Unmanned Systems, Vol. 10, No. 3 (2022) 1–10

DOI: 10.1142/S2301385023500085

UNMANNED SYSTEMS
https://www.worldscientific.com/worldscinet/us

# Augmented Convolutional Neural Network Models with Relative Multi-Head Attention for Target Recognition in Infrared Images

Billel Nebili*, Atmane Khellal*,‡, Abdelkrim Nemra*, Laurent Mascarilla†

*Ecole Militaire Polytechnique, UER SAI,
Algiers 16111, Algeria

†Laboratoire MIA, Université de La Rochelle,
Avenue Michel Crépeau, F-17042 La Rochelle Cedex, France

Deep convolutional neural network (CNN) models are typically trained on high-resolution images. When we apply them directly to low-resolution infrared images, for example, the performances will not always be satisfactory. This is due to CNN layers that operate in a local neighborhood, which is already poor in information for infrared images. To overcome these weaknesses and increase information of global nature, a hybrid architecture based on CNN with self-attention mechanism is proposed. This later provides information about the global context by capturing the long-range interactions between the different parts of an image. In this paper, we have incorporated a convolutional–attentional form in the top layers of two pre-trained networks VGGNet and ResNet. The convolutional–attentional form is a concatenation of two paths; the original convolutional feature maps of the pre-trained network, and the output of a relative multi-head attentional block. Extensive experiments are conducted in the FLIR starter thermal dataset, where we achieve a 97.07% overall accuracy in the four-class FLIR starter thermal dataset. Moreover, the proposed architectures exceed the state of the art in target recognition on two-class FLIR starter thermal dataset with a 3.5% improvement in overall classification accuracy. In addition, a study on the effect of different hyper-parameters and error analysis is carried out to give some research forward directions.

*Keywords*: Convolution; FLIR; infrared images; self-attention; target recognition; transfer learning.

US

## 1. Introduction

Convolutional neural networks (CNNs) have been used as a powerful tool for several pattern recognition applications such as classification, localization and detection [1–3]. Due to the availability of large datasets and computing resources, a series of more and more powerful convolutional architectures for vision tasks have been proposed: VGGNet [4], ResNet [5], GoogLeNet [6], etc. All of these networks are essentially based on convolutional layers. The design of the latter imposes two crucial inductive biases, locality via a limited reception field (the image's adjacent pixels are related to each other) and translation equivariance via weight sharing (the different parts of the image must be treated similarly independently of their absolute position). So, CNNs can preserve the spatial information i.e. the information related to the specific arrangement of the pixels is not lost. However, these properties remain local in nature which prevents the extraction of global contexts (GCs) in an image.

To overcome the above problem, the Vision Transformer (ViT), a new paradigm based on the self-attention mechanism, has been introduced to represent high-level concepts in images. The self-attention introduced by [7] has become the leader in natural language processing (NLP). It can operate in a GC by directly capturing long-distance interactions, and in a parallel way allowing the exploitation of the strengths of modern hardware resources. Contrary to convolution layers, the new value of a pixel depends on all the other pixels in the image, meaning that the receptive field is the whole image, and it is not a neighborhood grid.

CNN structure does not explicitly contain position information, unlike self-attention which adds absolute position

Received 23 January 2022; Accepted 27 April 2022; Published xx xx xx. This paper was recommended for publication in its revised form by editorial board member, Zhi Gao.
Email Addresses: ‡*khe.atmane@gmail.com; atmane.khellal@emp.mdn.dz*

embeddings to the input. Absolute position embeddings do not guarantee translation equivariance. This later can be achieved by introducing relative position embeddings (RPE) [8] instead of absolute position embeddings.

Instead of running a single attention model, it is advantageous to repeat its calculations several times in parallel with different parameters, each of them is called an attention head [7]. A self-attentive layer with enough heads and using relative position encodings can express any convolutional filter [9].

Multiple works have proposed attention schemes for visual tasks to address the limitations of convolutions [10–13]. Researchers have addressed the use of self-attention in three ways. The first is to replace the convolutions entirely by the attention mechanism [14–17]. These latter models require large datasets for training to obtain good performances [14]. The second view consists of models that alternate between self-attention layers and convolution layers [12,18–23]. The third way, which we have used in our work, is based on the combination of CNNs and the attention mechanism.

In [24], the authors introduced Squeeze-and-Excitation (SENets) block to enhance the representational power of CNNs with dynamic channel-wise feature recalibration. A GC block was designed in [25], it has a similar structure to the "Squeeze–Excitation block". Hu *et al.* [26] presented a pair of operators: gather and excite which perform a channel-wise reweighting for more efficient exploitation of the CNN context. Another attention blocks named "Bottleneck Attention Module (BAM)" and "Convolutional Block Attention Module (CBAM)" were proposed in [27,28], respectively. An attention map is generated along two distinct paths, spatial and channel, in a parallel manner for BAM and sequentially for CBAM. Bello *et al.* [29] proposed to augment convolutional feature maps with a set of feature maps generated by the self-attention mechanism. Srinivas *et al.* [30] replaced spatial convolutions in the last three bottleneck blocks of ResNet with self-attention mechanism.

The works mentioned above, related to the attention mechanism, try to integrate attention in pre-trained models to improve the classification performance in RGB images. In our work, we focus on infrared (IR) images that have low resolution. Models like VGGNet and ResNet are trained to efficiently learn feature maps from high-resolution images, which is not the case for thermal images. So, the objective is to adapt these models for target recognition in low-resolution images. The inferior layers of these CNN represent simple and general aspects of images, while the superior layers represent much more complex aspects of an image. So, to improve and adapt these CNN on another type of images (IR images in our case), we propose to modify the superior layers. We take two well-known models, pre-trained on large images (VGGNet and ResNet), and we modify the last stage by incorporating the attention mechanism inside and making the necessary modifications. This hybrid architecture allows to use inferior feature maps that are well optimized for general features, both for convolutions and self-attention; by entrusting convolutions for spatial learning and self-attention for GC learning. These adapted models improve the classification performances on FLIR starter thermal dataset [31].

