HAL open science

Bispectral Pedestrian Detection Augmented with Saliency Maps using Transformer

Mohamed Amine Marnissi, Ikram Hattab, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara

To cite this version:

Mohamed Amine Marnissi, Ikram Hattab, Hajer Fradi, Anis Sahbani, Najoua Essoukri Ben Amara. Bispectral Pedestrian Detection Augmented with Saliency Maps using Transformer. pp.275–284, 2022, VISIGRAPP (5: VISAPP), 10.5220/0010913000003124 . hal-03909902

HAL Id: hal-03909902

https://hal.sorbonne-universite.fr/hal-03909902v1

Submitted on 2 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bispectral Pedestrian Detection Augmented with Saliency Maps using Transformer

Mohamed Amine Marnissi[1,4,5], Ikram Hattab[2,5], Hajer Fradi[2,4], Anis Sahbani[5,6] and Najoua Essoukri Ben Amara[3,4]

[1]*Ecole Nationale d'Ingénieurs de Sfax, Université de Sfax, 3038, Sfax, Tunisie;*

[2]*Institut Supérieur des Sciences Appliquées et de Technologie, Université de Sousse, 4023, Sousse, Tunisie;*

[3]*Ecole Nationale d'Ingénieurs de Sousse, Université de Sousse, 4023, Sousse, Tunisie;*

[4]*LATIS- Laboratory of Advanced Technology and Intelligent Systems, Université de Sousse, 4023, Sousse, Tunisie;*

[5]*Sorbonne Université, CNRS, Ins titule for intelligent Systems and Robotics (ISIR), Paris, France*

[6]*Enova Robotics S.A.*

medamine.marnissi@eniso.u-sousse.tn

Keywords: Deep learning, object detection, YOLO, visible and thermal cameras, robotic vision, saliency map, transformer, features fusion.

Abstract: In this paper, we focus on the problem of automatic pedestrian detection for surveillance applications. Particularly, the main goal is to perform real-time detection from both visible and thermal cameras for complementary aspects. To handle that, a fusion network that uses features from both inputs and performs augmentation by means of visual saliency transformation is proposed. This fusion process is incorporated into YOLO-v3 as base architecture. The resulting detection model is trained in a paired setting in order to improve the results compared to the detection of each single input. To prove the effectiveness of the proposed fusion framework, several experiments are conducted on KAIST multi-spectral dataset. From the obtained results, it has been shown superior results compared to single inputs and to other fusion schemes. The proposed approach has also the advantage of a very low computational cost, which is quite important for real-time applications. To prove that, additional tests on a security robot are presented as well.

1 Introduction

Pedestrian detection plays a crucial role in the computer vision community. It has been extensively studied in a wide range of applications such as autonomous driving, video surveillance, human activity understanding, and robot vision (Liu et al., 2019). Over the past decade, a significant progress has been achieved in this field using deep learning models. Also, it has been mostly studied in the visible domain thanks to the availability of visible cameras (Lin et al., 2020; Ouyang et al., 2016; Liu et al., 2019; Fradi et al., 2018; Nagy and Czúni, 2021).

However, it is commonly known that visible cameras are not effective enough at nighttime, bad lighting conditions, total darkness, or in adverse weather conditions. In such situations, thermal cameras can be instead used since they could better discern warmer target objects than other surrounding ones (Marnissi et al., 2021b; Dai et al., 2021; Kieu et al., 2020). For the aforementioned reasons, earlier attempts in the field made use of both cameras, which is referred to as bispectral vision. This solution has been adopted to substitute the use of two detectors; each one is trained on the corresponding domain accordingly. Combining information from both cameras allows to deal with vast range of weather and lighting conditions.

Usually, these bispectral detectors are based on complex network architectures compared to the detection from one single spectrum (Li et al., 2018; Konig et al., 2017; Guan et al., 2019). Moreover, these detectors rely in most cases on aligned thermal and visible sensors at inference time (Hwang et al., 2015a). All these factors limit their feasibility in real-time applications. Because of the aforementioned reasons, we focus, in this paper, on the problem of pedestrian detection from both domains and for real-time applications. To handle that we propose a new approach based on feature fusion on YOLOv3 as base architecture. To further improve the detection results, an aug-

mentation with saliency maps using visual saliency transformation is applied. This augmentation is performed through a channel replacement for both inputs. The resulting fusion architecture is considered as the main contribution of this present paper. This has the advantage of improving the overall performance by considering both cues from single inputs with the corresponding saliency maps. Also, the use of visual saliency transformer on thermal images in order to generate deep saliency maps is novel. Moreover, the proposed approach is of low computational cost which is highly required for real-time applications.

