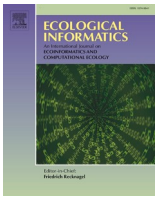
Ecological Informatics 82 (2024) 102733

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Trajectory-based fish event classification through pre-training with diffusion models

Noemi Canovi [a], Benjamin A. Ellis [b,c], Tonje K. Sørdalen [d], Vaneeda Allken [c], Kim T. Halvorsen [c], Ketil Malde [c,e], Cigdem Beyan [f,*]

[a] Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy
[b] School of Biological and Marine Sciences, University of Plymouth, Plymouth PL4 8AA, United Kingdom
[c] Institute of Marine Research, Ecosystem Acoustics Group, Bergen 5817, Norway
[d] Centre for Coastal Research, Department of Natural Sciences, University of Agder, Kristiansand 4604, Norway
[e] Department of Informatics, Faculty of Natural Sciences, University of Bergen, Bergen 5008, Norway
[f] Department of Computer Science, University of Verona, Verona 37134, Italy

ARTICLE INFO

Keywords:
Fish behavior
Underwater videos
Event recognition
Trajectory
Generative models
Autoencoder
Diffusion model
Corkwing wrasse

ABSTRACT

This study contributes to advancing the field of automatic fish event recognition in natural underwater videos, addressing the current gap in studying fish interaction and competition, including predator-prey relationships and mating behaviors. We used the corkwing wrasse (*Symphodus melops*) as a model, a marine species of commercial importance that reproduces in sea-weed nests built and cared for by a single male. These nests attract a wide range of visitors and are the focal point for behavior such as spawning, chasing, and maintenance. We propose a deep learning methodology to analyze the movement trajectories of the nesting male and classify the associated events observed in their natural habitat. Our approach leverages unsupervised pre-training based on diffusion models, leading to improved feature learning. Additionally, we introduce a dataset comprising 16,937 trajectories across 12 event classes, making it the largest in terms of event class diversity. Our results demonstrate the superior performance of our method compared to several deep architectures. The code for the proposed method and the trajectories can be found at https://github.com/NoeCanovi/Fish_Behaviors_Generative_Models.

1. Introduction

Animal behavior drives change and dynamics in populations and ecosystems through shaping predator-prey relationships, social networks, and the outcome of competition over mates and resources (Gurevitch et al., 2000; Shuster and Wade, 2003; Wey et al., 2008). An understanding of the key behavior and their cues and consequences is, therefore, a necessity for implementing effective conservation and management measures. Direct observations of behaviors are challenging for many species, but the recent decades' advances in camera and battery technology now offer the opportunity for collecting behavior data in remote locations that are difficult to access for human observers, or when human presence can disturb the animals (Caravaggi et al., 2017; Claridge et al., 2004; Ramsey et al., 2019).

Some behaviors matter more than others such that they can have a strong, often direct influence on the survival and reproduction of an individual. Survival-related behaviors, such as predation, can be a rare event and difficult to quantify using cameras. Reproduction often occurs in specific places and times across many species, which facilitates targeted observational studies. On the other hand, the nesting behavior in birds and fish, where the nest is the focal point for both mating and parental care can be relatively easily observed using cameras.

In this study, we focus on the nesting behavior of the corkwing wrasse (*Symphodus melops*), where only the males build nests and care for the eggs (Halvorsen et al., 2017a). In this species, the nesting males are larger and more colorful as compared to females (Halvorsen et al., 2016). However, a smaller proportion of males develop as sneaker males - a fixed alternative reproductive strategy. The sneaker males do not build nests and closely resemble females in appearance. Their strategy is rather to deceive the nesting male to get access to the nest and fertilize the eggs, and in that way avoiding the burden of parental care (Potts, 1974; Uglem et al., 2000). The nests also attract egg predators, which

