# KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving

Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, *Member, IEEE*, and In So Kweon, *Member, IEEE*

***Abstract*— We introduce the KAIST multi-spectral data set, which covers a great range of drivable regions, from urban to residential, for autonomous systems. Our data set provides the different perspectives of the world captured in coarse time slots (day and night), in addition to fine time slots (sunrise, morning, afternoon, sunset, night, and dawn). For all-day perception of autonomous systems, we propose the use of a different spectral sensor, i.e., a thermal imaging camera. Toward this goal, we develop a multi-sensor platform, which supports the use of a co-aligned RGB/Thermal camera, RGB stereo, 3-D LiDAR, and inertial sensors (GPS/IMU) and a related calibration technique. We design a wide range of visual perception tasks including the object detection, drivable region detection, localization, image enhancement, depth estimation, and colorization using a single/multi-spectral approach. In this paper, we provide a description of our benchmark with the recording platform, data format, development toolkits, and lessons about the progress of capturing data sets.**

***Index Terms*— Dataset, advanced driver assistance system, autonomous driving, multi-spectral dataset in day and night, multi-spectral vehicle system, benchmarks, KAIST multi-sepctral.**

## I. Introduction

ALONG with the start of the fourth industrial revolution, the expectations of and interest in autonomous systems have increased. A great deal of effort has been devoted to reaching human-level reasoning of sensing, mapping, and driving policies, which are referred to as the three components of autonomous driving. Because data-driven AI-based methods have enabled breakthroughs in both academia and industry, large-scale benchmarks have become one of the most important factors to advance this technology. For autonomous driving and advanced driver assistance systems (ADAS), the KITTI [1] and Cityscapes [2] datasets have made it possible to push the performance of visual perception methods to previously inconceivable levels. However, most large-scale datasets are mainly based on RGB-based images and thus are only feasible in will-lit conditions as opposed to ill-lit environments such as nighttime, dawn, sunrises, and sunsets.

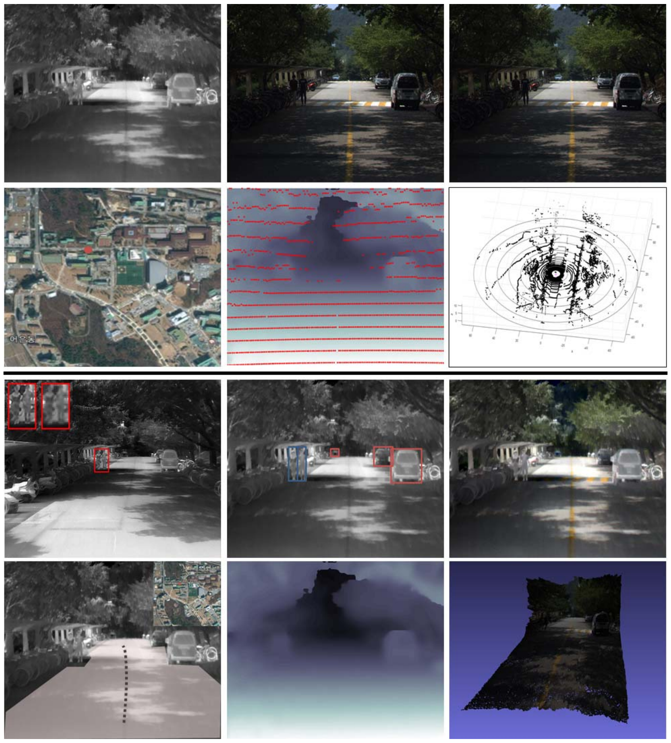


Fig. 1. **[Data & Tasks] (1st and 2nd rows)** The KAIST multi-spectral dataset for visual perception of autonomous driving in day and night. Dataset was repeatedly collected by the KAIST Multi-spectral All-day platform traversing in campus, urban and residential over several days. **(3rd and 4th rows)** The collected RGB stereo, thermal image, LiDAR and GPS data enable study into all-day vision problems such as image enhancement (red rectangles from top-left to top-right), pedestrian/vehicle detection, colorization, drivable region detection, driving path prediction with localization, dense depth estimation, 3D reconstruction.

To develop additional practical solutions, one of the main challenges is robustness to all-day conditions. For this purpose,

Manuscript received April 30, 2017; revised October 10, 2017 and December 5, 2017; accepted December 26, 2017. This work was supported by the Development of Autonomous Emergency Braking System for the Pedestrian Protection Project through the Ministry of Trade, Industry and Energy, South Korea, under Grant 10044775. The Associate Editor for this paper was Q. Wang. *(Corresponding author: In So Kweon.)*

Y. Choi, S. Hwang, and K. Park are with the Robotics and Computer Vision Laboratory, Department of Electrical Engineering, College of Information Science and Technology, Korea Advanced Institute of Science and Technology, Daejeon 307-701, South Korea (e-mail: ykchoi@rcv.kaist.ac.kr; smhwang@rcv.kaist.ac.kr; kbpark@rcv.kaist.ac.kr).

N. Kim is with NAVER LABS, Yongin 13494, South Korea (e-mail: namil.kim@naverlabs.com).

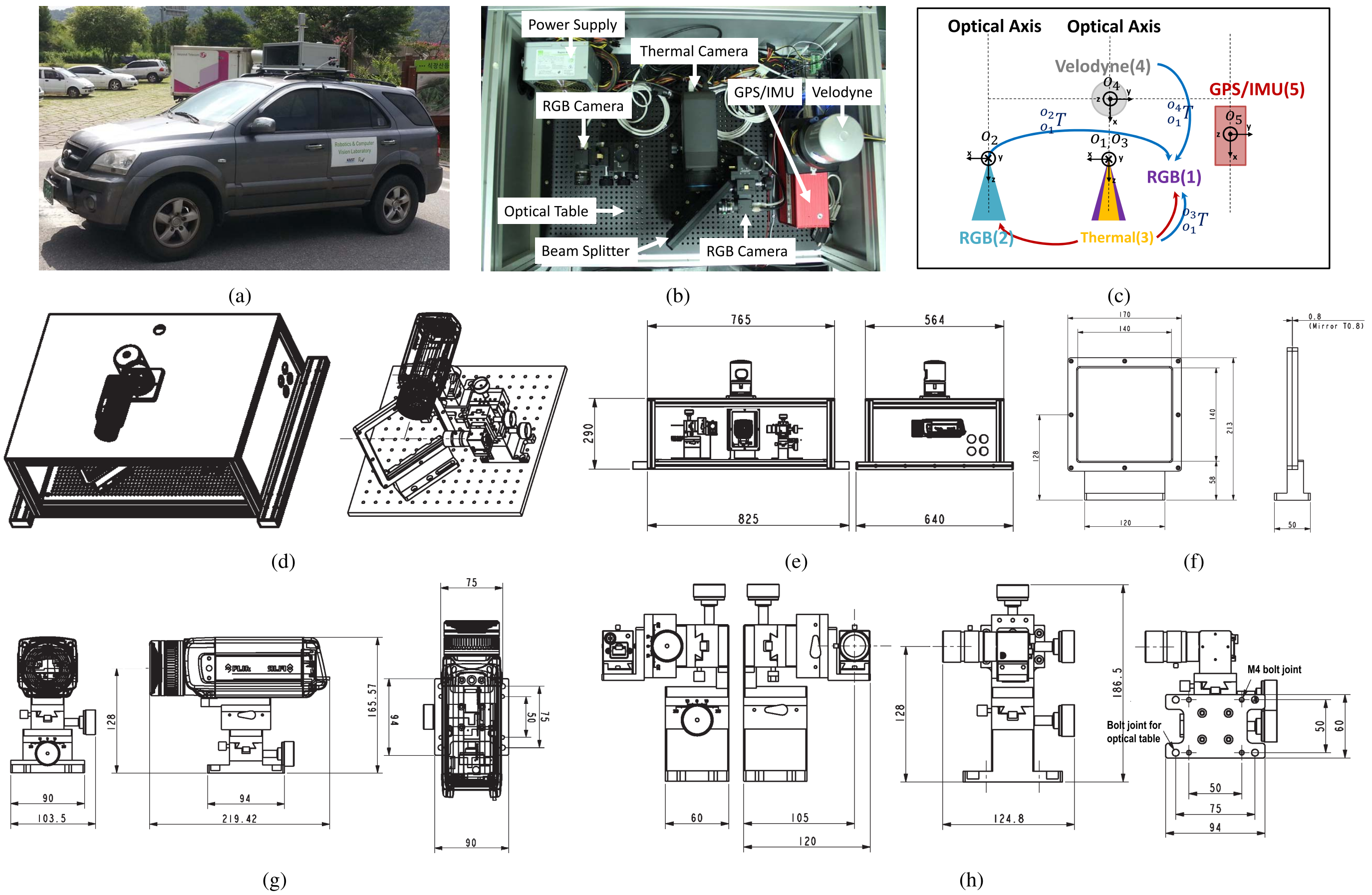
J. S. Yoon is with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: jsyoon4325@gmail.com).

K. An is with IT Convergence Technology Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 305-700, South Korea (e-mail: mobileguru@etri.re.kr).

I. S. Kweon is with the Department of Electrical Engineering, College of Information Science and Technology, Korea Advanced Institute of Science and Technology, Daejeon 307-701, South Korea (e-mail: iskweon77@kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2018.2791533



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 2. **Recording platform, sensor package, coordinates of sensors, and parts drawings.** (a) Our SUV is equipped with sensor pack on the roof. (b) The sensor package consists of two RGB and one thermal camera, one 3D LiDAR, and one integrated GPS/IMU device. All sensors are fixed on the optical table for the fine-level calibration. (c) All sensors in the figure are numbered in brackets. Blue lines indicate a coordinate transformation between sensors, and red lines indicate the connection of an external signal for the synchronization. In the calibration step, RGB(1) camera is used for the reference coordinate. (d) Figures show examples of the assembly drawing. (e)∼(h) All figures show details of parts and the unit of measurement is *mm*.

we created a new type of large-scale dataset covering various time slots in drivable areas. An example of data visualization in term of data & tasks is shown in Fig. 1. To do this, we used a thermal camera as a secondary vision sensor. Because a thermal sensor measures a long wave-length radiation emitted by subjects, it is highly advantageous for capturing scenes regardless of the amount of lighting. Compared to previous multi-spectral datasets [3], [4], we captured fully registered RGB and thermal images through a special optical device known as a *beam splitter*, which can reflect the visible spectrum (RGB) and transmit the thermal spectrum. Therefore, our multi-spectral image pairs undergo no loss of the rectified distortions caused by the multi-spectral stereo-based method, and they maintain the full resolution image. To this end, we designed a new multi-spectral recording platform which supports a co-aligned RGB/Thermal camera, RGB stereo, 3D LiDAR and inertial sensors (GPS/IMU) with calibration and synchronization techniques.

