frontiers

Take Good Care of Your Fish: Fish Re-identification with Synchronized Multi-view Camera System

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

[1]SANYA Oceanographic Institution, Ocean University of China, China, [2]Ocean University of China, China

Submitted to Journal:
Frontiers in Marine Science

Specialty Section:
Marine Fisheries, Aquaculture and Living Resources

Article type:
Original Research Article

Manuscript ID:
1429459

Received on:
08 May 2024

Journal website link:
www.frontiersin.org

Scope Statement

This manuscript primarily centers on the research of fish re-identification (re-ID). This is of paramount importance for monitoring fish and significantly contributes to the advancement of aquaculture and fish breeding. The manuscript elaborately details our initial endeavors in the field of fish re-identification. This encompasses the establishment of the first underwater fish re-identification benchmark dataset (FS48) and developing a robust and accurate re-identification baseline (FSNet) that integrates information from three camera positions. The research content aligns closely with the scope of the selected journal and is significant. It holds great significance in facilitating this domain's in-depth development and technological progress. By focusing on the application, this research not only provides valuable insights but also paves the way for more effective and efficient fish monitoring, which in turn enhances the management and productivity of aquaculture and the breeding of various fish species, ultimately benefiting the entire industry.

Conflict of interest statement

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

CRediT Author Statement

Suzhen Fan: Data curation, Software, Writing – original draft, Writing – review & editing. Haiyang Feng: Writing – review & editing. Chengyang Song: Writing – review & editing. Zhibin Yu: Funding acquisition, Resources, Writing – review & editing.

Keywords

Fish Re-identification, Multiple cameras, Within-view, Cross-view, Synchronized Multi-view

Abstract

Word count: 177

Fish re-identification(re-ID) is crucial for fish monitoring and can further promote aquaculture and fish breeding. Consequently, we have taken the first step in fish re-identification efforts.Synchronizing information from different cameras can accelerate or optimize re-identification performance. We constructed the first underwater fish re-identification benchmark dataset (FS48) under three camera conditions to promote the development of underwater re-identification.FS48 includes 48 different fish identities, 10,300 frames, and 39,088 bounding boxes, covering different lighting conditions during day and night and occluded and unoccluded background environments. We developed the first robust and accurate fish re-identification baseline, FSNet, which fuses information from three camera positions. FSNet extracts features from synchronized video frames from each camera position and fuses the synchronized information from the three positions. By combining information from three positions, FSNet achieves better re-identification performance. Our fish re-identification baseline helps improve overall re-test accuracy and evaluate the effectiveness of re-identification among detectors. Experimental results demonstrate that FS48 is universal and high-quality, and FSNet has an effective network design and good performance. Our dataset will be released upon acceptance of this paper.

Funding information

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 Grant No. SKJC-2023-01-004 and the finance science and technology Q19 project of 630 Hainan province
of China under Grant No. ZDKJ202017.

Funding statement

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article.

Ethics statements

Studies involving animal subjects

Generated Statement: Ethical approval was not required for the study involving animals in accordance with the local legislation and institutional requirements because 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. .

Studies involving human subjects

Generated Statement: No human studies are presented in the manuscript.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable images or data are presented in this study.

Data availability statement

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

frontiers

Take Good Care of Your Fish: Fish Re-identification with Synchronized Multi-view Camera System

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

Correspondence*:
Zhibin Yu
yuzhibin@ouc.edu.cn

ABSTRACT

Fish re-identification(re-ID) is crucial for fish monitoring and can further promote aquaculture and fish breeding. Consequently, we have taken the first step in fish re-identification efforts. Synchronizing information from different cameras can accelerate or optimize re-identification performance. We constructed the first underwater fish re-identification benchmark dataset (FS48) under three camera conditions to promote the development of underwater re-identification. FS48 includes 48 different fish identities, 10,300 frames, and 39,088 bounding boxes, covering different lighting conditions during day and night and occluded and unoccluded background environments. We developed the first robust and accurate fish re-identification baseline, FSNet, which fuses information from three camera positions. FSNet extracts features from synchronized video frames from each camera position and fuses the synchronized information from the three positions. By combining information from three positions, FSNet achieves better re-identification performance. Our fish re-identification baseline helps improve overall re-test accuracy and evaluate the effectiveness of re-identification among detectors. Experimental results demonstrate that FS48 is universal and high-quality, and FSNet has an effective network design and good performance. Our dataset will be released upon acceptance of this paper.