* Corresponding author.
E-mail addresses: ben.ellis@plymouth.ac.uk (B.A. Ellis), vaneeda.allken@hi.no (V. Allken), kim.halvorsen@hi.no (K.T. Halvorsen), ketil.malde@hi.no (K. Malde), cigdem.beyan@univr.it (C. Beyan).

https://doi.org/10.1016/j.ecoinf.2024.102733
Received 10 May 2024; Received in revised form 5 July 2024; Accepted 19 July 2024
Available online 28 July 2024

can be corkwing or other wrasse species. The corkwing wrasse plays a vital role in marine ecosystems by regulating populations of benthic invertebrates and helping to maintain ecological balance. Wrasse are also recognized as cleaner fish, as they consume ectoparasites from other fish, effectively cleaning them. Due to this behavior, they are utilized in salmon farming to manage lice infestations (Karaszkiewicz, 2020; Skiftesvik et al., 2014). Corkwing wrasse is also a commercially important marine fish in fisheries, and indeed concerns have been raised about intensive fishing practices that disproportionately target nesting males over females and sneaker males (Halvorsen et al., n.d.; Halvorsen et al., 2016; Halvorsen et al., 2017a). Such fishing practices, which alter sex ratios and size distributions, can indirectly affect reproductive behavior and population productivity (Kindsvater et al., 2020).

For corkwing wrasse and other nesting fish species, the reproductive behaviors can be identified by distinctive movement patterns of the individuals (e.g., spawning), enabling animal behavior scientists to accurately detect and quantify them in video using manual behavior annotation software. However, this procedure is very time-consuming and requires expert knowledge of the behavior of the species in question, limiting the potential for upscaling and expanding studies. Automation through machine/deep learning holds promise in addressing the challenge of analyzing these data-intensive videos. Advancements in computer vision for automated processing of data from animal studies now offer hope to many ecologists who struggle with this laborious task (Beyan and Browman, 2020; Dell et al., 2014; Goodwin et al., 2022). Furthermore, such advancements can provide new tools for marine biologists to observe behavioral changes in response to environmental changes. For example, they can use video monitoring to detect the onset and duration of the spawning period, as many fisheries are managed with temporal fishing closures during the reproductive period (Halvorsen et al., n.d.; Halvorsen et al., 2017b).

In this paper, we present a deep learning approach to analyze the movement trajectories of male corkwing wrasse to automatically detect their behavior in their natural habitat across various event categories. We introduce a novel unsupervised pre-training based approach (also called unsupervised feature learning) that harnesses the reconstruction capability of a type of generative model for enhanced feature learning. Traditionally, ecological problems have been addressed using computer vision and machine/deep learning approaches predominantly through fully supervised learning methods (Ditria et al., 2021; Frainer et al., 2023; Fundel et al., 2023; Marjani et al., 2023; Sujatha et al., 2023; Truong et al., 2023). However, unsupervised pre-training presents an effective alternative to this approach. Its advantages lie in data efficiency, transferability capability, the capacity to learn robust and generalizable feature representations, and its role as a regularization technique to mitigate overfitting to the labeled data as shown in several studies such as (D'incà et al., 2023; Erhan et al., 2010; Franceschini et al., 2022; Ge et al., 2023; Paoletti et al., 2022a; Suryawati et al., 2021; Zhang et al., 2022). Such attributes make unsupervised pre-training particularly appealing compared to solely relying on fully supervised methods.

While earlier works on various computer vision tasks employed Restricted Boltzmann Machines and their variants for unsupervised pre-training (Beyan et al., 2017; Katsageorgiou et al., 2017; Phan et al., 2016), as well as several types of autoencoders that have been gaining popularity (Erhan et al., 2010; Ge et al., 2023; Suryawati et al., 2021), recent advancements have highlighted the superior performance of diffusion models over autoencoders (Cao et al., 2024; Paoletti et al., 2021a; Paoletti et al., 2022b; Xiang et al., 2023). Therefore, it is now imperative to investigate the efficacy of diffusion models also for unsupervised feature learning, particularly for addressing ecological problems. In line with this, our study utilizes diffusion models, where the acquired features after unsupervised pre-training are subsequently employed to enhance the training of a classifier, facilitating effective fish event classification.