We gathered the dataset over an extended duration in August of 2015 using a roof-mounted recording platform on a sport-utility vehicle (SUV). Our dataset contains general traffic situations with many static/dynamic objects in urban, campus and residential areas. Essentially, we captured all scenarios at well-lit and ill-lit times (day and night) and collected fine time slots (sunrise, morning, afternoon, sunset, night and dawn) on the campus. Compared to other places, a campus is suitable for capturing similar perspectives according to illumination changes. Each frame is tagged with high-precision GPS measurements, (GPS combined) IMU accelerations and object annotations including the type, size, location and occlusion level. Based on multi-spectral data in all-day conditions, we designed various subset benchmarks for visual perception tasks, such as object detection, the drivable region detection, image enhancement, depth estimation, and colorization. In this paper, we provide a detailed description to help readers exploit our dataset and reproduce recording platform with the development toolkits.

The remainder of the paper proceeds as follows: we introduce our hardware configuration as used for data acquisition in section II. At the same time, we discuss several issues of importance when capturing high-quality data. The details of multi-spectral data and ground truth supported by this dataset are described in section III. We then, explain how to calibrate multiple sensors section IV. In section V, first we present development tools and explain how to operate them with a short summary, after which we introduce various benchmarks for visual perception tasks. In section VI and VII, we explain how our dataset differs from other datasets, and describe what we learned while building and operating our system. Lastly, we summarize the advantages of our dataset and suggest directions for future research.

## II. KAIST All-Day Platform

Our recording platform and hardware configuration are illustrated in Fig. 2. We designed the multi-spectral recording platform so that it holds all sensors on the optical table tightly, which makes fine-level calibration more efficient. We mounted our capturing system on top of a vehicle and housed a PC (Intel i7-4980K processor) in the trunk. Because we needed to maintain a transfer speed of at least 200MB/s for stable and rapid data collection, we used a Samsung pro-850 SSD (Solid-State Drive) which can secure sufficient bandwidth compared to a HDD. Additionally, to minimize write latency into storage, we attempted to keep memory usage under 80% of the total capacity. Moreover, we used one Giga-E card and one CAT7 cable in each sensor to prevent losses and to lower the transfer latency of the sensor data. Our platform in Fig. 2. (b) is equipped with the following sensors:

- **2**× PointGrey Flea3 RGB camera (FL3-GE-13S2C-C), 1.3 Mega-pixels, 1/3” Sony ICX445 CCD, 1280 × 960, 400 ∼ 750nm, GigE, Computar Optics 12*mm* Lens, 26°(H) × 22.1°(V)
- **1**× FLIR A655Sc thermal camera, ∼50Hz, 14bits data, 640 × 480, 17um detector pitch, 7.5 ∼ 14um, GigE, 25°(H) × 19°(V) with 24.3*mm* lens
- **1**× Velodyne HDL-32E 3D LiDAR, 10Hz, 32beams, 0.16 degree angular resolution, 2cm distance accuracy, collecting 0.7 million points/second, field of view: 360° horizontal, 41.34° vertical, range: ∼70m
- **1**× OXTS RT2002 inertial and GPS navigation system, 6 axis, 100Hz, L1/L2 RTK, resolution: 0.02m/0.1°

One of the main contributions of our system is capturing fully aligned RGB and thermal images simultaneously without a loss of the geometric distortion. This is made possible because of the optical device called *beam splitter*, which made of zinc oxide and silicon. Due to the use of a special coating, it can reflect RGB wavelengths (380nm ∼ 700nm) and transmit long-wave infrared light (8um ∼ 15um). Thus, with this device, we can achieve a parallax-free pair of RGB/thermal image with the proposed calibration technique. The details of proposed the calibration are given in section IV. With this procedure, we can create a fully aligned image simply by rectifying a middle-resolution thermal image into a high-resolution RGB(1) image. Note that the beam splitter should be carefully fastened onto an optical table to prevent bending, swaying, and vibration, as these factors can cause unintended image distortion.

We used an *A655sc camera*, which has a long-wavelength infrared (LWIR) sensor. The thermal sensor has the advantage of having less an effect on illumination changes, as shown in Fig. 4, and they have been used as night vision sensors in many academic and industrial applications [5]–[7]. One of the main challenges when handling a thermal sensor is the temporal change of the thermal contrast ratio according to the amount of heating energy. To handle such changes, various methods have long been studied. In our dataset, we provide a raw data format for thermal measurements to guarantee a proper level of user selection regarding how to deal with this in the preprocessing stage.

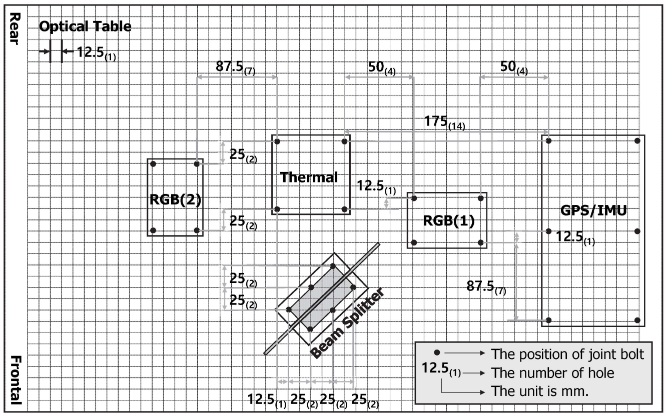


Fig. 3. **The top-view of the sensor configuration on the optical table.** The upper number is an actual measurement (*mm*) and the lower number is the number of the hole between devices.

We selected a *long focal lens*, which is favorable for obtaining a sufficient safety distance between driving cars and dynamic objects in actual traffic environments. At this time, the major consideration is that the RGB and thermal lens must be set in a similar field of view to guarantee the maximum overlapping region between both cameras.

*Optical table* and *camera jigs* are designed for the fine adjustment of each camera. The interval of holes in the optical table is 12.5*mm*, and the adjustment range of each axis of a camera jig is from −20*mm* to 20*mm*. The details of jigs installed on the optical table are shown in Fig. 3. The jig of the RGB camera has three axes ($x$, $y$, $z$) while that of the thermal camera has two axes ($x$, $z$). The $y$-axis of the thermal camera jig is excluded to prevent screws from becoming loosened due to the weight of the sensor, or the vibration of the driving car.

We employed high-accuracy *hardware-synchronization* using an external trigger, which is the reference signal of the thermal camera. Along with the rising pulse of the reference signal, all data from the sensors are captured. The details of the multi-sensor synchronization process are given in the following table.

- Trigger generator (Master: A655sc-Thermal camera)

  Trigger signal: 12V pulse with 25Hz
  Register: Vertical Sync
  GPIO: signal #3, power vcc #5, gnd #6

- Trigger receiver (Slave: Flea3-RGB camera)

  Mode: Overlapped Exposure/Readout(Mode14)
  GPIO: signal #3, power vcc #7, gnd #5

To obtain high quality images, it was necessary to control the exposure and shutter speed during the capturing step. Therefore, we set these values shorter than the sensor synchronization period (<40*ms*). Unlike cameras, Velodyne and GPS/IMU sensors are synchronized by software because these sensors do not support external triggering. To inte-

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

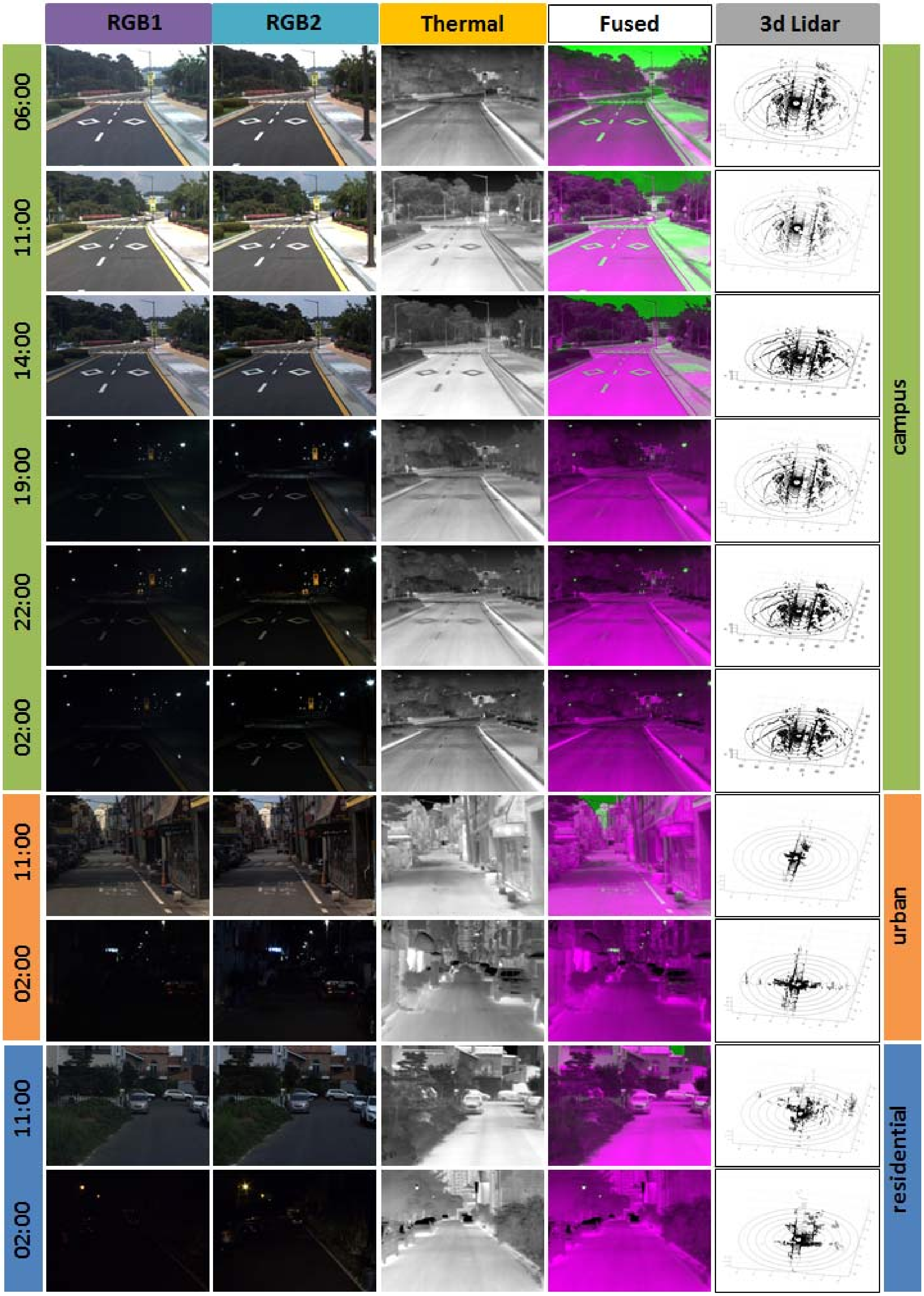


Fig. 4. **Examples from KAIST multi-spectral dataset** in day and night. From left to right, we show the RGB(1), RGB(2), thermal, fused images to overlay the thermal to RGB(1) images, and 3D points cloud. Compared to RGB sensors, the thermal sensor is advantageous in extreme lighting condition and can therefore be used both during the day and at night. They can also be used in special conditions such as foggy or otherwise poor weather. Compared to NIR sensors, the LWIR (thermal) type is not affected by headlights (known as the blooming effect), thus allowing their users to avoid dynamic objects in nighttime driving environments.

