Cross-Modality Attention and Multimodal Fusion Transformer for Pedestrian Detection

Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips

TELIN-IPI, Ghent University-imec, Gent, Belgium
{Weiyu.Lee, Ljubomir.Jovanov, Wilfried.Philips}@ugent.be

Abstract. Pedestrian detection is an important challenge in computer vision due to its various applications. To achieve more accurate results, thermal images have been widely exploited as complementary information to assist conventional RGB-based detection. Although existing methods have developed numerous fusion strategies to utilize the complementary features, research that focuses on exploring features exclusive to each modality is limited. On this account, the features specific to one modality cannot be fully utilized and the fusion results could be easily dominated by the other modality, which limits the upper bound of discrimination ability. Hence, we propose the Cross-modality Attention Transformer (CAT) to explore the potential of modality-specific features. Further, we introduce the Multimodal Fusion Transformer (MFT) to identify the correlations between the modality data and perform feature fusion. In addition, a content-aware objective function is proposed to learn better feature representations. The experiments show that our method can achieve state-of-the-art detection performance on public datasets. The ablation studies also show the effectiveness of the proposed components.

Keywords: Cross-Modality fusion, Multimodal pedestrian detection, Transformer

1 Introduction

Pedestrian detection is one of the most important challenges in computer vision due to its various applications, including autonomous driving, robotics, drones, and video surveillance. For achieving more accurate and robust results, thermal images have been widely exploited as the complementary information to solve the challenging problems that impede conventional RGB-based detection, such as background clutter, occlusions, or adverse lighting conditions. Radiated heat of pedestrians contains sufficient features to differentiate shape of humans from the background but lose visual appearance details in thermal images. On the other hand, RGB cameras can capture fine visual characteristics of pedestrians (e.g., texture). Hence, designing a fusion scheme to effectively utilize the different visual cues from thermal and RGB modalities has become a popular research interest.

In the existing methods, numerous fusion strategies have been exploited to utilize complementary features from color and thermal images [9, 28, 32]. In addition to simple feature concatenation [9], semantic information [13] and attention mechanism [29] are also introduced to improve the detection performance. Furthermore, in order to better exploit the characteristics of different modalities, the illumination condition is also considered during feature fusion [6, 14].

However, most previous studies only focus on performing feature fusion and exploiting the fused features instead of exploring features specific to each modality (i.e. modality-specific features) before fusion [31, 9, 11, 13, 20]. Specifically, most methods from the literature put emphasis on performing detection after the features from the color and thermal images are fused, while the features exclusive for each modality are not entirely utilized in the fusion process.

As a consequence, some of the features specific to one modality cannot be fully utilized. For instance, the texture or color of RGB images in bad illumination conditions might not be properly explored due to the domination of strong features from the other modality in the fusion process. However, in conventional RGB-based detection, the texture of the objects provides important cues that make the targets distinct from the background clutter. Without fully considering specific features, the fused information becomes the main discriminative cues, which limits the upper bound of discrimination ability.

In this paper, we propose a novel network architecture for multimodal pedestrian detection based on exploring the potential of modality-specific features to boost the detection performance.

The first key idea of this paper is better exploitation of modality-specific features by cross-referencing the complementary modality data in order to obtain more discriminative details. Instead of extracting features from modalities by independent feature extractors, we suggest that the aligned thermal-visible image pairs could act as a "consultant" for each other to discover potential specific features. The argumentation for such reasoning can be found in the set theory. Fused features are in fact represented as an intersection of features from thermal and RGB images, while modality-specific features remain unused in disjunctive parts of the feature sets of each modality.

This shows that a large amount of features remains unused. While pixel level fusion resembles finding intersection of two sets, we would like to introduce a union of **features** present in both modalities. This is accomplished by a unique multimodal transformer with a novel cross-referencing self-attention mechanism, called the Cross-modality Attention Transformer (CAT), to consider cross-modality information as keys to compute the weights on the current modality values.

Furthermore, after extracting modality-specific features, we propose a Multimodal Fusion Transformer (MFT) to perform the fusion process on every pair of ROIs. Our MFT further improves the detection performance by merging the multimodal features simultaneously with the help of the self-attention mechanism. Moreover, in order to learn distinct feature representations between foreground (pedestrian) and background, a novel content-aware objective function is proposed to guide the model, which shortens the intra-class variation and expands the inter-class variation.

Our main contributions can be summarized as follows:

- We have identified the disadvantages of the recent techniques relying on fused features for multimodal pedestrian detection, and introduced a new modality fusion scheme, which can effectively utilize modality-specific information of each modality.
- To our best knowledge, we are among the first to propose a Transformer-based network to enhance modality-specific features by cross-referencing the other modality. In feature extraction, we rely on the Cross-Modality Attention Transformer (CAT), which identifies related features in both modalities and consults the second modality in order to perform information aggregation based on the union of sets instead of the intersection type of fusion.
- Thanks to the ability of transformer networks to identify correlations between heterogeneous data, in this case RGB and thermal ROIs, our proposed Multimodal Fusion Transformer (MFT) performs association between detected regions in a more efficient way, compared to classical CNN methods.
- In our experiments, we qualitatively and quantitatively verify the performance of our model against recent relevant methods and achieve comparable or better detection results on the *KAIST* and *CVC-14* datasets.

