Machine Vision and Applications (2021) 32:71
https://doi.org/10.1007/s00138-021-01199-1

ORIGINAL PAPER

Maritime vessel re-identification: novel VR-VCA dataset and a multi-branch architecture MVR-net

Amir Ghahremani[1] · Tunc Alkanat[1] · Egor Bondarev[1] · Peter H. N. de With[1]

Received: 17 July 2020 / Revised: 14 January 2021 / Accepted: 25 March 2021 / Published online: 15 April 2021

Abstract
Maritime vessel re-identification (re-ID) is a computer vision task of vessel identity matching across disjoint camera views. Prominent applications of vessel re-ID exist in the fields of surveillance and maritime traffic flow analysis. However, the field suffers from the absence of a large-scale dataset that enables training of deep learning models. In this study, we present a new dataset that includes 4614 images of 729 vessels along with 5-bin orientation and 8-class vessel-type annotations to promote further research. A second contribution of this study is the baseline re-ID analysis of our new dataset. Performances of 10 recent deep learning architectures are quantitatively compared to reveal the best practices. Lastly, we propose a novel multi-branch deep learning architecture, Maritime Vessel Re-ID network (MVR-net), to address the challenging problem of vessel re-ID. Evaluation of our approach on the new dataset yields 74.5% mAP and 77.9% Rank-1 score, providing a performance increase of 5.7% mAP and 5.0% Rank-1 over the best-performing baseline. MVR-net also outperforms the PRN (a pioneering vehicle re-ID network), by 2.9% and 4.3% higher mAP and Rank-1, respectively.

Keywords Maritime surveillance · Deep learning · CNNs · Image retrieval · Maritime vessel re-identification

1 Introduction

In recent years, the demand for automated surveillance systems has grown rapidly. This is mainly due to the continuous decrease of the costs of cameras and sensors, leading to broadly available video material and the inefficiency and high labor costs to process this enormous amount of data by humans. To alleviate this, numerous algorithms have been proposed to automate the analysis of video surveillance material and the subsequent alerting for specific various events and dangerous situations. The automated analysis of video surveillance involves the automated detection of objects and their classification. These objects-of-interest include people, vehicles and maritime vessels. However, it is clear that regardless of the target entity, creating a continuous visual coverage of a large physical space is not feasible. In other words, the camera field-of-views are inevitably sparse compared to the full area where objects-of-interest may be present. As a result, intelligent surveillance systems need to find the correspondences between the identities of objects within multiple disjoint camera views. This makes the re-identification (re-ID) of specific objects and their identities one of the most important steps to achieve fully-automated surveillance and higher levels of event analysis.

Today, most security applications benefit from revealing the motion history of objects across disjoint camera views. This valuable information can be used to search for an object-of-interest in a larger database, to detect trajectory-based anomalies and to reveal inter-camera trajectories for observation, surveillance and flow statistics. In recent years, due to its prominent use-cases, the problem of re-ID increasingly attracted scientific attention. Starting from the area of person re-ID, the research has been extended to cover the problems of vehicle and maritime vessel re-ID, where different environments and added difficulties render the re-deployment of the person re-ID algorithms ineffective. As a result, specialized approaches to the maritime vessel re-ID are required and since maritime vessels are infamous tools for transporting illegal goods [1], piracy [27] and illegal fishing [7,30], would make an automated system for maritime vessel surveillance highly attractive for law enforcement.

✉ Amir Ghahremani
a.ghahremani@tue.nl; amirQahremani@gmail.com

[1] Department of Electrical Engineering, Video Coding and Architectures (VCA) Group, Eindhoven University of Technology (TU/e), Eindhoven, North Brabant, Netherlands

Similar to person and vehicle re-ID in traffic on public roads, maritime vessel re-ID is a difficult problem because of the associated challenges. Difficulties inherent to the re-ID task, such as occlusions, viewpoint variations, low-resolution images, target object similarities and variable lighting/weather conditions do also occur in maritime vessel re-ID. Furthermore, there are additional, specific difficulties that are unique to the maritime vessel re-ID. For example, a large variability of aspect ratio with changing pose and large distances between the camera and the target vessel introduce new challenges. However, thanks to modern computer vision techniques, faster hardware and deep learning, developing a practical, real-time re-ID system for automated surveillance is now possible.

In order to exploit and take advantage of deep learning algorithms, an excessive amount of labeled data is needed. However, to the best of our knowledge, only a few medium-scale datasets are publicly available to researchers for the problem of maritime vessel re-ID. To address this issue and in this study, we present a new maritime vessel re-ID dataset to promote further research. Our new dataset is named VR-VCA (Vessel Re-identification-Video Coding and Architectures Research Group). It includes 729 unique maritime vessel identities, pre-labeled for their vessel-type, orientation and identity. A total of 4,614 images are available in the dataset.

In order to assess the difficulty level of our dataset and to provide a baseline for further studies, we evaluate 6 different architectures with 2 different loss combinations and training settings. This quantitative evaluation of existing approaches reveals efficient training and network design strategies for vessel re-ID and provides further studies with useful basic design principles.

Lately, multi-branch architectures have gained importance in the fields of person and vehicle re-ID. State-of-the-art algorithms are now using spatially and/or channel-wise partitioned and pooled feature maps and train those individual branches with separate losses. Such an approach is shown to improve the re-ID accuracy significantly, since it takes advantage of combining the local features with global features in an end-to-end trained, deep learning framework. Motivated by the superior performance of branched networks in multiple image retrieval fields, we propose a new architecture that is carefully designed to address the problem of vessel re-ID. We summarize and present our contributions below.

- Large-scale maritime vessel re-ID dataset with annotated vessel orientation and vessel-type labels has been collected and created. This dataset is made publicly available[1].
- Comprehensive performance evaluation of deep learning architectures on the new dataset, to constitute a strong baseline.
- Novel, multi-branch architecture that is carefully designed to solve the difficulties of the maritime vessel re-ID problem.

The remainder is structured as follows. Section 2 reviews the person, vehicle and maritime vessel re-ID methods from literature. In Sect. 3, we propose a multi-branch deep learning method in detail. Section 4 introduces our new dataset and presents statistical information on its content. Section 5 discusses a quantitative performance evaluation of our method and selected baseline architectures. Lastly, in Sect. 6, concluding remarks are given.

2 Related work

To review the existing approaches to maritime vessel re-ID, we first briefly discuss the person and vehicle re-ID algorithms.

2.1 Person and vehicle re-identification

Person re-identification Person re-ID is the task of people identity matching across non-overlapping camera views. To the best of our knowledge, the problem of re-ID in general has been first introduced in [34], in the context of *person* re-ID. In this work, authors assign a latent label for each person and define a probabilistic relation between the labels and features. Then, the re-ID problem is solved by finding the posterior label distributions. Following [34], the field received immense scientific attention. In most studies, a fixed set of features are extracted from each of the person images, followed by the calculation of a mathematical measure of distance for each feature pair. Then, from the database of *known* people (gallery), likely matches that have small distances to an image under consideration (query) are retrieved. This methodology forms two main directions for person re-ID research. On one hand, researchers focused on developing better feature extraction approaches that are suitable for the problem. On the other hand, studies aimed to develop better distance metrics to yield better ranking.

In early feature extraction studies, a wide variety of different handcrafted features are proposed to address the problem. In [10], the authors use spatiotemporal over-segmentation to determine viewpoint invariant regions. Then, they compute a feature vector that uses color and structural information. In [2], Bazzani *et al.* introduce Symmetry-Driven Accumulation of Local Features (SDALF). SDALF features are extracted by first applying foreground-background segmentation on person bounding boxes, followed by a silhouette

[1] http://vca.ele.tue.nl/.

partitioning that identifies salient regions. Finally, the color and texture features that are extracted from each salient region are combined into a single feature vector. In [6], the authors apply a detector based on Histogram of Oriented Gradients (HOG) to locate the full body and other semantically meaningful partitions such as the top, the torso, legs, the left arm and the right arm of each pedestrian. Then, the covariance descriptor is extracted from each region, considering the position, color and gradients. Lastly, authors employ a pyramid matching scheme with multi-granularity features to compute distances between the detected people. In [16] and [12], scale-invariant feature transform (SIFT), and its modification, the speeded-up robust features (SURF) are used to characterize the person bounding boxes. Other popular feature extraction methods that have been used to solve the problem of person re-ID include maximally stable color regions (MSCR) [25], local binary patterns (LBP) [15] and local maximal occurrence (LOMO) [23].

