Enhancing Vessel Re-identification with RGB-Infrared Multi-Modal Techniques

F. van Abbema
f.vanabbema@student.utwente.nl

ABSTRACT

Vessel re-identification tries to identify a new ship as a ship that has been seen before or as an unknown ship. This field has made some good progress. However, all the literature only makes use of the visible light (RGB) modality. The use of the infrared (IR) modality has not yet been explored in this field. Using an IR modality next to the RGB modality adds new information to a sample. Exploiting this new information might result in a more generalised and expressive model and therefore a better-performing model. In this research, RGB-IR multi-modal models will be compared with the RGB-only models. In order to achieve this, a new RGB-IR vessel re-identification dataset is presented. Results show an increase of 0.023 of the weighted sum of rank-k accuracies and area under the precision-recall curve for the best RGB-IR model compared to the best RGB-only model. These results show that IR adds valuable information for vessel re-identification.

KEYWORDS

Vessel re-identification, Multi-modal learning, RGB, Infrared, Convolutional neural network, ResNet

1 INTRODUCTION

In this world where maritime traffic is a vital piece of our infrastructure, we require our ships to travel safely. The crew on a ship needs to monitor its surroundings. This can involve a lot of work if done manually. Cameras systems are one of the tools that can help to monitor the surroundings of a ship automatically. Especially when a ship sails in dangerous waters and needs to track multiple hostile ships simultaneously.

The current technology available today has the capability to identify ships. Take for example Thales Nederland B.V. which is a company that produces such camera systems for ships and where this research is performed. Their technology is a good basis among others for monitoring the surroundings of a ship. For example, with the Gatekeeper which is a camera surveillance system with 360 degrees horizontal field of view and visible light (RGB) and infrared (IR) cameras. However, certain situations are not fully explored in the literature yet. One of these situations occurs whenever a ship disappears and later reappears in the view of the camera. This can happen for example, when a ship sails behind another ship. The camera can not know if this ship is the same ship from before or if another ship has entered the view of the camera. This problem can be solved by the task of vessel re-identification.

The task of vessel re-identification is explained in Figure 1.

In the real world, the set of known ships for re-identification is defined as the ships seen in the last several minutes by a camera system. The model and threshold used by re-identification are gained by training and evaluating on a vessel re-identification dataset.

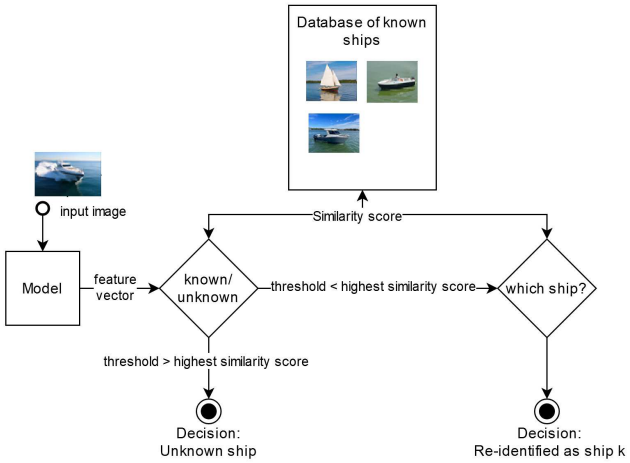

Figure 1: A schematic overview of how vessel re-identification is done. A system has a database of known ships, a model that can extract a feature vector from a ship and a threshold that decides given a similarity score gained from a pair of ships, whether the pair is from the same ship or not. With these components, a new ship is fed into the system and the system decides whether it is an unknown ship or it can be re-identified as one of the known ships.

The current literature focused on the use case of this problem in harbours or canals. Here, the harbour/canal management needs to monitor the space with cameras. However, the cameras do not cover the complete space and there are areas where a ship is not in the view of any camera. In this case, the use case behind re-identification stems from mapping ships to the same identity across multiple cameras. Furthermore, in these researches, there are no ships that are seen for the first time by the system. In other words, every sample in the dataset has at least one other sample from the same class. In the literature, this task is then seen as a classification problem during training and evaluation. In contrast, in this research, a ship might not belong to any seen class. However, with an extension these unknown boats can be incorporated into the model which is also seen in Figure 1 (when the threshold is higher than the similarity score of the most similar boat). Apart from this extension, the methodology of training and evaluating known boats for both use cases is the same. Therefore, the literature on vessel re-identification can be used and built upon.

When compared to the field of vessel re-identification, the field of person re-identification is more advanced. Therefore, some techniques used in this field can be used as inspiration for vessel re-identification.

One of the techniques used by person re-identification is the use of both RGB and IR data. This has not been explored yet in the field of vessel re-identification. This research will explore the use of these two modalities. The combined modalities will be compared to single-modality approaches to show if an improvement can be made. The modalities can be combined by focusing on modality-shared features (features seen in both modalities) or by focusing on modality-specific features (features seen in only one modality). These two approaches will be researched to see which one improves re-identification the most.

This paper has the following contributions:

- A new dataset with paired RGB-IR data is presented. This is the first dataset that contains vessels with both RGB and IR samples that can be used for vessel re-identification.
- The results show an improvement over the state-of-the-art by adding IR input next to the RGB input and focusing on the modality-specific features. With the dataset and models used in this paper, an increase of 4.5% rank-1 accuracy and 0.03 AUPRC is gained. When these metrics are combined, the combined score is increased by 0.023.

The structure of this paper is as follows. In section 2, the related work will be analysed. In section 3, the gap in the literature will be identified and a research question will be formulated. The approach to answering this question will be described in section 4. The results gained by following the approach are presented in section 5. The implications of the results on the research question will be discussed in section 6. Finally, the limitations of this research and the open issues are listed in section 7.

2 RELATED WORK

In 2019, Spagnolo et al. [19] started with vessel re-identification. They created a publically available dataset named Boat Re-Id containing 5523 images and 107 different identities. This dataset is still the basis for most vessel re-identification literature. As the first dataset was very small, this introduced a less robust model. So, other researchers improved on this by creating a larger dataset. The research of Qiao et al. [15] worked on a much larger dataset and proved that their method creates a more robust model for vessel re-identification.

Spagnolo et al. [19] used the pre-trained network ResNet50 [6]. This model is used for image classification and is a well-performing model on the ImageNet dataset [1]. The reason behind using a convolutional neural network (CNN) for processing the image is that a method is required to match similar ships with slight deviations. These deviations include a different viewpoint at the target ship, different lighting conditions, occlusions of the target ship et cetera. This is why classical image classification is not useful and machine learning comes in handy. The ResNet50 is adapted such that it maps an image to one of the 107 classes in the dataset. This is an adaption that can extract only the prominent features of each ship. This works for basic cases but when samples become harder this model struggles.

Other researchers noticed this struggle and amplified the model by adding local feature extraction (i.e. on parts of the image) next to global feature extraction (i.e. the whole image). Groot et al. [31] used an adaption of the Multiple Granularity Network (MGN) model [20]. This is a network with three branches and the ResNet50 model as backbone. The first branch captures the image fully, the second splits the image in an upper and lower part and the third splits the image in an upper, middle and lower part. The first branch then captures the global features and the other branches capture the features per split part of the image (i.e. the local features). Ghahremani et al. [4] applied the principle of MGN to create a similar network named Maritime Vessel Re-identification Network (MVR-Net). Next to splitting in the height dimension, MVR-Net also splits the image in the width dimension and channel dimension. The width branch has three branches that splits the image into one, two (left and right) and three parts (left, middle and right). The channel branch splits the feature maps into four. Ghahremani et al. [4] state that the height and width branches compensate for the varying input resolutions due to different viewpoints of the ship. Furthermore, they state that the channel branch captures internal correlations between feature maps better.

Qiao et al. [15] approached the split on local features differently. They detected semantic regions in an image using a YOLO network [16] and used the cropped regions as input for the specific branches.

Lastly, Groot et al. [31] employed some other interesting optimisations applicable to their dataset. One interesting optimisation is the use of vessel travel time filtering. In the used dataset, the ships appear in two cameras which are set up in a canal with a distance of a few kilometers between the cameras. The time a ship takes to travel between these two cameras is also considered. Only ships that appear between a minimum time or maximum time in the second camera after appearing in the first camera are considered for re-identification.

This summarises the state-of-the-art of vessel re-identification. Having a multi-branch network seems to perform the best as the network allows to focus on specific parts of the target and combine all the specific part features to result in a more expressive feature set. The state-of-the-art only focuses currently on the RGB modality. To explore the capabilities of RGB-IR multi-modal learning, another field needs to be analysed.

Take note that each of the above researches works on a closed set of vessels. In other words, there are no unknown ships that don't belong to any of the ships in the query set. This is different from the use case of this research.

2.1 RGB-IR multi-modal learning in person re-identification

In the field of vessel re-identification, there has been no usage of multiple modalities yet but when the field of person re-identification is analysed, some valuable lessons can be learned that can be used in the field of vessel re-identification.

Nguyen, Hong, Kim and Park [14] were the first ones to pioneer the use of both RGB modality and IR modality. According to them, the use of two different kinds of images helps us to reduce the effects of noise, background, and variation in the appearance of a human body. Wu, Zheng, Yu, Gong and Lai [21] continued their work. According to them, RGB images are not always suitable e.g. in a dark environment or at night. Therefore, IR imaging becomes necessary in many visual systems.

