frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 06 November 2024
DOI 10.3389/fmars.2024.1429459

EDITED BY
Narayanamoorthi R.,
SRM Institute of Science and Technology, India

REVIEWED BY
Peng Zhao,
Hainan University, China
Swakkhar Shatabda,
United International University, Bangladesh

*CORRESPONDENCE
Zhibin Yu
yuzhibin@ouc.edu.cn

RECEIVED 08 May 2024
ACCEPTED 14 October 2024
PUBLISHED 06 November 2024

CITATION
Fan S, Song C, Feng H and Yu Z (2024) Take good care of your fish: fish re-identification with synchronized multi-view camera system. *Front. Mar. Sci.* 11:1429459.
doi: 10.3389/fmars.2024.1429459

Take good care of your fish: fish re-identification with synchronized multi-view camera system

Suzhen Fan[1], Chengyang Song[2], Haiyang Feng[1] and Zhibin Yu[1,2*]

[1]Sanya Oceanographic Institution, Ocean University of China, Sanya, China, [2]College of Electronic Engineering, Ocean University of China, Qingdao, China

Introduction: Fish re-identification (re-ID) is of great significance for fish monitoring and can contribute to aquaculture and fish breeding. Synchronizing information from different cameras is beneficial for optimizing re-ID performance.

Methods: We constructed the first underwater fish re-identification benchmark dataset (FS48) under three camera conditions. FS48 encompasses 48 different fish identities, 10,300 frames, and 39,088 bounding boxes, covering various lighting conditions and background environments. Additionally, we developed the first robust and accurate fish re-identification baseline, FSNet, which fuses information from three camera positions by extracting features from synchronized video frames of each position and combining the synchronized information.

Results: The experimental results show that FS48 is universal and of high quality. FSNet has an effective network design and demonstrates good performance, achieving better re-identification performance by combining information from three positions, helping improve overall re-test accuracy, and evaluating the effectiveness of re-identification among detectors.

Discussion: Our dataset will be released upon acceptance of this paper, which is expected to further promote the development of underwater fish re-identification.

KEYWORDS

fish re-identification, multiple cameras, within-view, cross-view, synchronized multi-view

1 Introduction

Fish, as essential products of underwater agriculture, provide significant and sustainable nutrients for humans. Identifying fish individual precisely during their fry cultivation stage is crucial for precise agriculture and marine aquaculture. This ability enables us to provide tailored nutrition, optimize farming conditions, and boost sustainable and efficient farming practices Marini et al. (2018); Zhao et al. (2021). The development of information technology allows us to swiftly detect any unusual features or behaviors with a real-time monitor, making it possible to take timely actions to prevent disease outbreaks and ensure the overall health of the fish population Barbedo (2022); Gladju et al. (2022).

Fish identification technology is fundamental for increasing the importance of biological, ecological, and aquaculture studies because it involves tracing the fate of the organism under study Sandford et al. (2020). Unlike fish classification Alsmadi and Almarashdeh (2022); Chen et al. (2017); Alsmadi et al. (2019), which focuses on distinguishing fishes among different species, fish identification should identify a specified individual from other fishes even if they belong to one category. Traditional fish identification techniques depend heavily on fish tagging Macaulay et al. (2021); Runde et al. (2022); Musselman et al. (2017). Although tagging can provide relatively reliable measurements, the tagging process may inevitably bring detriment Runde et al. (2022). In addition, some sensitive species, such as delta smelt, can be more susceptible to accidental mortality Sandford et al. (2020). Since deep learning-based person/vehicle re-identification (re-ID) technologies have achieved great success Ahmed et al. (2015); Zakria et al. (2021), learning-based tagging-free fish re-identification technology has become a plausible solution. Re-ID technology aims to solve the problem of Re-identifying targets in different scenes or time points and identifying the identity of the same target in a multi-camera system. Re-id technology usually includes subtasks such as object detection, feature extraction, and similarity measurement and involves related technologies such as deep learning, image processing, and cross-camera matching. Fish re-identification technology can also support real-time monitoring and recording of fish growth, which is necessary for fish breeding and disease prevention.

However, the challenges for most general person/vehicle re-ID technologies mainly lie in cross-camera matching, lighting changes, and posture variations Zheng et al. (2023). Owing to the influence of underwater environments and the morphological differences between persons/vehicles and fish, person/vehicle re-ID technologies cannot be directly applied to fish re-ID. The experimental results clearly indicate that our fish re-identification technology can be used underwater and performs exceptionally well. Existing fish tracking and matching technologies are typically conducted via a single camera Chuang et al. (2016); Mei et al. (2022). In real-time monitoring, a single-view camera may unavoidably capture awkward poses detrimental to fish re-identification. Our fish re-identification method employs multi-view techniques (see Figure 1) to capture the visual characteristics of fish from different angles, which helps provide more precise and more accurate fish features and enhances the model's robustness to changes in lighting, water flow, and scenes. Even if we build a multi-view video system for fish re-identification, we still need to overcome the challenge of multi-view information utilization. To address these challenges, we captured synchronized video sequences from three cameras, capturing frontal, left, and top views. The synchronized video capture system allowed us to construct the first underwater fish re-identification benchmark (FS48). Additionally, we developed a robust and accurate multi-view fish re-identification framework called FSNet (see Figure 2).

Inspired by person re-identification techniques, we propose the first underwater fish re-identification network, FSNet, which combines features from three different views. Unlike existing underwater fish detection and tracking methods, FSNet enables synchronous information interaction when addressing occluded or blurry fish caused by reflections. FSNet adopts a traditional approach in which the information from the three views is separately fed into ResNet-50 backbones for feature extraction. We collected three-view video frames under various conditions, including occluded and unoccluded daytime and nighttime scenarios. By leveraging the fused information, FSNet can match and interact the occluded or blurry parts with any unoccluded or

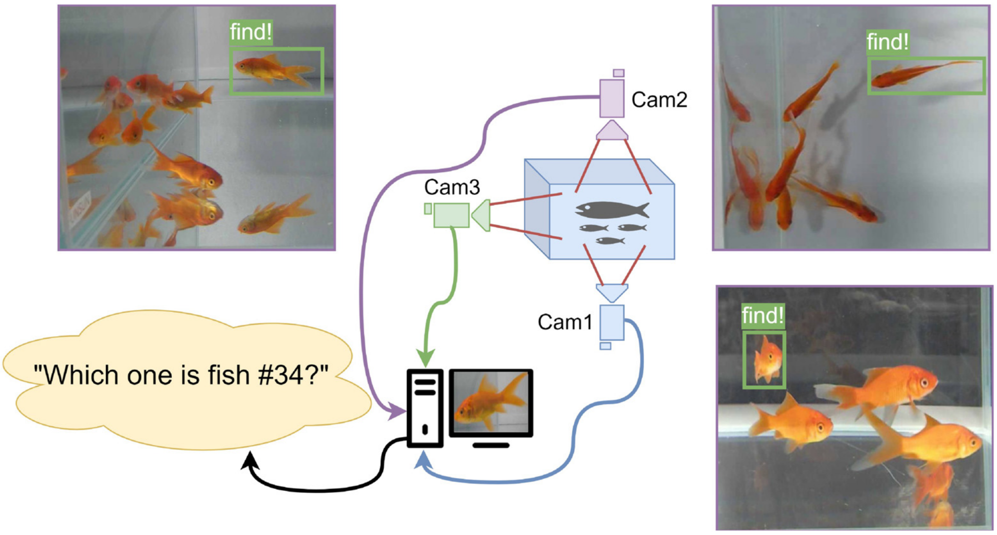

FIGURE 1
Overview of our three-view video capture system and the re-identification process. The characteristic information of the same fish from different visual directions is obtained from three positions to find the target fish (fish #34) after feature fusion.

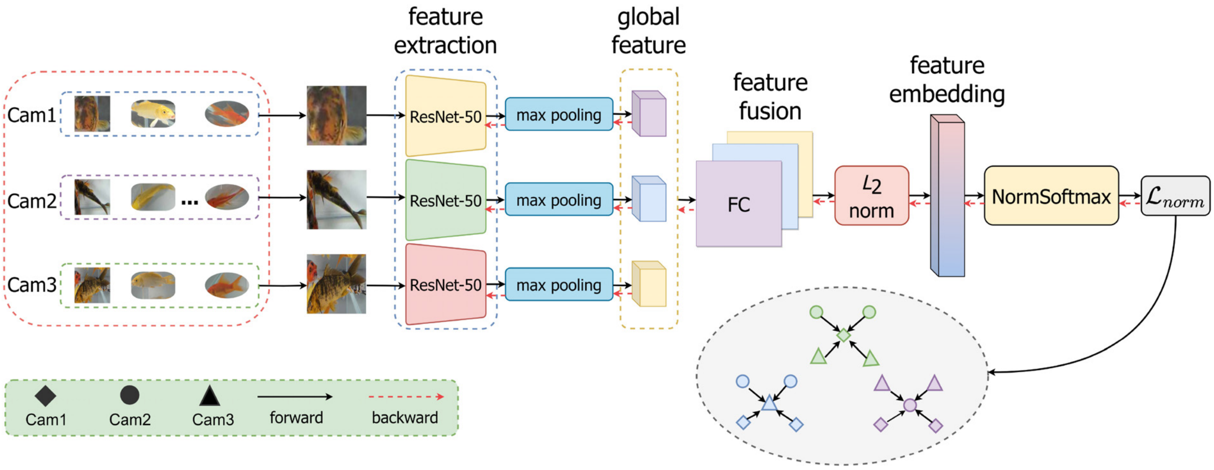

FIGURE 2
Pipeline of Mixed Three Views (FSNet). The image feature is extracted via the backbone. Then, overall feature fusion is performed.

clear parts in the fused feature space, achieving the highest quality fish re-identification performance. The main contributions of this paper are summarized as follows:

- We constructed the first multi-view synchronous underwater fish re-identification dataset, FS48, which consists of 109 video sequences, 10,300 frames, and 39,088 bounding boxes.
- We developed the first underwater multi-view fish re-identification benchmark, FSNet, which can efficiently utilize multi-view information for fish re-identification.
- We comprehensively evaluate the most classical and advanced detection and recognition methods on the FS48 dataset to construct a benchmark.

