Review Article

Pengfei Xu, Xianyi Liu, Jinping Liu, Meiling Cai, Ying Zhou*, Shanshan Hu*, and Minlian Chen

Survey on machine vision-based intelligent water quality monitoring techniques in water treatment plant: Fish activity behavior recognition-based schemes and applications

https://doi.org/10.1515/dema-2024-0010
received August 16, 2023; accepted April 7, 2024

Abstract: Water is a vital resource essential to the survival and development of all creatures. With the rapid growth of industry and agriculture, people face a severe threat of ecological destruction and environmental pollution while living earthly lives. Water pollution, in particular, harms people's health the most. As a result, water supply security has become a top priority. As a critical point in water supply safety, monitoring water quality effectively and forecasting sudden water contamination on time has become a research hotspot worldwide. With the rapid development and wide applications of artificial intelligence and computer vision technologies, biological activity identification-based intelligent water quality monitoring methods have drawn widespread attention. They were taking fish activities as the water-quality indicator has gained extensive attention by introducing advanced computer vision and artificial intelligence technologies with low cost and ease of carrying. This article comprehensively reviews recent progress in the research and applications of machine vision-based intelligent water quality monitoring and early warning techniques based on fish activity behavior recognition. In detail, it addresses water quality-oriented fish detection and tracking, activity recognition, and abnormal behavior recognition-based intelligent water quality monitoring. It analyzes and compares the performance and their favorite application conditions. Finally, it summarizes and discusses the difficulties and hotspots of water quality monitoring based on the fish's abnormal behavior recognition and their future development trends.

Keywords: machine vision-based water quality monitoring, water quality safety pre-warning, abnormal behavior identification, water pollution source tracing, fish object detection, fish object tracking

MSC 2020: 68T01

*** Corresponding author: Ying Zhou,** Data and Information Management Center, Hunan Children's Hospital, Changsha, Hunan, 410021, China, e-mail: yingzhou_hch@163.com
*** Corresponding author: Shanshan Hu,** Data and Information Management Center, Hunan Children's Hospital, Changsha, Hunan, 410021, China, e-mail: 27963038@qq.com
Pengfei Xu: College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, 410021, China, e-mail: xupf@hunnu.edu.cn
Xianyi Liu: College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, 410021, China, e-mail: liuxianyi@hunnu.edu.cn
Jinping Liu: College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, 410021, China, e-mail: ljp202518@163.com
Meiling Cai: College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, 410021, China, e-mail: caimeiling_hunnu@163.com
Minlian Chen: Data and Information Management Center, Hunan Children's Hospital, Changsha, Hunan, 410021, China, e-mail: hnch_gjhz@163.com

1 Introduction

As we all know, water plays an essential role in our lives and is an indispensable resource for the survival and development of human beings and all living things. The safety of water quality is directly related to people's lives and health, as well as the quality of daily life. However, with the rapid development of industrialization and urbanization, environmental pollution is worsening. Water pollution is the first to bear the brunt. Rivers, lakes and seas, groundwater, and so on often become the ultimate recipients of pollutants. The situation of water pollution has been quite severe [1].

The Bulletin of China's Marine Environmental Status in 2021 shows that the quality of the ecological environment of the near-shore local waters needs to be improved. On the one hand, there is an excessive discharge phenomenon. For instance, the total amount of sewage discharged from 458 direct sea pollution sources with a daily discharge capacity of greater than or equal to 100 tons approximates 72,778,000 tons, where individual points of total phosphorus, ammonia nitrogen, suspended solids, chemical oxygen demand, five-day biochemical oxygen demand exceed the standard. On the other hand, the estuary bay water quality needs to be further improved, i.e., with a sea area of 21,350 km^2 in 2021, the main exceedances are the inorganic nitrogen and active phosphate.

Given the seriousness of water pollution, ensuring the safety of drinking water and domestic water supply has become the most critical task of the people. As a key link to a safe water supply, it is essential to monitor the reliability of water quality quickly and effectively. Scientific researchers and environmental regulatory departments worldwide have widely valued the early warning of water pollution. Up to now, researchers have done much research on water quality monitoring [2], early warning [3], and traceability of water pollution [4]. Some methods have entered the stage of practical applications [5].

Since the 1980s, some institutions have started academic research and engineering applications [6] and developed some relevant online water quality monitoring (OWQM) systems or sensors for the early warning of water supply monitoring, considering the importance of water quality monitoring, especially the drinking water quality. For example, in 1994, the Japan Water Works Associate (JWWA) reported their progress in instrument development and technical applications in the OWQM [6].

The water treatment plant is the main object of the early research and technical application of OWQM. The primary purpose of water quality monitoring in the waterworks is to detect the acidity and basicity of the effluent and the content of any harmful substance in the water in real-time through corresponding detection instruments (sensors) to ensure quality water production for residents. The waterworks need to control and optimize the drinking water processing process with the help of OWQM results to ensure that the produced water meets the quality requirements. The results of OWQM can be compared and verified by the effects of water quality tests in the laboratory to improve the performance of the OWQM [6].

Traditional water quality monitoring, especially drinking water quality monitoring, is mainly based on the analysis and determination of the types of pollutants in the water with their content and change trends. Conventional indicators related to water quality include: **(1) Sensorial indicators**, such as chromaticity, turbidity, and smell; **(2) Common chemical indicators**, e.g., pH value, residual chlorine content, electrical conductivity, metal elements in water such as aluminum, iron, manganese, copper, zinc, water hardness, sulfate content, oxygen consumption, volatile phenols, synthetic anion detergents, and so on; **(3) Microbiological indicators**, such as total coliform group, thermostable coliform group, *Escherichia coli*, total colony count, and so on; **(4) Toxicological indicators**, such as arsenic, cadmium, chromium, lead, mercury, selenium, cyanide, fluoride, nitrate, formaldehyde, and so on, **(5) Radioactivity indicators**, including α total radioactivity, total β radioactivity, radon, radium, and so on.

Due to the many monitoring indicators, water quality monitoring is tedious and time-consuming. Researchers have experimented with various methods and techniques based on physical, chemical, or biological correlation to achieve water quality monitoring, including chemical, electrochemical, atomic absorption spectrophotometry, ion-selective electrodes, and chemical methods [7]. These monitoring methods are essentially based on physical and chemical detection methods. In practical applications, water sampling, laboratory testing, and calibration should finally be performed to determine the water quality test results. The physical and chemical water quality detection methods based on laboratory measurement and calibration are

challenging to achieve long-term and uninterrupted online monitoring because of the sampling frequency, testing time, and sample testing cost limitation.

Many water quality indicator sensors have been developed with the rapid progress of physiochemical sensors and wireless sensor network technologies. These sensors are sensitive to some specific pollutants or substances in the water body so that the number of corresponding physical and chemical factors in the water to be measured can be detected. Therefore, if enough sensors related to water quality indicators can be provided, it is theoretically possible to achieve the OWQM [8] in an area or sewage treatment in a factory. Distributed OWQM is the basis for realizing a regional early warning system (EWS) of water quality. An EWS can effectively predict water quality disasters and sudden water pollution and inform users to take corresponding measures in advance to ensure people's water safety.

The spectral analysis (SA)-based method attracts excellent attention. Compared with traditional chemical analysis, e.g., the electrochemical and chromatographic analysis, SA-based methods have many advantages, such as relatively simple operation, less reagent consumption, good repeatability, high measurement accuracy, fast detection speed, and easy-to-implement. In the literature, the authors have reviewed the UV-Vis spectroscopy analysis-based water quality monitoring in detail [9]. The author summarized that although SA-based water quality detection methods had many advantages in theory for the OWQM, they still required some specific pretreatments of water samples, which may, in turn, cause secondary pollution. In other words, for SA, it is necessary to study further the corresponding chemometrics signal processing algorithm combined with the absorption spectral characteristics of the substance to be measured to promote a broader application of this method.

Generally, water quality sensors can detect only one or a few indicators. In addition, many sensors for different water quality indicators are needed to realize omnidirectional water quality monitoring, which virtually increases the installation and maintenance complexity and influences the monitoring effectiveness, timeliness, and cost. For example, gas chromatography and high-performance liquid chromatography can accurately detect chemical contamination in water, but the cost of the sampling and test is too expensive. Studies have shown that the average cost of using this method to prioritize monitoring 126 pollutants by the Environmental Protection Agency reaches US $1,000/sample [10]. Therefore, we still have a long way to go toward developing water quality monitoring instruments and equipment to realize the real-time and automated detection of harmful components in water.

With the continuous progress of image processing, artificial intelligence, and pattern recognition in recent years, water quality monitoring methods based on biological activity analysis by advanced machine learning and artificial intelligence techniques have attracted increasing attention [11]. These methods determine the water quality by monitoring the physiological reactions and behavior characteristics of organisms in the water using intelligent monitoring techniques, especially machine vision-based schemes. Researchers have proposed various biological-related water quality monitoring methods for long-term and online monitoring. According to the activity differences of organisms in water by biological behavior recognition (incredibly abnormal behavior recognition), researchers can effectively achieve a natural early warning (BEW) of pollutants or toxic substances in water.

Bioactivity-based technologies are widely used in water quality monitoring because they can well reflect biological toxicity in water with high sensitivity. Compared with traditional chemical detection methods, physical activity detection can comprehensively evaluate water bodies' ecological safety. Using biological activity detection technology to detect water quality can detect pollution sources more accurately and take measures to control it in time. Commonly used bioactivity-based water quality monitoring methods involve a variety of related organisms, including bacteria, algae, water fleas, shellfish, and so on. Bae and Park [10] reviewed in detail the applications of commonly used tagged organisms in water quality monitoring, including the applications in drinking water, reservoir water, factory water, and agricultural water, and the biological behaviors (measurement variables) and detection frequency that need to be paid attention to when using corresponding types of biological monitoring. Physical activity analysis and abnormal behavior recognition based on machine vision are usually utilized in biological activity behavior recognition.

Water quality monitoring by machine vision-based abnormal behavior recognition of fish has attracted wide attention due to its merits, such as simple implementation, low maintenance cost, fast detection speed,

and so on. Some related commercial systems are on the market, such as the Kerren Aqua-Tox-Control system, the bbe Fish Toximeter system, and the bbe ToxProtect system [12]. In particular, the identification of fish activity behavior (such as hazard avoidance behavior, gill respiration, tail swing frequency, active area distribution, swimming mode, speed, acceleration, predation, death, and so on) is adopted to determine the water quality to realize an early warning of water quality monitoring [13].

The species and quantity of fish selected vary depending on the water environment (such as reservoir water, small fish pond water, and factory-treated water). The chosen fish should be easy to obtain, raise, and sensitive to specific pollutants. In the 1980s, the International Standards Organization (ISO) recommended a group of fish species specifically for freshwater quality monitoring. The existing water quality monitoring system mainly selects small, and medium-sized freshwater fish, such as zebrafish (daniorerio), swordtail fish, peacock fish, medaka, and goldfish. According to the physiological and toxicological experimental results of different fishes, the fish species selection for various pollutions was reviewed by Casebolt et al. [14].

Small-scale water quality monitoring is generally carried out by monitoring the activity characteristics of one or a few fish so that one or several monitoring cameras can be set up to achieve the water quality monitoring result. Machine vision in industrial process monitoring has been widely valued by academic and industrial communities in recent years [15–17]. The commonly used visual features related to fish include the position and body orientation of fish in each frame, the swimming trajectory, speed, moving direction, movement regularity, track curvature, and so on. In large-scale water quality monitoring, such as reservoir water quality monitoring, it is generally necessary to build multiple monitoring points around the reservoir at several specific locations on the water surface of the reservoir through comprehensive tracking and analysis of the behavior of fish groups at various monitoring sites to comprehensively monitor the water quality.

According to the species and quantity of fish, the commonly used fish behavior parameters or behavior state variables for water quality monitoring are as follows.

(1) Fish mortality-based methods [18]

It is mainly used to monitor sudden heavy toxic pollution in water, such as water pollution caused by premature toxic substance leakage. When poisonous substances pollute the monitoring water body, many fish die quickly. Through computer vision monitoring, the statistics of the death of fish populations in the water can be used to evaluate the deterioration of water quality. Although this method can achieve the purpose of water quality monitoring, once a large number of fish die, it shows that the water body has been highly polluted. It is usually challenging to achieve an early warning of water pollution.

(2) Fish rheotaxis-based methods [19]

The so-called rheotaxis refers to a biological behavior response of fish objects to water flow, usually divided into positive and negative chemotaxis. From the natural habits of fish, it can be observed that fish in the typical water quality environment swim against the current. That is, they tend to swim in the direction of water flow. When the water body is polluted, the rheotaxis of the positive current ability of fish may be destroyed. This method can monitor the water quality of large reservoirs and rivers and the water quality of small fish ponds according to the acting ability of fish. At the same time, it can also make early warnings of water pollution, sudden pollution, or gradual deterioration of water quality.

(3) Gill movement-based methods [20]

The respiration of fish gills is related to pollutants in the water. Gomes et al. [20] analyzed the possibility of abnormal gills movement of two different fish, the *Geophagus brasiliensis* and *Astyanax bimaculatus*, in various reservoir water quality. When the pollutants in the water reach a particular concentration (such as eutrophication), it will accelerate the gill respiration frequency of fish, and the fish Gill respiration is more likely to lack regularity, showing the frequent abnormal movement of gills. According to the frequency of Gill's fish respiration, the purpose of water quality monitoring can be achieved. This method has more stringent requirements for fish location and detection, gill location, and tracking in the water. Thus, it is often only suitable for water quality monitoring in small ponds.

(4) Swimming avoidance behavior-based methods [21]

Regular swimming fish will show apparent escape behavior when they encounter water quality not conducive to their living environment (such as a sudden leakage of toxic substances somewhere in the reservoir). Therefore, if it is observed that fish always gather at a particular end within a period, it can judge

the fish escape behavior, which means that the fish-free aggregation end of the water body to be tested may be polluted. The corresponding countermeasures should be taken as soon as possible. This method has multiple advantages through machine vision monitoring, e.g., simple algorithms with rapid processing speed, as long as the aggregation degree of fish in each position in a period is analyzed. This method is usually more suitable for monitoring water quality in large-scale reservoirs, rivers, and lakes.

(5) Weak electrical filed pulse-based methods [22]

The fish, such as gymniformes, generally emit weak electrical field pulses when swimming in the water. The frequency and waveform of the weakly electric pulse emitted by the fish in standard water quality are constant. However, once the water environment is polluted, the frequency or waveform of the weak electric pulse will change. Therefore, the change in water quality can be detected through the movement behavior of fish and the change, like the weak electric pulse of fish objects. It is evident that this method not only monitors the activity behavior of fish through visual sensors but also needs to build a corresponding thyristor circuit to detect the weak circuit pulse attributes emitted by fish. Hence, compared with the previous methods, the installation of this monitoring system is a little more complicated.

The fundamental reason why fish activity behavior can be used for water quality monitoring is that when the water quality is abnormal, the biological fish living in the water will have corresponding stress changes in their activity behavior. When the water is highly polluted or when there is sudden chemical (toxic) pollution, many fish will die, which can be used as a direct detection form of abnormal water quality [23–26]. Of course, according to the different pollutants in the water, it is necessary to select specific fish species for water pollution monitoring to obtain significant detection results.

Since fish have many species that can survive in a water environment with a certain level of water quality, using fish as a biological "indicator" for water quality monitoring has a natural advantage. Fish activities will vary due to environmental changes, which provide many scientific bases and suitable experimental materials for water pollution monitoring based on fish activity behaviors. Moreover, the fish responds very acutely to changes in the chemical composition of the water environment. Thus, choosing the activity behavior of fish as the parameter for monitoring water pollution cannot only study the function of a single pollutant in water but also reflect the toxicity of the mixture of multicomponent pollutants. These parameters can be used as a comprehensive indicator to evaluate water pollution. Compared with the traditional physio-chemical sensor analysis and monitoring method, the water quality monitoring methods based on the activity behavior analysis of fish have the following advantages:

(1) It can effectively monitor the impact of water quality on biological activities, and it is one of the most direct ways of comprehensively evaluating water quality.

Traditional physical and chemical monitoring methods can detect chemical components in effluent, but sensors vary with components and reagents, posing a risk of secondary water pollution. The toxic effects of interactions among different chemical components require further analysis. Deployed sensors may miss detection of unknown toxic substances, posing potential dangers to safe water supply. Biological monitoring methods compensate for these shortcomings. By intelligent analysis and identification of fish activities using computer vision technologies, it can effectively evaluate the effects of various chemical components on organisms and achieve OWQM.

(2) It has the merit of fast detection, easy to install, and low maintenance cost.

Through advanced intelligent monitoring techniques, especially computer vision-based monitoring techniques, it can realize the all-weather uninterrupted monitoring of the water quality.

(3) It has a wide application range and can realize the early prediction of sudden pollution and severe toxic pollution in water.

It can be widely applicable by deploying many visual sensors that can realize the water quality monitoring of large reservoirs and rivers, small pools, water purification systems, chemical wastewater treatment systems, and other water treatment systems. In addition, it can realize the early prediction of sudden pollution and severe toxic pollution of water by intelligently identifying fish activity behaviors in the water.

Considering the significance of fish activity behavior recognition-based OWQM in scientific research and practical applications [27], we review the research and application progress of fish activity recognition methods for water quality monitoring and early warning in water plants in recent years. Some key

technologies involving fish object detection, localization, fish (group) tracking, fish activity behavior representation, and abnormal behavior recognition are reviewed comprehensively, and their appropriate environment has been deeply analyzed and compared. In addition, we discuss and summarize the challenging issues that need to be further studied and make an outlook on its future development trend.

2 Machine vision-based fish activity behavior recognition

2.1 Basic diagram of intelligent water quality monitoring

When the water quality is abnormal or the water body is polluted, the fish in the monitored water will react and undergo a relatively noticeable change in behavior, such as life behavior, physiological activities, and community behavior. Machine vision technologies, online monitoring, and even early warning of water quality and pollution can be achieved.

The water treatment process mainly includes taking water from water sources (such as reservoirs), salvaging floating materials (such as leaves), removing sediment by sedimentation, chlorination sterilization, adsorption filtration by pharmaceutical addiction and other water purification process, and finally, the output of clean tap water. To achieve comprehensive water quality monitoring, the water quality monitoring points will be established generally in three crucial sites, i.e., each water source intake, the output of the chlorination sterilization, and the output of the water purification process, which are also referred to as “ecological fish pond” as shown in Figure 1.

The key technologies are fish activity behavior recognition-based water quality monitoring, fish detection, fish (group) tracking, behavior feature extraction, abnormal behavior recognition, and early warning of abnormal behavior of fish activities in ecological fish ponds. The primary process is shown in Figure 2.

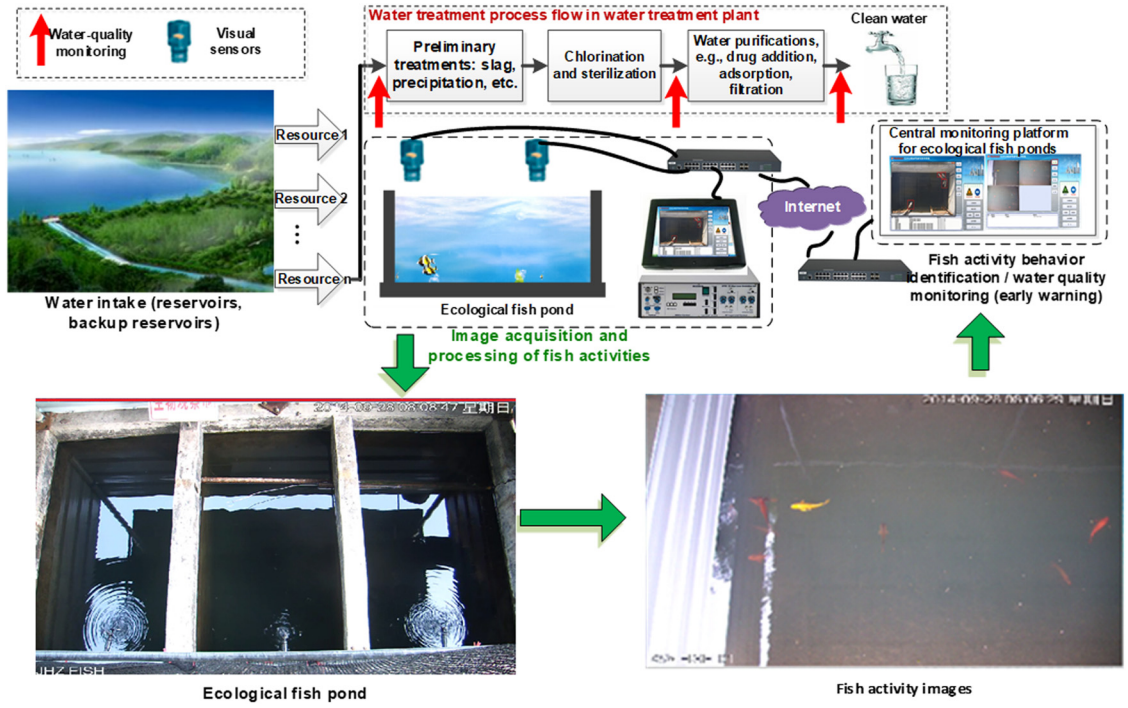

Figure 1: “Ecological fish pond” water quality monitoring diagram.