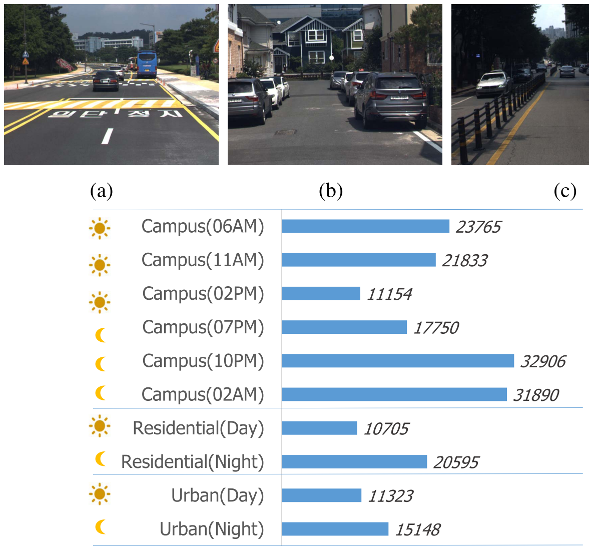


Fig. 5. **Statistics of captured images in various scenarios.** There is a statistical difference in the dataset for the same route depending on the driving style. (a) Campus. (b) Residential. (c) Urban.

grate hardware and software synchronization into our system, we adjusted the acquisition speed of software-based devices to that of hardware-based devices. In this way, we updated partially scanned LiDAR data and sampled the position information (GPS/IMU) using a timestamp to minimize a synchronization slip between devices with different cycles.

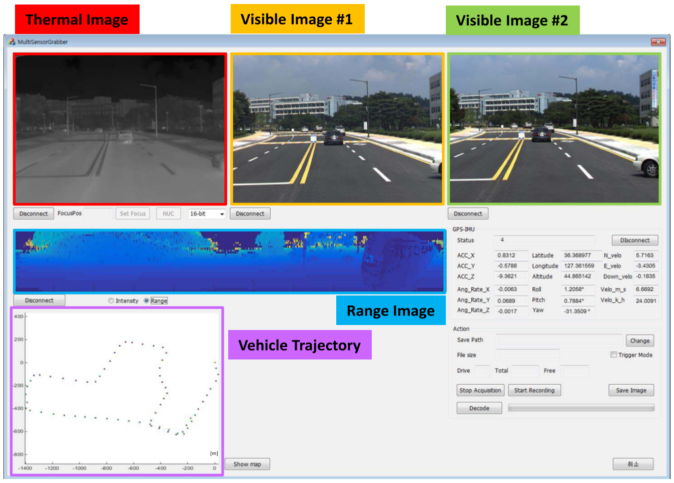
## III. Dataset Specifications

Our contribution is that our dataset contains large-scale multi-spectral sequences in various time conditions. An example of such a contribution is shown in Fig. 4, which was captured at the same location during a day and night. In this section, we provide a description of our benchmark with details of each modality from capturing to saving, grabbing, and annotating.

### *A. Multi-Modal Data*

*1) General Information:* Prior to capturing data, we manually tuned the shutter speed and exposure time of the RGB and thermal camera considering the surrounding environments and time of day. In addition, because the type of thermal sensor used is an uncooled camera, we utilized non-uniformity compensation (NUC) to ensure the image quality. By covering various regions such as *campus, city, and residential* areas, we attempted to prevent our dataset from becoming biased to a specific location. We principally captured each location in day and night conditions and also collected *campus* data in fine-time slots (sunrise, morning, afternoon, sunset, night and dawn). Compared to other locations, the *campus* is a superior location for capturing similar environments at different time slots without an interruptions by traffic, and the detection target can be properly balanced. The datasets for the locations used here are shown in Fig. 5.

*2) Images:* Image data is stored in raw form without any processing to minimize the write latency, and RGB and thermal images are respectively logged as 8-bit images on three channels and as 16-bit images on one channel. For better visualization of a thermal image which supports a high-dynamic resolution (HDR), post-processing steps such as histogram equalization can be employed. However, we simply logged untreated data considering expandability by the user. Note that because the reflectivity of the beam splitter is not perfect, RGB (1) and RGB (2) through the beam splitter have slightly different intensity levels in some sequences. Captured RGB and thermal images have a great advantage with regard to detecting pedestrians and vehicles that are farther away because we used a long focal lens.



| | | | | | |
|---|---|---|---|---|---|
| Status | 4 | | | | |
| ACC_X | 0.8312 | Latitude | 36.368977 | N_velo | 5.7163 |
| ACC_Y | -0.5788 | Longitude | 127.361559 | E_velo | -3.4305 |
| ACC_Z | -9.3621 | Altitude | 44.865142 | Down_velo | -0.1835 |
| Ang_Rate_X | -0.0063 | Roll | 1.2058° | Velo_m_s | 6.6692 |
| Ang_Rate_Y | 0.0689 | Pitch | 0.7884° | Velo_k_h | 24.0091 |
| Ang_Rate_Z | -0.0017 | Yaw | -31.3509 ° | | |
| Save Path | | | | | |
| File size | | | | | |
| Drive | | Total | | Free | |

Fig. 6. **Grabber** is a tool for real-time logging and visualization of incoming data from multiple sensors.

*3) GPS/IMU:* The inertial data is synchronized and logged along with other sensor data (25Hz). For every frame, we provide 19 different GPS/IMU values: global positions including the altitude, acceleration, angular rates, velocities, global orientation and signal quality information, among other types. Mercator projection is used to draw geographic coordinates based on the global position on a 2D google map, and related toolkits are also provided.

*4) Velodyne:* We provide range and reflectance images generated from Velodyne in sync with 2D vision sensors. For synchronization between devices with different cycles, it is common to operate on a slow cycle. However, our dataset is intended to collect data at high speeds. Therefore, for sensors having a slow cycle, the data is partially updated instead of being fully updated at every frame. In other words, we partially update only 144 degrees out of 360 degrees (40%) in time. Range and reflectance images are saved in the 16-bit/one channel and 8-bit/one channel formats at the same size, respectively. The size of a range image is $32 \times 12 \times 180$, the number of 3D points in each frame is approximately 0.7 million. To convert the image to 3D points, we also provide the individual angular information of all vertical scan lines.

### *B. Grabber and Ground Truth*

*1) Grabber:* As introduced in Fig. 6, our grabber has two roles: the logging and visualization of incoming data. Data is managed as a binary file, and the recording structure is shown in Table I. The most important role of the grabber is to

TABLE I

Overview of the Data Format in a Binary Video. The Length of One Data Cycle Is 8,199,184bytes, and This Includes Five Sensors Data Captured at the Same Time. Each Symbol H and D Indicate Header and Data. More Details Are in Info_Packet.txt

| | Sensors | Type | Length | Type | Description |
|---|---|---|---|---|---|
| | | H | 48 | mixed | packet header |
| Repeated Packet Format (/cycle) | RGB Splitter | H | 8 | int64 | timestamp |
| | | | 4 | uint | length |
| | | D | 1280x960x(3) | uchar | image (8bit,3ch) |
| | RGB Stereo | H | 8 | int64 | timestamp |
| | | | 4 | uint | length |
| | | D | 1280x960x(3) | uchar | image (8bit,3ch) |
| | Thermal Splitter | H | 8 | int64 | timestamp |
| | | | 4 | uint | reserved |
| | | | 4 | uint | length |
| | | D | 640x480x(2) | uchar | image (16bit,1ch) |
| | Velodyne | H | 8 | int64 | timestamp |
| | | D | 12x180x(2) | | horizontal angle (16bit) |
| | | | 32x12x180x(2) | uchar | range data (16bit,1ch) |
| | | | 32x12x180x(1) | | reflectance data (8bit,1ch) |
| | GPS IMU | H | 8 | int64 | timestamp |
| | | D | 248 | mixed | various data |

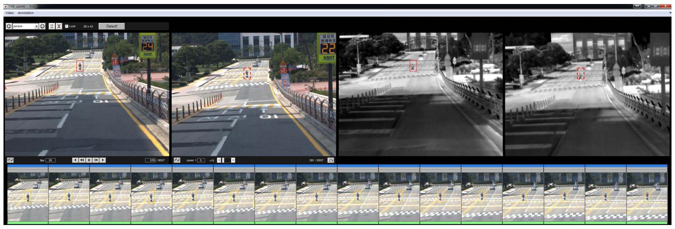
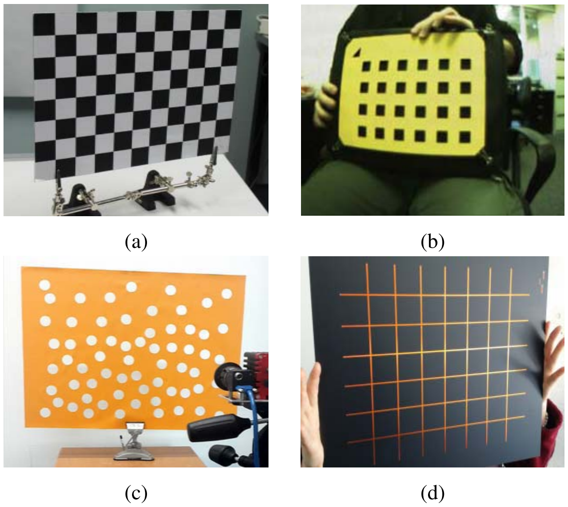


Fig. 7. With this **Annotator**, users can efficiently annotate all-day sequences contained in RGB and thermal images.

record data in keeping with the acquisition signal without any jitter or latency. If there is a GigE problem with the network transmission of certain sensor data, the incoming data is then discarded by the grabber. Additionally, the grabber uses a visualization tool to alert users if the incoming data is in an abnormal state. For example, if DGPS is unable to guarantee the number of receiving satellites due to the heights of the surrounding buildings or the weather, it will send a warning message about the signal quality.