To conduct the experimental analysis, we introduce a dataset comprising 16,937 trajectories corresponding to 12 different event classes. This dataset is the largest of its kind in terms of the diversity of event classes. The current literature lacks studies on automatic recognition of fish events based on trajectories in natural underwater videos such that the earlier research primarily focused on detecting unusual trajectories (Beyan and Fisher, 2012;Beyan and Fisher, 2013a; Beyan and Fisher, 2013b). For those considering RGB video frames instead of trajectories and apply deep learning, their scope is limited as they often focus on a single behavior class and analyze fish in restricted environments such as cages (Måløy et al., 2019; McIntosh et al., 2020). We argue that the absence of research like ours (i.e., conducted in fully natural settings and covering multiple classes) can primarily be attributed to the significant time and effort required for collecting and annotating event data in a detailed manner, typically done manually by fish behavior experts. Therefore, not only does the proposed method presented in this study stand as a pioneering contribution, but also the collected dataset represents an important initiative.

We further benchmarked our dataset (a) by incorporating state-of-the-art fish trajectory features (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b), (b) by applying fully supervised methods (i.e., based on one-dimensional Convolutional Neural Networks (1D-CNN)), and (c) by employing autoencoders (Tur et al., 2023a; Tur et al., 2023b). While such methods are established models for trajectory analysis, they have also been applied to imbalanced datasets, similar to the one we have on hand, for action/event classification and as solutions to ecological informatics problems. These additional analyses allow us to evaluate the efficacy of our proposed method in comparison to alternative methodologies, thereby confirming its superior performance.

The remainder of this paper is structured as follows: Section 2 delves into event/action recognition methods within computer vision, with a particular emphasis on trajectory-based methods, fish behavior, diffusion models, and the species under study in this paper. The dataset utilized in this paper is introduced in Section 3. Section 4 provides a detailed explanation of the proposed method along with its implementation details. The experimental results are presented in Section 5.2, followed by an in-depth discussion of the findings in Section 6. Finally, we conclude the paper with a summary, limitations, and potential future research directions in Section 7.

2. Related work

In this section, we provide a comprehensive discussion on event/action recognition studies in computer vision, along with diffusion models. Additionally, we compare our method, highlighting its distinctions in several key aspects compared to prior related studies. We also summarize the characteristics of the species under study in this paper.

2.1. Event recognition in computer vision

Event recognition, predominantly focused on *human* actions in the literature of computer vision, endeavors to detect and recognize a predetermined set of activity categories performed by one or more persons by analyzing images or videos. Key applications of action/event recognition include video surveillance, video content annotation and retrieval, healthcare, robotics, and scene modeling, among others.

Approaches of action/event recognition can be divided into two categories such that those employing hand-crafted features and those utilizing deep models. On the other hand, the problem is typically addressed by using fully supervised learning methods, although lately there are also approaches utilizing unsupervised feature learning (i.e., unsupervised pre-training) (Paoletti et al., 2021b; Paoletti et al., 2022c). When considering the data type, research on action/event recognition can be categorized into methods utilizing RGB data and those integrating color with depth data (i.e., RGBD) (Kong and Fu, 2022; Pareek and Thakkar, 2021). RGB and RGBD-based methods analyze the

appearance of individuals in the scene, whereas an alternative approach is to utilize trajectories, which offer a more compact representation. Trajectories have been defined in terms of the tracking path of human skeleton data (Lin et al., 2020; Paoletti et al., 2021a; Paoletti et al., 2021c; Paoletti et al., 2022b; Su et al., 2020; Zheng et al., 2018) or the center coordinates of detection bounding boxes encapsulating objects of interest, such as animals, over time (Beyan and Fisher, 2013c; Beyan and Fisher, 2013d; Palazzo et al., 2012; Schindler and Steinhage, 2021).

