Improved Object Detection in Thermal Imaging Through Context Enhancement and Information Fusion: A Case Study in Autonomous Driving

Junchi Bin[1,*], **Ran Zhang**[1,*], **Chengkai Zhang**[1,*]

Xiangrui Liu[1], **Shan Du**[1], **Erik Blasch**[2], **Zheng Liu**[1†]
[1]University of British Columbia, Okanagan Campus
[2]Air Force Research Lab
{zheng.liu}@ubc.ca

Abstract

With advances sensory technologies, autonomous driving systems incorporate more imaging sensors such as thermal cameras to enhance the capability of environmental perception beyond the visible spectrum. This paper proposes an integrated context enhancement and information fusion framework (CEIFF) to generate enhanced colorized synthetic visible (SVI) images from thermal images. The SVI and thermal images are fused for improved perception quality. The case study shows the effectiveness of the proposed CEIFF on a multimodal autonomous dataset.

1 Introduction

Object detection is one of the most important modules for environmental perception in autonomous driving. The object detection plays an essential role in classifying static roadblocks and dynamic intrusive objects that may cause severe incidents. Although the object detection is usually developed based on visible (VI) cameras, detection performance is poor under low-illuminated conditions such as cloud covers and lightning. On the contrary, the thermal imaging system can provide illumination-invariant images based on objects' temperature. Nonetheless, the thermal images usually lack texture and context details to classify foreground and background objects. Therefore, multimodal fusion-based object detection aims to combine the distinct advantages of VI and IR images for improving perception capability.

In past decades, several multimodal datasets such as KAIST [8] and FLIR ADAS [5] have been published to accelerate the research progress in the multimodal object detection in autonomous driving with visible (VI) and infrared (IR) image pairs. With given VI-IR image pairs, several multimodal detection frameworks are proposed to perceive pedestrians from aligned VI and IR images [1, 4]. For example, Chen et al. [1] proposed an MLF-CNN to conduct multi-feature fusion between visible (VI) and infrared (IR) backbones, which significantly improves the predictive performance on the aligned VI-IR image pairs. Despite the success in multimodal detection on aligned VI-IR pairs, the VI and IR cameras are difficult to calibrate in matching their images on pixel-level. Therefore, the aligned VI-IR pairs are too difficult to obtain in both industries and academic applications.

In this regard, Devaguptapu et al. [4] proposed a novel multimodal thermal object detection (MMTOD) to address the limitations by combining context enhancement (CE) and object detection. First,

[*]These authors contributed equally.
[†]Corresponding author.

Machine Learning for Autonomous Driving Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

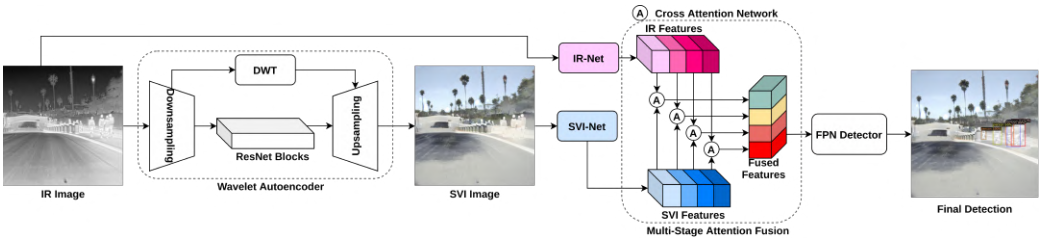


Figure 1: The illustration of the proposed context enhancement and information fusion framework (CEIFF) which includes wavelet autoencoder (WA) for generating synthetic VI images; the multi-stage attention fusion (MSAF) aims to fuse IR and SVI for final object detection.

the IR images are converted into colorized synthetic visible (SVI) images, which are trained on unaligned VI and IR images [15]. Then, the MMTOD employs a simple two-stream Faster-RCNN as detector with IR and SVI images. However, the design of MMTOD still has room to improve in both CE and detection. Therefore, this study proposes an improved context enhancement and information fusion framework, i.e., CEIFF. First, in the conversion between IR and SVI images, a new wavelet autoencoder (WA) is implemented to enrich the photo-realistic details of SVI during context enhancement. Finally, a multi-stage attention fusion (MSAF) network is proposed to achieve improved object detection performance. The experimental results show the effectiveness of the proposed CEIFF method on the FLIR ADAS dataset [5].