*2) Annotation of a Moving Object:* For various perception tasks, we manually labeled the dynamic objects of all sequences. Before annotating labels in moving objects, we defined the annotation targets and types, referring to previous datasets [1], [2]. At this time, individual targets and groups are separately managed, and if targets are difficult to recognize, they are excluded as evaluation targets by annotating then with the term *ignore*. Finally, while drawing bounding boxes (BBs), users can select one of the following labels: *person, car, cyclist, people, cars, or ignore*. The annotation toolbox (Fig. 7) for the process described above was created by modifying and supplementing a toolbox [8] distributed by Piotr. The modified toolbox uses both of RGB and thermal images to obtain tight and accurate BBs in poor lighting conditions or in the event of the halo effect.



(a) (b) (c) (d)

Fig. 8. **Pattern boards** for multi-spectral camera calibration. (a) Chessboard [10]. (b) Grid-board [11]. (c) Circle-board [12]. (d) Line-board (Ours).

*3) Dense Depth Map:* The depth data can be obtained by a depth sensor or stereo-type camera, and each method has its strengths and weaknesses. Velodyne-32E is a good range sensor to measure physical distances; however, it is insufficient to provide high-density depths such as those obtained by a stereo. Unlike the Velodyne device, a RGB pair (stereo-type cameras) can give a dense depth map, but the quality of the depth is not better than that of a range sensor. Furthermore, at night and under ill-lit conditions, it is difficult to provide reasonable depth values. Thus, we made an effort to provide high-density and accurate depth data in day time conditions with the two aforementioned approaches. In other words, we used the RGB patch-matching-based stereo method [9] combined with 3D point-based refinement. To obtain high-density depths even under ill-lit conditions, such as those in good lighting conditions, we are undertaking the reconstruction of 3D world as sparse 3D points combined with GPS/IMU. In the near future, we hope to release more accurate depth information for both day and night time.

## IV. Calibration

In this section, we describe multi-modal calibration methods between the RGB and thermal cameras (*cam-to-cam*) and for the RGB and Velodyne device (*cam-to-LiDAR*). To obtain precisely aligned multi-modal data, the calibration must be taken when computing the intrinsic parameters of each sensors and extrinsic parameters between sensors. The overall coordinates of the sensor calibration are illustrated in Fig. 2. (c). We define the RGB (1) as the reference camera coordinate. Because the RGB (1) and thermal cameras are adjusted to be aligned, the centers of the coordinates ($O_1$ and $O_3$) are shared.

### A. Multi-Spectral Camera Calibration

To capture the fully aligned RGB and thermal pairs, extrinsic calibration between the cameras should be done. Due to the different imaging properties (RGB: $450nm - 700nm$, thermal: $7um - 13um$), conventional methods to find the

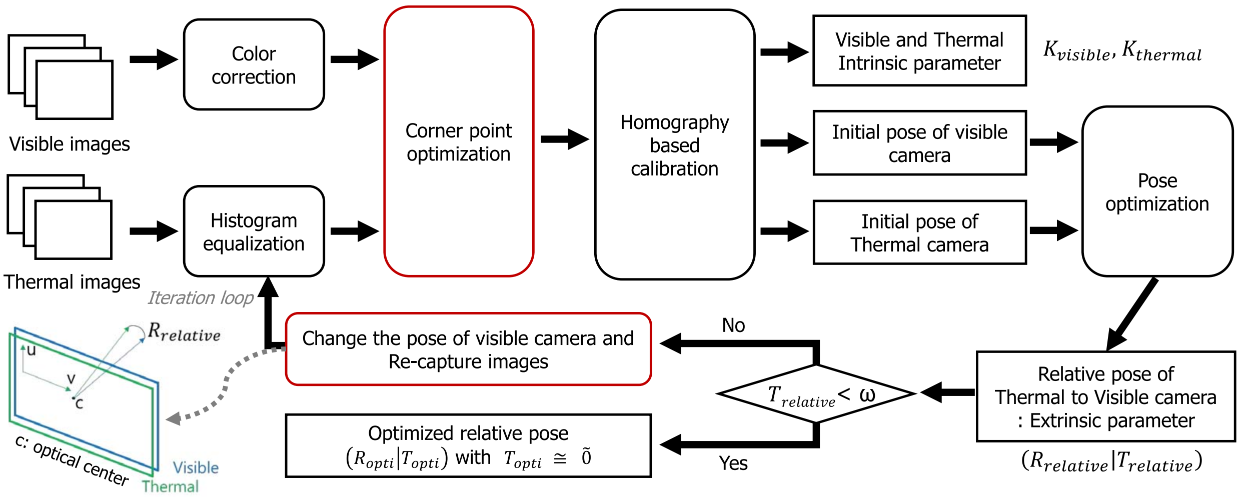

Visible images
Thermal images
Color correction
Histogram equalization
Corner point optimization
Homography based calibration
Visible and Thermal Intrinsic parameter
$K_{visible}, K_{thermal}$
Initial pose of visible camera
Initial pose of Thermal camera
Pose optimization
Relative pose of Thermal to Visible camera : Extrinsic parameter
$(R_{relative}|T_{relative})$
$T_{relative} < \omega$
No
Yes
Change the pose of visible camera and Re-capture images
Optimized relative pose $(R_{opti}|T_{opti})$ with $T_{opti} \cong \tilde{0}$
Iteration loop
$R_{relative}$
u
v
c
c: optical center
Visible
Thermal

Fig. 9. **Geometrical calibration process of multi-spectral cameras.** The overall process is similar to stereo calibration, and *corner point optimization* and *changing a pose* process are different from the conventional method.

correspondences are not accurate when used for multi-spectral calibration. Moreover, the blurry effect of thermal images which is caused by the diffusion of heat radiation and the shallower depth of field makes it difficult to extract accurate corner points. Therefore, one of the main issues is accurately extracting the corresponding points for computing relative poses between sensors.

A uniformly printed chess board, as shown in Fig. 8. (a), is the most popular pattern for RGB camera calibration. Although the black and white pattern is useful to extract corner points in RGB images, this pattern is rarely recognized in thermal images because the intensity of thermal images depends on the omitted energy rate from objects, not the visible colors. Therefore, various specially designed pattern boards have been proposed. The basic concept is that the different materials and radiated devices create temperature difference to facilitate recognition of the edges in thermal images. Moreover, most patterns can be used to calibrate RGB images for multi-spectral calibration. Mouats *et al.* [11] proposed a handmade grid-based pattern board which is punched onto a box board to allow heat to penetrate, as shown in Fig. 8. (b). During the calibration step, heating sources such as a laptop or steam are placed behind the board. Although this pattern is simple to use, it is easy to bend itself, and it is not easy to create a sharp grid in the manufacturing step. Jung *et al.* [12] was the first to propose pattern which pierces circular holes onto a metal board (Fig. 8. (c)) for calibration between a time of-flight (ToF) sensor and a RGB camera, and Hwang *et al.* [13] employed this pattern to calibrate RGB and thermal cameras. The circle-board pattern is used to extract the center points of each hole instead of corner points. However, the center points of circles can easily be distorted by dual circles which occur in the lateral view due to the thickness of the metal board. Moreover, because the board has a wider contact area, it is difficult to maintain the temperature on the board uniformly. To overcome these limitations, we proposed the line-based pattern board shown in Fig. 8. (d). A detailed description of this board and the multi-spectral calibration method are presented below.

| w/o optimization | | with optimization | |
|---|---|---|---|
| RGB | Thermal | RGB | Thermal |
| 0.73 | 0.78 | 0.31 | 0.38 |

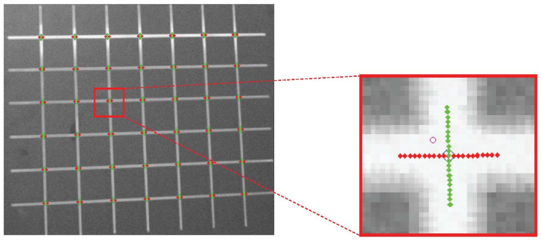


Fig. 10. **The results of corner points optimization** in Fig. 9. Magenta circle is the initial corner point, and blue circle is optimized point moved from magenta.

*1) Line-Based Pattern Board:* We designed a line-based pattern board which accurately calibrates RGB and thermal cameras. Copper lines are regularly milled onto a printed circuit board (PCB) 2*mm* wide and spaced at 40*mm*, and six and seven lines on each axis are printed to provide intersections. Note that we used different numbers of lines to eliminate the ambiguity of the corner point in the horizontal/vertical direction. Our line-grid pattern is more apt to maintain high contrast in thermal images because the copper line has good conductivity to maintain a uniform thermal distribution. Moreover, as the proposed pattern is identical in terms of the geometry to a conventional chess board, it can easily be adapted to many existing calibration techniques.