Given that our approach also relies on *a)* trajectories, *b)* unsupervised pre-training and specifically *c)* focusing on fish events, in this section, we delve into prior works on these topics. Other papers can be visited from the latest survey papers (Estevam et al., 2021; Kong and Fu, 2022; Pareek and Thakkar, 2021) on action recognition in computer vision.

2.1.1. Trajectory-based recognition

The methods based on hand-crafted features, as presented by Wang et al. introduce the dense trajectory (DT) and improved trajectory (IDT) techniques (Wang et al., 2011; Wang and Schmid, 2013). Initially, spatial feature points are identified on each frame of the image, and these points are then individually tracked to form trajectories of fixed length. These trajectories are subsequently described through descriptors. Notably, the IDT method offers an advantage in its ability to estimate camera motion by matching SURF descriptors (Bay et al., 2006) and dense optical flow feature points (Walker et al., 2015) between the previous and following frames, thereby mitigating the impact of camera movement. Following feature extraction, classification is performed using a Support Vector Machine (SVM). While in certain scenarios, these methods outperform deep learning-based approaches, they are limited by their slow processing speed and the necessity for accurate feature point tracking.

As a data-driven counterpart, a Convolutional Neural Network (CNN) based model was introduced in (Wang et al., 2015) which presents the Trajectory pooled Deep convolutional Descriptors (TDD) by integrating IDT features with two-stream depth features. That approach focuses on learning discriminative convolutional feature maps and utilizes trajectory-constrained pooling to aggregate the deep convolutional features into video descriptors. Differently, Shi et al. (Shi et al., 2017) propose a three-stream framework. These three streams, namely spatial, temporal, and sDTD streams are tailored to capture spatial, short-term, and long-term features, respectively. Specifically, the sDTD stream involves the extraction of simplified dense trajectories from each video, which are then transformed into a sequence of two-dimensional Trajectory Texture Images. These images are subsequently processed by a CNN + RNN (RNN stands for Recurrent Neural Network), enabling the learning of an efficient representation of long-term motion characteristics. In (Wang et al., 2018), Joint Trajectory Maps (JTM) were introduced as a technique to encode spatio-temporal information from 3D skeleton sequences into 2D images using color coding. These images are subsequently fed into CNNs to extract discriminative features for Human Action Recognition (HAR). In such a method, only adjacent joints within the convolution kernel are considered to learn co-occurrence features while potential correlations associated with all joints are not covered. Zhao et al. (Zhao et al., 2018a) propose a CNN architecture that explicitly predicts trajectories and integrates information along them using a convolution operation. Their design takes into account the changes in contents caused by motion or deformation. Specifically, they incorporate trajectory convolution into a Separable-3D ResNet18 architecture and employ either a variational method or an unsupervised method (MotionNet) to generate trajectory data from video. In contrast, the transformer architecture in (Patrick et al., 2021) does not explicitly predict trajectories, but provides an inductive bias that encourages the network to consider motion trajectories where useful. In detail, the study demonstrates that the joint attention mechanism for video transformers computes correlations between space and time. Furthermore, it elucidates how to guide the network in pooling information along motion paths.

In contrast to the methods discussed above, our approach relies on trajectories defined by the center points of detection bounding boxes over a fixed time segment, and we learn the feature representations by an unsupervised generative model. Similar to (Wang et al., 2011; Wang et al., 2018; Wang and Schmid, 2013; Zhao et al., 2018a), the detection bounding boxes are obtained as outputs of a standalone object tracker.