Besides the feature extraction-based approaches to person re-ID, metric learning methods have also received immense scientific attention. In such methods, the aim is to find a transformation of the feature space, such that the transformed feature space has better separation of different identity clusters. Such approaches typically attempt to minimize the intra-class variance of each identity, while maximizing the separation of different identities. For instance, in [20], the authors derive a metric learning method by exploiting an equivalence constraint in an efficient formulation. In [37], Zheng *et al.* introduce probabilistic relative distance comparison (PRDC) model that aims to maximize the probability of a matching pair, having a smaller distance than that of a non-matching pair. In [28], the authors reformulate the challenge as a ranking problem and learn a subspace where the potential true match is assigned the highest ranking. This approach effectively transforms the problem into a relative ranking problem, instead of an absolute scoring problem. In XQDA [23], Liao *et al.* formulate to learn a discriminant low-dimensional subspace by cross-view quadratic discriminant analysis.

Following the emergence of deep learning in 2012, most of the recent person re-ID research now utilizes deep models to solve the problem. Learning feature extraction and suitable distance metrics simultaneously from the available large-scale datasets, deep learning solutions to the person re-ID provide high accuracy and reasonable computational cost thanks to powerful modern hardware. This opens up new possibilities for the practical applications of re-ID, especially in the field of surveillance.

To take advantage of the potential of deep learning, numerous methods were proposed. Hermans *et al.* [14] introduce a mini-batch construction strategy that is fine-tailored for the use of triplet loss. Authors propose to use only the hardest positive and negative sample for a given anchor image in a carefully sampled mini-batch to improve the performance. In [22], authors propose a harmonious attention module that takes advantage of determining the regions-of-interest of a given person bounding box sample. In [5], authors derive a new loss function called the quadruplet loss. In this method, three images for each anchor image are sampled from the dataset, two of which are negative (different identity). In [38], authors propose a generative adversarial network (GAN) to enhance the training dataset with artificially generated data. In [17], Kalayeh *et al.* use a two-stream deep architecture, where one of the streams generates masks for semantically meaningful body partitions, and the other extracts features. Then, the final feature vector for a given bounding-box image is constructed by combining the feature vectors for each semantic region. In [31], Su *et al.* take advantage of the pose information to enhance the performance of re-ID. In this method, authors utilize the pose estimations to partition the bounding-box image, and learn robust global and local feature representations. In [32], authors use triplet and softmax losses in a multi-branch architecture called MGN. In this method, each branch divides the intermediate feature volume into multiple volumes before collapsing the spatial dimensions with pooling. Then, each divided part is trained with an independent loss to yield better feature extraction. In [26], authors first extract convolutional neural network (CNN) features for each person bounding box in a time sequence. Then, a Recurrent Neural Network (RNN) is used to combine the feature vectors of individual frames into one feature vector to be used for re-ID. Numerous other person re-ID methods were proposed for the problem. For further reading on the problem of person re-ID, the reader is referred to the surveys in [3] and [36].

Vehicle re-identification The problem of Vehicle re-ID is the task of identity matching of vehicles across disjoint cameras. Recently, this problem attracted increasing scientific attention, due to its valuable applications in the fields of surveillance and traffic-flow analysis. The vehicle re-ID problem includes additional challenges, such as motion blur, varying aspect ratio of bounding boxes, reflective surfaces of vehicles and only subtle differences between different identities with similar model/make/year.

Although the vehicle re-ID is a relatively new problem compared to its person variant, since research could take advantage of the already mature person re-ID literature, the performance has grown significantly in a short time. For instance, in [33], authors generate orientation-based region proposals to refine the global CNN features with respect to the viewpoint. In [35], Zapletal *et al.* first extract 3D bounding-box information of a given bounding-box image of a vehicle. Then, the image is *normalized* by mapping different visible sides of the vehicle into a fixed spatial location. In [24], authors introduce a multi-branch architecture called region-aware deep model (RAM). In this architecture,

multiple branches aim to extract better features by using different strategies, such as spatial feature volume partitioning, attribute learning and batch normalization. Similarly, in [4], authors propose a two-branch architecture that, in addition to spatial partitioning, employs partitioning of intermediate feature volumes in the channel dimension. In [21], authors discuss various mini-batch sampling strategies for the triplet loss. Further, Kumar *et al.* also comparatively evaluate the contrastive and triplet losses. For a detailed review of vehicle re-ID methods, the reader is kindly referred to [18].

Maritime vessel re-identification Compared to its person and vehicle variants, maritime vessel re-ID is a relatively new problem. In addition to the already challenging problems of person and vehicle re-ID, maritime vessel re-ID introduces additional challenges, such as low-resolution images due to the size of the vessel and imaging distance, and high variability of the bounding-box aspect ratios with the viewpoint.

Up to this point, the maritime vessel re-ID problem has received only fractional scientific attention compared to its person and vehicle variants. In [9], authors propose an architecture called the identity-oriented re-identification network that combines the triplet loss and softmax cross-entropy (CE) loss with a ResNet50 [13] architecture. In [11], authors base their work on [14] and extend the method with various multi-query strategies. In [29], authors introduce a new dataset, as well as a novel approach that employs global-and-local fusion-based discriminative feature learning. This method combines CE loss with a novel, orientation-guided quintuplet loss and performs multi-view representation learning for re-ID.

We conjecture that the field of maritime vessel re-ID suffers from the lack of a widely-adopted, large-scale dataset. To alleviate this, we introduce a new dataset called VR-VCA. In accordance with the current trends in re-ID literature, we also provide the viewpoint and vessel-type labels for each sample to promote further research. The detailed information and baseline analysis of our dataset is included in Sect. 4. Further, we also propose a novel, deep learning-based, multi-branch architecture in Sect. 3.

3 Maritime vessel re-identification

This section presents the proposed MVR-net method. In the following subsections, we provide an overview of the proposed method, after which we explain each element of the architecture in detail.

3.1 Architecture overview

Figure 1 illustrates the architecture of the MVR-net. The proposed network receives a mini-batch of labeled vessel images as input. Then, it extracts a feature embedding for the input images using a backbone feature-extraction network. Afterwards, the embedding is passed through three parallel convolutional branches. Each of these branches are carefully designed to further discriminate the extracted feature embedding and generate a more indicative representation of input images in feature space. Typical usage of branches for specific features are the processing of height, width, and channel information. This type of architecture is inspired by recent multi-branch methods like MGN [32] and PRN [4], which show significant improvement in re-identification performance for pedestrians and vehicles, respectively. The MGN network is an example that uses height as a guiding discriminating feature, while PRN employs height, width, and channel as discriminating features. The proposed MVR-net also exploits those three branches and uses a combination of a triplet and a softmax loss function to calculate the gradients required in the training procedure.

3.2 MVR-net description

This subsection explains each part of the proposed MVR-net in detail, as illustrated in Fig. 1.

Pre-processing of input images This work is based on the VR-VCA dataset for training. In this dataset, the majority of vessel samples are captured with cameras deployed on shorelines and therefore possess horizontally oriented structures. Therefore, the MVR-net first downsamples the training images into a size of 128×384 pixels (height $\times$ width). Then, it applies conventional standardization and augmentation techniques (e.g. normalizing, random-horizontal flipping, and random erasing [40]) on the resized input images. At the beginning of each training epoch, MVR-net randomly picks K pre-processed samples per individual vessel to generate an input batch. The pre-processing operations on input images and the batch generation are not depicted in the network architecture diagram.