Keywords: Fish Re-identification, Multiple Cameras, Within-view, Cross-view, Synchronized Multi-view

1 INTRODUCTION

Fish, as one of the essential product of underwater agriculture, contributes significant and sustainable nutrients for human. Fish identification technology is fundamental to boost biological, ecological, and aquaculture studies by tracing the organism's fate 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 a relatively reliable measurement, the tagging process may inevitably bring detriment Runde et al. (2022). Besides, 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 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 sub-tasks 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). Due 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. From the experimental results, it is clear that our fish re-identification technology can be used underwater and performs exceptionally well. Existing fish tracking and matching technologies are typically conducted using 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 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 FSNet (see Figure 3).

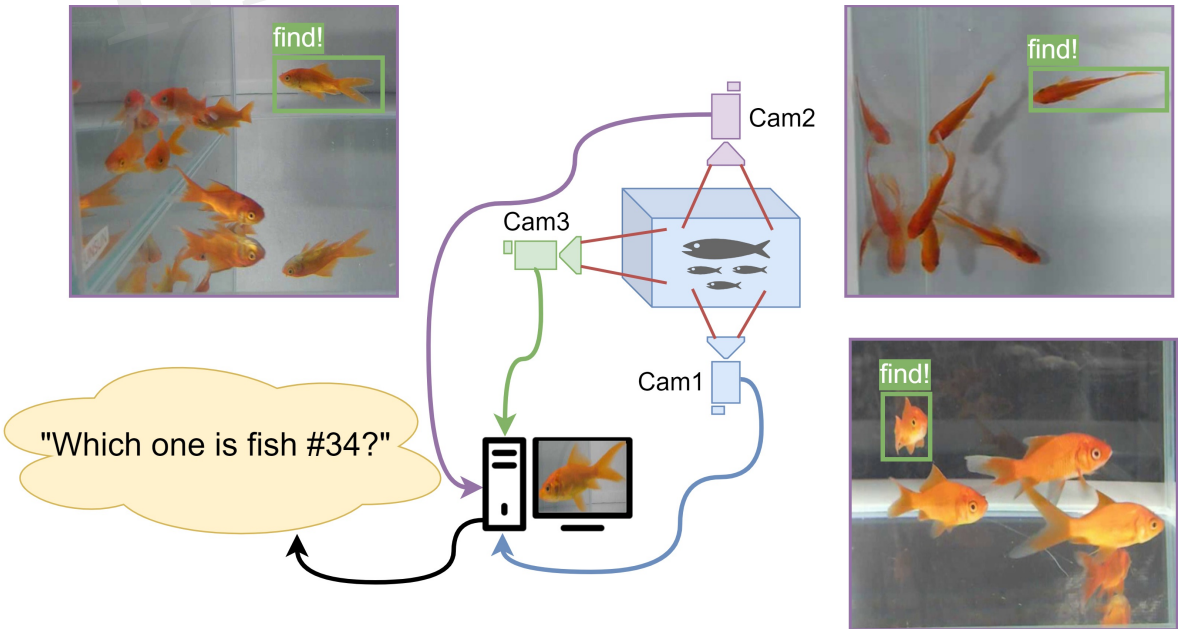

Figure 1. An overview of our three-view video capture system and the re-identification process. The characteristic information of the same fish from different visual directions is obtained from three positions to find the target fish (fish #34) after feature fusion.

