IR Reasoner: Real-time Infrared Object Detection by Visual Reasoning

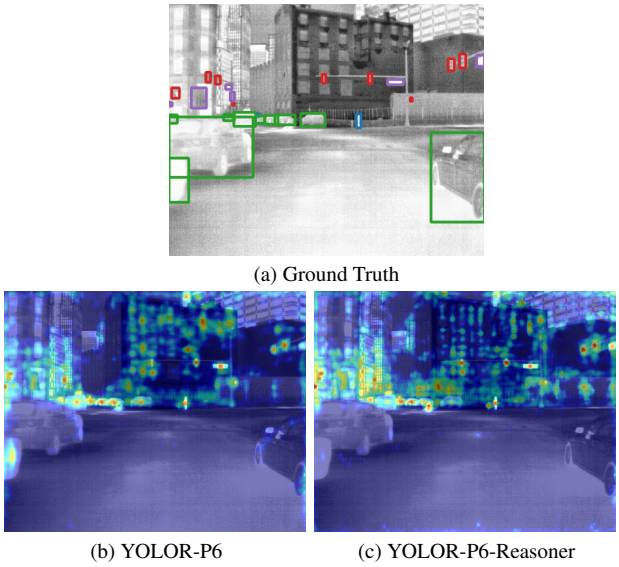
Meryem Mine Gündoğan[1,3*] Tolga Aksoy[2*] Alptekin Temizel[3] Ugur Halici[4]
[1]Aselsan Inc., Turkey
[2]Arizona State University, USA
[3]Graduate School of Informatics, Middle East Technical University, Turkey
[4]Department of Electrical and Electronics Engineering, Middle East Technical University, Turkey
megundogan@aselsan.com.tr, tolga.aksoy@asu.edu, {atemizel, halici}@metu.edu.tr

Abstract

Thermal Infrared (IR) imagery is utilized in several applications due to their unique properties. However, there are a number of challenges, such as small target objects, image noise, lack of textural information, and background clutter, negatively affecting detection of objects in IR images. Current real-time object detection methods treat each image region separately and, in face of these challenges, this sole dependency on feature maps extracted by convolutional layers is not ideal. In this paper, we introduce a new architecture for real-time object detection in IR images by reasoning the relations between image regions by using self-attention. The proposed method, IR Reasoner, takes the spatial and semantic coherency between image regions into account to enhance the feature maps. We integrated this approach into the current state-of-the-art one-stage object detectors YOLOv4, YOLOR, and YOLOv7, and trained them from scratch on the FLIR ADAS dataset. Experimental evaluations show that the Reasoner variants perform better than the baseline models while still running in real-time. Our best performing Reasoner model YOLOv7-W6-Reasoner achieves 40.5% AP at 32.7 FPS. The code is publicly available.[1]

1. Introduction

Imaging in the visible domain requires active illumination since cameras working in this band capture the reflected light. On the other hand, thermal infrared (IR) cameras capture the radiation emitted by objects, i.e., they do not require scene illumination, and hence they are not sensitive to illumination conditions [42]. This makes them suitable for applications that require detection of objects in complex environmental conditions [41]. Autonomous driving cars, smart farming, and surveillance systems are the major application fields where previously mentioned advantages are exploited by utilizing IR imagery standalone or fusing with visible imagery to detect objects of interest [14].



(a) Ground Truth

(b) YOLOR-P6 (c) YOLOR-P6-Reasoner

Figure 1. Ground truth (top) and visualization of class activation maps [35] before detection head for YOLOR-P6 (bottom left) and YOLOR-P6-Reasoner (bottom right)

Early efforts in IR object detection mostly relied on hand-crafted features [19, 24, 51]. However, the representation capacity of these models is limited and they struggle to perform well under arbitrary conditions. Like other visual tasks, deep learning has also revolutionized IR object detection. The existence of large-scale datasets such as the ImageNet [10] and Microsoft Common Objects in Context (MS-COCO) [31] facilitated research in visible imagery. This research has also been adapted into IR domain and several Convolutional Neural Networks (CNN) based methods, using similar architectures with state-of-the-art approaches

*indicates equal contribution
[1]https://github.com/tlgksy/ir-reasoner

in visible imagery, have been proposed [27]. While some of the IR object detectors combine the visible and IR imagery through a fusion architecture [29,48], the others work solely on IR images [14, 39]. In this work, we use only IR imagery since most of the applications contain only thermal cameras.

