**Gottfried Wilhelm Leibniz Universität Hannover**

Institut für Photogrammetrie und GeoInformation

# Thermal anomaly detection based on co-saliency analysis of optical and thermal images via deep learning

In the course Geodesy and Geoinformation

Master Thesis

of

Gurpreet Singh

**Examiner:**

Prof. Dr.-Ing. habil. Christian Heipke

**Supervisor:**

M.Sc Artuom Sledz

Hannover, October 2022

# Declaration

I declare that this Master Thesis is the result of independent research conducted by me under the guidance of my supervisor. It does not contain the results of any other scientific research that has been published or written by any other individuals or groups, except for those already cited in the thesis. Furthermore, I state that this work in the same or a similar form has not been submitted to an examination authority.

Hannover, 24-10-2022

Gurpreet Singh
(10017996)

# Abstract

In remote sensing, thermal anomaly detection plays a crucial role. In this work, thermal anomaly detection is formulated as a salient region detection, which is motivated by the assumption that a hot region often attracts attention of the human eye in thermal infrared images(TIR). Using TIR and optical images(RGB) together, our working hypothesis is defined in the following manner: a hot region that appears as a salient region only in the TIR image and not in the optical image is a thermal anomaly. By utilizing co-saliency technique, RGB images are used to reduce the number of false alarms that may occur when detecting thermal anomalies from TIR images alone. Recently, deep learning have been widely adopted to tackle this task. Initially in this thesis, the anomaly detection is done by adopting a Multi Interactive Dual Decoder(MIDD) approach and the problem is framed as a multi-class problem. In context, MIDD approach uses separate decoders and encoders for both modalities(RGB and TIR) to obtain the final saliency map. Subsequently, a simplified MIDD approach is presented to handle the same scenario, based on dual encoder and single fusion decoder. Moreover, a new dataset is generated established from orthomosics of RGB and TIR images. Further, obtained results are evaluated both on pixel level and object level using various metrics. Despite some limitations outlined in the thesis, the proposed method to identify the thermal anomaly has achieved up to 90 percent of recall for the large objects.

**Keywords:** thermal anomaly, saliency map, deep learning, encoder-decoder

# Contents

# List of Figures

# List of Tables

# 1 Introduction

During recent decades, as a reliable and economical way to transport energy, pipeline networks have become increasingly popular in recent decades. In light of the current energy requirements and environmental impacts, it becomes imperative to ensure energy resources are secured in an efficient way. Across the globe, the rising gas prices threaten livelihoods and social stability. In Germany, for example, the gas prices have been doubled in just a few months due to the Russia-Ukraine conflict. Meanwhile, poorly insulated roofs or leaks in the heating pipeline such as small cracks or pinholes can go unnoticed for long periods of time, causing irreversible environmental damage and energy waste. Therefore, ensuring the functioning of these pipelines is imperative to avert excessive financial losses due to the interruption of heating supply and, most importantly, to eliminate any potential threat to human lives and the ensuing detrimental aftermath on the environment.

Several conventional approaches were proposed to tackle with such problems. More recently, following the fourth Industrial revolution, Machine Learning data-driven approaches have gained popularity due to their high accuracy compared to other conventional methods and their efficient implementation due to recent advancements in tensor multiplication dedicated GPUs. In this vein, Deep Learning is widely employed to perform anomaly detection, it is the process of identifying conspicuous components, events, or observations that raise concerns because these elements differ significantly from the majority of data or expected behavior. Subsequently, the purpose of thermal anomaly detection is to localize the unusually temperatures those differ from their surroundings. In fact, thermal anomalies could be hot or cold anomalies depending upon high or low distinctiveness of the temperature than it's surroundings. In remote sensing, thermal anomaly detection plays a crucial role to capture abnormal heat signatures in a non-destructive manner. Due to its ability to capture heat signatures in the infrared portion of the electromagnetic spectrum, a Thermal Infrared (TIR) camera is one of the most extensively used equipment for this purpose. Thermal images can deal with complex conditions such as insufficient illumination, low contrast, noise as well as can also help to detect occluded heat signatures such as underground leakages. This technology have been widely used in many applications such as local surveillance, fire detection, and human inspections.

The primary goal of saliency analysis is to determine how distinct a certain region in an image is in relation to its surroundings. Initial saliency models were developed using image processing techniques, while current advances in deep learning approaches are showing great promise in this field. The scientific community has demonstrated that it is feasible to merge the complimentary information of visible and thermal modalities in order to correctly capture salient regions, in a method known as RGB-T saliency detection. Current developments in RGB-T saliency detection are mainly exploring the detection of common conspicuous objects in both modalities. However, this thesis will be focused on thermal anomaly detection based on co-saliency analysis.

In this study, thermal anomaly detection is formulated as a salient region detection, which is motivated by the assumption that a hot region often attracts attention of the human eye in thermal infrared images. Using TIR and RGB together, the working hypothesis is defined in the following manner: a hot region that appears as a salient region only in the TIR image and not in the RGB image is a thermal anomaly. The three channel RGB images contain more spatial information compared to of thermal images as RGB camera operates in the visible region (0.4-0.7 um). Colors and textures extracted from RGB images can be applied to identify individual objects which is not possible with TIR images, because TIR operates operates in the infra-red band, specifically in this work, the used dataset was acquired with a Long Wave Infra-Red (LWIR) band (7-14 um) camera. LWIR band enables the camera to capture heat signatures from the surface. Integrating RGB and TIR data have several advantages, as every hot item has the potential to be identified as a thermal anomaly, the purpose of utilising RGB images is to reduce on the amount of false alarms that occur throughout the process of detecting thermal anomalies from TIR images only. RGB images are quite helpful in this situation since they may make it possible to differentiate between hot items and other thermal abnormalities. On the other side, a cold object has the potential to induce false alarms as well, which manifest themselves as a high gradient on its borders. The used dataset is established from the orthomosic of RGB and TIR images, this dataset includes RGB and TIR images, as well as ground truth (GT). GT consists four classes: Background(BG), Thermal Anomalies(AN), Hot Objects(HO) and Cold objects(CO).

## 1.1 Thesis Objective

The main hypothesis proposed here is the detection of thermal anomalies, which are salient only in thermal images, but not in RGB images. In contrast, objects which are visible in both images should also be salient and these salient objects either could be a hot object or it could be cold object, i.e. objects that are salient in both modalities can not be thermal anomalies. In this way, the fusion of RGB-T modalities lead to reduction in false alarm rate.

Thesis goals are threefold as, examination of several RGB-T saliency detection methods in order to decide which one will be more adaptable to the main hypothesis of this study; instead of binary ground truth datasets, a multiclass ground truth dataset is generated from thermal and RGB orthomosaics, as SOD tasks only deal with foreground vs background classification but in this work a multiclass(BG, AN, HO and CO) problem is considered; deep learning model implementation using existing frameworks for deep learning such as Keras or Tensorflow. The evaluation criteria will be set by the commonly known methods in machine learning field, such as accuracy, precision and F1-score as well.

## 1.2 Thesis Contribution

Considering aforementioned problems, contributions of this work are as follow,

- Initially for the multi-class adaptation multi-interactive dual decoder (MIDD) proposed by Tu et al. (2021) (Chapter 4) is utilized. Instead of capturing common salient objects, MIDD network is adopted to detect thermal anomalies, which are not visible in RGB images as well as the hot and cold objects. The original MIDD model was proposed to detect common salient objects visible in both RGB-T pairs, i.e. they had considered the whole problem as a binary classification(background vs foreground).

- Secondly, a modified version of MIDD is purposed (Chapter 6). Instead of dual-decoders, a single multi-interactive decoder is inaugurate to speed-up and decrease the unnecessary parameters of the decoding network. Even the comparison of the evaluation results show that proposed network achieves slightly better results than MIDD, as well as reduced the size of the network.

- The evaluation for both the networks is carried out on both pixel and region level, as well as the comparison of the MIDD and proposed network is demonstrated(Chapter 6.4)).

## 1.3 Thesis Outline

Altogether, this work is organized as follow,

- **Chapter 2** will present current and state of the art methods for Saliency Object

Detection(SOD), including RGB-T fusion based saliency detection approaches.

- **Chapter 3** provides the theoretical background information about the specialized topic related to deep learning, such as loss function, optimization algorithms, Convolutional Neural networks(CNN) etc.

- **Chapter 4** is the detailed discussion of the MIDD network, comprising all the components of this network as well as the adopted loss function for the multi-class classification.

- **Chapter 5** will provide the information about the experimental setup, including the information from the used dataset to training and from training to evaluation of the results.

- **Chapter 6** demonstrates the purposed version of MIDD network, and also presents the outcomes this network in respect to the original MIDD network.

- **Chapter 7** is the final part of this thesis, which sums up whole work and also enlighten all the successes and failures of the outcomes. At the end, it will discuss the future possible future developments, which should deal with the disadvantages discussed earlier.

# 2 Literature Survey

Saliency model analyses is the distinctiveness of image regions with respect to their local neighbourhood (Borji et al. (2015)). Over the past decades, different theories and methods have been proposed for describing and creating saliency maps. Traditional salient object detection methods used low-level hand-crafted features such as color, contrast, and object prior. In the 1990s, Itti et al. (1998) developed the first saliency computational model, by implementing local centre-surround operations across multiscale image features. As this method was based on re-scaling images few times, which probably turns into the loss of frequency content. Further Achanta et al. (2008) proposed a salient detection method using low-level features of luminance and color, that defines pixel saliency, based on the color differences from the average color of the whole image. Subsequently, Achanta et al. (2009), compares five state-of-the-art salient region detection methods proposed by Achanta et al. (2008); Harel et al. (2006); Hou and Zhang (2007); Itti et al. (1998), and Ma and Zhang (2003) from a frequency domain perspective. In addition to low level features several approaches were proposed by integrating higher-level prior knowledge, such as Yang et al. (2013) instead of considering the contrast between the salient objects and their surrounding regions, they consider both foreground and background cues on the bases of rank of the similarity of the image elements (pixels or regions) with foreground cues or background cues via graph-based manifold ranking and Shen and Wu (2012) used as the center or semantic prior, for detecting salient objects.

Stemming from the conventional methods, incorporation of deep neural networks took saliency object detection to next level and deep learning based methods have yielded a qualitative leap in performances as compared to state of art models. As in 2015, Kümmerer et al. (2014) proposed DeepGaze-I , one of the earliest deep learning based saliency detection approach, it outperforms state-of-the-art models, by increasing the amount of "information gain explained" to 56% compared to 34% for state of the art models. "Information gain explained" relates the model's information gain to the gold standard information gain(Kümmerer et al. (2016)), here information gain is the information difference between baseline and image based saliency-model. DeepGaze-I used the well known deep network of Krizhevsky et al. (2012) to generate a high-dimensional feature space. The success of convolutional neural networks (CNN), has brought along a revolution for saliency models and the focus was shifted to CNN than handcrafted features. To further improve the performace, SALICON proposed by Huang et al. (2015) and DeepFix proposed by Kruthiventi et al. (2017) used much deep networks instead of using 5 layer network as Used in DeepGaze-I. DeepFix used VGG-16 as a backbone network, while SALICON was proposed with three different networks, AlexNet, VGG-16, and GoogLeNet networks, even these networks integrated transfer learning for further advancement in saliency detection. Subsequently, following the achievement of DeepGaze-I, Kümmerer et al. (2016) proposed DeepGaze-II and also DeepGaze-IIE(Linardos et al. (2021)). DeepGaze-II used the features from the VGG-19 deep neural network, trained to identify objects in images.Further by replacing the VGG19 with ResNet50 in DeepGaze II, it improve the performance on saliency prediction to 5%. On the other hand, DeepGaze-IIE proposed by combining multiple backbones in a principled manner, increased the information gain explained to 93%. On the other hand, Borji (2019) explores the landscape of the field emphasizing on new deep saliency models, benchmarks, and datasets. Additionaly, this paper addressed various question such as, in what ways current models fail, how to remedy them, what can be learned from cognitive studies of attention, how explicit saliency judgments relate to fixations, how to conduct fair model comparison, and what are the emerging applications of saliency models (Borji (2019)). Despite significant progress, visual saliency detection using visible spectrum camera remains a very challenging task in some complex scenarios, such as low illumination, background clutters, as well as bad weathers (rain, haze, smog, etc.). Li et al. (2018) proposed one of the earliest approaches using fusion of two modalities for saliency detection in the context of graph learning problem, but this approach was not time efficient. Further advancement was made by Zhang et al. (2020) using a intermediate module ADAC (Adjacent Depth Feature Combination) to integrate the multi-level features of single-modal images. While, Xu et al. (2022) proposed the first two-stream encoder-decoder network was designed to fuse the multilevel feature and as the low-level features contain more detailed saliency cues, and they gradually fade in the encoding process and to mitigate this fading Global Attention(GA) module was introduced. In the same stream, Tu et al. (2021) introduced multi interactive dual-decoder for the saliency detection which is discussed in details in Chapter 4 and in this work same network approach is adopted for the thermal anomaly detection. Specifically, in the context of thermal anomaly detection, Zhong et al. (2019) present a saliency-based District Heating System(DHS) leakage detection approach, an infrared saliency map is created to enhance the leakage targets, while the pipeline location integrated from a Geographic Information System (GIS). In 2020, Sledz et al. (2020) presented thermal anomaly detection for District Heating System(DHS) using image analysis techniques, in which they utilized TIR images for anomaly detection and localize their positions using Geographic Information System. Further, Sledz and Heipke (2021) proposed an update version of anomaly detection approach using multi-modal image sources, in which thermal anomaly detection was handled using information fusion of saliency maps derived from both, TIR and optical images. To the best of my knowledge, besides my work nobody has focused on using RGB-T saliency as a tool for thermal anomaly detection.

# 3 Theoretical Background

In this study, deep learning-based techniques are adopted for achieving the goal of thermal anomalies detecting and it is the recent adaptation in the context of Saliency Object Detetction(SOD). Deep learning includes various types of networks such as Recursive Neural Networks (RvNN), Recurrent Neural Network (RNN), and Convoulutional Neural Networks(CNN) etc. The structure of neural network was inspired by neural connections in human and animal brains. The idea of such networks was evaluated from a perceptron(Rosenblatt (1963)) a or single neuron(shown in Figure 2), which is a basic unit of computation in a neural network.

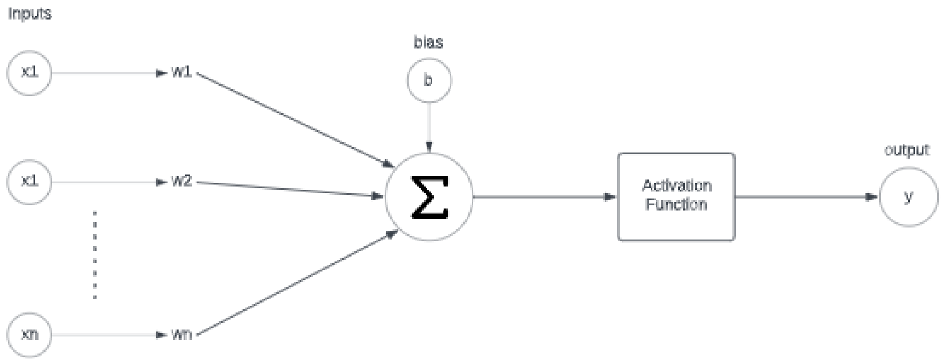


Figure 2: Basic model of Neurons

However, perceptrons are linear classifier and are limited to solving two-class problems. For more difficult problems, Multi-Layer Perceptrons (MLP)(Werbos (1974)) are used, which consist of distinct layers of perceptrons. There are mainly three types of layers: a single input layer, any number of hidden layers and a single output layer. MLP uses backpropagation as a supervised learning technique, where MLP tries to model the correlation between the input data and the ground truth by adjusting the weights and biases, by minimizing the error. Initially, this chapter will enlighten the basic concepts of neural networks and further it will explore some more relevant terms of computational neural networks(CNN) training such as loss functions, Optimization algorithms and Regularization.