*2) Overall Process:* The overall diagram for multi-spectral camera calibration is shown in Fig. 9. This process mainly consists of two stages (*calibration* and *alignment*). During the calibration, these stages are repeated until it meets the stop conditions. *Calibration* is the process of obtaining the intrinsic and extrinsic parameters of each camera using corresponding points, and *alignment* is the process of making the geometric distance between the optical axes of RGB and thermal cameras almost zero. As shown in Fig. 11, as RGB and thermal cameras have different imaging properties, conventional RGB-based

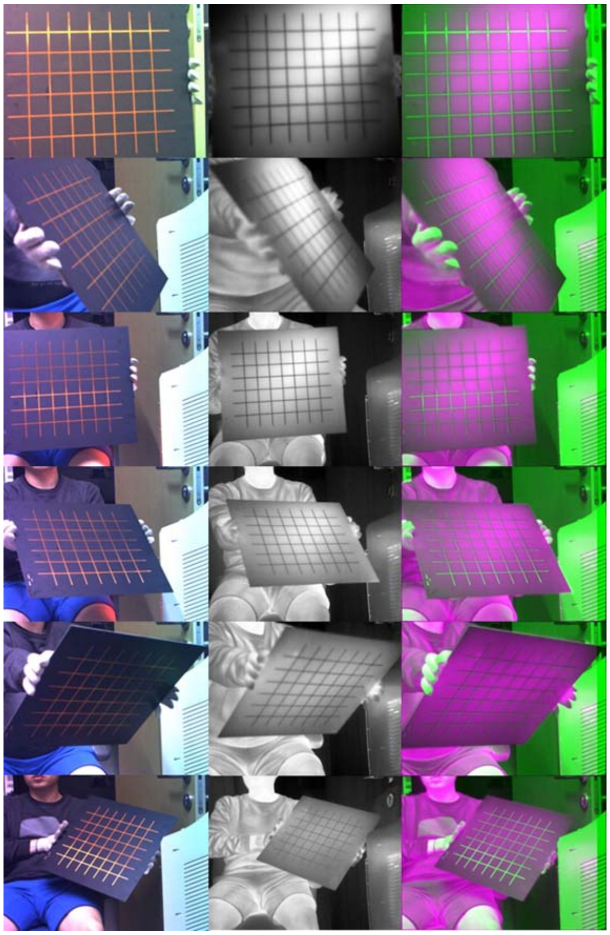
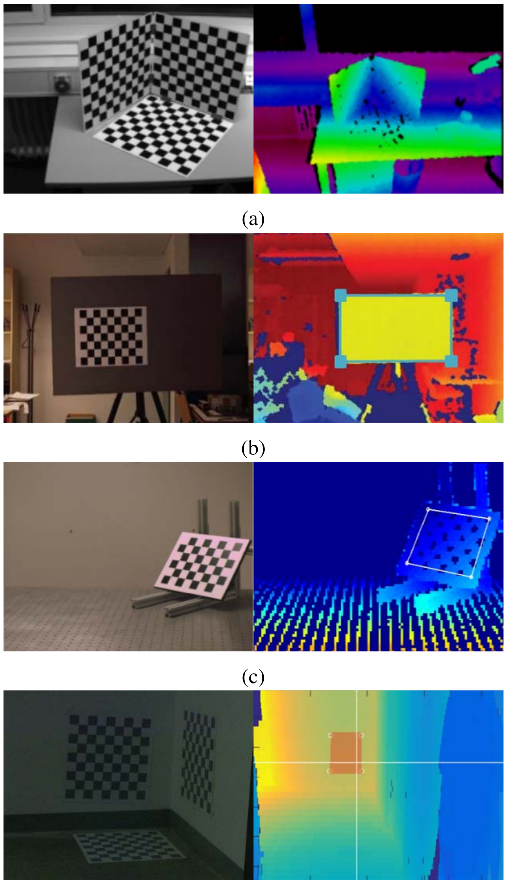


Fig. 11. **Examples of calibration results**: (left) RGB images, (middle) thermal images, and (right) fused images.

corner detection methods cannot guarantee an accurate result. Therefore, we proposed a greedy-based optimization approach to ensure accurate corner points in the thermal image.

To find the optimal solution, we initially define the following *guidelines*: $f_c(k, \theta) = [k + s \cdot cos\theta, k + s \cdot cos\theta \pm o \cdot sin\theta]$. The intersection of the guidelines (red and green lines in Fig. 10) defined by two axes is finally obtained by the optimal solution of an initial point (the magenta circle in Fig. 10). At this time, guidelines are constrained with a penalty to the boundary of the lines and with a constraint of orthogonality. In other words, the initial corner point (magenta circle in Fig. 10) should be guided to the pixel with the highest temperature on both axes (the highest intensity value), and these two axes should be close to a right angle, as indicated in Fig. 10. Note that $(x, y)$ are the position of the corner points, $\theta_r$ is the angle between the $x$ axis and the red guideline, and $\theta_c$ is the angle between the $y$ axis and the green guideline. The objective follows Eq. (1), and we use Levenberg-Marquardt optimization method.

$$
\begin{aligned}
\min_{x,y,\theta_r,\theta_c} \; & E_d(x, y, \theta_r, \theta_c) + \beta \cdot E_s(\theta_r, \theta_c) \qquad (1)\\
E_s(\theta_r, \theta_c) &= \|cos(\theta_r - \theta_c)\|\\
E_d(x, y, \theta_r, \theta_c) &= \|2 - V(x, y, \theta_r) - V(x, y, \theta_c)\|\\
V(x, y, \theta) &= I(f_c(x, \theta), f_c(y, \theta))
\end{aligned}
$$



(a)

(b)

(c)

(d)

Fig. 12. **Pattern boards and calibration toolkit**. (Left) pattern boards for calibration of RGB camera and Velodyne. (Right) depth examples captured by depth sensor and 3D points marked by a tool for extracting wall-planes. (a) Geiger *et al.* [14]. (b) Herrera and Heikkila [15]. (c) Unnikrishnan and Herbert [16]. (d) Ours.

After the each calibration step, we manually change the pose of the RGB (1) camera according to $T_{relative}$. Because our camera jigs are designed to make fine adjustments, repetition of this process is completed within a few times. Finally, we obtain the co-aligned RGB and thermal camera settings and fully registered multi-spectral images spatially.

*3) Calibration Result:* To verify the proposed method, we undertook calibration with the proposed greedy-based optimization method and without it. To do this, we captured RGB and thermal images (the left/middle of Fig. 11) and computed the intrinsic and extrinsic parameters. After calibration step, we obtained an aligned multi-spectral image, as shown on the right in Fig. 11. In the table in Fig. 10, the proposed corner -point optimization method can reduce re-projection errors in RGB and thermal images by half (RGB: 0.73 to 0.31, thermal: 0.78 to 0.38). We note that the initial magenta point moved to the peak point of the high-intensity region, which is equivalent to the intersection of the red and green lines (Fig. 10).

### *B. RGB and Velodyne Calibration*

For the RGB (1) and Velodyne calibrations, we took advantage of a wall structure in the real world. Our methods used

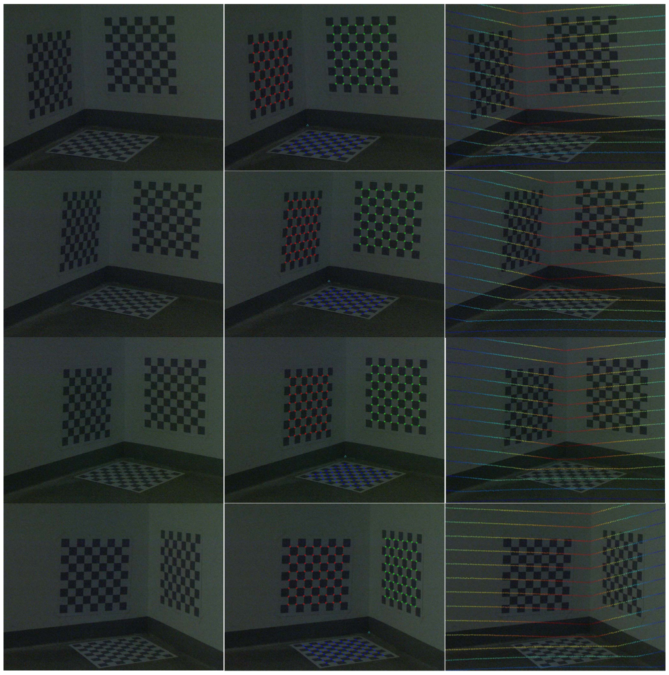


Fig. 13. **Calibration results of RGB camera and Velodyne.** (From left to right) RGB images, RGB images with reprojected pattern points, and RGB images with reprojected point-clouds.

the geometric relationship known as *plane-to-plane mapping* between sensors; therefore, it does not need specially manufactured calibration targets such as pattern boards or 3D structures. For practical calibration, we attached only printed chess-board patterns on three side walls and walls do not have to be located in an orthogonal structure (Fig. 13). In previous works, the authors presented point matching-based calibration methods [15], [16]. Users manually marked the corresponding points via the provided the toolkit in RGB and Velodyne images and the relative pose was computed by *point-to-point mapping* optimization. However, this method cannot guarantee that the points corresponding to RGB-based points are scanned in 3D LiDAR due to the angular resolution of 3D LiDAR. Geiger *et al.* [14] used multiple chess boards or a triaxial calibration target to find the geometric relationship between heterogeneous sensors. If the image takes all these targets, this method can be used with a single image to obtain the extrinsic parameter. To compute this parameter, the author undertook automatic *plane segmentation* to extract the calibration target from images. However, this type of method involves capturing the image in restricted conditions, such as a special purpose room with blackout windows. Compared to previous works, the proposed method is simple and practical in cases where it is difficult to prepare enough space with numerous calibration boards or to manually extract points in scanned 3D LiDAR and multiple RGB images.

*1) Wall-Attached Pattern Board:* The preparation steps to be taken when using the proposed method involve attaching the printed chess boards on a wall with three sides. First, we captured the scene, including the three side walls with the attached pattern boards in both RGB and Velodyne images. The pattern board is used to extract the corner points for the RGB camera and the scanned wall is used to extract the plane for the Velodyne device. To do this, we used a general chess board pattern of the type which is generally used for RGB calibration (a grid spacing of 40*mm)*, and used a printed sheet which prevents the effects of reflected light. Because the smoothness and flatness of the wall are the only considerations of our method, the three walls do not necessarily have to be orthogonal. Moreover, our method has the advantage of being extend to RGB and general depth sensors (e.g., Kinect and ToF).

*2) Overall Process:* The process can be divided into two stages: *plane extraction* and *plane registration*. We initially conduct plane extraction with the RGB and Velodyne data and then register the plane to obtain the relative pose between the sensors.