The remainder of the paper is organized as follows: in Section 2, an overview of the existing works for object detection, multispectral detection and salient detectors is presented. Then, our proposed approach of feature fusion from thermal and visible cameras by augmentation with deep saliency maps is detailed in Section 3. The conducted experiments and the obtained results are discussed in Section 4. Finally, we briefly conclude and give an outlook of possible future works in Section 5.

2 Related work

In this section, we give an overview of object detection methods, precisely single spectrum object detectors (visible or thermal) and multispectral object detectors. This overview includes salient object detectors as well.

2.1 Object Detection

The first object detection models were developed with a set of hand-crafted feature extractors such as Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005), Deformable Part-based Models (DPM) (Forsyth, 2014) and Viola-Jones detector (Viola and Jones, 2001). These models mostly suffer from poor performance on unfamiliar datasets. Afterwards, deep networks have promoted the research in many computer vision applications, including object detection in terms of results and inference time.

Current detectors based on deep models can be divided into two categories: two-stage and one-stage detectors. In the first category, the detection requires two stages: the first one consists of generating a set of regions of interest through the regional proposal network (RPN) and the second one aims at detecting objects of each proposal. It is the case of R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster-RCNN (Ren et al., 2015), and Mask R-CNN (He et al., 2017a) models. Different from two-stage models, other detectors such as YOLO family (Redmon and Farhadi, 2018a; Wang et al., 2021), SSD (Liu et al., 2016b), and RetinaNet (Lin et al., 2017) models allow to skip the region proposal step and to perform the detection directly on a dense sample of possible locations.

2.2 Pedestrian Detection in Single Spectrum

Most of the existing pedestrian detection models are trained on visible images captured under good lighting conditions. Some of these models are designed to address the problems of occlusion (Zhou et al., 2019; Wu et al., 2020) and scale variance (Jiao et al., 2020; Lin et al., 2018). To mitigate occlusion problems, in (Zhou et al., 2019) a discriminative feature transformation was employed, in which the distance between the occluded and non-occluded pedestrian examples is reduced. TFAN (Wu et al., 2020) combined features from every current image and nearby images to improve the representation of pedestrian features by exploiting temporal information from occluded pedestrians. Other studies have been carried out in the field to deal with the problem of scale variance, for example, MDFL (Lin et al., 2018) is a multi-grain deep feature learning that was adopted to solve occlusion and small scale challenges. In the same context, the Pose-Embedding Network (PEN) (Jiao et al., 2020) was introduced to enhance the visual representation of pedestrian through human pose information.

While these methods improved the performance of pedestrian detection, they have been applied in the visible spectrum. However, it is commonly known that the visible light spectrum is not convenient in bad light and weather conditions (Kim et al., 2021). For this reason, other attempts using thermal imagery have been conducted for pedestrian detection. For instance, in (Ghose et al., 2019) thermal images augmented with their saliency maps to serve as an attention mechanism for the pedestrian detector are employed. From the obtained results, it has been shown that the saliency maps provide complementary information to the pedestrian detector resulting in a significant improvement in performance over the baseline approach. Also, an enhancement architecture based on Generative Adversarial Networks, and composed of contrast enhancement and denoising modules is proposed in (Marnissi et al., 2021a). The proposed architecture has shown its advantage to enhance the overall thermal image quality and to further improve the detection results.