2 Related work

2.1 Fish detection and classification

While data-driven methods based on deep learning have made significant progress in various computer vision tasks such as classification, detection, segmentation, and image retrieval, most existing research on fish-related studies has focused on fish detection and classification. Fish detection aims to detect and recognize the presence and location of fish in images or videos. Zeng et al. improved the underwater object detection capability of the standard Faster-RCNN detection network by integrating adversarial networks and conducting joint training. The detection performance for fish was significantly enhanced Zeng et al. (2021). Liu et al. introduced an attention mechanism called efficient channel attention to enhance the YOLOX model. They also used Real-ESRGAN to handle multiple targets and blurry images in detection, significantly improving the fish detection accuracy Liu, Dongcai et al. (2022). Fish classification aims to train the model to automatically recognize and classify different fish species and accurately classify fish in various scenarios Spampinato et al. (2010). Automatic fish classification can provide helpful information for water monitoring, ecological research, and marine conservation. This information can be used to understand the distribution, abundance, and ecosystem health of the fish in the water Chen et al. (2017); Alsmadi et al. (2019); Saleh et al. (2022). Despite the recent impressive progress in fish detection and categorization, monitoring studies of individual fish have yet to be conducted. In this work, we perform underwater fish re-ID evaluation tests supported by detection and classification experiments on the FS48 dataset for fish re-identification.

2.2 Re-identification benchmark

Re-identification work has focused on person/vehicle re-identification in recent years Ren et al. (2023a), Ren et al. (2023b). Person re-identification research begins with multi-camera tracking. In 1997, Huang and Russell Huang and Russell (1997) proposed a Bayesian formula to estimate and predict the posterior probability of the appearance of objects in a camera on the basis of the information observed in other camera views. In 2006, Gheissari et al. Gheissari et al. (2006) used the spatiotemporal segmentation algorithm to segment images and then used human visual cues for foreground detection. This work begins with person re-identification and multi-target separation as independent computer vision tasks. In 2014, Xu et al. Xu et al. (2014) solved the impact of person detectors on re-ID accuracy by combining detection (commonness) and re-ID (uniqueness) scores. Owing to the significant changes in appearance and posture, person recognition is complex across cameras; therefore, it is used in security monitoring, personnel tracking, and other fields. Another critical area of re-identification is vehicle re-identification, in which the sensor-based approach phase occurred between 1990 and 1999 Fullerton et al. (1990). The historical stage of vehicle re-identification based on depth features occurred in 2017 and 2018 Liu et al. (2017). The vehicle's appearance is relatively stable, and it is easy to extract features for matching. The development process of two significant research fields has led to the rapid development of re-identification technology. Inspired by these two fields, this paper focuses on fish re-identification with a synchronized multi-view camera system.

2.3 Fish identification

In fisheries management research, individual or batch identification marking systems have been widely utilized for fish tracking, which is crucial for assessing fish growth, survival, mortality rates, and monitoring fish population sizes Buckley et al. (1994). However, the current labeling methods often cause damage to individual fish, thus affecting their survival rate of after labeling. These marking techniques include the use of tags or changing the parts of the fish (cutting off some fins, etc.) Dare (2003), and the relevant identification information, including species, sex, and length-related details, is typically recorded in databases. With the advancement of biological internal tagging technology Cooke et al. (2013); Wilder et al. (2016); Musselman et al. (2017), such as visible implant alphanumeric (VIA) tags Turek et al. (2014); Lindberg et al. (2013); Osbourn et al. (2011), passive integrated transponder (PIT) tags Castillo et al. (2014); Schumann et al. (2013); Hühn et al. (2014), and acoustic tags, the study of small fish individuals has become feasible. However, internal tagging still faces challenges, such as visual identification limitations applicable to small species and potential sublethal effects on fish behavior Murphy et al. (1996); Skalski et al. (2009). Recent studies on fish marking have focused primarily on visible implant alphanumeric (VIA) tags, adipose fin clips (suitable for subadult to adult stages), and calcium marks (ideal for the juvenile stage). However, currently, only VIA tags can provide effective individual identification. Analysis of different tagging methods and species combinations revealed significant differences in tag retention rates and survival rates. Methods such as adipose fin clips, particularly those affecting juvenile fish, have been shown to decrease postmortem survival rates. Even the currently most widely adopted VIA tags still involve harm to individual fish Sandford et al. (2020).

Based on the current research background, the fish re-identification technology proposed in this paper provides a new way to solve the problem of the impact of previous marking methods on individual fish. Our method, which relies on synchronized cameras, can effectively achieve individual identification without endangering individual fish. The proposed fish re-identification technology avoids the potential threat to survival caused by direct contact with fish, provides actual economic benefits and value for farmers, and avoids the risk of financial loss.

3 Fish re-ID

3.1 Problem formulation

Owing to the changes in underwater scenes and other factors, such as different water depths, underwater lighting, and rapid flow, underwater fish detection and identification are much more complex than human re-identification. We propose multi-view fish re-identification, which provides an innovative solution to solve these problems. Following the single image person re-ID settings Zheng et al. (2016), let us define $\mathcal{X}$ as a fish database composed of N images from M identities, denoted as $\mathcal{X} = \{(x_i, y_i)| y_i \in \mathcal{Y}\}_{i=1}^{N}$. Given a query image q, its identity is determined by:

$$\begin{aligned} i^* = \ & \arg\max sim(q, x_i) \\ & i \in 1, 2, \ldots, N \end{aligned} \tag{1}$$

where i^* represents the correct identity label of image q and where $sim(,)$ is a similarity measurement.

For a multi-view synchronized camera system, we can obtain P images for each identity as $\mathbf{x}_i = \left\{x_i^p\right\}_{p=1}^{P}$, simultaneously. Similar as the single image re-ID setting, we split the database $\mathcal{X} = \{(\mathbf{x}_i, y_i)|y_i \in \mathcal{Y}\}_{i=1}^{N}$ into training set $\mathcal{X}_{tra}$ and test set $\mathcal{X}_{test}$ with the same identity set $\mathcal{Y}$, in which there is no identity overlap between $\mathcal{X}_{tra}$ and $\mathcal{X}_{test}$ ($\mathcal{X}_{tra} \cap \mathcal{X}_{test} = \oslash$). Then, we obtain the training set as $\mathcal{X}_{tra} = \{(\mathbf{x}_i, y_i)|y_i \in \mathcal{Y}_{tra}\}_{i=1}^{M}$ and the test set as $\mathcal{X}_{test} = \{(\mathbf{x}_i, y)|y_i \in \mathcal{Y}_{test}\}_{i=1}^{M}$, separately. Both $\mathcal{X}_{tra}$and $\mathcal{X}_{test}$ include independent gallery and probe subsets. We train the multi-view fish re-ID model on $\mathcal{X}_{tra}$ and test it on $\mathcal{X}_{test}$.

3.2 Preliminary

We consider fish re-ID as an image retrieval problem that aims to recognize and match the identities of the same fish between different scenes and cameras. We conducted three different fish re-ID settings on the FS48 dataset and aimed to provide a valuable baseline for further research in this area (please refer to Section 5.1 for details). We comprehensively evaluate two CNN models, VGG-16 Simonyan and Zisserman (2014) and ResNet-50 Kaiming et al. (2016), for image feature extraction Kumar and Bhatia (2014). Our approach to fish re-identification mainly involves configuring two backbone networks with different loss functions. We explored the VGG-16 and ResNet-50 backbone networks on the FS48 dataset along with several loss functions of [SoftTriple Loss Qian et al. (2019), NormSoftMax Zhai and Wu (2019), ProxyAnchor Loss Kim et al. (2020), ArcFace Loss Deng et al. (2022), ProxyNCA Loss Yang et al. (2022)]. The fish re-ID baseline that is better for within-view and cross-view settings is a combination of the ResNet-50 backbone network with SoftTriple Loss, and the fish re-ID baseline that has better performance for synchronized multi-view settings is a combination of the ResNet-50 backbone network with SoftTriple Loss. The results of these experiments will provide helpful guidance and insights for future research and technology development. This paper aims to reveal the model's ability to adapt to other scenes, camera conditions, and changes in fish appearance. The generalization of the FS48 dataset to the fish re-ID problem is demonstrated through an in-depth analysis of the experimental results.