## 3.1 Activation Function

The activation function means to producing the input to non-linear output, it gives the ability to the network to learn extra-complicated things. There are numerous of activation functions available and the most commonly implemented activation functions are,

**Sigmoid** receives real numbers as inputs, and its output ranges from zeros to ones. Mathematically,

$$f(x)_{sigmoid} = \frac{1}{1+e^{-x}} \tag{1}$$

In the classification task, sigmoid outputs can be treated as normalized probabilistic interpretations,but it causes the gradient to become zero when it saturates at 0 or 1 and results into the gradient vanishing problem.

**Rectified Linear Unit (ReLU)**, simply converts the whole values of the input to positive numbers and it lower computational load of the network. Occasionally, a few significant issues may occur during the use of ReLU.

$$f(x)_{ReLU} = max(0, x) \tag{2}$$

**Leaky ReLU** is an alternative for ReLU, it down-scale the negative inputs, this activation function ensures these inputs are never ignored. Leaky ReLU can be represented mathematically as,

$$f(x)_{LeakyReLU} = \begin{cases} x & \text{if x>0} \\ \alpha x & \text{if x}\leq\text{0} \end{cases} \tag{3}$$

where $\alpha$ is a constant value$(0 < \alpha < 1)$, which is used to down-scale the negative inputs.

## 3.2 Loss Function

An error in prediction is computed by means of a loss function, which measures the deviation between the predicted output and the ground truth. The main aim is to obtain a set of parameters that minimize the difference between the prediction and the training dataset. There are various loss functions are available, but it is important to choose a loss function respective to the problem following an activation function. For example, In regression problem, where it is required to predict the value of a variable, the output layer has only one node and a linear activation is applied in the output, usually mean squared error(MSE) is implemented in this case. Let $y$ are the actual values and $\hat{y}$ are the predicted, and mathematically MSE can be given as,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{4}$$

On the other hand, In a binary classification problem, the output layer has one node and a sigmoid activation function(equation 1) is used in the output layer. In this case, the binary cross entropy(log loss) or logarithmic loss function is most suitable. Binary cross entropy(BCE) loss Jadon (2020) can be given as,

$$L_{BSC}(y, \hat{y}) = -ylog(\hat{y}) + (1-y)log(1-\hat{y}) \tag{5}$$

In a multiple-class classification problem, the output layer has nodes equal to the number of classes and the activation function used in the output layer is softmax and cross-entropy loss can be utilized. Other than classes, the type of training data should also be considered, whethere it is a balanced dataset or not. In these cases, some specific loss needs to be utilized such as weighted cross entropy, Dice loss(DL), or focal loss etc. Mathematically, dice loss can be give as,

$$DL = -\frac{2y\hat{y} + 1}{y + \hat{y} + 1} \tag{6}$$

where, 1 is added in numerator and denominator to ensure that the function is not undefined in edge case scenarios such as when y = $\hat{y}$ = 0.

## 3.3 Optimization Algorithm

An optimizer is a function or an algorithm that modifies the attributes of the neural network, such as weights and learning rate. Thus, it helps in reducing the overall loss and improve the performance. The main task of the neural network is to map a set of inputs to a set of outputs by iteratively adjusting the parameters. It is not possible to find the perfect set of parameters, since in neural networks there are usually millions of parameters to solve for, therefore the main focus here is to find the best set of parameters by using optimization algorithms. However, choosing the best optimizer depends upon the application.Usually, the gradient-based learning techniques are appear to be the standard selection and the network parameters should always update though all training epochs. The learning rate is defined as the step size of the parameter updating.

To update the parameters Gradient Descent or gradient-based learning algorithm, it needs to compute the objective function gradient (slope) by applying a first-order derivative with respect to the network parameters. Next, the parameter is updated in the reverse direction of the gradient to reduce the error. The parameter updating process is performed though network back-propagation, in which the gradient at every neuron is back-propagated to all neurons in the preceding layer. Backpropagation is the essence of neural network training. It is the method of fine-tuning the weights of a neural network based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows you to reduce error rates and make the model reliable by increasing its generalization. The Backpropagation algorithm in neural network computes the gradient of the loss function for a single weight by the chain rule. As suggested by Alzubaidi et al. (2021) final weight($w_{ij^t}$) with gradient decent can be represented as,

$$w_{ij^t} = w_{ij^{t-1}} - \Delta w_{ij^t}, \tag{7}$$

$$\Delta w_{ij^t} = \eta \frac{\partial E}{\partial w_{ij}} \tag{8}$$

where, the weight in the preceding (t-1) training epoch is denoted $w_{ij^{t-1}}$. Different alternatives of the gradient-based learning algorithm are available and commonly employed, such as Batch Gradient Descent, Mini-batch Gradient Descent, and Stochastic Gradient Descent(SGD). The learning rate is (step size) and the prediction error is E. In stochastic gradient descent, instead of taking the whole dataset for each iteration, it randomly select the batches of data. That means we only take few arbitrary samples from the dataset and the parameters are updated at each training sample in this technique. For a large-sized training dataset, this technique is more memory-effective and faster.

## 3.4 Convolutional Neural Network

Convolutional Neural Network (CNN) is the most famous and commonly employed algorithm in a range of different fields, including computer vision. A commonly used type of CNN, which is similar to the multi-layer perceptron (MLP), consists of numerous convolution layers preceding sub-sampling (pooling) layers, while the ending layers are FC layers. An example of CNN architecture for image classification is illustrated in figure 3.

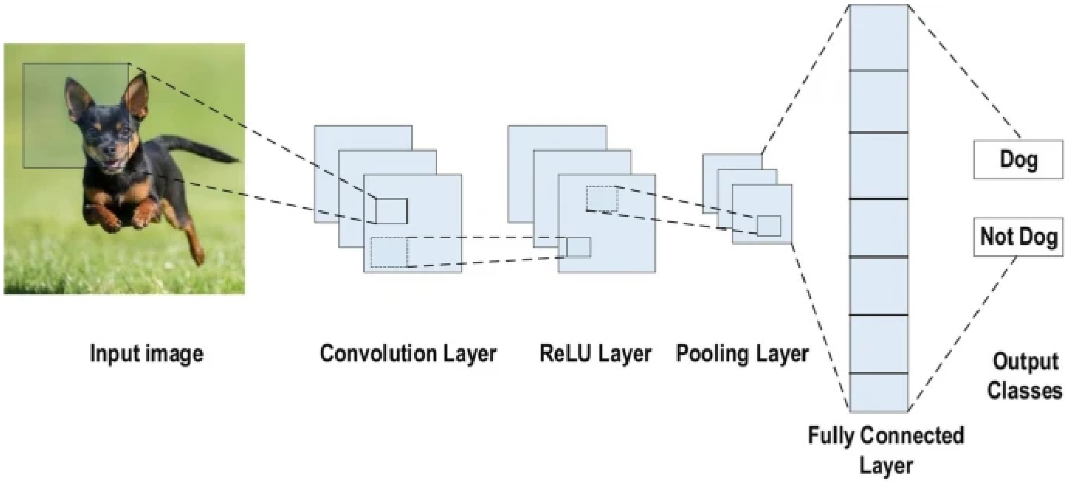


Figure 3: An example of CNN architecture for image classification with different layers (Alzubaidi et al. (2021))

**Convolutional layer** is the most significant component of CNN architecture. It consists of a collection of convolutional filters (kernels). The input image, is convolved with these filters using 2-dimensional convolution to generate the output feature map. Next, in ReLu layer the non-linearity or an activation function implemented to the convolution-layer output. Nonlinear activation functions are preferred as they allow the nodes to learn more complex structures in the data.

**Pooling layer** is primarily responsible for subsampling feature maps. In other words, this approach shrinks large-size feature maps to create smaller feature maps, but throughout the pooling stage, the majority of dominant information is maintained. Several types of pooling methods are available for utilization in various pooling layers such as average pooling, min pooling, max pooling, global average pooling, and global max pooling etc. In Fully Connected (FC) layer each neuron in this layer is connected to all neurons in the previous layer. It is the same as a traditional multilayer perceptron neural network (MLP). The flattened matrix goes through a fully connected layer to classify the images.

## 3.5 VGG16 Architecture

VGG16 is a one of the most commonly used convolutional neural network model proposed by Simonyan and Zisserman (2015). In this paper, VGG(Visual Geometry Group) was proposed with different depths(layers), in which VGG16 refers to 16 layers that have weights. In VGG16 there are thirteen convolutional layers, five Max Pooling layers, and three Dense layers which sum up to 21 layers but it has only sixteen weight layers i.e., learnable parameters layer. The most unique feature about VGG16 is that instead of using a large number of hyperparameters, they used convolution layers of 2x2 filter with stride 1 and padding and maxpool layers of 2x3 filter with stride 2. Throughout the architecture, convolution and max pool layers are consistently arranged.

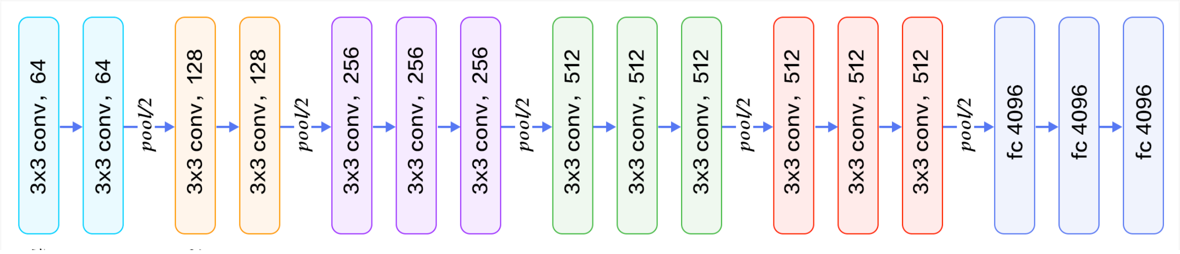


Figure 4: Architecture of VGG16

The overall structure includes 5 sets of convolutional layers, Conv-1 Layer has 64 number of filters, Conv-2 has 128 filters, Conv-3 has 256 filters, Conv 4 and Conv 5 has 512 filters. At the end of convolution layers, three Fully-Connected (FC) layers follow a stack of convolutional layers.

## 3.6 Evaluation Metrics

An evaluation metric is used to quantifies the performance of a predictive model, the most commonly used metrics are described through Chapters 3.6.1 to 3.6.4.

### 3.6.1 Accuracy

Accuracy also known as Rand index or pixel accuracy, is one or even the most known evaluation metric. It is defined as the number of correct predictions, consisting correct positive and negative predictions, compared to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where,

TP: True Positive, TN:True Negative, FP: False Positive, FN:False Negative

### 3.6.2 Intersection over Union(IoU)

The Intersection-over-Union (IoU), also known as Jaccard index or Jaccard similarity coefficient. IoU is essentially a method to quantify the percent overlap between the target mask and our prediction output. In general, the IoU metric measures the number of pixels common between the target and prediction masks divided by the total number of pixels present across both masks.

$$IoU = \frac{GT \cap S}{GT \cup S} \tag{10}$$

Where GT ∩ S represents the intersection between ground truth(GT) and predicted saliency map(S), while GT ∪ S represents the union of both. In terms of pixel-wise predictions IoU canbe expressed as,

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

### 3.6.3 Precision & Recall

The Precision(P) is the fraction of True Positive elements divided by the total number of positively predicted units.

$$P = \frac{TP}{TP + FP} \tag{12}$$

Precision is the indicator of the the quality of the positive predictions made by the model.On the other hand, recall(R)measures how many of the actual positive instances we were able to correctly predict (or recall). Mathematically, recall(R) canbe expressed as,

$$R = \frac{TP}{TP + FN} \tag{13}$$

### 3.6.4 F1 Score

F1 score also know as dice score, can also be used to evaluate the model, it combines both prediction and recall by taking their harmonic mean.

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{14}$$

F1 score gives same weightage to precision and recall, while $F_\beta$(see equation 15) can be used to give additional weightage to recall or precision.If $\beta$ is selected below 1, it will give more weightage to recall and if it is more than 1 it will prioritise recall. The selection of $\beta$ depends on the task. For example, if true positive are more important and we don't want to miss any of them then $\beta$ should be greater than 1, usually for recall prioritization $\beta = 2$ is selected.

$$F_\beta = \frac{(1 + \beta^2).Precision.Recall}{\beta^2.Precision + Recall} \tag{15}$$

# 4 Multi-Interactive Dual-Decoder(MIDD)

Early SOD methods work well in informative visible images, but they can't deal with the images with deficiency, semantic ambiguity or small objects. Basically, SOD tasks are used to differentiate between foreground and background. But Tu et al. (2021) shows that MIDD withstands with different challenges by combining two modalties(RGB and TIR), such as big salient object (BSO), bad weather (BW), cross image boundary (CIB), image clutter (IC), low illumination (LI), multiple salient object (MSO), small salient object (SSO), out of focus (OF), center bias (CB), similar appearance (SA), and thermal crossover (TC).

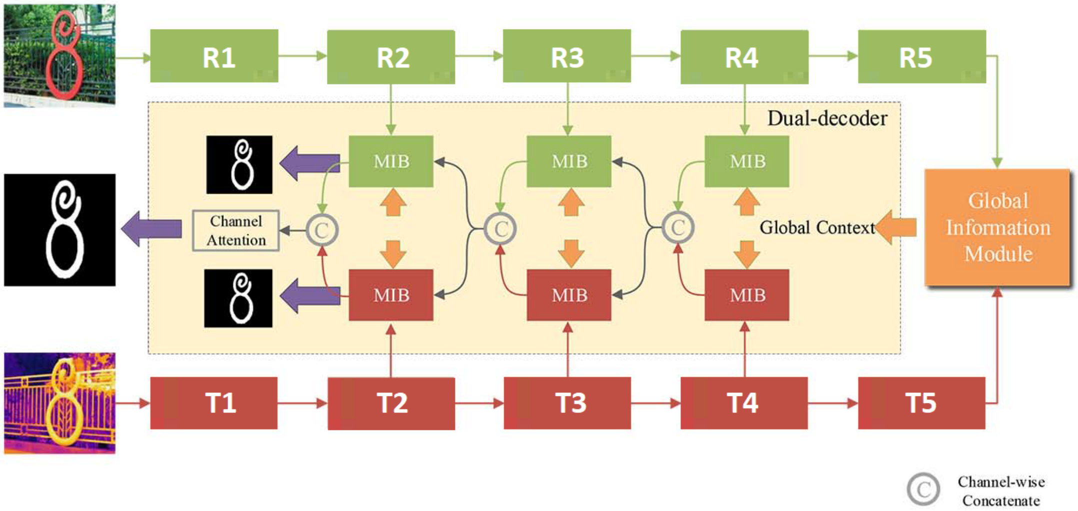


Figure 5: Framework of MIDD Network. RGB- encoders are painted in green and the thermal-decoder are painted in red.(Tu et al. (2021))

Figure 5 shows the MIDD network proposed by Tu et al. (2021). In order to extract features from RGB and thermal infrared images, this network uses two independent encoders. In addition, the global information(GI) module is prosposed to combine the highest-level features from two modalities(RGB and thermal) to find the global features that can be used to locate salient regions accurately using various receptive fields.