Previous works [14, 27, 29, 39, 48] focused on achieving good performance in public IR datasets such as FLIR ADAS [13] and KAIST Multispectral Pedestrian [20]. However, these approaches solely depend on convolutional features and lack the ability to consider possible semantic and spatial relations between different image regions. Extraction of high-quality convolutional features could not be possible at all times, especially in challenging conditions such as small-sized target object, image noise, and background clutter [42] and relying only on convolutional architectures results in unreasonable predictions. These observations lead the way to transformer-based detectors [2, 36] that have an ability to extract relations between image regions. However, as they are computationally demanding, they are infeasible in applications that require real-time processing. Autonomous driving is as an example application area where obtaining detection at high frame rates is vital.

To address the mentioned challenges, while still allowing real-time operation, we propose a visual reasoning based one-stage IR object detection architecture. The proposed architecture takes into consideration semantic and spatial coherency between image regions. Figure 1 shows the class activation maps before detection head for YOLOR-P6 and its proposed Reasoner variant. The baseline model have strong activations because of the visual similarity to traffic lights captured by the convolutional layers around the apartment windows. On the other hand, these activations are attenuated by the reasoner model and the ones on the left-hand side, where there are many objects of interest such as as traffic signs and lights, are amplified. We integrated our Reasoner approach to the recent state-of-the-art one-stage object detector models: YOLOv4-P6 [3], YOLOR-P6 [47] and YOLOv7-W6 [46]. All the models then have been trained from scratch on FLIR ADAS [13] dataset. Pre-trained weights were not used to eliminate any bias. We provide a comparative analysis of the proposed Reasoner variants with the baselines quantitatively and qualitatively.

2. Related Work

Object Detection: The use of Convolutional Neural Networks and the ImageNet challenge were major milestones in object detection [26]. Object detection tasks can be categorized into generation of region proposals and object classification for each region proposal. Detectors which handles these tasks in two separate frameworks are called two-stage object detectors. These architectures perform better in accuracy, but they are more computationally complex, prohibiting their use in real-time applications and embbedded devices. R-CNN [38] and its variants are pioneers of two-stage object detectors. One-stage object detectors perform both tasks in a single feed-forward fully convolutional network. These frameworks can achieve real-time inference speed, albeit with lower accuracy. Single Shot MultiBox Detector (SSD) [33] and You Only Look Once (YOLO) [37] are prominent examples [5]. EfficientDet [40] and YOLO variants [3, 45–47] are popular successors of one-stage networks. The YOLO family has been constantly being updated taking the state-of-the-art methods into account and they are optimized for an effective implementation. Hence, in this work, YOLO variants are selected as the baseline models.

Reasoning: There are several studies in the literature aiming to mimic human visual reasoning ability [4, 7, 8, 17, 34, 49, 52]. In [7], spatial and semantic relationships between objects are modelled using a spatial reasoning approach, where extracted object instances are fed into another CNN for context reasoning. Gated Recurrent Unit (GRU) is used to update spatial memory cells. In [17] a relation module, inspired by attention module [43], is used to improve instance recognition performance and to prevent duplicate bounding boxes. Spatial memory [7] is improved by attaching a global graph-reasoning module in [8]. DETR [4] approach is notable for being the first to successfully use transformers for object detection. A transformer encoder and decoder is added on top of a standard CNN model with a usage of bipartite matching loss. Several studies [30, 32, 50, 53] improves performance and efficiency of DETR which increases popularity of using attention-based deep learning models in literature. SWIN Transform is a general-purpose vision Transformer backbone that computes attention within a local window [34]. To capture overlapping regions between windows (image locations), window partitioning shifts gradually along the hierarchy of the network. This method achieves linear complexity instead of the original quadratical complexity with increasing image size. In this work, we used transformer encoder-like module for acquiring reasoning ability from learned feature maps.

IR Object Detection: In the earlier works, hand-crafted features were used [19,24,51] for IR object detection, while most of the recent work is based on deep learning. In multiple papers, pre-trained visible imagery networks are fused with IR imagery person detection datasets for successful transfer learning [16, 23, 25, 44]. Also, some studies shows that visible domain object detectors can be adapted to be used in IR domain [6, 11, 28]. In [14], authors augmented thermal images by extracting saliency maps from IR imagery which improved the IR detector performance. In [22], YOLO-based models' performances are compared for the FLIR ADAS dataset and UAV TIR video detection.