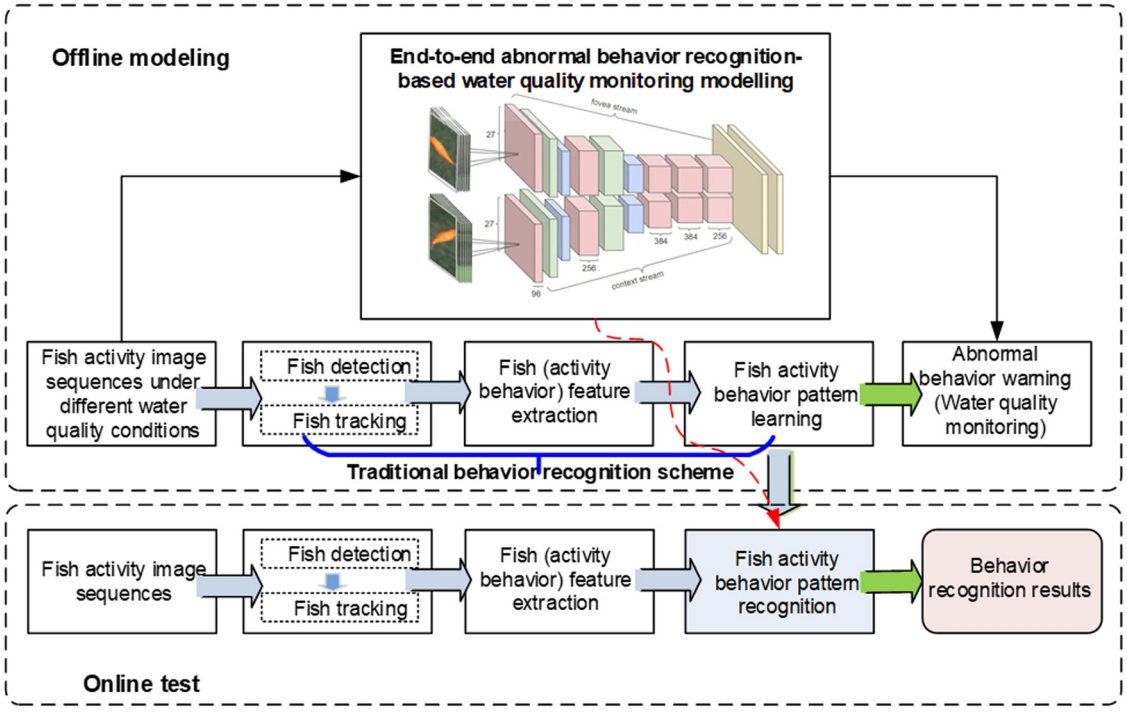

Figure 2: Fish activity behavior recognition process for water quality monitoring.

2.2 Machine vision-based fish object detection and tracking

2.2.1 Fish object representation

In machine vision-based dynamic object detection and tracking, a specific method should be first used to characterize the tracking objects. The commonly used object characterization methods can be divided into traditional and deep learning-based characterization methods.

Traditional object characterization methods mainly include shape-based methods and appearance-based representations [28]. Commonly used target representation methods include point-based, empirical geometry shape-based representation, and target contour or silhouette-based representation.

The point-based target representation method usually uses the tracking target's center (mass) point or the set of feature points on the tracking target to represent the target. This method is generally suitable for tracking small targets. The empirical geometry shape-based methods usually use a simple geometric frame (such as a rectangular or an elliptical frame) to represent the tracking target, which does not require complex image segmentation or target extraction and has good results for rigid body target tracking. It can also track non-rigid targets (whose deformation is not particularly significant) by considering various morphological changes described by mathematical models, such as geometry, affine, and perspective.

The object contour or silhouette-based methods usually choose the contour, silhouette, edge contour points, or skeleton to characterize the tracking objects. These methods have relatively high computational complexity because of the need to obtain the object's contour features. However, they have a better tracking effect on nonrigid bodies that undergo complex deformation. The schematic diagram of the object shape-based fish object characterization method is shown in Figure 3.

Another commonly used target representation method for tracking is the appearance-based representation method [29]. The appearance-based target characterization methods are probability density target appearance-based methods and template-based representation methods. Appearance-based characterization methods can use parametric (e.g., Gaussian model, hybrid Gaussian model) or nonparametric methods (Parzen window, histogram) [30] to characterize the appearance probability density of a target. The probability density of the

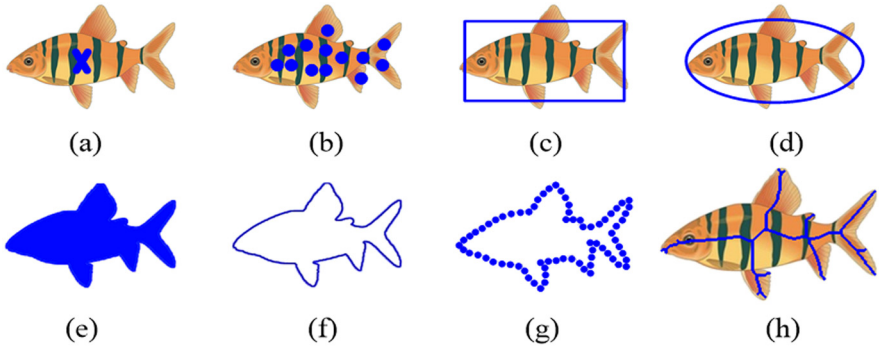

Figure 3: Schematic diagram of different representations of fish objects. (a) center point representation; (b) multipoint representation; (c) rectangular box representation; (d) elliptical representation; (e) revealed representation; (f) contour representation; (g) contour point representation; and (h) skeleton representation.

appearance of a target in an image is generally characterized by calculating the probability density of the appearance features (color [31], texture) of the internal regions of the target, which are usually described using the contours of target or directly using the areas contained in rectangles and ellipses. Template-based target representation is another appearance-based characterization method, which is simple and intuitive but only applies to target-tracking scenarios where only small deformations occur.

The target shape and appearance characterization methods can be combined, resulting in better tracking performance due to adequate target characterization, e.g., the active appearance model [32]. However, the target representation method should be closely related to the object tracking algorithm under the application scenario. For example, a point-based target representation method is effective when the tracking target is tiny. A rectangular box or elliptical box-based representation method is a better choice when geometric shapes can represent the target.

In fish target tracking, the appearance and shape of the fish target constantly change as the fish swim. Object representation based on the object's empirical geometry shape or integrating the shape and appearance of the fish objects can be used to obtain accurate tracking results of the fish target by simultaneously updating the geometry and appearance shape of the target. In addition, with the dramatic advances in artificial intelligence and big data technologies, deep learning techniques are good at discovering the intricate underlying structures in high-dimensional data. They have been widely used in many scenarios. Unlike traditional approaches that require domain expertise for feature representation, deep learning models represented by convolutional neural networks (CNNs) can automatically learn feature representation from raw data. Figure 4 shows a deep learning-based fish target representation scheme.

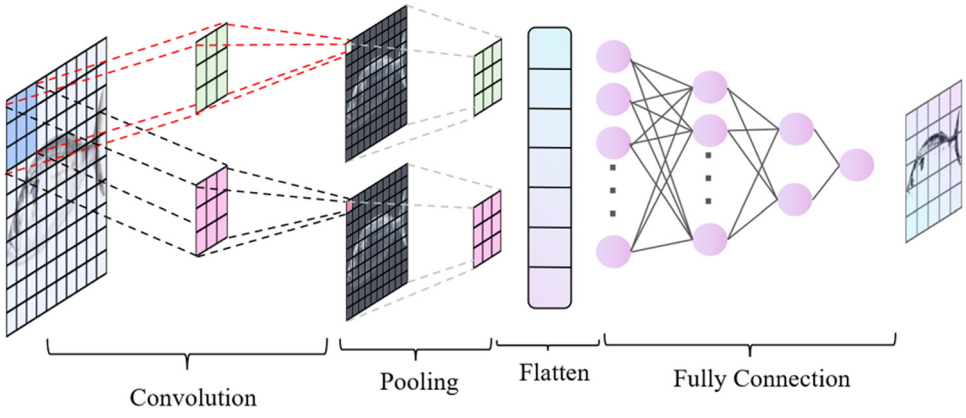

Figure 4: Fish object feature representation based on convolution neural network.

2.2.2 Fish object detection

Fish activity behavior recognition requires accurately detecting the corresponding fish target in image sequences. In the early years, the effective detection of fish objects is also the basis for accurately tracking fish targets. Motion target detection and tracking from dynamic image sequences is a fundamental research topic in computer vision [33].

In machine vision-based ecological fish pond monitoring, fish target detection also means effectively distinguishing and extracting the target of interest (fish target) from the background (water pattern) image for feature extraction of the fish (motion) state. The target detection methods are classified into motion information-based, target image segmentation-based, and classifier-based fish target detection.

The motion information-fused method performs fish (motion) target detection based on the changes in local areas of the image caused by the motion of fish targets in the image sequence. Image segmentation achieves target detection by performing image segmentation processing on each frame and effectively distinguishing foreground (tracking) targets in the image based on the segmentation results. While the third method builds a fish target classifier by learning the fish features in the sample image, which can detect fish targets in a single frame, the first method is powerless for fish targets that are not very active or dead. In contrast, the second and third methods are equally effective. Therefore, in actual fish target detection, several methods can be combined to detect effectively.

(1) Motion information-fused fish object detection

In the case of visual monitoring of "ecological fish ponds," a stationary monitoring camera is generally used so that the movement of fish targets typically causes the variation area in the adjacent frames. Therefore, calculating the pixel difference between adjacent frames is an effective method for fish target detection. It can be further subdivided into the pixel-frame difference methods, background subtraction methods, and optical flow methods with their variants for fish target detection.

(a) Inter-frame differential method

The inter-frame differencing method is a method used to obtain the moving object contour by performing differencing operations on successive frame images in a video image sequence [34]. Let $I(x, y)_t$ be the pixel values (possibly grayscale values or RGB color space values) of the pixel point (x, y) in the t frame image and T represents a pre-set threshold. The forward differential image $D(x, y)_{t+1}$ of the t frame image is described as follows:

$$(x, y)_{t+1} = \begin{cases} 1, & \text{if } \|I(x, y)_t - I(x, y)_{t+1}\| \geq T \\ 0, & \text{others} \end{cases} \tag{1}$$

The inter-frame difference method is straightforward in principle, fast in the calculation, and convenient in implementation. However, the limitation is that traditional inter-frame difference methods cannot detect a complete object. It is insensitive to slow-moving objects. In [35,36], the authors proposed a cumulative frame difference method, which obtained the region of moving targets by finding the part standard to the difference between the front and back frames in consecutive multiple sequential images, avoiding the problem of target motion (fish target) due to the situation that the target motion (fish target) cannot be detected due to its slow speed. Compared with the traditional inter-frame difference method, the continuous frame difference method can obtain better detection results with no increase in the computational effort [37].

(b) Background subtraction

Background subtraction is a method that uses the current image frame to subtract the image's background to obtain the moving targets. The key to this method is to be able to build the background model and achieve dynamic updates. Commonly used background models include Gaussian mixed scale models (Gaussian mixture mode (GMM)) [38], Vibe [39], CodeBook method [40], Sobs method [41], as well as multiframe averaging methods, and so on.

For slowly changing backgrounds, a weighted average of historical images over time can be used directly to approximate the background or a simple normal distribution can be used to characterize the change in pixel grayscale by subtracting the most recent background image from the current image. The region of the difference image that is larger than a certain threshold is caused by the motion of the active target.

In fish target detection, a breeze, water fog, and so on will cause dynamic scenes such as ripples on the water's surface. Then, the background may behave as a multipeaked distribution, and it is difficult to use a simple Gaussian distribution model to express the changes in such a dynamic background effectively. Therefore, the GMM [38] is a more effective background modeling approach. That is, the image pixel distribution is described as multiple weighted Gaussian components to illustrate the probability of the pixel values on the point (x, y) in the t frame image, namely,

$$P(I(x,y)_t) = \sum_{i=1}^{K} \omega_{i,t} N(I(x,y)_t;\ \mu_{i,t}, \zeta_{i,t}), \tag{2}$$

where $\eta(I(x,y)_t|\mu_{i,t}, \zeta_{i,t})$ represents the i Gaussian component, $\mu_{i,t}$ and $\zeta_{i,t}$ represents mean and covariance, and $\omega_{i,t}$ represents the weights of the i Gaussian component. For computational convenience, the covariance $\zeta_{i,t}$ matrix is often used simply, $\zeta_{i,t} = \sigma_i^2 \mathrm{I}$, which means that the individual color spaces of the pixels are considered independent and have the same variance.

The most commonly used method for GMM parameter estimation is the expectation maximization (EM) [42] algorithm. Since the EM algorithm needs to be executed for each pixel in the image, this imposes a significant computational overhead on the background update. Therefore, researchers often use the online K-means approximation (K-means approximation) method to update the GMM mixture model. Detailed specific steps are as follows:

For any pixel value $I(x,y)_t$ of the t frame image, first, determine whether $I(x,y)_t$ belongs to one of the current K Gaussian components. The judgment rule is that the pixel value is within 2.5 times the standard deviation of the corresponding Gaussian distribution. If there is no matching distribution, then the most unlikely distribution (the least probable distribution) of the K components is replaced by a new Gaussian distribution, and the mean value is the current pixel value and initializes a high variance and a low a priori weight. Then, the weights of K Gaussian components are updated at the same time,

$$\omega_{i,t} = (1 - \alpha)\omega_{i,t-1} + \alpha(M_{i,t}), \tag{3}$$

where α is the so-called learning rate and $M_{i,t}$ is 1 when the model matches and 0 otherwise. $1/\alpha$ defines a time constant that determines the background change, and $\omega_{i,t}$ is the average result of the posterior probability of a valid causal low-pass filter, that is, the posterior probability that a pixel-valued observation in the image matches the i Gaussian component in time 1 to frame t. The mean and variance update rules are as follows: for K Gaussian components, the mean and variance of Gaussian components without pixel-value matches remain at their original values, and Gaussian components with new observation matches are updated using the following formula,

$$\mu_t = (1 - \lambda)\mu_t + \lambda I_t \tag{4}$$

$$\sigma_t^2 = (1 - \lambda)\sigma_{t-1}^2 + \lambda(I_t - \mu_t)^T(I_t - \mu_t), \tag{5}$$

where $\lambda = \alpha N(I_t;\ \mu_i, \sigma_i)$ is the learning factor of the adaptive current distribution.

After obtaining the mixture Gaussian model of the pixels in the image sequence, these Gaussian components based on the background B in the image sequence should be the ones with the minimum variance and the most obvious evidence (the greatest supporting evidence). In the literature [38], it performs background acquisition by the following method.

$$B_t = \underset{b}{\operatorname{argmin}} \left(\sum_{k=1}^{b} \omega_{k,t} > T_{\mathrm{e}} \right), \tag{6}$$

where T_{e} is a predetermined threshold, representing the smallest component of the pixel that needs to be considered as the background. K Gaussian components are sorted using the indicator ω/σ, and the smallest number of them (satisfying the formula (6)) is selected to obtain the corresponding background. After obtaining the corresponding background image, the background B_t is subtracted from the current image frame I_t, and the region where the difference is greater than a certain threshold is the fish object region in the current image.

The Vibe method [43] is another background modeling method that has become more prevalent recently. Unlike GMM, this method will take the neighborhood history pixels of a point in the image to create the background model. A sample set is stored for each background point, and then each new pixel value is compared with the sample set to detect whether it belongs to the background. The background updating strategy mainly includes the memoryless update strategy, which randomly replaces one sample value of the sample set of the pixel point with a new pixel value each time if it is determined that the background model of the pixel point needs to be updated; the temporal sampling update strategy that updates the background model according to a specific update rate; and the spatial neighborhood update strategy that randomly selects a background model of the neighborhood of the pixel point and updates the background chosen model with the new pixel point. This method reduces the background model-building process and can effectively deal with sudden changes in the background. Still, it is easy to introduce artificial "ghosting" regions due to the possible use of the pixel initialization sample set of moving objects.

(c) Optical flow method

The optical flow method is another relatively prevalent motion estimation method. This method uses the optical flow equation to calculate the motion state vector for each pixel point. The optical flow field is the projection of the moving object in the 3D velocity vector on the 2D imaging plane, representing the instantaneous change of the object displacement in the image based on the moving object in the image that is detected [44]. Let the pixel point, (x, y, t) in the image video (sequence), and after some time (Δt), the pixel point becomes $(x + \Delta x, y + \Delta y, t + \Delta t)$, and since the pixel value of the pixel point changes independently from the position, $I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t)$, which is obtained by deriving for the time t:

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial x}V_y + \frac{\partial I}{\partial t} = 0, \tag{7}$$

where V_x and V_y represent the velocity of the pixel point in the x and y axes, respectively, which is the optical flow of $I(x, y, t)$.

However, the optical flow constraint equation described in equation (7) cannot determine the unique optical flow because the number of variables is too large, so it is necessary to introduce constraints to obtain a particular optical flow field. Commonly used constraints include the spatio-temporal gradient method, which uses the spatio-temporal differentiation of the image sequence grayscale to calculate the visual flow field at each pixel point on the image. This Horn-Schunck method assumes that the optical flow varies smoothly over the entire image, and the Lucas-Kanade method takes a constant motion vector over a small spatial neighborhood [45].

The velocity vector characteristics of each pixel point are obtained based on the motion vector field calculated by the optical flow method. The dynamic characteristics of objects can be achieved. If no objects with relative motion are occurring in the image sequence, the optical flow vector continuously changes throughout the image area. When there is a target object with the relative movement (target and background) in the image sequence, the velocity vector formed by the moving target object should be different from the velocity vector of the background, and the moving object's position can be calculated accordingly. However, although the optical flow method can detect various moving objects, the method is computationally intensive, and the process is seriously affected by illumination and ripples, which can quickly produce false detection objects.

Given the shortcomings of the traditional optical flow method, Li-chao et al. [46] proposed an improved optical flow method. First, restrictions are added to the optical flow algorithm to make different constraints used at other gradient points; then, the GMM is fused. Finally, the proposed fusion algorithm compares the number of object frames and their overlap area to display the fused fish detection information in the surveillance video. This method involves adding a constraint (parameter) to the optical flow calculation to use the luminance constancy constraint at the points with a large gradient. The visual flow field consistency constraint is used at the points with a slight gradient. Therefore, the binary weighting function can be defined as follows:

$$\delta(x, y) = \begin{cases} 0, & I_x^2 + I_y^2 > V \\ 1, & \text{Other} \end{cases} \tag{8}$$

$$S = \iint [(I_x u + I_y v + I_t)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2)]\mathrm{d}x\mathrm{d}y, \tag{9}$$

where V is a threshold, e.g., V is usually set as 0.5. When the sum exceeds the threshold, the function value is 0; in other cases, the function value is 1. After the restriction is incorporated into formula (9), we can obtain the following equation:

$$S = \iint [\delta(x, y)\cdot(I_x u + I_y v + I_t)^2 + \lambda(u_x^2 + u_y^2 + v_x^2 + v_y^2)]\mathrm{d}x\mathrm{d}y. \tag{10}$$

Equation (10) is used to calculate the optical flow vector, where $u_x^2 + u_y^2 + v_x^2 + v_y^2$ can be replaced by $k(\bar{u}_{i,j} - u_{i,j}) + k(\bar{v}_{i,j} - v_{i,j})$, k is 3, and $\bar{u}_{i,j}$ and $\bar{v}_{i,j}$ is a four-neighborhood mean value, on one hand, to facilitate the calculation, and on the other hand, according to the spatial correlation of the image, it can ensure the universality of the value taken. The specific calculation is shown in equation (11).

$$\begin{cases} \overline{u_{i,j}} = \dfrac{1}{4}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) \\ \overline{v_{i,j}} = \dfrac{1}{4}(v_{i-1,j} + v_{i+1,j} + v_{i,j-1} + v_{i,j+1}). \end{cases} \tag{11}$$

(2) Image segmentation-based fish detection

The representative image segmentation-based object detection methods are mean shift clustering (MSC), active contour (AC), and graph cut (GC)-based object segmentation methods.

The MSC is an effective nonparametric iterative clustering algorithm. By repeatedly iterating to search the region with the densest sample points in the feature space, it does not need to assume the nature of the distribution. It does not need to set the number of clusters in advance, so it has high iterative efficiency. Comaniciu and Meer [47] extended this method to image feature space analysis and successfully applied it to image segmentation, smoothing, non-rigid object tracking [48], and so on. The MSC-based image segmentation algorithm is based on the color and spatial information of the image ($[l, u, v, x, y]$, $[l, u, v]$ represents color information and $[x, y]$ represents coordinates). This algorithm randomly selects some clustering centers in the picture. Then it moves these clustering centers to the center of pixels in the multidimensional hyperellipsoid corresponding to the clustering center in the iterative procedure until the clustering center and the center point of the hyperelliptic converge.

The AC [49] method uses a continuous closed curve to express the object edge. By setting a rectangular box containing the object and the closed curve of energy functional evolution, when the energy reaches the minimum, the curve falls right on the object contour. The general energy function related to the object profile Γ is defined as follows:

$$E(\Gamma) = \int_0^1 E_{\text{int}}(\mathbf{v}) + E_{\text{im}}(\mathbf{v}) + E_{\text{ext}}(\mathbf{v})\mathrm{d}s, \tag{12}$$

where s is the arc length of the curve corresponding to the object contour Γ, and E_{int} is the constraint adjustment term, which usually includes the curve curvature item. The shortest object contour curve is found by the first-order ($\nabla\mathbf{v}$) or the second-order ($\nabla^2\mathbf{v}$) continuous term, E_{im} represents the apparent energy of the object, which can be calculated by local energy or global energy. The local energy is often expressed in the form of a gradient around the object contour. The global energy is calculated inside or outside the object area. E_{ext} stands for external energy. When the contour Γ is close to the edge of the object, the gray gradient of Γ will increase, and the energy of formula (12) is minimum when the curve stops evolving, the curve adheres to the edge of the object contour.