To fuse both the modalties a Multi-Interaction Block (MIB) is designed. Following subsections will provide specifics on each component of this network.

## 4.1 Encoder Network

In MIDD, VGG16 (Chapter 3.5) is proposed to use as an encoder, since using the more deep network had not shown any significant improvement in the performance. Further, Tu et al. (2021) had done an extensive research using different encoders and decided to stick to VGG16, because choosing deep networks like $ResNet50$ did not make any noticeable difference. So, to keep the network simple and as suggested by Tu et al. (2021), VGG16 is used as an encoder to extract hierarchical features from the input RGB-T pairs and the last pooling layer and two fully connected layers in VGG16 are removed. The features from the shallowest layer(R1 and T1) of the encoder are also not utilized, because these features contains high spatial information rather than semantic information, which are not conductive for saliency detection. The remaining features T2-T5 and R2-R5 extracted from RGB and TIR images,respectively, are considered for global information and decoder module.

## 4.2 Global Information Module

The Global Information(GI) module extract the coattention feature from the encoders. Global context, output from GI, assists to locate the regions in RGB-T Salient Object Detection(SOD) tasks by combining the highest-level features from two modalities. GI module is the modified version of Convolutional Block Attention Module (CBAM)(Woo et al. (2018)). As shown in figure 6, Channel Attention(CA) mechanism is used for selective recombination of the two features and is applied in the same way as proposed in CBAM(Woo et al. (2018)). while ,the Pyramid Pooling Module (PPM)(Zhao and Wu (2019)) embed into GI module for capturing multiple region contexts, instead of using Spatial Attention Module(SAM) used in CBAM and the multiple perceptrons in CBAM are replaced by 1 × 1 convolution layer.

GI module take top encoded features(R5 and T5) of RGB and TIR modalities as an input and then concatenate them in a channel-wise way. Then CA mechanism is used for combining these features. Mathematically, CA can be written as,

$$CA = \sigma(f_1(AvgPooling(X)) + f_1(MaxPooling(X))) * X \tag{16}$$

Where $f_1$, AvgPooling and MaxPooling are respectively represent 1 × 1 convolution layer, global average pooling and global max pooling. Top encoded features(R5,T5) are channel-wise concatenated and used as input as X. The $\sigma$is sigmoid function that maps the value to the range of 0 to 1.

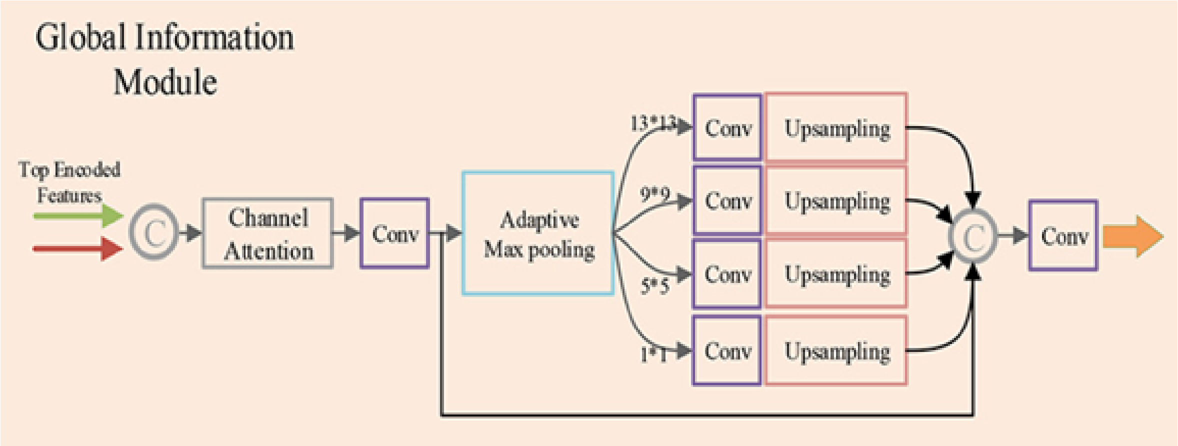
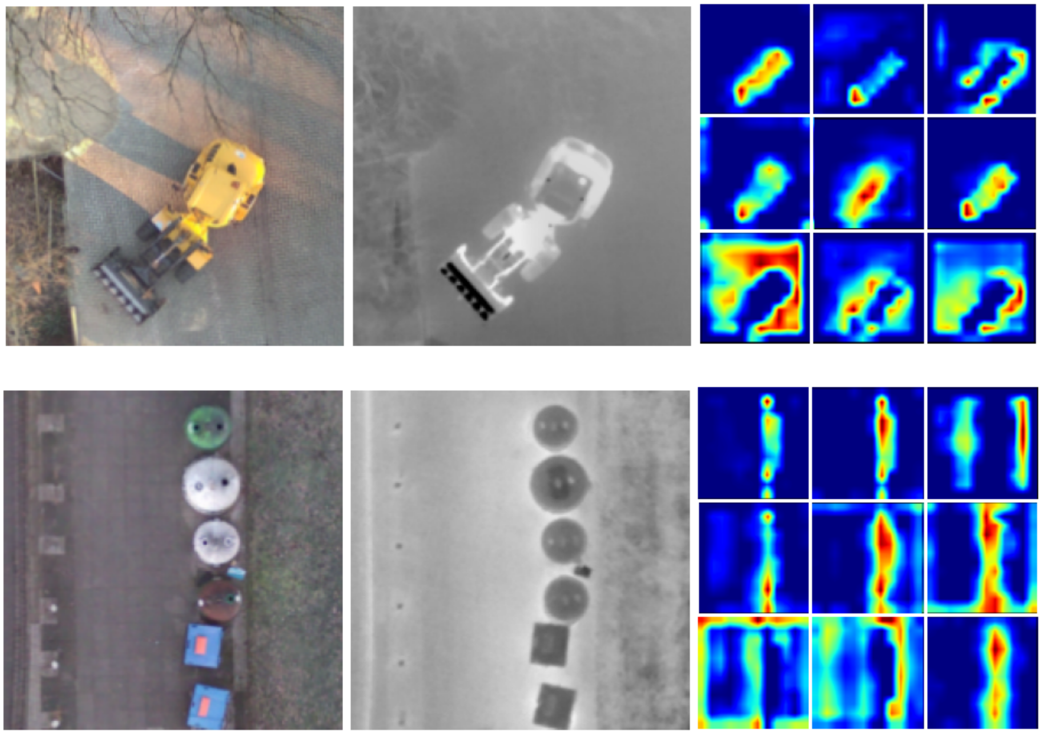


Figure 6: Global Information module(Tu et al. (2021))

The *Conv* block in GI module(figure 6) is the combination of convolution layer, batch normalization (Ioffe and Szegedy (2015)) and Relu (Nair and Hinton (2010)), and adopt this block to decrease the channel number to 256. The output from CA is processed through this Conv block as,

$$F = Conv(CA([R_5, T_5])) \tag{17}$$



(a) RGB image (b) TIR image (c) GI outputs

Figure 7: Visualisation of global context,where column (a) represents rgb image,(b) represents TIR image and (c), represents GI outputs

Further F (equation 17) is used as an input for the PPM part in GI, which is done using four operations of adaptive global max pooling with sizes of n=[1, 5, 9, 13], subsequently, these outputs go through four convolution blocks and up-sampling(UP).

$$F_i' = UP(Conv(MaxPooling_n(F))) \tag{18}$$

$$G = Conv([F_i']) \tag{19}$$

Finally, again a convolutional layer(equation 19) is used to the concatenated features and generate the reconstructed features (G) which contain the information from the global receptive field. Figure 7 visualize the global context, in the column (c), it shows randomly chosen 9 channels out of 512 channels of global context. Visualisation of the global context(Figure 7) shows how it helps to guide the decoder, as in shown outputs it shows various feature, some of them only shows the outline (borders) of the objects, some give only impression to the background, while some of them only highlight the objects. So, for each channel global context gives some specific information to the decoder, which further fused with RGB and TIR modalities using multi-interactive block(MIB).

## 4.3 Dual-Decoder Network

As shown in figure 5, MIDD network uses individual decoder for each RGB and thermal modalities. Multi-Interactive Block(MIB) is designed for getting the interactions between both modalities and embed it into the decoder in a cascade way, which can achieve the interactions of dual modalities, hierarchical features and global context.

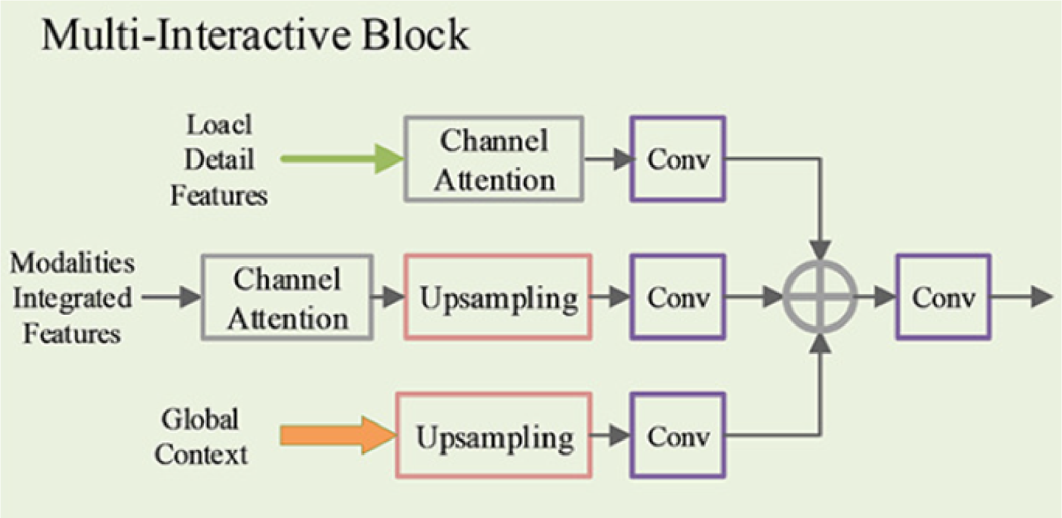


Figure 8: Multi-Interactive Block(MIB)(Tu et al. (2021))

Figure 8 shows the architecture of MIB, and it shows that MIB takes three features(Local detail features, Modalities Integrated features and Global context) as an input. Local detail features(A) are the inputs from the respective RGB and T encoder layer,

$$A_i = Conv(CA(Z_i)), i = 2, 3, 4 \tag{20}$$

Where, $Z_i$ are the local features for $R_i$ and $T_i$, CA is used to emphasize the more useful features, and then decrease the number of channels to 128. Subsequently, Modalities Integrated features(M) are the concatenated outputs($C_{MIB}$) from the previous MIB output from both modalities,

$$M_i = Conv(UP(CA(C_{MIB_i}))), i = 2, 3, 4 \tag{21}$$

Here, up-sampling(UP) is needed to reconstruct the features to match with the size with $A_i$ and adopt a convolutional block to reduce the number of channels of reconstructed features to 128. Last input to MIB is Global context(G),

$$G'_i = Conv(UP(G)), i = 2, 3, 4 \tag{22}$$

Same as last input($M_i$), GI needs to be upsampled to match the size with $A_i$ and convolutional block to decrease the number of channels to 128. At the end, after processing all these three inputs are summed up and reconstructed fused features through a convolutional block as the output of MIB,

$$MIB_i = Conv(A_i + M_i + G'), i = 2, 3, 4 \tag{23}$$

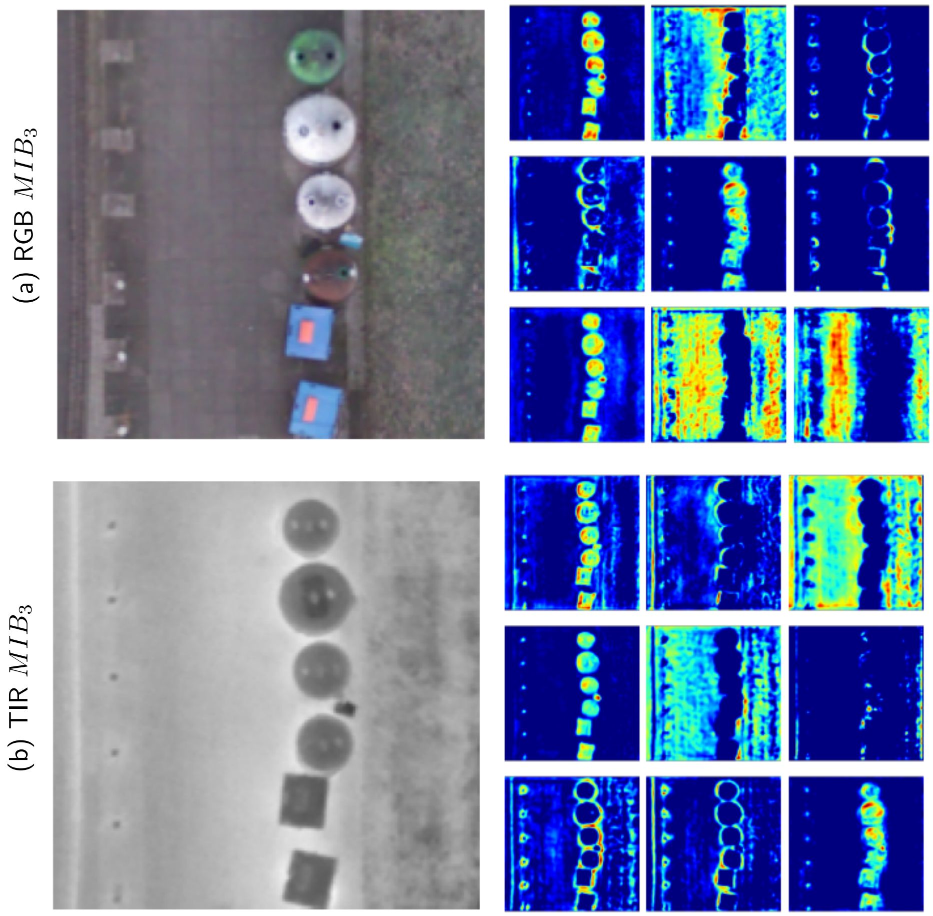


Figure 9: Visualisation of $MIB_3$, where first row represents the output for rgb,and second row represents thermal

Finally, final features are fused from MIBs in dual-decoder by the concatenation and a simple channel-wise attention to predict the final saliency map($S_f$). Figure 9 shows the randomly selected outputs for $MIB_3$(last MIB in the decoder) for individual RGB and TIR outputs. Similar to global context visualization, $MIB_3$ also shows specific information in each channel output, these channel include the visualization of the outline(borders) of the objects, some give only impression to the background, while some of them only highlight the objects. But if we compare visualization of global context(figure 7) and MIB outputs (figure 9), MIB output are much rich in the information as it shows much clear object and background information, because $MIB_3$ fused the information of global context and shallower layers from RGB and TIR encoders.