Mini-batch generation To benefit from the fast training and gradient smoothness advantages, we employ mini-batch gradient training with batch-hard sampling strategy [14]. For this, the second step in each epoch is to divide the generated batch into mini-batches of N samples. Each mini-batch will include P unique IDs, and has a size of N/K. This form of mini-batch generation is adopted, since our network uses triplet loss with hard batch mining as part of its total cost function. Finally, all mini-batches are supplied into the backbone network for training iterations.

Backbone network Generally, CNN-based re-identification methods extract features for input samples and further process these features to verify if a query sample belongs to a specific identity from the gallery database, based on feature similarities. To this end, a CNN network is required to extract the feature representations. In order to facilitate the process, an initial set of layers from reliable classification CNNs are

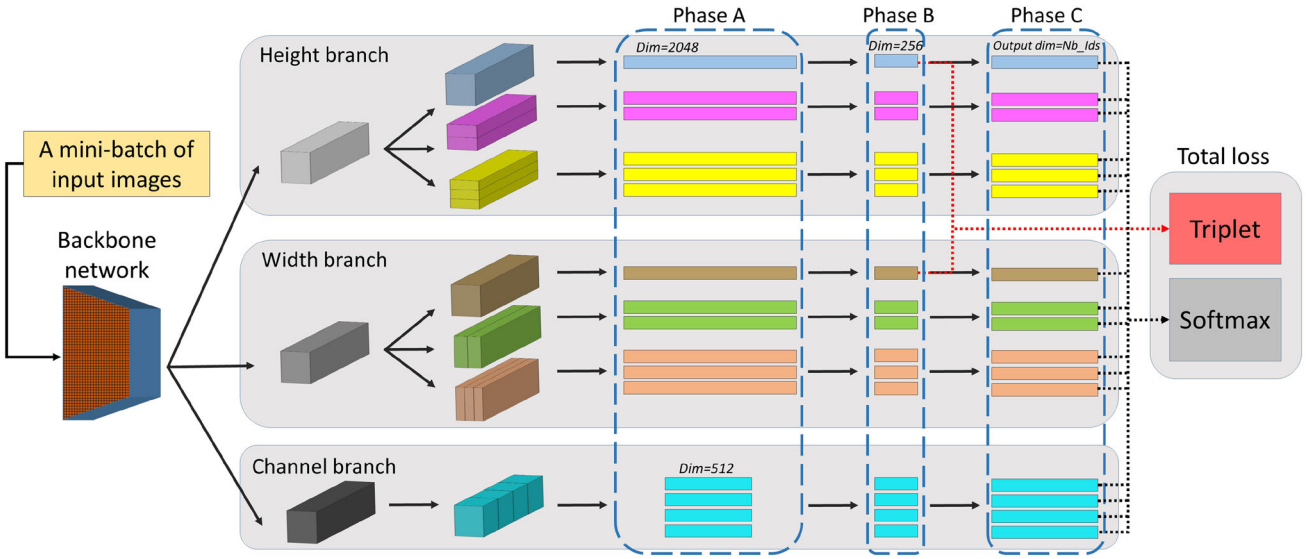

Fig. 1 Architecture of the proposed MVR-net

frequently employed to generate the coarse portion of the required feature representation. This will be followed by re-identification-oriented layers, designed specifically for the intended application.

One of the most frequently used backbones in the re-identification literature is ResNet50 [13]. This network is well known for its robust performance against the vanishing gradients problem. Therefore, we employ ResNet50 as our backbone network to extract the feature maps for each image of the mini-batch. Then, we duplicate all the extracted feature maps at the end of the Conv4_1 layer (see [13] for details) three times, and feed these features separately into our three re-identification branches. In other words, the layers of these branches will be constructed on top of separate copies of the Conv4_1 layer of the ResNet-50, provided by the same backbone network. We will refer to each of these duplication of feature maps as a feature embedding in the remainder of the paper. Each branch will then further process its embedding to generate more meaningful features for vessel re-identification in a parallel manner.

Multi-branch design As a core contribution, MVR-net proposes a 3-branch architecture, specified to address the maritime vessel re-identification problem. As illustrated in Fig. 1, these branches are called Height branch, Width branch, and Channel branch, focusing on the height-wise, width-wise, and channel-wise structures available in their input feature embeddings.

The Height branch performs three independent partitioning operations across the height dimension of its input feature embedding. These operations generate one, two, and three separate feature volumes, respectively. Each of these volumes contain the same spatial fraction of the input feature embedding. The Width branch applies the same partitioning operations on the horizontal axis of the spatial dimensions. In both of these branches, the operation of partitioning the input feature embedding into one volume means copying the whole input feature embedding into a separate volume, which aims at maintaining the global features of the embedding for the next steps. Such copied volumes will be referred to as global volumes in the remainder of this paper (e.g. the 3D block at the left in each branch). It is also important to highlight that these two branches spatially partition the input feature embedding into "vertical" and "horizontal" volumes. Finally, the Channel branch partitions its input feature embedding into four volumes of features across the embedding depth.

After generating these 16 partitioned volumes out of the input feature embeddings, each branch passes its feature volume through a three-phase pipeline to prepare the required feature vectors for the loss calculation block. These three phases are developed as follows.

Phase A: The generated feature volumes are separately shrunk into feature vectors of size 2, 048 (for the Height and Width branches) and 512 (for the Channel branch) using global max-pooling operations.

Phase B: 1x1 convolutional layers are employed to equalize the vector dimensions constructed in the previous phase to a size of 256. Additionally, a batch normalization operation is also applied on the feature vectors. At this point, separate copies of the two feature vectors resulting from the global volumes of the Height and the Width branches are supplied into a triplet loss block.

Phase C: Each of the obtained 16 vectors are transferred into separate fully-connected layers. The output dimensions of these fully-connected layers are equal to the number of unique vessel identities in the training set. Then, these outputs are carried into a softmax CE loss block. MVR-net combines

the triplet and the softmax CE losses to calculate the gradients required for updating the network parameters during training. These loss functions will be explained in detail in the next part of this subsection.

Loss function We employ softmax CE loss and triplet loss with the batch-hard sampling strategy as introduced in [14], to train our architecture. The batch-hard triplet loss is defined in Eq. (5) of [14] as:

$$\mathcal{L}_{\mathrm{BH}}(\theta; X) = \sum_{i=1}^{P}\sum_{a=1}^{K}\Big[m + \max_{p=1\ldots K} D\left(f_\theta(x_a^i), f_\theta(x_p^i)\right) - \min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(f_\theta(x_a^i), f_\theta(x_n^j)\right)\Big]_+, \tag{1}$$

where X, θ, P, K, m, f_θ and D denote an input mini-batch, learned network weights, number of different identities in a mini-batch, number of images per identity in a mini-batch, triplet loss margin, forward inference function and distance calculation function, respectively. In addition to the triplet loss, we use the softmax CE loss for identity classification. Mathematically, softmax CE loss for a mini-batch X can be specified by:

$$\mathcal{L}_{\mathrm{CE}}(\theta; X) = -\sum_{x\in X} log\left(\frac{e^{f_\theta(x)[\mathrm{id}(x)]}}{\sum_j e^{f_\theta(x)[j]}}\right), \tag{2}$$

where function id(.) returns the identity of a given image. Combining the two losses, the total loss is expressed as:

$$\mathcal{L}_{\mathrm{total}}(\theta; X) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\mathrm{BH}}^t(\theta; X) + \frac{\gamma}{C}\sum_{c=1}^{C}\mathcal{L}_{\mathrm{CE}}^c(\theta; X), \tag{3}$$

where γ, T and C are the loss weighting parameter, the number of triplet losses and the number of softmax CE losses, respectively.