The traditional AC method is time-consuming. Thus, some researchers have proposed the leaky integrate-and-fire (LIF) model [50]. The LIF model can segment the outline and texture of an object very well. The traditional AC model uses the Gaussian function to fit the probability distribution of each Ω_i pixel, which will lead to tedious calculations. The LIF model uses a simple function approximation method to fit the probability distribution, greatly reducing the amount of calculation. If $\chi_{E_i}(x)$ is expressed as a window function W_k, and

the coefficient a_i is taken as the mean value of the image on W_k, any measurable function can be approximated by a simple function as follows:

$$s_n(x) = \sum_{i=1}^{n} a_i \chi_{E_i}(x). \tag{13}$$

The GC method [51] considers image segmentation as a region segmentation problem of a graph. It regards all the pixels in image I as the vertices of graph G. The edge weight of the graph is generally obtained by calculating the similarity of color, texture, and brightness between vertices in the graph, and then the image segmentation problem is actually to divide the vertex set in graph G into N disjoint subgraphs by edge pruning. The trimmed edges between these subgraphs are called cuts. Region-based image segmentation has become one of the research hotspots of automatic or semi-automatic image segmentation because of its high efficiency and robustness. Given the uncertainty in the image region segmentation method of an ecological fish pond, a region segmentation method based on the cloud model is proposed in the literature [51]. Firstly, the growth criterion in regional growth is determined based on cloud transformation. Then, the reverse cloud algorithm is used to realize the transformation process of the segmented region from the quantitative pixel set to the qualitative cloud concept. Finally, the adjacent areas are merged based on the cloud synthesis algorithm, and the uncertain image segmentation based on region is realized [51]. Figure 5 shows the segmentation results of the fish image using three segmentation methods. Generally, these methods have high computational complexity and are mainly oriented to single-target image segmentation. The effect of these methods is not particularly ideal for scenes with multiple fish targets in the monitoring field of vision. However, these methods can still detect fish targets whose motion features are not prominent.

(3) Classifier-based fish detection

In ecological fish pond monitoring, the fish object detection in fish image videos can be converted into a two-class classification problem. This detection problem is similar to image segmentation, except that image segmentation generally considers only a single frame (static) image to distinguish the foreground target of interest from the background in a single frame [52].

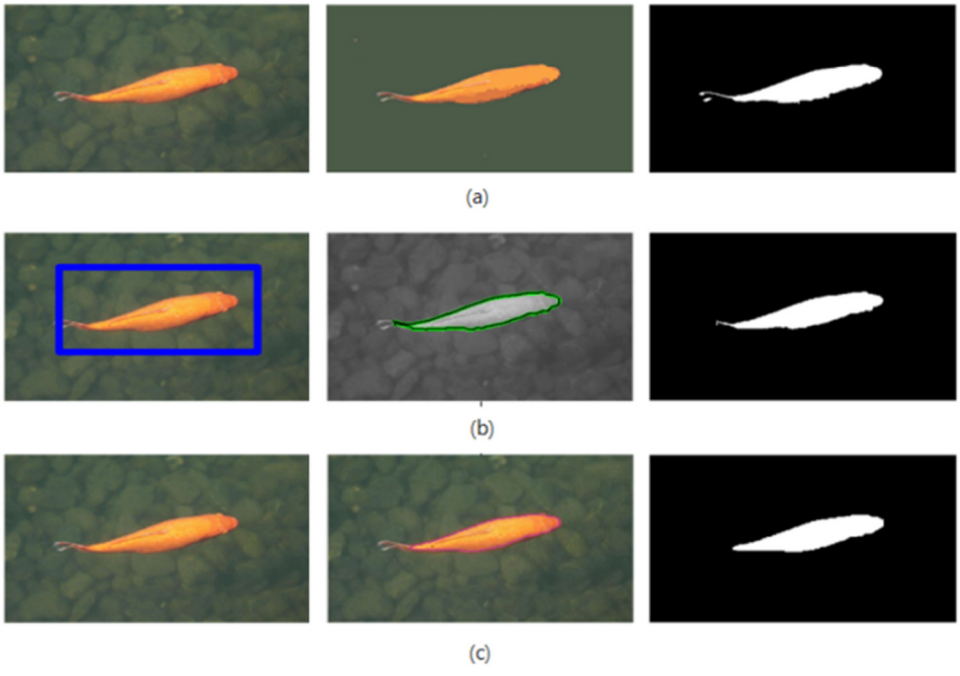

Figure 5: Different methods for fish object segmentation. (a) Segmentation of fish objects by MSC, (b) AC fish object segmentation, and (c) GC fish object segmentation.

Generally, object detection based on classifier learning includes object feature preparation and classifier establishment. Various classifiers can be used for classifier selection, including neural networks, support vector machines, random forests, and discriminative analysis classifiers. Each of these classifier models has its advantages and disadvantages, and it is difficult to determine what classifier should be selected to obtain better generalization performance and minimal training overhead.

Adaboost classifiers based on ensemble learning were mainly used before 2012 for various object detection applications, especially face detection, achieving extremely high detection performance. The basic idea is to train multiple weak classifiers for the same training set, and then the collection of these weak classifiers can form a final robust classifier [53].

Regarding feature selection, commonly used methods include color, texture, and contour. In earlier years, adaboost-based object detection methods usually used simple Haar features [54,55]. Recently, compared with the traditional manual feature method, deep learning-based approaches can avoid the tedious manual feature design process and automatically learn deeper features with more distinguishing power. In addition, the deep learning-based approach unifies feature extraction and classifier teaching in a single framework to facilitate end-to-end learning.

Since 2012, deep learning technologies have rapidly progressed, and many classic deep learning models have been proposed, as shown in Figure 6. This class of methods directly predicts a detection frame for each object in an end-to-end way. Based on the aforementioned technologies, computing power, and data, object detection has started to make significant progress in terms of accuracy and efficiency, with the emergence of region-based convolutional neural network (R-CNN) (Girshick et al.) [56], SSD (single shot detector) (Liu et al.) [57], YOLO (you only look once) (Redmon et al.) [58], and DETR (detection transformer) (Carion et al.) [59].

The YOLO (you only look once) series model is one of the representative's works on single-stage target detection methods. As shown in Figure 7, YOLO (Redmon et al.) [56] can directly divide the fish image into subregions of $N \times N$ size and predicts the probability, class, and position offset of the presence of objects in each subregion. In YOLO, the input fish image is divided into an $S \times S$ grid, and the network assigns each object to the cell where the center of that object is located. A grid cell predicts multiple bounding boxes. Each prediction array consists of five elements: the centers x and y of the bounding boxes, the dimensions w and h of the boxes, and the confidence score (the probability that the box contains the object). YOLO is inspired by the GoogLeNet model [60] of image classification, which uses a cascade module of smaller convolutional networks. It is pre-trained on ImageNet data [61] until the model reaches a high accuracy and then modified by adding randomly initialized convolutional and fully connected layers. Each grid cell predicts only one class during training because it converges better. YOLO greatly surpasses contemporary single-stage real-time models in terms of accuracy and speed, with processing speeds of up to 45 fps and even 155 fps for Fast YOLO in smaller models. However, it also has significant drawbacks, mainly the poor localization accuracy for small and dense objects and the limited number of objects predicted per cell. These problems have been solved in the subsequent versions of YOLO.

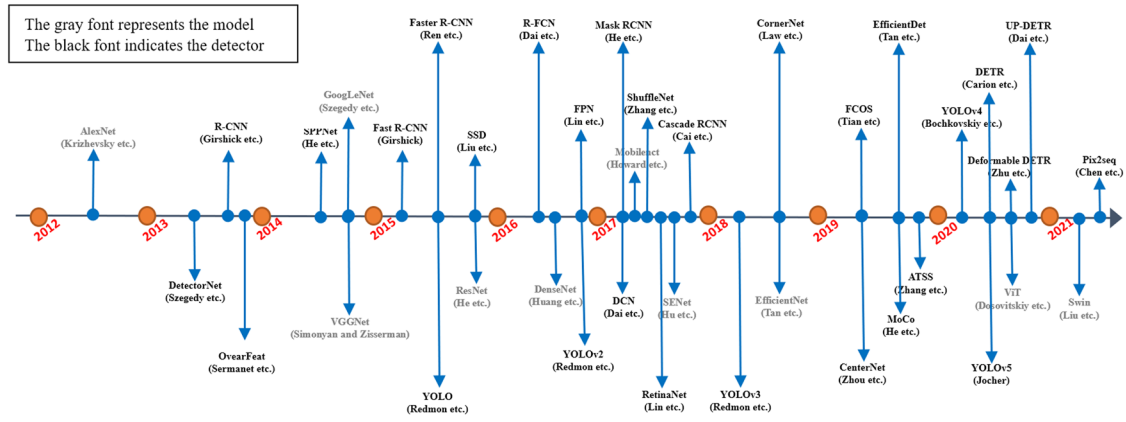

Figure 6: Development process of visual object detection based on deep learning.

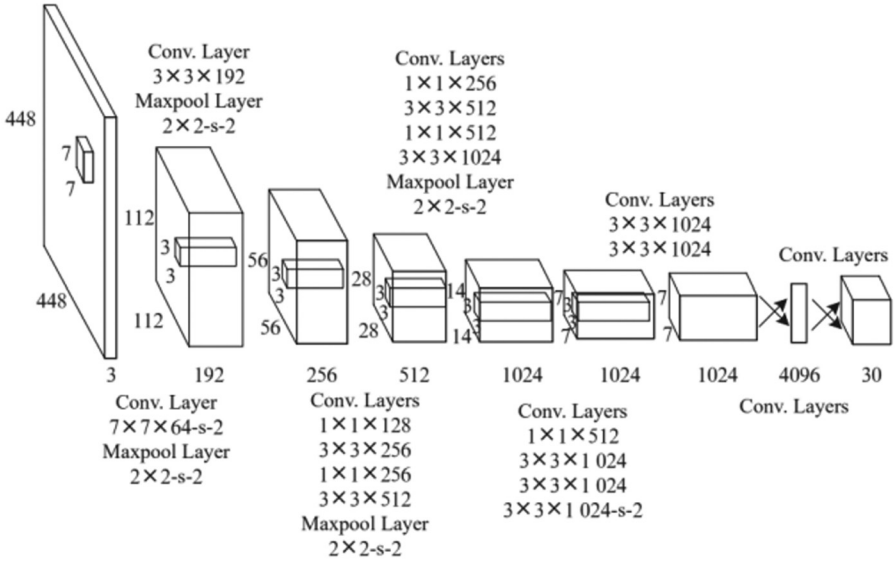

Figure 7: YOLO network model.

YOLOv2 [62] is an improvement on YOLO, providing a simple balance between speed and accuracy, while YOLOv2, also known as YOLO9000, can predict 9000 object classes in real time. YOLOv2 replaces the backbone architecture of GoogLeNet with Darknet-19. It combines many impressive technologies, such as batch standardization (batch normalization) [63] to improve convergence, joint training of classification and detection systems to increase the number of detection types (i.e., location training with object detection label data in COCO, object classification with data in ImageNet). It removes the full connection layer to improve speed and uses the anchor mechanism in Fast R-CNN [64] to provide a better prior frame and improve the recall rate. YOLOv2 provides better flexibility in choosing models in terms of speed and precision, and the new architecture has fewer parameters, and it is "better, faster, and stronger."

YOLOv3 [65] is an "incremental improvement" over previous versions of YOLO. Redmon and Farhadi [65] replaced the feature extractor network with a more extensive Darknet-53 network. At the same time, it also combines various technologies, such as data expansion, multiscale training, batch standardization, and so on. A binary cross-entropy classifier replaces the Softmax of the classifier layer. Although YOLOv3 is faster than YOLOv2, it lacks any breakthrough changes compared with its predecessor.

YOLOv4 [66] designed a fast and easy-to-train object detector that can work on existing simple equipment. It uses a "free package," that is, a method that only increases the training time without affecting the reasoning time. YOLOv4 uses data enhancement technology, regularization method, class label smoothing, CIoUloss [67], cross small batch normalized CmBN (cross mini-batch normalization), self-confrontation training, and other techniques to improve training. The network also adds a method that only affects the reasoning time called "special packet," including Mish activation [68], cross-stage partially connected network CSPNet (Cross Stage Partial Connections) [69], SPP-Block [70], PAN (path aggregation network) [71], path aggregation block, multi-input weighted residual connection (MiWRC), and so on. It has an ImageNet pre-trained CSPNetDarknet-53 backbone, SPP and PAN block neck, and YOLOv3 as the detection head. Most existing detection algorithms require multiple graphic processing units (GPUs) to train the model, but YOLOv4 can be easily trained on a single GPU. It is twice as fast as an efficient detector with the same performance. YOLOv5 (https://github.com/ultralytics/yolov5) provides four different sizes of object detectors to meet the needs of different applications.

SSD (single shot multibox detector) [57] is the first single-stage detector that can match the accuracy of two-stage detectors (such as faster RCNN) while maintaining real-time speed. SSD is based on VGG-16, as shown in Figure 8, with an auxiliary structure to improve performance. When the image features are not too rough, SSD detects smaller objects early in the network, while the deeper layer is responsible for the default frame and aspect ratio offset. During the training process, SSD matches each real box with the default box with the best

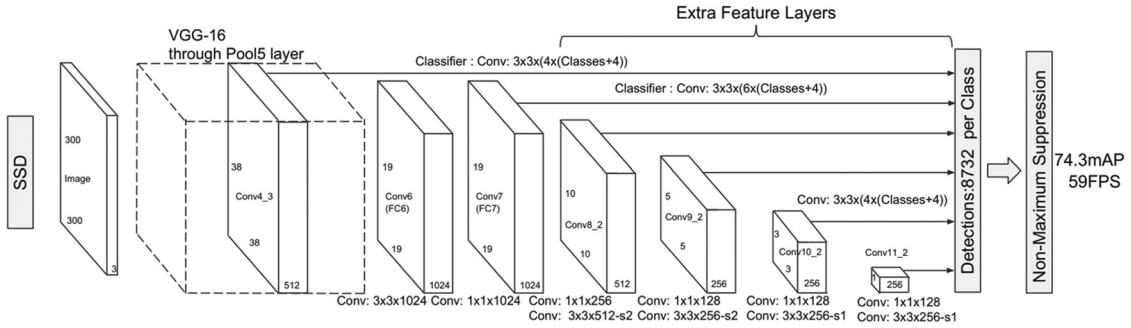

Figure 8: SSD model.

Jaccard overlap and trains the network accordingly, similar to Multibox [72]. At the same time, hard, harmful mining and massive data expansion are also used. Similar to DPM, it uses the weighted sum of localization and confidence loss to train the model, and the final output is obtained by performing non-maximum suppression. Although SSD is much faster and more accurate than state-of-the-art networks such as YOLO and R-CNN, it has difficulties detecting small objects.

Carion et al. proposed a transformer-based detection method DETR [59], which uses a CNN to extract features, and directly predicts the position and classification score of objects based on the transformer codec network. Figure 9 shows the DETR transformer architecture diagram.

However, DETR has shortcomings such as slow convergence speed, the relatively poor performance of small-scale object detection, and so on. Zhu et al. [73] proposed deformable DETR inspired by deformation convolution to overcome them. Unlike the global-based attention mechanism adopted by DETR, deformable DETR offers a deformable attention module based on local sparsity. Experimental results demonstrate that the deformable DETR has faster convergence speed and better performance for small-scale object detection. The

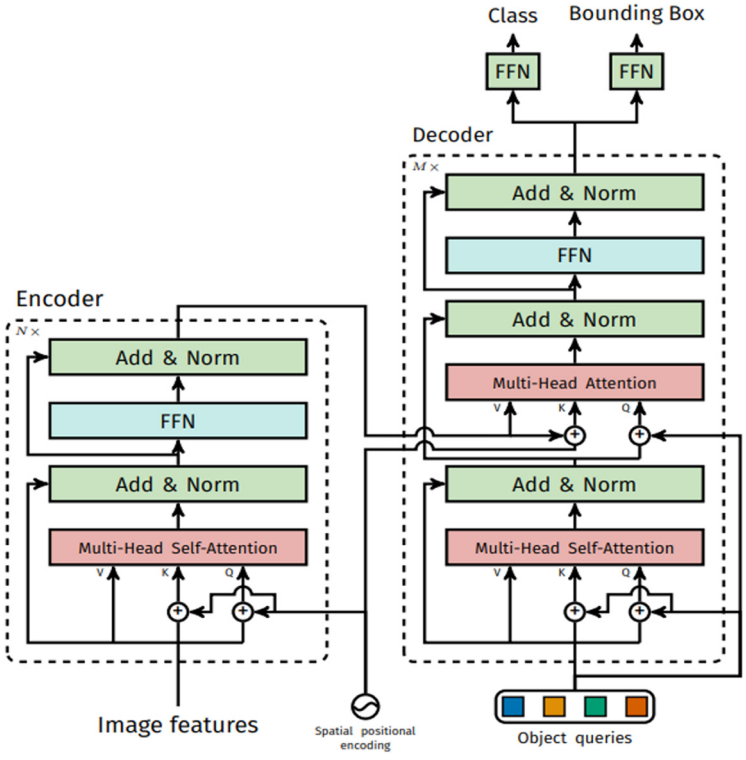

Figure 9: DETR transformer architecture.

deformable attention module used in deformable DETR is shown in Figure 10. Inspired by deformable convolution, the deformable attention module will only pay attention to a small number of crucial sampling points near the reference point, regardless of the space size of the feature map. The problems of convergence and feature space resolution are alleviated by allocating a few keys to each query. Given the input feature graph $x \in \mathbb{R}^{C\times H\times W}$, Q is used as the index of a query element, and its feature representation is z_q. The attention feature of the two-dimensional reference point p_q, the deformable attention, is calculated as follows:

$$\text{DeformAttn}(z_q, p_q, x) = \sum_{m-1}^{M} W_m \left[\sum_{k-1}^{K} A_{mqk} \cdot W'_m x(p_q + \Delta p_{mqk})\right], \tag{14}$$

where m is the index of attention head, k is the index of sampled keys, and K is the number of $(K \ll HW)$ of all sampled keys. Δp_{mqk} and ΔA_{mqk} represent the sampling offset and attention weight of the k sampling point in m attention head, respectively.

Studies have shown that DETR has poor performance in detecting small fish objects. The existing detectors usually have multiscale features, and small objects are usually detected on high-resolution feature images, but DETR does not use multiscale features. Hence, mainly high-resolution feature images will increase the unacceptable computational complexity of DETR. To solve this problem, it can apply the deformable attention module to multiscale feature maps. Let $\{x^l\}_{l=1}^{L}$ represent the input multiscale feature map, where $x^l \in \mathbb{R}^{C\times H_l\times W_l}$. $\hat{p}_q \in [0,1]^2$ represents the normalized coordinates of the reference point of each query element Q. The multiscale Deformable attention module can be expressed as follows:

$$\text{MSDeformAttn}(z_q, \hat{p}_q, \{x^l\}_{l=1}^{L}) = \sum_{m=1}^{M} W_m \left[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \cdot W'_m x^l(\Phi_1(\hat{p}_q) + \Delta p_{mlqk})\right]. \tag{15}$$

Using deformable DETR to achieve fish object detection can improve computing speed. Compared with traditional object detection, deformable DETR does not use NMS and CNNs but uses the idea based on a transformer to learn the object's position. deformable DETR uses a self-attention mechanism to learn the offset of each position, that is, selecting some points z_q in the fish image. At first, z_q only cares about the adjacent points. Through the deformable attention module, learn the offset to update the position of the z_q

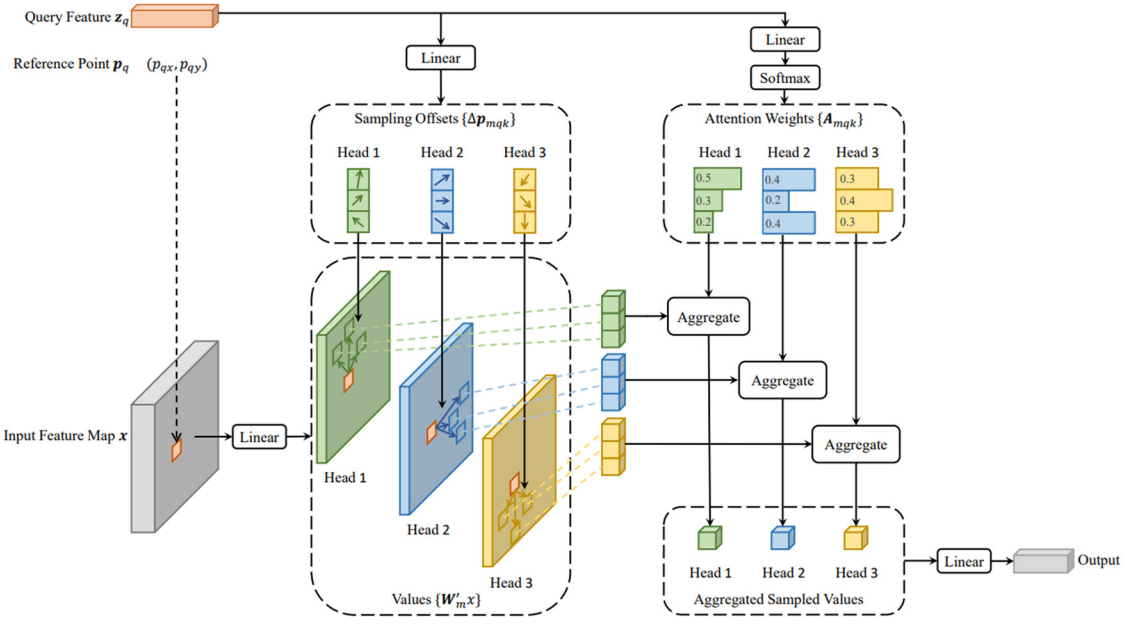

Figure 10: Deformable attention module.

adjacent points, so that we can dynamically update the position of the fish object, and finally predict the specific position of the fish.