2.1.2. Unsupervised pre-training for action recognition

The literature reviewed above employs a fully supervised learning paradigm, wherein each sequence is manually annotated with the associated event for feature learning. As a recent alternative, unsupervised feature learning approaches (Ben Tanfous et al., 2018; Gui et al., 2018; Holden et al., 2015; Kundu et al., 2019; Li et al., 2018; Lin et al., 2020; Martinez et al., 2017; Nie et al., 2020; Paoletti et al., 2021d; Rao et al., 2021; Su et al., 2020; Xu et al., 2023; Zanfir et al., 2013; Zheng et al., 2018), particularly based on trajectories of human body pose, are progressively narrowing the performance gap with their fully supervised counterparts. These approaches could yield more compact, less noisy, and more transferable feature representations across various datasets for HAR (Paoletti et al., 2021a; Paoletti et al., 2021c; Paoletti et al., 2022b). These methods refrain from utilizing action/event labels during feature learning and typically rely on the encoder-decoder recurrent architectures (Kundu et al., 2019; Lin et al., 2020; Rao et al., 2021; Su et al., 2020; Zheng et al., 2018).

For instance, Zheng et al. (Zheng et al., 2018) introduce the method utilizing Generative Adversarial Networks (GANs) with Gated Recurrent Units (GRUs) to learn temporal representations of skeletal body poses. Additionally, an adversarial loss supports an auxiliary inpainting task, enhancing the learning process. Similarly, Lin et al. (Lin et al., 2020) employ GRUs to integrate contrastive learning, motion prediction, and jigsaw puzzle recognition techniques. Furthermore, Kundu et al. (Kundu et al., 2019) integrate a GAN-based encoder into their recurrent architecture. Xu et al. (Xu et al., 2023) enhance a vanilla autoencoder, trained to reconstruct skeletal data using mean-squared error (MSE) loss, by incorporating an ad-hoc training mechanism based on expectation maximization with learnable class prototypes. Su et al. (Su et al., 2020) utilize an encoder-decoder RNN to learn representations for HAR in an unsupervised manner from skeletal joints while addressing classification with a 1-nearest neighbor predictor. On the other hand, Rao et al. (Rao et al., 2021) merge contrastive learning with momentum Long Short-Term Memory (LSTM), contrasting the similarity between augmented instances and the input skeleton sequence before encoding long-term actions with a momentum-based LSTM. Similarly, Guo et al. (Guo et al., 2022) extend contrastive methods, demonstrating robust representations derived from extreme augmentations and novel movement patterns. Recently, Paoletti et al. (Paoletti et al., 2021a; Paoletti et al., 2022b) demonstrated the utilization of raw trajectories as input for a convolutional autoencoder during unsupervised pre-training. Subsequently, the latent features extracted from the autoencoder's encoder are utilized to train a Multilayer Perceptron (MLP) for action recognition. This simpler approach (Paoletti et al., 2021a; Paoletti et al., 2022b) demonstrated a better classification concerning the above methods while illustrating that features learned through unsupervised pre-training are transferable across different datasets and even tasks.

The most recent and top-performing studies in this domain often capitalize on the reconstruction capability of various generative models like GANs and autoencoders. Also, contrastive learning has been used to improve the similarity between the augmented and original instances. Differently, our proposed method employs diffusion models when processing 2D trajectories over a specific time frame. The aim is to utilize the inherent noise sampling, corruption, and reconstruction characteristics of diffusion models to acquire more effective feature representations. Subsequently, the latent features extracted from the encoder structure of the proposed diffusion model serve as input for a dense neural network, specifically an MLP, to facilitate the training of this

classifier and execute the inference process. Furthermore, motivated by the improved results in (Paoletti et al., 2021a; Paoletti et al., 2022b) concerning other studies discussed in this section, we adapted the autoencoder structure in these studies to compare its performance against the proposed method.