3.3 FSNet

Following the baseline pipeline, we propose a fish re-ID framework named FSNet. In most cases, the information between synchronized video frames of the same fish in different orientations is closely related. We can utilize this relationship to improve fish re-identification performance by relying on globally fused features to infer obscured or blurred semantic information between synchronized frames. Inspired by the joint representation strategy Baltrušaitis et al. (2018), we input the images from each of the three viewpoints into separate ResNet-50 networks for feature extraction. As shown in Figure 2, three synchronized video frames are fed to the backbone for feature extraction. Next, the extracted features from the three frames are sent to the FC layer for feature fusion. With the help of fusion features, video frames affected by underwater environments or other adverse factors can be effectively re-identified.

3.4 Normalization functions

We comprehensively evaluated multiple losses for normalization and chose SoftTriplet Loss Qian et al. (2019) and NormSoftMax Loss Zhai and Wu (2019) as our objective functions to achieve a balance between maximizing the interidentity distance and minimizing the intraidentity distance for different tasks. Following the definition in Section 3.1, let $v_i = \phi(\mathbf{x}_i)$ denote the embedding vector extracted from a multi-view identity $\mathbf{x_i}$. The objective functions are as follows:

$$\mathcal{L}_{SoftTri}(v_i) = -\log \frac{\exp(\lambda(S'_{i,y_i} - \delta))}{\exp\,(\lambda(S'_{i,y_i} - \delta)) + \sum_{j \neq y_i} \exp\,(\lambda S'_{i,j})} \tag{2}$$

where the relaxed similarity $S'_{i,j}$ can be represented as follows:

$$S'_{i,c} = \sum_k \frac{\exp\left(\frac{1}{\gamma} v_i^\top w_c^k\right)}{\sum_k \left(\frac{1}{\gamma} v_i^\top w_c^k\right)} v_i^\top w_c^k \tag{3}$$

Here w represents the trainable weights from the FC layer, and c is the identity label. We follow the default setting and set $\lambda = 20, k = 10, \delta = 0.01, \gamma = 0.1$.

$$\mathcal{L}_{NormSoft}(v_i) == -\log \frac{\frac{\exp\,(v_i^\top p_y)}{t}}{\sum_{z \in \mathcal{Z}} \frac{\exp(v_i^\top p_z)}{t}} \tag{4}$$

where we follow the default setting Zhai and Wu (2019) and set the temperature $t = 0.05$; $\mathcal{Z}$ represents the set of all proxies; and p_y is the target proxy.

4 FS48 dataset

4.1 Camera setup

This study used two kinds of common freshwater fish, 17 crucian carp and 31 common carp, to construct the FS48 dataset. Crucian carp and common carp also belong to common ornamental fish. We chose these two categories since they are highly adaptable, omnivorous, and disease-resistant Li et al. (2022).

During the rearing process, we referred to some existing research; for example, the fish face recognition study based on rotated object detection considered 12 days for data collection Li et al. (2022). We adopted ten days from October 15, 2023, to October 25, 2023. During this period, each individual was removed from the rearing pool and placed in a small transparent fish tank for imaging every day. We used three cameras to construct a multi-view video recording system to obtain comprehensive data and ensure that the cameras were synchronized in time. Our experimental setup for multi-view re-identification, along with multi-view video synchronization and feature extraction techniques, is shown in Figure 3.

This research set up synchronization triggers to ensure that multiple cameras started capturing video simultaneously. In addition, during the acquisition process, we use timestamps for each frame to determine the temporal relationship between video frames. This setup allowed us to photograph the fish from three directions Yadav and Vishwakarma (2020): front, top, and side, providing a more three-dimensional and comprehensive dataset for subsequent experiments. We followed the camera angle setting of Wu et al.'s work Wu et al. (2022) to ensure that our methodology aligns with established standards and comparability. Additionally, we have made comprehensive adjustments to ensure clear images are obtained in our data sets. At the end of the experiment, we captured 109 video sequences, including 48 videos of individual fish instances and 61 videos of scenes containing multiple fish. During the manual labeling process, we obtained 10,300 images covering various angles, such as the front, left, and top of all the fish. A total of 39,088 bounding boxes were labeled Wei et al. (2018), supporting the accuracy and richness of the experimental results. This tedious and systematic data collection provided a solid foundation for our subsequent study. All labeled data from 10,300 sheets were used during the detection experiments.

4.2 Labeling

Before delving into fish re-identification (re-ID) tasks with the FS48 dataset, we should ensure accurate labeling of the fish in the images. This step was imperative for providing reliable ground truth data for model training and evaluation. This work presents a fish re-ID dataset named FS48, specifically labeled to support re-ID experiments.

The labeling process involved meticulous manual annotation of the bounding boxes around each fish in the video sequences Baltieri et al. (2011). We cropped these bounding boxes for the following re-ID experiments. The detailed structure of the fish re-ID process is illustrated in Figure 4. For the re-ID experiments, the training and gallery sets utilized manually labeled bounding boxes Li et al. (2012)

to emulate a manual query request, while the query set employed bounding boxes generated by the Co-DETR detector to simulate an automatic re-ID process. The details of the proposed FS48 dataset used in each experimental setup is presented in Table 1.

4.3 Comparison with existing fish datasets

We conducted a comparative analysis of existing typical datasets along five dimensions: 1) the number of images within the dataset; 2) the tasks intended to be accomplished by the image dataset; 3) whether the annotated entities in the dataset possess unique identifiers; 4) the quantity of bounding boxes (BBOX) present in the dataset; and 5) the number of cameras utilized during data collection. We compared these aspects between several established datasets and our proposed FS48 dataset.

In Table 2, a direct comparison is presented between these datasets and our FS48 dataset. Our FS48 dataset exhibits fundamental disparities compared with existing public datasets, with significant differences in the tasks it aims to fulfill. The unique feature within our proposed FS48 dataset is of particular importance, where each fish is endowed with a distinctive identifier, ensuring precise and accurate identification. This aspect holds paramount practical importance for future marine aquaculture and economic activities.

5 Experiments

5.1 Implementation details

5.1.1 Baselines

The main highlight of our baseline is the use of three different settings, and the excellent experimental results obtained under these settings prove the feasibility of our baseline method. Specifically, our baseline method first captures the video frames of the fish synchronously through three viewing angles and then inputs these video frames to the backbone for feature extraction. The extracted features are subsequently fused in the fully connected (FC) layer, followed by normalization. Finally, the loss function is used to reduce the distance between the corresponding images. We extract video frames every 5 seconds to avoid providing frame-to-frame semantic details. This comprehensive process enhances the accuracy and robustness of our baseline model in capturing multi-angle information about fish.

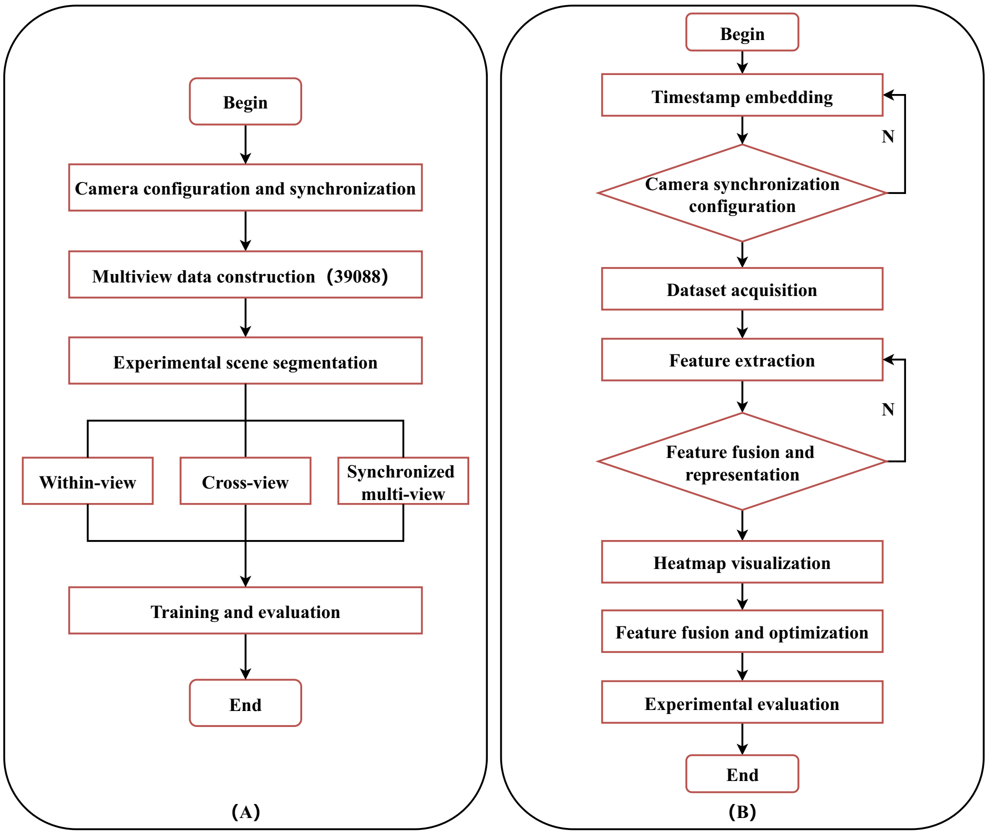
Begin
Camera configuration and synchronization
Multiview data construction (39088)
Experimental scene segmentation
Within-view
Cross-view
Synchronized multi-view
Training and evaluation
End
(A)
Begin
Timestamp embedding
N
Camera synchronization configuration
Dataset acquisition
Feature extraction
N
Feature fusion and representation
Heatmap visualization
Feature fusion and optimization
Experimental evaluation
End
(B)

FIGURE 3
Overall Fish Re-identification (Fish ReID) Flowchart: Experimental Setup for multi-view re-identification **(A)** and techniques for multi-view video synchronization and feature extraction **(B)**.