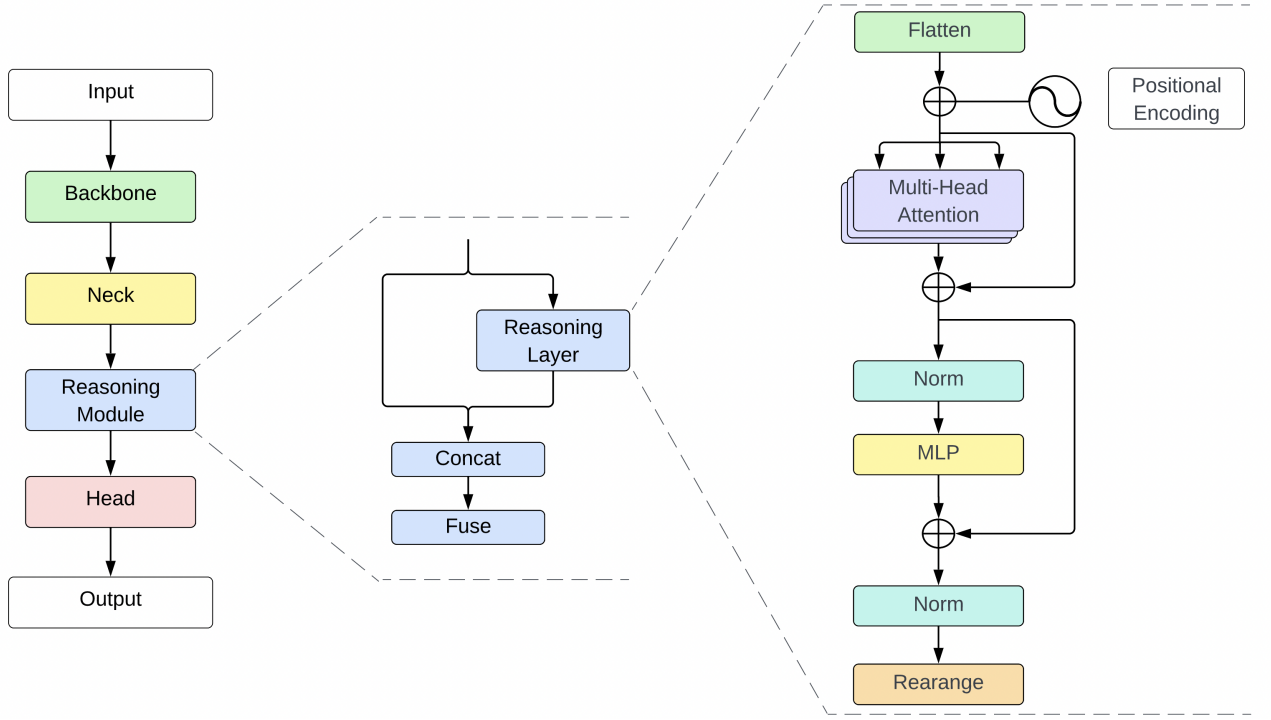


Figure 2. Overall architecture of the proposed reasoner model

3. Proposed Method

A general diagram of the proposed method is shown in Figure 2. First, convolutional features are extracted by the backbone. Then, multi-scale feature maps from different layers of the backbone are collected and concatenated at the neck. The reasoning layer extracts the semantic and spatial relationships between image regions subsequently. The reasoning features are concatenated with the only-convolutional ones and fused. Finally, class probabilities and bounding boxes are predicted by the head using improved feature maps.

3.1. Baseline Models

YOLO [37] is a single-stage lightweight, real-time, CNN based object detection model. It divides the image into $S \times S$ grids, where each grid is responsible for detecting objects in the corresponding grid. After the initial detections, non-maximal suppression is used to eliminate multiple bounding boxes for the same object [37]. YOLO achieves high accuracy and speed by detecting objects at multiple scales and using anchor boxes to simplify the bounding box prediction task.