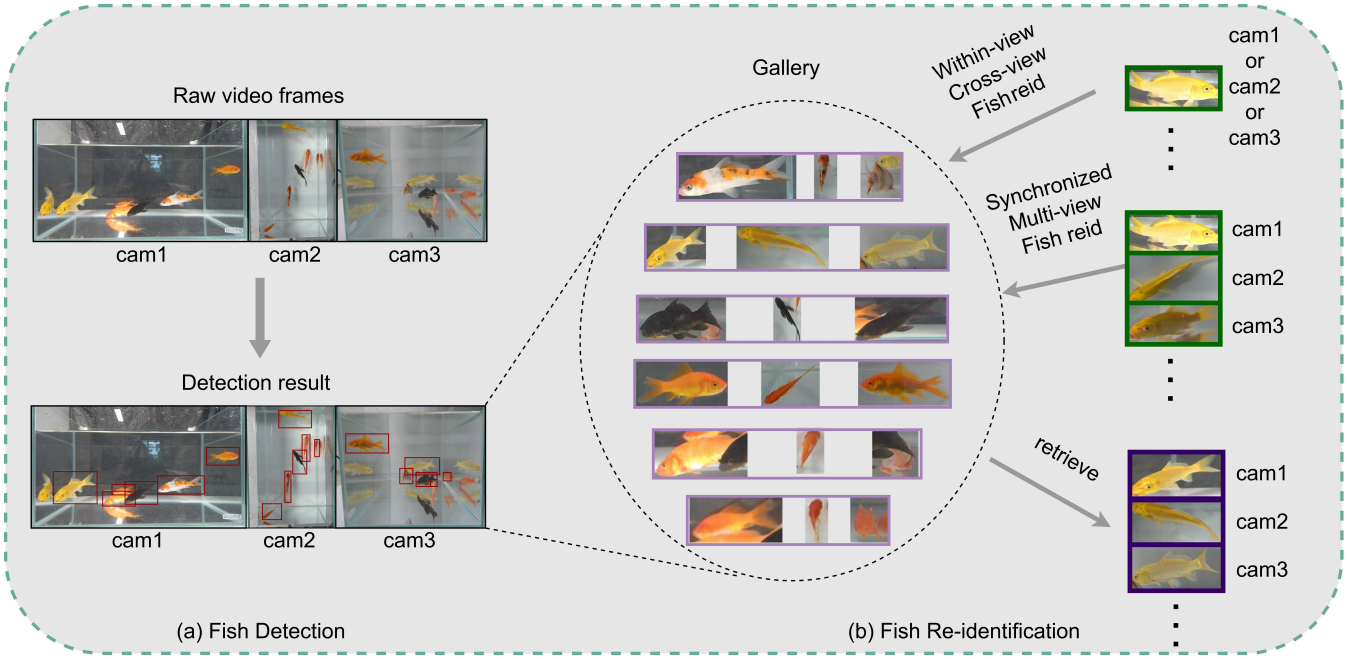

Figure 2. Pipeline of an end-to-end fish re-ID system. It consists of two modules: detection and fish recognition.

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 dealing with occluded or blurry fish caused by reflections. FSNet adopts a traditional approach where 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 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, consisting of 109 video sequences, 10,300 frames, and 39,088 bounding boxes.
- We have 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 build the benchmark.

2 RELETED 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 focuses 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 of 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 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. It helps to understand fish's distribution, abundance, and ecosystem health in waters 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 paper, 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. (2023b,a). 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 based on the information observed in other camera views. In 2006, Gheissari et al. Gheissari et al. (2006) used the spatiotemporal segmentation algorithm to segment the image and then used human visual cues for foreground detection. This work begins with person re-identification and multi-target separation as an independent computer vision task. 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. Due to the significant changes in appearance and posture, person recognition is complex to match across cameras, so it is used in security monitoring, personnel tracking, and other fields. Another critical area of re-identification is vehicle re-identification, the sensor-based approach phase between 1990 and 1999 Kell et al. (1990). Historical stage of vehicle re-identification based on depth feature 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 shows 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 the survival rate of fish 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, gender, 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 the potential sub-lethal effects on fish behavior Murphy et al. (1996); Skalski et al. (2009). Recent studies on fish marking have primarily focused on Visible Implant Alphanumeric (VIA) tags, adipose fin clips (suitable for sub-adult 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 reveals significant differences in tag retention and survival rates. Methods like adipose fin clips, particularly affecting juvenile fish, have been shown to decrease post-marking survival rates. Even the currently most widely adopted VIA tags still involve harm to fish individuals 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 fish individuals. Our method, which relies on synchronized cameras, can effectively achieve individual identification without endangering fish individuals. The fish re-identification technology we proposed 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

Due 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 proposed multi-view fish re-identification that provides an innovative solution to solve these effects. 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:

$$i^* = \underset{i \in 1,2,...,N}{\arg\max}\, sim(q, x_i) \tag{1}$$

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

For a multi-view synchronized camera system, we can have P images for each identity as $\mathbf{x}_i = \{x_i^p\}_{p=1}^{P}$, 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 overlapped 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 on $\mathcal{X}_{test}$.