3.3 Discussion on MVR-net architecture

Motivation for Height and Width branches As mentioned above, the Height and Width branches of the MVR-net separately partition their input feature embeddings into one, two, and three volumes. This idea of partitioning the feature embeddings into several volumes is inspired by the recent state-of-the-art methods. As an example, the well-known PRN network consists of only two branches. These branches partition their input feature embedding just once, and into four spatial volumes (one branch vertically and the other branch horizontally). However, this architecture is designed for the vehicle re-identification problem. Surveying the vehicle re-identification datasets shows that car samples contain similar image characteristics, although cars are different. This is due to the observation that surveillance cameras capturing cars are installed at fixed locations on top of roads and capture the cars in similar distances from the camera lenses, yielding car samples with similar image qualities and resolutions. However, in a vessel re-identification dataset (e.g. our VR-VCA dataset) captured in maritime environments, vessels appear in a diverse settings of camera viewing angles, distances to the camera, occlusions, etc. For a maritime surveillance system, it is essential to identify the vessels as soon as they enter the receptive field. This is important for the security of harbors. Consequently, the resolution of captured vessels in our vessel-oriented dataset varies in a wide range. This variation in input images diminishes the performance of a network that processes the feature embeddings for all input samples using only one limited number of partitions (e.g. only four partitions for the PRN). This motivates why our MVR-net has a broader partitioning to cover the various resolution scenarios. This is implemented in MVR-net by grasping features using a separate partitioning into different numbers of volumes.

In order to select the optimum number of partitioning for each spatial branch, we have trained several versions of the MVR-net with different numbers of partitioning (which will be discussed later in the experimental results). After this experimental investigation, we have decided to split each input embedding separately into one, two, and three volumes in each spatial branch. With this, for the vessels located at a close distance to the camera (vessels with higher resolution), the volumes obtained by splitting the input embeddings into three spatial partitions yield more detailed features. Similarly, the two-partition volumes will extract useful information for the low-resolution vessels (e.g. those located at a far-away distance to the surveillance camera). According to our experiments, including volumes that are obtained by splitting the feature embeddings into more than three partitions (like four in PRN) diminishes the re-identification performance for vessels. This occurs most probably because a higher degree of partitioning reduces the influence of coarser features (e.g. the global features) in the final feature representation. Therefore, we can conclude that the coarser features have a high impact on the vessel re-identification problem (compared to the vehicle re-identification), especially if the vessel samples lack resolution.

Motivation for channel branch We know from neural style transferring that identifying the correlations between outputs of different filters of a convolutional layer can help to grasp the existing style in a set of images (e.g. the common style of the Van Gogh paintings). This concept is also adopted in re-identification methods. To the best of our knowledge, PRN is the first to utilize these correlations by applying channel-wise partitioning on the feature embeddings. According to the original PRN paper [4], the authors target the extraction of distinct local features by these channel-wise operations.

However, these channel-wise partitionings are performed together with the spatial partitionings in the same channels. Each branch of PRN splits the input embedding into one and four spatial (one branch vertically and the other one horizontally) and four channel-wise volumes. We prefer to design a network with three branches, first one for detecting horizontal structures inside the input images (Height branch), the second one for detecting the vertical structures (Width branch), and the third one for detecting the internal correlations between different feature maps of the feature embedding.

4 Maritime vessel re-identification dataset: VR-VCA

In order to train the vessel re-identification model, we have recorded several videos at different day/year-times from various locations in the Netherlands. We have used two different cameras in our recordings. The videos contain a vast variety of viewpoints on vessels. Additionally, several vessel types with divergent sizes and distances to the cameras are represented in this dataset. Finally, challenging scenarios including vessel occlusion/truncation are also annotated. Figures 2 and 3 illustrate several examples of the VR-VCA dataset.

The dataset contains a total of 4614 vessel samples from 729 unique vessel identities. Each vessel identity is represented by several samples. Additionally, we have labeled each vessel with a bounding box, its vessel type, and vessel orientation (i.e. the approximate camera viewing angle) to facilitate future research. [2]. Vessel types include the following eight classes: sailing vessel, passenger ship, fishing vessel, river cargo, small boat, yacht, tug, and taxi vessel. The vessel orientations are described with the following five orientation labels: front view, front-side view, side view, back-side view, and back view. Figure 4 statistically analyzes the VR-VCA samples in terms of vessel types and orientations. Besides, we have provided the same unique ID to all samples corresponding to each specific vessel. Moreover, a label is assigned to each cropped sample showing whether the vessel is captured in its full body, or is truncated, or occluded. The dataset is split into training, gallery, and query datasets. The specifications of these datasets are as follows:

Training dataset The training dataset contains 2, 268 samples from 365 individual vessels. This dataset includes 184 sailing vessels, 442 passenger ships, 30 fishing vessels, 936 river cargos, 105 small boats, 22 yachts, 64 tugs, and 485 taxi vessels. These samples cover 170 vessels from front view, 727 vessels from front-side view, 736 vessels from side view, 556 vessels from back-side view, and 79 vessels from back view. The maximum and the minimum number of samples per unique ID are 38 and 2, respectively.

Gallery dataset The gallery dataset is comprised of 1, 667 samples from 364 unique vessel identities. The type statistics of the gallery dataset samples are as follows: 144 sailing vessels, 379 passenger ships, 13 fishing vessels, 722 river cargos, 57 small boats, 9 yachts, 29 tugs, and 314 taxi vessels. In this dataset, there are 124 front views, 569 front-side views, 498 side views, 412 back-side views, and 64 back views from vessels. The maximum and the minimum number of samples per unique ID are 26 and 1, respectively.

Query dataset The query dataset possesses 679 samples with 364 unique vessel identities. This dataset includes 68 sailing vessels, 152 passenger ships, 5 fishing vessels, 260 river cargos, 26 small boats, 5 yachts, 13 tugs, and 150 taxi vessels. These samples represent 32 vessels with front views, 157 vessels with front-side views, 243 vessels with side views, 213 vessels with back-side views, and 34 vessels with back views. In the query dataset, the maximum and the minimum number of samples per unique ID is equal to 7 and 1, respectively.

4.1 Discussion on VR-VCA characteristics

There is a fundamental difference between a vehicle re-identification dataset and our vessel re-identification dataset. In vehicle re-identification, cameras are installed at fixed locations on top or aside of roads, covering a specific background. The cars have no other way but to pass through the receptive fields of these cameras. Under such a setting, each car sample is also given a separate camera ID. While performing re-identification for each query sample (which means ranking gallery samples based on their similarities to the input query sample), it is common to discard the gallery images from the same car and the same camera as the query sample. The reason is that we need the query sample to be captured at a different camera location compared to the candidate gallery images. Otherwise, there is no need for a complex re-identification system and the task can be performed by a simple tracker. However, in a maritime environment, vessels move in arbitrary directions, practically making the fixed camera option infeasible. This problem becomes even more challenging if the maritime environment is a spacious harbor with different exit areas where vessels can maneuver easily. Such cases occur rather frequently in our dataset. Therefore, we have recorded our images by continuously chasing the moving vessels. With this, we have captured each individual vessel in different perspectives and with different backgrounds. Hence, we have discarded this camera ID concept in our dataset, since all our samples have a different camera and varying background settings. In other words, we have considered all our samples to virtually possess a different camera ID. It is possible that some samples have *similar*

[2] Besides the discussed dataset, there is also another separate multi-type and orientation vessel detection dataset [8].

