MAP: MULTISPECTRAL ADVERSARIAL PATCH TO ATTACK PERSON DETECTION

Taeheon Kim *Hong Joo Lee* *Yong Man Ro*[*]

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Recently, multispectral person detection has shown great performance in real world applications such as autonomous driving and security systems. However, the reliability of person detection against physical attacks has not been fully explored yet in multispectral person detectors. To evaluate the robustness of multispectral person detectors in the physical world, we propose a novel Multispectral Adversarial Patch (MAP) generation framework. MAP is optimized with a Cross-spectral Mapping(CSM) and Material Emissivity(ME) loss. This paper is the first to evaluate the reliability of a multispectral person detector against physical attack. Throughout experiment, our proposed adversarial patch successfully attacks the person detector and the Average Precision (AP) score is dropped by 90.79% in digital space and 73.34% in physical space.

***Index Terms*—** Multispectral adversarial patch, Thermal adversarial patch, Person detection, Cross-spectral mapping, Material emissivity loss

1. INTRODUCTION

Deep Neural Networks has shown remarkable success in the field of computer vision. With the development of DNNs, they have been applied into many real-world applications such as autonomous driving [1,2] and security systems [3–5]. In particular, in autonomous driving systems, multispectral data fusion-based person (pedestrian) detection model has shown great performance [6, 7]. By exploiting each spectral characteristic simultaneously, it can encode richer visual representations of objects than a single modality.

However, despite the great success of multispectral fusion models, recent studies show that DNN-based detection models are known to be vulnerable to adversarial attacks [8–10]. Especially, it has been shown that it is possible to generate adversarial patterns in the physical-world, and such patterns exist in forms of patches. These kinds of patches are called physical adversarial patches and they are used to evaluate the robustness of DNN-based detection models in physical world applications.

Therefore, many studies have generated adversarial patches to evaluate the reliability and robustness of the detection model [11–14]. Thys et al. [11] designed a printable adversarial patch that successfully fooled the person detection system. By optimizing the patch with objectness score loss and physical loss, they produce a physical adversarial patch in the real-world. Also, X.zhu et al. [14] proposed a physical adversarial patch for Thermal image. By placing small bulbs on a board and optimizing bulbs' location, they generate a thermal adversarial patch for infrared person detectors.

These existing studies about adversarial patches have focused only on single spectrum (RGB or Thermal) models, not evaluating the robustness for multispectral models. Recently, few studies have been published on the analysis of adversarial robustness in multispectral data fusion models [15–18]. However, these studies analyzed adversarial perturbations only in the digital space, not the vulnerability existing in the physical world. As we explained earlier, since the multispectral model is widely used in safety-related applications, it is necessary to develop an adversarial patch to evaluate the robustness of multi-spectral person detection models in physical world.

Therefore, in this paper, we generate a Multispectral Adversarial Patch(MAP) that could fool the multispectral detector in physical world. There are two major challenges in generating a MAP. First, two patterns' appearance at both sensors must be considered while generating the multispectral patch. In specific, the generated RGB pattern appears differently in the RGB camera and the thermal camera. Second, it is hard to control thermal pattern. The adversarial patch has a complex pattern. Therefore, in order to create such a complex pattern, we have to control the heat in each microscopic spot. However, it is difficult to keep the temperature of these microscopic spots constant.

To handle the aforementioned problems, we propose a novel Multispectral Adversarial Patch (MAP) generation framework. In the proposed generation framework, we exploit the emissivity of material as an alternative solution to represent a thermal pattern. A property of Stefan-Boltzmann law states that the intensity of a thermal image is related not only to temperature, but also to the emissivity of the material. Therefore, we predefine materials with different emissivity such that they are suitable for making various thermal and rgb patterns. With the predefined material emissivity, we proposed a Cross-spectral Mapping (CSM) such that the appearance of the patch on both sensors are considered while generating the patch. In the proposed CSM, the

[*]Corresponding author: Yong Man Ro, ymro@kaist.ac.kr.