On a high level, there are two ways to combine both modalities. The first is shared feature learning which aims to embed the features of both modalities into the same feature space. Therefore, the modality-specific features (features only seen in one modality and not in the other modality) are left out and the modality-shared features are only used (features shown in both modalities). This approach excels in filtering background noise. This is because the noise has to be present in both modalities to be used in the shared feature learned model.

The second approach to combining RGB and IR is feature compensation learning. This compensates for the missing modality-specific cues in the shared space. This can be done by considering each modality individually. This might bring noise from one modality in the feature space but modality-specific cues are also incorporated in the feature space.

Starting in the shared feature learning, Nguyen et al. [14] researched multiple feature vector extraction methods for only RGB, only IR and for both the RGB and IR frames. These feature vectors are concatenated followed by a Principal Component Analysis (PCA) to reduce the size of the feature vector. This resulted in combining RGB and IR frames in a CNN that had the best performance.

However, this requires two networks for both RGB and IR. Ye, Lan, Li and Yuen [23] took the first step into having combined weights for both modalities. The backbone CNN is not shared between the modalities. However, some fully connected layers are added with shared weights. Ye, Lan, Leng and Shen [22] improved this by sharing weights in the backbone CNN. Here, the first stage of the ResNet50 model is still modality-specific but the other 4 stages have shared weights such that the modality-shared feature space is learned earlier in the process. Liu, Tan and Zhou [13] noticed that sharing weights after the first stage of ResNet50 is not very grounded and experimented with how many stages should have shared weights.

Wu et al. [21] took another approach angle by converting RGB to grayscale. They appended a zero vector to the grayscale image and prepended a zero vector to the IR image. Both vectors are then used in a CNN to learn a feature vector to re-identify people.

Finally, Hao, Zhao, Ye and Shen [5] didn't supply the modality in the network. Therefore, the network is forced to learn the modality-shared feature space.

In cases where there are important modality-specific cues, a feature compensation learning approach is better as it focuses more on modality-specific cues. It has some interesting approaches.

Zhong et al. [30] used a complicated network resulting in IR features, RGB features and generated coloured features from the IR sample. Zhang, Zhao, Kang and Shen [27] introduced a modality synergy and a modality complement module to respectively learn both modality-shared features and modality-specific features. Huang et al. [9] generated a third modality from IR and RGB that contains information from both modalities. These 3 modalities are then used to classify.

Zhang et al. [27] state the advantage of their approach for person re-identification as follows. The visible features are discriminative enough to a large amount of identities. Infrared tends to capture the thermal features but due to this thermal sensitivity, it loses semantic value and becomes much more background robust. Therefore, infrared images become stable across the same identity and become very robust to noise. Therefore, learning to synergise these modality-specific features brings the strengths of both modalities together.

As a final note, after the addition of IR data, other researchers noted that the addition of video-based samples improves the re-identification process even more [3] [12]. This step is too big for this research but a good direction for improvements on the current methods. In subsection 4.1, this improvement will be taken into account.

2.2 Loss functions

In order to train re-identification models a loss function is required. The objective of this loss function is to keep similar ships together in the feature space and keep dissimilar ships distant in the feature space. Most researches mentioned above use one of the contrastive loss functions [4] [19] [27] [9] [13] [15] [22] [23] [30] [31] and/or a cross-entropy loss [9] [13] [15] [22] [23] [30] [31] [21] to obtain a good feature representation that achieves this goal. Since a loss function is required an overview of the loss functions used in literature is given below.

Triplet loss is a common contrastive loss function for re-identification models. Triplet loss extends on the basic contrastive loss by adding another sample. Instead of having a single pair that can be either similar or dissimilar, triplet loss has 3 samples and 2 pairs. There is an anchor, a positive sample and a negative sample. The anchor-positive pair belong to the same vessel, while the anchor-negative pair are different vessels. The triplet loss function minimises the distance between the anchor and positive sample and maximises the distance between the anchor and negative sample using a distance function. Different distance functions can be used for the triplet loss.

With a as anchor, p as positive sample, n as negative sample and α as margin, the triplet loss is then defined as:

$$L = d(a, p) - d(a, n) + \alpha \tag{1}$$

Triplet learning is popular in re-identification as it trains simultaneously on having similar outputs for similar ships and dissimilar outputs for dissimilar ships. Therefore, a ship that has never been seen by the 'database' of known ships, will have an output dissimilar to any known ship and therefore will be classified as an unknown ship. On the other hand, a ship similar to a known ship that has not been seen by training (and not completely different from any training ship) will have a similar output and therefore will be identified as that known ship.

If random triplets were selected the model would easily start to distinguish between these triplets. E.g. a container ship is easily distinguishable from a speedboat. The model has a harder time when it has to distinguish two different but similar-looking container ships. These harder samples are known as hard triplets which is introduced by Schroff, Kalenichenko and Philbin [17] in the notion of hard triplet mining. Here the positive and negative in the triplet are chosen based on their distance to the anchor. The positive is chosen such that it is the furthest from the anchor and the negative is chosen such that it is the closest to the anchor. From this hard triplet, the model should learn how to distinguish even the hardest samples.

Next to this Schroff et al. [17] suggest applying the following condition to combat the cases where the model converges to an early local minima because the model collapses, i.e. the output of every output neuron becomes nearly 0. The condition is that $d(f(x_a), f(x_p)) < d(f(x_a), f(x_n))$. In other words, it means that the anchor-positive distance should be smaller than the anchor-negative distance.

Next to a contrastive loss function, other researchers also use a cross-entropy loss to obtain a better optimisation [31]. Cross-entropy is usually given as:

$$L = -\log \frac{e^{(W_{y_i}^T x_{y_i} + b_{y_i})}}{\sum_{j=1}^{N} e^{(W_j^T x_i + b_j)}} \tag{2}$$

Deng, Guo, Xue and Zafeiriou [2] noticed that cross-entropy lacked the strength of grouping similar samples together which does happen for contrastive loss functions. Therefore, they created the Arcface loss which has the strength of cross-entropy and contrastive loss functions combined. The formula can be found in Equation 3.

Starting from Equation 2, the bias b is set to 0. The logit $W_j^T x_i$ is transformed to $\|W_j\| \|x_i\| \cos \theta_j$ where $\cos \theta_j$ is the angle between W_j and x_i. This converts the logit to an angular feature. Furthermore, $\|W_j\|$ is scaled to 1 and $\|x_i\|$ is scaled to s using l_2 normalisation. This results in a logit on the hypersphere. Finally, an additive angular margin penalty m is added to result in the following loss function [2]:

$$L = -\log \frac{e^{s\ cos(\theta_{yi} + m)}}{e^{s\ cos(\theta_{yi} + m)} + \sum_{j=1,\ j \neq yi}^{N} e^{s\ cos \theta_j}} \tag{3}$$

Following Zhang et al. [26], in this equation s can be set to the following:

$s = \log(\text{number of classes}) * \sqrt{2}$

By having cross-entropy as a basis, Arcface compares the ground truth label with other label outcomes and adapts the weights such that the boat will be classified as the ground truth boat and not as any other boat. Next to this, Arcface adapts an angular conversion. With this, the function puts all the classes in separate clusters which does not happen for cross-entropy. This strength of clustering the same boat to the same angle also happens in triplet loss.

This summarises the background on triplet, cross entropy and Arcface losses and how they are used to train re-identification models.

3 RESEARCH QUESTION

To conclude the previous section, vessel re-identification has made some good progress. The best results out there involve a global and local feature extraction network like MGN [20] or MVR-Net [4]. Using one of the networks described in the vessel re-identification papers is a good starting point for this research.

However, the use of only RGB data restricts the vessel re-identification task. Some reasoning on what IR adds to the field has been given in the section 2. The following three points summarise and complete the advantages of IR for vessel re-identification:

(1) Samples in low illumination or at night are hard or impossible to classify with only RGB as Wu et al. noted [21]. IR doesn't require light to work and therefore can help the model to also classify these harder samples.
(2) The physical nature of how RGB and IR are captured is different. Therefore, the things that are captured are also different. RGB is captured by filtering wavelengths that correspond to the RGB channels (630, 532 and 465 nanometers). IR captures infrared wavelengths which are from 700 nanometers to 1 millimeter. Specifically, in this research Long Wave IR (LWIR) is used which detects wavelengths between 8-12 micrometers. These wavelengths no longer capture light but capture heat. Due to the difference in what the wavelengths capture, a part of the image can be semantically different when looking at the RGB and IR modalities. Additional information can be gained to increase the performance by including the IR modality. [1]
(3) The use of two modalities can help to reduce the effect of noise and background distractions as noted by Nguyen et al. [14] and Zhang et al. [23].

Figure 2: Some examples of persons from a person re-identification dataset [23]

All these advantages can be used to increase the effectiveness of the model for vessel re-identification. On the other hand, the use of another modality introduces the complexity of another modality. The approaches on how to add this complexity to increase the effectiveness have been studied in the field of person-reidentification for some time, as seen in subsection 2.1. It would be ideal if these approaches could be used for vessel re-identification. However, the field of person re-identification is different from vessel re-identification. The differences in features that are learned are vast. (see Figure 2 and Figure 3 for examples between the fields)

The differences between these modalities are:

(1) A person's form consistently includes a head, a torso, two arms and two legs, so it doesn't vary much between individuals. In contrast, a ship's structure can vary greatly, making it a more important feature. For example, a sailing boat differs significantly from a motor boat or a container ship.