We can summarize the contribution offered by this paper in the following points:

- We propose two hybrid architectures based on CNN and self-attention mechanism for target recognition in IR images. These architectures allow capturing both types of information, local and global.
- We exceed the state of the art in target recognition on two-class FLIR starter thermal dataset.
- We report the results of classification in four-class FLIR starter thermal dataset, which will be the reference of comparison for next research works. As far as we know, there is no work for classification task on four-class FLIR starter thermal dataset in the literature.

This paper is organized as follows. In Sec. 2, we explain the background used for our approach, specifically the transfer learning method and the attention mechanism. In Sec. 3, we describe in detail the proposed augmented CNN models. Then, Sec. 4 describes the conducted experiments and the obtained results. Finally, the main findings of this work and some forward perspectives are given in Sec. 5.

## 2. Background

Before we start describing our approach, we give some backgrounds related to transfer learning and attention mechanism.

### 2.1. *Pre-trained CNN models*

Transfer learning has been very successful with the rise of deep learning. In fact, the CNN models used in this field often require high-computation times and important resources. However, by using the pre-existing models, trained on a huge dataset, as a starting point, transfer learning allows to quickly develop efficient models and to efficiently solve complex problems in computer vision. In this paper, we will cover two pre-trained models for image classification that are widely used.

#### 2.1.1. *VGGNet*

VGGNet is one of the most popular pre-trained models for image classification. It was trained using over one million

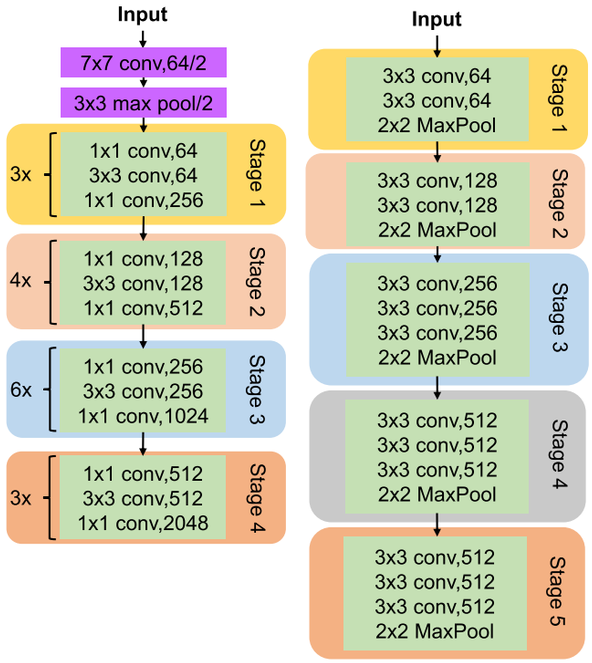


Fig. 1. The feature extraction stage; left: ResNet50; right: VGG16.

images from the ImageNet dataset [32]. It has a simple architecture (Fig. 1, right), it utilizes small 3×3 convolutional layers stacked on top of each other in increasing depth. The volume is reduced by max pooling. The final components are two fully connected layers, each with 4096 nodes, followed by a softmax classifier. The best performing architectures of VGGNet were VGG19 with 19 weight layers and VGG16 with 16 weight layers. We have conducted our experiments on VGG16.

#### 2.1.2. *ResNet*

The state-of-the-art CNN architecture requires going deeper and deeper. Networks like VGGNet stacked many convolutional layers. However, stacking layers on top of each other is not enough to increase networks' depth. Deep networks are hard to train because of some difficulties like optimization of networks, vanishing gradient problems, etc.

The residual network (ResNet) tries to solve these problems. Its core idea is introducing an "identity shortcut connection" that skips one or more layers. In this paper, ResNet50 (Fig. 1, left) network has been used. It had 50 layers of residual blocks.

ResNet50 is composed of four stages where each one contains building blocks: one convolution block and identify blocks. After all these blocks, there is a fully connected layer responsible for the classification task.

### 2.2. *Attention*

Attention is a mechanism used to model long-distance interaction, for example, across text in NLP [33]. It allows a model to attend to different parts by building shortcuts between a context vector and the input. This mechanism can be divided into several families, one of them is self-attention, which is the most popular type for computer vision tasks. Self-attention refers to attention applied to a single sequence to calculate its representation instead of multiple sequences.

#### 2.2.1. *Self-attention over images*

We have performed the self-attention mechanism as it was proposed in [7]. Let $H, W, D_{\text{in}}$ be the height, width and depth of the input tensor $X$. $d_v^h, d_k^h$ refer to the depth of values and the depth of queries and keys in single head attention, respectively.

First, the input tensor $X$ will be flattened to a matrix $X' \in R^{HW \times D_{\text{in}}}$. Then, $X'$ will be passed to the self-attention model with a single head $h$. The mechanism can be formulated with the following equation:

$$A_h = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k^h}}\right)V, \tag{1}$$

where $Q = X'W_q$, $K = X'W_k$ and $V = X'W_v$ are queries, keys and values, respectively. $W_q, W_k \in R^{D_{\text{in}} \times d_k^h}$ and $W_v \in R^{D_{\text{in}} \times d_k^v}$ are the weight matrices that we will be learned.

To improve the performance of the self-attention model, we often want to focus on several other positions when traversing the input tensor. The multi-head attention (MHA) mechanism allows us to have several subspace representations.

Let $N_h, d_h, d_k$ be the number of heads, the depth of values and the depth of queries and keys, respectively, in MHA with $N_h$ divided $d_h$ and $d_k$. From the input $X'$, we have three different groups of matrices: queries $\{Q_i\}_{i=1}^{N_h}$, keys $\{K_i\}_{i=1}^{N_h}$, values $\{V_i\}_{i=1}^{N_h}$. The attention for the head $h$ is computed by Eq. (1), using $Q_h$, $K_h$, $V_h$, $d_k^h, d_v^k$. The output of the MHA mechanism can be formulated as follows:

$$\text{MHA}(X) = \text{Concat}(A_1, \ldots, A_h, \ldots, A_{N_h})W^a, \tag{2}$$

where $W^a \in R^{d_v \times d_v}$ is the linear projection matrix. Finally, $\text{MHA}(X)$ is reshaped into a tensor of shape $(H, W, d_v)$.