In monitoring fish activity behavior in ecological fish ponds, fish activities are related to water quality and belong to random activities in standard (nontoxic) water quality. Fish activity ability is low under toxic water quality or severe pollution conditions and may even be close to death [196,197]. Therefore, to detect fish targets effectively, it is often necessary to combine dynamic target detection methods and static fish target detection methods based on classifier learning to meet the requirements of fish detection in various states.

2.2.3 Fish object tracking

The so-called object tracking estimates the trajectory of the moving object in the image plane. After the fish detection, tracking the fish objects in the video frame by frame is necessary to analyze the fish activity behavior. In fact, like object detection, object tracking is also a fundamental research topic in computer vision. Nowadays, the boundary between object tracking and detection is not so obvious. For example, for a single object, if the object's position is detected in each frame, the object tracking is completed. Of course, for multiobject tracking, in addition to detecting the corresponding object in each frame, the position of each object should be confirmed in the subsequent sequence. Therefore, object tracking mainly includes object detection, object model update, object searching and location association, and so on.

In the current research, the in-depth exploration of target-tracking algorithms for fish behavior recognition, particularly their learning processes, data acquisition methods, and data fusion strategies, has become a significant topic. The adoption of supervised and semi-supervised learning approaches, through the deep analysis and training of a large volume of annotated data, has proven to significantly enhance the accuracy of predicting fish movement trajectories. This highlights the crucial role of high-quality datasets in algorithm development. Simultaneously, the quality and diversity of data collection are vital for the performance of algorithms. High-resolution and varied image data captured by underwater cameras under different environmental conditions provide a rich training resource for the algorithms, strengthening the models' generalization capabilities and adaptability. Moreover, data fusion technology has shown great potential in improving the accuracy and reliability of fish behavior recognition systems. Integrating data from multiple sensors provides a more comprehensive and in-depth understanding of fish behavior and its environmental context, offering robust support for precisely identifying fish behavior.

According to the classification of object tracking methods by Ruize et al. [74], as shown in Figure 11, the main techniques of single target tracking can be divided into methods based on correlation filtering (CF) and twin network (Siamese network) framework. The CF target tracking algorithm was proposed in 2010. Because of its good balance between tracking accuracy and algorithm speed, it has rapidly developed into one of the mainstream object-tracking methods. Since 2014, in the mainstream object tracking challenge, i.e., visual object tracking challenge (VOT), the CF target tracking algorithm has obvious advantages in the number and performance of the competition. The target tracking algorithm based on the Siamese network appears later than the CF method, and the pioneering work is the SiameseFC [75] algorithm, which appeared in 2016. Since then,

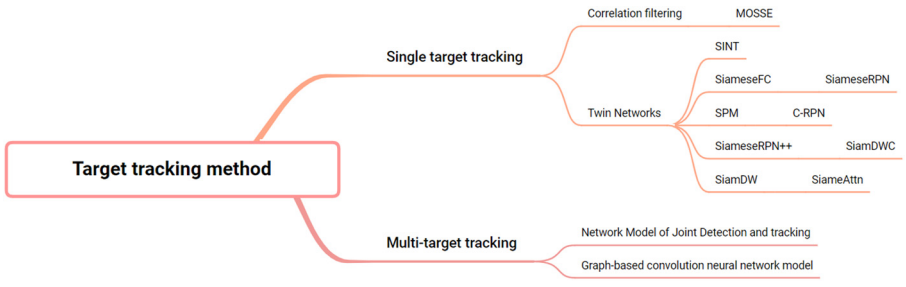

Figure 11: Object tracking method.

the target tracking method based on a twin filter has developed rapidly and achieved significant advantages in the algorithm performance.

(1) CF

The object tracking algorithm based on CF theory has made remarkable progress in the object tracking task. It has obvious advantages in tracking accuracy and running speed, so it has been one of the mainstream frameworks of video object tracking tasks in recent years. The application of CF in computer vision can be traced back to the eighties of the 20th century. In 2010, the MOSSE algorithm [76] applied CF to video object tracking for the first time and achieved excellent accuracy and ultra-high speed. The good performance of CF theory in object tracking is mainly due to the following two reasons: (1) the CF object tracking method implicitly uses cyclic translation operation to expand the training samples, thus greatly enriching the diversity of training samples. The robustness and accuracy of the algorithm are improved. (2) Fast Fourier transform (FFT) accelerates the calculation of complex convolution operations in the frequency domain, reduces the amount of calculation, and increases the efficiency of model solving. The MOSSE filter is solved as follows:

$$H^* = \frac{\sum_i G_i \odot F_i^*}{\sum_i F_i \odot F_i^*}, \tag{16}$$

where the numerator is the convolution of the input image with the desired input image, while the denominator is the energy spectrum of the input image.

Due to the excellent accuracy and speed, the object tracking algorithm based on CF theory has developed, and many related algorithms show promising results on the open dataset.

(2) Twin networks

A twin network (Siamese Network) [77] refers to two parallel networks with the same or similar structure. The related algorithms based on twin networks have been applied to template matching, similarity measurement, and other tasks as early as the 1990s. Because of its few parameters and fast running speed, the twin network is applied to many different tasks. Twin network was first used in a video object tracking task in SINT [78] and SiameseFC [75], published in 2016. For the first time, the algorithm transformed the object tracking problem into a matching problem between a given template and a candidate image. The earliest SiameseFC algorithm achieved high tracking accuracy, maintained an ultra-high algorithm speed (86 fps), and provided a good foundation for the follow-up methods. Figure 12 shows the network structure of the SiameseFC algorithm, where the upper branch z represents the object model image generated from the object region given in the first frame of the video sequence, the input of the lower branch is the current frame search region, x represents different object candidate images inside the search region, and z and x map the original image to the feature space through the feature mapping operation φ of the system. The feature vectors with the same channels are finally convolved by the * operation to obtain the response map, where the values at each

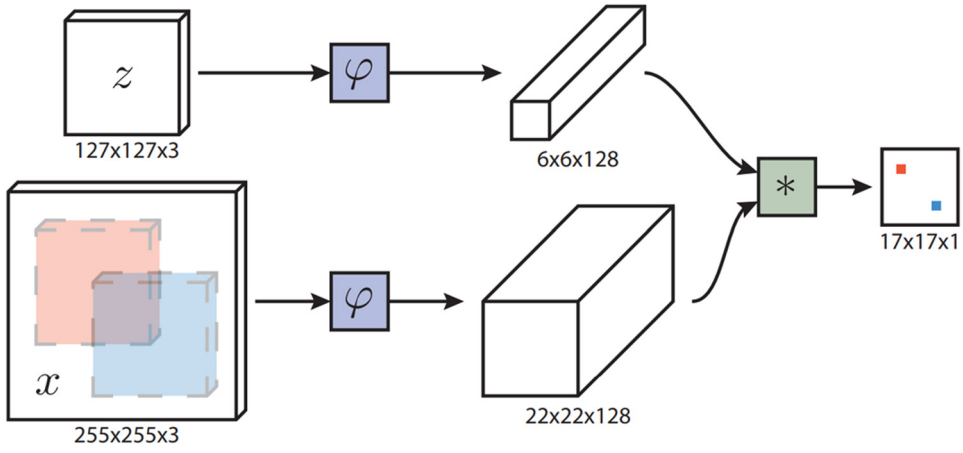

Figure 12: Fully convolutional twin network (SiameseFC) object tracking algorithm.

position represent the similarity between different object candidate images and the object template image. The most similar object candidate region is selected by taking the maximum value to complete the object localization tracking.

$$f(z, x) = \varphi(z) * \varphi(x) + b, \tag{17}$$

where $\varphi(z)$ is used as a convolution kernel, convolution is performed on $\varphi(z)$, where the similarity is the largest, then the natural response value is large, and the position of the target z in x can be obtained.

The network structure of SiameseFC contains only convolutional and pooling layers, so it is also a typical fully convolutional Siamese Network.

After the SiameseFC algorithm was used for video object tracking, it attracted wide attention. It brought a lot of research from related scholars due to its simple network structure and high algorithm speed, among which the classical representative is the SiameseRPN [79] series algorithm. Siamese region proposal network (RPN) algorithm is based on SiameseFC and incorporates the classical algorithm of object detector of Fast R-CNN [64] which is based on the SiameseFC, including the idea of RPN, i.e., region candidate network to extract object candidate frames. The loss function of Fast R-CNN is as follows:

$$L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda[u \geq 1]L_{\text{loc}}(t^u, v), \tag{18}$$

where L is the multitask loss, u is the true number of categories, and v is the regression loss of the true bounding box.

As shown in Figure 13, the RPN can generate candidate frames with different scales, thus mainly solving the problem of severe object deformation in object tracking. DaSiamRPN [80] improves the model generalization ability by the training set data augmentation. Both the recent SPM [81] and C-RPN (CascadeRPN) [82] algorithms are multistage extensions of the SiameseRPN, where SPM introduces the ideas of the classical object detection method, Faster R-CNN [83], to the object tracking network. SiameseRPN++ [84] and SiamDWC [85] focus on how to use a deeper backbone network in object tracking. SiameseRPN++ [84] improves the sample sampling strategy in SiameseRPN, prevents the problem that positive samples are located in the center of the image and affect object location, and shows good tracking accuracy and robustness on related data sets. SiamDW [85] studied how to improve the robustness and accuracy of the twin network by using deeper and wider convolution neural networks. The author has analyzed the reasons why the direct use of deeper networks can not improve the performance of the algorithm: (1) increasing the perception of neurons will also reduce the discrimination of features and the location accuracy of objects; (2) the filling operations in convolution network will cause location inaccuracies. To address the aforementioned problems, SiamDW proposed a new residual module to eliminate the negative impact of the filling operation and further built a new network structure to control the size of the receptive field and the network step size. The module is applied to

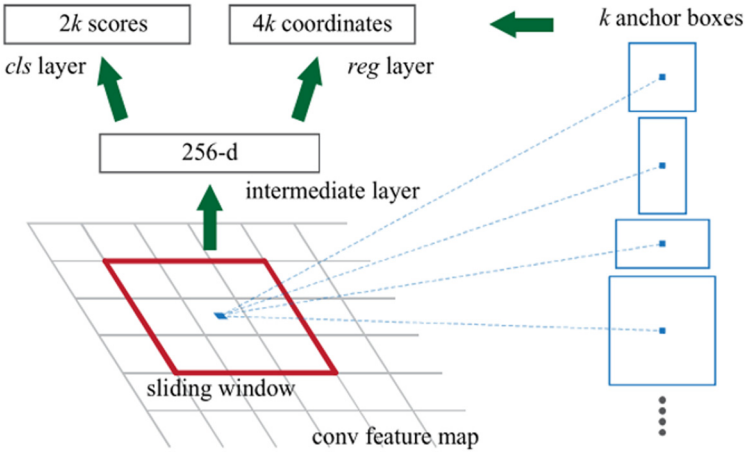

Figure 13: Region proposal network.

SiameseFC and SiameseRPN algorithms to obtain better tracking results and real-time running speed. On the basis of the aforementioned two studies, we can see that the application of a deeper backbone network for feature extraction in object tracking can further exert the effectiveness of the deep learning method in object tracking.

RASNet [86] introduced the attention mechanism into the twin network-based object tracking problem and proposed three models: general attention mechanism, residual attention mechanism, and channel attention mechanism, in which the general attention model is used to train the typical characteristics of different samples, the residual attention mechanism is used to capture individual features such as the shape and appearance of the object, and the channel attention mechanism is used to model the differences of feature channels. To select more effective features, SiameAttn [87] proposed a deformable twin attention network SiamAttn (Deformable Siamese Attention Networks), which improves the feature learning ability of twin network trackers. As shown in Figure 14, SiameAttn includes two parts: the deformable self-attention mechanism and the mutual attention mechanism. The self-attention mechanism can learn powerful context information through spatial attention and channel attention and selectively enhance the interdependence between channel features; the mutual attention mechanism can effectively aggregate and communicate rich information between templates and search areas. This attention mechanism provides an adaptive template feature implicit update method for the tracker. In [88], a pixel-guided spatial attention module and a channel-guided channel attention module are proposed to highlight the corner area for corner detection, and the attraction of the attention mechanism module improves the accuracy of corner detection. As a result, the accuracy of object location and rectangle detection is improved. The loss function for SiamAttn is as follows:

$$L = L_{\text{rpn-cls}} + \lambda_1 L_{\text{rpn-reg}} + \lambda_2 L_{\text{refine-box}} + \lambda_3 L_{\text{refine-mask}}, \tag{19}$$

where $L_{\text{rpn-cls}}$ and $L_{\text{rpn-reg}}$ correspond to the anchor classification loss and regression loss of the Siamese RPN stage, respectively. The latter two terms correspond to the prediction loss for the bounding box and mask in the region correction stage, respectively.

The aforementioned methods mainly focus on single-object tracking. In most twin network trackers in the past, their object template features are not updated in the tracking process, and the features of the object and the search area are independent in the calculation process. SiamAttn [89] is a better solution. It is a method to update template features adaptively and implicitly. At the same time, a region correction module is designed to modify the prediction results and finally get more accurate tracking results. There are generally multiple fish

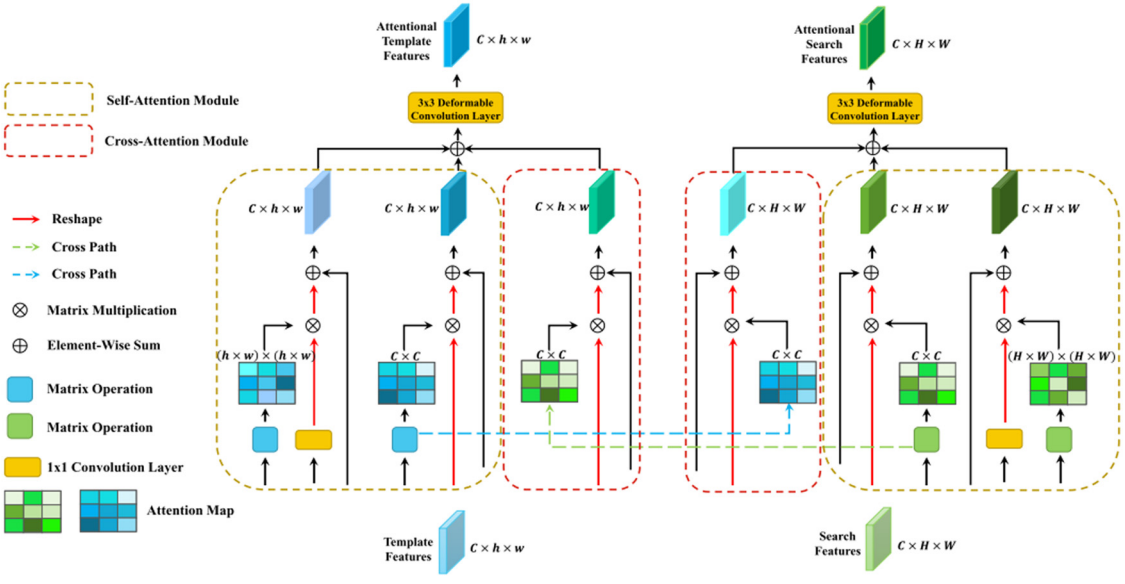

Figure 14: Deformable Siamese attention.

objects in the "ecological fish pond." Therefore, to achieve water quality monitoring, tracking and analyzing the activity behavior of multifish objects (schools of fish) in "ecological fish ponds" are often necessary.

A very important problem in multiobject tracking (MOT) [90] is realizing the effective association of different objects in adjacent frames. For example, in Kalman filter and particle filter tracking, it is necessary to determine the state of each object in adjacent frames. That is, the corresponding object correlation analysis is needed before these filters are applied. A simple method is to use the nearest neighbor method to realize the correlation of particles corresponding to different objects, but if the distance between different objects is relatively close, it is also easy to cause association failure so that the particle filter can not converge in time, resulting in tracking failure. MOT algorithms can be divided into two categories: filter-based and deep learning based. Recently, with the development of deep learning, the efficiency of MOT in practical applications has been dramatically improved. Deep learning-based MOT models include the network model of joint detection and tracking (JDT) and the network model based on graph convolution neural networks.

In a traditional tracking method, the detection and tracking are carried out separately, and only the number of high levels is combined in establishing a tracking relationship, which will lose the appearance information of the image. Thus, a feature extractor with a large amount of computation is needed. JDT [91] integrate some functional modules to a certain extent based on monitoring and tracking, reducing the algorithm's complexity, and increasing the coupling between functional modules. The function lies in (1) joint object detection and association learning, integrating tracking into the process of object detection, and taking the tracking results of the previous frame as input, it is more helpful to deal with occlusion and interruption; (2) using depth features to strengthen multiobject tracking, depth features instead of traditional manual features, and (3) fusion of single-object tracking algorithm.

To simplify the algorithm and improve the tracking performance, Peng et al. [92] proposed a unified frame, i.e., integrating object detection, feature extraction, and data association into an end-to-end model. The online tracking method chain tracker (CTracker) takes adjacent frame pairs as input and performs JDT in a single regression model. The regression model simultaneously regresses the paired bounding boxes of objects that appear in two adjacent frames. Through multitask learning, single-object tracking and metric learning are integrated into a unified triple network, improving the computational efficiency, reducing the memory requirement, and simplifying the training process.

As shown in Figure 15, the CTracker network structure mainly refers to Resnet50 [93] and FPN [94] to construct five scale-level multiscale feature representations, and the CTracker network structure represents them as $\{P_2, P_3, P_4, P_5, P_6\}$. With regards to the MOT of multiple fish objects in the water environment, in the past, it adopts a framework of the first detection and then tracking, usually including three modules: object detection, feature extraction, and object correlation. Because the three modules are independent, there are some problems in fish object tracking, such as high time overhead and the inability to carry out global

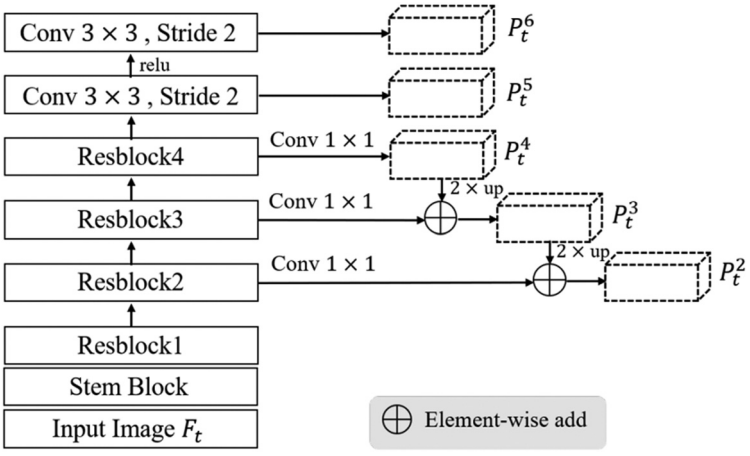

Figure 15: CTracker overall network structure.

optimization. By CTracker, the three modules can be perfectly integrated into one network, which dramatically improves the accuracy and reduces the tracking time.

In the literature [95], a graph neural network is introduced, and a joint feature extractor is proposed to learn appearance and motion features simultaneously from 2D and 3D space. At the same time, a new feature interaction mechanism is proposed. The features of each object are not obtained independently to solve the problem that the features of similar objects are easy to confuse in data association. Figure 16 shows that the previous work used a 2D or 3D feature extractor to obtain features independently from each object. Figure 17 shows a joint 2D and 3D feature extraction mechanism and feature interaction mechanism proposed in the literature [95] to improve differentiated feature learning for data association in MOT.

He et al. [96] pointed out that most graph-based optimization methods use separate neural networks to extract features, which leads to the inconsistency of training and reasoning. Therefore, a new learnable graph matching method GMTracker is proposed for multiobject tracking. The relationship between Tracklet and intra-detection is modeled as a general undirected graph, and then the association problem is transformed into a general graph matching problem between the trajectory graph and detection graph.

2.3 Identification of fish activity behaviors

2.3.1 Extraction of behavioral characteristics of fish activities

Under the stimulation of environmental change, the change in fish behavior shows a certain regularity, which can be described by the ecological stage pressure model, namely, the stepwise stress model (SSM) [97]. When the water environment changes, especially when the pollutant concentration changes, the change in fish activity behavior is affected by time and pollutant concentration and generally go through four stages: no effect, regulation, adaptation, and toxic effect [98]. The phased characteristics of fish behavior changes are the basis of biological water quality monitoring.