YOLOv4 [3] is a novel architecture that can maintain FPS rate with high accuracy. The succes of YOLOv4 relies on the CSPDarknet53 backbone, Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PAN) techniques. The novel backbone network is called CSPDarknet53 which combines a modified Darknet architecture with cross-stage partial connections. The SPP method pools information across multiple kernel sizes simultaneously, in order to acquire both fine and coarse information. Additionally, in order to increase detection efficiency, the PAN leverages information from layers near the input by transmitting features from various backbone levels. Also, several new Bag of Freebies (BoF) strategies are used such as CutMix and Mosaic data augmentation, DropBlock regularization, and class label smoothing. In this work, YOLOv4-P6 architecture has been used for implementation.

YOLOR [47] is a unified network that combines implicit and explicit knowledge conjointly. Explicit knowledge corresponds to the shallow layers of the network, whereas implicit knowledge corresponds to the deeper layers of the network. The authors states that implicit knowledge assists explicit knowledge to perform tasks effectively. Furthermore, the model is capable of learning general representations, namely multi-task learning model. The goal is to represent various task in a unified architecture. YOLOR performs comparable performance over other state-of-art models within real-time inference. In this work, we selected YOLOR-P6 model for reasoning layer fusion which is derived from YOLOv4-P6-light [3].

YOLOv7 is a recent update to YOLO family and it is reported to have an accuracy of 56.8% AP in MS COCO dataset while running at 30 FPS on NVIDIA V100 GPU

[46]. It has four principal improvements over the baseline YOLOv4 [3, 45] and YOLOR [47] as follows:

- Extended Efficient Layer Aggregation Network (ELAN) layer structure produces easier-to-optimize gradients where it facilitates learning various features.
- Network depth and width are scalable with a concatenation-based model that maintains the properties of the initial design and optimal parameters.
- Re-parameterization planning generates a more robust model by averaging the model weights.
- Auxilary heads are combined in the middle of the network to improve predictions.

In this work, we selected YOLOv7-W6 model as a baseline for reasoning integration.

3.2. Reasoning Module

3.2.1 Reasoning Layer

In this work, a transformer encoder-like module is used as a reasoning layer (Figure 2). Sub-layers of the reasoning layer are as follows:

Flatten: This layer converts the input grid $(H \times W \times C)$ into a sequence $(HW \times C)$ as expected by the multi-head attention layer.

Positional Encoding: Fixed sinusoidal positional encodings are calculated as shown in Eq. 1 and added to the input feature embeddings to encode order information of the image regions.

$$\begin{aligned} PE_{(i,2j)} &= \sin(\frac{i}{10000^{2j/d_{feature}}}) \\ PE_{(i,2j+1)} &= \cos(\frac{i}{10000^{2j/d_{feature}}}) \end{aligned} \tag{1}$$

where i is the position of the grid in the sequence, j is the feature depth index, and $d_{feature}$ is the same with the feature depth.

Multi-Head Attention: This layer models the semantic relationships between image regions. Multi-head attention employs parallel self-attention, which is based on query, key, and value.

Self-attention allows the query of a single grid cell to search the potential correlation with others in the sequence via the keys. The comparison of the query and key pairs yields the attention weight for the value, while the interaction of the attention weight and the value determines how much focus should be placed on other parts of the sequence, i.e., the image while representing the current cell.

Query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) matrices are calculated by weighing the input sequence with corresponding weight matrices ($\mathbf{W^Q}$, $\mathbf{W^K}$, $\mathbf{W^V}$). Then, attention is calculated using these values as shown in Eq. 2, where $\mathbf{X}$ is the input.

$$\begin{aligned} \mathbf{Q} &= \mathbf{X}\mathbf{W^Q} \\ \mathbf{K} &= \mathbf{X}\mathbf{W^K} \\ \mathbf{V} &= \mathbf{X}\mathbf{W^V} \\ Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) &= softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \end{aligned} \tag{2}$$

where d_k is the dimension of the query and key matrices. The scaled dot product is used to compare query and key matrices [43]. The attention weights indicate where to look in the value matrix ,i.e., they learn to distinguish valuable, informative and relevant parts of the image while encoding the current grid. Also, the grid cell, which represents a region of an image, is encoded by taking a summation of value matrix columns weighted by normalized attention weights.