As currently framed, the attention mechanism does not contain any encoding of position information of input elements. This makes it permutationally equivariant and therefore limited efficiency for vision tasks. The original publication [7] suggests the use of sinusoidal embeddings. However, recent research [8] shows that the use of RPE

allows obtaining much better accuracies. Next, we will give the necessary details for these later embeddings.

#### 2.2.2. *Relative position embeddings*

In this paper, we have used the attention with 2D RPE. The relative position will be encoded as follows.

The attention between two pixels $i = (i_x, i_y)$ and $j = (j_x, j_y)$ is computed as follows:

$$a_{i,j} = \frac{q_i^T}{\sqrt{d_k^h}}\left(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y}\right), \tag{3}$$

where $q_i$ is the $i$th row of $Q$, $k_j$ is the $j$th row of $K$ and $r^W_{j_x - i_x}$ and $r^H_{j_y - i_y}$ are the learned embedding for relative width $j_x - i_x$ and relative height $j_y - i_y$, respectively.

Now, for all input pixels, the two relative position matrices along height and width denoted $R_H, R_W \in R^{HW \times HW}$ are given by the following formulas $R_H(i,j) = q_i^T r^H_{j_y - i_y}$ and $R_W(i,j) = q_i^T r^W_{j_x - i_x}$.

According to Eqs. (1) and (3), the output of the attention module for a single head will be calculated as follows:

$$A_h = \text{Softmax}\left(\frac{QK^T + R_H + R_W}{\sqrt{d_k^h}}\right)V. \tag{4}$$

The heads of each layer share the same RPE $R_H$ and $R_W$. The overall architecture of attention and MHA mechanisms with RPE is shown in Fig. 2.

## 3. Augmented CNN Models

As mentioned before, visual attention forms can be divided into two types of attention: spatial and channel-wise. Our approach is based on spatial attention. In this section, the proposed convolution–attention block will be explained. Then, we incorporate this stage in VGG16 and ResNet50, which will be denoted AVGG16 and AResNet50, respectively.

### 3.1. *Convolutional–attentional form*

Given as input an intermediate feature map $X \in R^{H \times W \times D_{\text{in}}}$. The operations will follow two separate paths. The first path consists of $N$ succeeding standard blocks of convolutions. Let $X' \in R^{H \times W \times D_{\text{out}}}$ be the output of the last convolution block. The MHA block constitutes the second path. Let $X'' \in R^{H \times W \times d_v}$ be the output of the MHA block.

$$X'' = \text{MHA}(X) = \text{Concat}(A_1, \ldots, A_h, \ldots, A_{N_h}), \tag{5}$$

with $A_h$ is given by Eq. (4).

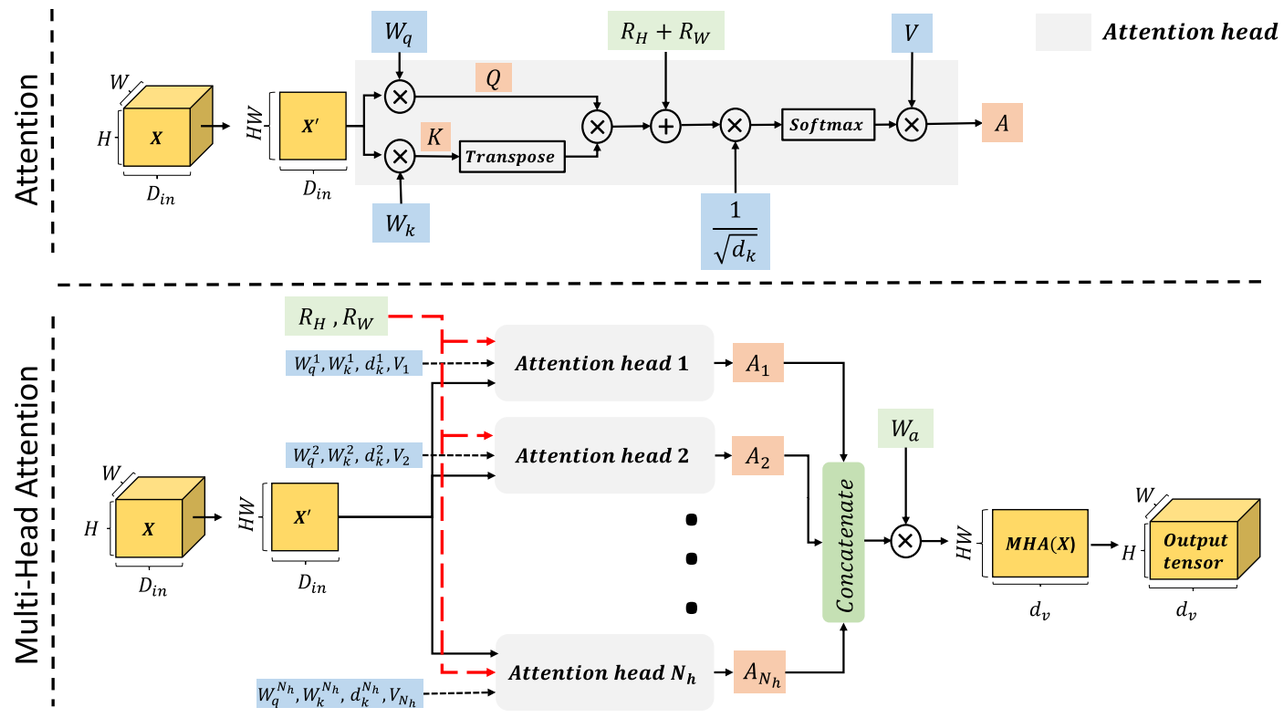


Fig. 2. Attention and MHA mechanisms with RPE.