The visual characteristics of fish (activity behavior) involved in fish behavior recognition can be divided into single-objective behavior characteristics, multi-objective behavior characteristics, and group object behavior characteristics according to the number of fish objects monitored. The characteristics of single fish object behavior include conventional motion indicators, such as fish object coordinates, swimming speed, angle, acceleration, fish trajectory, the curvature of trajectory curve, overall activity, and so on, and some special, abnormal motion behaviors, i.e., tail swing frequency, balance ability, steering, reverse swimming, belly turning, jumping, rest, sinking, and even death. The object behavior characteristics of multi-fish also include the routine movement index of fish and the special, abnormal behavior index of fish, but the object behavior characteristics of multi-fish record the related behavior characteristics of multiple fish objects. The

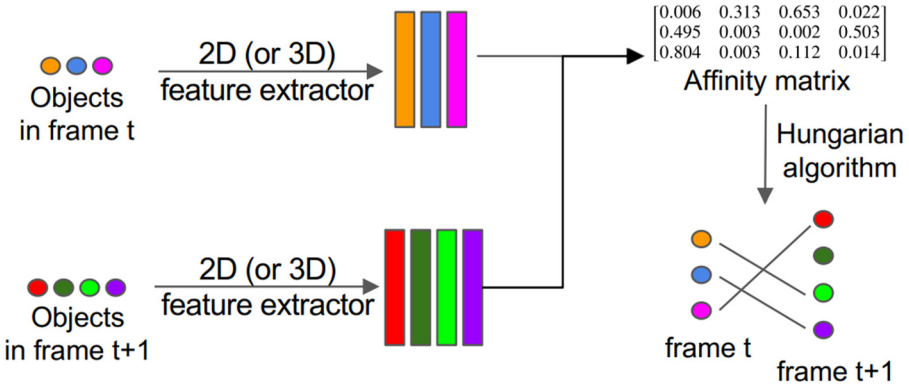

Figure 16: Uses a 2D or 3D feature extractor.

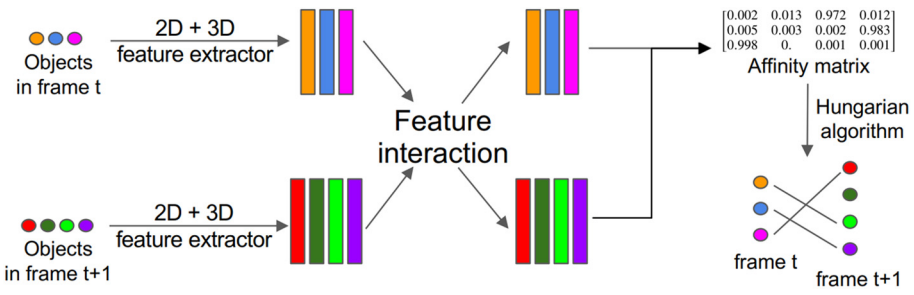

Figure 17: Joint 2D and 3D feature extractor and feature interaction mechanism.

characteristics of fish shoal object behavior can be divided into fish shoal routine characteristic data and fish shoal social behavior. The conventional characteristic data of fish shoals include the change of the center of gravity of fish shoal, the distribution of fish shoal, the area occupied by fish shoal (marking the degree of dispersion and concentration of fish shoal), and the social behavior of fish shoal, including fish clustering, rear-end collision, aggressive behavior, feeding behavior, and so on.

The feature extraction method based on deep learning automatically learns trainable features from input video data. According to the different designs of neural network structure, the feature extraction algorithms based on deep learning can be divided into feature extraction based on a double-flow convolution network, feature extraction based on a multi-flow convolution network, feature extraction based on a 3D convolution network, and feature extraction based on long-term memory network.

The double-stream model uses two convolution networks to model temporal information and spatial information respectively, alleviating the lack of dynamic features faced by the single-stream recognition network based on RGB data to some extent, but the dynamic features represented by an optical flow can only represent part of the time information, and it is an urgent problem to extract optical flow from video accurately and effectively. Although the motion recognition method based on a multistream convolution network can effectively capture the spatial features of the image and more comprehensively compensate for the lack of time information in a single video frame, the more types of input modes mean the more parameters needed to be trained in the depth model, which greatly reduces the effectiveness of the model. In addition, the increase in input patterns also means higher requirements for the design of feature fusion modules, which increases the complexity of multistream models.

The general method of feature extraction algorithm based on a 3D convolution network involves taking a small number of continuous video frames stacked spatio-temporal cube as model input and then adaptively learning the spatio-temporal representation of video information through hierarchical training mechanism under the supervision of a given action category label. The 3D convolution network captures differentiated video features from video data directly in both spatio-temporal dimensions, and there is no need to design a spatio-temporal feature fusion module that can effectively deal with the fusion of short-term spatio-temporal information.

To capture the action information over a long period, Varol et al. [99] designed a neural network with long-term temporal convolution (LTC) to obtain longer time features by convolution of more video frames at the same time, but the number of parameters is huge, so it is very difficult to train. A timeception module is proposed in [100]. As shown in Figure 18, the temporal-only convolution kernel ($T \times 1 \times 1 \times 1$) is constructed by deep separable convolution. The video is modeled for long time series by stacking multiple timeception modules. Still, this module chooses to exchange timing information at the expense of spatial information, which may lead to the compression or even loss of context semantic information in the process of long-term timing modeling. The model based on long short-term memory (LSTM) network specifically refers to adding LSTM or corresponding variant structure to the end of CNN. Thanks to its strong sequence modeling ability, this method has gradually become a research hotspot in the field of action recognition.

(1) Double-stream convolution networks

RGB data have rich appearance information, which can directly show the shape of fish and the appearance of objects. It compensates for the lack of apparent features in traditional methods and has been widely

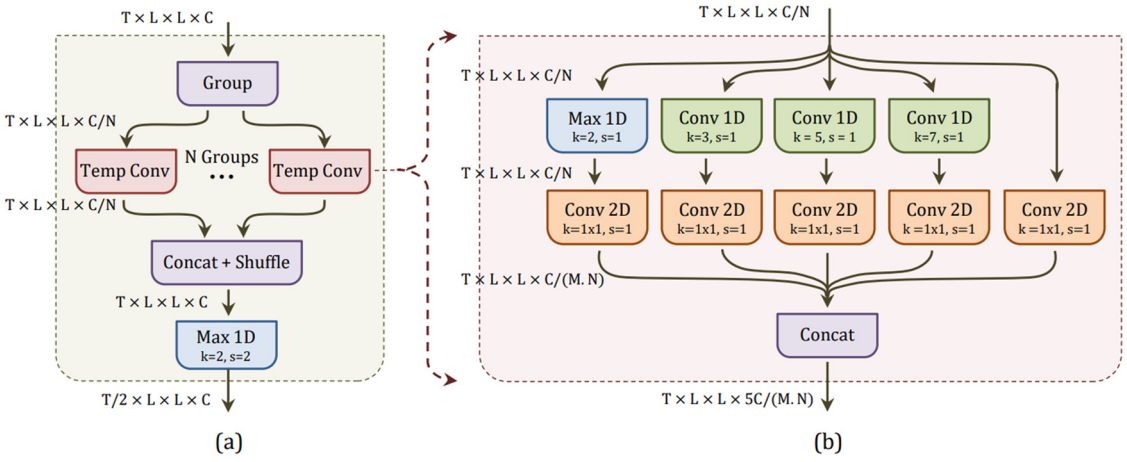

Figure 18: Timeception layer. (a) Timeception layer and (b) temporal conv module.

used in motion recognition. However, only using a single video frame as model input can represent the spatial information of a single time node. To mine the time information of the video stream, Karpathy et al. [101] proposed three methods, i.e., methods based on late fusion, early fusion, and slow fusion of RGB video frame. However, the recognition effect is still far from that of the traditional manual feature model. To address the problem that the lack of dynamic feature information limits the recognition performance of the RGB single-mode input model, the research has gradually entered into the exploration of the multimode input structure. Sanchez-Riera et al. [102] took both RGB and RGB-D as model inputs and discussed the impact of different data type fusion methods on the performance of the classifier. Zhu et al. [103] combined RGB and depth data for the segmentation and recognition of continuous gestures. In human activity recognition, since the bone information can be estimated quickly and accurately by RGB-D data, some researchers combined the RGB-D with human bone information to represent actionv [104], but RGB-D data collection is complicated, having additional noise information that interferes with the recognition effect.

The combination of optical flow information and RGB data is the most widely used dual-stream input mode. Goodale and Milner [105] proposed a well-known dual-stream assumption that visual information can be processed into two paths: a ventral stream for shape perception and a dorsal stream for motion perception. Derived from this clue, Simonyan and Zisserman [106] first applied the two-stream approach to the field of action recognition by dividing the video information into two parts: spatial information and temporal information. The basic idea is first to compute the dense optical flow between adjacent video frames and then to input the RGB video frames and optical flow features into the two-stream structure separately. The prediction accuracy exceeds that of traditional action recognition methods, verifying that the effect of temporal information is compensated by the optical flow signal and demonstrating the possibility of adopting a deep learning feature extraction method instead of traditional hand-designed features.

To make a better fusion of the two-way features of the dual-stream network, Feichtenhofer et al. [107] adopted the residual connectivity to establish an information connection between the spatiotemporal convolutional streams to facilitate their information interaction. In addition, some researchers [108] constructed temporal segment networks (TSN) based on the idea of remote temporal structure modeling and proposed a sparse sampling strategy to sparsely sample a series of video clips from a given video. Unlike the original dual-stream network structure, they use two input modes of the RGB data and the optical flow information as the input of different network streams, respectively.

To capture long-time dynamic information, some researchers [109] combine a dual-stream network with an LSTM network to capture the spatiotemporal global information. Xue [110] utilize a segmented sampling strategy for sampling and construct a spatiotemporal heterogeneous dual-stream network to achieve long-term temporal modeling. Miao [111] combined TSN networks with a temporal pyramidal pooling approach to model the dependencies between long-range video frames by constructing multiscale temporal features.

Zhi-qiang [112] built a deep residual LSTM network and combined it with a dual-stream network to extract global information. Unlike the traditional spatiotemporal dual-stream network that uses parallel arrangement, Biao [113] utilized serial connections in combination with the spatiotemporal stream network to save hardware resources. To avoid manual computation of optical flow features, Dosovitskiy et al. [114] proposed a multitask learning model, ActionFlowNet, which trained two convolutional flow networks separately from the original pixel points to perform action recognition while the model automatically estimates optical flow values. The model is trained on a dataset labeled with real optical flow values so that it adaptively learns optical flow information between consecutive video frames, thus reducing the computational effort while extracting motion information.

(2) Multi-stream convolution networks

To improve the model description ability, some researchers enrich the input patterns of the model and extend the two-flow network model to a three-stream network or even a multistream network. Different input patterns are processed separately and then fused. It is used for the subsequent classification and identification to obtain more discriminative fish action characterization.

Goodale and Milner [105] proposed a warped optical flow as an additional input mode based on optical flow and RGB input mode, and these three modes were inputs to the TSN network separately to explore the effect of multiple input modes on the discriminative power of the model. Wang and Deng [115] mapped the skeleton sequence features into RGB image features according to different orientations. They used them as the input of the three-flow network to achieve the information interaction between multiple features. Wang and Deng [116] proposed a three-stream convolutional network with the stacked motion stacked difference image (MSDI) on top of the optical stream and RGB data to constitute a three-mode input. The MSDI builds temporal features used to characterize the global action by fusing each local action feature to separate the three data forms feature learning through a CNN with the same setup (i.e., a sequential stack of five convolutional layers with two fully connected layers) to capture spatial epistemic information, local temporal features, and global temporal representation.

Bilen et al. [117] proposed the concept of dynamic images to encode RGB images and optical flow information using sequential pooling and approximate sequential pooling, which were trained to obtain RGB dynamic image flow network and dynamic optical flow network combined with the original RGB network and optical flow network to form a four-stream network structure. Wen-han [118] combined the advantages of multiple feature types to enhance model recognition using a multimodal input form of RGB data, optical flow, and depth information. To improve the learning ability of the model with limited training samples, Chenarlogh and Razzazi [119] extracted the optical flow and gradient information of the original video frames in both horizontal and vertical directions and fed them into the multistream convolutional network channel separately to increase the number of training samples.

Apart from the literature mentioned earlier, which uses the same design of convolutional flow approach for different input modes, Gu et al. [120] input the depth MHI and skeleton data into ResNet101 and ST-GCN,1 respectively, to extract the corresponding global motion and local motion information and combines RGB images to form a tri-modal input, considering the dependency between object and action. Sun et al. [121] defined the optical flow guided feature (OFF) from the feature-level optical flow orthogonal space by directly calculating the spatio-temporal gradient of the depth feature map by three sub-networks: feature generation sub-network, OFF sub-network, and classification sub-network.

(3) 3D convolutional networks

(1) Standard 3D convolution network models

Ji et al. [122] extend the two-dimensional convolutional networks to three-dimensional space while extracting video features from spatiotemporal dimensions. On this basis, various 3DCNN variations have been proposed, such as C3D [123], I3D [124], Res3D [125], and so on. Thanks to the development of GPU, 3DCNN-based methods have gradually become the mainstream method in video action recognition. Haiyang et al. [126] used multiview learning to extract multiple local shape descriptors. Then they combined them with 3DCNN to fuse multiple view descriptors to improve the description ability of classification features. Ming-li [127] added a buffer before the C3D network, enabling the model to perform real-time classification prediction while the video stream is being input. To present these methods more precisely, we have introduced

mathematical formulas to describe the working principles of 3D CNNs. Specifically, a 3D CNN model can be represented as $f(x; W, b)$, where x is the input video frame, and W and b, respectively, represent the network's weights and biases. The convolutional layer extracts features through $W*x + b$, where $*$ denotes the convolution operation. In this way, we can understand in more detail how 3D CNNs capture behavioral features across temporal and spatial dimensions, thereby enhancing the accuracy of fish behavior classification. It is calculated as follows:

$$s(i,j) = (X*W)(i,j) + b = \sum_{k=1}^{n_i n} (X_k * W_k)(i,j) + b, \tag{20}$$

where n_i is the number of input matrices, X_k represents the k input matrix, and W_k represents the k subconvolution kernel matrix of the convolution kernel. $s(i,j)$ is the value of the corresponding position element of the output matrix corresponding to the convolution kernel W.

Aiming at solving the problem that the shallow layers of the C3D network are not conducive to learning the depth characteristics, some researchers [128] combined the idea of residual error with the deep C3D network and introduced short-circuit connection into it, thus avoiding the defect that the deep C3D network that will lead to the degradation of its learning ability. However, compared with 2DCNN, the parameters of 3DCNN are doubled, which makes training its corresponding model on small datasets easy to lead to an overfitting effect. Therefore, Yang et al. [129] applied the dense connection method to 3DCNN and combined the spatial pyramid pooling method to reduce the difficulty of model training. In addition, researchers use the transfer learning method to fine-tune small datasets after pre-training the model with large public data sets.

Inspired by the fact that 2DCNNs greatly facilitate the acquisition of generic feature representations after pre-training on the ImageNet [61] dataset, Hara et al. [130] investigated whether the huge number of parameters of 3DCNNs causes overfitting problems during training. For the first time, they proposed to train multiple models of 3DCNNs from scratch on the kinetics [131] dataset (ResNet [93], Preactivation ResNet [132], Wide ResNet [133], ResNeXt [134], and DenseNet [135]) and investigated the deep structure that can be trained on this dataset without causing overfitting by the network structure from shallow (18 layers) to deep (200 layers). The upper limit of the number of layers demonstrated that training a deep 3DCNN using the kinetics dataset would retrace the successful history of 2DCNN and ImageNet. Pretraining mitigates the overfitting effect of commonly used small datasets and is an effective way to initialize and accelerate the convergence of the model. However, pretraining operations on large video datasets require expensive time costs. Hence, Huang et al. [136] utilized a 2DCNN model after pretraining on the image dataset ImageNet to construct a 3DCNN, which reconstructs 3D filters by stacking 2D convolution kernels of the same size along the time dimension and mimics the 2D convolution on the video stream by performing simultaneous 2D convolution on the frame sequence. The 3D convolution operation avoids the tedious pretraining process in large video datasets. However, video data contains a lot of useless information, and treating all features equally would lead to a large number of unnecessary features in the feature extraction process.

Thompson and Parasuraman [137] show that humans ignore the entire content when observing their surroundings but focus their attention on the salient regions of the environment. Inspired by this, some scholars have introduced the attention mechanism in feature extraction to help the model allocate more attention resources to the object regions during feature learning, suppressing redundant information and quickly filtering out critical information in complex video content. For instance, Woo et al. [138] proposed a convolutional block attention module (CBAM), which constructed a hierarchical dual attention mechanism based on a two-dimensional residual network structure that adds channel attention and spatial attention to each residual block sequentially. However, this network structure ignores the temporal information, which is crucial to the action recognition task. The overall formula is as follows:

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F'. \end{aligned} \tag{21}$$

The two formulas are channel attention and spatial attention operations. Here, $\otimes$ represents element-wise multiplication, that is, the multiplication of corresponding elements.

Cai and Hu [139] extend the two-dimensional residual attention structure to three-dimensional space and propose a 3D residual attention network (3DRANs) that sequentially infers the channel attention mapping and spatial attention mapping of the extracted features in each 3D residual block according to the channel and spatial attention mechanism submodules. Thus, the middle layer convolutional features can sequentially learn key cues in the channel and spatial domains. Liu et al. [140] constructed a dual-stream residual spatial-temporal attention network (R-STAN) based on residual networks, which branches from a stack of integrated spatio-temporal attentional residual blocks (R-STAB), giving R-STAN the ability to generate attentional features along the temporal and spatial dimensions with different discriminative power, which dramatically reduces redundant information.

Li et al. [141] proposed a spatio-temporal deformable 3D ConvNets with attention (STDA) module with an attention mechanism to overcome the small perceptual field of the 3 × 3 × 3 convolution kernel in the spatio-temporal domain that does not take into account the global information in the whole feature map and the whole frame sequence. It simultaneously performs inter-frame deformation operations and intra-frame deformation operations along the spatio-temporal dimension and autonomously learns the offsets in the spatio-temporal dimension to adaptively fit the immediate complex actions occurring in the video, thus producing a more discriminative video representation that compensates for the global information deficit problem and better captures long-term dependencies in the spatio-temporal domain and irregular motion information in the video.

Models based on the standard 3D convolutional structure have inherent advantages in extracting local Spatiotemporal fusion features due to their inherent structure, but at the same time, there are many limitations. For instance, the number of model parameters required to train the model based on the standard 3D convolutional structure is huge, which increases the computational complexity and storage of the model and is not conducive to the iterative optimization of the model, making it difficult for the model to converge to the optimal solution quickly.

(2) Variant 3D convolutional structure

The aim is to reduce the training parameters of the model, improve the computational speed, and reduce memory consumption. Various structural deformations based on standard 3D convolutional networks have been proposed. In an early related study, researchers approximated a common 3D convolutional layer with a convolutional kernel size of 3 × 3 × 3 as three cascaded convolutional layers with their filter sizes of 1 × 3 × 1, 1 × 1 × 3, and 3 × 1 × 1, respectively, to improve the effectiveness of the model [122], but this approach is equivalent to deepening the model by a factor of three, which leads to difficulty in training the model. To solve the aforementioned problem, an asymmetric 3D convolution was proposed in the literature [142] to approximate the traditional 3D convolution and improve the computational complexity of the traditional 3DCNN by approximating two layers of convolution kernel size 3 × 3 × 3 convolution layers into one layer of asymmetric 3D convolution layers with convolution kernel size 1 × 5 × 1, 1 × 1 × 5 with 3 × 1 × 1, and then stacking several different micro meshes to construct an asymmetric. The 3D convolutional deep model improves the feature learning capability of the asymmetric 3D convolutional layer without increasing the computational cost.

In addition, factorized spatio-temporal conventional networks [143] (FstCN) with pseudo-3D networks [144] (P3D) have also been proposed to alleviate the computational complexity of 3D convolutional networks. A spatio-temporal decomposition method based on 3D residual networks was presented in the literature [145], which decoupled the standard 3D convolutional operation into cascaded 2D spatial convolution and 1D temporal convolution, achieving good results with a more compact structure. Subsequently, a channel-separated conventional network (CSN) based on grouped convolution from a new perspective of channel separation was proposed, decomposing the standard 3D convolution into a channel interaction layer (filter size of 1 × 1 × 1) and a local spatio-temporal information interaction layer (filter size of 3 × 3 × 3). The former enhances the information exchange between different channels by reducing or increasing the channel dimension, while the latter utilizes the idea of depth-separable convolution, which discards the information transfer between channels and focuses on the interaction between local spatio-temporal information, hence reducing the computational effort of the model [146]. Each residual of P3D can be calculated by the following formula:

$$(I + F)\cdot x_t = x_t + F \cdot x_t \coloneqq x_t + F(x_t) = x_{t+1}, \tag{22}$$

where $F \cdot x_t$ represents the result of executing the residual function F on x_t.

Ji-yuan [147] connect two-dimensional spatial convolutional kernels with one-dimensional temporal convolutional kernels in three different ways and then construct pseudo-3D residual networks by concatenating the three networks to reduce the model training difficulty. Li-jun [148] proposed a Fake-3D module using tensor low-rank decomposition theory. The C3D network was selected as its base architecture and combined with the idea of residual connectivity to reduce the parameter size of the C3D model and improve the recognition performance. Xie et al. [149] demonstrated the performance gain of I3D over I2D while questioning the redundancy of the full 3D convolutional module and then proposed a lightweight model S3D-G that uses 2DCNN to extract spatial features in the underlying network and constructs separated 3D spatio-temporal convolutions using depth-separable convolutions in the top-level 3DCNN module. However, the aforementioned model is limited by the temporal dimension of the input data, which can only characterize local spatio-temporal information.