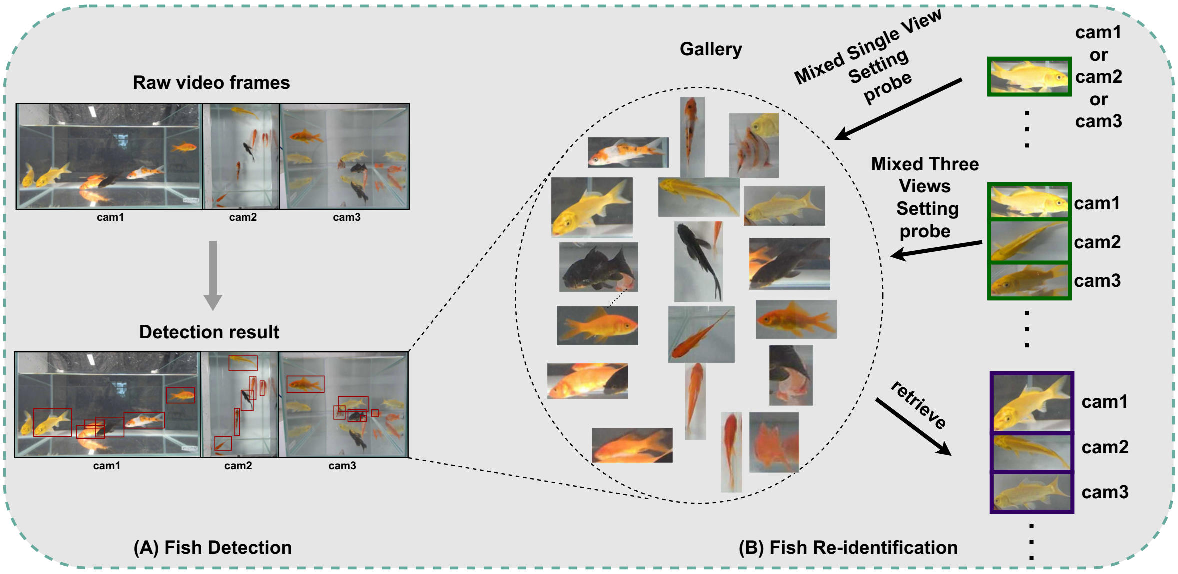

FIGURE 4
Pipeline of an end-to-end fish re-ID system. It consists of two modules: detection **(A)** and fish re-identification **(B)**.

5.1.2 Evaluation metrics

In the field of visual inspection, the indicators used to evaluate the detection accuracy usually include true positive (TP), false positive (FP), and false negative (FN), as well as the average precision (AP) and mean average precision (mAP) used in this paper. Among them, TP, FP, and FN are usually used to calculate the accuracy and recall rate of the detection model, and the accuracy (precision) is used to evaluate how many of the samples predicted by the model are real positive samples. Its calculation formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The recall rate (Recall) is used to assess how many actual positive samples are successfully detected by the model.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Through precision and recall, we can further obtain the evaluation index of average accuracy (AP). The AP considers the accuracy of the model under different confidence thresholds. The index corresponding to the AP is the mean average precision (mAP). In multi-category visual inspection, each category calculates an AP and then averages all categories of the AP to obtain mAP.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

This paper divides all the fishes into 48 categories; thus, N = 48. This paper divided all the fishes into 48 categories, where n is the total number of categories. In this study, all fishes were divided into 48 categories. Considering the number of 48 categories and the requirements for evaluation, we chose precision and mAP as the primary evaluation indicators of interest in this study.

TABLE 1 The details of the FS48 dataset under three re-ID experimental setups, C1 is the front view, C2 is the top view, and C3 is the side view.

Experimental setting		bbox_train	bbox_gallery	query
Within-view	C1	6,475	5,832	262
	C2	6,607	5,962	242
	C3	6,321	5,779	216
Cross-view		19,403	17,573	720
Synchronized Multi-view		18,558	14,955	612

5.1.3 Experimental setting

- Within-view fish re-ID:

Under this experimental setting, this setup is more straightforward since the captured images are from a known camera, and both the query and gallery images are from different known cameras. It allows three distinct experiments to be conducted.

- Cross-view fish re-ID:

We used the images captured from the three camera views for training and testing in this experimental setting. We performed experiments under the view-agnostic setting, indicating that both the query and gallery images could come from any camera view. In other words, the model was optimized to perform cross-view fish re-ID, learning feature representations that are robust to view changes. Notably, the input of the model is a single image.

- Synchronized multi-view fish re-ID:

Under this setting, we aimed to demonstrate that combining the three camera views could lead to more robust and accurate fish reidentification. In parallel, we concatenate the images from the

TABLE 2 Direct comparison between OzFish, Fish4-Knowledge (F4K), Fishnet Open Images, DeepFish, and our proposed FS48 dataset according to different properties.

Dataset	Images	Tasks	Bbox	ID	Camera
OzFish Al Muksit et al. (2022)	1,800	Detection	about 43k	–	1
Fish4-Knowledge (F4K) Qin et al. (2016)	27,320	Clf	–	–	1
Fishnet Open Images Kay and Merrifield (2021)	86,029	Detection, Clf	406,463	–	1
DeepFish Saleh et al. (2020)	39,726	Clf, Cnt, Loc, Seg	about 15k	–	1
FS48	10,300	re-ID	39,088	48	3

Clf, Cnt, Loc, Seg refers to the task of classification, counting, localization, and segmentation.

three camera views (front, side, and top views). To ensure consistent feature representations between the query and gallery, we set the input of both the query and the gallery as a mixture of the three images from the corresponding camera views. Notably, the three concatenated images were strictly time-synchronized. The pipeline under this setting is shown in Figure 2.

5.2 Fish detection

5.2.1 Fish only detection

We report the results using the classical detection models Faster-RCNN Ren et al. (2015), Reppoints Yang et al. (2019), Foveabox Kong et al. (2020), and YoloX Ge et al. (2021) in MMDetection, and the newest models, GFL Li et al. (2020), and Co-DETR Zong et al. (2023), are used as advanced object detection frameworks. These several detectors are fine-tuned via pre-trained models on the ImageNet dataset, and only one target, fish, is detected in this section. Note that these several detection models use default settings during training. We use average precision (AP) to measure the detection performance. Table 3 shows that in the detection experiments in which all the fish are treated as a single class, Faster-RCNN, Reppoints, Foveabox, YoloX, and the newest GFL and Co-DETR detectors achieve an average accuracy of approximately 70% at IoU 95, with Co-DETR performing the best with an AP of 0.838 at AP_{95}, 0.975 at AP_{50}. The visualization of Co-DETR is shown in Figure 5. Notably, multi-view fish re-identification makes the re-identification of high-density fish more accurate because the occlusion phenomenon becomes more severe in the case of high fish density.

TABLE 3 Fish-only detection results.

Detectors	AP_{95}	AP_{50}
Faster-RCNN Ren et al. (2015)	0.770	0.962
Reppoints Yang et al. (2019)	0.582	0.916
Foveabox Kong et al. (2020)	0.740	0.963
YoloX Ge et al. (2021)	0.618	0.917
GFL Li et al. (2020)	0.675	0.918
Co-DETR Zong et al. (2023)	**0.838**	**0.975**

We regard all the fish as one class and use several representative detection backbones to conduct detection experiments on our data sets. Bold highlighted as best.

5.2.2 Fish classification results with ID detection (48 fish)

In this section, the basic setup is the same as that in Section 5.1, except we further consider the identity information of the fish by treating individual fish as different categories. Specifically, we treat the 48 individual fish as 48 distinct categories during training Chang et al. (2018). The experimental results illustrated in Table 4 indicate that such a category-aware training approach significantly decreases the detection performance of the fish. The best Faster-RCNN AP_{95} for detection is only 0.082, with the AP_{50} reaching only 0.208.