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

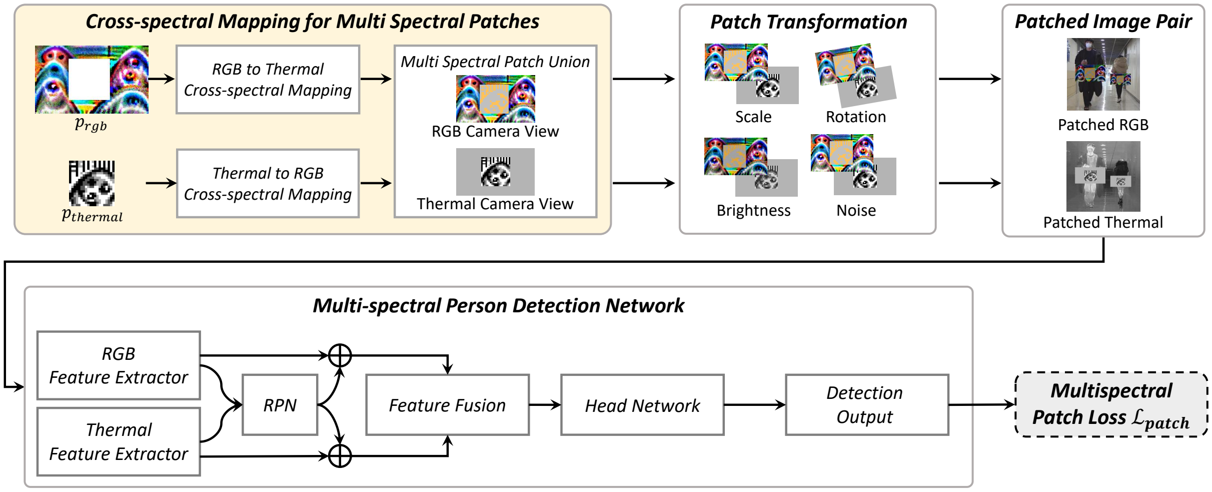
Cross-spectral Mapping for Multi Spectral Patches
p_{rgb}
RGB to Thermal Cross-spectral Mapping
Multi Spectral Patch Union
RGB Camera View
$p_{thermal}$
Thermal to RGB Cross-spectral Mapping
Thermal Camera View
Patch Transformation
Scale
Rotation
Brightness
Noise
Patched Image Pair
Patched RGB
Patched Thermal
Multi-spectral Person Detection Network
RGB Feature Extractor
Thermal Feature Extractor
RPN
Feature Fusion
Head Network
Detection Output
Multispectral Patch Loss $\mathcal{L}_{patch}$

Fig. 1. The overview framework for optimizing Multispectral Adversarial Patch(MAP). With the Cross-Spectral Mapping(CSM), the appearance of each patch is transformed into a different sensor view. Then, patches are transformed, fed into a multispectral person detector, and calculate the multispectral patch loss to optimize patches.

patch appearances are transformed into difference views of cross-spectral sensors. That is, the appearance of the thermal patch ($p_{thermal}$) is transformed into the appearance of the RGB sensor, and vice versa to the RGB patch (p_{RGB}) (details will be covered in Section 2.1). MAP is finalized by Multispectral Patch Union, which unifies the corresponding spectral pattern and the CSM output from the other sensor. Furthermore, we propose a Material Emissivity(ME) Loss so that the optimized thermal patch's pixel values have predefined emissivity values. The previous multispectral person detection datasets [19, 20] are not appropriate for generating adversarial patches, since the captured person size is too small to attach the patch. We collected a new dataset by capturing persons in various scenes such as inside a building, outdoors, on the street, in front of an entrance with a multi spectral camera. This dataset is expected to be used for future multispectral person(pedestrian) detection research. The major contributions of our paper are as follows:

- To the best of our knowledge, this is the first work generating adversarial patch for multispectral person detection model.
- To handle the complex pattern in a thermal sensor, we control the emissivity of the material instead of controlling the heat. By exploiting the emissivity, we optimize MAP with Cross-Spectral mapping(CSM) and Material Emissivity(ME) loss.
- We collected a new dataset for our task. It contains 1500 RGB-thermal image pairs of persons captured in diverse scenes, day and night.
- Through the extensive experiment, we show that the generated MAP could sufficiently fool the multispectral person detector in physical space as well as digital space.