## 4.4 Loss Function

In MIDD, total loss $L_t$ (equation 29) is the calculated as a combined loss of four losses,final loss $L_f$ (equation 24), global loss $L_g$ (equation 26), decoder loss $L_d$ (equation 25), and smoothing loss $L_s$ (equation 27). The Binary Cross Entropy (BCE) loss is used in this approach for calculating the losses, because in MIDD whole problem is considered as binary classification for differentiating foreground objects with respect to the background. The final loss $L_f$ can be calculated comparing the predicted output(S=$\{S_i$ $|i = 1, ..., T\}$) from the MIDD network and ground truth(GT=$\{GT_i$ $|i = 1, ..., T\}$) and formulated as,

$$L_f = -\sum_{i=1}^{T} (GT_i * log(S_i) + (1 - GT_i) * log(1 - S_i)) \tag{24}$$

While the individual predicted output from each decoder($S_{RGB}$ and $S_T$) are used to calculate the decoder loss as,

$$L_d = BCE(S_{RGB}, GT) + BCE(S_T, GT) \tag{25}$$

To make the global information module be learned better, saliency map $S_g$ from the global context(G) is also predicted. First size of GT is down-sampled to the size of global context(G). Then, a BCE loss is used,

$$L_g = -\sum_{i=1}^{Tg} (GT_{gi} * log(S_{gi}) + (1 - GT_{gi}) * log(1 - S_{gi})) \tag{26}$$

The output(global context) from the from GI, consists 512 channels with reduced size of factor 16 as compared to input size of images, instead of a single channel channel with orignal GT size. So, to find the loss first it reduced to the single channel output($S_g$) using the $1 \times 1$ convolution, on the other hand, the spatial dimensions of the binarised GT($GT_b$) are reduced by the factor of 16 using nearest interpolation to match with the dimensions of ($S_g$). In equation 26, $T_g$ is the number of total pixels of $S_g$. Furthermore, the smoothness loss proposed by Godard et al. (2017) is used as a constraint to achieve region consistency and obtain clearer edges. Smoothness loss is calculated as,

$$L_s = \sum_{i=1}^{T} \sum_{d \in \overleftarrow{x}, \overleftarrow{y}} \psi(|\partial_d S_{f_i}| e^{-\alpha |\partial_d Y_i|}) \tag{27}$$

$$\psi(s) = \sqrt{s^2 + 1e^{-6}} \tag{28}$$

where $\partial_d$ represents the partial derivatives on $\overleftarrow{x}$ and $\overleftarrow{y}$ directions and $\alpha = 10$ as Ye et al. (2021) does. Finally, all these losses are summed up to get the total loss as,

$$L_t = L_f + L_d + L_g + \beta L_s \tag{29}$$

where, $\beta$ empirically set to 0.5 to balance the effect of smoothness, and further details can be found in Tu et al. (2021) loss.

## 4.5 Thermal Anomaly Detection Based on MIDD

In MIDD, the entire problem was viewed as a binary classification (Salient object vs background), but, in this work, Loss function for MIDD is modified to tackle multiclass problem, so, instead of using BCE, Generalized DiceLoss (GDL) proposed by Sudre et al. (2017) is utilized for $L_d$ and $L_f$. As Dice loss is commonly used for unbalanced datasets and the dataset used in this work is also highly unbalanced(discussed in Chapter 5.1). But dice loss can be overly sensitive to incorrect classifications of small objects, i.e., mislabelling a few pixels of a small object would produce a large loss. So, instead of dice loss, Generaliszed Dice Loss (GDL) is consideed is this work. GDL helps to mitigates the limitation of dice loss by considering the weighting strategy for each class.

On the other hand, for global loss ($L_g$) binary dice loss is considered. The reason behind using a binary loss for $L_g$ is that, as Global Information(GI) module helps to guide the decoder in terms of locating the salient information. So, the output from GI, global context gives this salient information irrespective to the class of the region. Due to this reason binary loss is considered for $L_g$. To do so, first global context is reduced to single channel using $1 \times 1$ convolution and then, GT is down-sampled by factor of 16 as compared to input size to match with size of global context(G). This down-sampled GT further reduced to binary class dataset, in which all the objects (from classes AN, HO and CO) represents the foreground, while BG is considered as background. So, the global loss can be calculated as,

$$L_g = DL_2(S_g, GT_b) \tag{30}$$

where ($GT_b$) is binarised ground truth and Binary Diceloss($DL_2$) canbe calculated as,

$$DL_2 = 1 - 2\frac{\sum_{i=1}^{N} y_i \hat{p}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{p}_i + \epsilon} \tag{31}$$

Where y and $\hat{p}$ represent the $GT_b$ and predicted pixels, respectively, and sum runs over all the pixels from 1 to N(total number of pixels). To make it a decreasing function, it is subtracted from 1 and to handle the scenarios as $\hat{p}$=y=0, $\epsilon = 1e{-}14$ is added to both numerator and denominator.

Dual decoders give us two saliency maps for RGB and T decoder, and loss for the decoder canbe calculated as,

$$L_d = \frac{GDL(S_{RGB}, GT) + GDL(S_T, GT)}{2} \tag{32}$$

Generalized Dice loss(GDL) is used as proposed by Sudre et al. (2017),

$$GDL = 1 - 2\frac{\sum_{n=1}^{L} w_n \sum_{i=1}^{N} y_{ni}\hat{p}_{ni} + \epsilon}{\sum_{n=1}^{L} w_n \sum_{i=1}^{N}(y_{ni} + \hat{p}_{ni}) + \epsilon} \tag{33}$$

Where $w_n$ (Equation 34) is the assigned weight to each class(n represents the class number), is calculated using weighting strategies called as Inverse Frequency Weighting as proposed by Sudre et al. (2017). The weights are inversely proportional to the sum of the pixels for each batch of the respective label.

$$w_n = \frac{1}{(\sum_{i=1}^{N} y_{ni})^2} \tag{34}$$

In equation 33, n represents the number of class ($w_n = \{w_n | n = 1, ..., L\}$), in our case L=4,as we have 4 labels(BG,AN, HO and CO). Figure 10 shows, how weights are distributed for each class in each batch for the training process, where median of each class weight shown with an 'orange line', box shows the actual distribution and 'o' shows the outliers. As it can be clearly seen weights for the background(BG) are extremely low as compared to rest of the classes and remained in very skewed range under $10^{-10}$, while weights for AN and HO class ranges between a wide range. As these weights(equation 34) depend upon the presence of each class, if a class have high number of pixels, weights will be low or vice-versa. Outliers are basically small weight values as compared to the median, it shows if a batch have very small amount of pixels of a specific class.

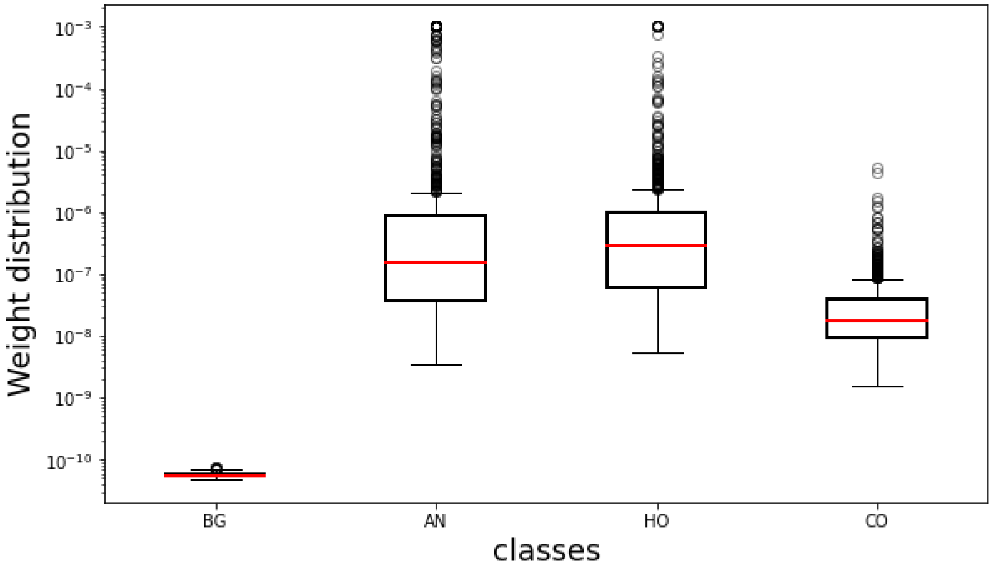


Figure 10: weight($w_n$) distribution for each class

The final saliency loss($S_f$) can be calculated as,

$$L_f = GDL(S_f, GT) \tag{35}$$

where, total loss can be calculated using equations 30, 32 and 35, as,

$$L_{total} = \frac{L_g + L_d + L_f}{3} \tag{36}$$

As compared to the original loss(equation 29) used in MIDD implementation, softness loss(equation 27) is dropped in the loss(equation 36) used in this work. The reason behind this is that dataset used in this work is created from co-registered orthomosaic images(RGB and TIR) with the patch size of 192 × 192, but the co-registration of the RGB and TIR images in orthomosaic may consist some errors, especially around the borders of the objects(further details can be found in Chapter 5.1). So, due to some inaccuracy around the borders of the objects use softness loss may worsen the situation. Moreover, there are only a few approaches are proposed to predict fine object boundaries, as proposed by Xu et al. (2022).

# 5 Experimental Setup

MIDD is implemented based on Pytorch and trained on Google colaboratory with Tesla-16GB GPU, using the stochastic gradient descent (SGD) to optimize parameters with the weight decay of 5e-4 and the momentum of 0.9 and trained it for 50 epochs with batch size of 4. For the inputs, image size of 192 x 192 is considered for both RGB and TIR.

## 5.1 Dataset

The data utilized in this study was acquired by Sledz et al. (2020). Specifically, this dataset includes 6772 pairs of RGB-T images which are prepared from the orthomosaic of TIR and RGB with the patch size of 192 x 192. But, the registration process has limitations, which result in errors. Further detail related to data acquisition and photogrammetric processing can be found in Sledz et al. (2020). Addition to the pairs of RGB-T, round truth(GT) dataset is also considered. Classes($C_i$) in the GT can be defined with frame of discernment($\Theta$),

$$\Theta = \{C_1, C_2, ....C_N\} \qquad 1 \leq i \leq N \tag{37}$$

In the current work, N is equal to four and $\Theta$ is described by:

- BG: class that represents the background candidates
- AN: class that represents the thermal anomalies

- HO: class that represents the Hot objects
- CO: class that represents the Cold objects

This dataset contains different type of challenges such as different size of objects(big salient objects(BSO) & small salient objects(SSO)), multiple salient objects(MSO), clustered objects(one class consists another) etc. As per our main objective the most important challenge is to tackle with thermal anomalies.

| ***Class combination*** | ***Overall count*** | ***Training count*** |
|---|---|---|
| BG | 502 | 0 |
| BG and HO | 534 | 371 |
| BG and CO | 1205 | 824 |
| BG and AN | 136 | 91 |
| BG and HO and CO | 1540 | 1061 |
| BG and HO and AN | 321 | 217 |
| BG and CO and AN | 830 | 566 |
| BG and HO and CO and AN | 1704 | 1070 |
| All other class combinations | 0 | 0 |
| **Total** | **6772** | **4200** |

Table 1: Count of different Class combinations in Ground-truth(GT) dataset

Table 1 shows the number of images consisting objects from different classes. As it can be seen from in the second column (overall count) of table 1 , only 2991 images include an anomaly class out of total 6772 images. Subsequently, table 2 shows the number of pixels for each class in the whole dataset, and it clearly depicts that the size (number of pixels) of the BG class as compared to others is unbalanced, as it alone contains the 93 % of the total labelled pixels. Firstly, to mitigate this problem, the training dataset is manually selected i.e. the appearance of the BG is down-sampled. For this purpose, images are sorted with respect to the number of pixels acquired by BG-class, with the top 10 percent(with the high number of BG-pixels) of images are eliminated and rest of the randomly select 70% data for training and rest of the data used for tesing and validation. This strategy helped to reduce the acquisition of BG class pixels by 4% for training as shown in table 2 and table 1 training counts(column 3) in shows the number of images with different combinations used for training. In total, training dataset contains 4200 image pairs, while, test and validation are performed on the remaining image pairs using 0.5:0.5 ratio.

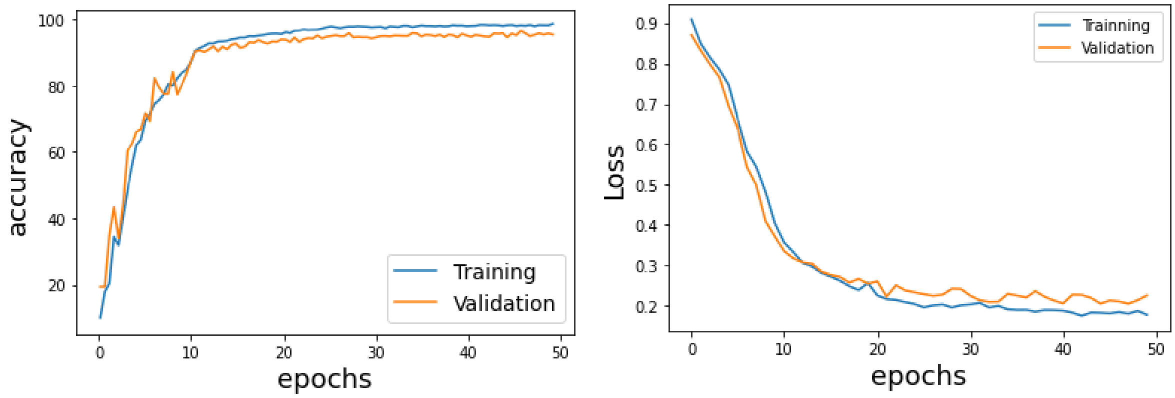
| ***Class*** | ***Total Pixel count*** | ***Training Pixel count*** |
|---|---|---|
| Background(BG) | 231491270(92.73%) | 137329392(88.70%) |
| Anomalies(AN) | 4316206(1.73%) | 4245755(2.74%) |
| Hot Objects(HO) | 3529607(1.41%) | 3208290(2.07%) |
| Cold Objects(CO) | 10305925(4.13%) | 10045363(6.49%) |

Table 2: Number of labeled pixels for each class in GT

This down-sampling approach did not significantly reduce the acquisition of BG class, but it helped to increase the possession of each class(AN, HO and CO) by approximately 1.5 times in the training dataset as compared to the whole dataset. However, evaluation of training and test process shows that BG class is dominating the results, further detailed discussion can be found in Chapter 5.2(Training & Validation) and in Chapter 5.3 (Results).

## 5.2 Training & Validation

As discussed above in chapter 5.1, the model is trained with 4200 pairs of images for 50 epochs and validation is performed on 900 images. Initially, to analyse the training and validation process accuracy metric is employed, the plot for both training and validation is shown in Figure 11a. From this plot, it can be clearly depicted that accuracy increases rapidly above 90%, and after about $10^{th}$ epochs, accuracy remains stable around 95-98% over the next epochs.