3.2 Preliminary

We consider fish re-ID as an image retrieval problem that aims to recognize and match the identity 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, VGG16 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 VGG16 and ResNet-50 backbone networks on the FS48 dataset along with several loss functions of(SoftTripleLoss Qian et al. (2019), NormSoftMax Zhai and Wu (2019), ProxyAnchorLoss Kim et al. (2020), ArcFaceLoss Deng et al. (2022), ProxyNCALoss Yang et al.

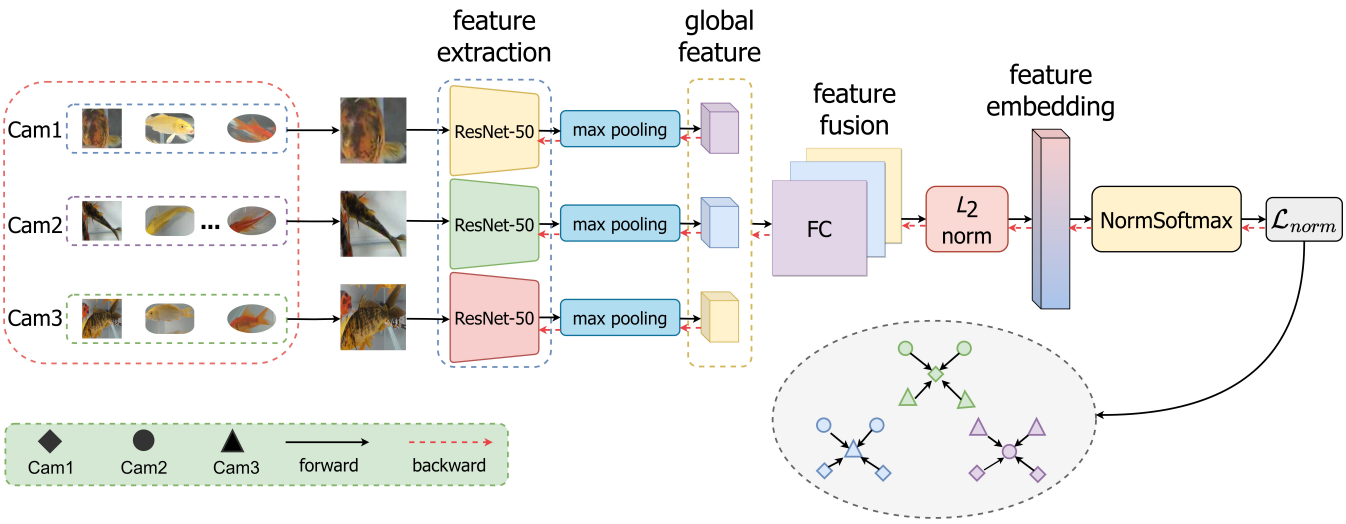

Figure 3. Pipeline of Mixed Three Views(FSNet). Directly splice the images together and perform feature extraction using Backbone. Then, perform overall feature fusion.

(2022)). The fish re-ID baseline that is better on 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 on 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 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 3, 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 evaluate multiple losses for normalization and choose the SoftTriplet loss Qian et al. (2019) and NormSoftMax loss Zhai and Wu (2019) as our objective functions to achieve a balance between maximizing the inter-identity distance and minimizing the intra-identity 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:

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

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

Table 1. FS48 Dataset usage under three re-ID experimental setups, C1 is the front view, C2 is the top view, 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

4 FS48 DATASET

4.1 Camera Setup

In this study, we chose two common freshwater fish, crucian carp and carp, to build the FS48 dataset. To increase the variety of the individuals, we chose different sizes of fish ranging from 6 to 15 cm. We used a total of 48 fish carefully reared under laboratory conditions Zhou and Xiang (2019). We set up partitions in the rearing tank to facilitate individual identification.

During the rearing process, we adopted ten days from October 15, 2023, to October 25, 2023. During the period, we took each individual out of the rearing pool and placed them in a small transparent fish tank for photographing every day. We used three cameras to build a multi-view video recording system to obtain comprehensive data and ensure cameras were synchronized in time. Realize synchronous shooting of the camera by programming. This setup allowed us to photograph the fish from three directions Yadav and Vishwakarma (2020): front, side, and bird, providing a more three-dimensional and comprehensive data set for subsequent experiments. 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 of 10,300 sheets were used during the detection experiments.