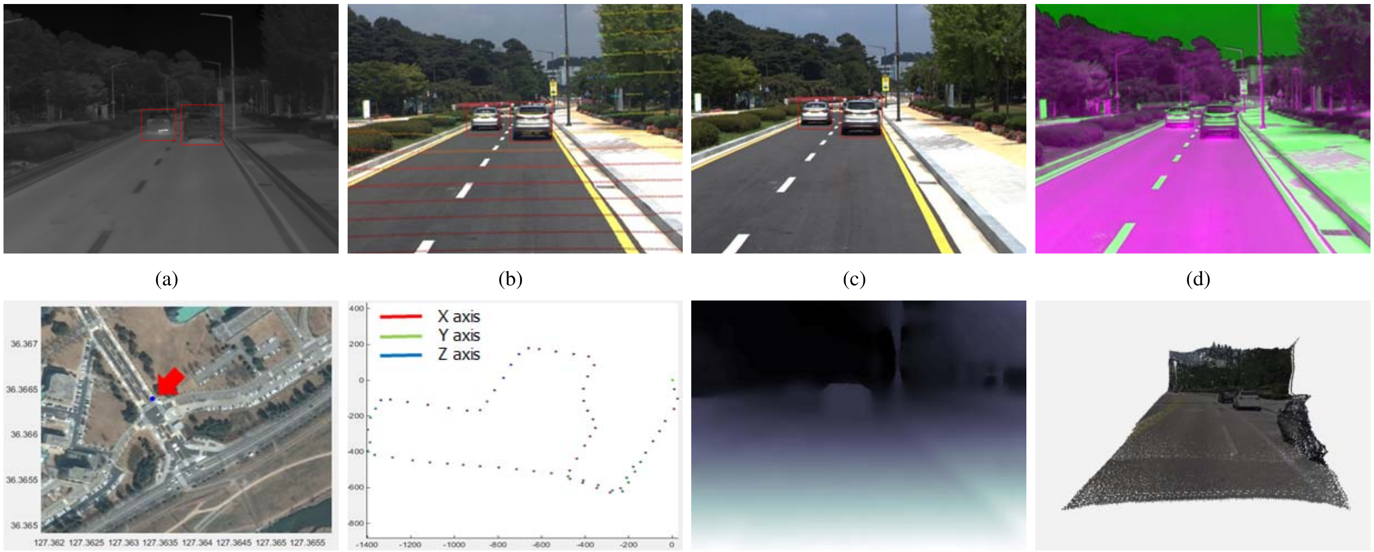
We assumed that the radial distortion of RGB images has already been removed with intrinsic parameters obtained through the previous camera calibration. The procedure to estimate three planes from the RGB images is as follows. First, we manually extract the corner points of the three wall-attached chess boards in the RGB images. Based on the RGB camera calibration, we compute the plane parameters of each wall using the extrinsic relationship between the points of the image coordinates and those of the world coordinates. To extract the plane from Velodyne, we convert 3D measurement into a range image consisting of the depth according to the vertical/horizontal angles. We then mark four points on each wall to compute the plane parameters and created the plane pair from that of the RGB image and the plane from Velodyne (Fig. 12. (d)-right).

Finally, to register the three planes extracted from each sensor, we use the Levenberg-Marquardt (LM) method to optimize the following objective function.

$$\begin{aligned}\min_{R,T} E(V_x, I_x) &= E_{dist} + \alpha \cdot E_{ang} + \beta \cdot E_{int} \qquad (2)\\ E_{int} &= f_i(V_o, I_o)\\ E_{dist} &= f_d(V_p^L, I_x^L) + f_d(V_p^R, I_x^R) + f_d(V_p^F, I_x^F)\\ E_{ang} &= f_\theta(V_l^{L,R}, I_l^{L,R}) + f_\theta(V_l^{R,F}, I_l^{R,F})\\ &\quad + f_\theta(V_l^{F,L}, I_l^{F,L})\end{aligned}$$

Here, $V_x$ and $I_x$ indicate the 3D points on the three walls within the captured range and the RGB image, and $V_p$ denotes the three walls within the captured range image. $V_l$ and $I_l$ are the intersection lines among the planes, and $V_o$ and $I_o$ are the origins of the sensors. The upper subscript denotes the type of wall, in this case left, right, and floor. $f_d$ returns a distance value between a plane and the 3D points, $f_\theta$ returns an angle value between the intersection lines, and $f_i$ returns a distance value between the origin points of each sensor.

*3) Calibration Result:* To verify the RGB and Velodyne calibration, we undertook a qualitative evaluation, as shown in Fig. 13. Fig. 13. (b) shows the reprojection of 3D pattern corners and the 3D intersection corner. Fig. 13. (c) shows the reprojection of 3D points from Velodyne on the 2D image after external calibration between RGB and Velodyne. As shown in Fig. 13, the calibration of the RGB camera and Velodyne are performed well from the plane-to-plane registration.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 14. **Samples from the MATLAB development tools.** (a) thermal video frames, (b) RGB(1) video frames and projected Velodyne points, (c) RGB(2) video frames and bold red dot-boxes indicate annotated bounding boxes of moving objects, and all video frames are undistorted and calibrated. (d) perfectly aligned RGB-thermal frame pairs (e,f) the current location and position of the vehicle, (g) depth video frames which are made by depth estimation method [9] using RGB stereo image and Velodyne 3d points. (h) 3D point cloud generation from a dense depth map. This development kit are available from the KAIST multi-spectral website.

## V. Dataset Access

### A. Development Kit

We provide a MATLAB toolbox for easy access and manipulation of the *raw dataset*. This toolkit supports a sequence decoder and several demos for each task. Several processed examples are shown in Fig. 14 and the provided functions are as follows.

*1) Sequence Decoder:* The function *GetFrame.m* reads binary sensor data from a specific raw sequence and returns manipulated images and meaningful data, such as 14-bit/8-bit thermal images, stereo-type color images, range and reflectance images of 3D LiDAR (Velodyne), and the latitude/longitude and relative 6D position with a timestamp.

*2) Undistorted and Rectified Images:* The function *GetRectifiedImages.m* removes lens distortion from each camera and rectifies the images between cameras. This function returns perfectly aligned RGB-thermal images, rectified RGB-RGB images, and rectified RGB-thermal images. The function *GetRGBTFusion.m* demonstrates precisely how to align the RGB and thermal images, as shown in Fig. 14. (d).

*3) Bounding Boxes and Labels:* The ground truth of the annotator is managed in the *video bounding box (VBB)* format. The function *GetBBLabel.m* decodes the VBB file and returns the class labels and positions of the bounding boxes in each frame of the video. VBB also provides an occlusion flag for individual objects, and red/yellow/green BB means no/partial/heavy occlusion, respectively.

*4) 3D Point Cloud Projection:* The function *GetVeloImage.m* returns 3D point cloud (360/FoV) and projected point cloud into the reference RGB (1). Note that 3D points are measured in units of meters and 3D points over 70*m* are removed. The local function *Convert2Dto3D.m* provides 3D points based on Velodyne. To project to the RGB (1), it is necessary to change the coordinates of the 3D points from Velodyne to RGB (1) using an extrinsic parameter, after which the converted points are projected by an intrinsic parameter.

*5) Vehicle Trajectory:* The function *ShowVehiclePath.m* shows how to read and display the vehicle trajectory using GPS/IMU data. Note that it does not use all of the data provided by RT2002 and instead uses latitude and longitude for translation, and the roll, pitch and heading information for rotation. The function *ConvertOxtsToPose.m* returns the 6D position of the vehicle in Euclidean space. For this conversion, we utilize the Mercator projection method [1].

*6) Dense Depth Images:* The function *GetDepthImage.m* returns two types of dense depth images depending on the option. As a first option, SGM [17] is a dense stereo matching method that can be used for accurate 3D reconstruction from a pair of calibrated images. As a second option, MC-CNN [9] undertakes stereo matching by training a convolutional neural network based on image patch matching. We fine-tuned the network to our dataset; the image-based depth results are refined by 3D Velodyne points.

*7) Integrated Visualizer:* The function *Viewer.m* shows an integrated visualizer. We can demonstrate all functions of development tools using a graphical user interaction (GUI). We expect that the user can extend our functions to various purposes to the greatest extent possible.

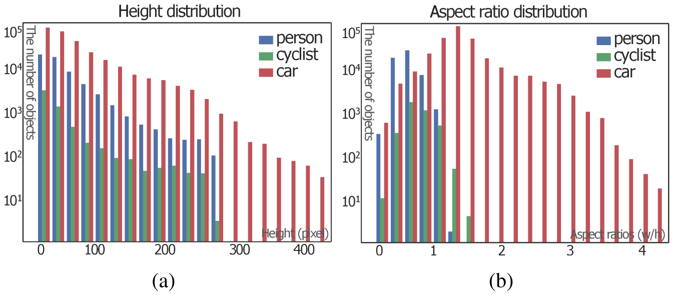
### B. Subset Benchmarks

We designed the subset of the visual perception tasks with the KAIST multi-spectral dataset. As a canonical imaging sensor, RGB images have been used as inputs in many computer vision and robotics researches. These works focused on mining more useful information from only RGB image to be robust in various conditions ( [18]–[20]). While the vast majority of camera networks still employ traditional RGB sensors, recognizing objects in case of illumination variation, shadows, and low external light is still challenging open issue. Hence, this led to the question of how it would be possible to robustly perceive the world for all-day. We believe that the answer will rely on the use of alternatives to RGB sensors such as depth or thermal imaging devices. In recent times, there have been two main streams using multi-modal information. The first is the methodology using multi-modal information during training and inference. In fact, these works have shown that the multiple image modalities can be used simultaneously to produce better recognition models than either modality alone in the challenging scenarios to RGB images ( [13], [21]–[23]). Another approach is using multi-modal information during training, which is used to learn to hallucinate features from RGB images, and the trained model performed well on RGB images alone as input during inference. For these purposes, our dataset can be useful to deal with many tasks of autonomous system with respect to multi-modal data fusion approaches as the first way, and RGB-based approaches from multi-modal feature learning. According to the tasks, an additional data can be used in conjunction with the subset benchmarks. For details about the benchmarks and evaluation metrics, we refer the reader to several earlier works [13], [24]–[27], and our project website.

*1) Object Detection:* Object detection is a crucial part of building intelligent vehicles and an advanced driver assistance system (ADAS). Specifically, because autonomous emergency braking (AEB) for pedestrian protection is to be deemed as an important evaluation item in Euro NCAP starting in 2018, the interest in object detection is growing. To assure pedestrian protection for day and night, additional sensors such as a thermal camera and LiDAR are required. Hence, we created an object detection benchmark to encourage researchers to develop accurate detection algorithms for day and night. For this benchmark, we provide bounding boxes with occlusion flags as the ground truth on RGB-Thermal images for the *Campus, Residential* and *Urban* settings. There are *52,826*, *5,205* and *250,882* bounding boxes (BBs) for *person*, *cyclist* and *car*, respectively. A statistical analysis of the BBs for object detection provides a scale distribution and an aspect ratio distribution, as indicated in Fig. 15.