2 Related Work

2.1 Multimodal Pedestrian Detection

Although deep learning methods have significantly advanced and dominated conventional RGB-based detection, detecting pedestrians in adverse weather and lighting conditions, background clutter or occlusions, is still a non-trivial problem. Motivated by this, numerous researchers have developed different fusion schemes relying on an additional modality to improve detection performance. Hwang et al. [7] have proposed a widely used pedestrian dataset with synchronized color and thermal image pairs. Next, Liu et al. compared various fusion architectures and proposed an important baseline model based on the halfway fusion [9] and Faster R-CNN [22]. Then, in the papers of Konig et al. [11] and Park et al. [20], the fusion models based on Faster R-CNN were also proposed. Moreover, in order to distinguish pedestrian instances from hard negatives, additional attributes have been introduced. For instance, Li et al. [13] leveraged the auxiliary semantic segmentation task to boost pedestrian detection results. Zhang et al. [29] also utilized the weak semantic labels to learn an attentive mask helping the model to focus on the pedestrians. In addition, the illumination condition of the scenes is also studied to improve the fusion results. For example, Guan et al. [6] and Li et al. [14] estimate the lighting condition of the images to determine the weights between thermal and color features. Furthermore, the misaligned and unpaired problems between the modalities have been investigated by Zhang et al. [31] and Kim et al. [10]. However, in most of the aforementioned methods, pedestrian detection is usually performed after the features have been fused. The discussion about exploring or enhancing the specific features of each modality is quite limited.

2.2 Multimodal Transformers

The self-attention mechanism of transformers has shown its advantages in many natural language processing and computer vision tasks [25, 3], and recently, it has been also applied to various multimodal fusion problems, such as image and video retrieval [4, 1], image/video captioning [18, 24, 8, 15], visual tracking [27] and autonomous driving [21].

Typically, multimodal inputs to transformers are allowed free attention flow between different modalities [19]. For instance, spatial regions in the image and audio spectrum would be considered and aggregated simultaneously without limitation. However, unlike audiovisual learning tasks, aligned thermal and color image pairs have more features in common, such as the shape of objects. Directly applying traditional self-attention to the image pairs might have difficulties to extract useful features due to the redundant information. In addition, the literature on fusing thermal and color images relying on transformers for pedestrian detection is quite limited. Hence, the strategies for utilization of shared information and exploring specific cues of each modality remains an important issue.

3 Proposed Method

The objective of our proposed fusion scheme is to explore potential modality-specific features and perform multimodal pedestrian detection using thermal and color image pairs as input. Our model consist of three main parts: (1) two-stream feature extractor with cross-modality attention, (2) modality-specific region proposals, and (3) multimodal fusion. As illustrated in Fig. 1, we rely on two independent feature extractors to obtain the features from color and thermal image pair I_c and I_t.

Simultaneously with the extraction, we feed feature maps from feature extractors to our proposed Cross-modality Attention Transformer (CAT), for considering cross-modality information used to enhance modality-specific features. The attended results are concatenated with each original input feature maps and forwarded to the corresponding region proposal network $\mathrm{RPN_c}$, and $\mathrm{RPN_t}$ to find modality-specific ROIs. Finally, the proposed Multimodal Fusion Transformer (MFT) merges the ROI pair R_c and R_t from ROI pooling module and output classification and bounding box predictions. In the following subsections, we explain the details of each contribution.

3.1 Cross-Modality Attention Transformer

In previous studies, numerous fusion methods have been proposed to merge and utilize the attributes of each modality in a proper manner. However, the

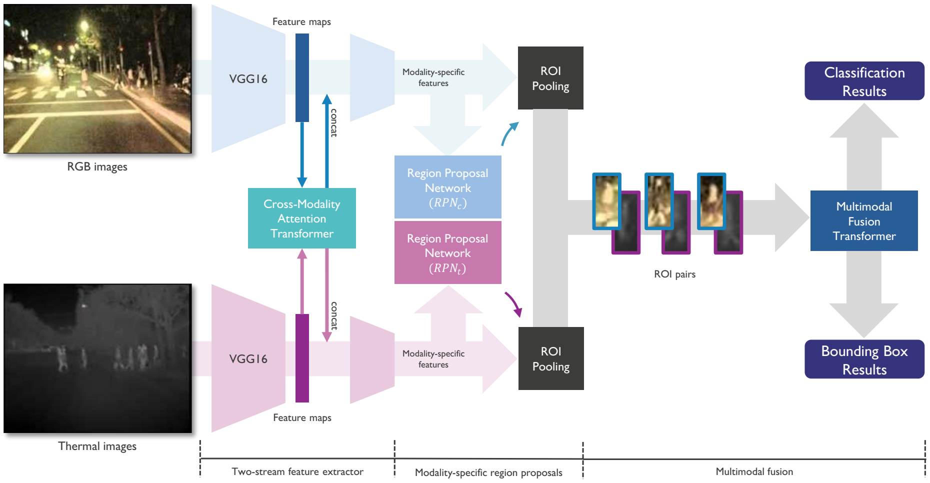
RGB images
VGG16
Feature maps
concat
Cross-Modality Attention Transformer
Modality-specific features
Region Proposal Network (RPN_c)
Region Proposal Network (RPN_t)
ROI Pooling
ROI pairs
Multimodal Fusion Transformer
Classification Results
Bounding Box Results
Thermal images
VGG16
Feature maps
concat
Modality-specific features
ROI Pooling
Two-stream feature extractor
Modality-specific region proposals
Multimodal fusion

Fig. 1. Our proposed multimodal pedestrian detection network. We propose a two-stream feature extractor with Cross-modality Attention Transformer (CAT) to extract modality-specific features. After fetching the modality-specific region proposals, Multimodal Fusion Transformer (MFT) is introduced to merge the ROI pairs for class and bounding box predictions.

