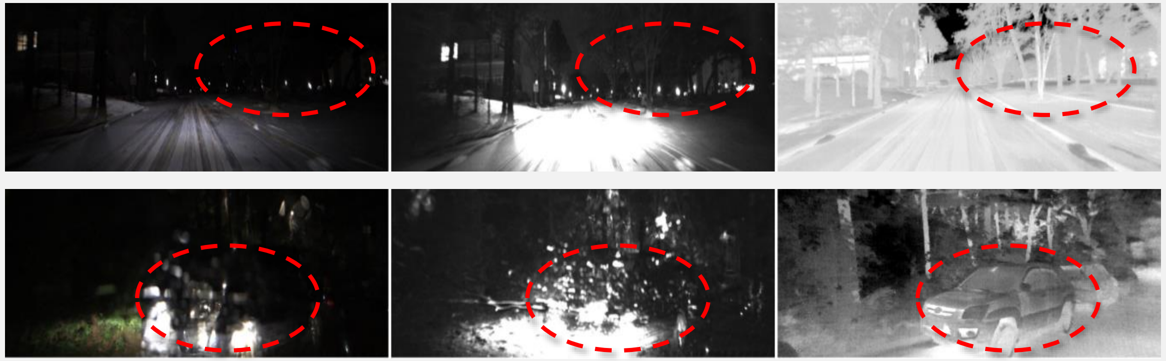
Devansh Dhrafan, et al, "FIReStereo: Forest InfraRed Stereo Dataset for UAS Depth Perception in Visually Degraded Environments", Under-review (U. Shin: Co-author)

Part 1. Takeaway message

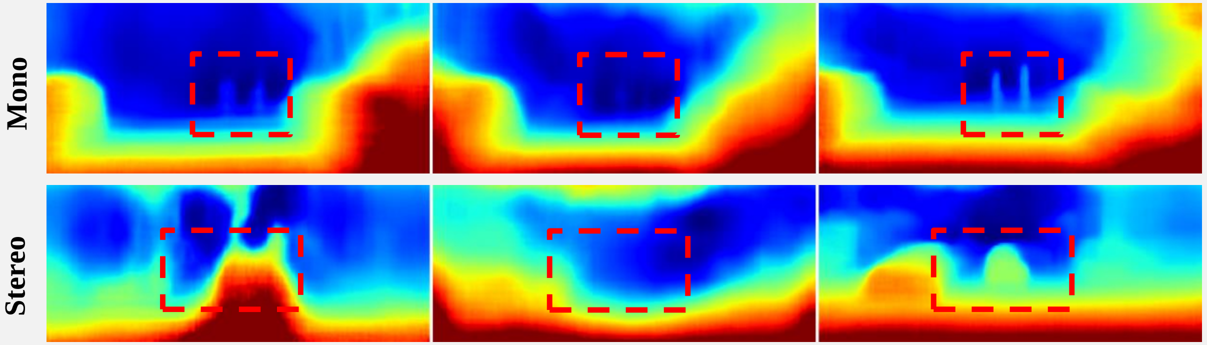
[Benchmark] Deep Depth Estimation from Thermal Image

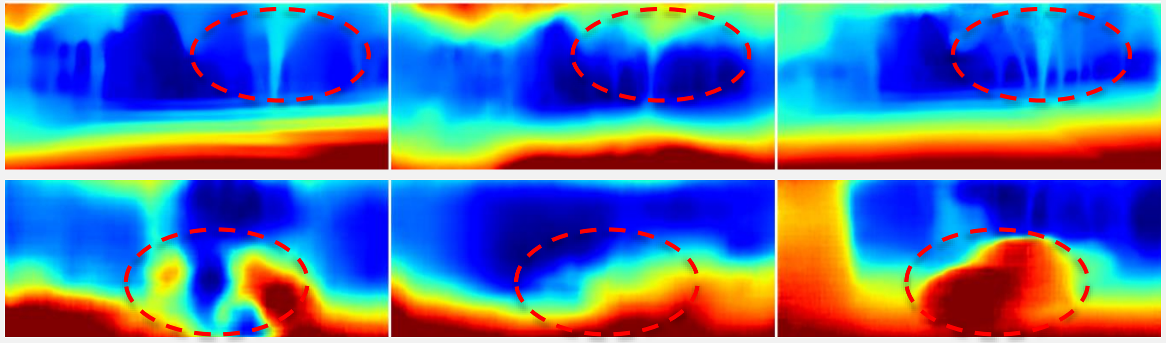
- **Thermal camera** is a potential rescue for **robust spatial perception in challenging conditions**

Unique information & Safety



Clean visibility against *low-light, snowy, rainy conditions*





Depth from thermal images shows the best accuracy, robustness, and generalization performance

Part 2. Takeaway message

[Take-home message]

- **Self-supervision from thermal image enable scalable and label-free spatial perception in adverse weather/lighting conditions.**

However, we have lots of unexplored part, disadvantages, and unique property of thermal cameras:

- **[GT label] Investigate better form of supervision generation.**
- **[Image quality] resolve disadvantage of thermal images: low-resolution, noise, thermal homogeneity, …**
- **[Traversable area detection] Able to see traversable area in challenging environments.**
- **[Detecting transparent objects] Thermal image is effective for transparent objects.**
- **[Exploration] Needs extensive exploration in spatial perception tasks (odometry, scene flow, NeRF, …)**
- **[Selective sensor fusion] Thermal camera is not the one-fit-to-all solution.**
- **[Modality bias problem] Naïve sensor fusion network bias toward one of modality.**

Q & A