2. PROPOSED METHOD

Fig. 1 shows the overview of the proposed Multispectral Adversarial Patch (MAP) generation framework. In the proposed framework, a Cross-Spectral Mapping (CSM) transforms multispectral patterns (p_{rgb} and $p_{thermal}$) into appearances on opposite camera sensor views. The patches on both sensor views are augmented with various transformations (scale, rotation, brightness, and noise) and attached to each person in the paired images. They are applied in a random location near the center of the person for each spectral image. After patches are applied, the patched RGB-thermal image pair is fed into the pretrained multispectral person detector. The MAPs are optimized to minimize the multispectral patch loss $\mathcal{L}_{patch}$. In the following section, we describe the proposed CSM and multispectral patch loss in detail.

2.1. Cross-Spectral Mapping for Multispectral Patterns

2.1.1. Exploiting Material Emissivity

The challenges when generating a MAP are that we have to consider the patterns seen from both types of sensors at once, and we have to control a complex thermal pattern. In particular, it is hard to control the pattern that appears in the thermal sensor. To solve this problem, we focus on exploiting the material emissivity rather than controlling heat. According to Stefan-Boltzmann law, the amount of energy radiated by an object is proportional to the emissivity of the material at the same temperature. Therefore, various intensities on the thermal sensor can be expressed just by arranging the materials of different emissivity without any temperature control. Following the proposed CSM, each spectral patttern can be easily determined by the investigated RGB and thermal pixel values of predefined materials. Also, our proposed method disposes of different materials to represent thermal patterns, which ulti-

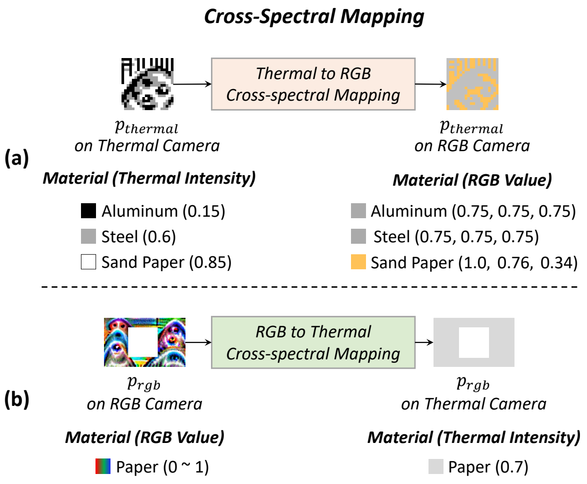
Cross-Spectral Mapping
(a)
Thermal to RGB Cross-spectral Mapping
$p_{thermal}$ on Thermal Camera
$p_{thermal}$ on RGB Camera
Material (Thermal Intensity)
Aluminum (0.15)
Steel (0.6)
Sand Paper (0.85)
Material (RGB Value)
Aluminum (0.75, 0.75, 0.75)
Steel (0.75, 0.75, 0.75)
Sand Paper (1.0, 0.76, 0.34)
(b)
RGB to Thermal Cross-spectral Mapping
p_{rgb} on RGB Camera
p_{rgb} on Thermal Camera
Material (RGB Value)
Paper (0 ~ 1)
Material (Thermal Intensity)
Paper (0.7)

Fig. 2. Visual explanation of Cross-spectral Mapping

mately prevents patttern breakage due to thermal conduction. This allows the implementation of precise thermal adversarial pattterns thus boost the attack performance dramatically.

2.1.2. Cross-Spectral Mapping

In the proposed Cross-spectral Mapping (CSM), the patch of each sensor is transformed into views in the opposite sensors. In other words, a thermal patch ($p_{thermal}$) is transformed into the pattern shown in the RGB senor, and RGB patch (p_{RGB}) is also transformed into the pattern shown in the thermal sensor. To this end, we predefine three materials that have different emissivity (Aluminum, Steel, and Sand paper). Then, we investigate the intensity and rgb values of each material. We measured these under room temperature $25°C$ and bright illumination condition. Aluminum, steel, and sand paper have 0.15, 0.6, 0.85 intensity values on the thermal camera, respectively. Also, in the case of paper which the rgb patch is printed on, the intensity is 0.7. Fig. 2 shows the visual explanation of CSM. As shown in the figure, $p_{thermal}$ is converted in to a value that appears on the RGB sensor through CSM. For example, if the pixel intensity is close to 0.85 (sand paper), it is converted to (1.0, 0.76, 0.34) pixel value. Then it is unified with the RGB patch with Multi Spectral Patch Union process shown in Fig. 1.