The computational complexity and memory consumption limit the length of the input video data. Therefore, the feature extraction model based on the 3D convolutional network can only characterize the action in the short-term time range, and it is not easy to handle the video data information for an extended period, thus affecting the model performance. Whether the long-term spatio-temporal sequence information can be adequately analyzed is critical to improve the accuracy of video action classification.

(4) LSTM network

(1) Standard LSTM model

The introduction of LSTM liberates the input length and better captures the dependencies between long-term video data. Yue-Hei Ng et al. [150] used the features of each video frame by a CNN. Then the features of the CNN are fed into the LSTM sequentially in the temporal order in the video to obtain temporal correlation features to compensate for the temporal dynamics lacked by the CNN. In addition to exploring the correlation between video frames, LSTM can also be used to model the semantic relationships between different video clips. LSTM is introduced on the I3D model that has been pre-trained on the kinetics dataset, where the I3D network is used to extract the local spatio-temporal features of the input video clips at different moments and model the temporal dependencies between different clips to achieve advanced temporal features with local spatio-temporal fusion features.

Analogous to the aforementioned approach, Zhang et al. [151] feed both video frames and optical streams into a 3DCNN network with a feature fusion module to obtain fusion features of both modes and finally use a deep LSTM to model the sequential fusion features temporally to emphasize the ability of the model to characterize coherent actions. Ouyang et al. [152] introduced multitask learning based on 3DCNN with LSTM networks to model the temporal relationships between video frames while emphasizing the rich information in the relevant tasks. To solve the overfitting problem caused by deepening the number of layers in the LSTM network, Yu et al. [153] introduced residual connections in the recurrent network to construct a pseudo-recursive residual neural network for extracting Spatiotemporal features. The various gates of LSTM are calculated as follows:

$$\begin{aligned} I_t &= \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i), \\ F_t &= \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f), \\ O_t &= \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o), \end{aligned} \tag{23}$$

where $W_{xi}, W_{xf}, W_{xo} \in \mathbb{R}^{d\times h}$, and $W_{hi}, W_{hf}, W_{ho} \in \mathbb{R}^{h\times h}$ are the weight parameters and $b_c \in \mathbb{R}^{1\times h}$ is the bias parameter.

LSTM can be used as an encoding-decoding network in addition to time series modeling. A motion map network based on 3DCNN is proposed in the literature [154], where the motion information contained in the whole video is integrated into the motion map by iterative means, and then the LSTM encoding network encodes the extracted feature map into the corresponding hidden activation form and then reconstructs the approximate output through the decoding network in the input layer to explore the hidden patterns among the video sequences.

Despite its powerful sequence modeling capability, LSTM still suffers from various shortcomings. The standard LSTM only considers sequence information in a single direction and uses vectorized one-dimensional data as model input, which is prone to the problem of loss of crucial information. Thus the combination of variant CNN and LSTM structures has also started to be favored by researchers.

(2) Variant LSTM variant structure

The one-way LSTM considers only past sequence information to classify and recognize actions with high similarity (e.g., running vs triple jump). Researchers have used Bi-LSTM networks to model temporal information [155,156]. The bi-directional LSTM consists of two standard LSTM networks stacked in different directions with two pathways, forward and backward. The features extracted by the convolutional network are fed into the subsequent deep Bi-LSTM network for dependency exploration, which can help the model effectively extract contextual semantic information about the past and future of the action occurrence and thus distinguish similar movements more effectively. Khaled et al. [157] combine a dual-stream 3DCNN network with a bidirectional LSTM to model long-term dependencies in both the front and back directions of the video stream. However, vectorizing the convolutional layer features and feeding them directly into the LSTM can disrupt the inherent spatial location correlation between feature planes, thus interfering with the recognition effect.

To preserve the spatial topology of the feature maps, Zhu et al. [103] combine 3DCNN and ConvLSTM and feed the 2D features captured by these two networks into 2DCNN for learning deeper features to achieve action recognition of arbitrarily long video sequences. Jin et al. [158] combined multilayer dense bidirectional ConvLSTM after generating feature maps of corresponding sampled frames with rich spatio-temporal information, which are then fed into the 3DDenseNet network together with the original sampled frames, preserving the spatial topology of the convolutional layer feature plane while considering the correlation of different video clips.

Wang et al. [159] designed a lightweight architecture that only uses RGB image data. By using ConvLSTM and FC-LSTM to model the time series information in different visual perception layers, it is beneficial to integrate local spatial details and global semantic features and enhance the comprehensive representation ability of the model. However, the ConvLSTM structure uses its internal convolution structure to explicitly code the relationship between the input space position and the long-term time dependence in the input state and state-to-state transition, and its parameters are large. Hence, it is difficult to get sufficient training on small data sets, which is prone to lead to over-fitting of the model.

To address the aforementioned issues, a deep self-coding network combining 3DCNN and ConvGRU structures is proposed in the literature [160] for learning spatio-temporal dimensional features of videos with comparable performance to ConvLSTM, but it has fewer parameters and is easier to train. Zhu et al. [161] replace the traditional convolutional structure in ConvLSTM with the help of computational decomposition as well as the idea of sparse connectivity to obtain redundancy analysis using deep separable convolution, grouped convolution, and mixed convolution.

The action recognition algorithm based on the combination of CNN and LSTM network can maximize the advantages of both models; correlate the apparent information, motion information, and long and short-term spatio-temporal information in an irregular period; and provide a more comprehensive spatio-temporal representation for the subsequent classification and discrimination stage. However, the aforementioned model still requires a large amount of video data for model training, which has a high demand on the data set used for training. The time cost of the training process is more extended, which increases the difficulty of model training.

In summary, when feature extraction is performed on fish objects, the traditional method requires manual participation and has problems of low efficiency and high cost. Deep learning-based feature extraction methods can automatically learn trainable features through network models, which significantly improves the model efficiency and reduces the computational cost. For example, the LSTM model used for feature extraction can handle fish image sequences and captures well the dependencies in long-term video data. The combination of advanced temporal features and local spatio-temporal fusion features is achieved.

2.3.2 Fine-grained behavioral recognition of fish activity

Fine-grained behavior identification of fish activities refers to the accurate identification of different fine-grained fish behavior states based on fish activity behavior characteristics, such as swimming, foraging, rest, and so on, to evaluate and monitor the water quality more accurately. Activity behavior recognition of moving objects is a fundamental problem in computer vision. In recent years, researchers have conducted considerable research on human behavior recognition to achieve safety monitoring, and many research results have

been achieved. Activity behavior recognition involves activity behavior modeling and feature representation, activity behavior inference, and recognition.

However, researchers do not have a more general model for the behavioral representation of motion objects, and most of the studies are based on a specific application scene by some corresponding behavioral representation methods. Xin et al. [162] extracted the motion vector field of fish objects based on the optical flow method. Then they used statistical methods to count the two behavioral features of the motion of the fish object, namely, the velocity and turning angle, to calculate the corresponding joint histogram and joint probability density distribution features and then applied the normalized mutual information (NMI) and local distance anomaly factor to detect the abnormal behavior of fish objects.

At present, hierarchical structural models are usually used in the behavioral representation of various movement object activities. For example, in human behavior recognition, Moeslund et al. [163] reviewed the progress of human behavior analysis and recognition and described the action/motion primitives, action, and activities by hierarchical representation models.

After obtaining the object activity characteristics, nonlinear modeling methods, such as neural networks and support vector machine methods can be used to classify and identify the activity behavior of surveillance objects to detect abnormal behaviors of motion objects. For example, Shu-hong et al. [27] extracted several characteristic parameters of velocity, acceleration, curvature, and neighborhood characteristics of red carp in normal and abnormal water quality by introducing fish motion trajectory curvature and neighborhood characteristics parameters to establish a corresponding recognition model.

To achieve an effective prediction of object activity behavior trends, researchers have now widely used graph model-based inference methods to analyze the activity behavior of moving objects. Commonly used methods include probabilistic models, such as hidden Markov model [164], dynamic Bayesian networks [165], and conditional random fields [166]. Some researchers use rule-based methods such as decision trees [167] for activity behavior analysis.

Fine-grained image recognition is a subfield that lies between semantic-level and instance-level tasks. Fine-grained image subcategories have only subtle local differences, while the appearance varies significantly between the same categories and is susceptible to uncertainties such as pose and occlusion. Therefore, fine-grained image recognition is highly challenging.

In traditional fine-grained image recognition tasks, background noise in the image is eliminated by annotating boxes to locate the object, and the local area feature extraction is achieved by position annotation. These algorithms rely excessively on manual annotation [168]. Manual annotation information is labor-intensive. The ability to rely on manual annotation methods to feature extraction and feature representation is weak and has certain limitations. CNN-based fine-grained image recognition methods [169,170] have become more and more mature in recent years. Their extracted features possess more powerful representation capabilities and usually achieve good results in fine-grained image recognition tasks. However, the essential parts are too fine to obtain all the critical information by traditional CNNs. Researchers have started to work on improvements within the framework to locate the key parts further and enrich the feature representation. Some scholars believe that the CNN-based fine-grained image recognition methods still have loopholes in the mastery of global information, so their visual transformer is introduced to fine-grained image recognition. It proves that in the field of fine-grained visual recognition, although learning local area features plays a crucial role, the complement of global information will further improve the accuracy of recognition.

(1) CNN-based fine-grained behavior recognition

With the continuous improvement of deep learning techniques, CNNs have been rapidly developed and applied in computer vision [171], natural language processing [172], and so on. In fine-grained image recognition, commonly used CNN structures are VGGNet [173] and ResNet [93]. There are many types of variant CNNs, such as inflated convolution, which increases the perceptual field and maintains the width and height of the input features. The depthwise separable convolution first performs convolution operation on each channel

and then performs point-by-point convolution, which has the advantages of fewer parameters and low operation cost. The general formula for ResNet is as follows:

$$\begin{aligned} y_l &= h(x_l) + F(x_l, W_l) \\ x_{l+1} &= f(y_l), \end{aligned} \tag{24}$$

where h function is a direct mapping, and f function is an activation function, generally using ReLU.

Fine-grained behavior recognition based on the CNN aims to identify and distinguish subtle differences in fish behavior by in-depth analysis of high-level features of images. This process usually consists of the following key steps:

(a) **Feature extraction**: CNNS automatically learns image feature representations through multiple convolutional and pooling layers. In the context of fine-grained action recognition, the first convolutional layers may capture simple features such as edges and corners. In contrast, deeper convolutional layers can recognize more complex patterns, such as a fin's shape or a fish's swimming posture. These features become more and more abstract as the network goes deeper and can reveal subtle differences in fish behavior.

(b) **Feature integration and classification**: After feature extraction, CNNS usually contains one or more fully connected layers that integrate and use the features learned by the convolutional layers for classification. In fine-grained action recognition, these thoroughly combined layers aim to map the extracted features onto specific action categories.

(c) **Challenges and strategies of fine-grained recognition**: A significant challenge in fine-grained action recognition is dealing with intra-class variation and similarity. Methods such as attention mechanism, data augmentation, and feature fusion can be adopted to overcome this challenge.

Through these steps, the CNN-based fine-grained behavior recognition technique can accurately distinguish and identify subtle differences in fish behavior, which provides a powerful tool for water quality monitoring and fish behavior research.

(2) Fine-grained action recognition with transformer

Most of the fine-grained image recognition methods are based on CNNs. These methods inevitably complicate the recognition channel and produce a lot of redundancy in the local area of location. To address this problem, researchers proposed to complete the fine-grained image recognition task based on transformer. The transformer is a classical natural language processing model proposed by the Google team in 2017. It incorporates the self-attention mechanism and does not use the sequential structure of a recurrent neural network (RNN), allowing the model to be trained in parallelization. Vision transformers (ViT) have made breakthroughs in traditional recognition tasks in recent years. They have also demonstrated their ability to capture global and local features in areas such as object detection [59] and semantic segmentation [174]. Compared to CNNs, transformer's image serialization is an entirely new form. The attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{25}$$

He et al. [175] proposed a framework for fine-grained image recognition based on transformer. The framework pools the original attention weights before the last layer of the transformer into an attention graph to guide the network in selecting the accurate discriminative region image blocks. Specifically, the method uses a self-attentive mechanism to capture the most discriminative regions and image blocks to process the internal relationships between regions, and a contrast loss function to expand the distance between similar subclass feature representations. The network structure of the method is shown in Figure 19. Although this scheme has overlapped the input image blocks to avoid damage to the local neighborhood structure, it is still computationally expensive and has low recognition accuracy on some benchmark datasets.

Although He et al. [175] improved the efficiency of fine-grained image recognition, their method with fixed image block size and deep class token concentrated in the global perceptual field cannot generate multiscale fine-grained recognition features. In response, Zhang et al. [176] proposed a new adaptive attentional

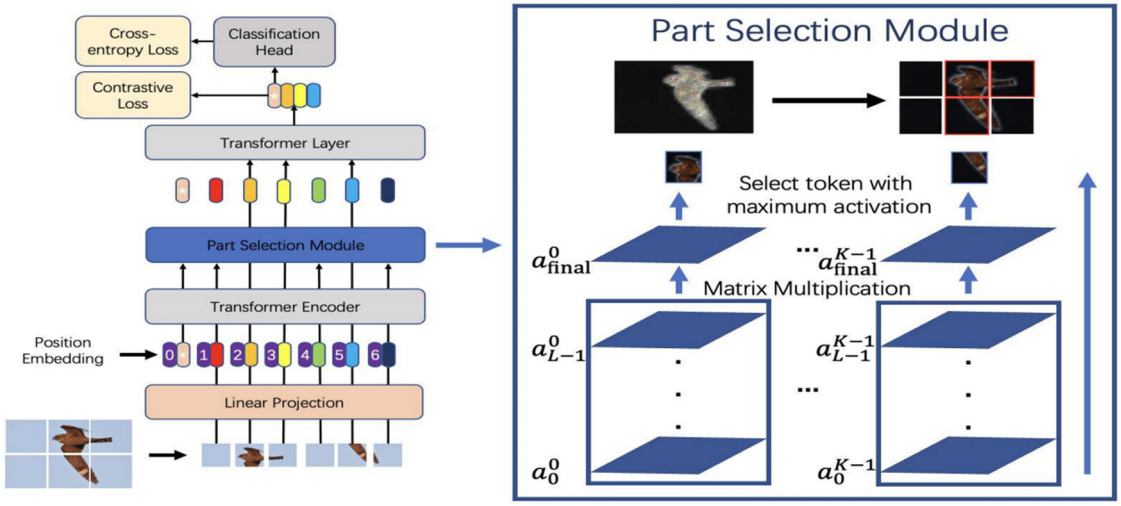

Figure 19: Fine-grained transformer network framework.

multiscale fusion transformer method. In this method, the attention collection module is selected to use attention weights to filter out relatively important input blocks adaptively. The multiscale (global and local) channels are supervised by weight-sharing encoders that can be trained end-to-end.

The identification clues of fine-grained image recognition methods are usually piecemeal, ignoring the details of additional areas and lacking consideration for other relevant image clues. To solve the aforementioned problems, Liu et al. [177] proposed a transformer structure with a peak suppression module and knowledge guidance module. The peak suppression module removes the most distinguishing marks according to the attention weight value, forcing the network to pay attention to the additional neglected information areas to ensure the diversity of fine-grained representation. The knowledge guidance module compares the image representation generated by the peak suppression module with the learnable knowledge embedding to obtain the knowledge response coefficient. In the training process, the scheme updates knowledge embedding and image representation simultaneously, embeds many identification clues of different images in the same category, and embeds the acquired knowledge into the image representation as a comprehensive representation, thus significantly improving the recognition performance.

Aiming at the problem that the self-attention mechanism weights the information aggregation of all image blocks to the classification token, which makes the deep classification token focus more on the global information and lacks the local and bottom features, Wang et al. [178] proposed a feature fusion vision transformer (FFVT) framework, which collected the important tokens of each transformer layer to complete the local, bottom and middle information. Specifically, through a token selection method, a representative token of each layer is selected as the input of the last layer. The experimental results show that this method improves the accuracy of fine-grained image recognition. Conde and Turgutlu [179] proposed a fine-grained image recognition framework with multistage ViT, which captures outstanding image features from multiple different local regions using an inherent multiheaded self-attentive mechanism. Various attention-guided enhancements are used to enhance the model to learn more distinct discriminative features, thus improving the model's generalization ability. However, the approach still has shortcomings in that it cannot be trained completely end-to-end and needs to be trained in a sequential (multistage) manner; it requires high computational power. The future goal is to make the framework end-to-end trainable.

The transformer-based methods are summarized in Table 1, and it can be seen that the transformer achieves high accuracy in fine-grained image recognition tasks. The method of He et al.'s [175] is highly accurate but less applicable because the input image block size is fixed. The existing drawbacks of the transformer as a newly introduced technique are a large number of parameters and the long computation time. The computation time length can be considered for future explorations.

Table 1: CUB-200-2011 Innovation points and accuracy rates of different methods on the dataset

Method	Innovations	Accuracy/%
He et al. [175]	The first network framework for fine-grained image recognition based on transformer is proposed	91.7
Zhang et al. [176]	By using the attention weight, the discriminant input block is screened adaptively	91.5
Liu et al. [177]	The feature representation is learned by the peak suppression module, and the most discriminating part is punished to obtain detailed information	91.3
Wang et al. [178]	Bring together the key token of each transformer layer to complete local, underlying, and middle-level information	91.6
Conde and Turgutlu [179]	Using a multihead self-attention mechanism to capture different image features from several different local regions	91

3 Industrial applications of intelligent water quality monitoring

Many countries have developed the corresponding water quality standards to effectively carry out the prevention and control of pollution and achieve a safe water supply and water quality criteria (WQC). WQC refers to the maximum dose or concentration of pollutants in the water that does not harm specific objects and is an essential basis for evaluating, predicting, and managing water pollution. On the basis of WQC, researchers have carried out a large number of studies on the OWQM or BEW in recent years.

Since fish survival is very sensitive to changes in the water body environment, and fish objects can undergo significant phase changes in activity behavior when water quality changes, it is possible to perform online monitoring of pollutants or early warning of water quality based on the trend of fish activity behavior if the phases of fish activity can be effectively identified. Saberioon et al. [180] reviewed the application of machine vision-based monitoring systems in aquaculture, where some research advances were analyzed in particular.

When sudden pollution occurs, the fish in the water show abnormal behavior. Researchers often monitor the water quality based on the identification results of the abnormal behaviors of fish. The advanced OWQM or BEW systems mainly include the Fish taximeter and Toxprotect64 systems of BBE Company in Germany. This system uses zebrafish and tiger skin fish as marked fish to monitor water quality online by tracking and recording the behavioral characteristics of marked fish objects in the water. It has many advantages, such as fast response speed, a wide range of monitoring pollutants, 24-hour continuous monitoring, and so on. However, the monitoring fish selected by the EWS of water quality varies with the monitoring environments (such as waterworks, domestic sewage, and industrial wastewater environments). Most researchers choose fish activity characteristics such as respiratory rate, respiratory intensity, movement rate, and even fish cough rate to establish a model of the relationship with the concentration of pollutants to achieve the OWQM or BEW.

On the basis of the analysis of the respiratory movement of zebrafish, Hong-jun et al. [181] discussed the effect of exogenous heavy metal exposure on the respiratory response of zebrafish and tried to use the changes in the respiratory parameters of fish to warn against water pollution. Biosensor [182] produced by Biosensors Company of Virginia in the United States also carries out an early warning of water pollution through the breathing behavior of fish. It is reported that the system has been widely used in the United States, and the commonly used species are Rainbow trout and Bluegill.

The OWQM or BEW based on fish activity behavior analysis has many advantages, such as high sensitivity and fast calculation speed. When the water's environmental quality changes, the indicator fish living in the water body can reflect the pollution status continuously. Therefore, the fish activity recognition-based EWS can be used to monitor the pollution of rivers, lakes, reservoirs, and other water sources, especially sudden water pollution accidents. For instance, the International Committee for the Protection of the Rhine (ICPR), which consists of nine countries in the Rhine River Basin, has been effective in reducing sudden water pollution events since it constructed the Warning and Awarning Platform (WAP) in the RiverRhine. It mainly uses the biological monitoring system to perform the OWQM [183]. The National Environmental Protection

Agency Environmental Assessment Center and the U.S. Army Environmental Health Research Center have jointly used the Fish Respiration Early Warning System for the online monitoring of surface water treatment facilities, with an effective monitoring rate of 99% over 12 months of the year. Cunha et al. [184] even successfully applied the biological EWS to seawater quality monitoring by building the Marine On-line Biomonitor System (MOLBS), which automatically records the behavioral responses of marine and freshwater fish activities to achieve a BEW of marine water quality.

In summary, the OWQM or BEW based on fish activity behavior has gradually become an effective and rapid detection tool for water quality monitoring. However, since the activity behavior of fish is affected by other environmental factors such as light and temperature in addition to water pollution, the effective detection of gill respiration characteristics of fish objects is a difficult task. To summarize, intelligent OWQM or BEW based on fish activity behavior, therefore, needs to effectively detect the subtle activity characteristics of fish objects (e.g., gill respiration characteristics) and effectively to predict the stage and trend performance of fish activity behaviors to realize online monitoring and even early warning of water qualities.