2 Methodology

Overview. Figure 1 presents the overall architecture of the proposed two-stage CEIFF. First, the original IR images are fed into the wavelet autoencoder (WA) to generate the SVI image [14]. Unlike contemporary image translation methods [10], the proposed WA employs discrete wavelet transform (DWT) as a shortcut to deliver the high-frequency components from downsampling blocks to upsampling blocks. The DWT aims to reserve the photo-realistic texture during CE. Then, the IR images and photo-realistic SVI images are fed into corresponding backbones, i.e., IR-Net and SVI-Net. Unlike the late fusion strategy in MMTOD [4], the proposed multi-stage attention fusion (MSAF) aims to fuse the features from IR and SVI on multiple levels of these backbones, which can maximize the fused feature qualities. On the other hand, the MSAF employs a modified non-local attention network (MNLN) [12] which can aggregate global context during fusion. Finally, the fused features are fed into the standard FPN detector for final detection [11].

Wavelet Autoencoder for IR2VI Translation. The proposed wavelet autoencoder (WA) is designed based on the UNIT [10] which is developed based on combination of variational autoencoder (VA) and generative adversarial networks (GANs) [14]. The original design of VA does not have shortcuts to link the downsampling and upsampling blocks, which causes the information degradation through propagation. Inspired by recent advances in neural style transfer [13], the DWT is used to bypass these blocks with high-resolution components while the low-resolution components are passed to ResNet blocks for process low-frequency components. More details can be found in [14] and Appendix.

Multi-stage Attention Fusion for Detection. The proposed MSAF aims to fuse IR and SVI features in each level of the corresponding backbones. In order to aggregate the global context during information fusion, the modified non-local network (MNLN) is designed for this purpose [12]. The MNLN is defined as:

$$Y = X_{\text{IR}} + f(X_{\text{IR}}, X_{\text{SVI}})g(X_{\text{SVI}}), \quad f(X_1, X_2) = \text{Softmax}(\theta(X_1)\phi(X_2)^T) \tag{1}$$

where $\theta(.)$, $\phi(.)$ and $g(.)$ are the $1 \times 1 \times 1$ convolution. The MNLN aims to regard the generated SVI images as reference attention to guide the IR feature generation by drawing cross correlation maps between IR and SVI through $f(.)$. Although the SVI image has colorized context, the SVI may still have noisy textures after CE. Therefore, the cross-direction design is modified to the proposed one-direction design [2], which only regards SVI as reference features.

3 Experimental Results

(a) Original IR (b) CycleGAN (c) MUNIT (d) StarGAN (e) UGATIT (f) WA

Figure 2: The examples of methods in context enhancement.

Experimental Setup. In this study, the FLIR ADAS [5] is implemented to validate the proposed framework for pedestrian detection. The dataset has been randomly split into a training set (8862 IR-VI pairs) and a testing set (1257 IR-VI pairs). Although the IR and VI images are synchronized on time, the IR and VI cameras are unaligned. The backbone is ResNet-18 [6] for all the the baseline detectors under Faster-RCNN framework [11]. The optimizer is stochastic gradient descent (SGD) with 2 as batch size. The learning rate of SGD is set to 0.001, which decays 10 times after 60000 iterations. The momentum is set to 0.9 for smoothing the training process.

Table 1: The comparative results of context enhancement methods.

IR	SVI	CE	Framework	AP50	AP75	mAP
✓	–	–	Faster-RCNN	49.97	18.67	23.38
–	✓	WA	Faster-RCNN	47.64	19.70	23.61
✓	✓	CycleGAN	MMTOD	52.34	20.62	25.23
✓	✓	MUNIT	MMTOD	52.59	21.03	25.44
✓	✓	**WA**	**MMTOD**	**55.18**	**21.20**	**26.10**

Results on Context Enhancement. In this experiments, recent advanced context enhancement methods, *i.e.*, CycleGAN [15], MUNIT [7], StarGAN[3], UGATIT[9] and WA, are implemented on the IR images. The examples of generative synthetic visible (SVI) images are illustrated in Fig. 2. The original IR image lacks detailed imagery textures which makes objects indistinguishable. However, StarGAN and UGATIT can not render the image texture from unaligned IR-VI image pairs. Specifically, the StarGAN only amplifies the edges of objects with false rendered colors while UGATIT generates fragmented images. Therefore, the StarGAN and UGATIT are not feasible in pedestrian detection. On the other hand, CycleGAN, MUNIT, and WA can have proper SVI outputs after enhancement. These methods enlighten the IR image with colors and detailed textures. Besides, the objects are much clear such as vehicles and buildings after CE. Compared with CycleGAN and MUNIT, the generated SVI images are more photorealistic in colors and texture from WA. Besides, the SVI images of WA have less noise and higher resolution, as shown in Fig. 2.

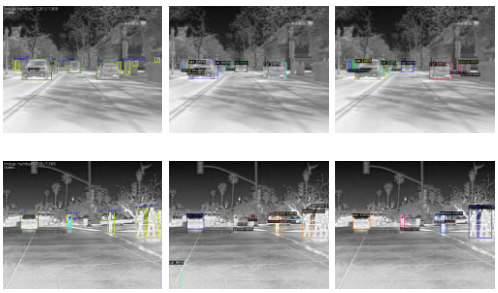
Table 2: The comparative results of context enhancement methods.