To get the final output feature map $Y \in R^{H\times W\times(D_{\text{out}}+d_v)}$, we will combine $X, X'$ and $X''$ using the following formulas:

For nonresidual networks

$$Y = \text{Concat}(X', X''). \tag{6}$$

For residual networks

$$Y = \text{Concat}(X + X', X''). \tag{7}$$

The resulting augmented map $Y$ is followed by batch normalization [34] to scale the convolution feature maps and the MHA feature maps. To define $d_v$ and $d_k$, we note two parameters $\alpha$ and $\beta$, where $d_v = \alpha D_{\text{out}}$ and $d_k = \beta D_{\text{out}}$.

### 3.2. *Proposed architectures: AVGG16, AResNet50*

We propose to expand VGG16 and ResNet50. As mentioned before, the convolutional–attentional form is integrated in the top layers. The details of the proposed hybrid architectures are given below.

#### 3.2.1. *AVGG16*

VGG16 is composed of five stages. The last stage stacks three blocks of "$3 \times 3$ Conv, 512" and one block of max pooling "$2 \times 2$". The MHA mechanism will be applied on the output of stage 4, in parallel to the standard convolution in stage 5. The outputs of MHA and stage 5 will be concatenated along the spatial dimension (Eq. (6)). The architecture of the proposed stage 5 for VGG16 is described in Fig. 3.

#### 3.2.2. *AResNet50*

ResNet50 is composed of four stages. Each stage is built with residual blocks. The last stage contains three residual blocks. As VGG16, we take an intermediate feature map (stage 3 output), and process it through two parallel paths: residual block and MHA block. We concatenate the two paths outputs along the spatial dimension (Eq. (7)).

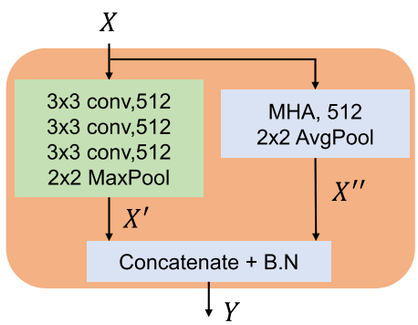


Fig. 3. The proposed stage 5 for VGG16; B.N: Batch Normalization.

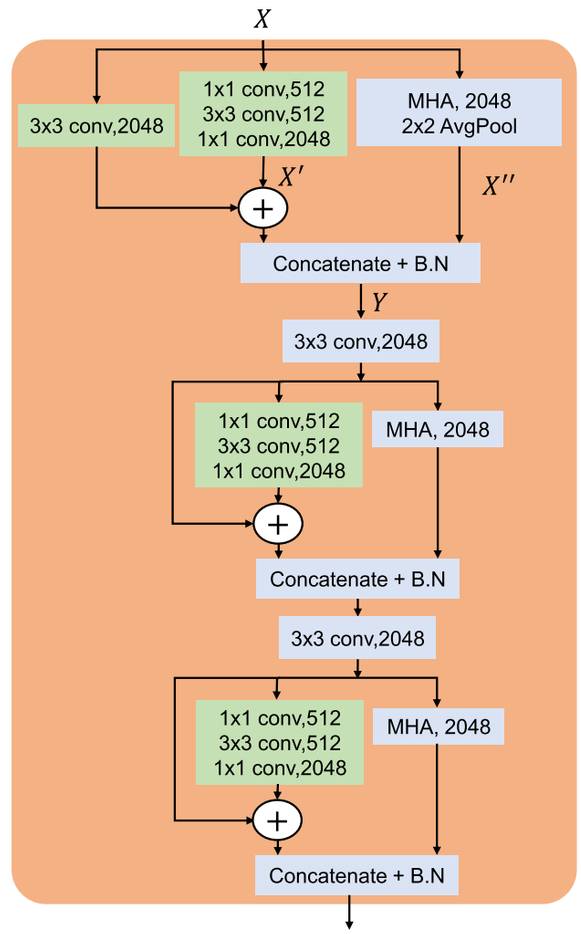


Fig. 4. The proposed stage4 for ResNet50; B.N: Batch Normalization.

We perform the same procedure for the two other residual blocks. The architecture of the proposed stage 4 for ResNet50 is described in Fig. 4.

To respect the sizes when performing concatenation and addition operations, we resort to: (1) in some cases downsampling the self-attention outputs before concatenating them with the convolutional feature maps. The downsampling is done by applying a $3 \times 3$ average pooling with a step of 2. (2) Add a convolution block for ResNet50 ($3 \times 3$ Conv, 2048) between the residual blocks.

## 4. Experiments and Results

In this section, we will describe the experimented work and the obtained results. First, FLIR starter thermal dataset will be briefly described. Then, we will give the experimental details for implementation. Next, experiments and results are reported and discussed. Finally, our results will be

Table 1. Number of training and test samples using the FLIR dataset.

| Class | Training | Test |
|---|---|---|
| People | 22,372 | 5779 |
| Vehicle | 41,260 | 5432 |
| Bicycle | 3986 | 471 |
| Dog | 226 | 14 |

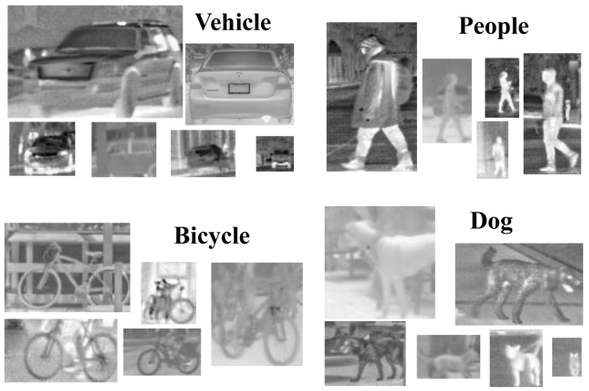


Fig. 5. Few example images of FLIR dataset.

compared with the state of the art for target recognition in FLIR dataset.