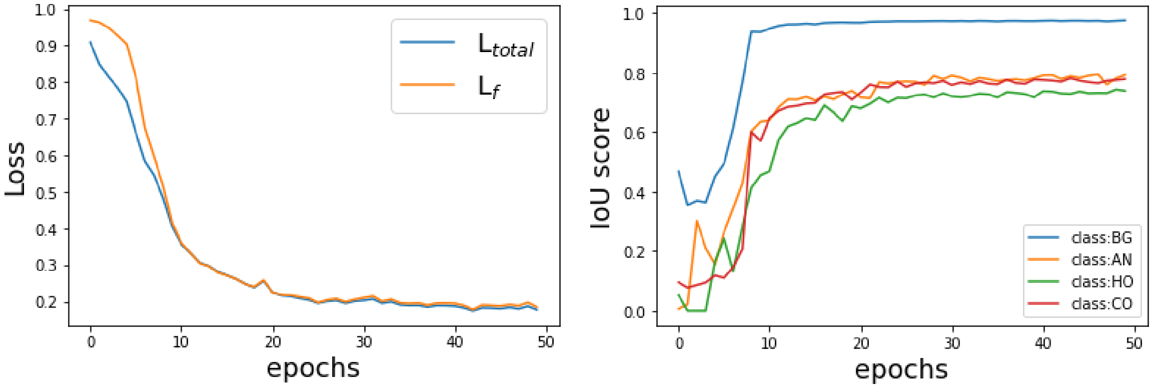
(a) Overall accuracy

(b) Loss

Figure 11: Overall accuracy and loss for training and validation

The reason behind this is that, as we discussed in chapter 5.1 the ratio of background pixels is 88% in training dataset, so, this class can alone outweighs the overall accuracy. However, it is not a good evaluation metric for analysing unbalance dataset because one class can outweigh the overall performance. As region of interest for this work is thermal anomaly, which takes only a small percentage($\sim$ 2%) of pixels in the image, whereas the $\sim$ 88% image is all annotated as background. Because of the true negative inclusion, the accuracy metric will always result in an illegitimate high scoring. Even predicting the segmentation of an entire image as background class, accuracy scores are often higher around 90% or even sometime close to 100%.

Subsequently, Figure 11b shows the loss plot for training and validation processes, the training loss indicates how well the model is fitting the training data, while the validation loss indicates how well the model fits new data. Despite the fact that validation loss is slightly higher than training loss, but Figure 11b shows validation loss follows the training loss for the whole training process, and it shows no sign of under or over-fitting, so it shows that model is learning effectively. As discussed in chapter 4.4, generalized dice loss is used for calculating the loss, which means it gives high weight to the classes which have small size, as in this work the size thermal anomaly class is very small as compared to background class, so with respect to the equation 34, much higher weight will be assigned to thermal anomaly class as compared to background class .



(a) Total loss vs decoders loss(mean)

(b) Individual IoU

Figure 12: Loss and IoU comparision

For better understanding about the learning process of the individual decoder outputs, loss for the each decoder is also analysed as shown in figure 12a the plot of the losses for RGB and thermal decoders(for the both parts of the equation 25). It shows decoder loss and total loss follows the same trend and after the initial 10 epochs, both remained indiscriminable from each others. Further evaluation is done using IoU score (Intersection over Union) and $F_\beta$ score. Both of these metrics are used for class-wise evaluation.

Figure 12b shows the plot of intersection over union, calculated with equation 11 for each class, it shows BG class outperforms all classes as it reaches above 0.90 just in few epochs while rest of the classes just managed to touch 0.80 after about 30 epochs and are remained stable around this score for the rest of the epochs.

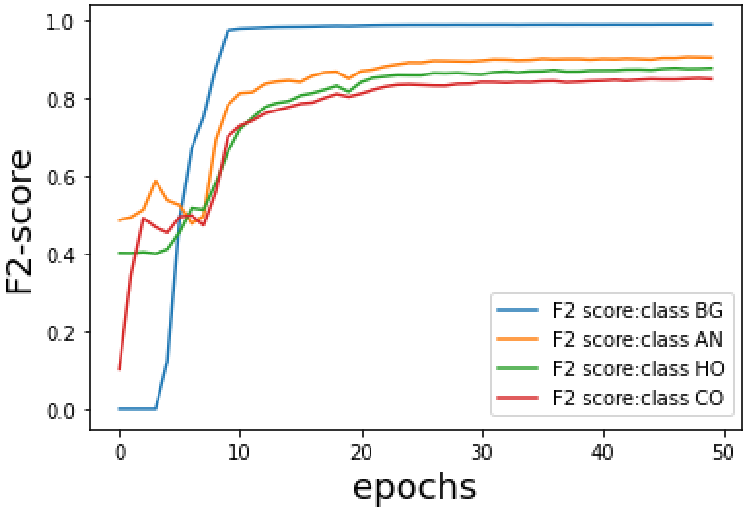


Figure 13: Individual F2-Score

Figure 13 shows the $F_2$ score for the training with respect to each class using equation (15), where $\beta^2$=2 is used to emphasizes the importance of recall as discussed in chapter 3.6.4. Motive behind this is that, this work is focused on thermal anomalies(AN) detection and cost of missing AN is consider higher than predicting preciously, i.e recall(R) utilizes(equation 13) false negatives(FN) instead of false positives as precision(P) do. Same as IoU score, BG class again outweighs all other classes in $F_2$ score, for BG class $F_2$ score is nearly hitting 1.0, while $F_2$ score for all other classes(AN,HO and CO) is ranging approximately between 0.80 to 0.90.

## 5.3 Results

After training the model, testing is performed on the set of 900 images. Below table 5 shows the confusion matrix for tested images and it depicts that all the four classes achieves true predictions above 80%, but if we compare the performance of BG class, same as training it again outperforming all the classes.

| | BG | AN | HO | CO |
|---|---|---|---|---|
| BG | 29590230 (98.75%) | 72308 (0.24%) | 75193 (0.25%) | 227102 (0.76%) |
| AN | 89584 (12.12%) | 643075 (87.01%) | 819 (0.11%) | 5633 (0.76%) |
| HO | 72912 (12.04%) | 507 (0.08%) | 513595 (84.81%) | 18580 (3.07%) |
| CO | 193384 (10.35%) | 6106 (0.33%) | 6793 (0.36%) | 1661779(88.96%) |

Table 3: Confusion matrix, where vertical axis shows true labels and horizontal axis shows predicted results, brackets show percentages of each prediction

Moreover, if we closely analyze the confusion matrix then it shows the majority of failed predictions(false predictions) for AN, HO and CO gone to BG class, as around 10-12% of each class labeled as BG. Below figure 14 and 15 show some of the best and worst predictions. Figure 14 shows the top 4 results, with input images(RGB and TIR), ground truth(GT), prediction and the differnece between GT and prediction as well as the outputs from both decoders(RGB and TIR).

In Figure 14 and Figure 15, row (e) shows the binary difference between GT and predictions, where 0 (black color) represents the correct class prediction, and 1 (white color) represents the incorrect one. Both of these shows that generally predictions are not accurate around the borders of the objects, as nearly all the difference images(row e) show an outline around the objects. For this outcome there are two reasons, first used ground truth dataset is prepared using co-registered RGB and TIR orthomosaic images, which consists error because of the registration and secondly, specifically no precaution was implemented to preserve the edges of the objects. However, this work is intended to accomplish a specific objective, the primary objective of our analysis is to detect thermal anomalies on the object level but not on the pixel level. In the event, if predictions are getting an overlap of 70% with GT for thermal anomalies, that is enough for locating anomalies.

For the failed predictions in the figure 15, most of these cases arise for the very tiny objects, as discussed before predictions are not accurate enough around the borders. For instance, if a prediction for a small object losses just a few pixels around the border, the recall will be very low as it will lose high number of true predictions as compared to total size of the ground truth, as shown in $3^{rd}$ image(column 3) in figure 15. Secondly, in figure 15, $2^{nd}$ image(column 2) shows some good prediction with respect to the each label, but if we closely analyze thermal anomaly(AN, yellow color) prediction it shows some discontinuity in the outcome, because very thin parts of the object are eroded. Predictions are also somehow failed to differentiate between two objects of same class if both the objects are placed very close to each other as shown in $4^{rd}$ image(column 4) in figure 15 areas of two different objects are merged with each others. Further, in figure 15, $1^{st}$ image (column 1), it was failed to maintain the complexity of the object, as a result complex object is dilated. Moreover, small details can also not be preserved as shown in $1^{st}$ image.