2.3 Multispectral Pedestrian Detection

To deal with vast range of weather and lighting conditions e.g. rain, fog, daytime and nighttime, multispectral detectors that use complementary information from thermal and visible images in order to enhance the visual representation of pedestrians have been proposed. (Hwang et al., 2015a) is one of the first works that took advantage of aligned image pairs for pedestrian detection through a multispectral aggregation channel features (ACF). Following the fast development of deep learning, the performance of multi-spectral pedestrian detectors is greatly improved. For instance, MSDS-RCNN (Li et al., 2018) is fusion method composed of a multispectral proposal network (MPN) and a multispectral classification network (MCN). Park *et al.* designed in (Park et al., 2018) a convolutional neural network architecture that has as inputs color, thermal and fusion features. To combine them, a channel weighting fusion layer and accumulated probability fusion by highlighting the most robust features are used. In (Cao et al., 2019), semantic segmentation is performed by taking advantage of the box-level supervision that guides the networks to distinguish between the pedestrian and the background. An illumination-aware network is added to the detection framework by adjusting the weight according to the illumination score in (Li et al., 2019). In (Zhou et al., 2020), an illumination-aware feature alignment that selects features based also on an illumination score is proposed. The resulting Modality Balance Network (MBNet) facilitates the optimization process in a balanced manner.

2.4 Salient Object Detection

During the last decades, several methods have been proposed for saliency detection, which aim to highlight the most apparent objects in the image. Traditional saliency object detectors include some processing methods such as global contrast, local contrast and other features e.g. color and texture (Cheng et al., 2014; Borji et al., 2019). Recent works use instead CNN architectures. (He et al., 2017b) is one of the first algorithms that used U-Net architecture by including compression and expansion paths. BASNet (Qin et al., 2019) is a prediction and refinement architecture that uses hybrid loss for boundary-aware salient object detection. The contextual architecture of spatial attenuation proposed in (Hu et al., 2020) is used to allow contextual features with different attenuation factors to be translated independently and repeatedly. In order to decompose RGB images into highlight and detail streams, a label decoupling method has been presented in (Wei et al., 2020). The model was also supervised and a feature interaction network was designed. The two branchs iteratively exchange information in order to generate accurate saliency maps. Related work includes (Liu et al., 2020), where a dynamic feature integration approach to automatically study feature combinations in relation to input tasks is proposed. R3-Net (Deng et al., 2018) presented a refinement network that comprises a succession of residual refinement blocks (RRBs) for saliency detection on a single image. The main idea is to take advantage of RRBs to recurrently learn the difference between the coarse saliency map and the ground truth by leveraging both low-level and high-level features.

3 PROPOSED APPROACH

In this paper, we investigate the use of bispectral visible and thermal images for pedestrian detection in various lighting conditions. It is about bispectral fusion augmented with saliency maps for better detection. Particularly, we employ a fusion scheme that uses features from paired visible and thermal images with their corresponding deep saliency maps. The visible and thermal image channels are meant to be complementary, where visible images tend to provide color and texture details, while thermal images are sensitive to object temperature, which can be very useful in different light conditions. The overall architecture is shown in Figure 1. In the following we are detailing each of these architecture components.

3.1 Base architecture

YOLO (You Only Look Once) is one of the most used real-time detectors that belongs to the category of one-stage. Our proposed approach is developed using YOLOv3 (Redmon and Farhadi, 2018b) as base architecture to build a deep neural network combining information from both images and to perform detection in real-time applications. In this model, Darknet53 which acts as feature extractor is mainly composed of 3x3 and 1x1 filters with skip connection as residual network and consists of five blocks. Inspired by Feature Pyramid Network (FPN) structure, YOLOv3 makes three prediction heads: small (grid size of 13 x 13), medium (26 x 26) and large (52 x 52) for multi-scale detection. The output result includes bounding box coordinates, confidence scores for each class, and object confidence (1 for object and 0 for non-object). The YOLO head outputs are post-processed by non-maximum suppression to remove