### 4.1. *FLIR dataset*

FLIR starter thermal dataset is a multi-spectral dataset released by the thermal camera manufacturer FLIR. It consisted of 10,288 RGB/IR annotated frames, distributed over four classes: people, bicycles, cars and dogs. FLIR is used for training and validation of detection algorithms. Our work is for the classification task. For this, we have created a classification dataset, which consists of the four classes. This dataset (hereby referred 'four-class FLIR') comprises more than 79,000 images (Table 1).

To compare with the state of the art, we have created a dataset as described in [35] (hereby referred 'two-class FLIR'), which consists of two classes: vehicles and people. Figure 5 gives example images from the FLIR dataset.

### 4.2. *Implementation*

All models are initialized with "ImageNet weights" and fine-tuned for 300 epochs. We perform data augmentation with several variations, including random horizontal flipping, rescaling in $[0, 1]$, random rotation, width and height shift, brightness shift, shear intensity and random zoom. Optimization is performed with a mini-batch size of 32 using: SGD with momentum 0.9, 1e$-$4 learning rate and 1e$-$5 learning rate decay for VGGNet (Adam with 1e$-$4 learning rate and 1e$-$5 learning rate for ResNet). When testing, the images are rescaled in $[0, 1]$. All images are resized to $60 \times 60$ pixels.

### 4.3. *Experiments*

As mentioned above, the experiments were conducted with two basic CNNs: VGG16 and ResNet50. We specify that for all architectures used in this work, we have used the network top shown in Fig. 6.

First of all, we have evaluated the Vision Transformer (ViT) [14] on the two variants of FLIR dataset (two and four classes). Second, we test standard VGG16 and ResNet50 architectures. Then, we evaluate the approach proposed in [29]. We augment VGG16 by augmenting the three blocks "$3 \times 3$ Conv, 512" in the stage 5 (named AAVGG16). We augment ResNet50 by augmenting the "$3 \times 3$ Conv, 512" blocks of each residual block in the stage 4 (named AAResNet50). Finally, the two proposed architectures (AVGG16, AResNet50) are evaluated.

All attention networks use $N_h = 8$, $\alpha = 0.125$, $\beta = 0.125$. The impact of the parameters $N_h, \alpha, \beta$ will be discussed below. Table 2 shows the experimental results which the mean and the standard deviation of the accuracy averaged over three runs are reported.

As expected, we obtained poor performance using the ViT, which is due to the problem of overfitting caused by the lack of training data in IR images-based recognition problems. The main issue with ViT is the need to pre-train on large datasets. Although our dataset contains 60 k images which is a small size dataset.

In all experiments, the proposed architectures (AResNet50, AVGG16) increase performance over the nonaugmented

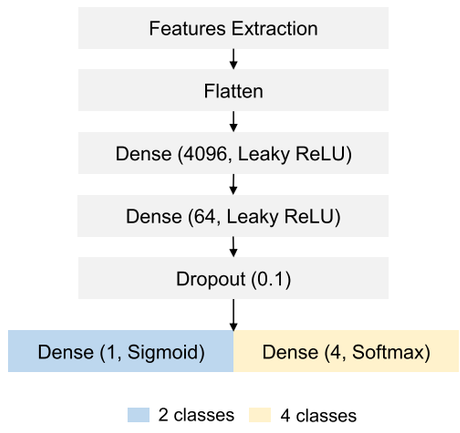


Fig. 6. Network top for classification.

Table 2. Image classification performance of different attention networks on FLIR dataset (two classes and four classes).

| Architecture | Acc (two classes) | Acc (four classes) |
|---|---|---|
| ViT [14] | $87.63\% \pm 0.05$ | $83.81\% \pm 0.06$ |
| VGG16 [4] | $97.36\% \pm 0.06$ | $95.98\% \pm 0.10$ |
| AAVGG16 [29] | $97.91\% \pm 0.06$ | $96.40\% \pm 0.06$ |
| AVGG16 (Our) | $98.10\% \pm 0.03$ | $96.71\% \pm 0.07$ |
| ResNet50 [5] | $97.88\% \pm 0.04$ | $96.21\% \pm 0.08$ |
| AAResNet50 [29] | $98.00\% \pm 0.04$ | $96.60\% \pm 0.06$ |
| AResNet50 (Our) | $98.33\% \pm 0.03$ | $97.07\% \pm 0.06$ |

baseline models (ResNet50, VGG16) and notably outperform the augmented models (AAResNet50, AAVGG16). We also observe that convolutional–attentional form brings an improvement in performance for nonresidual networks. As expected, the performance of residual networks is better than nonresidual ones.

Our AResNet50 and AVGG16 offer a competitive accuracy in "four-class FLIR" with 97.07%. Considering the issue that an evaluation based on accuracy for "four-class FLIR" is not very significant (since the number of examples for the four classes is unbalanced), we use precision, recall and $F1$ score metrics. Figures 7 and 8 show precision, recall and $F1$ score values and the confusion matrix for AResNet50 on four-class FLIR.

For easier comparison, we want to calculate an overall $F1$ score. We can use the weighted-average method. The weighted $F1$ score will be equal to 97.00%.

The classification performance of people and car classes was satisfactory. Values superior to 0.96, for the recall and precision of the "people" and "car" classes indicate that the images belonging to the "people" and "car" classes are correctly classified. However, a confusion was observed in the "bicycle" class, where many examples are misclassified as "car" or "people" samples, leading to low values of precision and recall for the class "bicycle". The worst performance is seen in the case of the "dog" class which can be explained by the reduced number of samples. This result can be improved by training the model with a larger number of samples containing representative images.