Instead of using a single self-attention function, multi-head attention is implemented [43]. Parallel projections of query, key and value matrices are used to compute multiple self-attention functions. Using multi-head attention, the model can jointly attend to catch different types of information from various representation subspaces. The attention of $head_i$ is calculated by Eq. 3.

$$head_i = Attention(\mathbf{Q}\mathbf{W_i^Q}, \mathbf{K}\mathbf{W_i^K}, \mathbf{V}\mathbf{W_i^V}) \tag{3}$$

Then, each head's attentions are concatenated and projected once again with a weight matrix $\mathbf{W^O}$ to calculate multi-head attention, as shown in Eq. 4.

$$MultiHead(\mathbf{Q},\mathbf{K},\mathbf{V}) = Concat(head_1, ...)\mathbf{W^O} \tag{4}$$

Skip Connections: There are two skip connections in the reasoning layer. Residual skip connections improve backpropagation [15], and the original information is propagated to the subsequent layers.

Normalization: There are two normalization layers in the reasoning layer. Besides skip connection, Layer normalization [1] is utilized in the reasoning layer. This is another key factor helping with easier convergence by alleviating the internal covariate shift [21].

MLP: Output of the multi-head attention layer is fed into the MLP layer, which creates a new representation of the reasoning information. MLP layer consists of two linear layers and a ReLU non-linearity in between as in Eq. 5

$$MLP(x) = max(0, x\mathbf{W_1} + b)\mathbf{W_2} + b_2 \tag{5}$$

where $\mathbf{W1}$ and $\mathbf{W2}$ are weight matrices and b_1 and b_2 are biases.

Rearrange: This layer converts the flatted sequence data $(HW \times C)$ into three-dimensional grid format $(H \times W \times C$ as object detection head expects, where H is height, W is weight and C is channel.

3.2.2 Concat Layer

The output of the reasoning layer is concatenated with the original neck output only-convolutional features in this layer wich ensures the reusability of only-convolutional features [18]. With the help of this compound usage strategy, the network can exploit complementary information available in features extracted by only-convolutional layers and the the reasoning layer.

3.2.3 Fuse Layer

This layer fuses the concatenated features by a 1×1 convolutional layer. Following this layer, feature maps improved by visual reasoning are fed into the detection head to generate bounding boxes and class probabilities for each grid region.

4. Experiments

4.1. Dataset

This research utilizes RAW IR images, on the other hand, only 1.4% frames in amongst the whole IR datasets are RAW frames [9]. Publicly available Teledyne FLIR ADAS v2 Dataset provides both 8-bit enhanced and 14-bit RAW IR images for the development of thermal object detection systems. This dataset contains 26,442 fully annotated 640×512 frames with 520,000 bounding box annotations across 15 different object categories [13]. Additionally, this dataset comes with pre-classified train, validation and test splits. We used the 14-bit RAW thermal images during training and testing. We used only the *car*, *person*, sign, light, and bike classes as the instance numbers of the other classes are negligible such as the *dog* class has only four instances and the *deer* class has only eight instances in the train set.

4.2. Implementation Details

The experiments have been conducted on a system having 4 NVIDIA Tesla V100 GPUs with a batch size of 8 images per GPU. During the training, the default hyperparameter set and default augmentations (except color augmentation) have been used. 14-bit RAW thermal images have been normalized into the range [0-1] at the preprocessing stage by dividing all the pixel values with $2^{14} - 1$. All models are trained from scratch without using any pretrained weights until they converge.

4.3. Evaluation Metrics

The most frequently used evaluation metric in object detection is "Average Precision" [12] which is a measure of average detection performance under various recall values. However, the mean AP method is used as a final evaluation metric for predicted box localization and performance over whole object categories. The localization performance of the predicted box is evaluated using the Intersection over Union (IoU). If the IoU between the ground truth box and the predicted box is greater than a predefined threshold, the prediction is marked as a successful detection. Otherwise, it is missed detection [54]. Mean Average Precision metrics across three different threshold values AP_{50}, AP_{75} and AP and three different scales AP_S, AP_M and AP_L for small ($< 32^2$ pixels), medium (32^2 to 96^2 pixels) and large objects ($> 96^2$ pixels) have been used for evaluation. AP_{50} and AP_{75} indicates that the predicted bounding box is considered a correct detection if it has an IoU greater than or equal to 0.50 and 0.75 respectively, with the ground truth bounding box. AP indicates that precision is averaged between 0.50 (coarse localization) and 0.95 (perfect localization) IoU thresholds.