2.2. Multispectral Patch Loss

To optimize the Multispectral Adversarial Patch(MAP), we propose multispectral patch loss, which can be written as follow:

$$\mathcal{L}_{patch} = \mathcal{L}_{thermal} + \mathcal{L}_{rgb} + \mathcal{L}_{obj}, \qquad (1)$$

where $\mathcal{L}_{thermal}$ is a loss function that optimizes $p_{thermal}$, $\mathcal{L}_{rgb}$ is a loss function that optimizes p_{rgb}, and $\mathcal{L}_{obj}$ is a loss function that minimize the objectness scores of RPN and Head network in the Faster-RCNN network [11, 12]. In the case of $\mathcal{L}_{thermal}$, the loss function can be written as follows in detail:

$$\mathcal{L}_{thermal} = \sum_{p^{i,j} \in p_{thermal}} \min_{m \in M} \left| p^{i,j} - m \right| + \mathcal{L}_{tv}, \qquad (2)$$

where $p^{i,j}$ denotes the (i, j) pixel of the thermal patch $(p_{thermal})$. The left term of the equation represents the Material Emissivity(ME), and the right term represents the total-variation loss that reduces the noisy pattern. Minimizing ME loss makes sure that the intensity in the thermal pattern can converge to the predefined material intensity. Our goal is to generate a thermal pattern that can only consist of the intensities of our predefined materials. Since we select three materials that have 0.15, 0.6, 0.85 thermal intensity, M can be represented as $M = \{0.15, 0.6, 0.85\}$. Minimizing the ME loss makes the pixel values of the generated thermal pattern converge to these intensity values so that it is possible to manufacture with our predefined materials.

In the case of $\mathcal{L}_{rgb}$, it consists of non-printable score loss and total-variation loss proposed in [11]. By minimizing $\mathcal{L}_{rgb}$, we can print the optimized RGB patch on paper.

3. EXPERIMENTS

3.1. Experiment Settings

3.1.1. Dataset Collection

We collected a multispectral person detection dataset. Unlike existing public datasets [19,20], the majority of images in our dataset contains large size persons taller than 115 pixels to be suitable for generating patches. To this end, we collected 1,500 rgb-thermal image pairs in various scenes at day and night time to consider different light conditions. The camera equipment is FLIR duo Pro R which has a RGB-thermal dual sensor. The resolution of the RGB camera is 4000×3000, and the resolution of the thermal camera is 640×512. The field of view (FoV) is $56° \times 45°$ for the RGB camera and $32° \times 26°$ for the thermal camera. The spectral band of thermal camera is $7.5 \sim 13.5 \mu m$. Each pair of images is cropped so it has the same field of view ($32° \times 26°$) and size (640×512). Then, we calibrate each image pair. We used 1200 image pairs (day/night 600:600) for training and 300 image pairs (day/night 150:150) for testing. Also, we captured 200 image pairs (day/night 100:100) with physically generated MAP to evaluate the robustness of person detector in the physical space.

3.1.2. Object Detection Model

Before we generate MAP, we train the multispectral person detection network. We use the detection model proposed in [6] with VGG16 backbone network. After training the detection network, we optimize a MAP and then we apply the patch digitally and physically in the image.

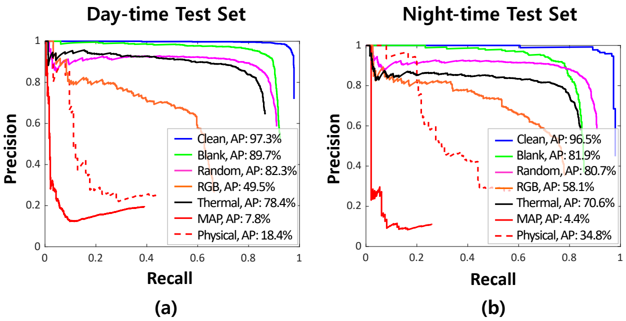

Fig. 3. PR-curve of the different patch types on Daytime images and Nighttime images.