To understand more the confusion between the different classes, we have visualized some misclassified test samples, as shown in Fig. 9. The first and the second rows show some bicycle samples confused with people and car, respectively. As shown in the last two lines, we can obviously see that it is difficult to classify some samples correctly, even by the human eye. These examples are completely unrecognizable or blurred due to low-image resolution or noise. Therefore, the exclusion of these unrecognizable examples may be a reasonable strategy to improve the performance of our algorithm because it is not meaningful to require an algorithm to correctly classify unrecognizable examples.

The pre-trained VGG and ResNet models' performance confirms the successful transfer learning, i.e. the quality of the inferior layers' features. In addition, the architectures of [29] have been performing as well as the fully convolutional models (VGG16, ResNet50). So, we can infer that the incorporation of attentional forms in the superior CNN layers, clearly improves the performance, especially when the feature maps of the lower layers are well trained. Indeed, the last layers of a deep network generally have a global receptive field, so the global embeddings are more directly accessible through the attention mechanism compared to classical CNN which always focus on local and spatial context. Therefore, the use of this global information mainly in the top layers has a significant influence on the models' performances.

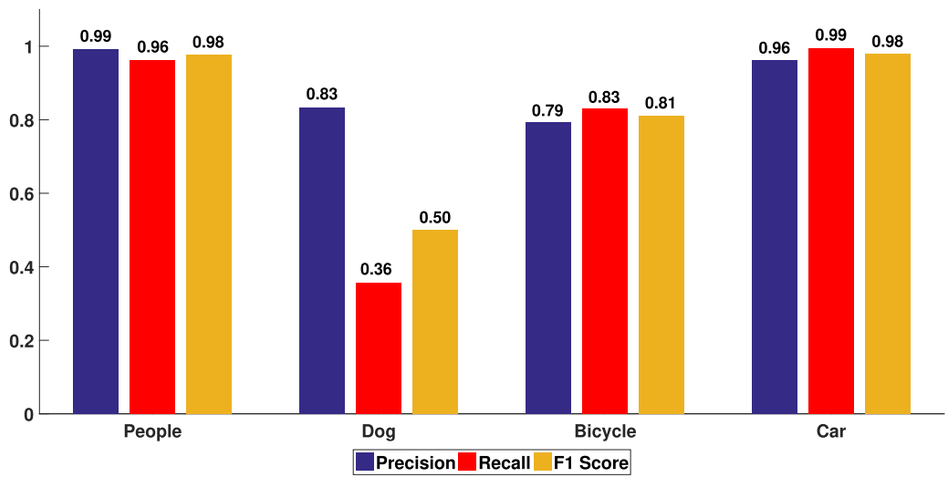


Fig. 7. Precision, recall and $F1$ score values obtained for AResNet50 on "four-class FLIR".

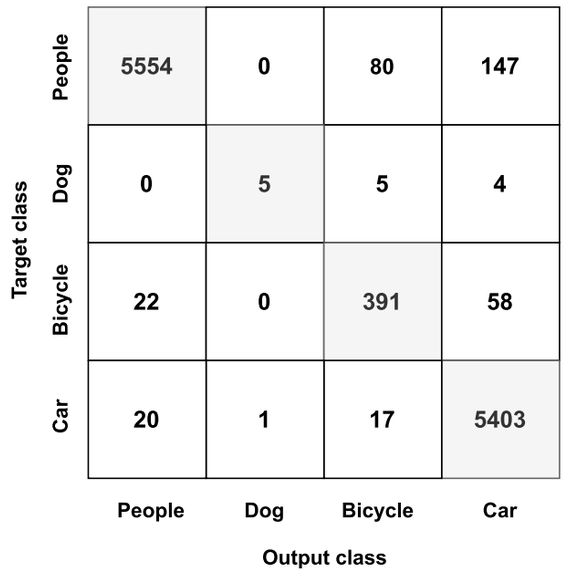


Fig. 8. Confusion matrix obtained for AResNet50 on "four-class FLIR".

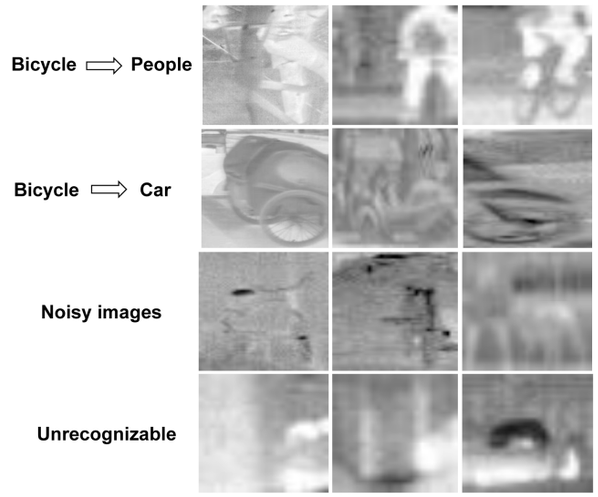


Fig. 9. Some misclassified FLIR test samples using our approach.

### Effect of hyper-parameters $N_h, \alpha, \beta$

Using MHA instead of single head attention allows each attention module to focus only on one set of features, which gives additional encoding power. We can predict that increasing the heads number improves features quality. However, a large number of heads means more memory consumption. Therefore, we test our models in two-class FLIR for a reduced number of heads to see how they affect classification performance.

The performance of augmented models (AVGG16, AResNet50) also depends on the ratio of attentional channels. This ratio is controlled by $\alpha$ and $\beta$. We search $\alpha$ and $\beta$ in $\{0.125, 0.25, 0.5, 1\}$, $\{0.125, 0.25\}$, respectively. Figure 10 shows the performance of the augmented models by varying $\alpha$ and $\beta$.

Figure 10 clearly shows that the model performs better with $\alpha = 0.125$ and $\beta = 0.125$ for all numbers of heads. It can be seen that the lowest performances are obtained for $N_h = 1$, this is explained by the fact that an additional head allows the model to expect more information and encodes multiple relationships between the features. However, increasing the number of heads does not necessarily mean a performance improvement. These are related to the three parameters $N_h, \alpha, \beta$. Therefore, to get the best performance, we have to look for the optimal triplet $(N_h, \alpha, \beta)$.