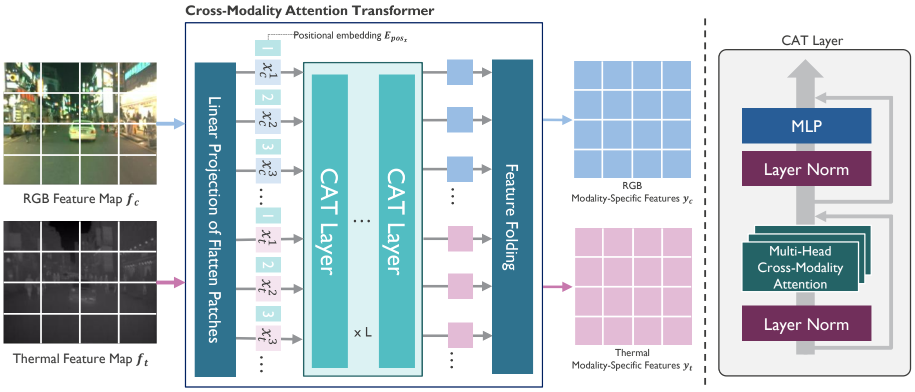
Cross-Modality Attention Transformer
Positional embedding E_{pos_x}
1
x_c^1
2
x_c^2
3
x_c^3
1
x_t^1
2
x_t^2
3
x_t^3
RGB Feature Map f_c
Thermal Feature Map f_t
Linear Projection of Flatten Patches
CAT Layer
CAT Layer
x L
Feature Folding
RGB Modality-Specific Features y_c
Thermal Modality-Specific Features y_t
CAT Layer
MLP
Layer Norm
Multi-Head Cross-Modality Attention
Layer Norm

Fig. 2. Our proposed Cross-modality Attention Transformer (CAT). We divide feature maps f_c and f_t from the two feature extractors into patches as input, and fold the output patches to form the attended modality-specific feature maps y_c and y_t. Different from previous works, we introduce cross-modality attention to utilize the other modality information for encouraging the model to focus on the regions ignored by the current modality but highlighted by the other.

discussion about enhancing modality-specific features before fusion is quite limited. Modality-specific information plays an important role in single modality detection, such as textures in color images. Therefore, in this part, we aim at

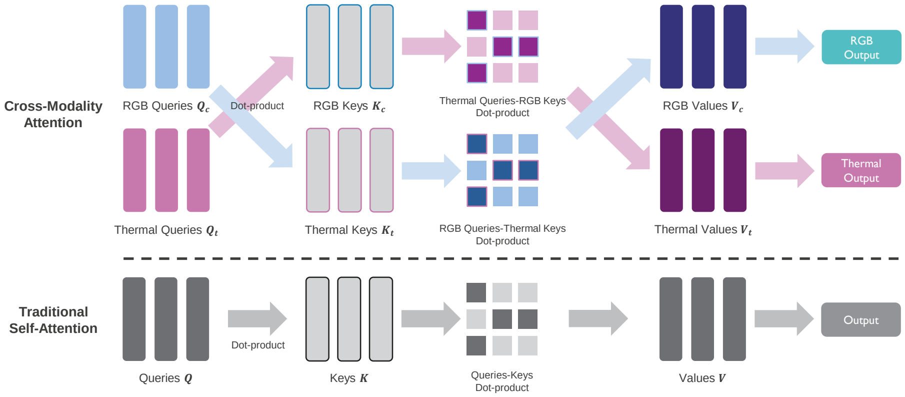

Fig. 3. Illustration of cross-modality attention. Instead of computing the queries with all modality keys, we introduce a novel way to only reference the complementary keys for introducing new perspectives from the other modality.

enhancing and preserving the modality-specific features before fusion by our proposed cross-modality attention mechanism. Specifically, we argue that with our enhanced modality-specific features it is possible to achieve more accurate region proposals.

Cross-Modality Attention As we have already learned, fused features usually cannot fully exploit modality-specific information and become dominated by strong cues from one of the modalities. In order to explore specific features of each modality, we introduce a novel attention mechanism to reference cross-modality data and to enhance our feature extraction. The main idea behind this design is to utilize the cross-modality information in order to encourage the model to focus on the regions ignored by the current modality but highlighted by the other. For this purpose, we use cross-modality attention transformers, which significantly outperform conventional CNN in discovering subtle and distant correlations in different modalities.