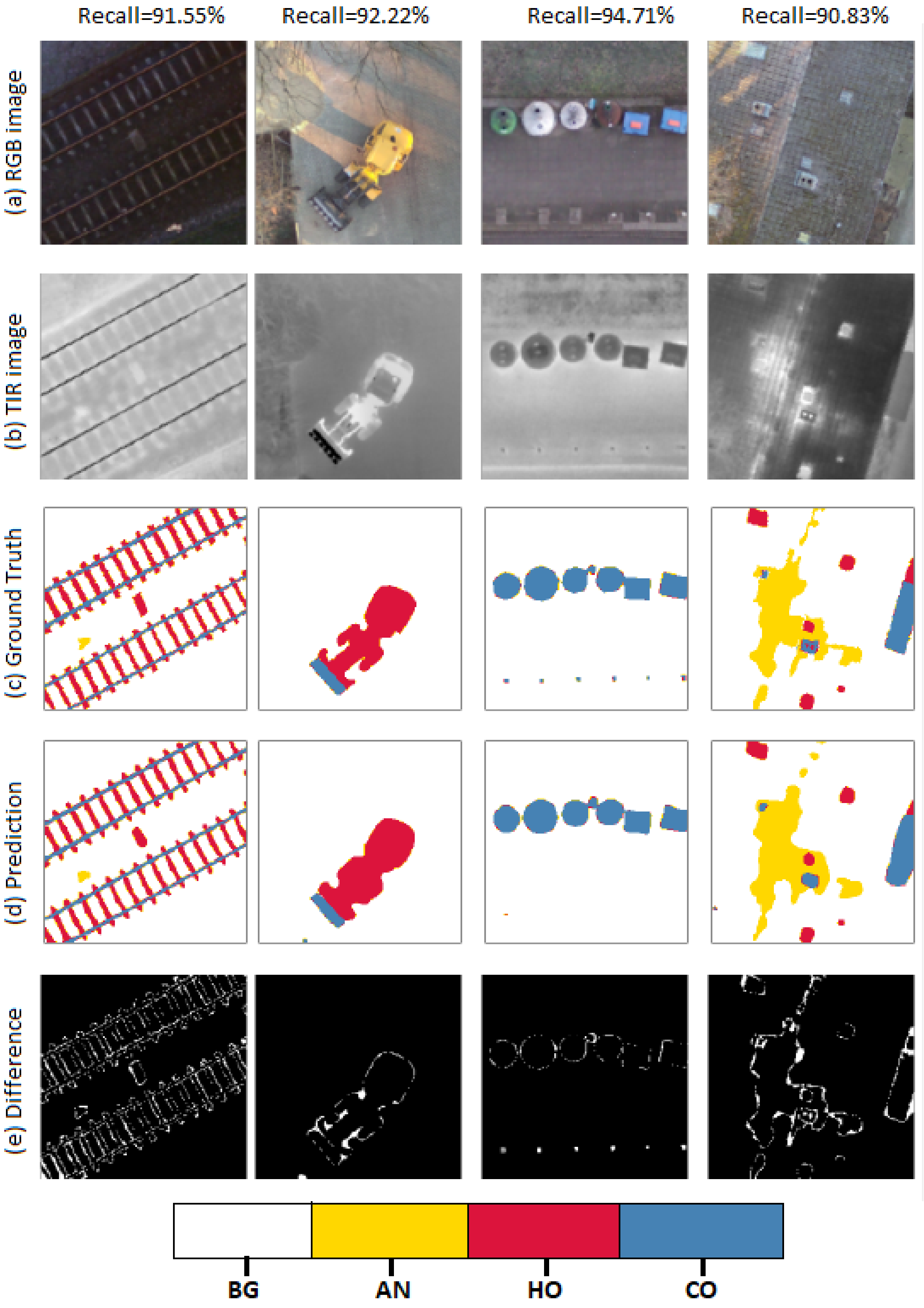


Figure 14: Successful predictions based on Recall value

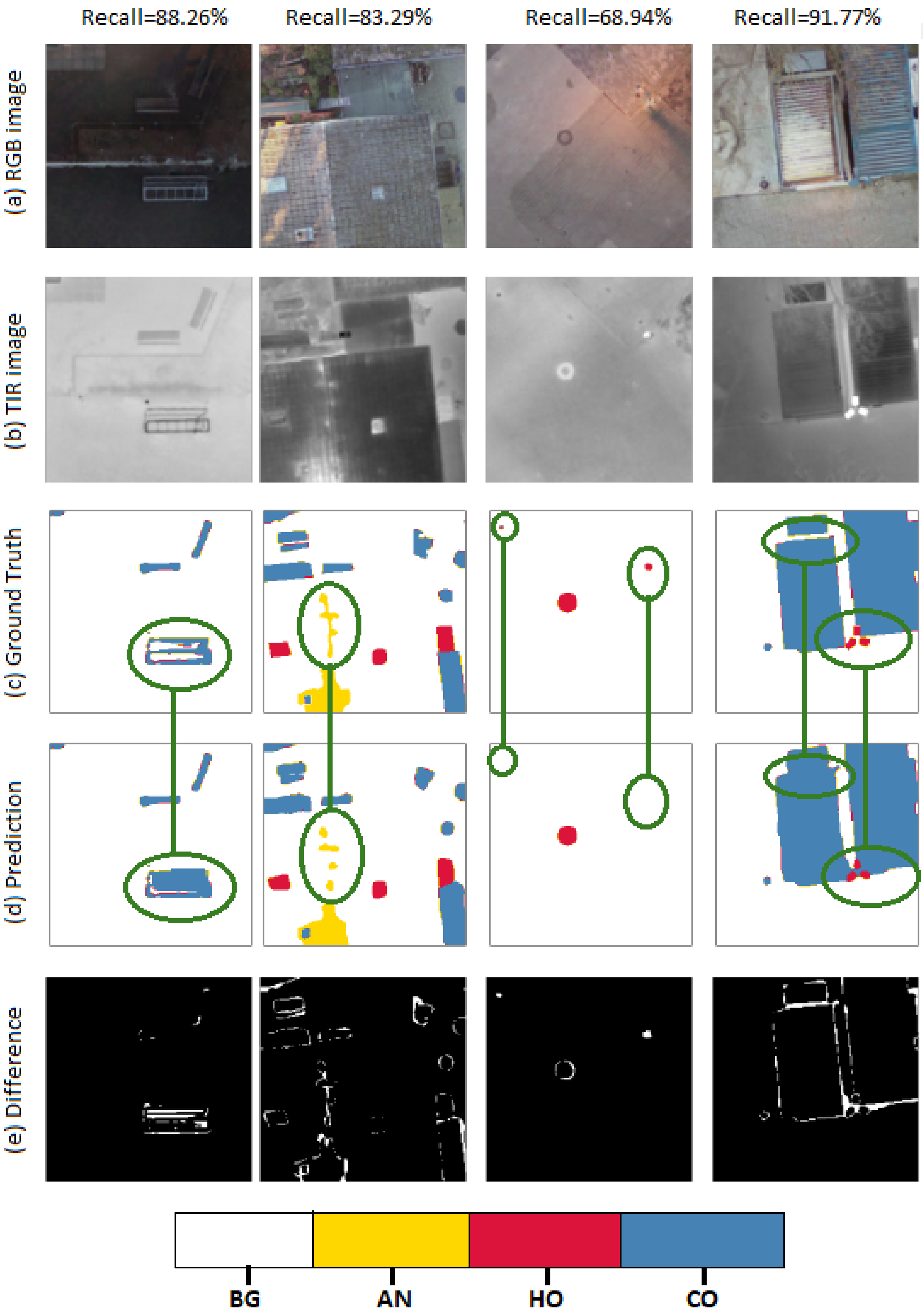
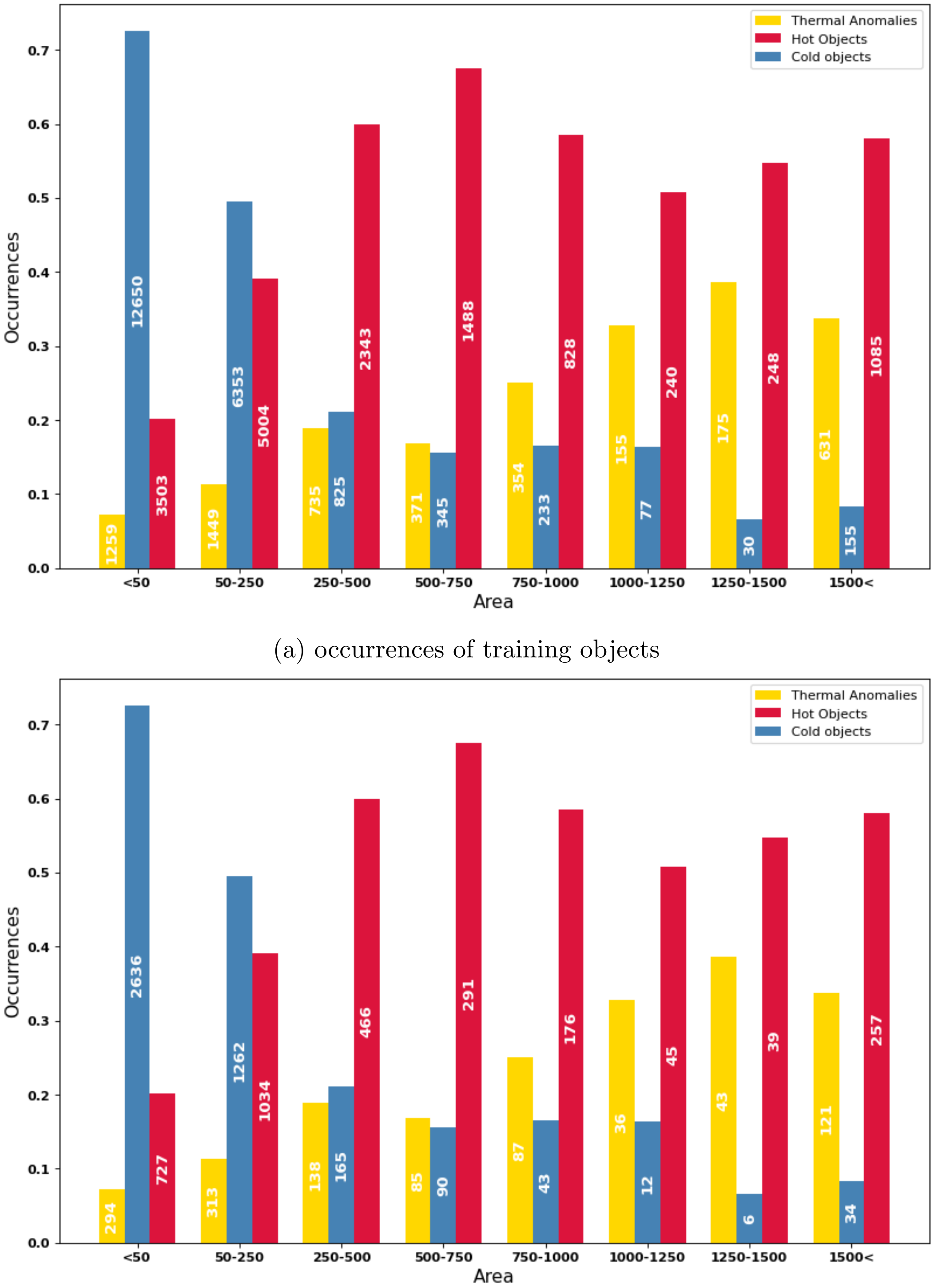


Figure 15: Failed predictions based on visual analyses



(a) occurrences of training objects

(b) occurrences of test objects

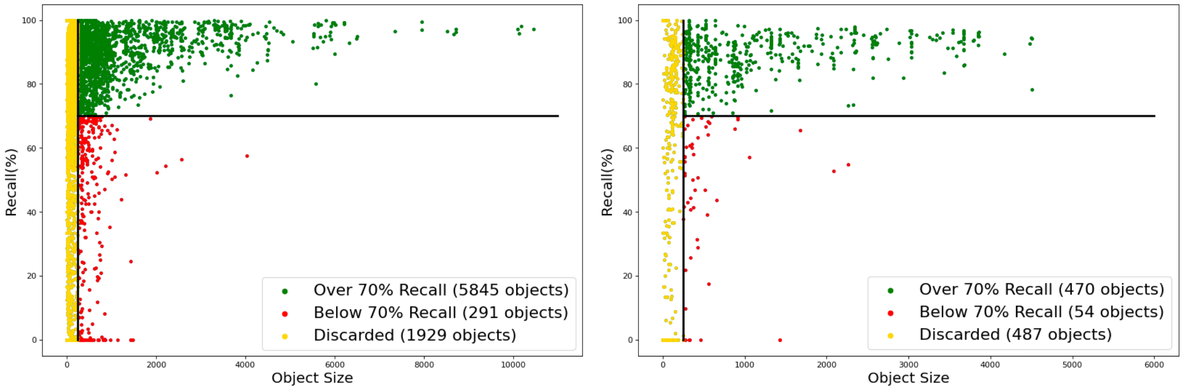
Figure 16: occurrences of training and test objects with respect to the size, numbers on the bars are the counts of objects of their respective class and size

So far, the training and testing have been evaluated based on pixel-wise analysis but as discussed before we are interested in the thermal anomalies on object level not on pixel level, so now instead of pixels evaluation, region based evaluation will be adopted. The used dataset have high occurrence of small sized objects(area $< 250$ pixels) as shown in figure 16, further as discussed in chapter 5.1, 10% of the total images were dropped from the whole dataset which have very high number of BG pixels or had very tiny objects and this approach of dropping 10% of images did not affected the number of objects. Figure 16a and figure 16b show the occurrences of the objects for training and test,respectively. The y-axis shows the ratio for each class for each object size segment individually. In addition, numeric values on bars shows the number of objects for each object size segment.

| | Class:1(AN) | Class:2 (CO) | Class:3 (HO) |
|---|---|---|---|
| Precision(pixel) | 87.73% | 88.39% | 83.41% |
| Precision(object) | 77.02% | 76.91% | 75.06% |
| Recall(pixel) | 85.76% | 85.36% | 85.28% |
| Recall(object) | 76.16% | 80.35% | 72.45% |

Table 4: Precision and recall based on pixel and object analyses

For the object-wise evaluation, recall and precision are utilized as main metrics. Table 4 shows the comparison of the object vs pixel evaluation with respect to each class. But these results also include the small objects as shown in figure 16a, so, these minor object affects the overall outcome. In this table it can be noted that, the precision and recall for the objects is much lower(approximately by 10%) as compared to the pixel level. The reason behind such a behaviour is that, the test dataset contains very high number of small objects($> 250$) as shown in Figure 16b, so if the prediction fails to predict only few pixels for a small object recall will be very low, On the other hand for the large objects it would not not affect much.To get the more clear view, below figure 17 shows how the recall is affected by the size of objects. As it can be seen that for the very tiny objects recall is low, and as the object size increases recall also gets higher.



(a) for all classes

(b) For thermal anomaly class

Figure 17: Recall vs Object size

As per the objective of this work, recalling objects with the overlap of more than 70% can be considered as a success, as in this work main interest is to locate anomalies instead of predicting them precisely. Figure 17a shows the recall vs object size for the all classes, it depicts that if the objects below 250 are discarded(yellow points), and setting a threshold for 70%, it shows that 95%(green points) of the objects are lies above this threshold. Further, as per the main objective of this work thermal anomalies are also evaluated solely, figure 17b shows the recall for the AN class in comparison to object size, it shows if the objects below area of 250 are neglected, 90% of the objects crosses the threshold limit. Further, figure 18 shows some of the events which have recall below 50% for thermal anomalies(AN). As it is clear from the visual inspection of figure 18 that most of these cases(below 50%) occurs, when an object lies along the edge or corner of the image. The reason for this behaviour is that, for the thermal anomaly detection it needs to be salient from all directions, but around the corners or along the edges it misses the surrounding from these sides.

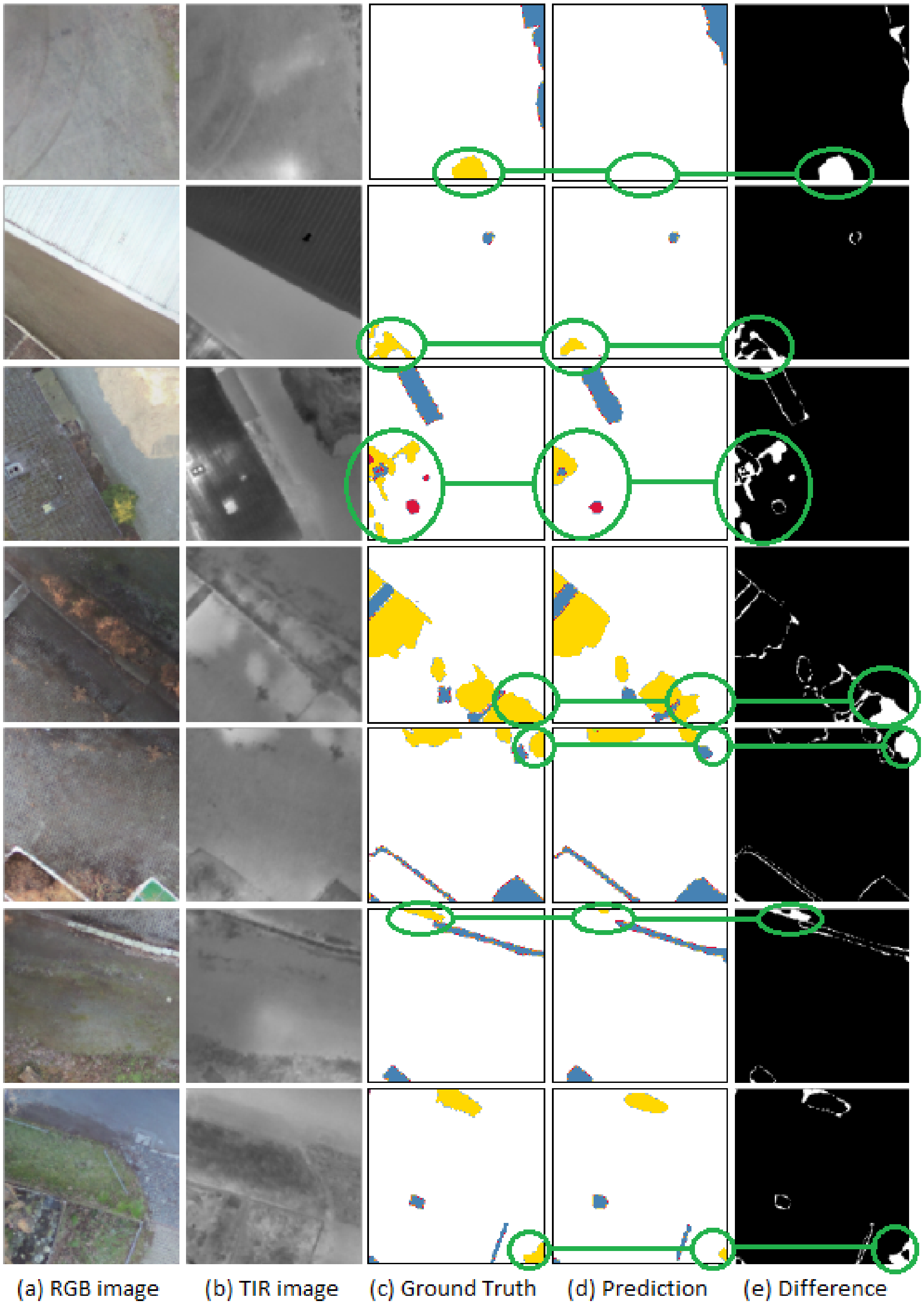


Figure 18: Thermal anomalies with less than 50% Recall for MIDD

# 6 Modified MIDD

MIDD (Tu et al. (2021)) performed as expected on the given dataset and objective, but if we analyse the outputs from the both decoders, then it arise attention, as both the decoders gives approximately same output maps. To validate this point, let's consider top outcomes from MIDD with individual outputs of both(RGB and TIR) decoders (figure 19).

In figure 19, row (e) shows the binary difference between both RGB and TIR decoder, where 0 (black color) represents correct class prediction, 1 (white color) represents incorrect one. This difference do not shows any significant disparities, it only shows negligible variations around the predicted objects and difference(denoted by diff on Figure 19) shows very minute values. So, in this experiment we modified the MIDD network, in which instead of using two separate decoders we will use a single decoder in order to simplify the overall architecture. To do so, decoder and the multi-interactive block (MIB) are modified, while the encoder and global information (GI) are kept same as MIDD, i.e. VGG16 is used for the encoder. In overall, a simplified network is proposed, which reduces the size of decoder network by half, as it only uses three Modified MIBs(MMIB), unlike MIDD which uses six MIBs(3 for each decoder).