Michel *et al.* [36] confirmed this conclusion through several experiments. They have shown that in some cases a high number of heads are not necessarily better than a single head. They found that it is possible to remove a large percentage of heads at test time because they are redundant. Their absence does not affect the performance and in some cases even increases it.

The performance of AResNet50 degrades slightly as $\alpha$ and $\beta$ increase. This may be due to the downsampling applied after the first MHA block (Fig. 4).

### Effect of relative position encodings

We additionally perform experiments to see the efficiency of position encodings. Two scenarios are considered: (1) None, i.e. the position encodings are removed. (2) RPE. Table 3 presents the obtained comparison results.

As expected, experiments demonstrate that position embeddings provide improvement. There is a gap (2% in average) between the performances of the model with no position embeddings and models with position embeddings. This is attributed to the position embeddings that allow to take into consideration the distances between the features in different locations, i.e. an efficient association of the objects' information with the consideration of their positions.

### Comparison with state of the art

We have compared our approach with the state of the art on two-class FLIR dataset, in terms of classification performance. Table 4 shows the comparison results for two-class FLIR dataset.

The above results show that our model performs well and is very efficient for target recognition in IR images. It outperforms other works with a rate of 3.5%. For four-class FLIR, no previous classification work is being carried out.

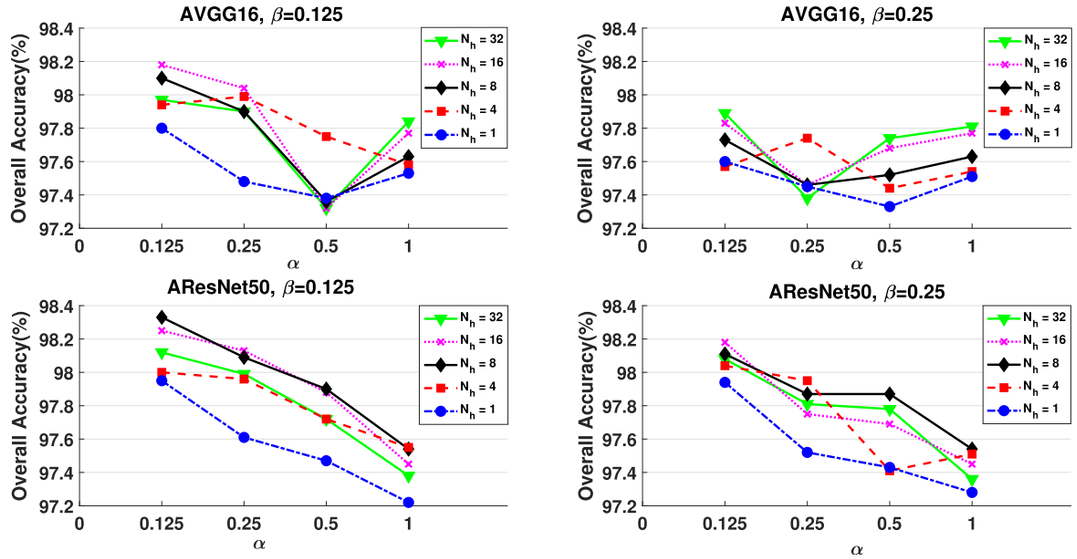


Fig. 10. Results of the variations on $\alpha$, $\beta$ and $N_h$ values.

Table 3. Results of the ablation study on position embeddings.

| Architecture | PE | Acc (two classes) | Acc (four classes) |
|---|---|---|---|
| AVGG16 | None | 96.21% | 93.06% |
| | RPE | 98.10% | 96.71% |
| AResNet50 | None | 96.44% | 93.59% |
| | RPE | 98.33% | 97.07% |

Table 4. The performance comparison of our approach with the state of the arts on FLIR (two classes) dataset.

| Two-class FLIR | |
|---|---|
| Approaches | Test accuracy (%) |
| MSER [35,37] | 80.0 |
| WingerMSER [35] | 84.6 |
| AlexNet_conv4 [38] | 88.8 |
| VGG19_conv5_3 [38] | 94.1 |
| HE_SIFT [39] | 94.8 |
| The proposed approach | 98.3 |

As mentioned before, we have achieved 97.07% overall classification accuracy.

## 5. Conclusion

In this work, we have presented two hybrid architectures AVGG16 and AResNet50 to improve the representational power of deep networks in target recognition in IR images. We have introduced a convolutional–attentional form in the last stage of VGG16 and ResNet50. The attention mechanism located at the top layers allows the proposed networks to capture more specific and global features. Extensive experiments on FLIR starter thermal dataset have shown that the proposed architectures have improved the classification performance over standard models [4,5], ViT [14] and augmented models [29].

We believe that improving CNN requires going deeper. However, our results suggest that incorporating attentional forms with CNN layers is preferable to simply make networks deeper, which supports our hypothesis that self-attention captures long-range dependencies better than stacking CNN layers. In future work, we want to integrate the self-attention mechanism at lower levels to evaluate its performance in the representation of general aspects.

## References

[1] S. Khan, M. Tufail, M. T. Khan, Z. A. Khan, J. Iqbal and A. Wasim, A novel framework for multiple ground target detection, recognition and inspection in precision agriculture applications using a UAV, *Unmanned Syst.* **10**(1) (2022) 45–56.

[2] K. Feng, W. Li, J. Han and F. Pan, Low-latency aerial images object detection for UAV, *Unmanned Syst.* **10**(01) (2022) 57–67.

[3] L. Dai, Y. Zhu, G. Luo, C. He and H. Lin, A real-time visual tracking approach using sparse autoencoder and extreme learning machine, *Unmanned Syst.* **3**(4) (2015) 267–275.

[4] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *3rd Int. Conf. Learning Representations, Conference Track Proceedings*, San Diego, CA, USA, 2015.

[5] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA (IEEE, 2016), pp. 770–778.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA (IEEE, 2015), pp. 1–9.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, California, USA (Curran Associates, 2017), pp. 6000–6010.