As illustrated in Fig. 3, instead of computing the scaled dot-product of the query with all modality keys, we only compute the scaled dot-product with the complementary keys. To be more specific, following [3]'s standard ViT architecture, we divide the input feature map pairs f_c^i and f_t^i from a certain layer i of the feature extractors into N_x patches for each modality, and then we flatten the patches and project them into N_x d_x-dimensional input sequences as $(x_c^1, ..., x_c^{N_x})$ and $(x_t^1, ..., x_t^{N_x})$.

Further, we add N_x d_x-dimensional learnable positional embeddings E_{pos_x} to each modality, and define the new matrices as X_c and X_t of size $\mathbb{R}^{N_x \times d_x}$. Different from the traditional transformer layers in ViT [3], we introduce CAT layers containing cross-modality attention module to utilize the other modality information. Our cross-modality attention operates on the queries and values from color patches as Q_c, V_c, keys from thermal patches as K_t. Hence, in every CAT layer, the cross-modality attention matrix of the color sequence for single head can be written as:

$$\text{Attention}(Q_c, K_t, V_c) = \text{softmax}(\alpha Q_c K_t^T) V_c, \tag{1}$$

where α is the scaling factor, and $Q_c = \boldsymbol{W}_q X_c$, $K_t = \boldsymbol{W}_k X_t$, and $V_c = \boldsymbol{W}_v X_c$ are linear transformations. $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d_x \times \frac{d_x}{N_h}}$ are the weight parameters for query, key, and value projections and N_h is the number of heads. As same as above, the attention matrix of the thermal sequence is: $\text{Attention}(Q_t, K_c, V_t) = \text{softmax}(\alpha Q_t K_c^T) V_t$. Different from the other multimodal transformers [21, 26, 16], we do not consider all the modality keys to find the attention weights. Instead, we introduce the perspective from the other modality by cross-referencing the keys to see if there is any target sensed by the other sensor and enhance the current sensor's features.

The output of the cross-modality attention module is then passed into Layer Normalization and MLP layers to get the attended features for color and thermal modalities. Next, we repeat several CAT layers and fold the attended output patches to form the feature maps y_c^i and y_t^i, which represent the additional modality-specific features learned from the other modality. Note that we still use the values from each modality to form the outputs, which means we do not directly fuse the multimodal features here. Then, we concatenate the output feature maps to the corresponding input features. Network-in-Network (NIN) [17] is applied to reduce the dimension and merge them with the input features. Furthermore, we apply our CAT on three different scales for considering coarse to fine-grained modality-specific features in practice.

Modality-Specific Region Proposal In order to fully exploit the modality-specific features, we propose two independent region proposal networks to find the proposals separately. Different from the previous works [31, 9, 11, 13, 20], we suggest that using the fused features to perform region proposal might limit the discrimination ability because the fused features might be dominated by one modality without considering the other. Therefore, in our work, we perform region proposal separately relying on our enhanced modality-specific features to explore the potential candidates. Afterwards, we use IoU threshold to match the proposals from the two modalities to fetch ROI pairs. In order to maximize the recall rate, we form the union of the proposals to involve all the possible candidates and to avoid mismatches. In other words, if a proposal from thermal sensor is not matched, we still use the bounding box to fetch the ROI from the RGB sensor to form a ROI pair.

3.2 Multimodal Fusion

In order to optimally use the modality-specific cues from thermal and color images to perform detection, we propose a Multimodal Fusion Transformer (MFT)

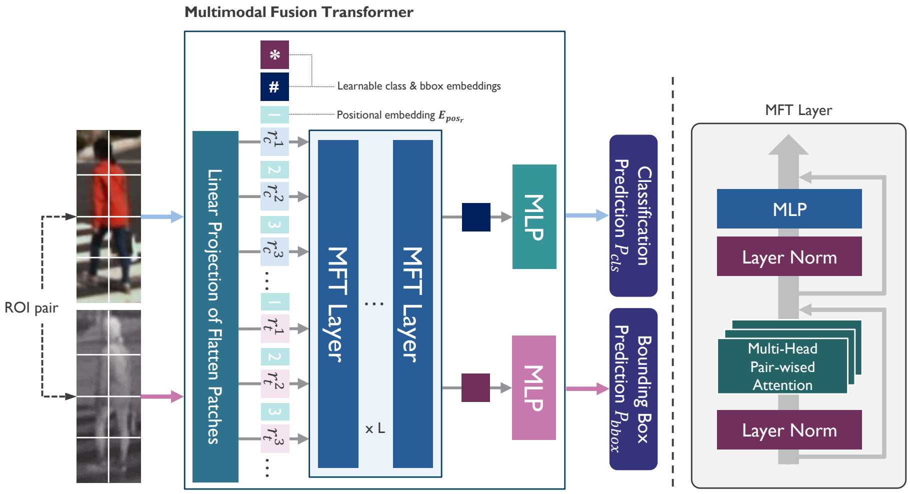
Multimodal Fusion Transformer
*
#
Learnable class & bbox embeddings
Positional embedding E_{pos_r}
ROI pair
Linear Projection of Flatten Patches
1
r_c^1
2
r_c^2
3
r_c^3
1
r_t^1
2
r_t^2
3
r_t^3
MFT Layer
…
MFT Layer
x L
MLP
MLP
Classification Prediction P_{cls}
Bounding Box Prediction P_{bbox}
MFT Layer
MLP
Layer Norm
Multi-Head Pair-wised Attention
Layer Norm

Fig. 4. Our proposed Multimodal Fusion Transformer (MFT). We divide ROI pairs from the two modalities into patches and prepend *class* and *bbox* for learning image representations. Different from CAT, we treat all the modality input sequences equally to apply pair-wised attention for computing attention weights.

to perform the fusion of the features. Instead of considering the whole image pair with background clutter, our fusion scheme only focuses on combining each ROI pair separately for lower computation time. In addition, to learn more discriminative feature representations, we introduce content-aware loss to enforce MFT to group the features based on the content.