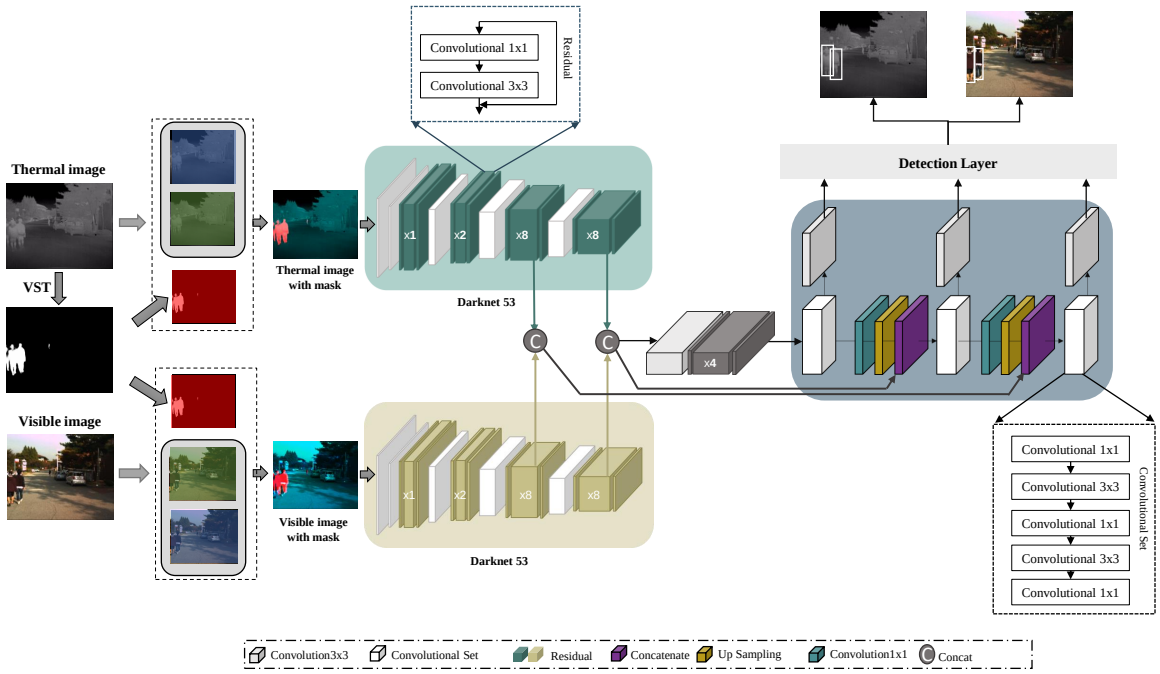

Figure 1: The architecture of the proposed fusion scheme of visible and thermal images augmented by saliency maps for pedestrian detection.

the overlapping bounding boxes.

3.2 Augmentation with saliency maps

We propose to enhance visible and thermal images with their saliency maps by giving better cues for object detection during day and night time. These additional saliency maps are considered to guide the detection process by highlighting some information at pixel-level. This highlight allows to illuminate the most visible parts of the image, while preserving the textural information. Since in our particular case, input images are aligned, we choose to calculate the saliency maps of thermal images. Practically, we replace one duplicated channel of the 3-channel thermal image by the corresponding saliency map. The same map is introduced to the visible input through a channel replacement before feeding them to the network as can be seen in Figure 1.

In order to highlight some particular regions of the thermal images, an unified Visual Saliency Transformation (VST) model is adapted. It was originally proposed in (Liu et al., 2021), with color images and their corresponding depth maps as inputs. In our case, VST is used to generate a deep saliency map based on the thermal image and its static saliency map as shown in Figure 2. The model first uses an encoder to generate multi-level tokens from the input image patch sequence using the T2T-ViT model (Yuan et al.,

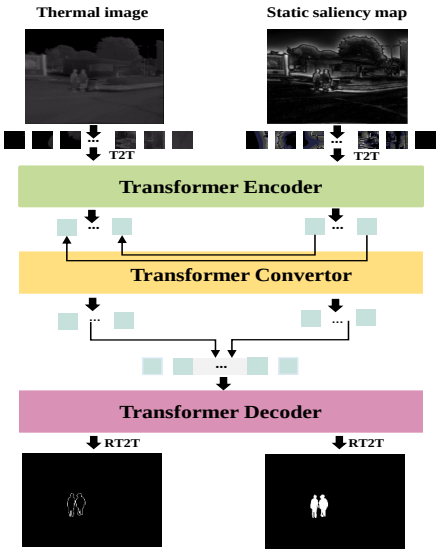

Figure 2: Architecture of Visual Salient Transformer (VST) for generating deep saliency maps.