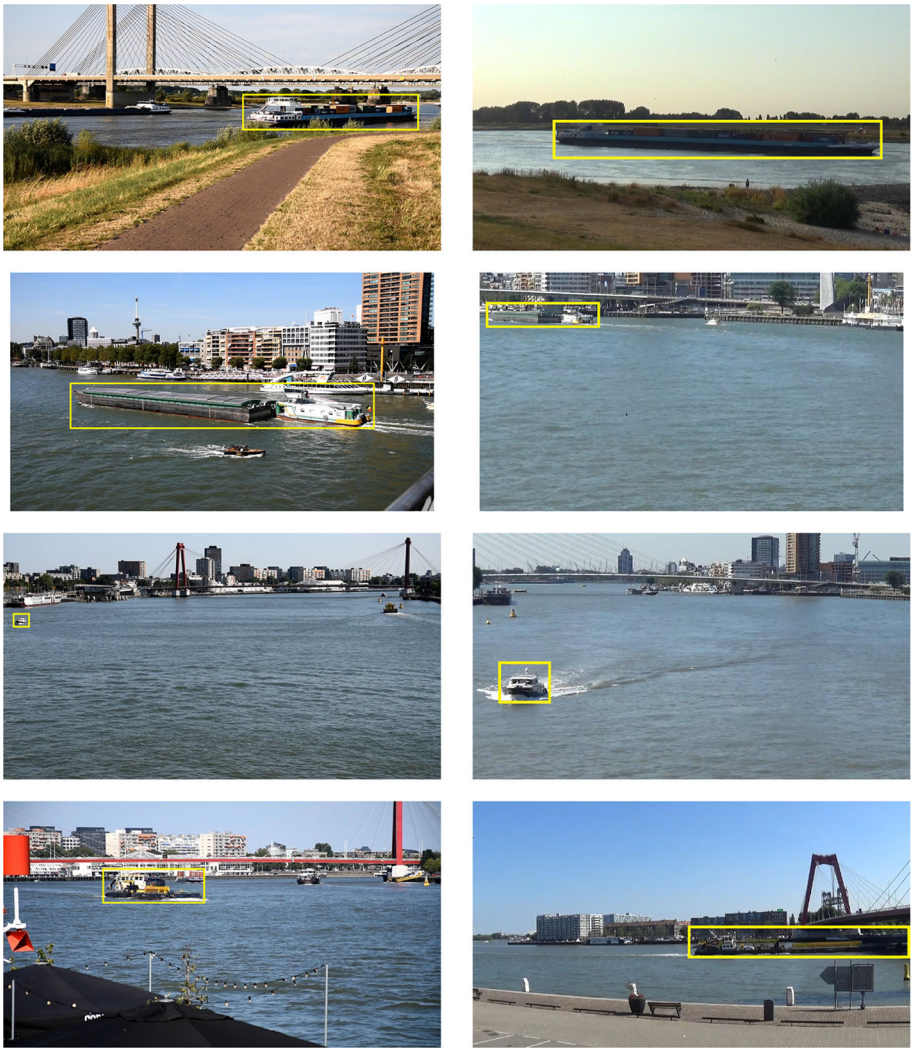

Fig. 2 Four VR-VCA examples. Each row presents two samples of an individual vessel in different locations

backgrounds, which comes from the fact that the vessels with the same identity are always captured in the same neighborhood.

Besides this difference to vehicle re-identification datasets, there is a resemblance too. In vehicle re-identification, there is always a possibility that cars with similar appearances (model, color, etc.) pass through views of the same cameras. Therefore, a vehicle re-identification dataset is generally containing samples with different identities, but with almost the same appearance (although these samples may slightly differ because of tiny stickers or human passengers or other details). This scenario becomes even more challenging in our vessel re-identification dataset captured in city/river-type harbors. Only a limited number of vessel models appear in such environments, and consequently the probability and frequency of finding vessels with the same appearance is much higher than for cars. For example, in a considered scenario, a harbor is located next to a wide river passing through a city and passenger vessels of the same model belonging to a specific company are continuously transferring people across the river. Thus, another aspect making our dataset challenging is the problem of having appearance-wise overlap between the vessel samples.

Last but not least, vessel samples of the VR-VCA dataset are cropped from outdoor surveillance images, having different sizes and aspect ratios. Therefore, CNN-based re-ID systems need to resize them into a fixed footage in a data pre-processing stage (based on system requirements and architectural design). However, vessels appear with more divergent aspect ratios in image frames, compared to other conventional re-ID targets (e.g. pedestrians with mostly vertical and cars with mostly squared-shape bounding boxes). Thus, the vessel samples of the VR-VCA dataset vary from very narrow- yet long- (horizontally or vertically) shaped samples to quite square-shaped ones, depending on their types and orientations. This implies a need for decision making on the proper size for the input samples, making VR-VCA an even more challenging dataset. This challenge

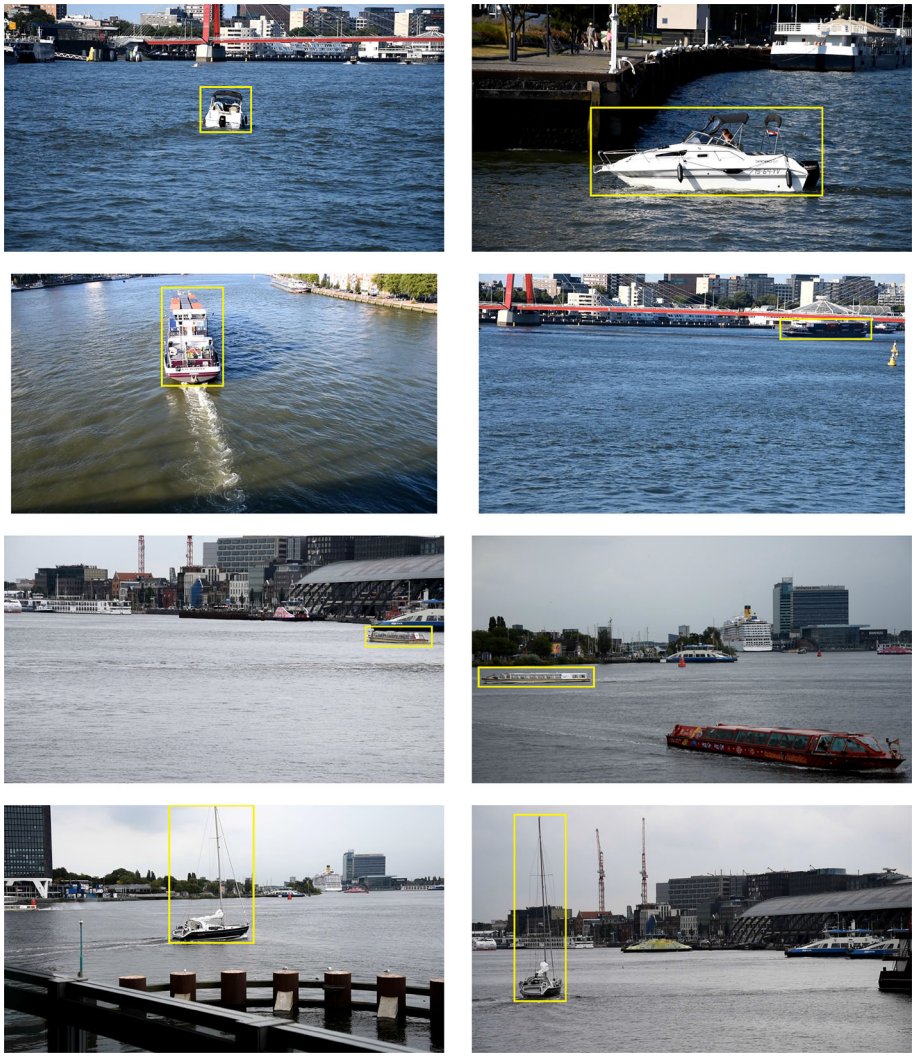

Fig. 3 Another four VR-VCA examples. Each row presents two samples of an individual vessel in different locations

may not hold for other vessel re-identification datasets, since we deliberately capture videos such that yields vessels in a divergent range of resolutions and aspect ratios (by covering vessels also in far-away locations from different viewpoints).

5 Empirical validation

This section presents the empirical evaluations and discussions on the outcomes in the following subsections.

5.1 VR-VCA performance analysis on baselines

This subsection benchmarks the VR-VCA data by testing several re-ID methods on that. For choosing the methods for comparison, we have adopted models that vary in terms of complexity by having a different network architecture and a different number of network layers. Table 1 presents the performance of the selected models on VR-VCA. The table compares six baseline models with three different losses, both with and without re-ranking technique. First, the implementation process of designing the deep learning models is explained. Afterwards, the results are analyzed.

Implementation details for baseline methods In order to match each baseline architecture to the vessel re-ID problem, we have adapted the architectures to optimize their performances for this problem. The chosen optimization measures are generic and apply to all baseline architectures, so that each baseline architecture is modified in the same way for a fair comparison.