Under this experimental setup, the detection model faces more significant challenges in localizing and identifying different fish while considering different fish identities, thus leading to an overall decrease in performance. Furthermore, we demonstrated the intrinsic limitations of the above detection-based algorithms, which cannot detect unseen fish while assigning the correct IDs. This fact indicates that the model's performance will be severely constrained when encountering new, untrained fish in real-world scenarios, thus significantly reducing its utility. Therefore, to remedy this shortcoming, the following section focuses on fish re-identification experiments by introducing the re-ID technique, in which an individual fish's identity is considered essential for re-identification experiments. In this research direction, we are committed to the re-ID model to capture fish identity information effectively and thus improve the model's recognition accuracy in multiple viewpoints and scenarios. By introducing re-ID based on the detection results, we expect to realize a more accurate and robust re-identification of fish identity and provide theoretical support and practical guidance for constructing a multi-view fish re-identification system. The exploration of this research direction will hopefully overcome the limitations of detection models in practical applications and provide more powerful technical support for the real-time identification of multiple unknown fish species.

5.3 Fish re-identification

In the re-ID experiments, the bounding boxes were first calibrated manually to perform the experiments, and the fish re-ID was performed under three different settings, namely, within-view, cross-view, and synchronized multi-view. We fine-tuned the entire model for 40 epochs using the Adam optimizer. We set the initial learning rate to 1e-4 and measured the fish re-ID performance via the checking accuracy and mean average precision (mAP).

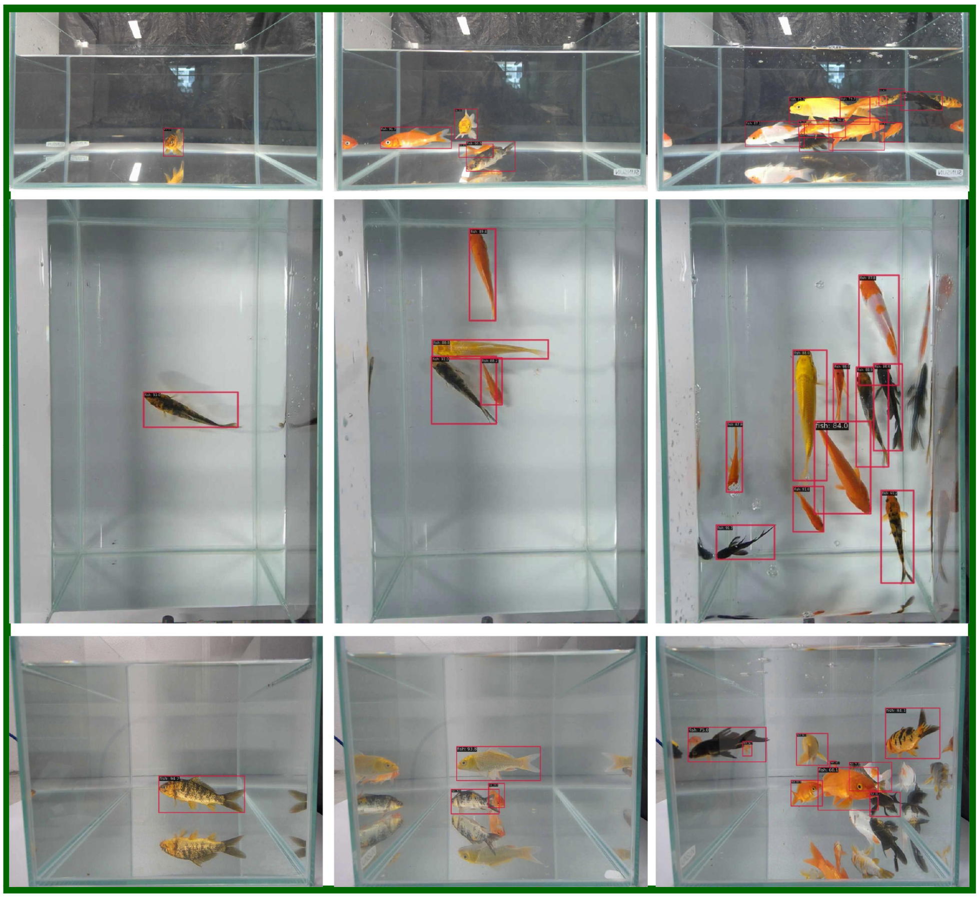

FIGURE 5
Visualization of Co-DETR under the “fish only” setting.

We present the fish re-ID performance under the within-view setting in Table 5. For the front view, the combination of SoftTriplet loss and ResNet-50 achieved the highest mAP@all value of 39.31. From a top view, the SoftTriplet Loss and ResNet-50 combination also achieved the highest mAP@all value of 37.51. For the side view, the combination of NormSoftMax and VGG-16 achieved the highest mAP@all value of 38.14. Hence, the benchmark model that performs well across all views combines SoftTriplet loss and ResNet-50. We can conclude that the front view provides more semantic information about the fish than the top and side views.

We present the fish re-ID performance under the cross-view setting in Table 6. The SoftTriplet Loss and ResNet-50 combination achieved the best performance with a mAP@all value of 32.78. As shown in Tables 5 and 6, we can conclude that SoftTriplet Loss combined with ResNet-50 achieved the best performance in two different experimental settings.

We performed time-independent in within-view and cross-view, respectively, in Tables 7 and 8, where we trained FSNet using video frames captured in the first five days and evaluated video captured in the next five days. Specifically, we choose the loss functions with the highest recognition performance and backbone for evaluation in Tables 5 and 6. The experimental results in Tables 7 and 8 show that our model still has high recognition performance in time-independent data sets.

We observe that multiple fish may easily trigger occlusions under the within-view (Figure 5). The occlusions would hinder some key features and reduce the performance. The re-identification system cannot perform robust and accurate feature extraction and identification. Thus, we propose mixing the information of the images captured from the three camera views. We believe that different views can provide more complementary information through information fusion, which can lead to more accurate re-ID performance. We used the query images under cross-view and multiple synchronized multi-view settings to verify our idea. By comparing the experimental results of Tables 6

TABLE 4 Fish classification results with ID detection (48 fish).

Detectors	AP_{95}	AP_{50}
Faster-RCNN Ren et al. (2015)	**0.082**	**0.208**
Reppoints Yang et al. (2019)	0.045	0.098
Foveabox Kong et al. (2020)	0.042	0.109
YoloX Ge et al. (2021)	0.042	0.160
GFL Li et al. (2020)	0.052	0.091
Co-DETR Zong et al. (2023)	0.003	0.010

We further consider the identity information of the fish by treating each individual as a different category. Bold highlighted as best.

TABLE 5 Within-view fish re-ID.

Loss functions	Backbone	Front view				Top view				Side view			
		P@5	mAP@10	P@10	mAP@all	P@5	mAP@10	P@10	mAP@all	P@5	mAP@10	P@10	mAP@all
SoftTriplet Qian et al. (2019)	VGG16	**58.55**	**49.17**	**56.41**	35.52	52.15	42.33	49.30	34.11	59.17	48.38	55.42	35.42
ArcFace Loss Deng et al. (2022)		50.15	39.17	46.68	22.23	50.25	36.8	45.25	23.96	53.61	40.65	47.92	24.65
ProxyNCA Yang et al. (2022)		52.67	42.04	49.43	30.03	47.77	37.05	44.92	29.55	55.09	44.01	51.16	31.81
ProxyAnchor Kim et al. (2020)		57.33	48.44	55.50	34.39	52.98	42.81	49.92	32.73	57.50	48.08	55.56	33.95
NormSoftMax Zhai and Wu (2019)		58.55	48.39	55.42	**35.69**	**53.47**	**43.21**	**51.24**	**34.67**	**61.85**	**52.13**	**58.84**	**38.14**
SoftTriplet Qian et al. (2019)	ResNet-50	**67.71**	**57.94**	**64.43**	**39.31**	62.07	**52.67**	**59.42**	**37.51**	65.28	54.78	61.62	**37.19**
ArcFace Loss Deng et al. (2022)		65.73	55.83	62.63	36.13	56.69	45.36	52.44	30.84	63.33	52.73	59.4	35.31
ProxyNCA Yang et al. (2022)		61.30	51.12	58.17	36.12	55.62	45.88	53.18	35.68	59.35	50.14	56.71	36.88
ProxyAnchor Kim et al. (2020)		65.11	56.56	63.28	34.95	**62.23**	52.08	57.98	34.45	**67.04**	**56.44**	**62.27**	33.80
NormSoftMax Zhai and Wu (2019)		67.40	56.19	63.17	37.48	60.99	50.12	57.60	35.76	64.72	54.68	61.67	36.88

The experimental results of fish re-identification on video frames captured from the same perspective suggest that front views provide more informative features during the re-identification process. Bold highlighted as best.

and 9, we can find that the mAP@all scores of the synchronized multi-view experimental setting are much higher than those of the cross-view experimental setting under various loss restrictions and backbones. The results indicate that synchronized multi-view images can provide richer information than cross-view images can provide to support fish re-identification tasks. Our FSNet can effectively utilize the multi-view information for fish identification.