4.2 Labeling

In this work on fish re-ID, we present a fish re-ID dataset named FS48. The data in the experiment are cropped based on manually labeled bounding boxes Baltieri et al. (2011), and some of the cropped data are randomly excluded from the data in this experiment. See Figure 2 for a detailed structure of fish re-ID. Combined with the detection experiments, train and gallery still use manually labeled bounding boxes Li et al. (2012), the query uses bounding boxes generated by the Co-DERT detector, and the re-ID usage data under each experimental setup is shown in Table 1.

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
OzFishAl Muksit et al. (2022)	1,800	Detection	about 43k	-	1
Fish4-Knowledge (F4K)Qin et al. (2016)	27,320	Clf	-	-	1
Fishnet Open ImagesKay and Merrifield (2021)	86,029	Detection, Clf	406,463	-	1
DeepFishSaleh 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.

4.3 Comparison with Existing Fish Datasets

We have 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 to existing public datasets, with significant differences in the tasks it aims to fulfill. The unique feature within our proposed FS48 dataset is of particular significance, where each fish is endowed with a distinctive identifier, ensuring precise and accurate identification. This aspect holds paramount practical significance for future marine aquaculture and economic activities.

5 EXPERIMENTS

5.1 Implementation Details

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. Subsequently, the extracted features are 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 strictly 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.

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:

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

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

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

This paper divides all fishes into 48 categories, so $N = 48$. This paper divides all fishes into 48 categories, where n is the total number of categories. In this paper, all fishes are 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 this paper.

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 query and gallery images are from different known cameras. It allows for conducting three distinct experiments.

- Cross-view fish re-ID:

We use the images captured from the three camera views for training and testing in this experimental setting. We perform 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 is optimized to perform the cross-view fish re-ID, learning the feature representations robust to the view changes. It is worth noting that the input of the model is a single image.

- Synchronized multi-view fish re-ID:

Under this setting, we aim to demonstrate that combining the three camera views could lead to more robust and accurate fish re-identification. In parallel, we concatenate the images from the three camera views (front, side and top view). To ensure consistent feature representations between the query and gallery, we set the input of both the query and gallery as the mixture of the three images from the corresponding camera views. It is worth noting that the concatenated three images are strictly time-synchronized. The pipeline under this setting is shown in Figure 3.

5.2 Fish Detection

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), YoloX Ge et al. (2021) in MMDetection, and the newest ones GFL Li et al. (2020), and Co-DERT Zong et al. (2023) as advanced object detection frameworks. These several detectors are fine-tuned using 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. From Table 3, we can see that in the detection experiments where all fish are treated as a single class, Faster-RCNN, Reppoints, Foveabox, YoloX, and the newest GFL and Co-DERT detectors achieve an average accuracy of about 70% at IoU 95, with Co-DERT performing the best with an AP of 0.838 at AP_{95}, 0.975 at AP_{50}. The visualization of Co-DERT is shown in Figure 4. It is noteworthy that multi-view fish re-identification makes the re-identification of high-density fish more accurate because the occlusion phenomenon will become more severe in the case of high fish density.

Table 3. Fish-only detection results. We regard all the fish as one class and use several representative detection backbones to conduct detection experiments on our data sets.

Methods	AP_{95}	AP_{50}
Faster-RCNNRen et al. (2015)	0.770	0.962
ReppointsYang et al. (2019)	0.582	0.916
FoveaboxKong et al. (2020)	0.740	0.963
YoloXGe et al. (2021)	0.618	0.917
GFLLi et al. (2020)	0.675	0.918
Co-DETRZong et al. (2023)	**0.838**	**0.975**

Table 4. Fish classification results with ID detection (48 fish). We further consider the identity information of the fish by treating each individual as a different category.