Modality-Specific Features Fusion As illustrated in Fig. 4, we divide each ROI into N_r patches and project the flatten patches into N_r d_r-dimensional input sequences as $(r_c^1, ..., r_c^{N_r})$ and $(r_t^1, ..., r_t^{N_r})$. In addition, we also add N_r d_r-dimensional learnable positional embeddings E_{pos_r}, and similar to [3], we introduce two d_r-dimensional learnable embeddings *class* and *bbox*, whose output state serve as the image representation for classification and bounding box predictions. Then, we concatenate input sequences with the embeddings as input and feed them into the MFT layers.

Different from the traditional ViT [3], we propose pair-wised attention module inside the MFT layers to apply the self-attention to all the input sequences to perform prediction. In particular, for merging the ROI pairs, we need to consider all the pair-wised modality patches to transfer the complementary information. During the pair-wised attention, we mix all the paired sequences and compute the scaled dot-product of queries with **all** the modality keys to fuse the features from the two sensors. After several MFT layers, finally, we forward the output of *class* and *bbox* token to two independent MLP layers for the classification prediction P_{cls} and bounding box prediction P_{bbox}. We argue the differences from the previous studies [6, 14] as follows: without designing external network to learn balancing parameters, we utilize the self-attention mechanism to find the attention weightings and perform information fusion on each ROI pair.

Content-Aware Loss For learning better feature representations, we propose content-aware loss to utilize the label information of each ROI pair. We impose content-aware loss $\mathcal{L}_{ca}$ on each output state of *class* token $Y_{cls,c}$, where c indicates the class label, 0 for background and 1 for foreground, to maximize the inter-class discrepancy and minimize intra-class distinctness. Specifically, for each input batch, we average the all the $Y_{cls,1}$ to fetch the representative feature of foreground: $Y_{fg} = \sum_{batch} Y_{cls,1}/N_f$, where N_f represents the number of outputs with pedestrian annotation. As the same way, we can fetch representative feature of background Y_{bg}. Then, the distance between each $Y_{cls,c}$ and the corresponding center feature can be calculated as: $d_{bg} = \sum_{batch} \|Y_{cls,0} - Y_{bg}\|$, $d_{fg} = \sum_{batch} \|Y_{cls,1} - Y_{fg}\|$. With the above definitions, the proposed content-aware loss $\mathcal{L}_{ca}$ can be written as:

$$\mathcal{L}_{ca} = \max\left(d_{bg} + d_{fg} - \|Y_{bg} - Y_{fg}\| + m, 0\right), \tag{2}$$

where $m > 0$ is the margin enforcing the separation between foreground and background features. By this way, we shorten the intra-class feature distance and expand the inter-class distance for better discrimination ability. For each training iteration, we optimize the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{RPN_c} + \mathcal{L}_{RPN_t} + \mathcal{L}_{cls} + \mathcal{L}_{bbox} + \lambda\mathcal{L}_{ca}, \tag{3}$$

where $\mathcal{L}_{RPN_c}$ and $\mathcal{L}_{RPN_t}$ represent the loss from region proposal networks [22] of color and thermal modality, and $\mathcal{L}_{ca}$ is weighted by a balancing parameter λ. Similar to the Faster R-CNN [22], we use cross entropy and smooth L1 loss as the classification $\mathcal{L}_{cls}$ and bounding box regression loss $\mathcal{L}_{bbox}$ of MFT. Moreover, for better understanding, we also show the pseudo-code of our proposed method in Alg. 1 to illustrate the whole process.

4 Experiments

In order to evaluate the performance of our proposed method, we conduct several experiments on the KAIST [7] and CVC-14 [5] datasets to compare with the previous methods. Furthermore, we also conduct ablation studies to demonstrate the impact of the proposed components. In all the experiments, we follow [31]'s settings and use log-average Miss Rate over the range of $[10^{-2}, 10^{0}]$ false positive per image (FPPI) as the main metric to compare the performance.