4.4. Results

The comparison of the performances of baseline methods and Reasoner variants are presented in Table 1 in terms of average precision (AP) and frames-per-seconds on a single Tesla V100 GPU in a single batch. According to the results, Reasoner variants incorporating semantic relationship information increase the performance in terms of AP, AP_{50}, and AP_{75} for all the models in question. In particular, AP_{50} increases by 4.5, 4.4 and 1.0 percentage points in YOLOv4-P6, YOLOR-P6 and YOLOv7-W6 models respectively with the proposed Reasoner variants. AP analysis in terms of different object sizes are also provided in this table. While the Reasoner variants exhibit better performance for all object size types impact of the Reasoner variants is the highest for small objects as indicated by the AP_S metric. As the proposed model increases the number of parameters, its computational complexity is higher, reflected by their lower FPS values in comparison to their counterparts. On the other hand, the impact is not large to disallow real-time operation for most applications. Among different models, YOLOv7-W6 is the best performing one amongst the baseline models. The Reasoner variant of this model improves the accuracy, and it is the best-performing model overall, with 40.5% AP and running at 32.7 FPS.

Figures 3, 4 5 show sample detection results on 14-bit RAW test images of FLIR ADAS dataset for YOLOv4-P6, YOLOR-P6 and YOLOV7-W6 and their Reasoner variants. All sample RAW images enhanced with Contrast Limited Adaptive Histogram Equalization (CLAHE) [55] for visualization. While baseline YOLOv4 could not detect the rightmost two signs in purple bounding boxes in Figure 3, reasoner alternative successfully detects the both. As the IR object signature of such signs are not obvious the reasoner module helps detection by the incorporation of semantic and spatial properties. In Figure 4, baseline YOLOR-P6 mispredicts a traffic light on the building exterior, while the

Table 1. Comparison of Reasoner models with their baseline detectors.

Detector	Size	FPS	# params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv4-P6	1280	25.0	126M	26.3%	40.4%	28.2%	21.5%	41.9%	91.8%
YOLOv4-P6-Reasoner	1280	20.6	146M	**29.1%**	**44.9%**	**31.0%**	**24.5%**	**43.8%**	**92.9%**
YOLOR-P6	1280	33.7	37M	36.4%	53.1%	39.9%	32.2%	49.1%	92.9%
YOLOR-P6-Reasoner	1280	28.1	43M	**39.4%**	**57.5%**	**44.2%**	**35.2%**	**50.8%**	**93.6%**
YOLOv7-W6	1280	44.1	80M	39.5%	57.5%	44.0%	35.9%	**49.1%**	94.7%
YOLOv7-W6-Reasoner	1280	32.7	96M	**40.5%**	**58.5%**	**44.8%**	**37.2%**	47.8%	**94.9%**

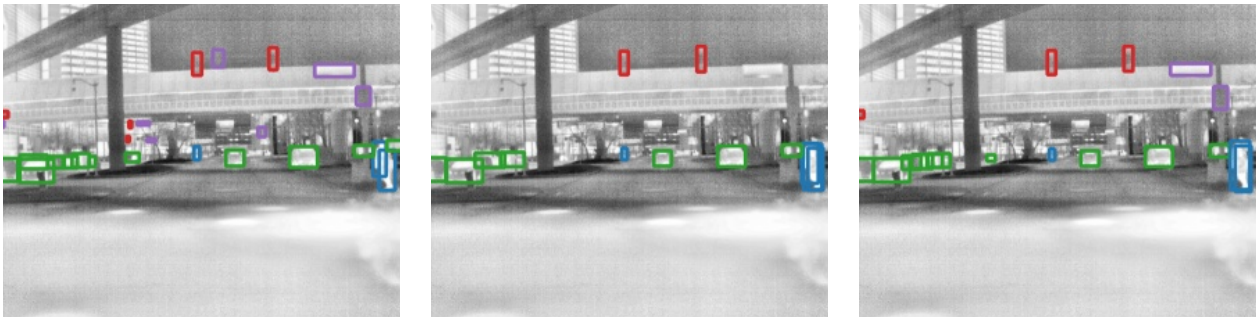


Figure 3. Ground truth (left), YOLOv4-P6 (middle) and YOLOv4-P6-Reasoner results (right) for a sample image.