2.1.3. Fish behavior analysis in underwater videos

While the majority of studies focus on HAR, there is a growing interest in comprehending and analyzing animal behaviors. However, research on fish behavior using computer vision is relatively less explored compared to other vertebrates. A recent survey paper on fish detection and behavior analysis mentions that fish detection, species identification, and counting are topics that have received significantly more attention compared to fish behavior analysis (Yang et al., 2021). This disparity is largely attributed to the challenges inherent in collecting and analyzing data in underwater environments (Ditria et al., 2021), as well as the difficulty in annotating event classes in natural settings where the occurrence and conditions of events are unpredictable. Indeed, most research on fish behavior analysis has been conducted in controlled settings such as water tanks (Wang et al., 2021; Zhao et al., 2018b).

Appearance-based (e.g., color-based, oriented gradients) and motion-based (e.g., frame differencing, background subtraction) detection methods have been frequently employed in fish detection, followed by tracking methods to establish motion trajectories. Various features have been extracted from these trajectories, such as velocity, acceleration, turns, centroid distance function, local features, loop features, and displacement, and utilized for the analysis of individual fish behavior in underwater videos (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b). Within the Fish4Knowledge project (Boom et al., 2014; Fisher et al., 2016), three methods were developed for the damselfish *Dascyllus reticulatus*'s behavior recognition in unconstrained underwater videos, all emphasizing the distinction between abnormal (defined as rarely seen behavior classes) and normal behaviors. The first model (Beyan and Fisher, 2012) employs a filtering method that eliminates normal trajectories through a cascade of 21 rule-based filters, focusing solely on consecutive detections belonging to the same fish. The second method (Beyan and Fisher, 2013a) utilizes labeled and clustered data, applying outlier detection to clusters. The third method (Beyan and Fisher, 2013b) constructs a hierarchy using clustered and labeled data based on data similarity, employing different feature sets at various hierarchy levels.

The accuracy of detection and tracking techniques, particularly those based on appearance and motion analysis (i.e., without deep learning), plays a critical role in effective fish behavior analysis in underwater environments (Beyan et al., 2018; Beyan and Fisher, 2012; Beyan and Fisher, 2013a; Beyan and Fisher, 2013b; Palazzo et al., 2012). As noted in (Beyan et al., 2018), approximately 25% of the data may suffer from detection and tracking errors, rendering it impossible to ensure complete cleanliness of the entire dataset due to outliers from false detection and incorrect trajectory assignments. Beyan et al. (Beyan et al., 2018) employed an effective deep learning-based clustering algorithm based on mean-covariance restricted Boltzmann machines to clean noisy tracking data, encompassing a dataset of 4 million fish trajectories. Nevertheless, abnormal tracking trajectories are rare even in outlier detection, presenting an ongoing challenge in dealing with small sample sizes.

When it comes to deep learning-based methods, those utilizing trajectories still rely on hand-crafted features similar to those mentioned earlier. However, fish detection and tracking have been enhanced by adopting models like YOLOv3 and Mask RCNN (Hu et al., 2021; Lopez-Marcano et al., 2021). In contrast to trajectory-based methods, there have been a few attempts to analyze RGB video frames using deep neural network backbones (Ditria et al., 2021) combined with RNNs and variations (Måløy et al., 2019; McIntosh et al., 2020; Zhao et al., 2018c), or by applying 3D CNNs (Long et al., 2020; Måløy et al., 2019). Nonetheless, such studies are significantly limited in several aspects. For example, (Ditria et al., 2021) focuses solely on grazing behavior, (McIntosh et al., 2020) specializes in startle behavior, (Long et al., 2020) is tested only in laboratory environments, (Måløy et al., 2019) analyzes fish in sea cages with a focus only on feeding behavior, and (Zhao et al., 2018c) detects unusual behavior of fish school in aquaculture tanks.

To sum up, there is no study that focuses on multiple behavioral events by modeling raw trajectories and/or video frames, and being validated on natural underwater videos where it is not possible to stimulate the fish to capture specific behaviors. Therefore, our study, which addresses all these aspects, is unique and fills an important research gap. Furthermore, our study is the first to employ unsupervised pre-training of fish trajectories to obtain effective feature representations for fish event recognition. We also test the effectiveness of several deep models as well as mainstream trajectory features (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b).