The following measures have been implemented. (1) We have substituted the fully-connected classifier and pooling layers at the end of each network, and added a max-pooling layer to incorporate both spatial dimensions in concentrated form. (2) The latter layer produces a fixed-size feature vector which is then used for the triplet loss training. (3) Additionally, if softmax CE loss is employed during training, we have added a fully-connected layer at the end that outputs the class

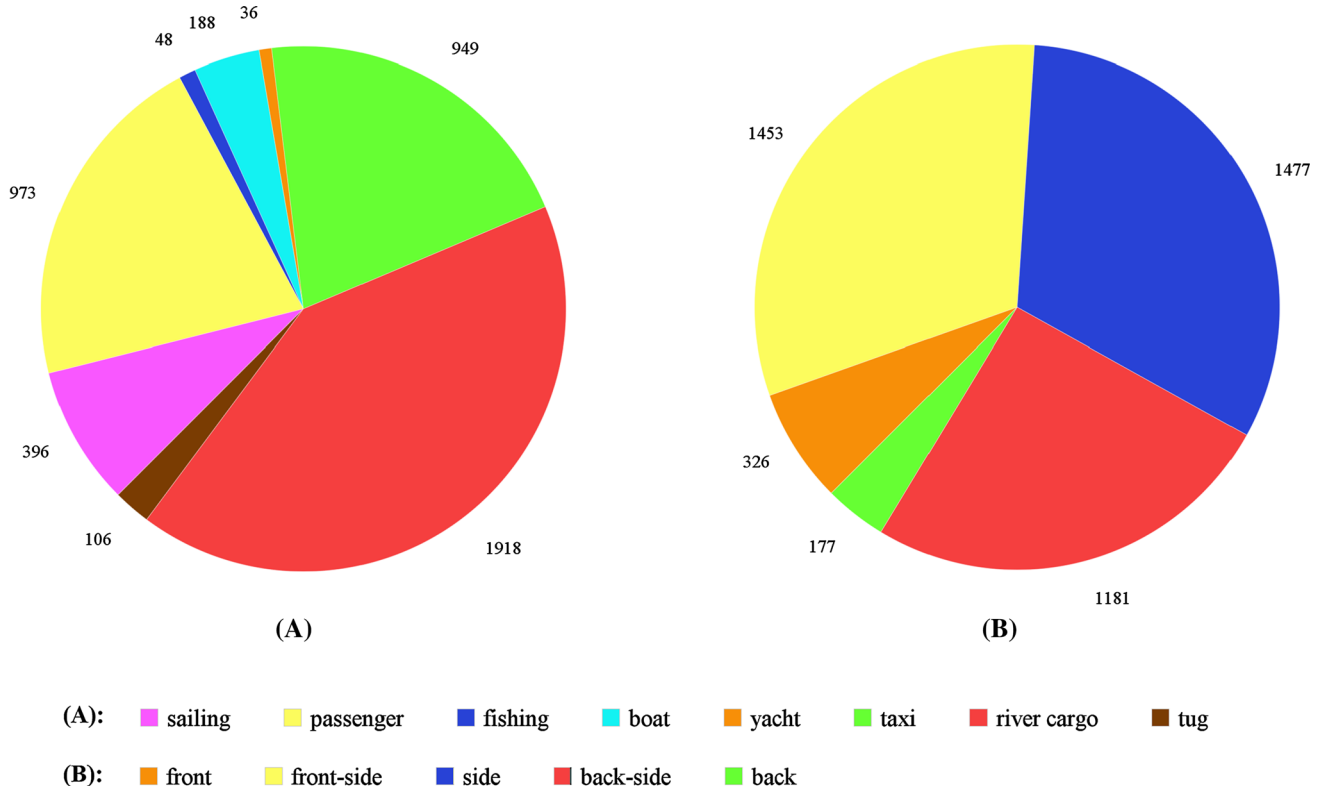

Fig. 4 VR-VCA representation in terms of: **a** vessel types, and **b** vessel orientations

probabilities for each vessel identity. (4) We have employed re-ranking to improve performance of the re-identification. Re-ranking works by post-processing the initial ranking results. Following the common practice in literature, we have used the k-reciprocal encoding re-ranking algorithm [39]. For a given query sample, this algorithm assumes that a correct match in the list of top-k_1 retrieved samples is likely to have the query itself retrieved within the top-k_2 positions when queried to the re-identification module ($k_2 < k_1$). Thus, the re-ranking step computes a derived distance metric by separately analyzing the k-reciprocal neighbors for all samples and computing their Jaccard distance. Finally, the final distances are computed by a weighted sum of the original and Jaccard distances, where individual contributions are weighted by λ and $1 - \lambda$, respectively. (5) Our ImageNet pre-trained baseline architectures are trained with the Adam optimizer [19] for 25 epochs with an initial learning rate of 3×10^{-4}, which is reduced to one-tenth of this value after every 10^{th} and 15^{th} epochs. Batch-hard parameters P and K are set to $P = 5$ and $K = 4$, while γ is set to unity where applicable, the weight decay is set to 5×10^{-4} and the triplet loss margin, m is set to unity. For data augmentation, random horizontal flipping is employed during training and all images are resized into a fixed size of 128×384 pixels. During testing, we calculate the features for both the original images and their horizontally flipped versions and average them to compute the final feature vector for each image. The re-ranking parameters, k_1, k_2, and λ are set to 20, 6, and 0.3, respectively.

Result analysis According to Table 1, increasing the number of layers and thereby, the complexity of the deep models improves the performance. For example, the slope of increment is sensible going from ResNet18 RR to ResNet50 RR, where the mAP improves by 2.1%. However, the higher number of layers in the ResNet121 RR eventuates in only a slight growth of mAP, compared to the ResNet50 RR. Therefore, it can be concluded that a re-identification network based on ResNet50 can provide a more reliable performance on this dataset. Additionally, according to our experiments, using only a softmax CE loss cannot provide adequate discrimination in feature space. However, combining this loss with triplet loss improves the performance of the baseline architectures. For example, ResNet50 RR employing only triplet loss generates 1.1% lower mAP compared to the combined loss function version. It is also important to mention that employing the re-ranking technique always improves the mAP, while decreasing the rest of the metrics (Rank-1, R-3, R-5, R-10). Our explanation for obtaining lower ranks when utilizing the re-ranking technique is the high frequency of having vessels with the same appearance but of different identities in VR-VCA, as explained in Sect. 4.1.

Table 1 Performances of various network architectures, given by percentage scores of mAP, Rank-1 (R-1), Rank-3 (R-3), etc. for ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, DenseNet121, DenseNet161, DenseNet169, DenseNet201, and MobileNet

Network Arch.	Training: Triplet Loss					Training: Triplet+Softmax CE Loss				
	mAP	R-1	R-3	R-5	R-10	mAP	R-1	R-3	R-5	R-10
ResNet18	55.0	67.3	80.6	85.9	91.5	57.3	68.9	84.7	89.1	92.6
ResNet18 (RR)	58.5	61.3	77.6	83.5	88.8	62.3	66.4	80.3	85.1	90.3
ResNet34	54.2	67.9	80.1	85.7	91.3	56.4	70.3	82.0	86.3	90.7
ResNet34 (RR)	59.1	63.9	76.9	81.9	88.2	60.6	63.8	76.0	82.5	89.3
ResNet50	58.0	71.6	84.5	89.7	94.0	58.4	71.7	84.1	88.4	93.8
ResNet50 (RR)	63.3	68.3	80.6	86.3	91.0	64.4	67.9	81.3	86.3	92.5
ResNet101	57.0	70.1	84.2	89.1	92.5	57.9	71.4	82.9	87.2	91.8
ResNet101 (RR)	62.7	65.1	80.4	85.7	91.6	63.3	68.3	78.7	84.1	89.8
ResNet152	57.4	71.6	84.0	88.5	92.9	59.8	73.1	85.6	90.0	92.9
ResNet152 (RR)	64.4	69.1	81.7	85.6	90.3	64.8	68.0	81.0	86.2	91.5
DenseNet121	59.9	73.1	85.1	90.0	94.4	61.8	75.4	86.3	91.2	95.7
DenseNet121 (RR)	66.9	70.8	84.0	88.1	92.3	68.2	72.8	83.2	87.6	92.3
DenseNet161	59.2	73.1	83.7	88.2	92.9	62.8	75.7	87.3	91.5	95.1
DenseNet161 (RR)	64.5	67.8	81.6	85.6	90.3	68.8	72.9	83.7	87.6	92.5
DenseNet169	59.9	72.2	86.2	89.7	94.1	62.8	75.9	86.5	90.6	95.3
DenseNet169 (RR)	65.3	68.8	80.3	86.2	92.8	68.7	71.4	83.2	87.3	91.3
DenseNet201	59.6	73.8	84.5	88.8	92.9	64.1	77.2	87.5	91.3	94.9
DenseNet201 (RR)	66.0	68.3	82.3	87.0	90.1	70.1	72.9	84.4	88.5	92.3
MobileNet	54.6	65.8	80.4	85.7	91.2	60.1	71.0	85.6	90.3	94.3
MobileNet (RR)	57.7	59.5	74.8	80.3	88.8	65.2	66.3	81.4	86.5	92.1