4.1 Dataset and Implementation Details

KAIST dataset The KAIST dataset [7] consists of 95,328 visible-thermal image pairs captured in urban environment. The annotation includes 1,182 unique

Algorithm 1: Multimodal Fusion for Pedestrian Detection

```
Input  : Color Image: $I_c$, Thermal Image: $I_t$
Output : Classification and bounding box predictions $P_{cls}, P_{bbox}$
// Step 1: Feature extraction
for three different scales $i$ during feature extraction do
    Extract color and thermal feature maps $f_c^i$ and $f_t^i$ from $\mathrm{VGG_c}(I_c)$ and $\mathrm{VGG_t}(I_t)$.
    // Enhance modality-specific features by Cross-modality Attention Transformer
       (CAT)
    $y_c^i = \mathrm{CAT}(f_c^i)$, $y_t^i = \mathrm{CAT}(f_t^i)$
    // Concatenate with $f_c$ and $f_t$, and reduce the dimension with NIN [17]
    $f_c^{i+1} = \mathrm{NIN}(f_c^i, y_c^i), f_t^{i+1} = \mathrm{NIN}(f_t^i, y_t^i)$
end

// Step 2: Modality-specific region proposals
// Use feature extractor outputs $F_c$ and $F_t$ as inputs
$R_c = \mathrm{ROIpooling}(\mathrm{RPN_c}(F_c)), R_t = \mathrm{ROIpooling}(\mathrm{RPN_t}(F_t))$

// Step 3: ROI matching
$R_p = \mathrm{ROImatch}(R_c, R_t)$

// Step 4: Multimodal Fusion
// Perform feature fusion and prediction by Multimodal Fusion Transformer (MFT)
forall ROI pairs $R_p^k$ and learnable embeddings $class$ and $bbox$ do
    $P_{cls}^k, P_{bbox}^k = \mathrm{MFT}(R_p^k, class, bbox)$
end
```

pedestrians with $103,128$ bounding boxes. After applying the standard criterion [7], there are $7,601$ training image pairs and $2,252$ testing pairs. We train our model on the paired annotations released by [31] and apply horizontal flipping with single scale 600 for data augmentation. For fair comparison with the reference state-of-the-art methods, we evaluate the performance on "reasonable" day, night, and all-day subsets defined by [7] with sanitized labels [13].

CVC-14 dataset The CVC-14 dataset [5] consist of $7,085$ and $1,433$ visible (grey) and thermal frames captured in various scenes at day and night for training and testing. Different from the KAIST dataset, the field of views of the thermal and visible image pairs are not fully overlapped and calibrated well. The authors provided separated annotations for each modality and cropped image pairs to make thermal and visible images share the same field of view. In our experiments, we use the cropped image pairs and annotations to evaluate our method. The data augmentation strategy is as same as the KAIST dataset.

Implementation In our proposed method, we use two independent VGG-16 [23] pretrained on ImageNet [12] to extract color and thermal modality feature. Subsequently, we apply our proposed Cross-modality Attention Transformer (CAT) on the last three different scales with patch size 16, 3, and 3 without overlap to reference the other modality. Each of the transformer contains 3 CAT layers. For the proposed Multimodal Fusion Transformer (MFT), we also use 3 MFT layers with patch size 3, and each input patch dimension is $1,024$. Except the attention modules, we follow [3]'s architecture to design CAT and MFT

Table 1. Comparisons with the state-of-the-art methods on the KAIST dataset.

Methods	Feature extractor	Miss Rate (IoU = 0.5)		
		All	Day	Night
ACF [7]	-	47.32%	42.57%	56.17%
Halfway Fusion [9]	VGG-16	25.75%	24.88%	26.59%
Fusion RPN + BF [11]	VGG-16	18.29%	19.57%	16.27%
IAF + RCNN [14]	VGG-16	15.73%	14.55%	18.26%
IATDNN + IASS [6]	VGG-16	14.95%	14.67%	15.72%
CIAN [30]	VGG-16	14.12%	14.77%	11.13%
MSDS-RCNN [13]	VGG-16	11.34%	10.53%	12.94%
AR-CNN [31]	VGG-16	9.34%	9.94%	8.38%
MBNet [32]	ResNet-50	8.13%	8.28%	7.86%
MLPD [10]	VGG-16	7.58%	7.95%	6.95%
Ours	VGG-16	**7.03%**	**7.51%**	**6.53%**

Table 2. Comparisons with the state-of-the-art methods on the CVC-14 dataset.

Methods	Feature extractor	Miss Rate (IoU = 0.5)		
		All	Day	Night
MACF [20]	-	69.71%	72.63%	65.43%
Choi et al. [2]	VGG-16	63.34%	63.39%	63.99%
Halfway Fusion [20]	VGG-16	31.99%	36.29%	26.29%
Park et al. [20]	VGG-16	26.29%	28.67%	23.48%
AR-CNN [31]	VGG-16	22.1%	24.7%	18.1%
MBNet [32]	VGG-16	21.1%	24.7%	13.5%
MLPD [10]	VGG-16	21.33%	24.18%	17.97%
Ours	VGG-16	**20.58%**	**23.97%**	**12.85%**

layers. The λ parameter for $\mathcal{L}_{ca}$ is 0.001. For more details, please refer to the Supplementary Materials.

4.2 Quantitative Results

Evaluation on the KAIST dataset As illustrated in Table 1, we evaluate our method and compare it with other recent related methods. Our method achieves 7.03% MR, 7.51% MR, and 6.53% MR on day, night, and all-day subsets under the 0.5 IoU threshold. This table clearly shows that our method can achieve superior performance on all the subsets.