2021). Then, a converter is applied to transform the patch tokens into the decoder space. In addition, cross-modal information fusion for the thermal image and the static saliency map is performed. Finally, a decoder predicts the deep saliency map through the patch-task-attention mechanism. Following (Liu et al., 2021), RT2T transformation is also used to pro-

(a) Input images (b) Deep saliency maps (c) Concatenated channels

Figure 3: Example results of the generated deep saliency maps concatenated with the input images.

gressively upsample the patch tokens. Figure 3 shows some results of the generated deep saliency maps and a concatenation with the input image from each domain.

3.3 Fusion scheme

At this stage, we intend to apply a fusion scheme of the two input images from both cameras for producing bispectral images augmented with saliency maps in order to enable better detection. After adding the saliency maps as input, we need to choose at which level the fusion can be performed. Since the backbone network as any CNN architecture embed different representations at different abstraction levels, it is not trivial to choose the most relevant layers or blocks for feature extraction. It has been demonstrated in some previous studies in image retrieval and texture classification (Jun et al., 2019; Fradi et al., 2021) that intermediate layers, mainly the layer before the last one mostly performs better than other layers or blocks. Following these studies, each visible and thermal input after augmentation is passed through the first four residual blocks. Once concatenated, the model operates on shared feature maps at the fifth block of the backbone network. To further justify this choice, we consider other fusion schemes for comparisons. As a first solution, the images could be merged at the input stage, i.e. combining information of the two raw images. It results in a new 6-channel image by concatenating visible and thermal images simultaneously. This fusion scheme is referred to as input fusion. Afterwards, we realize the early-fusion architecture that implements fusion operation before the first block. Then, the feature maps are fused after the mid-level blocks, namely block1, block2, block3, for halfway fusion. At the last stage, we implement the late-fusion architecture, where we combine the whole blocks of Darknet architecture for each input.

4 Experimental results

4.1 Dataset and Experiments

The proposed approach is evaluated on KAIST (Hwang et al., 2015a), a publically available dataset commonly used in the field of object detection. It is one of the largest multi-spectral pedestrian dataset composed of temporally and spatially aligned visible and Long-Wave Infrared (LWIR) images under adverse illumination conditions, day and night. The visible camera generates 640 x 480 pixels resolution with a 103.6 vertical field of view. However, the thermal camera has 320 x 256 pixels resolution with a 39 vertical field of view. In total, the dataset approximately consists of 95k frames on urban traffic environment and of dense annotations for 1182 different pedestrians. It is divided into a training set of 50.2k images from Set 00 to Set 05, and a test set of 45.1k images from Set 06 to Set 11.

In our case, we select every 3 frames from training sets and every 20 frames from testing sets, and we only consider the non-occluded, non-truncated and large instances (> 50). This results in a training set of 7601 images for both thermal and visible sets, and a testing set of 2252 (1455 day and 797 night). For training the deep saliency network, pixel-level annotations as masks are required. Hence, we used a subset of KAIST published in (Ghose et al., 2019), where mask annotations are provided for 1702 images; 913 day images and 789 night images. Mean Average Precision (mAP) is used to evaluate the performance of the proposed detector at Intersection Over Union (IOU) equal to 0.5 regarding the ground truth. The ob-

tained results are compared to those obtained in each domain separately. Also, comparisons to different fusion schemes (input, early, halfway, and late) and existing methods are considered.

4.2 Implementation details

Experiments were conducted on NVIDIA TITAN RTX GPU with 24 GB RAM. For YOLOv3, we used PyTorch framework which supports GPU computing. Our custom model YOLOV3 was implemented based on Darknet framework but with some configurations and modifications in order to adapt the model to KAIST dataset, to consider the fusion schemes and to include saliency maps as well. We trained our model on 30 epochs with a mini-batch size equal to 8. As optimizer, we used stochastic gradient descent (SGD) with an initial learning rate of 0.0001. For deep saliency maps generation, VST is trained on batch size of 8, and 40.000 as total training steps.

4.3 Comparison results

We evaluate our proposed approach on the test set of KAIST dataset and we make comparison with single input from thermal or visible cameras, with the different fusion methods suggested for producing the bispectral images and with/without saliency maps as presented in Table 1. As reported in the table, using our proposed approach which consists of fusing feature maps at the fourth block with saliency map augmentation, we got the best results 75.8% in terms of mAP, with a margin of 17.7% and 9.9% to visible and thermal inputs, respectively. This performance increase compared to single spectrum is added to the advantage of using one single model for both domains. This is of significant interest since it allows faster execution time and less consumption of resources, which are highly relevant in real-time applications.