The result in Table 10 illustrates the impact of the backbone (VGG-16 and ResNet-50) and five loss functions in Table 9 on all metrics (P@5, mAP@10, P@10, mAP@all), with the most notable effects observed in P@5 and mAP@10. In this case, the p-value is used as the evaluation criterion. A p-value less than 0.05 helps strengthen the experimental conclusions in Table 9, while a p-value less than 0.01 strengthens the experimental conclusions to a greater extent. Furthermore, the loss functions significantly influence mAP@10 and P@10, emphasizing their pivotal role. These results validate the critical significance of the backbone architecture and affirm the crucial importance of method selection, thus validating the decision to employ the ResNet-50 and NormSoftMax (the highest mAPall) in Table 9 to build the proposed FSNet.

Automatically detecting and re-identifying an unknown fish is a fundamental requirement of fish re-ID. Using the best detection model trained under the detection experiment (Co-DETR), which only trains fish as a category of the model, we can obtain bounding boxes as queries under within-view and cross-view settings. To maintain the same scale and variety of Table 1, the single-fish query dataset also includes 612 samples, while the train and gallery datasets still used manually annotated data under the synchronized multi-view setting. Notably, the bounding boxes generated by the detection model do not include fish IDs. Thus, we only consider the single-fish scene case (only one fish exists in the tank) for cross-view and synchronized multi-view automatically fish re-ID task with Co-DETR detector. We evaluate and locate the best loss functions and backbones for different settings in Table 11. Since we did not consider the multi-fish scene for the automatic re-ID task, the scores in Table 11 are even higher than those in Tables 6 and 9.

The heatmap visualizations in Figure 6 demonstrate that the front view contains the key information, including areas surrounding the fish eyes for identification. In contrast, the semantic information in the top and side views is related mainly to contours (such as the dorsal fin, pectoral fin, ventral fin, anal fin, and caudal fin). In other words, the front view plays a vital role in providing feature information for judgment in recognition. The top and side views compensate for the information deficiency the front view. Our FSNet network extracts features from the three views and builds feature embedding to complete the fish re-identification process. We show four examples in Figure 6; the images in line one show an example of carp (#05), and the other three examples are carp (#00, #02, and #14). To verify the proposed FSNet can capture the key features, we use the Grad-CAM method to obtain the heat map of FSNet, focusing on the convolutional layers to visualize the important regions identified by the network Selvaraju et al. (2017). From top to bottom, there are two different species of fish. From the cam1 column, we can find that the ventral fin of the carp is highlighted on the heat map of FSNet, and the dorsal fin is

TABLE 6 Cross-view fish re-ID. Experiments are carried out under view-independent settings.

Loss functions	Backbone	P@5	mAP@10	P@10	mAP@all
SoftTriplet Qian et al. (2019)	VGG-16	54.75	44.78	52.10	**30.30**
ProxyNCA Yang et al. (2022)		47.67	37.89	45.46	27.71
ProxyAnchor Kim et al. (2020)		53.58	43.73	51.19	26.48
NormSoftMax Zhai and Wu (2019)		**55.28**	**45.33**	**52.72**	29.82
ArcFace Loss Deng et al. (2022)		46.94	34.95	43.44	18.00
SoftTriplet Qian et al. (2019)	ResNet-50	**66.72**	**57.40**	**64.04**	**32.78**
Proxynca Yang et al. (2022)		56.75	47.34	54.46	31.10
ProxyAnchor Kim et al. (2020)		64.28	55.15	61.85	29.66
NormSoftMax Zhai and Wu (2019)		63.72	53.52	60.74	31.36
ArcFace Loss Deng et al. (2022)		60.11	50.48	57.36	30.49

Bold highlighted as best.

TABLE 7 Time-independent within-view fish re-ID.

Loss functions	Backbone	Frontal view				Bird of view				Side of view			
		P@5	mAP@10	P@10	mAP@all	P@5	mAP@10	P@10	mAP@all	P@5	mAP@10	P@10	mAP@all
NormSoftMax Zhai and Wu (2019)	VGG-16	54.24	46.99	55.21	**33.46**	52.13	41.98	50.01	**32.84**	58.06	50.12	57.22	**36.87**
SoftTriplet Qian et al. (2019)	ResNet-50	65.37	58.04	61.43	**37.66**	60.11	49.85	57.31	**36.13**	64.79	55.46	60.33	**35.29**

We train the FSNet with video frames captured in the first five days and evaluate videos captured in the next five days. Specifically, we choose the loss functions with the highest recognition performance as well as backbones in Table 5 for evaluation. Bold highlighted as best.

TABLE 8 Time-independent cross-view fish re-ID.

Loss function	Backbone	P@5	mAP@10	P@10	mAP@all
SoftTriplet Qian et al. (2019)	VGG-16	54.98	44.11	51.67	28.87
	ResNet-50	64.12	55.03	60.91	30.12

We train the FSNet with video frames captured in the first five days and evaluate videos captured in the next five days. Specifically, we choose the loss functions with the highest recognition performance as well as backbones in Table 6 for evaluation.

TABLE 9 Synchronized multi-view fish re-ID.

Loss functions	Backbone	P@5	mAP@10	P@10	mAP@all
SoftTriplet Qian et al. (2019)	VGG-16	**91.27**	**81.40**	**83.58**	49.22
ProxyNCA Yang et al. (2022)		86.50	73.04	75.56	39.49
ProxyAnchor Kim et al. (2020)		87.84	77.61	79.82	45.94
NormSoftMax Zhai and Wu (2019)		90.62	80.93	82.88	**50.61**
ArcFace Loss Deng et al. (2022)		85.13	73.81	76.83	42.66
SoftTriplet Qian et al. (2019)	ResNet-50	**79.18**	**74.03**	**77.43**	52.07
Proxynca Yang et al. (2022)		73.53	67.21	71.63	47.06
ProxyAnchor Kim et al. (2020)		77.32	72.52	76.00	49.50
NormSoftMax Zhai and Wu (2019)		77.22	72.94	76.52	**52.17**
ArcFace Loss Deng et al. (2022)		77.68	71.01	74.75	50.17

We concatenate the images from three camera views. To ensure that the feature representation between the query and the gallery is consistent, we set the input of the query and the library to a mixture of three images from the corresponding camera view. Note that these three images are strictly synchronized. Bold highlighted as best.

TABLE 10 An analysis of variance was performed on the loss functions and backbone cross-performance indicators in Table 9.

Metric	Source	Sum Sq	df	F	P(>F)
P@5	Loss functions	33.479	4	2.880	0.165
P@5	Backbone	318.434	1	109.555	0.000**
mAP@10	Loss functions	80.162	4	9.563	0.025*
mAP@10	Backbone	84.565	1	40.354	0.003**
P@10	Loss functions	64.422	4	10.050	0.023*
P@10	Backbone	49.908	1	31.142	0.005*
mAP@all	Loss functions	86.686	4	5.652	0.061
mAP@all	Backbone	53.13	1	13.857	0.020*

*P<0.05; **P<0.001; df, degrees of freedom; F, F-statistic; P(>F), p-value for the F-statistic.

TABLE 11 Performance of the bounding boxes generated by the Co-DETR detection model as a query.

Loss functions	Backbone	View	P@5	mAP@10	P@10	mAP@all
SoftTriplet Qian et al. (2019)	ResNet-50	Cross-view	77.82	67.57	71.74	34.77
NormSoftMax Zhai and Wu (2019)		Synchronized multi-view	77.52	72.46	76.18	52.27

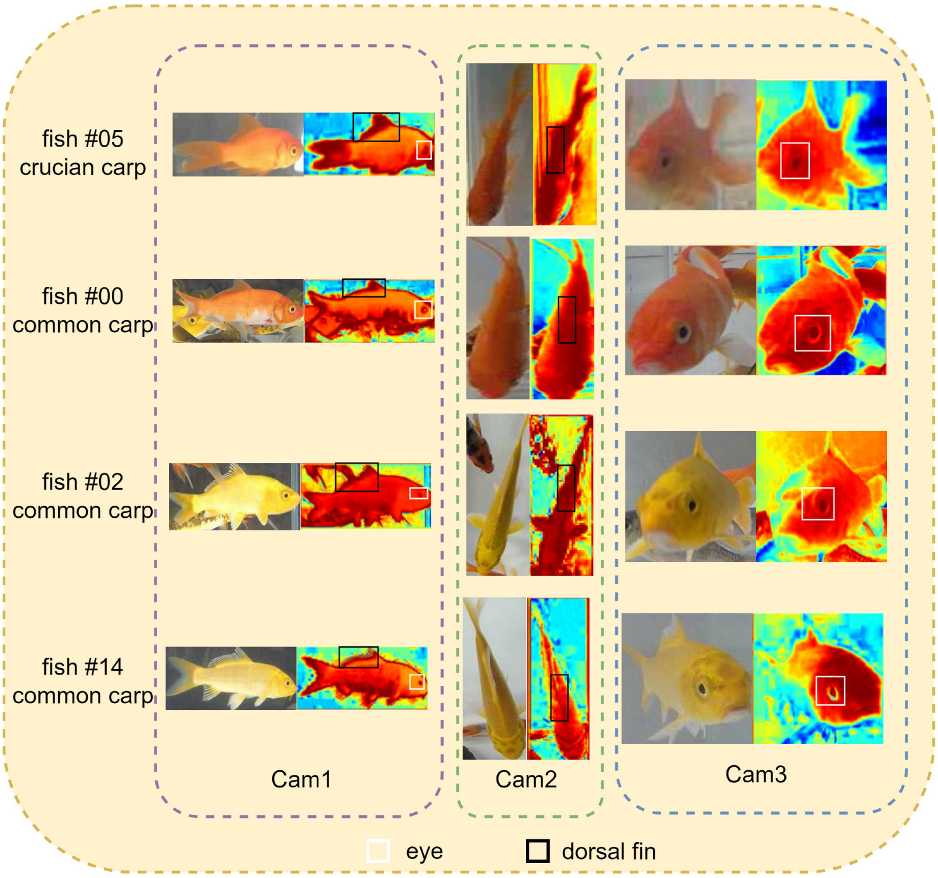

FIGURE 6
Heatmap visualization of images from three views, with the image in the left column of the same shot being the original image and the heatmap image in the right column. The first one is the crucian carp, and the last three are the common carp.