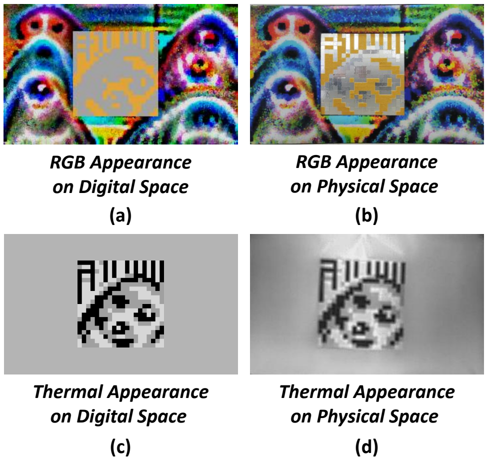
RGB Appearance on Digital Space
(a)
RGB Appearance on Physical Space
(b)
Thermal Appearance on Digital Space
(c)
Thermal Appearance on Physical Space
(d)

Fig. 4. Generated MAP. (a) the appearance of the pattern shown in the RGB camera in the digital space. (b) result of physically generated patch captured by the RGB camera. (c) the appearance of the pattern shown in the thermal camera in the digital space. (d) result of physically generated patch captured by the thermal camera.

3.2. Quantitative Result

Fig. 3 shows the precision-recall (PR) curve according to the patch type. Average Precision (AP) is calculated by computing the area under the PR curve. (a) is a PR curve comparison with day-time images. (b) is a PR curve comparison with night-time images. As shown in Fig. 3 (a), without patch (Clean) result shows high AP score (97.3). Also, when applying blank patch that has 0.5 pixel value, it does not reduce the AP score dramatically. Similar results are shown with random patch that has random pixel values. Furthermore, when applying only a RGB or a Thermal patch, AP is reduced by 43.2 and 22.4 respectively. It is necessary to generate MAP, since a single sensor adversarial patch would not be enough to fool the multispectral person detector.

Therefore, we generate MAP and apply it both in physical and digital space. As shown in Fig 3 (a) , when applying the MAP, the AP score is decreased dramatically (97.3 to 7.8) in digital space. There is an AP score drop of 89.5. Furthermore, when we physically applied the patch (Physical), AP score still decreased dramatically (97.3 to 18.4). It can be interpreted that the generated MAP can effectively hide persons both in the digital space and physical space. Also, the result of Fig. 3 (b) shows that the generated MAP works well even in the low light condition.

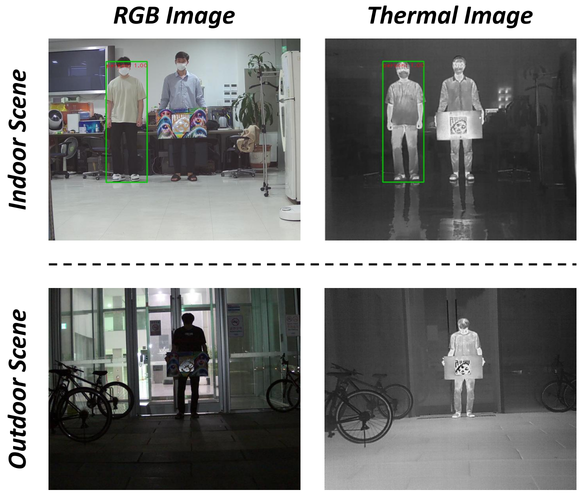

Fig. 5. Detection results with physical MAP. The person who holds the patch can not be detected. The MAP can successfully hide persons from the detector.

3.3. Qualitative Result

Fig. 4 shows the optimized MAP. (a) is the appearance of the pattern shown in the RGB camera in the digital space. (b) is a result of the physically generated patch captured by the RGB camera. As shown in the figure, since (a) and (b) are similar, it can be interpreted that our proposed CSM work so that the pattern is well expressed in the physical space. (c) is the appearance of the pattern shown in the thermal camera in the digital space. (d) is the result of the physically generated patch captured by the thermal camera. As shown in the figure, the thermal pattern is sophisticatedly expressed. Also, the pattern is mapped well into thermal camera view.