Detectors	AP50	AP75	mAP	Person	Bicycle	Car
MMTOD	55.18	21.20	26.10	26.53	8.237	43.54
DenseFuse	53.36	20.75	25.54	25.69	7.928	42.99
IR2VI	55.83	21.61	26.76	27.17	9.222	43.88
MSAF	**59.38**	**23.76**	**28.79**	**31.63**	**10.71**	**44.03**

Detection Results with Context Enhancement. This section aims to verify the feasibility of SVI images for improved detection as shown in Table 1. "✓" means the corresponding inputs are used during the experiment. The bold font indicates the best result in the column. Firstly, the classical Faster-RCNN [11] is implemented to validate the effectiveness of SVI images from WA. Compared with the Faster-RCNN trained by pure IR images, the detector of using SVI has fewer values in AP50. Nonetheless, the scores of AP75 and mAP are higher than the Faster-RCNN trained by pure IR images. The results indicate that the SVI can help the detector to have more precise proposals to localize the persons and vehicles on the road. In contrast, the lower values of AP50 reveal that the pure

SVI-based Faster-RCNN may miss some objects. To summarize, there is room to improve in object detection beyond single-modal detectors. Then, a recent multi-modal detector, *i.e.*, MMTOD [4], is implemented to validate the effectiveness of information fusion with SVI images from WA. As shown in Table 1, the IR-SVI fusion can improve the general detection quality. Especially for AP50, the fusion scheme brings around 15% improvement. On the other hand, SVI images from CycleGAN and MUNIT are also used to validate the effectiveness of the fusion mechanism. The results also suggest that the feasibility of IR-SVI fusion via context enhancement. Compared with CycleGAN and MUNIT, implemented WA achieves better detection accuracy by generating SVI images of higher quality.

Comparison with Recent Multimodal Detectors. The previous experiments demonstrate the effectiveness of applying IR-SVI fusion via recent MMTOD [4] and proposed WA. This section aims to compare the AP for each class and the mAP of our proposed MSAF against the recent multimodal detectors. As shown in Table 2, we observe that the proposed MSAF outperforms the baseline detectors across all the classes. Compared with the second-best detector IR2VI, the MSAF achieves 28.79% on mAP and brings around 16% improvement on Person and Bicycle detection and around 10% improvement on AP50 and AP75. On the other hand, the Fig. 3 shows the qualitative examples from groundtruth (GT), MMTOD and MSAF, which also indicates that the MSAF has more accurate detection. In conclusion, both qualitative and quantitative results demonstrate the effectiveness of the proposed MSAF for detection.



(a) GT (b) MMTOD (c) MSAF

Figure 3: The examples of multimodal detection. From the left to right in columns, the images are of (a) groundtruth (GT), (b) MMTOD and the proposed (c) MSAF.

Discussion on Selection of Fusion Operators. The previous experiments demonstrate the effectiveness of the proposed MSAF for detection by applying IR-SVI fusion. This section aims to compare the performance of different fusion operators in the proposed MSAF. Two fusion operators have been implemented, i.e., Embedding and MNLN. The embedding is the a 3×3 convolution neural network after concatenation of IR and SVI features. As shown in Table 3, it reflects that MNLN achieves the best performance on AP75, mAP, Person, and Bicycle detection. The main reason is the attention-based fusion operator which suppresses the noises brought by aggregating global information during feature fusion. Therefore, it is recommended to apply attention-based fusion operators to achieve better performance.

Table 3: The comparative results of fusion operators.

Methods	AP50	AP75	mAP	Person	Bicycle	Car
Embedding	41.18	19.88	21.51	24.40	2.008	38.14
MNLN	**59.38**	**23.76**	**28.79**	**31.63**	**10.71**	**44.03**

4 Conclusion

In this study, a context enhancement and information fusion framework (CEIFF) is proposed to improve the quality of multimodal environmental perception. The CEIFF employs wavelet autoencoder (WA) to generate synthetic VI (SVI) images from original IR images compared with the novel solution. The SVI images have more details in textures and context, making the objects discriminate from the background. Then, a multi-stage attention fusion (MSAF) is designed to fuse IR and SVI features with consideration of global information. Finally, the features are fed into the detector for final detection. The proposed solution demonstrates its effectiveness on a public dataset in autonomous driving.

References

[1] Y. Chen, H. Xie, and H. Shin. Multi-layer fusion techniques using a cnn for multispectral pedestrian detection. *IET Computer Vision*, 12(8):1179–1187, 2018.

[2] L. Chi, G. Tian, Y. Mu, and Q. Tian. Two-stream video classification with cross-modality attention. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4511–4520, 2019.