*2) Vision Sensor Enhancement:* We designed a multi-spectral benchmark which is used for up-sampling or detail enhancement. By solving these problems, we can offer a competitive price of the thermal sensor. In addition, we expect greater sensor popularization. We provide two types of subsets. For the first subset (RGBT-67), RGB-thermal pairs are sampled from multi-spectral sequences. This consists of a train set (57) and test set (10), all of which are available at a resolution of *640 x480*. In addition, we offer an additional test set (T-137) captured in a different scenario (urban). It was collected using FLIR-MSX technology, which supplies thermal and RGB images together.

*3) Depth Estimation:* We designed the first multi-spectral stereo benchmark from day to night in various real-driving conditions. From our benchmarks, we initially select a RGB stereo pair with dense depth maps and thermal images, i.e., co-aligned to the left-view RGB image. For depth estimation, we separate the urban area into downtown and suburbs for a more diverse dataset of drivable regions. In total, we provide *8,970* stereo/thermal pairs, consisting of training (*4,534*) and testing (*4,436*). Compared to the previous depth dataset, our benchmark supports co-aligned RGB and thermal pairs with high-quality depth maps and even test samples at night time.



(a) (b)

Fig. 15. **Statistics for object bounding boxes.** (a) is shown in the number of BB according to the height size of each object, and (b) is shown in the number of BB according to the aspect ratio width to height size.

*4) Multi-Spectral Colorization:* Many colorization methods have been proposed for various purposes such as image restoration, synthesis, and virtual data generation. Recently, deep-learning-based techniques have showed high performance in the RGB domain [28], [29] and in the multi-spectral domain [30]–[32]. With completely aligned RGB and thermal images, we can create a new large-scale RGB and thermal colorization benchmark which considers the driving conditions. For compatibility with other tasks, the colorization subset contains four scenarios (campus, urban, suburb, and residential) and both the train and test samples are identical to the stereo benchmark. We provide raw images with no pre-processing methods such as tone mapping or HDR applied. In addition, we provide compressed thermal images (8-bit) for visualization.

## VI. Comparison With Other Datasets

In the section, we compare our benchmark to other datasets in terms of ADAS and autonomous driving as well as multi-spectral configurations. As numerous works about ADAS and autonomous driving have been conducted, many datasets have been published. However, this paper only deals with recent large-scale studies. In addition, it deals with studies similar to ours in terms of data acquisition or sensor installation environments.

### *A. Datasets for ADAS and Autonomous Driving*

*1) KITTI:* Our benchmark is inspired by the KITTI [1] dataset, influential benchmark which has been widely used for quantitative comparisons of various computer vision and robotics tasks. We intend to mimic KITTI to provide a rich source of multi-modal data in normal conditions in addition to challenging conditions with a thermal imaging sensor. Therefore, we can provide the first multi-spectral dataset for use in actual driving conditions which contains co-aligned high-resolution RGB and thermal pairs, object annotations, 3D information (depth, point), vehicle positions and driving paths (GPS, IMU) in every synchronized frame for all-types of days. Moreover, we made an effort to prevent bias in certain object classes.

*2) Cityscapes:* Cityscapes [2] is the first large-scale dataset for pixel-level and instance-level semantic segmentation in urban street scenes. Compared to our recording platform, RGB stereo video is the main instrument used to capture the dataset in normal conditions. In this version, we focus on all-day perception for object-level information, which is better than pixel-level annotation. However, we expect that our dataset can be extended to pixel and instance-level segmentation in a similar manner to Cityscapes, which uses RGB stereo pairs.

*3) RoboCar:* RoboCar dataset [33] focuses on long-term autonomous driving. It was collected in various weather conditions, including heavy rain, night, direct sunlight and snow. Compared to our dataset, cameras provide a full 360-degree visual coverage of scenes around vehicles. In addition, although they do not use an all-around 3D LiDAR system such as Velodyne, they support a 3D point-cloud jointly generated from 2D LiDARs and cameras. Our dataset is also created to cover various driving conditions. Particularly, when we captured the dataset, we attempted to capture more challenging time slots, such as sunrises, mornings, afternoons, sunsets, and night and dawn scenes.

**TorontoCity** TorontoCity dataset [34] provides different perspectives of the world captured from airplanes, drones, as well as cars driving around a city. This dataset contains aerial images, street-view panoramas, street-view LiDAR and airborne LiDAR. These data sources are aligned in the maps to generate accurate ground truth. Compared to our dataset, the TorontoCity dataset is focused on static environment recognition instead of the dynamic objects (e.g., pedestrians, vehicles) mainly handled in existing datasets, including ours. Moreover, this dataset does not consider the various capturing conditions mentioned above.

**VirtualKITTI** VirtualKITTI [3] is large-scale synthetic dataset which provides photo-realistic virtual worlds with accurate ground truth outcomes for object detection, tracking, scene and instance segmentation, depth, and optical flows. It is similar to our benchmark in that it also provides a dataset which can be used in many different environmental conditions (e.g., morning, sunset, overcast, fog, rain), but it differs from the proposed dataset in that it does not use additional robust sensors in a virtual environment.

### *B. Multi-Spectral Datasets for ADAS*

*1) Multi-Modal Stereo Dataset (CVC-15):* CVC-15 [35] [36] is stereo dataset which attempts to be a solution to the locating of correspondences between multi-spectral sensors. This dataset consists of 100 pairs of RGB-thermal images which were captured in different urban scenarios but not driving environments. Although CVC-15 was captured using a Bumblebee stereo camera, it is difficult to use it for various applications because it does not collect stereo image pairs or depth images. On the other hand, because our dataset consists of RGB-RGB-thermal images, we can provide pairs of stereo images and depth images in addition to RGB-thermal images.

*2) Color and Thermal Stereo Dataset (CATS):* The CATS [37] dataset focuses on the stereo matching of various spectra, including those of RGB and thermal images. It consists of stereo thermal, stereo RGB, and cross-modality image pairs with high accuracy ground truth outcomes (<2mm) generated from a LiDAR system. Compared to the proposed dataset, CATS is smaller, containing approximately 1400 images in various environmental conditions (e.g., day, night, and foggy scenes).

*3) Multi-Spectral Pedestrian Dataset:* Multi-spectral Pedestrian Dataset [13] is the first large-scale multi-spectral dataset for pedestrian detection in day and night. This dataset consists of RGB-thermal image pairs captured in a driving environment. However, the multi-spectral pedestrian dataset supports grayscale thermal images (8-bit) for image processing instead of raw images (14-bit). This type of grayscale thermal image has an advantage in that it can easily be handled in the form of the RGB image in terms of a visualization and applied algorithms. However, pre-processed thermal images cannot be directly converted with regard to accurate temperatures, such as those on the Kelvin or Celsius scales. Moreover, our benchmark has directly advantage of being applicable to various areas related to computer vision and intelligent vehicles.

## VII. Lessons Learned

In this section, we summarize some of the issues to consider when capturing a dataset, from preparation to finishing. Through our lessons, we hope that the difficulties of researchers who want to build a similar system can be resolved.

### *A. Preparing Collection*

*1) Camera Resolution:* It is the common choice to select a high-resolution camera, as larger images contain more visual information. However, before choosing the maximum resolution, several issues must be considered, such as the bandwidth for data transmission and the writing speed of the SSD used for image storage. Taking all of these factors into consideration, it becomes possible to select a camera successfully within the allowable range.

*2) Camera Lens:* In our case, we select a long focal lens to observe remote objects. If wanting to a wider field of view, a short focal lens may suffice. Recently, Tesla and Mobileye devised trifocal camera system (HoV−20°, −50°, −150°) as a new hardware configuration. A trifocal camera at the top of the windscreen can help identify pedestrians or various targets (e.g., forward vehicles, traffic lights, traffic signs) that stand at the different depths.

*3) Camera Synchronization:* There are two ways to synchronize multiple cameras: *hardware triggers* and *software triggers*. As software methods have some latency time, hardware synchronization via an external trigger is typically used wherever possible. In recent years, a feature called a *signal generator* has been added to the camera to match the shutter times between devices. We also used this function to capture images simultaneously. However, despite the use of a signal generator, there is drift due to the differences in the exposure times between the devices. Therefore, there remains an asynchronous phenomenon wherever excessive movement of the vehicle occurs, such as a corner or a bump.

*4) 3D LiDAR:* The selection of 3D LiDAR depends on the vertical fields of view of the vision cameras used together. Because Velodyne provides sparse 3D points (from 16 to 64 vertical lines), we recommend the use of 3D LiDAR with a dense angular resolution or the use of several units together to compensate for the scanning spaces of individual units.

*5) Collecting Environment:* One important aspect for data collection is the road condition and surrounding complexity. Unpaved roads, construction sites, and areas with speed bumps are not good places to capture data. Unless deliberately choosing such places, it is best to avoid them. Another important aspect of data collection is the level of difficulty according to the number of objects. Therefore, it is necessary to choose locations which have a suitable level of difficulty depending on the task.

### *B. Before Collection*

*1) Periodic Calibration:* Although our sensors are tightly equipped in a rigid and sophisticated system, it is not easy to maintain high-quality calibration between devices due to vibrations and shocks generated by bumpers and unpaved roads. For these reasons, sensor calibration is periodically required. Going one step further, in order for the sensor package to be commercialized and attached to the vehicle, an *auto calibration* feature must be included. Compared to other papers [1], [37] which fixed sensors by bar types structure, our system has an advantage of relatively less tolerance of the calibration because we fixed all sensors on the optical table. Before data collection, the list that requires a periodic inspection is as follows: the screw statuses of the camera jig and splitter jig. In addition to periodic calibration, calibrating the device according to power on/off settings is required, and DGPS is a typical example. Before data capturing, it is necessary to check the current position of the vehicle through DGPS. If there is a position drift, DGPS calibration should be done.

*2) Storage Space Check:* SSDs are known to operate at low writing speeds when using more than 80% of the total storage space. Therefore, it is preferable to secure sufficient space on SSDs before collecting data.

### *C. During Collection*

*1) Camera Parameters:* Because the camera is synchronized by an external trigger, we set the exposure time such that it is not longer than the trigger time. At night, we set the camera parameters to capture images as brightly as possible.