Methods	AP_{95}	AP_{50}
Faster-RCNNRen et al. (2015)	**0.082**	**0.208**
ReppointsYang et al. (2019)	0.045	0.098
FoveaboxKong et al. (2020)	0.042	0.109
YoloXGe et al. (2021)	0.042	0.160
GFLLi et al. (2020)	0.052	0.091
Co-DETRZong et al. (2023)	0.003	0.010

Fish classification results with ID detection (48 fish): In this section, the basic setup is the same as 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 AP_{50} reaching merely 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 the overall performance drop. Furthermore, we demonstrated the intrinsic limitation of the above detection-based algorithms,

which cannot detect unseen fish while assigning the correct IDs. This fact reveals 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, where an individual fish's identity is considered essential for re-identification experiments. Under this research direction, we are committed to the re-ID model to effectively capture fish identity information 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 the detection model in practical applications and provide more powerful technical support for the real-time identification of multiple unknown fish species.

Figure 4. Visualization of the Co-DERT under the "fish only" setting.

5.3 Fish Re-identification

In the experiments of re-ID, the bounding boxes were firstly calibrated manually by humans to do 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 whole model for 40 epochs using the Adam optimizer. We set the initial learning rate to 1e-4 and measured the fish re-ID performance using the checking accuracy and Mean Average Precision (mAP).

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 VGG16 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.

Table 5. Within-view fish re-ID. 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.

Methods	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
SoftTripletQian 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
ArcFaceLossDeng 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
ProxyNCAYang 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
ProxyAnchorKim 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
NormSoftMaxZhai 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**
SoftTripletQian 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**
ArcFaceLossDeng 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
ProxyNCAYang 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
ProxyAnchorKim 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
NormSoftMaxZhai 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

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

Methods	Backbone	P@5	mAP@10	P@10	mAP@alll
SoftTripletQian et al. (2019)	VGG16	54.75	44.78	52.10	**30.30**
ProxyNCAYang et al. (2022)		47.67	37.89	45.46	27.71
ProxyAnchorKim et al. (2020)		53.58	43.73	51.19	26.48
NormSoftMaxZhai and Wu (2019)		**55.28**	**45.33**	**52.72**	29.82
ArcFaceLossDeng et al. (2022)		46.94	34.95	43.44	18.00
SoftTripletQian et al. (2019)	ResNet-50	**66.72**	**57.40**	**64.04**	**32.78**
ProxyncaYang et al. (2022)		56.75	47.34	54.46	31.10
ProxyAnchorKim et al. (2020)		64.28	55.15	61.85	29.66
NormSoftMaxZhai and Wu (2019)		63.72	53.52	60.74	31.36
ArcFaceLossDeng et al. (2022)		60.11	50.48	57.36	30.49

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 Table 5 and Table 6, we can conclude the performance of SoftTriplet loss combined with ResNet-50 achieved the best performance in two different experimental settings.

Table 7. Synchronized multi-view fish re-ID. 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.

Methods	Backbone	P@5	mAP@10	P@10	mAP@alll
SoftTripletQian et al. (2019)	VGG16	**91.27**	**81.40**	**83.58**	49.22
ProxyNCAYang et al. (2022)		86.50	73.04	75.56	39.49
ProxyAnchorKim et al. (2020)		87.84	77.61	79.82	45.94
NormSoftMaxZhai and Wu (2019)		90.62	80.93	82.88	**50.61**
ArcFaceLossDeng et al. (2022)		85.13	73.81	76.83	42.66
SoftTripletQian et al. (2019)	ResNet-50	**79.18**	**74.03**	**77.43**	52.07
ProxyncaYang et al. (2022)		73.53	67.21	71.63	47.06
ProxyAnchorKim et al. (2020)		77.32	72.52	76.00	49.50
NormSoftMaxZhai and Wu (2019)		77.22	72.94	76.52	**52.17**
ArcFaceLossDeng et al. (2022)		77.68	71.01	74.75	50.17

We observe that multiple fish may easily trigger occlusions under the within-view (Figure 4). 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 to mix the information of the images captured from the three camera views. We believe different views can provide more complementary information be 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 Table 6 and Table 7, we can find that the mAP@all scores of the synchronized multi-view experimental setting are much higher than that 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 to support fish re-identification tasks. Our FSNet can effectively utilize the multi-view information for fish identification.

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-DERT), 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 keep 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. Please note that 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. We evaluate and locate the best loss functions and backbones for different settings in Table 8. Since we did not consider the multi-fish scene for the automatically re-ID task, the scores in Table 8 are even higher than in Tables 6 and 7.

Table 8. Performance of the bounding boxes generated by the Co-DERT detection model as a query.