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] C. Devaguptapu, N. Akolekar, M. M Sharma, and V. N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[5] FLIR Systems Inc. Flir thermal dataset for algorithm training. `https://www.flir.ca/oem/adas/adas-dataset-form/`, 2020. (Accessed: April 9, 2021).

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[8] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] J. Kim, M. Kim, H. Kang, and K. H. Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020.

[10] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 700–708, 2017.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (6):1137–1149, 2017.

[12] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[13] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha. Photorealistic style transfer via wavelet transforms. In *International Conference on Computer Vision (ICCV)*, 2019.

[14] R. Zhang, J. Bin, Z. Liu, and E. Blasch. Chapter 13 - wggan: A wavelet-guided generative adversarial network for thermal image translation. In A. Solanki, A. Nayyar, and M. Naved, editors, *Generative Adversarial Networks for Image-to-Image Translation*, pages 313–327. Academic Press, 2021. ISBN 978-0-12-823519-5. doi: https://doi.org/10.1016/B978-0-12-823519-5.00015-4.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

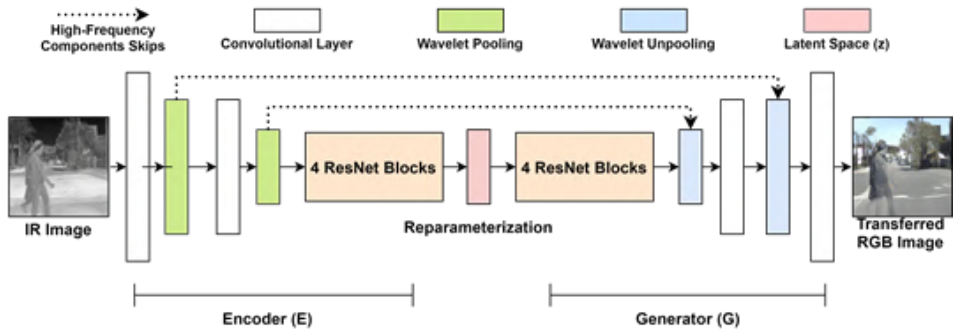


Figure 4: The structure of wavelet autoencoder (WA) [14]

A Wavelet Autoencoder

The Fig. 4 shows the architecture of the implemented wavelet autoencoder (WA), which is developed based on variational autoencoder and discrete wavelet transformation (DWT) as shown in Fig. 5. As shown in Fig. 4, the WGVA consists of two sub-networks: an encoder for converting IR image to latent space; and a decoder to recover the VI image from latent space. The design of WA follows standard residual autoencoder architecture with two convolutional layers, two pooling layers, and four residual blocks at the encoder. Wavelet pooling and wavelet unpooling are designed to substitute conventional pooling layers in standard residual autoencoder [13]. Then, the high-frequency components skips bridge the corresponding pooling and unpooling layers for improving generative resolution [13] as shown in Fig.6. The details of training the network can be found in [10, 14]. More generative examples are illustrated in Fig. 7.

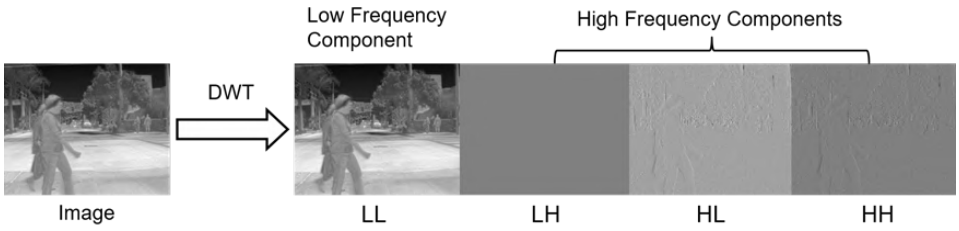


Figure 5: The example of DWT [14]

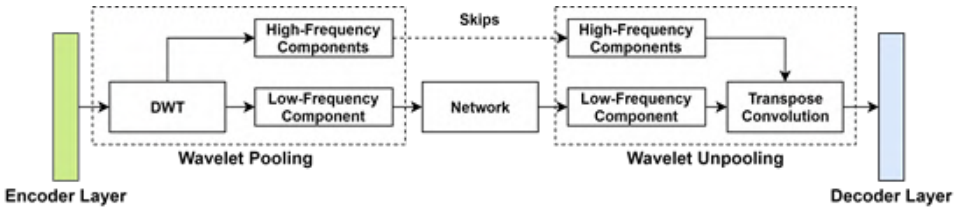


Figure 6: The example of wavelet pooling [14]

Figure 7: The generative examples from WA [14]. (a) source IR images; (b) proposed WGGAN; (c) CycleGAN; (d) MUNIT; (e) StarGAN and (f) UGATIT.