[8] P. Shaw, J. Uszkoreit and A. Vaswani, Self-attention with relative position representations, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana (Association for Computational Linguistics, 2018), pp. 464–468.

[9] J.-B. Cordonnier, A. Loukas and M. Jaggi, On the relationship between self-attention and convolutional layers, *Int. Conf. Learning Representations*, Addis Ababa, Ethiopia, 2020.

[10] Y. Chen, Y. Kalantidis, J. Li, S. Yan and J. Feng, A2-Nets: Double attention networks, in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montreal, Canada (Curran Associates, 2018), pp. 350–359.

[11] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku and D. Tran, Image transformer, in *Proc. 35th Int. Conf. Machine Learning*, Stockholmsmässan, Stockholm, Sweden (PMLR, 2018), pp. 4055–4064.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, End-to-end object detection with transformers, *16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 213–229.

[13] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, *9th Int. Conf. Learning Representations*, Virtual Event, Vienna, Austria, 2021.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, *9th Int. Conf. Learning Representations*, Virtual Event, Vienna, Austria, 2021.

[15] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya and J. Shlens, Stand-alone self-attention in vision models, in *Advances in Neural Information Processing Systems 32: Annual Conf. Neural Information Processing Systems*, Vancouver, BC, Canada (Curran Associates, 2019), pp. 68–80.

[16] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille and L.-C. Chen, Axial-deeplab: Stand-alone axial-attention for panoptic segmentation, *16th European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 108–126.

[17] H. Zhao, J. Jia and V. Koltun, Exploring self-attention for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Virtual Event, Seattle, WA, USA (IEEE, 2020), pp. 10076–10085.

[18] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena, Self-attention generative adversarial networks, in *Proc. 36th Int. Conf. Machine Learning*, Long Beach, California, USA (PMLR, 2019), pp. 7354–7363.

[19] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang and X. Tang, Residual attention network for image classification, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Honolulu, HI, USA (IEEE, 2017), pp. 6450–6458.

[20] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan and L. Zhang, CvT: Introducing convolutions to vision transformers, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Virtual Event, Montreal, QC, Canada (IEEE, 2021), pp. 22–31.

[21] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, CCNet: Criss-cross attention for semantic segmentation, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Seoul, Korea (South) (IEEE, 2019), pp. 603–612.

[22] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li and H. Shi, Escaping the big data paradigm with compact transformers, preprint (2021), arXiv:2104.05704.

[23] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy and T. Kipf, Object-centric learning with slot attention, in *Advances in Neural Information Processing Systems 33: Annual Conf. Neural Information Processing Systems*, Virtual Event (Curran Associates, 2020), pp. 11525–11538.

[24] J. Hu, L. Shen and G. Sun, Squeeze-and-excitation networks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA (IEEE, 2018), pp. 7132–7141.

[25] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, GCNet: Non-local networks meet squeeze-excitation networks and beyond, in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops*, Seoul, Korea (South) (IEEE, 2019), pp. 1971–1980.

[26] J. Hu, L. Shen, S. Albanie, G. Sun and A. Vedaldi, Gather-excite: Exploiting feature context in convolutional neural networks, in *Advances in Neural Information Processing Systems 31: Annual Conf. Neural Information Processing Systems*, Montreal, Canada (Curran Associates, 2018), pp. 9423–9433.

[27] J. Park, S. Woo, J.-Y. Lee and I.-S. Kweon, BAM: Bottleneck attention module, *British Machine Vision Conf.*, Newcastle, UK, 2018, p. 147.

[28] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, CBAM: Convolutional block attention module, *15th European Conf. Computer Vision*, Munich, Germany, 2018, pp. 3–19.

[29] I. Bello, B. Zoph, Q. Le, A. Vaswani and J. Shlens, Attention augmented convolutional networks, in *Proc. IEEE/CVF Int. Conf. Computer Vision*, Seoul, Korea (South) (IEEE, 2019), pp. 3285–3294.

[30] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel and A. Vaswani, Bottleneck transformers for visual recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Virtual Event (IEEE, 2021), pp. 16519–16529.

[31] Flir thermal starter dataset (2018), https://www.flir.com/oem/adas/adas-dataset-form/.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, USA (IEEE, 2009), pp. 248–255.

[33] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, Attention-based models for speech recognition, in *Proc. 28th Int. Conf. Neural Information Processing Systems - Volume 1*, Montreal, Canada (Curran Associates, 2015), pp. 577–585.

[34] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Proc. 32nd Int. Conf. Machine Learning*, Lille, France (PMLR, 2015), pp. 448–456.

[35] A. Akula, R. Ghosh, S. Kumar and H. K. Sardana, WignerMSER: Pseudo-Wigner distribution enriched MSER feature detector for object recognition in thermal infrared images, *IEEE Sens. J.* **19**(11) (2019) 4221–4228.

[36] P. Michel, O. Levy and G. Neubig, Are sixteen heads really better than one? in *Advances in Neural Information Processing Systems 32: Annual Conf. Neural Information Processing Systems*, Vancouver, BC, Canada (Curran Associates, 2019), pp. 14014–14024.

[37] N. V. A. Akula, R. Ghosh, N. Guleria, S. Kumar and H. K. Sardana, Towards an optimal bag-of-features representation for vehicle type classification in thermal infrared imagery, in *Optics and Photonics for Information Processing XII* (SPIE, 2018), pp. 191–207.

[38] A. Akula and H. K. Sardana, Deep CNN-based feature extractor for target recognition in thermal images, in *IEEE Region 10 Conf. (TENCON)*, Kochi, India (IEEE, 2019), pp. 2370–2375.

[39] B. Nebili, A. Khellal and A. Nemra, Histogram encoding of SIFT based visual words for target recognition in infrared images, in *Int. Conf. Recent Advances in Mathematics and Informatics*, Tebessa, Algeria (IEEE, 2021), pp. 1–6.