[1] Take note that this advantage can impact the problem very differently in person re-identification versus vessel re-identification. The content of an image becomes very different. So, IR can possibly show much more differences for vessels than persons

Figure 3: Some examples of the images in the boat re-id dataset [19]

(2) The colour of the clothing of a person can be used to identify the person. Ships mostly have the same colour but can be identified by markings, logos or structural characteristics which is more subtle than the colour of the clothing of a person
(3) Ships emit heat in their engine part which might be a feature visible only in the IR modality. Therefore, the IR modality contains specific information not visible in the RGB modality.
(4) If a person rotates, the width isn't drastically changed. If a vessel rotates, the width of the ship in the image changes a lot, especially in cases where the length-to-width ratio of a ship is very high.
(5) Two ships can be of the same model and therefore look exactly the same. The only distinguishment that can be made is the contents on the ship or maybe a visible name on the side of a ship. Persons can look the same but this occurence is less common than two ships being of the same model.
(6) Images of ships are captured in open sea, harbours or canals while images of persons are captured in more controlled settings (e.g. by security cameras in buildings, streets etc.). Therefore, person re-identification usually has a static background while vessel re-identification has a dynamic background which can be distracting.
(7) Persons are captured by cameras that are relatively close range while ships are captured by cameras that are much further away. Therefore, the level of detail for person images is higher while ships have a lesser level of detail.

As the differences are many between the two fields, the approaches cannot be mapped one-to-one. A deeper understanding of the addition of the combination of RGB and IR modalities for vessels is required.

From the above differences, it is at least known that the RGB modality contains useful information. Therefore, the model should at least contain the features from the RGB modality. The addition of IR modality can be done by approaches that embed the modality-shared RGB-IR features or embed the modality-specific RGB and modality-specific IR features.

The research question is then: What specific features of the infrared modality enhance the feature representation for vessel re-identification?

To answer this question, the addition of IR to a network has to be measured. As seen in subsection 2.1, IR can be added by focussing on the modality-specific features or on the modality-shared features. The effect of these features on vessel re-identification is unknown so in order to answer the research question, the modality-specific and modality-shared features need to be analysed individually. Therefore, the research question is divided into two subquestions:

- How much do the modality-specific IR features enhance the feature representation for vessel re-identification?
- How much do the modality-shared RGB-IR features enhance the feature representation for vessel re-identification?

3.1 Evaluation metrics

A good feature representation results in better re-identification. So, by measuring the re-identification task, the strength of the feature representation is measured as a consequence. Therefore, by measuring re-identification task performance the research question can be answered. The task of re-identification defined in Figure 1 can be split into two subtasks. The first subtask is to decide whether the new ship is a previously identified ship or an unseen ship. The second subtask is to decide which of the previously identified ships this new ship is, given that the new ship is a previously identified ship. The first task will be referred to as the known/unknown classification task and the second task will be referred to as the identification task. To measure the known/unknown classification, the Area Under the Precision-Recall Curve (AUPRC) will be used. In the literature, a common metric used to measure identification is the rank-k accuracy which will also be used in this paper. By combining both metrics, the re-identification task is measured.

To determine the AUPRC, the model needs to classify whether some ship is known or unknown by comparing it with the known ships. If the most similar known ship (compared with the new ship) has a higher similarity score than some defined threshold it will be classified as a known ship. By managing stricter or more lenient thresholds, the precision and recall of this binary classification problem change and the precision-recall curve can be determined. The area under this curve determines how well the model predicts known and unknown ships under different thresholds. The implementation of this metric will be further explained in subsection 4.5.

To calculate the rank-k accuracy, a sample from the query set is compared with every sample from the gallery set. The similarities are ranked from most similar to most dissimilar. For every query sample, it is evaluated if the k most similar gallery samples contain the ground truth. The percentage of query samples that contain the ground truth in the top k predictions is the rank-k percentage. The rank-1 metric focuses solely on whether the top prediction is correct, which is analogous to the accuracy metric. The rank-k metric with $k > 1$, also gives insight into how well the model performs beyond the top prediction. This is interesting because once the model is deployed, the database of known boats might contain fewer boats than the gallery set making the prediction easier. This easier prediction resembles the rank-k accuracy with $k > 1$. The rank-k accuracy states how good the model is in matching a new ship with a known ship. So, given that the model knows the ship has already been seen before, it indicates how good the model is in identifying it with the right ship. The AUPRC tells how

Name	no. identities	no. images	background	trajectories/ single images	annotations	publicly available	has IR
Boat Re-ID [19]	107	5K	static	single images	none	yes	no
VesselID-539 [15]	539	149K	dynamic	single images	ship name, hull color, vessel type, orientation	no	no
VR-VCA [4]	729	5K	static	single images	vessel type, orientation	no	no
Zwemer et al.'s dataset [31]	1237	137K	static	trajectories	none	no	no
Zhang et al.'s dataset [25]	1248	31K	dynamic	single images	none	On request	no
VeRis [24]	2904	151K	dynamic	single images	vessel type	no	no
Marvel [11]	2665	25K	dynamic	single images	none	only at Thales	no
Own created dataset	70	5K	dynamic	trajectories	none	only at Thales	yes

Table 1: Comparison between different vessel re-identification datasets. A K means thousand, e.g. 151K = 151.000. A static background means that the camera that took the images is stationary and the background is mostly the same. A dynamic background could be any background. Trajectories mean that the images came from a video and sequential images can be interpreted as a video. Annotations are extra information for each sample that can be used for re-identification. The last dataset is the dataset that is curated during this research to solve the absence of an RGB-IR dataset.

good the model is in classifying known versus unknown ships. A higher score gives better discriminative thresholds and decreases the chances that an unknown ship is classified as known or a known ship is classified as unknown.

4 APPROACH

Before the proposed models can be described, which can be evaluated to answer the research subquestions, the availability of datasets and the new dataset creation need to be discussed.

4.1 New dataset creation

An overview of vessel re-identification datasets is shown in Table 1. The last dataset in the table is created during this research and the other datasets stem from literature papers.

As can be seen, from all the datasets used in the literature, only one dataset is publicly available. Furthermore, none of them have IR samples in their datasets. This creates another challenge for the research, namely that there are almost no datasets available and none have IR samples.

To solve this problem a new dataset has to be created. Thales already has available video data that can be used to create such a dataset. This video data has timestamps and RGB and IR cameras directed to the same view. Therefore, from the video data, paired RGB and IR images can be extracted. The only part required is to label the videos.

In order to label this data, an automatic pipeline is created. This approach is also used in literature for example by Du et al. [3], Zheng et al. [28], Zheng et al. [29] and Zwemer et al. [31]. The idea is to use an object detection module followed by a labelling module. The object detection module is the YOLOv8 network which can detect 80 different objects including boats [16]. The labelling module then takes detected objects from successive frames in a video that are close to the same position and gives the objects the same label.

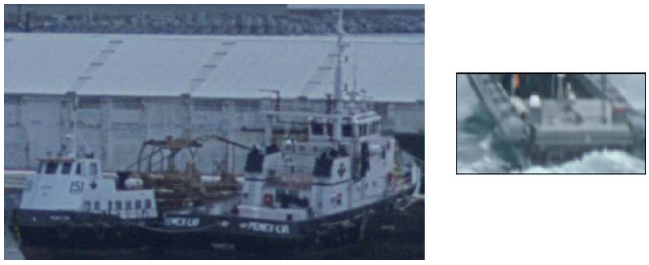

Figure 4: Two examples of incorrectly cropped images by the YOLO network

This approach has the advantage over manual labelling in that labelling goes mostly automatically and a much larger dataset can be created.

However, there are some technical difficulties arising from using this method. The YOLO network might misclassify a lot of objects. With some semantic knowledge and the labelling module, some issues can be alleviated (e.g. by only saving a detected object if it still is detected after 1.25 seconds which removes incidental random objects that are detected). Still, the resulting dataset contains a lot of non-ship samples, bad-cropped samples and different sequences of the same ship.

To fix the first and third issues, manual filtering needs to be applied. Due to the fact that a detected object needs to be at the same position to get the same label, it is assumed that a sequence can only contain one object. Therefore, if some images in a sequence of objects are not a ship, the whole sequence can be removed. Afterwards, every sequence of ships can be inspected. If two sequences are from the same ship, these sequences can be put under the same label. The resulting dataset still has the shortcoming that a ship can be badly cropped. Two examples can be seen in Figure 4

Another issue is that YOLO also cannot classify in low or no illumination. As a consequence, all the resulting samples in the dataset will have normal illumination and the performance difference of RGB-IR against RGB-only in low illumination will not be assessed in this paper.

However, if it is assumed that in no illumination every pixel value of a RGB image is 0, a RGB model cannot extract any meaningful features and only adds noise to the model. Therefore, it is better to use an IR-only model. This can be simulated by only supplying IR images and not any RGB images[2]. In this research, the networks are also assessed with IR-only input. These results can then also be reflected to the cases where there is a setting with no illumination.