Methods	Backbone	Cross-view fish re-ID			
		P@5	mAP@10	P@10	mAP@all
SoftTripletQian et al. (2019)	ResNet-50	77.82	67.57	71.74	34.77
Methods	Backbone	Synchronized multi-view fish re-ID			
		P@5	mAP@10	P@10	mAP@all
NormSoftMaxZhai and Wu (2019)	ResNet-50	77.52	72.46	76.18	52.27

The heatmap visualizations in Figure 5 demonstrate that the front view contains the key information, including areas surrounding fish eyes for identification. In contrast, the semantic information in the top and side views is mainly related to contours (such as 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 the front view for the information deficiency. Our FSNet network extracts features from the three views and builds feature embedding to complete the fish re-identification process. Please note the first two yellow fish in Figure 5 are very similar in the raw images. Looking closely, we can see that their fins and eyes are different (shape and detail). Their fins and eyes are also highlighted on the thermal map, which indicates that our proposed FSNet can efficiently identify the key features for fish re-ID tasks.

Table 9. Comparison of baseline and our elevated pipeline.

Synchronized multi-view fish re-ID	P@5	mAP@10	P@10	mAP@all
baseline	77.22	72.94	76.52	52.17
FSNet	79.24	71.98	80.15	53.33

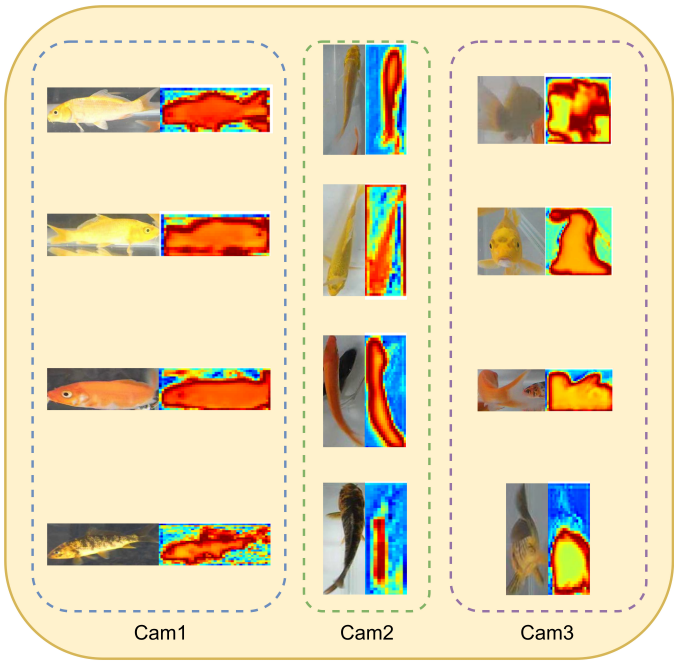

Figure 5. Heatmap visualization of images from three views, with the image in the left column of the same shot being the original image and the heatmap image in the right column.

The 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 9. The baseline used a shared backbone to extract information from three perspectives and perform feature fusion. As shown in Table 9, 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.

6 CONCLUSIONS

To the best of our knowledge, this work takes the first step towards deep learning based fish re-identification, and the reliability of this work is consolidated, mainly due to the large number of experiments we have 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 propel research in fish monitoring and aquaculture.

FUNDING

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 Grant No. SKJC-2023-01-004 and the finance science and technology Q19 project of 630 Hainan province of China under Grant No. ZDKJ202017.

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)* (Los Alamitos, CA, USA: IEEE Computer Society), 3908–3916. 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. *Ecological Informatics* 72, 101847

Alsmadi, M. K. and Almarashdeh, I. (2022). A survey on fish classification techniques. *Journal of King Saud University-Computer and Information Sciences* 34, 1625–1638

Alsmadi, M. K., Tayfour, M., Alkhasawneh, R. A., Badawi, U., Almarashdeh, I., and Haddad, F. (2019). Robust feature extraction methods for general fish classification. *International Journal of Electrical & Computer Engineering (2088-8708)* 9, 5192–5204

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* (New York, NY, USA: Association for Computing Machinery), J-HGBU '11, 59–64. doi:10.1145/2072572.2072590

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 423–443

Buckley, R. M., West, J. E., and Doty, D. C. (1994). Internal micro-tag systems for marking juvenile reef fishes. *Bulletin of marine science* 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 American Journal of Fisheries Management* 34, 30–38