Evaluation on the CVC-14 dataset As for the KAIST dataset, we show the evaluations of our method and compare it with other state-of-the-art models in Table 2. In this table, we follow the setting introduced in [20] to conduct the experiment. Our method achieves 20.58% MR, 23.97% MR, and 12.85% MR on

Table 3. Ablation studies of proposed cross-modality attention and content-aware loss on the KAIST dataset. The baseline model is Halfway fusion [9], and we evaluate the model with or without the proposed components to verify the improvements.

Methods	CAT		MFT	Miss Rate (IoU = 0.5)		
	Cross-modality Attention	Pair-wised Attention	Content-aware Loss	All	Day	Night
Baseline	-	-	-	25.75%	24.88%	26.59%
Ours	-	✓	-	13.54%	14.87%	13.01%
	-	✓	✓	12.42%	13.75%	12.11%
	✓	-	-	10.14%	10.87%	9.74%
	✓	-	✓	**7.03%**	**7.51%**	**6.53%**

day, night, and all-day subsets under the 0.5 IoU threshold. We can observe that our method outperforms the other method under all the subsets.

4.3 Ablation Study

Effects of cross-modality attention For further analysis of the effect of our proposed cross-modality attention, we conduct an experiment to compare the results with or without the cross-modality attention mechanism. In Table 3 we list four models to demonstrate the advantages of our method. Instead of using cross-modality attention in CAT, we use pair-wised attention to allow free attention flow between the modalities to compute the attention matrix (computing queries with all the modality keys) and to compare it with our proposed method. We find that cross-modality attention improves the performance of the reference models by a large margin. The performance can be improved by referencing the complementary information from the other modality rather than allowing free attention flow and directly fusing all the modality data in the early stage.

In addition, we also show the feature maps of the models with and without cross-modality attention in Fig. 5 to demonstrate our advantages qualitatively. Heat maps in this figure are generated by averaging the final convolution layer of RGB feature maps and superimposing them on the RGB image. We observe that with our proposed cross-modality attention, the model can correctly identify the pedestrians. In contrast, without cross-modality attention, the model can hardly focus on the targets.

Effects of content-aware loss In Table 3, we also compare the the results with and without the content-aware loss to discuss the effect of content-aware loss. We can see that our proposed loss further improves the detection performance irrespective of the cross-modality attention or pair-wised attention. Especially when the content-aware loss is applied to the model with cross-modality attention, the result shows the best performance and largest improvement among all the combinations, which can demonstrate the effectiveness this component.

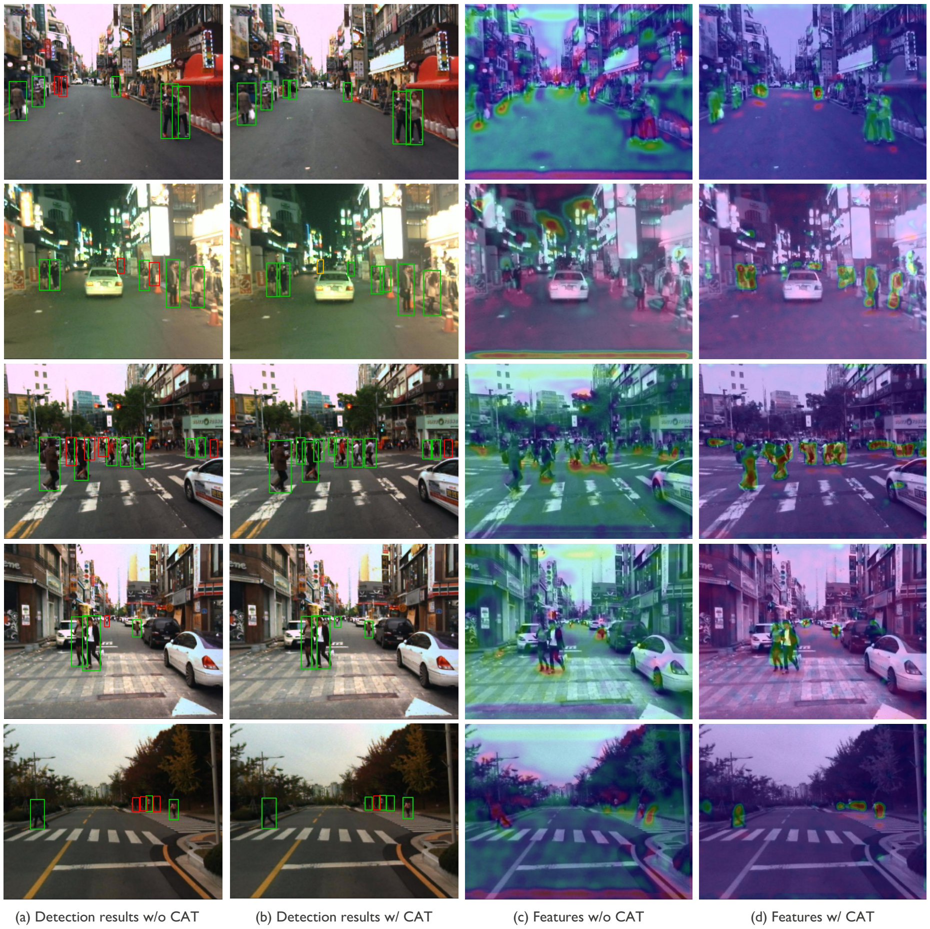

Fig. 5. The effects of cross-modality attention transformer. The left column (a) and (b) show the RGB image with detection results comparison, and the right column (c) and (d) show the feature maps with/without our proposed cross-modality attention transformer. The green boxes represent the correct detection results, and the red and orange boxes represent false negative and false positive samples respectively.