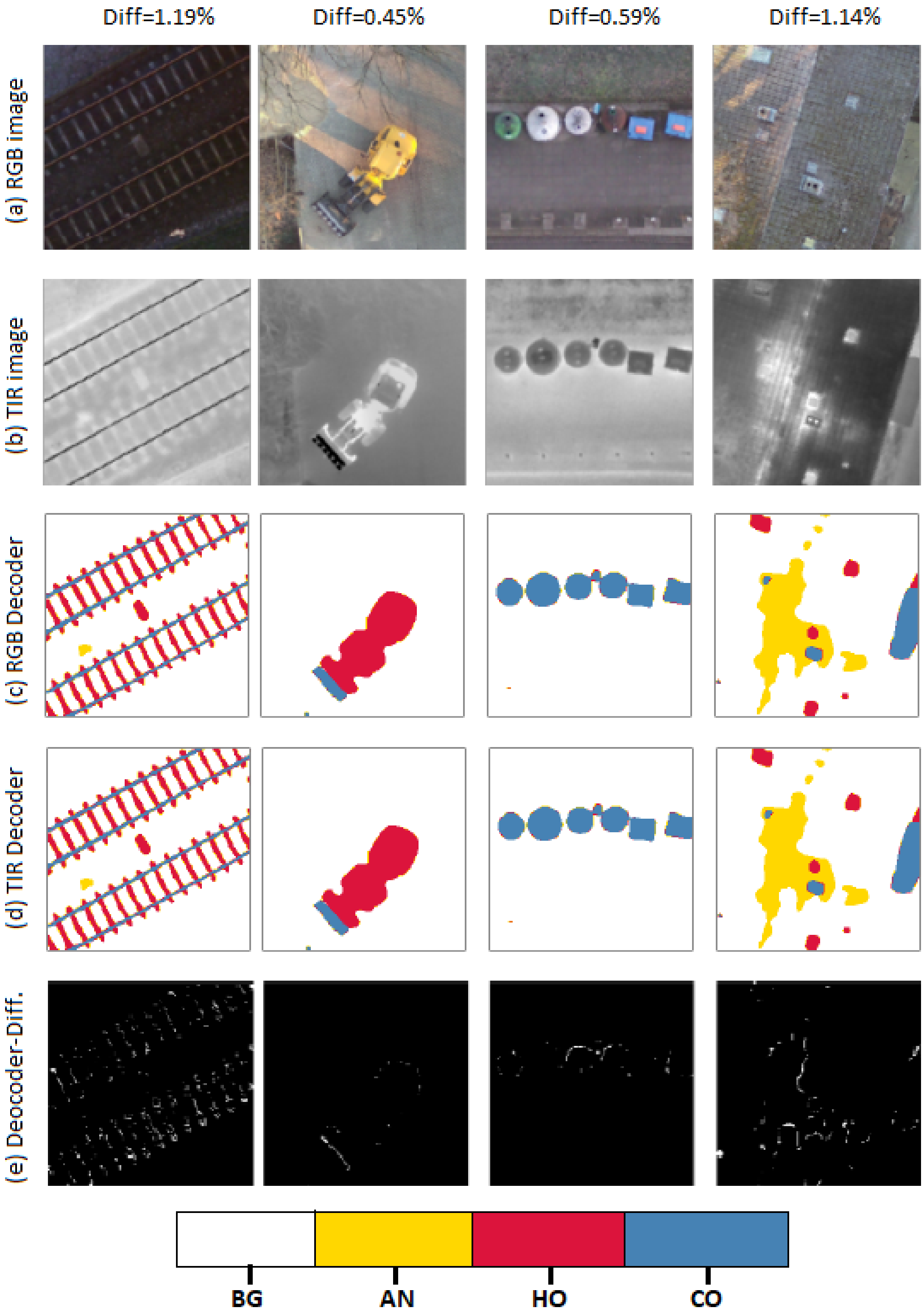


Figure 19: Differnce between RGB-TIR decoders,(a)RGB Image (b) TIR image (c) RGB-Decoder(d) Thermal-Decoder (e) difference between RGB decoder and TIR Decoder outputs

Overall architecture of the proposed network is shown below in figure 20, which take pair of RGB-TIR images as inputs. Then, from both of these inputs, features are extracted using VGG16 as backbone network. Same as MIDD, last convolution is used to get global context using global information module (shown in figure 6), and same GI module was adoted as disussed in chapter 4.2. In figure 20,marked with R1-R5(yellow) shows extracted RGB features, T1- T5(green) shows TIR encoded features, GI is global information(red) and multi-interactive block is marked with MMIB(orange).

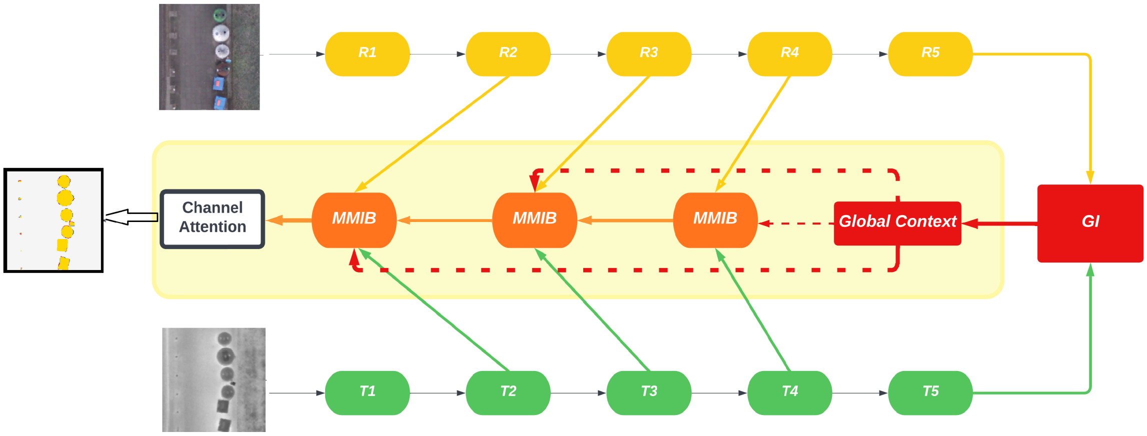


Figure 20: Modified MIDD

As discussed in chapter 4.2, GI module take top encoded features(R5 and T5) of RGB and thermal modalities as an input and then concatenate them in a channel-wise way. Considering all the parameters as Global Information module as in MIDD(chapter 4.2), mathematically, global context(G) from the GI can be written same as equation 19,

$$G = Conv([F_i']) \tag{38}$$

## 6.1 Modified Decoder Network

Modified MIDD network uses a single decoder instead of using individual decoder for each RGB and thermal modalities as shown in figure 20. For the interactions between both modalities modified multi-interactive block(MMIB) is designed to achieve the interactions of dual modalities, hierarchical features of MIBB and global contexts. Figure 21 shows the architecture of MIB, and it shows that MMIB takes four input features(RGB and TIR features, Modalities Integrated features and Global context). The main idea of using MMIB here is to use direct interaction of both modalities instead of first going through separate MIBs and then interact.

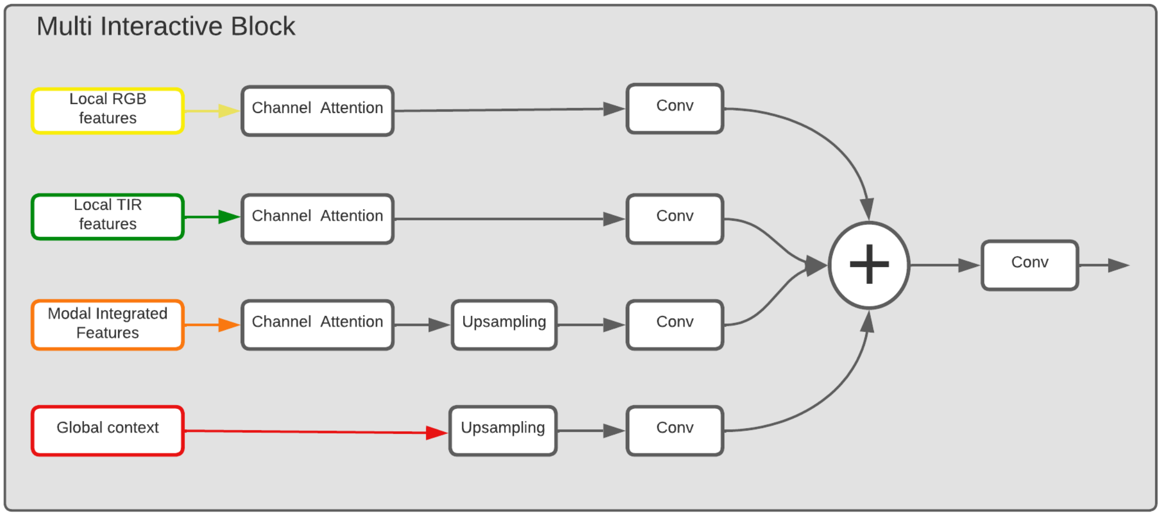


Figure 21: Modified MIB

Initially,features from both RGB ($R_j$)and TIR($T_j$)modalities are taken as the inputs from the respective RGB and T encoder layer,

$$R1_i = Conv(CA(R_j)), i = 2, 3, 4 \, and \, j = 4, 3, 2 \tag{39}$$

$$R2_i = Conv(CA(T_j)), i = 2, 3, 4 \, and \, j = 4, 3, 2 \tag{40}$$

Where, channel attention (CA) is used to emphasize the more useful features, and then decrease the number of channels to 128. Subsequently, Modalities Integrated features(M) are the outputs from the previous MMIB,

$$M_i = Conv(UP(CA(MIB_i))), i = 2, 3, 4 \tag{41}$$

Here, up-sampling the reconstructed features is needed to match the size with $R_j$ and adopt a convolutional block to reduce the number of channels of reconstructed features to 128. Global context(G) is taken as a final input to MMIB,

$$G'_i = Conv(UP(G)), i = 2, 3, 4 \tag{42}$$

To match the size with $R1_i$, G also needs to be upsampled and convolution block to decrease the number of channels to 128. At the end, after processing all these three inputs are summed up and reconstructed fused features through a convolution block as the output of MIB,

$$MIB_i = Conv(R1_i + R2_i + M_i + G'), i = 2, 3, 4 \tag{43}$$

In the Decoder network, output from the last MMIB($MMIB_4$) used to get final saliency map($S_f$) after processing it through CA and reducing the channels to 4 using a 2D convolution layer.

$$S_f = Conv2D(CA(MIB_4)) \tag{44}$$

## 6.2 Adopted Loss Function

Adopting generalized dice loss(GDL) and dice loss(DL) as discussed in chapter 4.4, here only final and global loss are considered, this network don't have any intermediate outputs unlike MIDD which also considers the individual losses from both decoders. So, using binary dice loss (equation 31)the global loss($L_g$) can be calculated as,

$$L_g = DL_2(S_g, GT_b) \tag{45}$$

$S_g$ is the global context reduced to single channel and $GT_b$ is the binarised ground truth. Final saliency loss is,

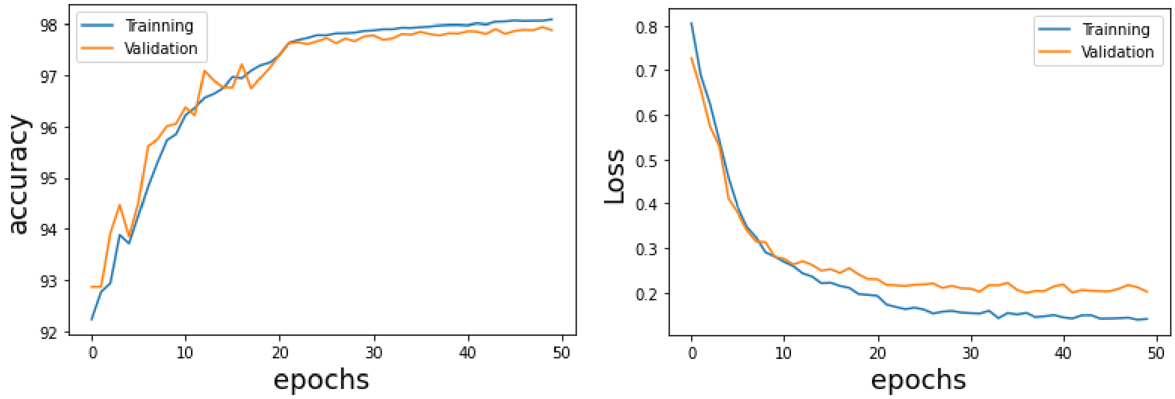
$$L_f = GDL(S_f, GT) \tag{46}$$

And the total loss($L_total$) can be calculated using $Lg$ and $L_f$,

$$L_{total} = \frac{L_g + L_f}{2} \tag{47}$$

## 6.3 Modified MIDD Training & Validation

For comparing and analysing the results, the modified model is also trained with same 4200 pairs of images and validation is performed on 900 images for 50 epochs as done in MIDD(Chapter 5.2). Again, to analyse the training and validation process for the initial stage accuracy metric is employed, the plot for both training and validation is shown in figure 22a. Similar to MIDD, accuracy reaches accuracy increases rapidly above 90% in just 10 epochs, and after around 25th epochs, accuracy remains stable between 95-98% over the remaining training process. While, the loss for the training and validation(Figure 22b) shows a gap as validation is the set of unseen data, but still both the losses are moving along each other, so there is no sign of over-fitting.

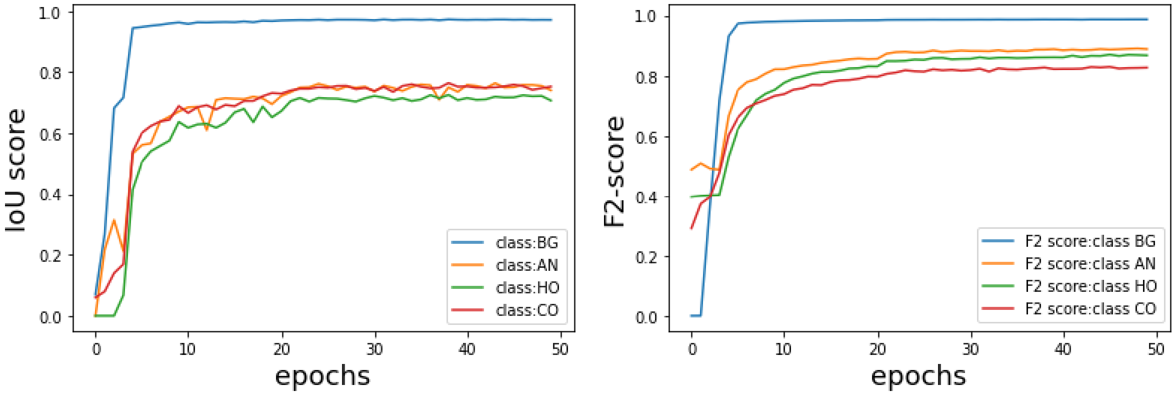


(a) Overall accuracy

(b) Loss

Figure 22: Overall accuracy and loss for training and validation

To analyse how this modified MIDD network behaves on different classes, evaluation is done using IoU score (Intersection over Union) and $F_\beta$ score. Both of these metrics are used for class-wise evaluation.



(a) IoU score

(b) Individual F2-Score

Figure 23: IoU and F2 score

Figure 23a shows the individual IoU score for each class, it shows the dominance of the BG class with just under 1 while other classes hardly touches 0.80. Similarly, figure 23 shows the $F_2$ score with respect to the each class, for the initial few epochs BG was trailing, with a rapid gain BG again shows it's dominance.Figure 23b and 23a show the F2 score and IoU for the training with respect to each class. As per the training, both of these plot shows compatible results to MIDD but this network is again biased to BG class as per the pixel level analyses.