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 (Cham: Springer International Publishing), 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)* (Los Alamitos, CA, USA: IEEE Computer Society), 24–29. 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 Transactions on Systems, Man, and Cybernetics: Systems* 47, 2467–2477

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. *Animal Biotelemetry* 1, 1–19

Dare, M. R. (2003). Mortality and long-term retention of passive integrated transponder tags by spring chinook salmon. *North American Journal of Fisheries Management* 23, 1015–1019

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., and Zafeiriou, S. (2022). Arcface: Additive angular margin loss for deep face recognition. *IEEE transactions on pattern analysis and machine intelligence* 44, 5962—5979. doi:10.1109/tpami.2021.3087709

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

Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., and Hartley, R. (2006). Person reidentification using spatiotemporal appearance. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2, 1528–1535

Huang, T. and Russell, S. (1997). Object identification in a bayesian context (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.), IJCAI'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 American journal of fisheries management* 34, 322–334

Kaiming, H., Xiangyu, Z., Shaoqing, R., and Jian, S. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and 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. *ArXiv* abs/2106.09178

Kell, J. H., Fullerton, I. J., and Mills, M. K. (1990). *Traffic detector handbook*. Tech. rep.

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)* (Los Alamitos, CA, USA: IEEE Computer Society), 3235–3244. 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 Transactions on Image Processing* 29, 7389–7398

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* (Los Alamitos, CA, USA: IEEE Computer Society), 5–12

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* (Berlin, Heidelberg: Springer-Verlag), ACCV'12, 31–44. doi:10.1007/978-3-642-37331-2_3

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. *Advances in Neural Information Processing Systems* 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 American Journal of Aquaculture* 75, 186–196

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

Liu, Dongcai, Wen, Xianhui, and Zhou, Youling (2022). Research on an improved fish recognition algorithm based on yolox. *ITM Web Conf.* 47, 02003. doi:10.1051/itmconf/20224702003

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. *Reviews in Aquaculture* 13, 1565–1582

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. *Computers and Electronics in Agriculture* 201, 107335

Murphy, B. R., Willis, D. W., and Society, A. F. (1996). *Fisheries techniques* (American Fisheries Society), 2nd ed edn.

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. *Journal of Fish and Wildlife management* 8, 353–364

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 Review* 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)* (Los Alamitos, CA, USA: IEEE Computer Society), 6449–6457. 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

Ren, H., Zheng, Z., Wu, Y., and Lu, H. (2023a). Daco: domain-agnostic contrastive learning for visual place recognition. *Applied Intelligence* , 1–14

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 Transactions on Circuits and Systems for Video Technology*

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28, 1137–1149

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. *Journal of Fish Biology* 101, 419–430

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. *Scientific Reports* 10, 14671

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

Sandford, M., Castillo, G., and Hung, T.-C. (2020). A review of fish identification methods applied on small fish. *Reviews in Aquaculture* 12, 542–554

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. *The Open Fish Science Journal* 6, 41–47

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR* abs/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. *Reviews in Fisheries Science* 17, 391–401

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* (New York, NY, USA: Association for Computing Machinery), 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. *Journal of Fish Biology* 84, 971–981

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)* (Los Alamitos, CA, USA: IEEE Computer Society), 79–88. 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 American Journal of Fisheries Management* 36, 1167–1177

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* (New York, NY, USA: Association for Computing Machinery), 937–940

Yadav, A. and Vishwakarma, D. K. (2020). Person re-identification using deep learning networks: A systematic review. *arXiv preprint 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)* (Los Alamitos, CA, USA: IEEE Computer Society), 449–458. 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)* (Los Alamitos, CA, USA: IEEE Computer Society), 9656–9665. doi:10.1109/ICCV.2019.00975

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

Zeng, L., Sun, B., and Zhu, D. (2021). Underwater target detection based on faster r-cnn and adversarial occlusion network. *Engineering Applications of Artificial Intelligence* 100, 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* (Cardiff,UK: BMVA Press), 91

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

Zheng, Z., Ren, H., Wu, Y., Zhang, W., Lu, H., Yang, Y., et al. (2023). Fully unsupervised domain-agnostic image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*

Zhou, K. and Xiang, T. (2019). Torchreid: A library for deep learning person re-identification in pytorch. *ArXiv* abs/1910.10093

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