4 Challenging issues and further trends

The leading water quality monitoring and assessment methods in water treatment plants involve physical and chemical analysis and biological monitoring. As physical and chemical analysis requires regular collection of water samples for the quantitative analysis of pH value, oxygen content, heavy metal content, organic content, and so on, these methods are time-consuming, labor -intensive, too costly, and difficult to achieve continuous online monitoring. They cannot reflect water pollutants' comprehensive nature and toxic effects on water organisms.

The biological monitoring method integrates the theoretical knowledge of environmental science and biological monitoring technologies by observing the response of aquatic organisms, such as individuals, communities, or populations, to pollution and then clarifying the status of environmental water pollution. The machine vision monitoring of water quality based on fish behavior is the most widely studied and fastest-growing species, and it is also the current research focus.

The identification based on the activity ability of fish can more accurately evaluate the ecological security of the water body and the living environment of fish. Identifying fish activity can be used as a sensitive indicator to detect potential harmful substances and pollution sources in water bodies. By analyzing the activity capacity of fish, we can better understand the biodiversity and ecosystem stability in water bodies, which has important scientific significance.

The water quality monitoring system with fish as the biological carrier of water quality monitoring has been applied to practical engineering. Therefore, water quality monitoring based on fish activity identification is an effective OWQM method with low cost, fast detection speed, and early warning of water quality. This article reviews the progress in the research and application of water quality monitoring and early warning based on fish activity recognition in recent years.

Due to the random nature of fish activities, fish activity behavior recognition for water quality monitoring in water treatment plants still has the following challenging issues.

(1) Multi-fish object detection and tracking

Multi-object tracking has been a fundamental and challenging research area in computer vision. Unlike the conventional rigid multi-object tracking, in the detection and tracking of multiple fish objects, the relative positions of fish and fixed monitoring cameras are randomly changed due to the random swimming of fish. Therefore, the fish objects in the video image sequence are affected by various factors, such as object occlusion, different depths of the object floating in the water, and so on. The fish objects in different monitoring moments will have corresponding geometric, perspective, and affine changes due to the reflection of the water surface, and the influence of the water surface ripples, bringing great difficulties to the online monitoring and effective tracking of multiple fish objects. The effective modeling of complex rippled water surface backgrounds, the accurate detection of numerous fish objects, and the effective association of various objects in

adjacent image frames (to achieve multi-object tracking) are the key research directions for future breakthroughs in fish object activity behavior recognition.

(2) Modeling and characterization of abnormal behaviors, such as "avoidance" of fish

Modeling and characterization of fish activity behaviors (especially abnormal behaviors, such as avoidance behaviors exhibited by fish that are directly related to water quality safety) are the basis for achieving water quality monitoring results. However, there is no more general object activity behavior modeling and characterization method. Research shows that when water quality changes (especially sudden pollution), fish object activity behavior will appear as noticeable stage changes, such as no effect period, regulation period, adaptation period, toxicological effect period, and so on. In different stages, the fish activity behavior should be characterized by different motion characteristics (such as speed, acceleration, angular velocity, trajectory, and so on), yet, this is not definitive. Therefore, modeling and characterizing various stages of fish behavior (especially abnormal behavior) is a challenging problem to solve to realize accurate fish activity behavior recognition.

(3) Accurate identification of water quality based on "fine-grained" behavioral recognition of fish activities

Fine-grained behavior identification helps to identify the nuances of fish activities further. Thus, compliance with the safety standards of water quality may still include different quality standards, such as different minerals in the water, oxygen content, and so on, and the fish activity behavior will be slightly different accordingly. To achieve the fine identification of water quality, it is natural to identify the nuances of fish activities effectively. Although fine-grained image and behavior recognition is a popular research area in computer vision, fine-grained recognition has been a complex problem in various fields. In the future, researchers can continue to explore the possibilities in the fine-grained domain. For example, local and global information plays an essential role in fine-grained image recognition tasks, and a combination of both can be considered. In addition, the transformer network can improve the accuracy of fine-grained image recognition. Suppose the accuracy of the network does not meet the requirement. In that case, the accuracy can be improved by increasing the "width" of the network, and the increase in computational effort by increasing the width is minimal compared to the rise in the number of layers of the deep network. Therefore, the application of broad learning to fine-grained images can be considered in the future to improve computational speed.

(4) Biological early warning of water monitoring through an intelligent understanding of abnormal fish activity behavior and analysis of activity trends

When the quality of the monitored water body changes, especially when sudden water pollution occurs, the fish in the water will reflect the corresponding stress due to the diffusion time, degree, and toxicity of pollutants in the water. The stage-wise behavior response will appear until the final possible fish toxicity effect occurs (such as fish mass mortality). With the fish activity recognition-based intelligent water quality monitoring, we can not only achieve the current water quality but also, more critically, achieve an early warning of water quality through the fish activity of the phased stress behaviors. Therefore, monitoring the fish object activity characteristics, especially the intelligent understanding of abnormal activity behavior and the intelligent prediction of fish activity behavior trends, is a crucial research direction in the future.

This is despite significant advances in deep learning in recent years, such as multimodal biometrics, which combines a variety of biometric information to achieve more accurate biometric identification and verification. At the same time, in fish object detection, a lot of research work is still needed to realize the distance detection and recognition level of distance from humans [185]. In the future, fish behavior recognition oriented to water quality monitoring can be concerned but not limited to the following points:

(1) Multimodal technology

Based on the multimodal technology, the motion tracking, sound, physiological indicators, and other aspects of fish are analyzed to realize the assessment and monitoring of water quality safety and ecological environment. By using a variety of sensor technologies, the technology can comprehensively and accurately judge different behavioral states of fish, including swimming speed, posture, sound, gill breathing rate, fin swing frequency, etc., thus improving the accuracy and real-time monitoring of water quality.

(2) Lightweight detection

Many applications need to improve the speed of algorithms to continuously and steadily execute on mobile devices, such as unmanned driving, smart cameras, face recognition, intelligent robots, etc. Although object detection has been greatly improved in model performance in recent years, most model networks have a large number of parameters and the computing power of embedded devices is limited, so it is difficult to run smoothly on embedded devices [186]. In recent years, many lightweight detection networks have also been proposed, such as SqueezeNet [187] for implementing model compression, MobileNet [188] for using a single-stage model with deep convolution separation, ShuffleNet [189] for mobile devices, etc., and many objected improvement methods have also appeared, such as literature [190] Lightweight SSD detection method based on feature fusion, face detection algorithm based on lightweight attention mechanism proposed in the literature [191], real-time human key point detection algorithm improved by lightweight network proposed in the literature [192], and so on. Future improvements can be made in the direction of lightweight model optimization, balancing the speed and accuracy of the model to make it run smoothly and fast on mobile devices with limited memory [193].

(3) Detection satisfying automatic machine learning (AutoML)

The detection model based on deep neural networks is becoming more and more complicated, and its modeling and application are faced with significant bottlenecks and constraints, such as heavy dependence on artificial design and long modeling cycles. Automated machine learning (AutoML) technology, which has emerged internationally, has gained wide attention in academic and industrial fields at home and abroad. It uses machines instead of humans to automatically complete model selection and super-parameter optimization, to automate model design [194]. For example, in the literature [195], the authors describe that AutoML can realize super-parameters in semantic segmentation using engineering thinking optimization, transfer learning, and neural architecture search. The possible future direction of object detection is to design an intelligent detection model by deep research on neural network structure, to reduce manual intervention on the model, such as how to select an appropriate anchor frame according to the image automatically and how to select a good optimization scheme for the model. AutoML will be a vital object detection technology in the future.

5 Conclusions

This review article meticulously examines and encapsulates the significant advancements in identifying fish activity behaviors for OWQM. It encompasses a broad spectrum of topics, including the detection and tracking of fish targets, the identification of fish activity behaviors, and the practical deployment of these methodologies in biological water quality assessment. We thoroughly catalog the principal techniques employed at each stage of fish activity behavior analysis, critically evaluating the merits and limitations inherent to various prevalent approaches within the realm of ichthyology.

Furthermore, the review article delves into the intricacies of fish object detection and tracking facilitated by machine vision technologies, the nuanced identification of aberrant behaviors, and the implications of such identifications for water quality surveillance. A comprehensive discourse is presented, synthesizing current algorithms' efficacy, strengths, and potential drawbacks, alongside a comparative analysis of their suitability across diverse environmental settings.

Concluding with a forward-looking perspective, we articulate the prevailing challenges that confront accurately recognizing anomalous fish behavior as an indicator of water quality. This contemplation extends to prognostications about the evolutionary trajectory of this field. We hope this survey article will serve as a practical and exhaustive resource for those endeavoring to harness artificial intelligence in aquatic quality control, thereby facilitating the expedited selection and application of productive technological solutions in engineering projects.

Funding information: This work was supported by the National Natural Science Foundation of China under Grant No. 62371187, partially supported by the Inclusive Policy and Innovation Environment Construction

Program of Hunan Province: Science and Technology Innovation Decision Consulting Research Project under Grant No. 2022ZL3020, and the Natural Science Foundation of Hunan Province, China under grant No. 2023JJ50495.

Author contributions: Pengfei Xu directed the project; Xianyi Liu and Jinping Liu conceived of the presented idea, wrote the original manuscript and proofread the final article; Meiling Cai, Ying Zhou, Shanshan Hu, and Minlian Chen contributed to the revised version of the manuscript. All authors provised critical feedback and helped the research, data analysis, and mansucript.

Conflict of interest: The authors declare no conflict of interest.

References

[1] L. E. D. Smith and G. Siciliano, *A comprehensive review of constraints to improved management of fertilizers in China and mitigation of diffuse water pollution from agriculture*, Agricult. Ecosyst. Environ. **209** (2015), 15–25.

[2] W. Kang-Lin, C. Ming, W. Zhi-Yu, and X. Yin-ke, *Research on signal processing for water quality monitoring based on continuous spectral analysis*, Spectroscopy Spectral Anal. **34** (2014), no. 12, 3368–3373.

[3] C. Tang, Y. Yi, Z. Yang, and J. Sun, *Risk analysis of emergent water pollution accidents based on a Bayesian network*, J. Environ. Manag. **165** (2016), no. 2, 199–205.

[4] L. Qing, X. Shi-Qin, G. Jun-Qiang, W. Shi-feng, W. Jing, C. Cheng, et al., *Pollution source identification of water body based on aqueous fingerprint-case study*, Spectroscopy Spectral Anal. **36** (2016), no. 08, 2590–2595.

[5] E. N. Knudsen, J. W. Howell, and R. L. Knudsen, *Water Quality Monitoring Device and Method*, US, 2009.

[6] M. H. Banna, S. Imran, A. Francisque, H. Najjaran, R. Sadiq, M. Rodriguez, et al., *Online drinking water quality monitoring: Review on available and emerging technologies*, Critical Rev. Environ. Sci. Technol. **44** (2014), no. 12, 1370–1421.

[7] C. Gonzalez, R. Greenwood, and P. Quevauviller, *Rapid Chemical and Biological Techniques for Water Monitoring*, Wiley, 2009.

[8] S. H. Kim, M. M. Aral, Y. Eun, J. J. Park, and C. Park, *Impact of sensor measurement error on sensor positioning in water quality monitoring networks*, Stoch. Environ. Res. Risk Assess., 2016, 1–14.

[9] W. Kanglin, Z- Y. Wen, W. Xin, Z. W. Zhang, and T. L. Zeng, *Research Advances in Water Quality Monitoring Technology Based on UV-Vis Spectrum Analysis*, Spectroscopy Spectral Anal. **31** (2011), no. 04, 1074–1077.

[10] M. J. Bae and Y. S. Park, *Biological early warning system based on the responses of aquatic organisms to disturbances: A review*, Sci. Total Environ. **466–467** (2014), no. 1, 635–649.

[11] B. A. Akinnuwesi, S. G. Fashoto, E. Mbunge, A. Odumabo, A. S. Metfula, P. Mashwama, et al., *Application of intelligence-based computational techniques for classification and early differential diagnosis of COVID-19 disease*, Data Sci. Manag. **4** (2021), 10–18.

[12] L. K. Nüßer, O. Skulovich, S. Hartmann, T. -B. Seiler, C. Cofalla, H. Schuettrumpf, et al., *A sensitive biomarker for the detection of aquatic contamination based on behavioral assays using zebrafish larvae*, Ecotoxicol. Environ. Safety **133** (2016), 271–280.

[13] X. Zhu, D. Li, D. He, J. Wang, D. Ma, and F. Li. *A remote wireless system for water quality online monitoring in intensive fish culture*, Comput. Electronics Agriculture **71** (2010), no. 1, S3–S9.

[14] D. B. Casebolt, D. J. Speare, and B. S. Horney, *Care and use of fish as laboratory animals: current state of knowledge*, Laboratory Animal Sci. **48** (1998), no. 2, 124–36.

[15] J. Liu, W. Gui, Z. Tang, C. Yang, J. Zhu, and J. Li, *Recognition of the operational statuses of reagent addition using dynamic bubble size distribution in copper flotation process*, Minerals Eng. **45** (2013), no. 1, 128–141.

[16] J. Liu, Z. Tang, W. Gui, W. Liu, P. Xu, and J. Zhu, *Application of statistical modeling of image spatial structures to automated visual inspection of product quality*, J. Process Control **44** (2016), no. 1, 23–40.

[17] J. Liu, W. Gui, Z. Tang, and J. Zhu, *Dynamic bubble-size-distribution-based health status analysis of reagent-addition in froth flotation process*, Control Theory Appl. **30** (2013), 492–502.

[18] D. L. Breitburg, K. A. Rose, and J. H. Cowan, *Linking water quality to larval survival:predation mortality of fish larvae in an oxygen-stratified water column*, Mar. Ecol. Prog. **178** (1999), no. 3, 39–54.

[19] S. C. Cary, K. J. Coyne, A. Rueckert, S. A. Wood, S. Kelly, and C. E. C. Gemmill, et al., *Development and validation of a quantitative PCR assay for the early detection and monitoring of the invasive diatom Didymosphenia geminata*, Harmful Algae **36** (2014), no. 6, 63–70.

[20] I. D. Gomes, A. A. Nascimento, A. Sales, and F. G. Araújo, *Can fish gill anomalies be used to assess water quality in freshwater neotropical systems?*, Environ. Monit. Assess. **184** (2012), no. 184, 5523–31.

[21] F. J. Kroon and G. P. Housefield, *A fluviarium with controlled water quality for preference-avoidance experiments with fish and invertebrates*, Limnol. Oceanogr. Methods **1** (2003), no. 1, 39–44.

[22] M. Thomas, A. Flroion, and D. Chretien, *A new warning biomonitor using a weakly electric fish, Apteronotus albifrons (Gymnotiformes), and the effect of temperature on the bioelectric responses*, Environ. Monit. Assess. **51** (1998), no. 3, 605–620.

[23] J. Liu, J. He, Z. Tang, W. Gui, T. Ma, H. Jahanshahi, et al., *Frame-dilated convolutional fusion network and GRU-based self-attention dual-channel network for soft-sensor modeling of industrial process quality indexes*, IEEE Trans. Syst. Man. Cybernetics Syst. **52** (2022), no. 9, 5989–6002.
[24] J. Liu, J. Wu, Y. Xie, J. Wang, P. Xu, Z. Tang, et al., *Toward Robust process monitoring of complex process industries based on denoising sparse auto-encoder*, J. Industr. Inform. Integrat. **30** (2022), 100410.
[25] J. Liu, L. Xu, Y. Xie, T. Ma, J. Wang, Z. Tang, et al., *Toward Robust fault identification of complex industrial processes using stacked sparse-denoising auto-encoder with softmax classifier*, IEEE Trans. Cybernetics **53** (2023), no. 1, 428–442.
[26] J. Liu, S. Zhao, Y. Xie, H. Jahanshahi, S. Wei, and A. Mohammadzadeh, *Fault monitoring-oriented transition process identification of complex industrial processes with neighbor inconsistent pair-based attribute reduction*, J. Process Control **121** (2023), 30–49.
[27] C. Shu-hong, L. Jie, and L. Lei-hua, *Study on anomaly water quality assessment factor based on fish movement behavior*, Chinese J. Scientif. Instrument **36** (2015), no. 8, 1759–1766.
[28] A. Yilmaz, O. Javed, and M. Shah, *Object tracking: a survey*, ACM Comput. Surv. **38** (2006), no. 4, 1–45.
[29] J. Ding, Y. Tang, H. Tian, and Y. Huang, *Robust Appearance Learning for Object Tracking in Challenging Scenes*, Springer, Berlin Heidelberg, 2014.
[30] Y. Nan, J. Cui, Z. Zheng, Z. Shanyong, Z. Liufeng, L. Dichen, et al., *Research on nonparametric kernel density estimation for modeling of wind power probability characteristics based on fuzzy ordinal optimization*, Power Electron. Technol. **40** (2016), 335–340.
[31] J. Liu, J. He, Y. Xie, W. Gui, Z. Tang, T. Ma, et al., *Illumination-invariant flotation froth color measuring via Wasserstein distance-based cycleGAN with structure-preserving constraint*, IEEE Trans. Cybernetics **51** (2021), no. 2, 2168–2275.
[32] Y. Wang, L. Jiang, Q. Liu, and M. Yin, *Optimal appearance model for visual tracking*, PLos One **11** (2016), no. 1, e0146763.
[33] Y. Pang, J. Cao, and X. Li, *Learning sampling distributions for efficient object detection*, IEEE Trans. Cybernetics **47** (2016), no. 1, 117–129.
[34] K. C. Hui and W. C. Siu, *Extended analysis of motion-compensated frame difference for block-based motion prediction error*, IEEE Trans. Image Process. **16** (2007), no. 5, 1232–1245.
[35] X. Yue-lei, Z. Ji-zhang, Z. Xion, and B. Du-yan, *A video segmentation algorithm based on accumulated frame differences*, Opto-Electronic Eng. **7** (2004), 69–72.
[36] Z. Feng-yan, G. Sheng-fa, and H. Jian-yu, *Moving object detection and tracking based on weighted accumulative difference*, Comp. Eng. **35** (2009), no. 22, 159–161.
[37] Q. Jing-jing and X. Yun-hong, *Combined continuous frame difference with background difference method for moving object detection*, Acta Photonica Sinica **43** (2014), no. 07, 219–226.
[38] C. Stauffer and W. E. L. Grimson, *Learning patterns of activity using real-time tracking*, IEEE Trans. Pattern Anal. Machine Intelligence **22** (2000), no. 8, 747–757.
[39] O. Barnich and D. M. Van, *ViBe: a universal background subtraction algorithm for video sequences*, IEEE Trans. Image Process. Pblication IEEE Signal Process. Soc. **20** (2011), no. 6, 1709.
[40] J. M. Guo, C. H. Hsia, Y. F. Liu, and M. H. Shih, *Fast background subtraction based on a multilayer codebook model for moving object detection*, IEEE Trans. Circuits Syst. Video Technol. **23** (2013), no. 10, 1809–1821.
[41] L. Maddalena and A. Petrosino, *The SOBS algorithm: What are the limits?* In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012.
[42] C. Ying-xia and Y. Yi-biao, *Non-parallel Corpora voice conversion based on structured Gaussian mixture model under constraint conditions*, Acta Electronica Sinica **44** (2016), no. 9, 2282–2288.
[43] O. Barnich and M. Van Droogenbroeck, *ViBE: A powerful random technique to estimate the background in video sequences* In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.
[44] B. K. P. Horn and B. G. Schunck, *Determining optical flow: a retrospective*, Artif. Intell. **59** (1993), no. 93, 81–87.
[45] S. S. Beauchemin and J. L. Barron, *The computation of optical flow*, ACM Comput. Surv. **27** (1995), no. 3, 433–466.
[46] C. Li-chao, X. Dan, C. Jian-fang, and Z. Rui, *Research on vehicle real-time detection algorithm based on improved optical flow method and GMM*, CAAI Trans. Intelligent Syst. **16** (2021), no. 02, 271–278.
[47] D. Comaniciu and P. Meer, *Mean shift: A Robust approach toward feature space analysis*, IEEE Trans. Pattern Anal. Machine Intell. **24** (2002), no. 5, 603–619.
[48] D. Comaniciu, V. Ramesh, and P. Meer, *Kernel-based object tracking*, IEEE Trans. Pattern Anal. Machine Intelligence **25** (2003), no. 5, 564–575.
[49] Y. Li, Y. Li, H. Kim, and S. Serikawa, *Active contour model-based segmentation algorithm for medical robots recognition*, Multimedia Tools Appl. **77** (2017), 1–16.
[50] K. Zhang, H. Song, and L. Zhang, *Active contours driven by local image fitting energy*, Pattern Recognition **43** (2010), no. 4, 1199–1206.
[51] X. Kai, Q. Kun, H. Bo-he, and D. Yi, *A new method of region based image segmentation based on cloud model*, J. Image Graph. **15** (2010), no. 05, 757–763.
[52] L. Jin-ping, C. Qing, Z. Jin, and T. Zhao-hui. *Interactive image segmentation based on ensemble learning*, Acta Electr. Sinica **44** (2016), no. 07, 1649–1655.
[53] C. Ying, M. Qi-guang, L. Jia-Cheng, and G. Lin, *Advance and prospects of AdaBoost algorithm*, Acta Autom. Sin. **39** (2013), no. 6, 745–758.
[54] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*, In Proceedings of the Computer Vision and Pattern Recognition, 2001 CVPR 2001 Proceedings of the 2001 IEEE Computer Society Conference on, 2001.