2.2. Diffusion models

Diffusion models, a type of generative deep learning model, have demonstrated potential across a range of computer vision tasks, encompassing image generation (Liu et al., 2023), denoising (Saharia et al., 2022), and segmentation (Gu et al., 2024). Later on, these models have been extended to tackle discriminative tasks such as image classification (Yang and Wang, 2023), object detection (Chen et al., 2023), and anomaly detection in videos (Tur et al., 2023a; Tur et al., 2023b), showing very promising results.

Diffusion models operate by iteratively refining an input signal until it aligns with a target distribution. When applied to discriminative tasks, the input distribution represents noisy probabilities across class labels, while the conditioning signal usually comprises an image (Chen et al., 2023; Yang and Wang, 2023). This iterative process entails applying a sequence of transformations to the input signal, progressively bringing the output closer to the target distribution.

We are pioneering the application of diffusion models to the domain of fish event recognition, particularly in scenarios characterized by significant class imbalance in the data. Instead of utilizing images, our focus lies in extracting spatio-temporal features from the trajectories of individual fish. It is crucial to note that during the feature learning phase, we avoid incorporating any class labels associated with fish events. Consequently, the pre-training with diffusion models is carried out entirely in an unsupervised manner. Through unsupervised feature learning, we demonstrate the extraction of effective features at a certain stage, which are then utilized to recognize even relatively scarce event classes. Our approach demonstrates superior classification performance compared to using autoencoders replaced by diffusion models, as well as fully supervised methods: based on hand-crafted fish trajectory features (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b), and 1D-CNN.

2.3. Study species

The corkwing wrasse is an ecological and commercially important coastal fish, common in shallow waters of the North-east Atlantic (Halvorsen et al., 2016). This species has a complex life history and mating system, with most males developing into colorful nesting males that are strongly territorial and build seaweed nests during the spawning period in May–July (Potts, 1985). Several females visit and spawn in the nest, but only the male defends the nest and provides care for the eggs, where he performs a vigorous continuous movement of his fins and body close to the nest entrance. In addition, some males develop as sneaker males which closely resemble the females in morphology to sneak-fertilize eggs and, in that way, they avoid the struggle for territory acquisition and parental care. When the nesting male identifies a sneaker male, he will chase him off or block the entrance to his nest with his body. The nesting males may also chase off other species, such as the smaller goldsinny wrasse (*Ctenolabrus rupestris*) and conspecific females

and immature nesting males that may predate on eggs whenever he is away from the nest (Ellis, 2023). In this study, we analyze the trajectories of male corkwing wrasse associated with several event classes.

3. Dataset

The dataset utilized in this paper features observational recordings of 20 corkwing wrasse nests in Austevoll, Norway, during the summer months of June and July 2022. This is a large video dataset spanning several weeks of the spawning season of the focal species, and several different individual nesting males. Because of this, there is a natural variability in both the social and physical environment that could influence behavior variation between video samples, and we judge the videos to be representative of the behaviors in question for this species. The overarching purpose of the dataset is to assess the impact of consistent individual differences in agonistic and risk-taking behaviors on the parental care behaviors of nest-tending males. To assess the agonistic responses of these males (e.g., attack latency and frequency), a fish model depicting another male is briefly deployed within the nest. Thus, a model fish appears in some of the videos used in this paper. The videos were manually annotated by marine scientists to indicate the start and end times of events using the Behavioral Observation Research Interactive Software (BORIS) (Friard and Gamba, 2016).