To further prove the overall performance of the proposed approach, we compare the obtained results to the different fusion schemes. In this table Halfway fusion-i refers to fusion at i-th block. The corresponding results vary from 59.4% to 63.9% in terms of mAP at different blocks. Precisely, it is shown that halfway fusion-4 gives better results compared to other halfway fusion levels and to input and late fusions, which corresponds to our observation stated from the beginning of the paper. Hence, this fusion level is selected to be coupled with saliency maps, which results in a overall performance of our proposed approach.

Table 1: Detection performance comparison using single input (thermal or visible), using different fusion levels, with/without augmentation with saliency maps. All in terms of mAP on KAIST dataset.

Methods	**Day**	**Night**	**All**
Visible input	64.4	42.3	58.1
Thermal input	62.7	72.6	65.9
Input-fusion	69.2	46.0	62.8
Early-fusion	67.8	42.1	60.4
Halfway fusion-2	68.5	39.2	59.4
Halfway fusion-3	69.6	44.0	62.1
Halfway fusion-4	69.3	49.5	63.4
Late-fusion	67.7	48.8	62.4
Ours	72.6	79.0	**75.8**

The corresponding qualitative results on some sample images from KAIST datasets are shown in Figure 4. These results also indicate the performance increase by our detector compared to other single inputs. Precisely, in the sample visual results, it is shown that some false positives and false negatives results are corrected by the proposed detector compared to single visible and thermal models.

It is important to highlight that in addition to the increase of the performance compared to other methods, the inference time on a test image using our proposed approach is equal to 0.019s which is a quite interesting result known that the inference time on one single input is equal to 0.013s. Moreover, we consider other state-of-the-art methods for comparisons, namely, ACF(Hwang et al., 2015b), Halfway Fusion (Faster RCNN) (Liu et al., 2016a), Fusion RPN+BF(Konig et al., 2017), IAF R-CNN (Li et al., 2019), and AR-CNN (Zhang et al., 2019). The comparison results are shown in Table 2, with the corresponding inference time of each method.

As depicted in the table, the obtained MR of our proposed approach is better in some cases and competitive in other cases, however the computational cost is significantly reduced compared to other methods. As a result, we conclude that our obtained results are quite satisfactory while keeping a limited runtime. Thus our approach is more efficient compared to other two-stage detectors using complex fusion architecture in most cases.

4.4 Deployment on a security robot

By means of different sessions of acquisition, we built our dataset using "Pearl Guard"[1] security robot that has autonomous capability to navigate in different environments and is equipped by two cameras; ther-

[1] https://enovarobotics.eu/pguard/

Figure 4: Qualitative results of our proposed method on four sample images from KAIST. From top to bottom rows, visible images with their corresponding thermal images and the resulting saliency maps are shown. For each sample image, the detection results are shown in different color and the corresponding annotated bounding boxes in white color.

mal and visible cameras. This dataset is composed of 4615 pairs of visible and thermal images, that are captured in different lighting conditions by means of optical zoom x32 and thermal cameras. The dataset is split to 2956 training images and 1659 test images. Figure 5 presents some examples of the robot dataset. It is worth to note that this dataset will be publically available for research directions.

Figure 5: Visible and thermal samples from the robot dataset

We annotated these images using a semi-automatic strategy. We first apply Faster R-CNN (Ren et al., 2015) to get some preliminary bounding boxes. This strategy helps the annotation process since it is time consuming. These first annotations are corrected and completed to generate the ground truth. It is important to note that acquired images suffer from a misalignment problem due to the spatial shift between cameras, different zooms and inference time. Figure 6 illustrates this problem by displaying the two bounding boxes at the same position in the two images in order to highlight the shift.