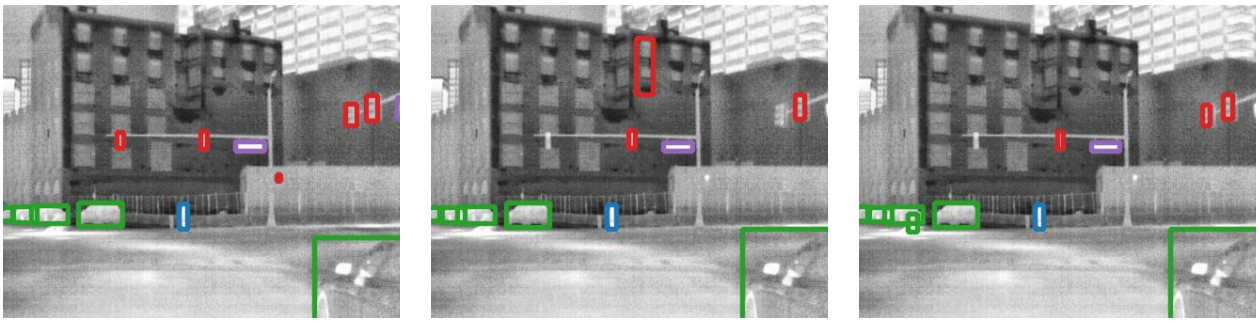


Figure 4. Ground truth (left), YOLOR-P6 (middle) and YOLOR-P6-Reasoner results (right) for a sample image.

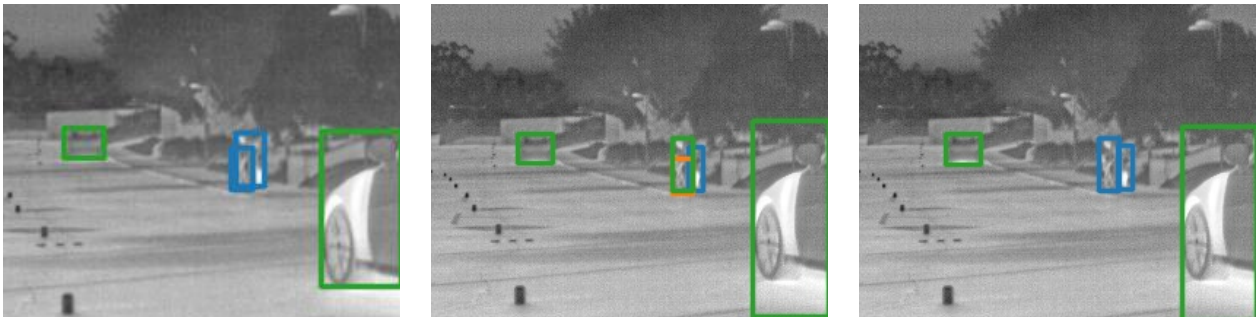


Figure 5. Ground truth (left), YOLOv7-W6 (middle) and YOLOv7-W6-Reasoner results (right) for a sample image.

Reasoner variant is not mistaken for this particular case, as YOLOR-P6-Reasoner may not be expecting a traffic sign whose size is bigger than other signs and traffic lights. In Figure 5, there is a partially occluded group of people in blue bounding boxes on the right-hand side. YOLOv7-W6 predicts them as a bike (orange), a car (green) and a person (blue), while YOLOv7-W6 successfully predicts two persons. On the other hand, it fails to detect the child as IR

signatures of the persons get mixed up.

5. Conclusion

While deep learning approaches have shown great success in object detection, they have limitations when it comes to considering semantic and spatial relations between image regions. This is particularly true in cases where images contain small objects, image noise, and background clutter, as is often the case with IR imagery. To overcome these limitations, transformer-based detectors have been introduced, which have an ability to extract relations between image regions. In this study, we proposed a novel reasoning module that incorporates a transformer encoder-like module to capture reasoning ability from feature maps extracted from convolutional layers. Experimental evaluation of the proposed method on the thermal 14-bit RAW FLIR ADAS dataset indicates that the proposed reasoner module increases the performance in terms of AP metrics over baselines while still running in real-time and the increase is particularly pronounced for small objects.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *NIPS 2016 Deep Learning Symposium*, 2016. 4

[2] Neelanjan Bhowmik, Jack W Barker, Yona Falinie A Gaus, and Toby P Breckon. Lost in compression: the impact of lossy image compression on variable size object detection within infrared imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2022. 2