At the end of section 2, the addition of video-based samples is briefly touched upon. When automatically creating this dataset, images are stored in a sequence. This allows the use of video-based techniques instead of image-based techniques for this dataset. As noted at the end of subsection 2.1 this should improve the performance of the model. To not overcomplicate this research, the models will use single images. This is considered a limitation of this approach as video-based techniques are better.

With the newly created dataset, the networks can be trained and evaluated. The networks will be described in the upcoming subsections.

4.2 Modality-specific feature learning network

To answer the first research sub-question, a network is proposed that will test the improvement gained from modality-specific IR features. In order to do this, a network is designed that learns both RGB and IR features in separate branches. This is done because the branches do not depend on each other and can go their own way in learning their best representation. The first branch learns the best representation for RGB and the second branch learns the best representation for IR. Both output feature vectors are then concatenated. Combined they produce a feature vector of modality-specific RGB and modality-specific IR features. Take note that modality-shared features are still present in this network but since both branches can embed these features, they can be embedded twice in the resulting feature vector and therefore become disadvantageous for the representation strength since it contains redundant information.

Figure 5a shows a simple representation of the proposed model.

4.2.1 RGB and IR only. In order to compare the added value of the other modality, the networks also need to be trained and evaluated on a specific modality alone. Therefore, a RGB-only and an IR-only approach is also required. Figure 5b and 5c show a simple overview of the network for the modality-only approaches. When these modality-only approaches are evaluated and a decrease in the performance metrics is measured it indicates that the other modality adds valuable information to the network. However, if an increase in the performance metrics is measured, it means that the other modality distracts from the problem and adds more noise than meaningful information to the resulting feature vector.

[2]Technically speaking, there is a small semantical difference in the IR modality during the night. Since the night is cooler, objects are cooler as well and the IR sensor should detect lower values than during the day. For this research, this is not taken into account.

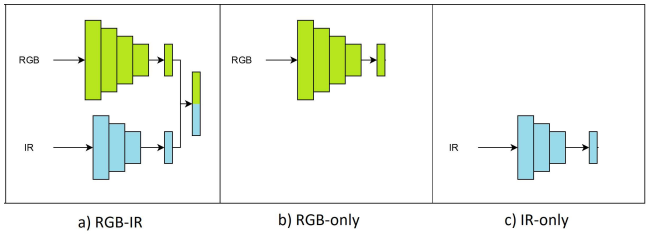

Figure 5: The proposed system for a modality-specific learning network. In a) the model uses both RGB and IR input, in b) only RGB input is used and in c) only IR input is used.

In order to make these modality-only networks, the other modality branch is removed and the output of the remaining branch is also used as the output of the network.

4.2.2 RGB and IR branch. For the RGB branch, a ResNet50 model[6] initialised with pre-trained weights is used as a starting point. The advantage of having pre-trained weights is that it has a good backing knowledge of objects from the ImageNet dataset [1]. By fine-tuning this knowledge (i.e. training only on the last layers of ResNet), it keeps the general knowledge while also learning the fine details of ships in the area of vessel re-identification. The ResNet is configured and fine-tuned differently depending on the hyperparameters which are discussed in subsection 4.6 (output sizes, number of fully connected layers etc.).

Since the research question is about the difference between RGB-only and RGB-IR it is not necessary to use the best RGB network as long as the RGB network is consistent such that a comparison can be made. The literature showed that there are better options than the ResNet50 model but to keep the approach simple, these more complex options will not be used.

For the IR branch, four options are proposed. The first option is a pre-trained network while the other three options are untrained networks.

The pre-trained network is trained on the ImageNet dataset [1]. Therefore, it has the advantage that it is pre-trained on a lot of data and therefore has a good general basis for all kinds of objects. The disadvantage is that this basis is based on the RGB feature space and not on the IR feature space. IR-specific cues can therefore be disregarded as unimportant due to that the cues are not existent in the RGB space.

The other untrained IR networks have the advantage that the model trains on IR from scratch. Therefore, the model can capture all IR-specific features as long as it has the capacity to capture the complexity. The disadvantage is that in order to train a model from scratch, a good dataset needs to be available. Having not enough data results in a less robust model when doing re-identification. As seen in Table 1 our dataset has 5K images which is a reasonable size but might not be enough to learn as much as a pre-trained ResNet.

All these options will be implemented and evaluated. For every option, some fully connected (FC) layers are added in order to learn the complexity of the output. This is configured according to the hyperparameters which will be further discussed in subsection 4.6.

Layer name	Output size	Convolutions
Conv1	48x48	7x7, 16, stride=2
Stage1	24x24	$\begin{pmatrix}1\text{x}1\ 8\\3\text{x}3\ 8\\1\text{x}1\ 32\end{pmatrix}$ 4x
Stage2	12x12	$\begin{pmatrix}1\text{x}1\ 16\\3\text{x}3\ 16\\1\text{x}1\ 64\end{pmatrix}$ 4x
Stage3	6x6	$\begin{pmatrix}1\text{x}1\ 32\\3\text{x}3\ 32\\1\text{x}1\ 128\end{pmatrix}$ 4x
Average pool, flatten	128	

Table 2: Stages in the ResIRNet. A convolution of kxk, d means a kernel of size k and output feature map depth of d. The first 3x3 convolution layer of each stage has a stride of 2 and halves the output size. The rest of the convolution layers in the stage have a stride of 1. At the end of each bottleneck block, the residual part is added to the vector just like in ResNet (see also Figure 6)

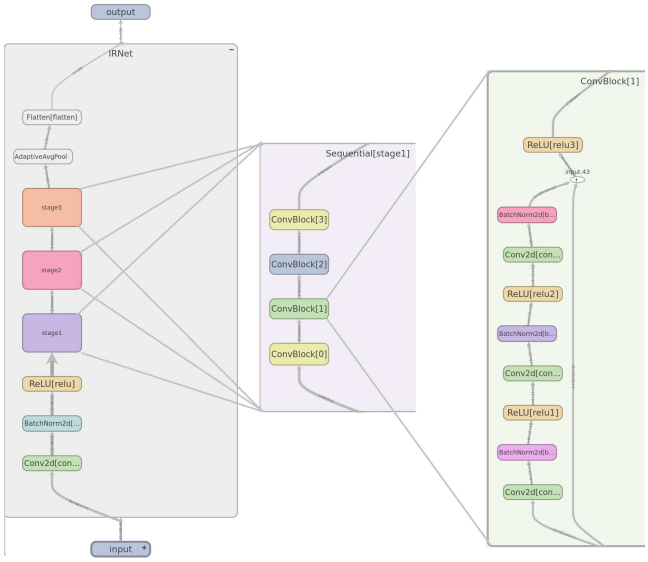

Figure 6: Overview of stages in the ResIRNet. On the left, the complete network is shown. In the middle, a stage is shown. Note that the 3 stages are the same apart from the input and output dimensions. On the right, a convolution block is shown. The first convolutional block in a stage has a downsample in the residual part before the residual and convoluted parts are added together which is not shown here. Apart from that the convolutional blocks are the same. After the network, some fully connected layers are added depending on the hyperparameters which is not shown in this figure. The exact input and output shapes of each layer can be found in Table 2

4.2.3 Pre-trained ResNet. The first option for the IR network, is to use a pre-trained ResNet50 model and fine-tune the network to learn the IR features instead of RGB features.

4.2.4 Untrained ResNet. The second option for the IR network, is to take an untrained ResNet50 network with 1 channel input instead of 3 channel inputs by changing the first convolutional layer and keeping the rest of the architecture the same. In essence, this is the previous network without pre-trained weights.

4.2.5 New network: ResIRNet. For the third option, a new IR network is created from scratch. The first reason behind this is that ResNet requires 3 input channels while IR only has 1 channel. So, creating a network with one input channel instead of three removes the initial redundancy of additional channels. The second reason is that the input IR images in the dataset have a lower resolution than the RGB images. The ResNet50 network is trained and evaluated on an input resolution of 224x224. To achieve this resolution some upsampling is always required for the IR images. A lot of pixels are therefore reused and a lot of information becomes duplicated. To alleviate the redundant information, a network is designed with an input resolution of 96x96 and a similar output width and height resolution. Due to the lower amount of input channels and lower resolution, the output size is chosen to be 8 times lower than the ResNet50 (128 output neurons). The network will be called ResIRNet. The complete architecture is found in Figure 6 and Table 2.

For the model, an architecture similar to ResNet is used. Just like ResNet, the network starts with 7x7 convolution followed by batch normalisation and relu activation. The 7x7 convolution is applied for a quick increase of the receptive field while the input channel dimension is still one[3]. Note that the 2x2 maxpool layer is not present in ResIRNet but the first convolutional block of the first stage does have a stride of 2 instead of 1. This changes the layer from a simple downsample with a pooling layer to a more complex convolution layer allowing for a bit more complexity to be learned.

Afterwards, some stages with bottleneck blocks are used which are also present in ResNet50. Four bottleneck blocks are combined in a stage where the first block halves the input size and doubles the feature map depth. The other 3 blocks keep the width and height dimensions. The role of the first convolutional block is to learn to convolute the matrix to a lower width and height and higher depth. The role of the other identity blocks is to learn to adapt the input slightly for a better feature space representation. The choice for a total of 4 blocks is because the model then has some capacity to learn complexity. Using only 1 or 2 convolutional blocks results in too few parameters that can be tuned which results in a less complex feature representation. If the model doesn't have enough complexity it might not be able to distinguish between similar boats.