*2) Guidelines for Driving:* It does not matter who drives the experimental vehicle, but the following aspects should be considered in advance. When capturing data, it is necessary to follow guidelines related to the average speed during straight ahead operations and during rotation. In addition, the more similar the pattern is for acceleration and deceleration, the better the dataset to be obtained. It is good to reduce the speed as much as possible when passing over speed bumps or through construction sites.

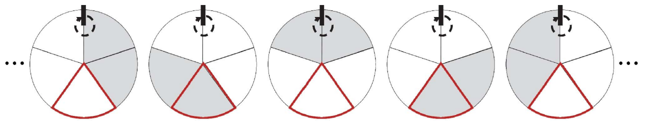


Fig. 16. **The updated periodic pattern of data.** According to this pattern (X O X O X), frontal 3D point clouds (10Hz, red region) are updated in 2 times out of 5 times. This figure is top-view, and field of view (FoV) of the thermal camera (25Hz) is overlapped on a red region. Gray color indicates updated angular regions according to the cycle and black-bar is a reference coordinate of Velodyne.

### *D. After Collection*

*1) Post-Processing:* All data is stored in a raw format for research scalability. In other words, we did not use a specific format for image storage, such as the bmp or jpg formats. Sensor data is collected and managed in video form rather than in each frame because many fragmented files can be a burden on the file system such that the risk of file corruption is increased.

*2) Velodyne Parsing:* As mentioned above, all sensors are synchronized to the reference signal generated by the thermal camera. The RGB and thermal cameras of a sensor system operate at 25Hz and LiDAR operates at 10Hz. When one RGB/thermal image is stored, LiDAR data is updated only in 72 degrees of 360 degrees. Therefore, the presence or absence of LiDAR synchronized with the RGB/thermal image repeats the following pattern: (X O X O X). The perfectly synchronized RGB/thermal image and Velodyne points can be used by parsing, referring to the following Fig. 16.

## VIII. Summary and Future work

In this paper, we have presented the *KAIST multi-spectral dataset*, which is focused on *all-day vision and extreme illumination changes* for autonomous driving. With our dataset, we intend to challenge current approaches to all-day vision tasks, and this advance enables research for all-day and lifelong learning for ADAS and autonomous vehicles. Because many researchers in industry and academia are interested in day and night problems, we will provide a benchmark service similar to KITTI using a common ground truth and evaluation criteria. Moreover, we will provide more sophisticated benchmark algorithms, such as off-the-shelf deep learning architectures. Finally, we hope that numerous researchers can create their own specific applications with our multi-spectral dataset.

## Acknowledgements

The authors' data set and toolkits are available online at http://multispectral.kaist.ac.kr.

## References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] M. Cordts *et al.*, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Visison Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3213–3223.

[3] F. Barrera, F. Lumbreras, and A. D. Sappa, “Multispectral piecewise planar stereo using Manhattan-world assumption,” *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 52–61, 2013.

[4] W. Maddern and S. Vidas, “Towards robust night and day place recognition using visible and thermal imaging,” in *Proc. Robot., Sci. Syst. Conf. Workshop (RSS)*, 2012. [Online]. Available: https://eprints.qut.edu.au/52646/

[5] L. Jingjing, Z. Shaoting, W. Shu, and M. Dimitris, “Multispectral deep neural networks for pedestrian detection,” in *Proc. (BMVC)*, 2016, pp. 73.1–73.13.

[6] J. Han and B. Bhanu, “Human activity recognition in thermal infrared imagery,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Sep. 2005, p. 17.

[7] D. A. Torabi, G. Massé, and G.-A. Bilodeau, “An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications,” *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 210–221, 2012.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: A benchmark,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 304–311.

[9] J. Žbontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, Apr. 2016.

[10] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, “A mask-based approach for the geometric calibration of thermal-infrared cameras,” *IEEE Trans. Instrum. Meas.*, vol. 61, no. 6, pp. 1625–1635, Jun. 2012.

[11] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, “Multispectral stereo odometry,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1210–1224, Jun. 2015.

[12] J. Jung, J.-Y. Lee, Y. Jeong, and I. S. Kweon, “Time-of-flight sensor calibration for a color and depth camera pair,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1501–1503, Jul. 2015.

[13] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[14] A. Geiger, F. Moosmann, O. Car, and B. Schuster, “Automatic camera and range sensor calibration using a single shot,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2012, pp. 3936–3943.

[15] D. H. C, J. Kannala, and J. Heikkilä, “Joint depth and color camera calibration with distortion correction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2058–2064, Oct. 2012.

[16] R. Unnikrishnan and M. Hebert, “Fast extrinsic calibration of a laser rangefinder to a camera,” School Comput. Sci., Robot. Inst., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-05-09, 2005.

[17] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[18] Q. Wang, J. Gao, and Y. Yuan, “A joint convolutional neural networks and context transfer for street scenes labeling,” *IEEE Trans. Intell. Transp. Syst.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/8012463/

[19] C. L. Zitnick and P. Dollár, “Edge boxes: Locating object proposals from edges,” in *Proc. ECCV*, 2014, pp. 391–405.

[20] J. Gao, Q. Wang, and Y. Yuan, “Embedding structured contour and location prior in siamesed fully convolutional networks for road detection,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May/Jun. 2017, pp. 219–224.

[21] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell, “Cross-modal adaptation for RGB-D detection,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 5032–5039.

[22] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, “Learning rich features from RGB-D images for object detection and segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 345–360.

[23] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, “Large-margin multi-modal deep learning for RGB-D object recognition,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1887–1898, Nov. 2015.

[24] Y. Choi, N. Kim, K. Park, S. Hwang, J. S. Yoon, and I. S. Kweon, “All-day visual place recognition: Benchmark dataset and baseline,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRw)*, Jun. 2015, pp. 1–9. [Online]. Available: https://roboticvision.atlassian.net/wiki/spaces/PUB/pages/9633810/CVPR+2015+Workshop+on+Visual+Place+Recognition+in+Changing+Environments

[25] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon, “Thermal image enhancement using convolutional neural network,” in *Proc. Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 223–230.

[26] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, “Multispectral transfer network: Unsupervised depth estimation for all-day vision,” in *Proc. Conf. Artif. Intell. (AAAI)*, Feb. 2018.

[27] J. S. Yoon *et al.*, “Thermal-infrared based drivable region detection,” in *Proc. IEEE Intell. Veh. Symp. (4)*, Jun. 2016, pp. 978–985.

[28] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification,” *ACM Trans. Graph.*, vol. 35, no. 4, p. 110, 2016.

[29] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.

[30] V. Sharma and L. Van Gool, “Does V-NIR based image enhancement come with better features?” *CoRR*, vol. abs/1608.06521, Aug. 2016. [Online]. Available: https://arxiv.org/abs/1608.06521

[31] M. Limmer and H. P. A. Lensch, “Infrared colorization using deep convolutional neural networks,” in *Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 61–68.

[32] M.-Y. Liu, T. Breue, and J. Kautz, “Unsupervised image-to-image translation networks,” in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–9.

[33] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 Year, 1000km: The oxford robotcar dataset,” *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[34] S. Wang *et al.*, “Torontocity: Seeing the world with a million eyes,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3028–3036.

[35] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtualworlds as proxy for multi-object tracking analysis,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4340–4349.

[36] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, “Multispectral image feature points,” *Sensors*, vol. 12, no. 9, pp. 12661–12672, 2012.

[37] W. Treible *et al.*, “Cats: A color and thermal stereo benchmark,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 134–142.

**Yukyung Choi** received the B.S. degree in electronic engineering from Soongsil University in 2006 and the M.S. degree in electrical and electronic engineering from Yonsei University in 2008, and the Ph.D. degree in robotics program from Korea Advanced Institute of Science and Technology in 2018. From 2008 to 2010, she was with KIST. Her research interests include deep learning, computer vision, and robotics. She was a recipient of the Samsung Human Tech Paper Award.

**Namil Kim** received the B.S. degree (*summa cum laude*) in electrical engineering from Sungkyunkwan University, Suwon, South Korea, in 2014 and the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2016. He is currently with the Autonomous Driving Group, NAVER LABS.

**Soonmin Hwang** received the B.S. degree in electric engineering and computer science from Hanyang University, Seoul, South Korea, in 2012 and the M.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012. He is currently pursuing the Ph.D. degree with the Robotics and Computer Vision Laboratory, Department of Electrical Engineering, College of Information Science and Technology, KAIST.

**Kibaek Park** received the B.S. degree in electrical engineering from Kyungpook National University, Daegu, South Korea, in 2013 and the M.S. degree in robotics program from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016. He is currently pursuing the Ph.D. degree with the Division of Future Vehicle, Robotics and Computer Laboratory, Department of Electrical Engineering, College of Information Science and Technology, KAIST.

**Jae Shin Yoon** received the B.S. degree in electric engineering and computer science from Hanyang University, Seoul, South Korea, in 2015 and the M.S. degree in robotics program from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the University of Minnesota.

**Kyounghwan An** received the B.S., M.S., and Ph.D. degrees in computer science from Pusan National University, Pusan, South Korea, in 1997, 1999, and 2004, respectively. Since 2004, he has been with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. He is currently serving as a Principal Researcher for the Intelligence Robotics Research Division, ETRI, and a Professor with the University of Science and Technology, Daejeon. His main research interests include enhanced map building and provisioning platforms for automated vehicles, navigation services, and localization-based services.

**In So Kweon** received the B.S. and M.S. degrees in mechanical design and production engineering from Seoul National University, Seoul, South Korea, in 1981 and 1983, respectively, and the Ph.D. degree in robotics from Carnegie Mellon University, Pittsburgh, PA, USA, in 1990. He was with the Toshiba Research and Development Center, Kawasaki, Japan. In 1992, he joined the Department of Automation and Design Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, where he is currently a Professor with the Department of Electrical Engineering, College of Information Science and Technology. His research interests include camera and 3-D sensor fusion, color modeling and analysis, visual tracking, and visual simultaneous localization and mapping. He is a member of the Korea Robotics Society. He was a recipient of the Best Student Paper Runner-Up Award at the 2009 IEEE Conference on Computer Vision and Pattern Recognition. He was the Program Co-Chair for the 2007 Asian Conference on Computer Vision (ACCV) and the General Chair for the 2012 ACCV. He is on the Editorial Board of *International Journal of Computer Vision*.