Effects of proposed components In order to demonstrate the contributions of all the proposed components, we evaluate the model with/without the components to verify the improvements. In Table 4, we use Halfway fusion [9] as our baseline model, and then apply the proposed components gradually to show the performance difference.

First, we only apply CAT to enhance modality-specific features and concatenate the features for single region proposal network. There is only a marginal improvement because the potential ROIs might not be fully explored. In addition, for different proposals, an advanced fusion process is also required to

Table 4. Ablation studies of our proposed components on the KAIST dataset. The baseline model is Halfway fusion [9], and we evaluate the model with/without the proposed components to verify the improvements. The results show that the proposed transformer significantly improves the detection performance and outperforms the previous models to achieve the best performance among all the combinations.

Methods	CAT	Modality-specific RPNs	MFT	Miss Rate (IoU = 0.5)		
				All	Day	Night
Baseline	-	-	-	25.75%	24.88%	26.59%
Ours	✓	-	-	20.45%	21.66%	20.47%
	✓	✓	-	13.12%	14.45%	12.87%
	✓	✓	✓	**7.03%**	**7.51%**	**6.53%**

handle various illumination scenes. Secondly, we apply two independent region proposal networks for each modality to find proposals and merge the ROI pairs by concatenation. This leads to a larger improvement, demonstrating that the enhanced modality-specific features and independent RPNs can truly help the model to find more potential proposals.

Furthermore, we apply MFT to fuse the ROI pairs instead of feature concatenation to verify the effect of fusion by the attention mechanism. The results show that the proposed transformer significantly improves the detection performance and outperforms the previous models to achieve the best performance among all the combinations.

5 Conclusions

In this paper, we propose a novel fusion scheme to combine visible and thermal image pairs to perform multimodal pedestrian detection. We introduce the Cross-modality Attention Transformer (CAT) to reference complementary information from the other modality during feature extraction to investigate the potential of modality-specific features to improve detection performance. Instead of directly fusing all the modality data, by our proposed cross-modality attention, we can extract more discriminative details. Moreover, we propose modality-specific region proposal networks to explore the potential candidates and merge the modality features pair-wisely by our proposed Multimodal Fusion Transformer (MFT) to make better predictions. Finally, a novel content-aware loss is proposed to separate the foreground and background features to increase the discrimination ability. The experiment results on the public KAIST and CVC-14 datasets confirm that our method can achieve state-of-the-art performance, and the ablation studies also clarify the effectiveness of the proposed components.

Acknowledgments This work was funded by EU Horizon 2020 ECSEL JU research and innovation programme under grant agreement 876487 (NextPerception) and by the Flemish Government (AI Research Program).

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
2. Choi, H., Kim, S., Park, K., Sohn, K.: Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In: International Conference on Pattern Recognition (ICPR) (2016)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
4. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision (ECCV) (2020)
5. González, A., Fang, Z., Socarras, Y., Serrat, J., Vázquez, D., Xu, J., López, A.M.: Pedestrian detection at day/night time with visible and fir cameras: A comparison. Sensors (2016)
6. Guan, D., Cao, Y., Yang, J., Cao, Y., Yang, M.Y.: Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Information Fusion (2019)
7. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baselines. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
8. Iashin, V., Rahtu, E.: Multi-modal dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
9. Jingjing Liu, Shaoting Zhang, S.W., Metaxas, D.: Multispectral deep neural networks for pedestrian detection. In: Proceedings of the British Machine Vision Conference (BMVC) (2016)
10. Kim, J., Kim, H., Kim, T., Kim, N., Choi, Y.: Mlpd: Multi-label pedestrian detector in multispectral domain. IEEE Robotics and Automation Letters (2021)
11. Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., Teutsch, M.: Fully convolutional region proposal networks for multispectral person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems (2012)
13. Li, C., Song, D., Tong, R., Tang, M.: Multispectral pedestrian detection via simultaneous detection and segmentation. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
14. Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition (2019)
15. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
16. Li, Z., Li, Z., Zhang, J., Feng, Y., Zhou, J.: Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021)

17. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
18. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems (2019)
19. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems (2021)
20. Park, K., Kim, S., Sohn, K.: Unified multi-spectral pedestrian detection based on probabilistic fusion networks. Pattern Recognition (2018)
21. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (2015)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)
26. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2021)
27. Xiao, Y., Yang, M., Li, C., Liu, L., Tang, J.: Attribute-based progressive fusion network for rgbt tracking (2022)
28. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
29. Zhang, H., Fromont, E., Lefèvre, S., Avignon, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
30. Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., Hussain, A.: Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion (2019)
31. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
32. Zhou, K., Chen, L., Cao, X.: Improving multispectral pedestrian detection by addressing modality imbalance problems. In: European Conference on Computer Vision (ECCV) (2020)