Three stages are used to end up at a matrix of 128x6x6 (ResNet50 ends up with 2048x7x7). Take note, that 3 stages are used while

[3]As an alternative, applying 3 times a 3x3 convolution would give the same receptive field. However, this would result in more FLOPs and learnable parameters which is not beneficial. The 7x7 convolution has $7 * 7 * 1 * 16 = 784$ learnable parameters and the 3 times 3x3 convolutions have $3 * 3 * 1 * 16 + 2 * (3 * 3 * 16 * 16) = 4752$ learnable parameters. The 7x7 convolution has $784 * H_{out} * W_{out}$ FLOPs and the 3 times 3x3 convolutions has $4752 * H_{out} * W_{out}$ FLOPs (where H_{out} and W_{out} are the height and width of the output). Note that there is no bias in this layer just like in ResNet.

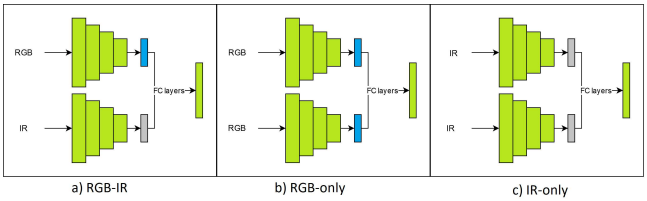

Figure 7: The modality-shared network. In a) the model uses both RGB and IR input, in b) only RGB input is used and in c) only IR input is used. The green model for both branches means that this is the same model for both branches. The double blue and double grey bars in b and c mean two times the exact same feature vector (as the input and the model are the same for both branches)

ResNet uses 4 stages. This is due to the lower input resolution. Within 3 stages the ResIRNet reaches an output size of 6x6 which is similar to the 7x7 output size of ResNet. Finally, an average pooling layer and flattening layer similar to ResNet are used to end up with a feature vector of 128 values.

4.2.6 Untrained MobileNet. The ResIRNet has far fewer parameters and FLOPs than the ResNet50. This is due to lower input resolution and lower depth of the feature maps in the model. Due to this difference in the number of parameters, another network is added as an option to compare the performance difference between the ResIRNet and a network of equal size.

For this fourth option, the MobileNetV3_small is chosen [7]. The MobileNet networks [8] are smaller networks made to be able to run on machines with low computational power (e.g. mobile phones). For MobileNetV3_small (from now on MobileNet), an input resolution of 96x96 is used. Since MobileNet usually operates on input resolutions of 224x224, this changes the global average pooling layer functionality. Instead of pooling a 7x7 matrix, it pools a 3x3 matrix resulting in a less dense pooling operation.

The MobileNet and ResIRNet are both trained from scratch. The networks can be pre-trained with ImageNet data, but to limit the approach scope, this will not be done. It is expected that a pre-trained ResNet will have better performance metrics than a pre-trained ResIRNet or pre-trained MobileNet due to the ResNet being bigger and therefore having the capacity to capture more complex features.

4.3 Modality-shared feature learning network

For the second sub-research question, a network is required that learns the modality-shared features. This can be achieved by using one model for both modalities. Then the model is forced to learn both modalities together. A simple model overview is given in Figure 7a.

The model will consist of a single ResNet50 network with pre-trained weights initialised. Both RGB and IR input will flow through this model with the same shared weights. Consequently, the weights are then also trained on both RGB and IR together. Both outputs are then merged using some FC layers which are dependent on the hyperparameters which are described in subsection 4.6. In order to adapt IR from 1 channel to 3 channels, the channel contents are copied to all three channels. In the end, the network learns to embed both RGB and IR in the same feature space.

In order to assess the effect of adding another modality to this network, the RGB-only and IR-only approach is again applied to measure the differences. In order for the network to accommodate one modality only, the input has to be copied to both branches. This is shown in Figure 7b and 7c. For a more efficient network, the output of ResNet which is the input of the FC layers, is copied twice instead of having two times the same input for the same model.

4.4 Used loss functions

The proposed models have to be trained with a loss function. Each model described above will be trained with a triplet loss function or an Arcface loss function as described in subsection 2.2.

The triplet loss requires a distance function. Since the similarity score for re-identification is also determined using cosine similarity, the loss function uses cosine distance as distance function. Cosine similarity is given as:

$$\text{cosine similarity}(x, y) = \frac{x \cdot y}{\|x\| \times \|y\|} \tag{4}$$

Furthermore, the margin α is set to 1. The triplet loss with cosine distance as used in this paper is then given as:

$$\begin{aligned} L &= d(a, p) - d(a, n) + 1 \\ d(x, y) &= 1 - \text{cosine similarity}(x, y) \end{aligned} \tag{5}$$

Hard triplet mining as described in subsection 2.2 by Schroff et al. [17] will be used to select better triplets. Furthermore, Schroff et al. simplify the process by picking every possible anchor-positive pair within the mini-batch. In this paper, this is adapted to picking a random positive for each anchor (if available).

The Arcface loss function (found in Equation 3) also requires a margin which is set as a hyperparameter which is further described in subsection 4.6. The s in the loss function is set to log(number of classes) $*\sqrt{2}$ following Zhang et al. [26].

4.5 Measuring known/unknown ship classification

To measure the known/unknown classification, the AUPRC metric is used. To do this, the problem is considered as a binary classification problem. Something is either a known ship or an unknown ship. In the gallery set, 50% of the boats are removed. If a sample in the query set still has the corresponding boat in the gallery set, then it is a known boat. Otherwise, it is an unknown boat. The known boats are positives and the unknown are negatives. Furthermore, the query set is sampled such that the number of known and unknown boat images are both 200 such that the known/unknown boats have an equal distribution. It is out of the scope of this research to analyse how the PR curve and AUPRC adapt when the gallery and query sets contain different ratios of known and unknown boats.

A sample in the query set can then be compared with all the known boats in the gallery set using cosine similarity. The highest cosine similarity is then used as a comparison value to assess whether the sample is a known or unknown ship. Given a certain threshold, some ships will be classified as positive and others as negative. With this binary classification, the precision and recall

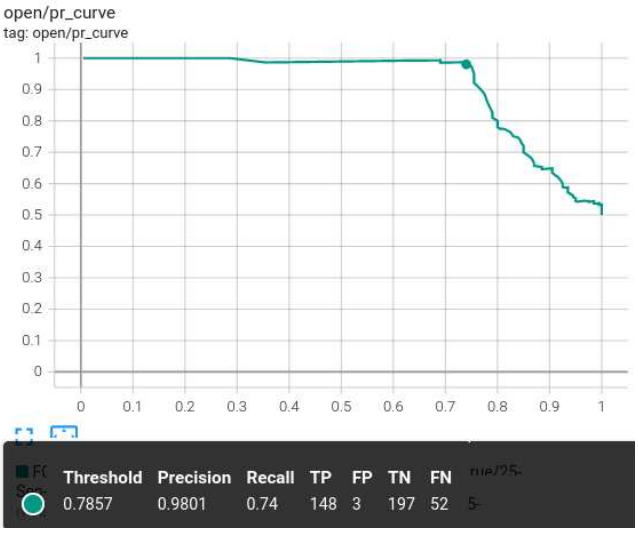

Figure 8: An example of a precision-recall curve. At the point in the graph, the exact threshold, precision and recall are given. Given some specifications on how good the precision and recall should be, a threshold can be defined for some model

can be calculated. Setting a higher threshold will yield more predicted unknown boats, while a lower threshold will yield more predicted known boats. By using the threshold as a variable, a precision-recall curve can be defined. This curve can give insight into how the threshold changes the precision and recall. With specific precision and/or recall requirements, an exact threshold can be determined using this curve. An example of the precision-recall curve can be seen in Figure 8.

The area under the precision-recall curve (AUPRC) gives an indication of how well the model generally performs under different thresholds. An area of 1 would mean that there is a threshold that perfectly classifies the unknown and known ships. An area around 0.5 would indicate that the precision stays around 0.5 while the recall increases i.e. the model classifies randomly between the two classes.

4.6 Hyperparameters, setup and best model selection

A set of hyperparameters are tuned to gather the best results for re-identification. A grid search is applied for the applicable parameters to exhaust every possible option. The reason behind grid search is because the options per hyperparameter are not extensive (either 2 or 3 options) so smart searching within these few options is not very helpful. Furthermore, a grid search results in a complete set of results and a good hyperparameter set is never accidentally skipped.

4.6.1 General hyperparameters. For every model, the following hyperparameters are used:

- The loss function: Arcface or triplet loss
- In case of Arcface, the margin m in Equation 3: 0.05 or 0.2. According to Zhang et al. [26], a too large m (e.g. 1) won't allow the model to converge. Therefore, these much smaller margins are chosen.