The networks are trained with (1) triplet loss only, and (2) the combination of triplet loss with softmax CE loss. RR stands for re-ranking and is most important

5.2 Validation of MVR-net

This subsection specifically evaluates the MVR-net performance on our vessel re-identification dataset. To this end, the MVR-net is compared with two state-of-the-art re-identification networks, PRN and MGN. The implementation details are first provided and then the obtained results are analyzed. Finally, two separate topics are discussed: our network design, and the batch-hard sampling strategy utilized by the triplet loss function.

Implementation details of MVR-net The MVR-net architecture is trained with the Adam optimizer for 25 epochs with an initial learning rate of 2×10^{-4}, which is reduced to one-tenth after the 15^{th} and 20^{th} epochs. Batch-hard parameters P and K are set to $P = 5$ and $K = 4$. Parameter γ is set to $\gamma = 2$, the weight decay is set to 5×10^{-4} and the triplet loss margin m is set to unity. For data augmentation, random horizontal flipping and random erasing are employed during training and all images are resized to a fixed size of 128×384 pixels. During test time, we calculate the features for both the original images and their horizontally flipped versions and average them to compute the final feature vector for each image.

Validation results for MVR-net The network is compared with the MGN and PRN re-ID architectures. Table 2 illustrates the results, both with and without the re-ranking (RR) technique. Here, the results are only analyzed and discussed based on the models with the re-ranking technique. According to the table, the proposed MVR-net outperforms the PRN and MGN with 2.9% and 8.7% mAP, respectively. The improved performance holds also for other evaluation metrics of re-identification. For example, MVR-net generates 4.3% and 6.6% higher Rank-1 compared to PRN and MGN, respectively. This performance clearly proves the efficiency of the MVR-net design for maritime surveillance applications.

Table 2 Performance comparison of network architectures and our MVR-net

Network arch.	mAP	R-1	R-3	R-5	R-10
MGN	62.0	75.1	87.3	91.2	95.0
MGN (RR)	65.8	71.3	82.8	87.2	93.8
PRN	67.8	76.9	89.7	92.9	95.6
PRN (RR)	71.6	73.6	84.1	87.9	92.6
MVR-net	70.5	80.9	90.0	90.1	96.3
MVR-net (RR)	**74.5**	**77.9**	**87.6**	**90.9**	**95.0**

Results expressed in percentage scores for mAP, Rank-1, R-3, etc. RR stands for re-ranking and bold numbers for best score

Side experiment using triplet loss According to common approaches in literature, we have also empowered our triplet

loss function with the batch-hard sampling strategy. Implementing this strategy, the triplet block treats all N vessel samples of a mini-batch as an anchor sample once. This means in practice for each of the N samples, we randomly select an anchor image, a positive pair (i.e. another sample of the same vessel), and a negative pair (i.e. a sample from another vessel). For example, this finds the most similar negative pairs (i.e. other samples of the mini-batch with different identities to the anchor sample). Likewise, a similar statement can be made for the most dissimilar positive pairs. The intuition behind utilizing this technique is to minimize the distance between samples with the same identity and to maximize the distance between samples with different identities in the feature space as much as possible. Normally, this happens by comparing the feature similarities of mini-batch samples using metrics like Euclidean distance.

In this work, we have attempted to construct better triplets by choosing most dissimilar positive and most similar negative pairs according to their orientation labels. The motivation is that employing an appearance-based differentiating metric, like Euclidean distance, can result in selecting positive and negative pairs only because of reasons like differences in scene lighting or having extra background-pixels around the vessel (inside the cropped image). Moreover, we manipulated the input mini-batch generation block to force this block create each mini-batch using samples having the same or at least similar type labels. With this, we aimed at increasing the ability of the triplet block in choosing more similar negative pairs. However, after applying the explained type and orientation-based strategy to the triplet loss-calculation process, we have noticed that there is no improvement compared to the MVR-net with the standard batch-hard sampling strategy. We think this happens because for training of the MVR-net, more than 35k random mini-batches have been used. Therefore, the triplet block is already trained sufficiently for different combinations of input images to the network, which explains the lack of improvement. The conclusion of this side experiment is that with batch-hard sampling and the applied triplet loss, the re-identification of the same vessel at another harbor position does not improve from incorporating the viewing angles, but the proposed network finds already sufficient occurrences of the same vessel in the dataset.

Discussion on the MVR-net design As mentioned in Sect. 3.3, we have designed and empirically evaluated several network architectures to find the optimized network architecture for the vessel re-identification problem (i.e. the MVR-net). This part briefly reflects on these experiments with all possible candidate designs, with a limit of splitting the feature embedding up to four partitions. Table 3 illustrates the obtained results. The tested architectures can include two or three branches, depending on whether the channel-wise partitioning is implemented inside the spatial branches (as is the case for PRN) or in a third independent channel branch. According to the table, the architecture of MVR-net yields higher performance both in terms of mAP and R-1. This motivates our preferred architecture that is illustrated in Fig. 1 as our selected re-identification network for maritime surveillance.

Table 3 Performance scores (%) in mAP and Rank-1 for candidate architectures with different branches and partitionings

Network architecture	mAP	R-1
$B_1(H^1, H^2), B_2(V^1, V^2)$	73.1	74.8
$B_1(H^1, H^2, C^4), B_2(V^1, V^2, C^4)$	71.3	75.0
$B_1(H^1, H^2, H^3), B_2(V^1, V^2, V^3)$	73.0	75.7
$B_1(H^1, H^2, H^3, C^4), B_2(V^1, V^2, V^3, C^4)$	71.1	74.7
$B_1(H^1, ..., , H^4), B_2(V^1, ..., V^4)$	71.6	75.7
$B_1(H^1, ..., H^4, C^4), B_2(V^1, ..., V^4, C^4)$	66.1	68.2
$B_1(H^1, H^2), B_2(V^1, V^2), B_3(C^4)$	72.9	75.8
MVR-net	74.5	77.9
$B_1(H^1, ..., H^4), B_2(V^1, ..., V^4), B_3(C^4)$	69.7	72.9

Parameters B_k, H^i, V^i, and C^4 stand for the k^{th} Branch, horizontal (H) partitioning into i equal splits, vertical (V) partitioning into i equal splits, and channel-wise (C) partitioning into 4 equal splits, respectively. Finally, separate branches are combined in a parallel manner after the Conv4_1 layer of ResNet50, represented in table by commas. As an example of such a parallelism, MVR-net is illustrated in Fig. 1 and would be specified as $B_1(H^1, H^2, H^3), B_2(V^1, V^2, V^3), B_3(C^4)$

6 Conclusions

In this paper, we have introduced two main contributions for addressing the vessel re-identification problem. First, we have captured, annotated, and hereby publish a novel vessel re-identification dataset, referred to as VR-VCA. This dataset contains 4, 614 vessel samples from 729 unique vessel identities. Additionally, we have provided eight vessel types and five vessel orientation labels for each dataset sample. The images of the VR-VCA dataset are captured at different locations in the Netherlands. A divergent set of weather conditions, water region types, and backgrounds are represented in VR-VCA. In our dataset, multiple vessels occur with very similar appearances (i.e. model, etc.). Additionally, vessels appear in various aspect ratio distributions and are captured in different distances and orientations to the cameras. These broad variations make the VR-VCA a challenging vessel re-ID dataset.