Fig. 5 shows person detection results. The person in the figure holds the physically generated MAP. As shown in the figure, the generated MAP successfully hides the person even in the poor light condition.

4. CONCLUSION

In this paper, we evaluated the robustness of multispectral person detectors in the physical world. A novel Multispectral Adversarial Patch (MAP) generation framework has been proposed. In the framework, we optimized MAP with Cross-Spectral Mapping(CSM) and Material Emissivity(ME) loss. Extensive experiments with our collected dataset show that the generated MAP could fool the multispectral person detector both in physical and digital spaces. We believe that our novel framework could be used to evaluate the robustness of DNNs based detection models for physical world applications.

5. REFERENCES

[1] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis, “A survey on 3d object detection methods for autonomous driving applications,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[2] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang, “Gs3d: An efficient 3d object detection framework for autonomous driving,” in *CVPR*, 2019, pp. 1019–1028.

[3] Shizhou Zhang, De Cheng, Yihong Gong, Dahu Shi, Xi Qiu, Yong Xia, and Yanning Zhang, “Pedestrian search in surveillance videos by learning discriminative deep features,” *Neurocomputing*, vol. 283, pp. 120–128, 2018.

[4] Xudong Li, Mao Ye, Yiguang Liu, Feng Zhang, Dan Liu, and Song Tang, “Accurate object detection using memory-based models in surveillance scenes,” *Pattern Recognition*, vol. 67, pp. 73–84, 2017.

[5] Tianrui Liu, Jun-Jie Huang, Tianhong Dai, Guangyu Ren, and Tania Stathaki, “Gated multi-layer convolutional feature extraction network for robust pedestrian detection,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3867–3871.

[6] Jung Uk Kim, Sungjune Park, and Yong Man Ro, “Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[7] Sungjune Park, Jung Uk Kim, Yeon Gyun Kim, Sang-Keun Moon, and Yong Man Ro, “Robust multispectral pedestrian detection via uncertainty-aware cross-modal learning,” in *International Conference on Multimedia Modeling*. Springer, 2021, pp. 391–402.

[8] Xiyu Yan, Xuesong Chen, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Feng Zheng, “Hijacking tracker: A powerful adversarial attack on visual tracking,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2897–2901.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[10] Gengxing Wang, Xinyuan Chen, and Chang Xu, “Adversarial watermarking to attack deep neural networks,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1962–1966.

[11] Simen Thys, Wiebe Van Ranst, and Toon Goedemé, “Fooling automated surveillance cameras: adversarial patches to attack person detection,” in *CVPR Workshops*, 2019, pp. 0–0.

[12] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin, “Adversarial t-shirt! evading person detectors in a physical world,” in *ECCV*. Springer, 2020, pp. 665–681.

[13] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein, “Making an invisibility cloak: Real world adversarial attacks on object detectors,” in *ECCV*. Springer, 2020, pp. 1–17.

[14] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu, “Fooling thermal infrared pedestrian detectors in real world using small bulbs,” in *AAAI*, 2021, vol. 35, pp. 3616–3624.

[15] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro, “Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning,” *arXiv preprint arXiv:2005.10987*, 2020.

[16] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro, “Towards robust training of multi-sensor data fusion network against adversarial examples in semantic segmentation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4710–4714.

[17] Hong Joo Lee and Yong Man Ro, “Adversarially robust multi-sensor fusion model training via random feature fusion for semantic segmentation,” in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 339–343.

[18] Sheeba Lal, Saeed Ur Rehman, Jamal Hussain Shah, Talha Meraj, Hafiz Tayyab Rauf, Robertas Damaševičius, Mazin Abed Mohammed, and Karrar Hameed Abdulkareem, “Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition,” *Sensors*, vol. 21, no. 11, pp. 3922, 2021.

[19] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López, “Pedestrian detection at day/night time with visible and fir cameras: A comparison,” *Sensors*, vol. 16, no. 6, pp. 820, 2016.

[20] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.