4.6.2 Pre-trained Resnet hyperparameters. When RGB or RGB-IR input is used, the ResNet50 model introduces the following hyperparameters:

- The number of output neurons at the end of the ResNet50 model: 128, 256 or 2048 output neurons. Having too many neurons in the feature vector makes some features redundant which is bad for the distinctiveness of the feature vector and therefore makes re-identification worse. Removing too many neurons removes too much present generality in the pre-trained network and makes the network too biased towards the trained boats. It keeps the information that can still distinguish the boats in the train set but it loses information that can distinguish boats outside the train set. Therefore, some numbers are chosen that retain the information (2048) or compress the information (128 and 256).
- After the Resnet50 the amount of fully connected layers that are appended: 1 or 2. More layers result in a more complex understanding of the output features of ResNet but could also lead to a less general interpretation of the output features.
- The amount of unfrozen ResNet stages: 1 or 2 stages. The trainable parameters in the pre-trained Resnet50 model are initially frozen (untrainable). As a hyperparameter, the last stages can be unfrozen (made trainable) in order to fine-tune the Resnet50 model on the problem. Shermin et al. [18] noted that unfreezing the final few layers is beneficial for fine-tuning but unfreezing too many layers results in performance loss.

4.6.3 Modality-specific IR network hyperparameters. In the case of the modality-specific system, there are different IR networks which bring their own hyperparameters:

- In the case of a ResNet50, the hyperparameters that are used for the RGB ResNet model are also used for the IR ResNet model (number of output neurons, FC layers and unfrozen stages). This is done to reduce the amount of possibilities for the grid search.
- In the case of a ResNet50, the IR network can be initialised with pre-trained weights or not. If the network is randomly initialised nothing is frozen.
- In the case of ResIRNet, the number of output neurons at the end of the IR network can be 64 or 128. (The output size before the FC layers is 128)
- In the case of MobileNet, the number of output neurons at the end of the IR network can be 128 or 256 (The output size before the FC layers is 576)
- In the case of ResIRNet and MobileNet, the amount of FC layers after the last convolutional block: 1 or 2. For fewer possibilities in the grid search, the RGB ResNet50 network has the same amount of FC layers.

The experiments are performed on a Nvidia RTX A5000. The batch size is 32 for Arcface and 128 for triplet loss. The higher batch size for triplet loss is due to the hard triplet mining. The hard triplets are searched within the batch. That means there has to be room for positive pairs. There are 53 ships in the train set (which

will be further described in subsection 5.1). With 53 ships and 128 samples, there is room enough for positive pairs. Adam [10] is used as optimiser with hyperparameters learning rate=0.001, β_1=0.9 and β_2=0.999. The models are trained for max 100 epochs.

To check the model's performance, the gallery and query set are evaluated for the rank-k accuracies and AUPRC metrics after each epoch. The performance of the model is then measured by the following formula:

$$\text{argmax}_\theta \begin{pmatrix} \text{rank-1}(\theta, G_k, Q_k) + \dfrac{\text{rank-3}(\theta, G_k, Q_k)}{3} + \\ \text{AUPRC}(\theta, G_u, Q_u) \end{pmatrix} \quad (6)$$

Where θ are the model parameters, G_k and Q_k are the gallery and query sets where all boat IDs in the query set are known (i.e. in the gallery set) and G_u and Q_u are the gallery and query sets where half of the boat IDs from the query set are unknown (i.e. not in the gallery set).

Empirically, the 'good' rank-3 accuracy lies between 0.9 and 1.0 while the 'good' rank-1 accuracy lies between 0.75 and 0.95. Logically, this is due to rank-3 accuracy being easier to optimise than rank-1 accuracy. Therefore, the rank-3 accuracy is divided by 3 in Equation 6 to make it less important in selecting the best model.

This formula optimises the rank-k accuracies and PR curve. By optimising this, it optimises both matching the input with the correct ship and recognising whether the input is a known or unknown ship. If more training results in a higher score, the model was still underfitted. If more training doesn't result in higher scores for many epochs, the model is probably overfitting on the training set.

If a model doesn't improve for 20 epochs according to Equation 6, the training is early stopped and the model's performance of 20 epochs ago is selected as the best model. [4]

Since the gallery set G_k contains 17 boats, a high rank-k accuracy will be meaningless. A rank-3 accuracy is almost always above 90% in the best performances of each hyperparameter combination. A higher k will give even higher percentages up to a point where the difference between two models becomes marginal. Therefore, only the rank-1 and rank-3 accuracies will be used in selecting the best model.

5 RESULTS

In this section, the resulting dataset and usage of this dataset are described. Furthermore, the performance metrics of the models and the computational metrics are presented. Finally, given these metrics, an ablation study is performed on the gathered results.

5.1 Resulting dataset

The dataset is constructed by applying the detection module and labelling module described in subsection 4.1, to recordings from the Gatekeeper of Thales (see section 1). These labelled objects are then filtered by removing sequences of objects that are not ships. Afterwards, it is filtered by merging sequences of ships that are the same. Remember that individual images within sequences are not considered while filtering. Therefore, it can happen that a ship is badly cropped as shown in Figure 4.

[4]The use of early stopping could result in underfitting models which is less ideal.

The initial filtered dataset contains 111 different ships. However, some ships only have 1 or 2 images making the ships not very useful for re-identification. Furthermore, there are some images of ships that are far away and have a low resolution in the IR modality (number of pixels in the length $*$ number pixels in the width < 500). Therefore, these ships are also left out of the dataset. This results in 70 different ships with 5169 images.

In order to train and evaluate the models the boats will be split into seen and unseen data sets. This split will be made on the IDs of the boats, e.g. a boat with samples in the seen set will not have samples in the unseen set. The split will be 75% of the IDs in the seen set and 25% of the IDs in the unseen set. The seen set contains 53 ships and 3816 images and the unseen set contains 17 ships and 1353 images.

The seen set will be split into a training and validation set with a 90/10% split. The samples within a boat identity are very similar because every 1.25 seconds in a recording, a cropped image is extracted (as described in subsection 4.1). Due to this lack of diversity, the validation set can have less data and keep the same validation power.

Due to the classification nature of the Arcface loss, when this loss function is used, the model becomes biased on the validation data because these boat IDs have been seen before. Though the validation samples themselves are not seen before, the training samples only have slightly different viewpoint angles, lightning, background changes etc. than the validation samples . Due to this bias, the model is very good in re-identifying the boats from the validation set[5]. Due to this strong bias, the validation set can not be used for the gallery and query sets. Otherwise, the metrics become biased and less meaningful. To keep the gallery and query set clean, only the unseen boats are used which are unbiased since they are unseen by training.

The samples in the unseen set will be split into the query set and gallery set such that samples in the query set can be matched with the samples in the gallery set to gain a rank-k accuracy (i.e. G_k and Q_k in Equation 6). Next to this, a second gallery set is created where not every ID is available (i.e. G_u in Equation 6). With this second gallery set and by sampling the query set (to obtain Q_u as described in subsection 4.5) the AUPRC metric can be determined. In a deployed re-identification system, the time since the ship was seen last seen would decide whether a ship is known (i.e. in the gallery set). However, to make the gallery set and query set challenging enough, the filtering on the time aspect will not be taken into account.

The whole unseen set has been manually analysed if for every boat a boat of the same type exists in a training set. For example, if a container ship is in the unseen set but there are no other container ships in the training set, the model has a hard time re-identifying container ships. This is not the case, so this effect doesn't need to be taken into account.

[5]This can also seen in the metrics when these validation boats are re-identified with extremely high accuracies. Furthermore, this is seen in the weaker correlation between validation loss and re-identification accuracies (rank-k and AUPRC) on the unseen set. These statistics will not be shown further on in this research.

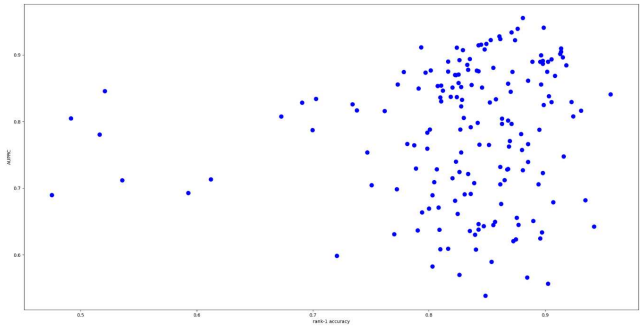

Figure 9: The rank-1 accuracy (y-axis) plotted against the AUPRC (x-axis). Every point represents a model run with a different set of hyperparameters or a different approach. As can be seen, the data seems to be uncorrelated as most points seem to be grouped in a vertical bar. The Pearson correlation coefficient of this plot is 0.06.

5.2 Performance metrics

Before the performance metric results are presented, the correlation between rank-k and AUPRC will be analysed. With this, the modality-specific approach as described in subsection 4.2 is evaluated. Within this approach, the different IR networks are also compared. Finally, the modality-shared approach as described in subsection 4.3 is evaluated.

5.2.1 Correlation between rank-k accuracy and AUPRC. Evaluating the models only on the combined score from Equation 6 is simple and gives a good overview of how the model performs generally. However, the metrics correspond to different tasks that the model performs. In Figure 9, a plot is given of every best result of each hyperparameter and model architecture combination where the AUPRC is plotted against the rank-1 accuracy. The plot and the Pearson correlation coefficient of 0.06, show that the metric results gained are not very correlated. Good discriminative thresholds do not indicate good rank-k accuracies and vice versa. Therefore, in order to show a more complete picture, the rank-k accuracies and AUPRC are also separated in the upcoming results. When separated, the best results are chosen based on only the respective metrics, either the rank-k accuracies or the AUPRC.

5.2.2 Comparison between modality-specific approach. First, the RGB-only, IR-only and RGB-IR approaches of the modality-specific approach are compared.