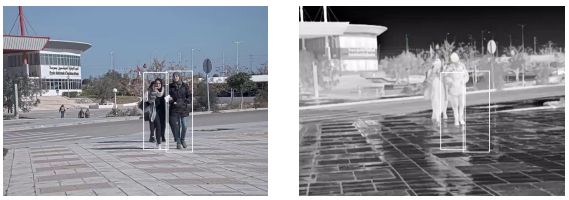

Figure 6: Illustration of the misalignment problem between thermal and visible cameras of the robot

To conduct real-tests on the robot, the trained model on KAIST dataset is fine-tuned on a subset of the data acquired by the robot.We made this choice of transfer learning to harness the advantage of the

Table 2: Comparisons with the state-of-the-art methods on KAIST dataset in terms of miss rate (MR) and speed according to a given hardware specification. For MR values, lower is better

Methods	**MR**	**Speed (Platform)**
ACF (Hwang et al., 2015b)	47.32	2.73 (MATLAB)
Halfway Fusion (Faster RCNN) (Liu et al., 2016a)	25.75	0.43 (TITAN X)
Fusion RPN+BF (Konig et al., 2017)	18.29	0.80 (MATLAB)
Ours	17.24	0.019 (TITAN X)
IAF R-CNN (Li et al., 2019)	15.73	0.21 (TITAN X)
IATDNN + IASS(Guan et al., 2019)	14.95	0.25 (TITAN X)
MSDS-RCNN (Li et al., 2018)	11.34	0.22 (1080 Ti)
AR-CNN (Zhang et al., 2019)	9.34	0.12 (TITAN X)

large-scale of KAIST dataset and to limit the problem of misalignment using the robot as well. As shown on the user interface in Figure 7, once the target persons are detected using our proposed fusion scheme, the robot is controlled through a set of commands to follow them. The tracking process depends on the position and the size of the detected bounding boxes in the frame. The distance between the frame center and the center of the detected bounding box is considered to determine the the robot movements. Precisely, the green arrows indicate the horizontal shift to follow the two detected persons.

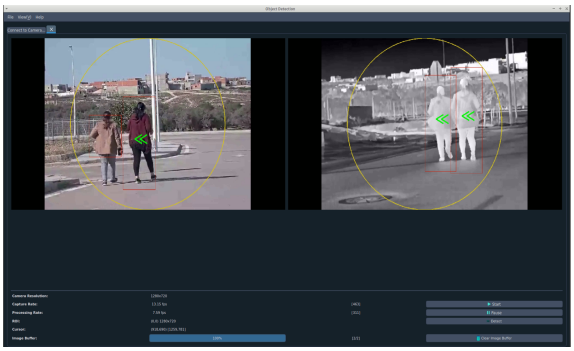

Figure 7: Real-time streaming from the thermal and visible cameras of the security robot with the detected boundary boxes.

5 Conclusion

In this paper, we proposed a novel fusion architecture based on bispectral images and augmented with saliency maps using transformer. By means of tests on KAIST dataset, the effectiveness of the proposed architecture is proven by obtaining better quantitative and qualitative results compared to single inputs and to other fusion schemes. By comparisons to the state-of-the-art methods for fusion, it has been demonstrated that our proposed approach achieves competitive results while keeping a very low computational cost. In addition, real tests of detection on a security robot have been conduced in order to follow target persons. As perspectives, other object detectors could be investigated but always from the one-stage category such as YOLOr, YOLOx, SDD and RetinaNet. Also, the results on the robot could be improved by augmenting the amount of data and by handling the misalignment problem.

Acknowledgments

This work has been supported by the DGVR research fund from the Tunisian Ministry of Higher Education and Scientific Research that is gratefully acknowledged.

REFERENCES

Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., and Li, J. (2019). Salient object detection: A survey. *Computational visual media*, 5(2):117–150.

Cao, Y., Guan, D., Wu, Y., Yang, J., Cao, Y., and Yang, M. Y. (2019). Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS journal of photogrammetry and remote sensing*, 150:70–79.

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., and Hu, S.-M. (2014). Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582.

Dai, X., Yuan, X., and Wei, X. (2021). Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., and Heng, P.-A. (2018). R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 684–690. AAAI Press.

Forsyth, D. (2014). Object detection with discriminatively trained part-based models. *Computer*, 47(02):6–7.

Fradi, H., Bracco, L., Canino, F., and Dugelay, J.-L. (2018). Autonomous person detection and tracking framework using unmanned aerial vehicles (uavs). In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1047–1051. IEEE.