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2, 3, 4

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[5] Manuel Carranza-García, Jesús Torres-Mateo, Pedro Lara-Benítez, and Jorge García-Gutiérrez. On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data. *Remote Sensing*, 13(1):89, Dec 2020. 2

[6] Bingwen Chen, Wenwei Wang, and Qianqing Qin. Robust multi-stage approach for the detection of moving target from infrared imagery. *Optical Engineering*, 51(6):067006, 2012. 2

[7] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 4086–4096, 2017. 2

[8] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7239–7248, 2018. 2

[9] Kevser Irem Danaci and Erdem Akagunduz. A survey on infrared image and video sets. *arXiv preprint arXiv:2203.08581*, 2022. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[11] Tarek Elguebaly and Nizar Bouguila. A nonparametric bayesian approach for enhanced pedestrian detection and foreground segmentation. In *CVPR 2011 WORKSHOPS*, pages 21–26, 2011. 2

[12] Mark Everingham, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2009. 5

[13] FLIR. Free flir thermal dataset for algorithm training, 2022. *Available at* https://www.flir.com/oem/adas/adas-dataset-form/. 2, 5

[14] Debasmita Ghose, Shasvat Mukeshkumar Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 988–997, 2019. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for ımage recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] Christian Herrmann, Miriam Ruf, and Jürgen Beyerer. Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, pages 38–43. SPIE, 2018. 2

[17] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018. 2

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[19] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 1, 2

[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*, pages 1037–1045, 2015. 2

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal co-

variate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 4

[22] Chenchen Jiang, Huazhong Ren, Xin Ye, Jinshun Zhu, Hui Zeng, Yang Nan, Min Sun, Xiang Ren, and Hongtao Huo. Object detection from uav thermal infrared images and videos using yolo models. *International Journal of Applied Earth Observation and Geoinformation*, 112:102912, 2022. 2

[23] Shu Wang Jingjing Liu, Shaoting Zhang and Dimitris Metaxas. Multispectral deep neural networks for pedestrian detection. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 73.1–73.13. BMVA Press, September 2016. 2

[24] Mohammad Nazmul Alam Khan, Guoliang Fan, Douglas R. Heisterkamp, and Liangjiang Yu. Automatic target recognition in infrared imagery using dense hog features and relevance grouping of vocabulary. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 293–298, 2014. 1, 2

[25] Daniel Konig, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. 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, 2017. 2

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[27] Eun Ju Lee, Byoung Chul Ko, and Jae-Yeal Nam. Recognizing pedestrian's unsafe behaviors in far-infrared imagery at night. *Infrared Physics & Technology*, 76:261–270, 2016. 2

[28] Alex Leykin, Yang Ran, and Riad Hammoud. Thermal-visible video fusion for moving target tracking and pedestrian classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[29] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 2

[30] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[32] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 2

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[35] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 1

[36] Farzeen Munir, Shoaib Azam, and Moongu Jeon. Sstn: Self-supervised domain adaptation thermal object detection for autonomous driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 206–213. IEEE, 2021. 2

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2, 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian 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. 2

[39] Iain Rodger, Barry Connor, and Neil M Robertson. Classifying objects in lwir imagery via cnns. In *Electro-Optical and Infrared Systems: Technology and Applications XIII*, volume 9987, pages 152–165. SPIE, 2016. 2

[40] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 2

[41] Michael Teutsch, Thomas Muller, Marco Huber, and Jurgen Beyerer. Low resolution person detection with a moving thermal infrared camera by hot spot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 209–216, 2014. 1

[42] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. Computer vision in the infrared spectrum: challenges and approaches. *Synthesis Lectures on Computer Vision*, 10(2):1–138, 2021. 1, 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[44] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016. 2

[45] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038, 2021. 2, 4

[46] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 2, 4

[47] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021. 2, 3, 4

[48] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5371, 2017. 2

[49] Hang Xu, ChenHan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6419–6428, 2019. 2

[50] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[51] Li Zhang, Bo Wu, and Ram Nevatia. Pedestrian detection in infrared images based on local shape features. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 1, 2

[52] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo:transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2799–2808, 2021. 2

[53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[54] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. 5 2019. 5

[55] Karel J. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems*, 1994. 5