The results can be found in Table 3. The first thing that pops out is the clear improvement of RGB-IR compared to RGB-only and IR-only. For both the combined and individual metric scores, the RGB-IR performs better. This performance increase is seen in an increase of 4.5% rank-k score which indicates around 4.5% rank-1 accuracy increase, an increase of 0.03 AUPRC and an increase of 0.023 combined score.

Another interesting point is when RGB-only and IR-only are compared next to each other, the RGB performs better. This confirms the claim in section 3 that RGB contains useful information, and if available, the RGB modality should be present in the model as it is more discriminative than IR.

	Combined score	Rank-k score	AUPRC score
RGB-IR	**2.163 (T, R)**	**1.286 (A, N)**	**0.955 (T, R)**
RGB-only	2.140 (T)	1.241 (T)	0.924 (T)
IR-only	1.857 (A, R)	1.055 (A, R)	0.888 (A, M)

Table 3: The performances of the modality-specific networks. The combined score is both rank-k and AUPRC combined while the other two columns consider these metrics individually. The letters after the score indicate the loss function and IR model used to obtain this best result (A=arcface, T=triplet, R=pre-trained ResNet, N=ResIRNet, M=MobileNet)

5.2.3 Comparison between IR networks. In subsubsection 5.2.2, the best IR networks differ per table cell though the pre-trained ResNet seems to have the best models for most results. To properly find out which IR network has better performance metrics, the metrics are compared when different IR networks are used. Since the results above showed that RGB-IR is better, only those results are presented. The results are shown in Table 4.

	Combined score	Rank-k score	AUPRC score
Pre-trained ResNet	**2.163 (T)**	1.259 (A)	**0.955 (T)**
Untrained ResNet	2.064 (A)	1.216 (A)	0.907 (A)
Untrained ResIRNet	**2.162 (A)**	**1.286 (A)**	0.941 (A)
Untrained MobileNet	2.140 (A)	1.252 (A)	0.939 (A)

Table 4: The modality-specific IR network performances compared next to each other. All performances are measured with RGB-IR enabled. The letter after the score indicates the loss function used that resulted in this best result (A=arcface, T=triplet).

The results are simple, the pre-trained ResNet and the ResIRNet are the better performers for re-identification. Untrained ResNet lags behind, possibly due to the lack of enough IR data and the corresponding bigger network which requires more data to learn more complex feature representations.

The individual rank-k and AUPRC metrics do not present additional insights.

5.2.4 Comparison modality-shared approach. Now for the second approach that focuses on modality-shared features. The results can be seen in Table 5.

RGB-IR now performs worse than RGB-only. This is interesting. Apparently, the addition of IR in the modality-shared network next to the RGB distracts the model from re-identification. It is better for a model to focus only on the RGB features instead of a good focus on the RGB-IR modality-shared features.

	Combined score	Rank-k score	AUPRC score
RGB-IR	2.115 (A)	1.232 (A)	0.899 (A)
RGB-only	**2.161 (T)**	**1.253 (A)**	**0.946 (A)**
IR-only	1.990 (A)	1.132 (A)	0.905 (A)

Table 5: The performances of the modality-specific networks. The combined score is both rank-k and AUPRC combined while the other two columns consider these metrics individually. The letter after the score indicates the loss function used that resulted in this best result (A=arcface, T=triplet).

The individual rank-k and AUPRC metrics do not present additional insights.

5.3 Computational metrics

Different models have different sizes. These sizes have a high impact on speed and memory usage. Therefore, the computational metrics are also measured to compare the options. For this, the amount of FLoating point OPerations (FLOPs) and the number of parameters in the model are measured. FLOPs indicate how many operations are needed to do one forward pass through the network and therefore measure the speed of the network. The number of parameters indicates the amount of memory required to store the network. Together, these metrics indicate how fast and how big a model is. The metrics can be seen in Table 6

	FLOPs	Parameters
ResIRNet	22.33M	136.43K
MobileNetV3_small	21.81M	1.00M
ResNet50	8.21G	22.33M
ResNet50 + ResIRNet	8.23G	22.71M
ResNet50 + MobileNetV3_small	8.23G	23.33M
ResNet50 + ResNet50	16.26G	45.18M
Modality-shared ResNet RGB/IR-only	8.21G	23.11M
Modality-shared ResNet	16.43G	30.45M

Table 6: Comparison of the number of FLOPs (computational cost) and number of parameters (memory cost) between different models. In this table, K means times 10^3, M means times 10^6 and G means times 10^9. All entries belong to the modality-specific approach apart from the two entries that specifically state modality-shared. A model can be used in different settings. For example, the ResNet50 is used as a pre-trained RGB-only network but also as a pre-trained and untrained IR-only network. These numbers are slightly dependent on the amount of FC layers and output sizes of a network which are hyperparameters. These hyperparameters are selected based on the corresponding best models in the previous section. Due to these hyperparameters, the results might differ from what is in the papers [6] [7]

As expected the ResIRNet and MobileNet are much smaller than the ResNet50 in terms of FLOPs and parameters. ResNet has at least 350 times more FLOPs than ResIRNet and MobileNet making it a much slower model. The number of parameters for ResNet is also 22 times bigger than MobileNet and around 160 times bigger than ResIRNet.

However, as seen in the previous section, the IR-only approach (ResIRNet and MobileNet in Table 6) is missing important features compared to the RGB-only or RGB-IR approach. Therefore, to get some good performance results, a ResNet50 + IR network is required.

So, for the RGB-IR approach, a ResNet is always added. The choice of IR network impacts the FLOPs and parameters as well. ResNet is a big network compared to the other networks. The IR network is then either just as big as the RGB network (ResNet) or has negligible size compared to the RGB network. Therefore, the impact of the IR network is only seen in the double amount of FLOPs and parameters when the ResNet + ResNet is used compared to ResNet + ResIRNet/MobileNet.

It is out of scope to study different RGB models as stated in subsubsection 4.2.2. But at the cost of some accuracy, the ResNet can be swapped out for a smaller RGB network (like a pre-trained MobileNet). If speed and/or memory are critical, this can be considered as an alternative to ResNet. The other way around is also possible as other researches have shown that networks with global and local feature extraction obtain higher accuracies. Because these networks have more branches, it follows that the speed is even slower and the model size is even higher.

5.4 Ablation study

To quickly summarise the previous result sections, an increase in performance is gained when RGB and IR are both used in a modality-specific network. That means that given the correct network for the IR modality, the feature vector representation is strengthened which results in better re-identification.

The best performance for the modality-specific learning approach comes from using the ResNet as an IR network which is logical as it is the biggest and therefore can capture the most complexity. However, an almost similar performance is gained when the ResIRNet as an IR network is used. This indicates that by removing the unnecessary 3-channel input and one stage from the model, the performance is nearly the same. All the while by removing these redundancies, the model size is decreased and the inference speed is increased. Therefore, the only case where the pre-trained ResNet is better as an IR network is when computational metrics are of no importance. Otherwise, the speed advantage gained from using ResIRNet is very beneficial at the cost of nearly no performance. This makes the ResIRNet a good choice. Furthermore, this finding paves the way for designing better IR networks. Since the use of 1-channel inputs removes additional redundancies compared to 3-channel inputs this becomes a useful insight for designing IR networks.

There are two limitations in the untrained IR networks compared to the pre-trained network. First, in subsubsection 4.2.2 the discrepancy in the dataset size was identified. Therefore, this approach of untrained ResIRNet might have suffered from insufficient amounts of data. The scope of this research didn't allow to analyse if the dataset is big enough to learn sufficient unseen IR representations. Therefore, this research is limited in this knowledge. However, other researches (see Table 1) used datasets with 25-30 times more

images which indicates there might be a discrepancy in the amount of data in the dataset. If the dataset can indeed be improved by more IR data, it would indicate that the RGB-IR approaches can learn even more.

Second, there is the limitation of lower IR resolution in the dataset as identified in subsubsection 4.2.5. If a dataset is used with IR resolutions similar to the RGB input resolutions, the input resolution of ResIRNet can be the same as ResNet. The removed stage of ResIRNet makes less sense then, which follows that a bigger ResIRNet can be created with more convolutions and therefore more potential complexity. If this is exploited, IR might result in even more added value than what was already shown here at the cost of some more FLOPs and parameters.

When the modality-shared learning approach is analysed, it was seen that RGB and IR combined resulted in lesser performance than RGB-only. That means that modality-shared features are already very present in the RGB branch. The modality-specific RGB features are cancelled out too much and the strengthened modality-shared RGB-IR features do not compensate for this. That is why the RGB-only approach performs better due to the better access to the modality-specific RGB features. So, focusing on the modality-shared features is not beneficial.

The state-of-the-art approaches were missing the use of infrared as a modality. This research shows that by focusing on the modality-specific IR features, the feature representation strength is increased. So, by incorporating this in the state-of-the-art approaches, the feature vectors are even better and a stronger vessel re-identification model can be created.

6 DISCUSSION

The research question was divided into two parts which analysed the contribution of respectively modality-specific features and modality-shared features of IR. In order to answer these subquestions, two architectures were designed that measured the strength of respectively modality-specific features and modality-shared features. To measure the difference of the added modality, the results were presented with only one modality and with both modalities. By measuring performance differences the strength of the feature vectors is found and the research question can be answered.