Fradi, H., Fradi, A., and Dugelay, J.-L. (2021). Multi-layer feature fusion and selection from convolutional neural networks for texture classification. In *VISIGRAPP (4: VISAPP)*, pages 574–581.

Ghose, D., Desai, S. M., Bhattacharya, S., Chakraborty, D., Fiterau, M., and Rahman, T. (2019). Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Guan, D., Cao, Y., Yang, J., Cao, Y., and Yang, M. Y. (2019). Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017a). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

He, S., Jiao, J., Zhang, X., Han, G., and Lau, R. W. (2017b). Delving into salient object subitizing and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1059–1067.

Hu, X., Fu, C.-W., Zhu, L., Wang, T., and Heng, P.-A. (2020). Sac-net: Spatial attenuation context for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1079–1090.

Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015a). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045.

Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015b). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045.

Jiao, Y., Yao, H., and Xu, C. (2020). Pen: Pose-embedding network for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):1150–1162.

Jun, H., Ko, B., Kim, Y., Kim, I., and Kim, J. (2019). Combination of multiple global descriptors for image retrieval. *arXiv preprint arXiv:1903.10663*.

Kieu, M., Bagdanov, A. D., Bertini, M., and Bimbo, A. D. (2020). Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *Computer Vision - ECCV*.

Kim, Y.-H., Shin, U., Park, J., and Kweon, I. S. (2021). Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504.

Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., and Teutsch, M. (2017). Fully convolutional region proposal networks for multispectral person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 49–56.

Li, C., Song, D., Tong, R., and Tang, M. (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference, BMVC*.

Li, C., Song, D., Tong, R., and Tang, M. (2019). Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171.

Lin, C., Lu, J., Wang, G., and Zhou, J. (2020). Graininess-aware deep feature learning for robust pedestrian detection. *IEEE transactions on image processing*, 29:3820–3834.

Lin, C., Lu, J., and Zhou, J. (2018). Multi-grained deep feature learning for robust pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3608–3621.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Liu, J., Zhang, S., Wang, S., and Metaxas, D. (2016a). Multispectral deep neural networks for pedestrian detection. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Liu, J.-J., Hou, Q., and Cheng, M.-M. (2020). Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *IEEE Transactions on Image Processing*, 29:8652–8667.

Liu, N., Zhang, N., Wan, K., Shao, L., and Han, J. (2021). Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4722–4732.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016b). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y. (2019). High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196.

Marnissi, M. A., Fradi, H., Sahbani, A., and Amara, N. E. B. (2021a). Thermal image enhancement using

generative adversarial network for pedestrian detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6509–6516. IEEE.

Marnissi, M. A., Fradi, H., Sahbani, A., and Amara, N. E. B. (2021b). Unsupervised thermal-to-visible domain adaptation method for pedestrian detection. *Pattern Recognition Letters*.

Nagy, A. M. and Czúni, L. (2021). Detecting object defects with fusioning convolutional siamese neural networks. In *VISIGRAPP (5: VISAPP)*, pages 157–163.

Ouyang, W., Zeng, X., and Wang, X. (2016). Learning mutual visibility relationship for pedestrian detection with a deep model. *International Journal of Computer Vision*, 120(1):14–27.

Park, K., Kim, S., and Sohn, K. (2018). Unified multispectral pedestrian detection based on probabilistic fusion networks. *Pattern Recognition*, 80:143–155.

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., and Jagersand, M. (2019). Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489.

Redmon, J. and Farhadi, A. (2018a). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Redmon, J. and Farhadi, A. (2018b). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE.

Wang, C.-Y., Yeh, I.-H., and Liao, H.-Y. M. (2021). You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*.

Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., and Tian, Q. (2020). Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13025–13034.

Wu, J., Zhou, C., Yang, M., Zhang, Q., Li, Y., and Yuan, J. (2020). Temporal-context enhanced detection of heavily occluded pedestrians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13430–13439.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., and Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., and Liu, Z. (2019). Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5127–5137.

Zhou, C., Yang, M., and Yuan, J. (2019). Discriminative feature transformation for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9557–9566.

Zhou, K., Chen, L., and Cao, X. (2020). Improving multispectral pedestrian detection by addressing modality imbalance problems. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 787–803. Springer.