[55] R. Lienhart, *An extended set of Haar-like features for rapid object detection*, In Proceedings of the 2002 IEEE International Conference on Image Processing, vol 1, 2002, pp. 900–903.
[56] R. Girshick, J. Donahue, T. Darrell, and J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[57] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. -Y. Fu, et al., *SSD: Single Shot MultiBox Detector*, 2015, arXiv:1512.02325.
[58] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, 2015, arXiv:1506.02640.
[59] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. *End-to-End Object Detection with Transformers*, 2020, arXiv:2005.12872.
[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., *Going deeper with convolutions*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet A large-scale hierarchical image database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[62] J. Redmon and A. Farhadi, *YOLO9000: better, faster, stronger*, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.
[63] K. He, X. Zhang, S. Ren, and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[64] R. Girshick, *Fast R-CNN, 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
[65] J. Redmon and A. Farhadi, *Yolov3: An Incremental Improvement*, 2018, arXiv preprint arXiv:1804.02767.
[66] A. Bochkovskiy, C.-Y. Wang, and H-Y. M. Liao *Yolov4: Optimal Speed and Accuracy of Object Detection*, 2018, arXiv preprint arXiv:2004.10934.
[67] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, *Distance-IoU loss: Faster and better learning for bounding box regression*, Proc. AAAI Confer. Artif. Intel. **34** (2020), no. 7, 12993–13000.
[68] M. D. Mish, *A self regularized non-monotonic activation function*, 2019, arXiv preprint arXiv:1908.08681.
[69] C. Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, *CSPNet: A new backbone that can enhance learning capability of CNN*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391.
[70] K. He, X. Zhang, S. Ren, and J. Sun, *Spatial pyramid pooling in deep convolutional networks for visual recognition*, IEEE Trans. Pattern Anal. Machine Intell. **37** (2015), no. 9, 1904–1916.
[71] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, *Path aggregation network for instance segmentation*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
[72] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, *Scalable object detection using deep neural networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2147–2154.
[73] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, *Deformable DETR: Deformable Transformers for End-to-End Object Detection*, 2020, arXiv:2010.04159.
[74] H. Rui-ze, F. Wei, G. Qing, and H. Qing-hua, *Single object tracking research: a survey*, Chinese J. Comput. **45** (2022), no. 9, 1877–1907.
[75] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, *Fully-Convolutional Siamese Networks for Object Tracking*, 2016, arXiv:1606.09549.
[76] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, *Visual object tracking using adaptive correlation filters*, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2544–2550.
[77] J. Bromley, I. Guyon, Y. Lecun, E. Säckinger, R. Shah, *Signature Verification Using a Siamese Time Delay Neural Network*, 1993, p. 6.
[78] R. Tao, E. Gavves, and AW M. Smeulders, *Siamese Instance Search for Tracking*, 2016, arXiv:1605.05863.
[79] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, *High performance visual tracking with Siamese region proposal network*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8971–8980.
[80] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, *Distractor-aware Siamese Networks for Visual Object Tracking*, 2018, arXiv:1808.06048.
[81] G. Wang, C. Luo, Z. Xiong, and W. Zeng, *SPM-Tracker: Series-Parallel Matching for Real-Time Visual Object Tracking*, 2019, arXiv:1904.04452.
[82] H. Fan and H. Ling, *Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking*, 2018, arXiv:1812.06148.
[83] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, 2015, arXiv:1506.01497.
[84] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, *SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks*, 2018, arXiv:1812.1.1703.
[85] Z. Zhang and H. Peng, *Deeper and Wider Siamese Networks for Real-Time Visual Tracking*, 2019, arXiv:1901.01660.
[86] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, *Learning attentions: residual attentional siamese network for high performance online visual tracking*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4854–4863.
[87] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, *Deformable Siamese Attention Networks for Visual Object Tracking*, 2020, arXiv:2004.06711.
[88] F. Du, P. Liu, W. Zhao, and X. Tang, *Correlation-guided attention for corner detection based visual tracking*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 6835–6844.
[89] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, *Deformable Siamese Attention Networks for Visual Object Tracking*, 2020, arXiv:2004.06711.

[90] B. Keni and S. Rainer, *Evaluating multiple object tracking performance: the CLEAR MOT metrics*, Eurasip J. Image Video Process. **2008** (2008), no. 1, 246309.
[91] Z. Yao, L. Huan-zhang, Z. Lu-ping, and H. Mou-fa, *Overview of visual multi-object tracking algorithms with deep learning*, Comput. Eng. Appl. **57** (2021), no. 13, 55–66.
[92] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, et al., *Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking*, 2020, arXiv:2007.14557.
[93] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[94] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, *Feature Pyramid Networks for Object Detection*, 2016, arXiv:1612.03144.
[95] X. Weng, Y. Wang, Y. Man, and K. Kitani, *GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with Multi-Feature Learning*, 2020, arXiv:2006.07327.
[96] J. He, Z. Huang, N. Wang, and Z. Zhang, *Learnable Graph Matching: Incorporating Graph Partitioning with Deep Feature Learning for Multiple Object Tracking*, 2021, arXiv:2103.16178.
[97] A. Gerhardt, D. B. L. Janssens, and A. M. Soares, *Evidence for the stepwise stress model: Gambusia holbrooki and Daphnia magna under acid mine drainage and acidified reference water stress*, Environ. Sci. Technol. **39** (2005), no. 11, 4150–8.
[98] Z. Jin-song, H. Yi, H. Xiao-bo, and H. Ting-lin, *Application of changes of the fish behavior in the water quality monitoring*, Water Wastewater Eng. **49** (2013), no. 7, 166–170.
[99] G. Varol, I. Laptev, and C. Schmid, *Long-term Temporal Convolutions for Action Recognition*, 2016, arXiv:1604.04494.
[100] N. Hussein, E. Gavves, and A. W. M. Smeulders, *Timeception for Complex Action Recognition*, 2018, arXiv:1812.01289.
[101] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, *Large-scale video classification with convolutional neural networks*, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[102] J. Sanchez-Riera, K.-L. Hua, Y.-S. Hsiao, T. Lim, S. C. Hidayati, and W.-H. Cheng, *A comparative study of data fusion for RGB-D based visual recognition*, Pattern Recognit. Lett. **73** (2016), 1–6.
[103] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah, and M. Bennamoun, *Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM*, IEEE Trans. Multimedia **21** (2019), no. 4, 1011–1021.
[104] M. Li, H. Leung, and H. P. Shum, *Human action recognition via skeletal and depth based feature fusion*, Proceedings of the 9th International Conference on Motion in Games, 2016, pp. 123–132.
[105] M. A. Goodale and A. D. Milner, *Separate visual pathways for perception and action*, Trends Neurosci. **15** (1992), no. 1, 20–25.
[106] K. Simonyan and A. Zisserman, *Two-Stream Convolutional Networks for Action Recognition in Videos*, 2014, arXiv:1406.2199.
[107] C. Feichtenhofer, A. Pinz, and R. P. Wildes, *Spatiotemporal Residual Networks for Video Action Recognition*, 2016, arXiv:1611.02155.
[108] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, et al., *Temporal Segment Networks: Towards Good Practices for Deep Action Recognition*, 2016, arXiv:1608.00859.
[109] M. Zhi-qiang, M. Cui-hong, C. Jin-long, and W. Yi, *Human action recognition model based on spatio-temporal two-stream convolution and LSTM*, Software **39** (2018), no. 9, 9–12.
[110] B. Xue, *Human action recognition based on two-stream network*, Zhengzhou University, 2019.
[111] A. Miao, *Research on video action recognition based on deep learning*, North China Electric Power University, 2019.
[112] M. Zhi-qiang, *Research on human abnormal behavior analysis technology in video sequences*, North China University of Science and Technology, 2019.
[113] W. Biao, *A series-stream deep network model for video action recognition*, Jiangxi University of Science and Technology, 2019.
[114] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, et al., *FlowNet: Learning optical flow with convolutional networks*, 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2758–2766.
[115] X. Wang and H. Deng, *A multi-feature representation of skeleton sequences for human interaction recognition*, Electronics **9** (2020), no. 1, 187.
[116] L. Wang, L. Ge, R. Li, and Y. Fang, *Three-stream CNNs for action recognition*, Pattern Recognit. Lett. **92** (2017), 33–40.
[117] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, *Action recognition with dynamic image networks*, IEEE Trans. Pattern Anal. Machine Intel. **40** (2017), no. 12, 2799–2813.
[118] Y. Wen-han, *Research on gesture recognition algorithm based on multi-stream three dimensions convolutional neural network*, Xidian University, 2017.
[119] V. A. Chenarlogh and F. Razzazi, *Multi-stream 3D CNN structure for human action recognition trained by limited data*, IET Comput. Vision **13** (2019), no. 3, 338–344.
[120] Y. Gu, X. Ye, W. Sheng, Y. Ou, and Y. Li, *Multiple stream deep learning model for human action recognition*, Image Vision Comput. **93** (2020), 103818.
[121] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, *Optical flow guided feature: A fast and robust motion representation for video action recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1390–1399.
[122] S. Ji, W. Xu, M. Yang, and K. Yu, *3D convolutional neural networks for human action recognition*, IEEE Trans. Pattern Anal. Machine Intel. **35** (2012), no. 1, 221–231.
[123] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning spatiotemporal features with 3d convolutional networks*, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.

[124] J. Carreira and A. Zisserman, *Quo vadis, action recognition? a new model and the kinetics dataset*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
[125] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, *Convnet Architecture Search for Spatiotemporal Feature Learning*, 2017, arXiv preprint arXiv:1708.05038.
[126] H. Hai-yang, D. Jia-min, H. Hua, C. Jie, and L. Zhong-jin, *Workflow recognition method based on 3D convolutional neural networks*, Comput. Integrated Manuf. Syst. **24** (2018), no. 7, 1747–1757.
[127] Y. Ming-li, *Research on real-time video action classification based on three-dimensional convolutional*, Beijing University of Posts and Telecommunications, 2019.
[128] X. Xin, *Research on dynamic gesture recognition method based on three-dimensional deep neural network*, Xidian University, 2018.
[129] W. Yang, Y. Chen, C. Huang, and M. Gao, *Video-based human action recognition using spatial pyramid pooling and 3D densely convolutional networks*, Future Internet **10** (2018), no. 12, 115.
[130] K. Hara, H. Kataoka, and Y. Satoh, *Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?* Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.
[131] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, et al., *The Kinetics Human Action Video Dataset*, 2017, arXiv preprint arXiv:1705.06950.
[132] K. He, X. Zhang, S. Ren, and J. Sun, *Identity mappings in deep residual networks*, European Conference on Computer Vision, 2016, pp. 630–645.
[133] S. Zagoruyko and N. Komodakis, *Wide Residual Networks*, 2016, arXiv:1605.07146.
[134] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, *Aggregated residual transformations for deep neural networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.
[135] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[136] Y. Huang, Y. Guo, and C. Gao, *Efficient parallel inflated 3D convolution architecture for action recognition*, IEEE Access **8** (2020), 45753–45765.
[137] J. Thompson and R. Parasuraman, *Attention, biological motion, and action recognition*, Neuroimage **59** (2012), no. 1, 4–13.
[138] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, *Cbam: Convolutional block attention module*, Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[139] J. Cai and J. Hu, *3D RANs: 3D residual attention networks for action recognition*, Visual Comput. **36** (2020), no. 6, 1261–1270.
[140] Q. Liu, X. Che, and M. Bie, *R-STAN: Residual spatial-temporal attention network for action recognition*, IEEE Access **7** (2019), 82246–82255.
[141] J. Li, X. Liu, M. Zhang, and D. Wang, *Spatio-temporal deformable 3d convnets with attention for action recognition*, Pattern Recognit. **98** (2020), 107037.
[142] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, et al., *Asymmetric 3d convolutional neural networks for action recognition*, Pattern Recognit. **85** (2019), 1–12.
[143] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, *Human action recognition using factorized spatio-temporal convolutional networks*, Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4597–4605.
[144] Z. Qiu, T. Yao, and T. Mei, *Learning spatio-temporal representation with pseudo-3d residual networks*, Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.
[145] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, *A closer look at spatiotemporal convolutions for action recognition*, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.
[146] D. Tran, H. Wang, L. Torresani, and M. Feiszli, *Video classification with channel-separated convolutional networks*, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5552–5561.
[147] Z. Ji-yuan, *Research of Human Action Recognition Based on Deep Learning*, Chongqing University Of Technology, 2019.
[148] M. Li-jun, *Research on Behavior Recognition Algorithm Based on 3D Convolutional Neural Network*, China University Of Geosciences, 2018.
[149] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, *Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification*, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 305–321.
[150] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, *Beyond short snippets: Deep networks for video classification*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694–4702.
[151] Y. Zhang, K. Hao, X. Tang, B. Wei, and L. Ren, *Long-term 3D convolutional fusion network for action recognition*, 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2019, pp. 216–220.
[152] X. Ouyang, S. Xu, C. Zhang, P. Zhou, Y. Yang, G. Liu, et al., *A 3D-CNN and LSTM based multi-task learning architecture for action recognition*, IEEE Access **7** (2019), 40757–40770.
[153] S. Yu, L. Xie, L. Liu, and D. Xia, *Learning long-term temporal features with deep neural networks for human action recognition*, IEEE Access **8** (2019), 1840–1850.
[154] S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, *3D-CNN-based fused feature maps with LSTM applied to action recognition*, Future Internet **11** (2019), no. 2, 42.
[155] H. Yang, J. Zhang, S. Li, and T. Luo, *Bi-direction hierarchical LSTM with spatial-temporal attention for action recognition*, J. Intel. Fuzzy Syst. **36** (2019), no. 1, 775–786.

[156] T. Yu, C. Guo, L. Wang, H. Gu, S. Xiang, and C. Pan, *Joint spatial-temporal attention for action recognition*, Pattern Recognit. Lett. **112** (2018), 226–233.
[157] N. Khaled, M. Marey, and M. Aref, *Temporal action detection with fused two-stream 3d residual neural networks and bi-directional LSTM*, 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 130–140.
[158] T. Jin, Z. He, A. Basu, J. Soraghan, G. Di Caterina, and L. Petropoulakis, *Dense convolutional networks for efficient video analysis*, 2019 5th International Conference on Control, Automation and Robotics (ICCAR), 2019, pp. 550–554.
[159] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, *Human action recognition by learning spatio-temporal features with deep neural networks*, IEEE access **6** (2018), 17913–17922.
[160] X. Wang, W. Xie, and J. Song, *Learning spatiotemporal features with 3DCNN and ConvGRU for video anomaly detection*, 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018, pp. 474–479.
[161] G. Zhu, L. Zhang, L. Yang, L. Mei, S. A. A. Shah, M. Bennamoun, et al., *Redundancy and attention in convolutional LSTM for gesture recognition*, IEEE Trans. Neural Networks Learn. Sys. **31** (2019), no. 4, 1323–1335.
[162] Y. Xin, H. Xiao-jiao, L. Huang-da, Y. Xin-jie, F. Liang-Zhong, and L. Ying, *Anomaly detection of fish school behavior based on features statistical and optical flow methods*, Trans. Chinese Soc. Agricult. Eng. **30** (2014), no. 2, 162–168.
[163] T. B. Moeslund, A. Hilton, and V. Krüger, *A survey of advances in vision-based human motion capture and analysis*, Comp. Vision Image Understanding **104** (2006), no. 2–3, 90–126.
[164] W. Jeon, S. H. Kang, J. B. Leem, and S. H. Lee, *Characterization of fish schooling behavior with different numbers of Medaka (Oryzias latipes) and goldfish (Carassius auratus) using a Hidden Markov Model*, Phys. A Stat. Mech. Appl. **392** (2013), no. 10, 2426–2433.
[165] A. Mihoub, G. Bailly, C. Wolf, and F. Elisei, *Graphical models for social behavior modeling in face-to face interaction*, Pattern Recognit. Lett. **74** (2016), 82–89.
[166] I. Fatima, M. Fahim, Y. K. Lee, and S. Lee, *A unified framework for activity recognition-based behavior analysis and action prediction in smart homes*, Sensors **13** (2013), no. 2, 2682–99.
[167] Y. Zhang, S. Wang, P. Phillips, and G. Ji, *Binary PSO with mutation operator for feature selection using decision tree applied to spam detection*, Knowledge-Based Sys. **64** (2014), 22–31.
[168] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, *Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization*, Pattern Recognit. **76** (2018), 704–714.
[169] G.-S. Xie, X.-Y. Zhang, W. Yang, M. Xu, S. Yan, and C.-L. Liu, *LG-CNN: From local parts to global discrimination for fine-grained recognition*, Pattern Recognit. **71** (2017), 118–131.
[170] Q. Wang, P. Li, and L. Zhang, *G2DeNet: Global Gaussian distribution embedding network and its application to visual recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2730–2739.
[171] A. Simonelli, F. De Natale, S. Messelodi, and S. R. Bulo, *Increasingly specialized ensemble of convolutional neural networks for fine-grained recognition*, 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 594–598.
[172] Y. Chen, *Convolutional Neural Network for Sentence Classification*, University of Waterloo, 2015.
[173] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-scale Image Recognition*, 2014, arXiv preprint arXiv:1409.1556.
[174] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., *Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.
[175] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, et al., *TransFG: A Transformer Architecture for Fine-grained Recognition*, 2021, arXiv:2103.07976.
[176] Y. Zhang, J. Cao, L. Zhang, et al., *A free lunch from ViT adaptive attention multi-scale fusion Transformer for fine-grained visual recognition*, ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3234–3238.
[177] X. Liu, L. Wang, and X. Han, *Transformer with peak suppression and knowledge guidance for fine-grained image recognition*, Neurocomputing **492** (2022), 137–149.
[178] J. Wang, X. Yu, and Y. Gao, *Feature fusion vision transformer for fine-grained visual categorization*, 2021, arXiv preprint arXiv:2107.02341.
[179] M. V. Conde and K. Turgutlu, *Exploring Vision Transformers for Fine-Grained Classification*, 2021, arXiv preprint arXiv:2106.10587.
[180] M. Saberioon, A. Gholizadeh, P. Cisar, A. Pautsina, and J. Urban, *Application of machine vision systems in aquaculture with emphasis on fish: State-of-the-art and key issues*, Rev. Aquaculture **110** (2016), no. 2, 466–469.
[181] W. Hong-jun, L. Si-Xin, Z. Lian-feng, Z. Jin-xiu, and L. You-guang, *The effect of exposure to five kinds of heavy metals on respiratory movement of zebra fish (Brachydanio rerio)*, J. Agro-Environ. Sci. **29** (2010), no. 09, 1675–1680.
[182] A. Tsopela, A. Laborde, L. Salvagnac, V. Ventalon, E. Bedel-Pereira, I. Séguy, et al., *Development of a lab-on-chip electrochemical biosensor for water quality analysis based on microalgal photosynthesis*, Biosens. Bioelectron. **79** (2015), 568.
[183] P. Diehl, T. Gerke, A. Jeuken, J. Lowis, R. Steen, J. V. Steenwijk, et al., *Early Warning Strategies and Practices Along the River Rhine*, Springer, Berlin Heidelberg, 2006.
[184] S. R. Cunha, R. Gonçalves, S. R. Silva, and A. D. Correia, *An automated marine biomonitoring system for assessing water quality in real-time*, Ecotoxicology **17** (2008), no. 6, 558–564.
[185] Z. Ze-miao, H. Huan, and Z. Feng-yu, *Survey of object detection algorithm based on deep convolutional neural networks* J. Chinese Comput. Syst. **40** (2019), no. 9, 1825–1831.

[186] Q. Rong, J. Ruisheng, X. Zhifeng, and M. Qichao, *Lightweight object detection network based on YOLOV3*, Comput. Appl. Softw. **37** (2020), no. 10, 208–213.
[187] F. N. Iandola, S. Han, and M. W. Moskewicz, *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size*, 2016, arXiv:1602.07360.
[188] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017, arXiv preprint arXiv:1704.04861.
[189] X. Zhang, X. Zhou, M. Lin, and J. Sun, *Shufflenet: An extremely efficient convolutional neural network for mobile devices*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[190] W. Tian-cheng, W. Xiao-quan, C. YI-jun, J. You-bo, and C. Cheng-ying, *Lightweight SSD object detection method based on feature fusion*, Chinese J. Liquid Crystal Displays **36** (2021), no. 10.
[191] G. Liuya, S. Dong, and L. Yixiang, *Face detection algorithm based on a lightweight attention mechanism network*, Laser Optoelectronics Progress **58** (2021), no. 2, 0210010.
[192] H. Jianghao, W. Hongyu, Q. Wenchao, and M. JingXuan, *Real-Time Human Keypoint Detection Algorithm Based onLightweight Network, Computer Engineering*, 2021.
[193] T. Schröder and M. Schulz, *Monitoring machine learning models: A categorization of challenges and methods*, Data Sci. Manag. **5** (2022), no. 3, 105–116.
[194] F. Xin, *Algorithm research and system implementation of automatic machine learning for typical scenarios*, Nanjing University, 2020.
[195] L. Guixiong, H. Jian, L. Siyang, and L. Pul, *AutoML method for semantic segmentation of machine vision*, Laser J. **40** (2019), no. 6, 1–9.
[196] A. Emadi, T. Lipniacki, A. Levchenko, and A. Abdi, *Single-cell measurements and modeling and computation of decision-making errors in a molecular signaling system with two output molecules*, Biology **12** (2023), no. 12, 1461.
[197] A. Emadi, T. Lipniacki, A. Levchenko, and A. Abdi, *A decision making model where the cell exhibits maximum detection probability: Statistical signal detection theory and molecular experimental data*, 2023 57th Annual Conference on Information Sciences and Systems (CISS), 2023, pp. 1–4.