TABLE 12 Ablation study of the proposed FSNet.

Methods	P@5	mAP@10	P@10	mAP@all
baseline	63.72	53.52	60.74	31.36
baseline (multi-view)	77.22	72.94	76.52	52.17
FSNet	79.24	71.98	80.15	53.33

TABLE 13 We used the T-test method to verify the results in Table 12 is reliable.

Metric	T-statistic	P-value
P@5	-11.171	0.000**
mAP@10	-1.527	0.205
P@10	-16.02	0.001*
mAP@all	-3.918	0.017*

The experiment calculated T statistics and degrees of freedom from three independent sets of data from baseline and FSNet.
*P<0.05, **P<0.001.

highlighted in a relatively straight shape on the heat map. The abdominal fin of the crucian carp is highlighted at a higher position on the heat map, and the dorsal fin is highlighted in a relatively round shape on the heat map. The two yellow fish have similar contours and colors but different texture shapes. The heat map shows that the FSNet's attention usually focuses on the eyes and fin areas to distinguish different fish identities. These results prove that our model can distinguish between similar but different fish based on small but critical features.

The detection and re-identification experiments perform well on the FS48 dataset, indicating that our FS48 dataset is reliable and practical. To evaluate the efficiency of the strategy, we set a baseline model (shared backbone) and designed an ablation study in Table 12. The baseline used a shared backbone to extract information from three perspectives and perform feature fusion. As shown in Table 12, our network FSNet, which can utilize the information from the three viewpoints more effectively, performs better than the baseline. The results verify the effectiveness of the joint representation strategy for multi-view fish re-identification tasks. The T-test results in Table 13 further verify the above conclusions.

6 Conclusions

To the best of our knowledge, this work takes the first step toward deep learning based fish re-identification, and the reliability of this work is confirmed, mainly because of the large number of experiments we conducted. We present FS48, a fish re-identification multi-view dataset comprising 10,300 three-view images from 48 crucian carp and carp, accompanied by 39,088 manually labeled bounding boxes. Using the FS48 dataset, we have developed a robust and accurate fish identification framework called FSNet to facilitate the advancement of aquatic species identification and promote research in fish monitoring and aquaculture.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Animal ethics approval is not required for the use of bivalves, yabbies, crayfish or other aquatic creatures not considered to be animals under the Animal Care and Protection Act 2001 (Qld) (the Act). Observation and basic husbandry of fish in a classroom aquarium is a Category 1 activity and does not require animal ethics approval.

Author contributions

SF: Data curation, Software, Writing – original draft, Writing – review & editing. CS: Writing – review & editing. HF: Writing – review & editing. ZY: Funding acquisition, Resources, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Key Special Project of the National Key Research and Development Program of China under Grant No. 2022YFD2401304; the Project of Sanya Yazhou Bay Science and Technology City under Grant No. SKJC-2023-01-004 and the finance science and technology Q19 project of 630 Hainan Province of China under Grant No. ZDKJ202017.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ahmed, E., Jones, M., and Marks, T. K. (2015). "An improved deep learning architecture for person re-identification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3908–3916 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/CVPR.2015.7299016

Al Muksit, A., Hasan, F., Emon, M. F. H. B., Haque, M. R., Anwary, A. R., and Shatabda, S. (2022). Yolo-fish: A robust fish detection model to detect fish in realistic underwater environment. *Ecol. Inf.* 72, 101847. doi: 10.1016/j.ecoinf.2022.101847

Alsmadi, M. K., and Almarashdeh, I. (2022). A survey on fish classification techniques. *J. King Saud University-Computer Inf. Sci.* 34, 1625–1638. doi: 10.1016/j.jksuci.2020.07.005

Alsmadi, M. K., Tayfour, M., Alkhasawneh, R. A., Badawi, U., Almarashdeh, I., and Haddad, F. (2019). Robust feature extraction methods for general fish classification. *Int. J. Electrical Comput. Eng. (2088-8708)* 9, 5192–5204. doi: 10.11591/ijece.v9i6.pp5192-5204

Baltieri, D., Vezzani, R., and Cucchiara, R. (2011). "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the 2011 joint ACM workshop on human gesture and behavior understanding*, vol. J-HGBU '11. (Association for Computing Machinery, New York, NY, USA), 59–64. doi: 10.1145/2072572.2072590

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Barbedo, J. G. A. (2022). A review on the use of computer vision and artificial intelligence for fish recognition, monitoring, and management. *Fishes* 7, 335. doi: 10.3390/fishes7060335

Buckley, R. M., West, J. E., and Doty, D. C. (1994). Internal micro-tag systems for marking juvenile reef fishes. *Bull. Mar. Sci.* 55, 848–857.

Castillo, G., Morinaka, J., Fujimura, R., DuBois, J., Baskerville-Bridges, B., Lindberg, J., et al. (2014). Evaluation of calcein and photonic marking for cultured delta smelt. *North Am. J. Fisheries Manage.* 34, 30–38. doi: 10.1080/02755947.2013.839970

Chang, X., Huang, P.-Y., Shen, Y.-D., Liang, X., Yang, Y., and Hauptmann, A. G. (2018). "Rcaa: Relational context-aware agents for person search," in *Computer vision – ECCV 2018*. Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Springer International Publishing, Cham), 86–102.

Chen, G., Sun, P., and Shang, Y. (2017). "Automatic fish classification system using deep learning," in *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. 24–29 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/ICTAI.2017.00016

Chuang, M.-C., Hwang, J.-N., Ye, J.-H., Huang, S.-C., and Williams, K. (2016). Underwater fish tracking for moving cameras based on deformable multiple kernels. *IEEE Trans. Systems Man Cybernetics: Syst.* 47, 2467–2477. doi: 10.1109/TSMC.2016.2523943

Cooke, S. J., Midwood, J. D., Thiem, J. D., Klimley, P., Lucas, M. C., Thorstad, E. B., et al. (2013). Tracking animals in freshwater with electronic tags: past, present and future. *Anim. Biotelemetry* 1, 1–19. doi: 10.1186/2050-3385-1-5

Dare, M. R. (2003). Mortality and long-term retention of passive integrated transponder tags by spring chinook salmon. *North Am. J. Fisheries Manage.* 23, 1015–1019. doi: 10.1577/M02-106

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., and Zafeiriou, S. (2022). Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 5962—5979. doi: 10.1109/tpami.2021.3087709

Fullerton, I. J., Kell, J. H., and Mills, M. K. (1990). "*Traffic detector handbook.*," in *Tech. rep* (Federal Highway Administration, United States).

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). "Yolox: Exceeding yolo series in 2021," in *arXiv preprint arXiv:2107.08430*.

Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., and Hartley, R. (2006). "Person reidentification using spatiotemporal appearance," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Massachusetts, United States: IEEE) Vol. 2. 1528–1535.

Gladju, J., Kamalam, B. S., and Kanagaraj, A. (2022). Applications of data mining and machine learning framework in aquaculture and fisheries: A review. *Smart Agric. Technol.* 2, 100061. doi: 10.1016/j.atech.2022.100061

Huang, T., and Russell, S. (1997). Object identification in a bayesian context. *In IJCAI (Citeseer)* 97, 1276–1282.

Hühn, D., Klefoth, T., Pagel, T., Zajicek, P., and Arlinghaus, R. (2014). Impacts of external and surgery-based tagging techniques on small northern pike under field conditions. *North Am. J. fisheries Manage.* 34, 322–334. doi: 10.1080/02755947.2014.880762

Kaiming, H., Xiangyu, Z., Shaoqing, R., and Jian, S. (2016). Deep residual learning for image recognition. *2016 IEEE Conf. Comput. Vision Pattern Recognition (CVPR)* 1, 770. doi: 10.1109/cvpr.2016.90

Kay, J., and Merrifield, M. (2021). "The fishnet open images database: A dataset for fish detection and fine-grained categorization in fisheries," in ArXiv *abs/2106.09178*.