Performance of different baseline methods is benchmarked with the described dataset. Based on this benchmarking, we have adopted ResNet50 as the backbone network for the vessel re-ID problem. In addition to the dataset, we have introduced a re-identification deep network, MVR-net, specifically designed for maritime surveillance domain. This network architecture achieves reliable re-ID performance on maritime vessels by combining their spatial and channel-wise

features. For extracting a better representation of vessels in spatial dimensions, MVR-net employs two separate height-wise and width-wise branches. Since the vessels are captured at different resolutions, the spatial branches partition the feature embedding into three different sets to detect more useful features for each resolution scenario. The proposed network outperforms PRN and MGN, two well-known re-identification networks, with 2.9% and 8.7% mAP, and 4.3% and 6.6% higher Rank-1, respectively. We have validated the MVR-net efficiency by testing several alternative candidate network designs, where it is shown that the adopted architecture yields the highest scores.

For future work, the implementation parameters of baseline networks and the MVR-net can be further tuned. Additionally, we aim at improving the MVR-net performance on VR-VCA, using pose and class information of vessels, and multi-resolution feature pyramids. Moreover, to address the challenge that is imposed by having different vessels with similar appearances, a model refinement focusing on detecting local features of vessel images could be explored.

Acknowledgements The authors appreciate the nVIDIA Corporation gift of two GPUs for this research. The research is funded by the European H2020 Interreg PASSAnT Project and the Provincial Government of Noord-Brabant, The Netherlands.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Atkinson, M.P., Kress, M., Szechtman, R.: Maritime transportation of illegal drugs from south america. Int. J. Drug Policy **39**, 43–51 (2017)
2. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)
3. Bedagkar-Gala, A., Shah, S.K.: A survey of approaches and trends in person re-identification. Image Vis. Comput. **32**(4), 270–286 (2014)
4. Chen, H., Lagadec, B., Bremond, F.: Partition and reunion: a two-branch neural network for vehicle re-identification. In: Proceedings of the CVPR Workshops, pp. 184–192 (2019)
5. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 403–412 (2017)
6. Corvee, E., Bremond, F., Thonnat, M., et al.: Person re-identification using spatial covariance regions of human body parts. In: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 435–440. IEEE (2010)
7. Detsis, E., Brodsky, Y., Knudtson, P., Cuba, M., Fuqua, H., Szalai, B.: Project catch: a space based solution to combat illegal, unreported and unregulated fishing: Part i: vessel monitoring system. Acta Astron. **80**, 114–123 (2012)
8. Ghahremani, A., Bondarev, E., de With, P.H.: Toward robust multi-type and orientation detection of vessels in maritime surveillance. Electronic Imaging (2020)
9. Ghahremani, A., Kong, Y., Bondarev, E., et al.: Towards parameter-optimized vessel re-identification based on iornet. In: International Conference on Computational Science, pp. 125–136. Springer (2019)
10. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1528–1535. IEEE (2006)
11. Groot, H.G., Zwemer, M.H., Wijnhoven, R., Bondarau, E., et al.: Vessel-speed enforcement system by multi-camera detection and re-identification. In: 15th International Conference on Computer Vision Theory and Applications 2020 (2020)
12. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Interest points harvesting in video sequences for efficient person identification. In: The Eighth International Workshop on Visual Surveillance-VS2008 (2008)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
15. Hirzer, M., Roth, P.M., Bischof, H.: Person re-identification by efficient impostor-based metric learning. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, pp. 203–208. IEEE (2012)
16. Jüngling, K., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: CVPR 2011 WORKSHOPS, pp. 55–61. IEEE (2011)
17. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1062–1071 (2018)
18. Khan, S.D., Ullah, H.: A survey of advances in vision-based vehicle re-identification. Comput. Vis. Image Underst. **182**, 50–63 (2019)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295. IEEE (2012)
21. Kuma, R., Weill, E., Aghdasi, F., Sriram, P.: Vehicle re-identification: an efficient baseline using triplet embedding. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. IEEE (2019)
22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2285–2294 (2018)

23. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
24. Liu, X., Zhang, S., Huang, Q., Gao, W.: Ram: a region-aware deep model for vehicle re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
25. Ma, B., Su, Y., Jurie, F.: Local descriptors encoded by fisher vectors for person re-identification. In: European Conference on Computer Vision, pp. 413–422. Springer (2012)
26. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1325–1334 (2016)
27. Orji, U.J., et al.: Tackling piracy and other illegal activities in nigerian waters. JoDRM **4**(2), 65–70 (2013)
28. Prosser, B.J., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. BMVC **2**, 6 (2010)
29. Qiao, D., Liu, G., Dong, F., Jiang, S.X., Dai, L.: Marine vessel re-identification: a large-scale dataset and global-and-local fusion-based discriminative feature learning. IEEE Access **8**, 27744–27756 (2020)
30. Sander, K., Lee, J., Hickey, V., Mosoti, V.B., Virdin, J., Magrath, W.B.: Conceptualizing maritime environmental and natural resources law enforcement-the case of illegal fishing. Environ. Dev. **11**, 112–122 (2014)
31. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3960–3969 (2017)
32. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
33. Wang, Z., Tang, L., Liu, X., Yao, Z., Yi, S., Shao, J., Yan, J., Wang, S., Li, H., Wang, X.: Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 379–387 (2017)
34. Zajdel, W., Zivkovic, Z., Krose, B.: Keeping track of humans: Have i seen this person before? In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 2081–2086. IEEE (2005)
35. Zapletal, D., Herout, A.: Vehicle re-identification for automatic video traffic surveillance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–31 (2016)
36. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
37. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR 2011, pp. 649–656. IEEE (2011)
38. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762 (2017)
39. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)
40. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)

Amir Ghahremani received his BSc degree in Electrical Engineering with an emphasis on telecommunications from Azad University, Urmia, Iran, in 2009. He finished his BSc education working on superconductivity concepts for his thesis. This was followed by a MSc degree in Electrical Engineering with an emphasis on Electronics at the Khajeh Nasir Toosi University of Technology (KNTU), Tehran, Iran, in 2014. He did his MSc thesis project on image processing and computer vision topics. Since 2015, he has worked as a PhD candidate in the Video Coding and Architectures (VCA) Research Group at the Eindhoven University of Technology (TU/e), The Netherlands. His PhD research interests included computer vision, machine learning, and deep learning.

Tunc Alkanat received the BS and MS degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2013 and 2016, respectively. He is currently working toward the PhD degree at the Video Coding and Architectures Group, Eindhoven University of Technology, Eindhoven, The Netherlands. His research interests include image retrieval, anomaly detection, computer vision for surveillance and computational spectral imaging.

Egor Bondarev received his MSc degree in robotics and informatics at the State Polytechnic University, Belarus Republic, in 1997. In 2009, he has obtained his PhD degree in Computer Science Department at Eindhoven University of Technology (TU/e), The Netherlands, in the research domain of performance predictions of real-time component-based systems on multiprocessor architectures.

Peter H.N.de With received his PhD degree (1992) from University of Technology Delft, The Netherlands. From 1984 to 1997, he worked for Philips Research Eindhoven on video compression and chaired a cluster for programmable TV architectures as senior TV Systems Architect. From 1997 to 2000, he was full professor at the University of Mannheim, Germany, Computer Engineering, and chair of Digital Circuitry and Simulation. From 2000 to 2007, he was with LogicaCMG in Eindhoven as a principal consultant and distinguished business consultant. He was also part-time professor at the TU/e, heading the chair on Video Coding and Architectures. From 2008 to 2010, he was VP Video (Analysis) Technology at Cyclomedia Technology. Since 2011, he has been a full professor at Eindhoven University of Technology, faculty EE. From 2011 to 2018, he was scientific director of the Centre for Care and Cure Technologies.