In the previous section, the modality-specific approach had better performances when RGB and IR were combined in one architecture. This translates that these modality-specific IR features are important for a stronger feature vector. Furthermore, in the modality-shared approach, the RGB and IR combination resulted in worse performances. This then translates to that modality-shared features in the IR modality do not strengthen the feature vector and are therefore less important.

So by combining these findings, it can be stated that the modality-specific IR features enhance the performance of the vessel re-identification task.

The improvements result in a combined score of 2.163 on vessel re-identification (following Equation 6). The relative improvement compared to the RGB-only approach is a 0.023 combined score. But when rank-k accuracy (from the identification task) is individually analysed an increase of 4.5% in rank-1 accuracy is found. When the AUPRC (from the known/unknown classification task) is individually analysed, an increase of 0.03 is found.

So what does this research contribute to the context of vessel re-identification? First, the first RGB-IR vessel re-identification dataset has been created which can be used to continue the multi-modal RGB-IR vessel re-identification research.

Secondly, the state-of-the-art models can be enhanced even further when IR is added to the network. So, combining these findings with the state-of-the-art advances the field of vessel re-identification. (Note that in this research the models are limited and not state-of-the-art. These limitations are listed in the next section.)

7 LIMITATIONS AND OPEN ISSUES

In this paper, some limitations and open issues are described. These are collected in this subsection. For the limitations the first four will (possibly) increase performances for re-identification and the last two will give more extensive results for this research. The limitations are:

(1) The use of single images as input to the model. According to the literature, video-based input should improve the model
(2) The dataset used is impure. Higher performances can be gained if a method that creates less impure data is used.
(3) The model for the RGB is a ResNet50 model. In the literature, better-performing models are used. A different RGB model will not change the outcome of this research but will result in higher performances overall (for example using MVR-net [4] or an adaption of MGN [31] as described in section 2).
(4) The dataset has not been analysed on whether it contains enough versatile data to capture all the necessary features for re-identification. If this is the case, more data can improve performances. Other researchers (see Table 1) used datasets with 25-30 times more images. This might indicate that an improvement can be made on the dataset.
(5) The AUPRC is bound to change when different ratios are used for known/unknown gallery ships and known/unknown query ships. By adapting these parameters more extensive results on AUPRC can be presented.
(6) The ResIRNet and MobileNet have only been evaluated when it was trained from scratch. A pre-train on grayscale ImageNet [1] can be done to have a basis of modality-shared features just like the pre-trained ResNet50. It is expected that this will have lower performance metrics than ResNet because the model is smaller but it does have better computational metrics.

There is also an open issue in finding a better IR network for vessel re-identification. In the approach, some IR networks were compared in the modality-specific approach. Since this approach gave promising results, it is interesting to research how to improve this network further. This can be done by improving the RGB network following the existing state-of-the-art but also by improving the IR network. Since this is the first RGB-IR research in the field, the way forward is still an open issue. The way forward can be inspired by studying networks for the IR modality in other fields than re-identification.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2019-June. IEEE Computer Society, 4685 – 4694. https://doi.org/10.1109/CVPR.2019.00482

[3] Yunhao Du, Cheng Lei, Zhicheng Zhao, Yuan Dong, and Fei Su. 2024. Video-Based Visible-Infrared Person Re-Identification with Auxiliary Samples. *IEEE Transactions on Information Forensics and Security* 19 (2024), 1313 – 1325. https://doi.org/10.1109/TIFS.2023.3337972

[4] Amir Ghahremani, Tunc Alkanat, Egor Bondarev, and Peter H. N. de With. 2021. Maritime vessel re-identification: novel VR-VCA dataset and a multi-branch architecture MVR-net. *Machine Vision and Applications* 32, 3 (2021). https://doi.org/10.1007/s00138-021-01199-1

[5] X. Hao, S. Zhao, M. Ye, and J. Shen. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. 16383–16392.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. 2019. Searching for MobileNetV3. arXiv:1905.02244 https://arxiv.org/abs/1905.02244

[8] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861 https://arxiv.org/abs/1704.04861

[9] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. 2022. Alleviating Modality Bias Training for Infrared-Visible Person Re-Identification. *IEEE Transactions on Multimedia* 24 (2022), 1570 – 1582. https://doi.org/10.1109/TMM.2021.3067760

[10] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 https://arxiv.org/abs/1412.6980

[11] J. Koopmans. 2023. *Improving Tracking Continuity in Maritime Camera Surveillance Systems through Deep Representation Learning*. Master's thesis. Radboud University.

[12] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. 2022. Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2022-June. IEEE Computer Society, 20941 – 20950. https://doi.org/10.1109/CVPR52688.2022.02030

[13] Haijun Liu, Xiaoheng Tan, and Xichuan Zhou. 2021. Parameter Sharing Exploration and Hetero-Center Triplet Loss for Visible-Thermal Person Re-Identification. *IEEE Transactions on Multimedia* 23 (2021), 4414 – 4425. https://doi.org/10.1109/TMM.2020.3042080

[14] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors (Switzerland)* 17, 3 (2017). https://doi.org/10.3390/s17030605

[15] Dalei Qiao, Guangzhong Liu, Feng Dong, She-Xiang Jiang, and Likun Dai. 2020. Marine Vessel Re-Identification: A Large-Scale Dataset and Global-and-Local Fusion-Based Discriminative Feature Learning. *IEEE Access* 8 (2020), 27744 – 27756. https://doi.org/10.1109/ACCESS.2020.2969231

[16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2016-December. IEEE Computer Society, 779 – 788. https://doi.org/10.1109/CVPR.2016.91 Cited by: 32412; All Open Access, Green Open Access.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 07-12-June-2015. IEEE Computer Society, 815 – 823. https://doi.org/10.1109/CVPR.2015.7298682

[18] Tasfia Shermin, Shyh Wei Teng, Manzur Murshed, Guojun Lu, Ferdous Sohel, and Manoranjan Paul. 2019. Enhanced Transfer Learning with ImageNet Trained Classification Layer. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11854 LNCS (2019), 142 – 155. https://doi.org/10.1007/978-3-030-34879-3_12

[19] Paolo Spagnolo, Francesco Filieri, Cosimo Distante, Pier Luigi Mazzeo, and Paolo D'Ambrosio. 2019. A new annotated dataset for boat detection and re-identification. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2019*. https://doi.org/10.1109/AVSS.2019.8909831

[20] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*. Association for Computing Machinery, Inc, 274 – 282. https://doi.org/10.1145/3240508.3240552 Cited by: 905; All Open Access, Green Open Access.

[21] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. 2017. RGB-Infrared Cross-Modality Person Re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2017-October. Institute of Electrical and Electronics Engineers Inc., 5390 – 5399. https://doi.org/10.1109/ICCV.2017.575

[22] Mang Ye, Xiangyuan Lan, Qingming Leng, and Jianbing Shen. 2020. Cross-Modality Person Re-Identification via Modality-Aware Collaborative Ensemble Learning. *IEEE Transactions on Image Processing* 29 (2020), 9387 – 9399. https://doi.org/10.1109/TIP.2020.2998275

[23] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. 2018. Hierarchical discriminative learning for visible thermal person re-identification. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. AAAI press, 7501 – 7508. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055722326&partnerID=40&md5=ba0329820c7aed75e844c19c2bd4e5bd

[24] Zijun Yu, Jin Liu, Shenjie Zou, and Yuetian Cao. 2023. VesselNet: A Large-Scale Dataset and Efficient Mixed Attention Network for Vessel Re-identification. In *Proceedings - 2023 2nd International Conference on Machine Learning, Cloud Computing, and Intelligent Mining, MLCCIM 2023*. Institute of Electrical and Electronics Engineers Inc., 437 – 441. https://doi.org/10.1109/MLCCIM60412.2023.00070

[25] Qian Zhang, Mingxin Zhang, Jinghe Liu, Xuanyu He, Ran Song, and Wei Zhang. 2023. Unsupervised Maritime Vessel Re-Identification With Multi-Level Contrastive Learning. *IEEE Transactions on Intelligent Transportation Systems* 24, 5 (2023), 5406 – 5418. https://doi.org/10.1109/TITS.2023.3243591

[26] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. 2019. AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations. arXiv:1905.00292 https://arxiv.org/abs/1905.00292

[27] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. 2022. Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13674 LNCS (2022), 462 – 479. https://doi.org/10.1007/978-3-031-19781-9_27

[28] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. 2016. Mars: A video benchmark for large-scale person re-identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9910 LNCS (2016), 868 – 884. https://doi.org/10.1007/978-3-319-46466-4_52

[29] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2015 International Conference on Computer Vision, ICCV 2015. Institute of Electrical and Electronics Engineers Inc., 1116 – 1124. https://doi.org/10.1109/ICCV.2015.133

[30] Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. 2022. Grayscale Enhancement Colorization Network for Visible-Infrared Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2022), 1418 – 1430. https://doi.org/10.1109/TCSVT.2021.3072171

[31] Matthijs H. Zwemer, Herman G. J. Groot, Rob Wijnhoven, Egor Bondarev, and Peter H. N. de With. 2021. Multi-camera vessel-speed enforcement by enhancing detection and re-identification techniques. *Sensors* 21, 14 (2021). https://doi.org/10.3390/s21144659