Kim, S., Kim, D., Cho, M., and Kwak, S. (2020). "Proxy anchor loss for deep metric learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3235–3244 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/CVPR42600.2020.00330

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., and Shi, J. (2020). Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* 29, 7389–7398. doi: 10.1109/TIP.83

Kumar, G., and Bhatia, P. K. (2014). "A detailed review of feature extraction in image processing systems," in *2014 Fourth international conference on advanced computing & communication technologies*. 5–12 (Los Alamitos, CA, USA: IEEE Computer Society).

Li, D., Su, H., Jiang, K., Liu, D., and Duan, X. (2022). Fish face identification based on rotated object detection: dataset and exploration. *Fishes* 7, 219. doi: 10.3390/fishes7050219

Li, W., Zhao, R., and Wang, X. (2012). "Human reidentification with transferred metric learning," in *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part I*, Vol. ACCV'12. 31–44 (Berlin, Heidelberg: Springer-Verlag). doi: 10.1007/978-3-642-37331-23

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., et al. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* 33, 21002–21012.

Lindberg, J. C., Tigan, G., Ellison, L., Rettinghouse, T., Nagel, M. M., and Fisch, K. M. (2013). Aquaculture methods for a genetically managed population of endangered delta smelt. *North Am. J. Aquaculture* 75, 186–196. doi: 10.1080/15222055.2012.751942

Liu, D., Xianhui, W., and Youling, Z. (2022). Research on an improved fish recognition algorithm based on yolox. *ITM Web Conf.* 47, 2003. doi: 10.1051/itmconf/20224702003

Liu, X., Liu, W., Mei, T., and Ma, H. (2017). Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multimedia* 20, 645–658. doi: 10.1109/TMM.2017.2751966

Macaulay, G., Warren-Myers, F., Barrett, L. T., Oppedal, F., Føre, M., and Dempster, T. (2021). Tag use to monitor fish behaviour in aquaculture: a review of benefits, problems and solutions. *Rev. Aquaculture* 13, 1565–1582. doi: 10.1111/raq.12534

Marini, S., Fanelli, E., Sbragaglia, V., Azzurro, E., Del Rio Fernandez, J., and Aguzzi, J. (2018). Tracking fish abundance by underwater image recognition. *Sci. Rep.* 8, 13748. doi: 10.1038/s41598-018-32089-8

Mei, Y., Sun, B., Li, D., Yu, H., Qin, H., Liu, H., et al. (2022). Recent advances of target tracking applications in aquaculture with emphasis on fish. *Comput. Electron. Agric.* 201, 107335. doi: 10.1016/j.compag.2022.107335

Murphy, B. R., Willis, D. W., and Society, A. F. (1996). *Fisheries techniques (American fisheries society), 2nd.* (New Jersey, U.S.: Citeseer).

Musselman, W. C., Worthington, T. A., Mouser, J., Williams, D. M., and Brewer, S. K. (2017). Passive integrated transponder tags: review of studies on warmwater fishes with notes on additional species. *J. Fish Wildlife Manage.* 8, 353–364. doi: 10.3996/122016-JFWM-091

Osbourn, M. S., Hocking, D. J., Conner, C. A., Peterman, W. E., and Semlitsch, R. D. (2011). Use of fluorescent visible implant alphanumeric tags to individually mark juvenile ambystomatid salamanders. *Herpetological Rev.* 42, 43–47.

Qian, Q., Shang, L., Sun, B., Hu, J., Tacoma, T., Li, H., et al. (2019). "Softtriple loss: Deep metric learning without triplet sampling," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6449–6457 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/ICCV.2019.00655

Qin, H., Li, X., Liang, J., Peng, Y., and Zhang, C. (2016). Deepfish: Accurate underwater live fish recognition with a deep architecture. *Neurocomputing* 187, 49–58. doi: 10.1016/j.neucom.2015.10.122

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* 28, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Ren, H., Zheng, Z., Wu, Y., and Lu, H. (2023a). Daco: domain-agnostic contrastive learning for visual place recognition. *Appl. Intell.* 1–14. doi: 10.1007/s10489-023-04629-x

Ren, H., Zheng, Z., Wu, Y., Lu, H., Yang, Y., Shan, Y., et al. (2023b). Acnet: Approaching-and-centralizing network for zero-shot sketch-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* 9, 5022–5035. doi: 10.1109/TCSVT.2023.3248646

Runde, B. J., Buckel, J. A., Bacheler, N. M., Tharp, R. M., Rudershausen, P. J., Harms, C. A., et al. (2022). Evaluation of six methods for external attachment of electronic tags to fish: assessment of tag retention, growth and fish welfare. *J. Fish Biol.* 101, 419–430. doi: 10.1111/jfb.v101.3

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., and Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* 10, 14671. doi: 10.1038/s41598-020-71639-x

Saleh, A., Sheaves, M., and Rahimi Azghadi, M. (2022). Computer vision and deep learning for fish classification in underwater habitats: A survey. *Fish Fisheries* 23, 977–999. doi: 10.1111/faf.12666

Sandford, M., Castillo, G., and Hung, T.-C. (2020). A review of fish identification methods applied on small fish. *Rev. Aquaculture* 12, 542–554. doi: 10.1111/raq.12339

Schumann, D. A., Koupal, K. D., Hoback, W. W., and Schoenebeck, C. W. (2013). Evaluation of sprayed fluorescent pigment as a method to mass-mark fish species. *Open Fish Sci. J.* 6, 41–47. doi: 10.2174/1874401X01306010041

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017 IEEE/CVF International Conference on Computer Vision (ICCV) (Los Alamitos, CA, USA:IEEE Computer Society). 618–626.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*. doi: 10.48550/arXiv.1409.1556

Skalski, J. R., Buchanan, R. A., and Griswold, J. (2009). Review of marking methods and release-recapture designs for estimating the survival of very small fish: examples from the assessment of salmonid fry survival. *Rev. Fisheries Sci.* 17, 391–401. doi: 10.1080/10641260902752199

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.-H. J., Fisher, R. B., and Nadarajan, G. (2010). "Automatic fish classification for underwater species behavior understanding," in *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams* (Association for Computing Machinery, New York, NY, USA), 45–50.

Turek, K. C., Pegg, M. A., and Pope, K. L. (2014). Short-term evaluation of visible implant alpha tags in juveniles of three fish species under laboratory conditions. *J. Fish Biol.* 84, 971–981. doi: 10.1111/jfb.2014.84.issue-4

Wei, L., Zhang, S., Gao, W., and Tian, Q. (2018). "Person transfer gan to bridge domain gap for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 79–88 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/CVPR.2018.00016

Wilder, R. M., Hassrick, J. L., Grimaldo, L. F., Greenwood, M. F., Acuña, S., Burns, J. M., et al. (2016). Feasibility of passive integrated transponder and acoustic tagging for endangered adult delta smelt. *North Am. J. Fisheries Manage.* 36, 1167–1177. doi: 10.1080/02755947.2016.1198287

Wu, X., Huang, J., Wang, Y., and Wang, L. (2022). Pose estimation-based experimental system for analyzing fish swimming*. *bioRxiv* 2022–09, 2022–2009. doi: 10.1101/2022.09.07.507033

Xu, Y., Ma, B., Huang, R., and Lin, L. (2014). "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proceedings of the 22nd ACM international conference on Multimedia*. 937–940 (New York, NY, USA: Association for Computing Machinery).

Yadav, A., and Vishwakarma, D. K. (2020). Person re-identification using deep learning networks: A systematic review. *arXiv preprint arXiv:2012.13318*. doi: 10.48550/arXiv.2012.13318

Yang, Z., Bastan, M., Zhu, X., Gray, D., and Samaras, D. (2022). "Hierarchical proxy-based loss for deep metric learning," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 449–458 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/WACV51458.2022.00052

Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. (2019). "Reppoints: Point set representation for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9656–9665 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/ICCV.2019.00975

Zakria,, Deng, J., Hao, Y., Khokhar, M. S., Kumar, R., and Cai, J. (2021). Trends in vehicle re-identification past, present, and future: A comprehensive review. *Mathematics* 9, 3162. doi: 10.3390/math9243162

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on faster r-cnn and adversarial occlusion network. *Eng. Appl. Artif. Intell.* 100, 104190. doi: 10.1016/j.engappai.2021.104190

Zhai, A., and Wu, H.-Y. (2019). "Classification is a strong baseline for deep metric learning," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, Vol. 91 (Cardiff,UK: BMVA Press).

Zhao, S., Zhang, S., Liu, J., Wang, H., Zhu, J., Li, D., et al. (2021). Application of machine learning in intelligent fish aquaculture: A review. *Aquaculture* 540, 736724. doi: 10.1016/j.aquaculture.2021.736724

Zheng, Z., Ren, H., Wu, Y., Zhang, W., Lu, H., Yang, Y., et al. (2023). Fully unsupervised domain-agnostic image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* 34, 5077–5090. doi: 10.1109/TCSVT.2023.3335147

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016). Person re-identification: Past, present and future. *CoRR abs/1610.02984*. doi: 10.48550/arXiv.1610.02984s

Zong, Z., Song, G., and Liu, Y. (2023). "Detrs with collaborative hybrid assignments training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 6725–6735 (Los Alamitos, CA, USA: IEEE Computer Society). doi: 10.1109/ICCV51070.2023.00621