## 6.4 Modified MIDD Results

After training the model, testing is performed on the set of 900 images. Below table 5 shows the confusion matrix for tested images and it depicts that all the four classes archives true predictions above 85%.

| | BG | AN | HO | CO |
|---|---|---|---|---|
| BG | 98.39% | 0.36% | 0.39% | 0.85% |
| AN | 11.69% | 87.53% | 0.21% | 0.57% |
| HO | 10.01% | 0.21% | 86.97% | 2.81% |
| CO | 11.50% | 0.57% | 1.03% | 86.90% |

Table 5: Confusion matrix, where vertical axis shows true labels and horizontal axis shows predicted results, brackets show percentages of each prediction

In comparison to MIDD, modified MIDD shows minor variation, it improves the true predictions for AN,CO and HO by minute margin, as well as the percentage of misclassification for all classes as BG also decreased relative to MIDD . But similar to MIDD, most of the false prediction are assigned to BG class, and BG is again dominating other classes with highest number of true predictions with 98.39%. Now, instead of using pixel-based analyses, evaluation will be done using region based analyses. Same testing dataset used shown above in bar graph(figure 16b). For the object-wise evaluation,

| | Class:1(AN) | | Class:2 (CO) | | Class:3 (HO) | |
|---|---|---|---|---|---|---|
| | MIDD | MMIDD | MIDD | MMIDD | MIDD | MMIDD |
| Precision(pixel) | 87.73% | 85.70% | 88.39% | 84.66% | 83.41% | 83.86% |
| Precision(object) | 77.02% | 77.62% | 76.91% | 77.29% | 75.06% | 71.06% |
| Recall(pixel) | 85.76% | 85.14% | 85.36% | 84.70% | 85.25% | 83.94% |
| Recall(object) | 76.16% | 76.70% | 80.35% | 83.20% | 72.45% | 79.33% |

Table 6: Comparision of MIDD & MMIDD on the bases of Precision and Recall for both pixel-wise and object-wise

recall and precision are utilized. Table 6 shows the comparison MIDD & MMIDD on the bases of Precision and Recall for both pixel-wise and object-wise. But these results also include the small objects as shown in figure 16a, so, these minor object affects the overall outcome. Considering all the different sized objects both MIDD and MMIDD shows some good predictions, but on the object level our proposed MMIDD network shows better results in comparison to MIDD. While to make these results look more clear, very tiny objects are dropped and subsequently, figure 24 below visualize the affect of object size on the predicted recall. It can be seen that as the object size increases, the recall value also increases.

| | Class:AN | |
|---|---|---|
| | MIDD | MMIDD |
| Recall (for all objects) | 77.02% | 77.62% |
| Recall(for object above area of 250) | 89.69% | 91.22%% |

Table 7: Comparision of MIDD & MMIDD for thermal anomalies(AN) on the bases of object size using Recall values

As discussed above in MIDD chapter, recalling objects with an overlap of more than 70% can be considered as a success, as in this work main interest is to locate anomalies instead of predicting them precisely. Further, as per the main objective of this work thermal anomalies are also evaluated solely, figure 24 shows the recall for the AN class in comparison to object size, it shows if the objects below area of 250 are neglected, 91.22% of the objects crosses the threshold limit, for MIDD it was 89.69%. Below table 7 shows the comparison of MMIDD and MIDD for thermal anomaly(AN) detection, as it clearly shows that, MMIDD performs better than MIDD for both instances, i.e for for all sized objects and for object below area of 250 pixels.

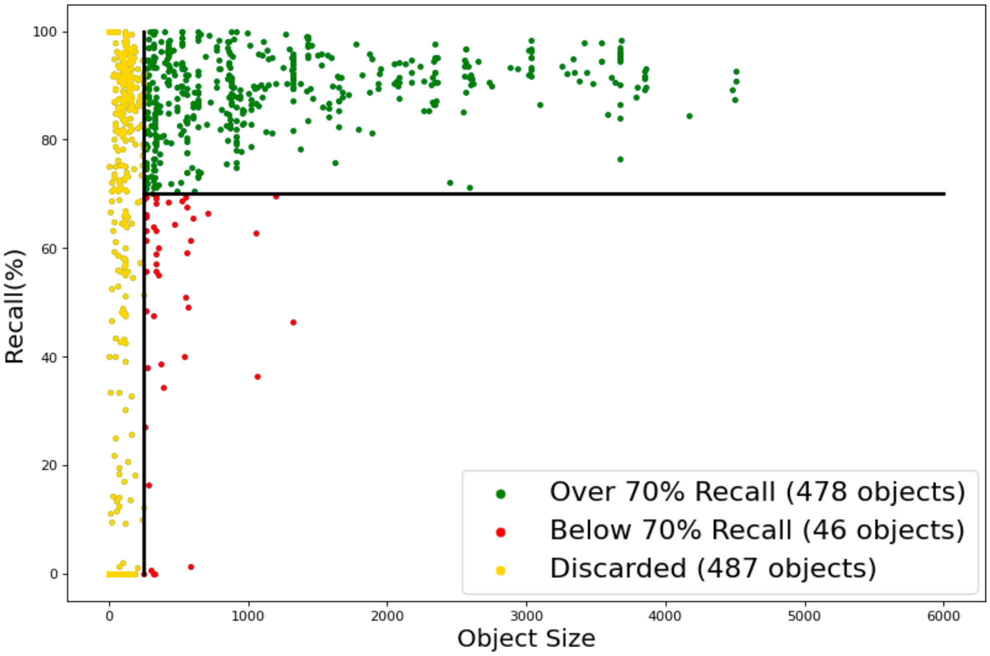


Figure 24: For thermal anomaly class

Further, in figure 25, detected thermal anomalies for MMIDD below 50% recall are analyzed. Similar to MIDD these predictions shows same problem, as approximately all these cases (below 50% recall) arises for the objects, which are either on the edges of the images or on the corners.

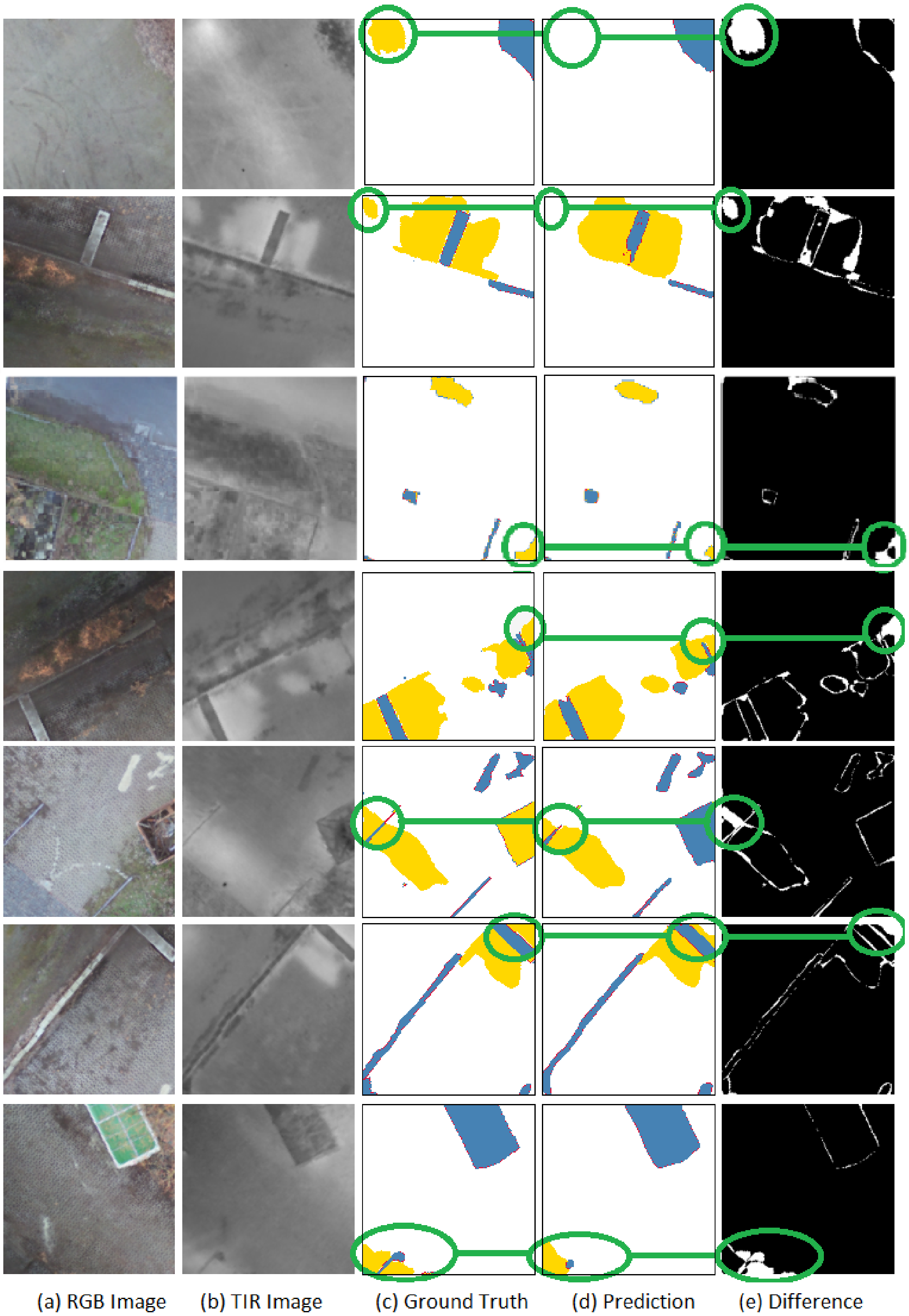


Figure 25: Thermal anomalies with less than 50% Recall for MMIDD

With respect to the evaluation of results, MMIDD do not make any significant improvement in the performance. But, if we compare the size of the both networks(MIDD and MMIDD), by simplifying the decoder of MIDD, MMIDD also reduced the number of trainable parameters. MIDD had 52429197 trainable parameters but MMIDD reduced these numbers to 37567237. i.e. MMIDD reduced the number of trainable parameters by 28.34% as comapred to MIDD. Consisting much smaller size of the network MMIDD still getting the compatible results to MIDD, in other words, MMIDD successfully reduced the complexity of the network.

# 7 Summary

The objective of this study is to detect thermal anomalies using a combination of thermal and optical images. Thermal anomalies are not discernible in RGB images, but only in thermal images. For this purpose, initially MIDD proposed by Tu et al. (2021) is adopted, except some specific cases discussed in 5.3 this approach helped to achieve good results. In order to further improve the results for the thermal anomaly detection, Modified-MIDD(MMIDD) is proposed. The proposed network simultaneously combines the information of two modalities directly from the encoder, and secondly, attempted to reduces the complexity of the network as dense connectivity requires significant amounts of GPU memory. Results show that the MMIDD gives compatible output to MIDD and significantly reduced the size of the trainable parameters. As per the results, after ignoring the small objects, both models show that nearly 90% of the detected objects for AN class, have an overlap of 70% or more with the ground truth.

Future research should be focused on the solution for imbalance in the training. As the predictions for MIDD and MMIDD are biased towards BG class. Next, to improving the pixel-wise performance of the network, as tiny object fades away and complex object are merged with each other in current predictions. Doing so, it will be easy to detect anomalies at initial stage of heat leakage. In addition, maybe inclusion of R1 and T1, which present high-frequency details, may lead to a better performance. Further we are also interested in taking measures to detect salient objects with fine boundaries, as it is also a active topic of research. Moreover, using large patch size could be solve the problem of failed predictions on the borders of images, as currently we are using relatively small patches of size 196x196. Subsequently, analysis should be focused on temperature analysis, as it maybe the case that failed results are coming from the low temperature difference between thermal anomalies and their surroundings.

# References

Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In Antonios Gasteratos, Markus Vincze, and John K. Tsotsos, editors, *Computer Vision Systems*, pages 66–75, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-79547-6.

Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009. doi: 10.1109/CVPR.2009.5206596.

Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-dujaili, Ye Duan, Omran Al-Shamma, Jesus Santamaría, Mohammed Abdulraheem Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 2021.

Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, PP, 08 2019. doi: 10.1109/TPAMI.2019.2935715.

Ali Borji, Dicky N. Sihite, and Laurent Itti. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24:5706–5722, 2015.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. volume 2017-January, 2017. doi: 10.1109/CVPR.2017.699.

Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Pro-*

*cessing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/file/4db0f8b0fc895da263fd77fc8aecabe4-Paper.pdf.

Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383267.

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2 2015.

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. doi: 10.1109/34.730558.

Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020. doi: 10.1109/CIBCB48159.2020.9277638.

Alex Krizhevsky, Sutskever Ilya, and Hinton Geoffrey E. “imagenet classification with deep convolutional neural network”, in advances in neural information processing systems, p. 1097-1105. *Elsevier Ltd*, 2012.

Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. doi: 10.1109/TIP.2017.2710620.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. 11 2014.

Matthias Kümmerer, Thomas Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. 10 2016.

Chenglong Li, Guizhao Wang, Yunpeng Ma, Aihua Zheng, Bin Luo, and Jin Tang. A unified rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. *ArXiv*, abs/1701.02829, 2018.

Akis Linardos, Matthias Kümmerer, Orion Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12899–12908, 2021.

Yu-Fei Ma and Hong Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. MULTIMEDIA '03, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137222. doi: 10.1145/957013.957094. URL `https://doi.org/10.1145/957013.957094`.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. *American Journal of Psychology*, 76:705, 1963.

Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 853–860, 2012. ISBN 9781467312264.

doi: 10.1109/CVPR.2012.6247758. 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012 ; Conference date: 16-06-2012 Through 21-06-2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

Artuom Sledz and Christian Heipke. Thermal anomaly detection based on saliency analysis from multimodal imaging sources. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-1-2021:55–64, 2021. doi: 10.5194/isprs-annals-V-1-2021-55-2021. URL `https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/V-1-2021/55/2021/`.

Artuom Sledz, Jakob Unger, and Christian Heipke. Uav-based thermal anomaly detection for distributed heating networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B1-2020:499–505, 2020. doi: 10.5194/isprs-archives-XLIII-B1-2020-499-2020. URL `https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B1-2020/499/2020/`.

Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer International Publishing, 2017. doi: 10.1007/978-3-319-67558-9_28. URL `https://doi.org/10.1007%2F978-3-319-67558-9_28`.

Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021.

Paul Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Harvard: Harvard University.*, 1974.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Chang Xu, Qingwu Li, Mingyu Zhou, Qingkai Zhou, Yaqin Zhou, and Yunpeng Ma. Rgb-t salient object detection via cnn feature and result saliency map fusion. *Applied Intelligence*, 52, 08 2022. doi: 10.1007/s10489-021-02984-1.

Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. pages 3166–3173, 06 2013. doi: 10.1109/CVPR.2013.407.

Xinchen Ye, Xin Fan, Mingliang Zhang, Rui Xu, and Wei Zhong. Unsupervised monocular depth estimation via recursive stereo distillation. *Trans. Img. Proc.*, 30:4492–4504, jan 2021. ISSN 1057-7149. doi: 10.1109/TIP.2021.3072215. URL `https://doi.org/10.1109/TIP.2021.3072215`.

Qiang Zhang, Nianchang Huang, Lin Yao, Dingwen Zhang, Caifeng Shan, and Jungong Han. Rgb-t salient object detection via fusing multi-level cnn features. *Trans. Img. Proc.*, 29:3321–3335, jan 2020. ISSN 1057-7149. doi: 10.1109/TIP.2019.2959253. URL `https://doi.org/10.1109/TIP.2019.2959253`.

Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection.

*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3080–3089, 2019.

Yanfei Zhong, Yao Xu, Xinyu Wang, Tianyi Jia, Guisong Xia, Ailong Ma, and Liangpei Zhang. Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 151:207–222, May 2019. doi: 10.1016/j.isprsjprs.2019.02.021.