3.1. Motion trajectories

The trajectories used in this study were acquired from the data collection mentioned above. We used the multi-object tracking method known as Tracktor (Bergmann et al., 2019) to obtain the trajectories of individuals. Tracktor (Bergmann et al., 2019) stands out because it transforms an object detector such as Faster-RCNN (Ren et al., 2015), equipped with a bounding box regressor, into a tracker. This enables the estimation of positions across frames through linear interpolation between consecutive boxes while ensuring identity consistency via association heuristics (see (Bergmann et al., 2019) for details). We favor this multi-object tracker due to our previous experience with the Faster R-CNN (Ren et al., 2015), which demonstrated effective performance in our earlier studies for fish detection in underwater videos (Allken et al., 2021; Knausgård et al., 2022). The Faster R-CNN was fine-tuned using manually annotated bounding boxes of male corkwing wrasse provided by marine experts, which are also used in the study (Knausgård et al., 2022). In each trajectory, the male corkwing wrasse is detected every 5 frames. The videos have a frame rate of 30 fps, resulting in 6 detections per second.

In this study, we define a trajectory as $Tra = (Tra_1, Tra_2, \ldots Tra_N)$ consisting of detections Tra_i composed by three coordinates: x_i and y_i, coordinates of the center of the bounding box of the tracked fish, and z_i, which is the ratio between the width and the height of the bounding box. This ratio serves to capture the orientation of the fish (i.e., a value of z_i closer to one can indicate that the fish's shape is more elongated horizontally, suggesting a horizontal orientation. Conversely, a value significantly different from one indicates a non-horizontal orientation.) and, in some cases, allows for an approximation of the depth information in the scene (e.g., if a fish swims closer to the camera, it will appear larger in the frame, leading to a decrease in the z_i ratio. Conversely, if the fish moves farther away, it will appear smaller, resulting in an increase in the z_i ratio.).In mathematical terms, we have a curve in a 3-dimensional space, parameterized by the time information.

We conducted data cleaning to eliminate events with incorrect detections (e.g., the cases in which severe occlusions happen and the tracker fails), trajectories with duplicated frames (wherein objects other than the targeted fish are mistakenly detected), or trajectories with insufficient number of detections (i.e., the ones having up to six detections only). We also applied trajectory interpolation to address instances of missing detections, which occur due to artifacts, rapid movement of the tracked fish, or occlusion by other objects in the scene. Interpolation aids in filling these gaps, resulting in more cohesive trajectories. We have experimented with different polynomial degrees of spline interpolation, such as linear, quadratic, and cubic, ultimately determining that linear spline interpolation was visually the most effective.

3.2. Description of events

Each trajectory consists of instances where a single male corkwing wrasse is detected in every five consecutive frames. Additionally, each trajectory is associated with a single label that represents a specific fish event. The description of each event is summarized as follows, with example images belonging to each category provided in Fig. 1. From an ecological standpoint, it is noteworthy that events such as spawning, egg-caring, antagonistic interactions, and nest maintenance represent common behavioral classes observed in the corkwing male nest during the breeding season. These behaviors are of particular interest to wrasse experts (Karaszkiewicz, 2020; Uglem and Rosenqvist, 2002) and also exist in our trajectory dataset.

Spawning and egg-caring events:

- *Spawning (S)* is characterized by the female and the nesting male corkwing wrasses following each other (with the female leading) and performing a distinct and rapid cycling movement at the nest entrance. Initially, the female deposits eggs, followed by the male fertilizing them. Often, several of these spawning cycles occur consecutively within a short timeframe (i.e., seconds).
- *Solo Spawning (SSP)* is the behavior where the nesting male corkwing wrasse spawns alone in the nest, likely to deposit more sperm to ensure a higher fertilization success of eggs laid.
- *Nest Blocking (NB)* occurs when the nesting male corkwing shuts down spawning (and sneaking) by blocking the nest entrance with its body for several seconds. It can be an adaption to reduce sneak fertilization (extra-pair paternity), at the cost of reduced spawning opportunities with females.
- *Fanning (FD)* involves the nesting male corkwing wrasse facing the nest and continuously moving its tail to aerate the eggs laid within. This behavior is crucial for the survival of the eggs.

Antagonistic events:

- *Egg Predation (EP)* refers to instances when the nesting male corkwing wrasse reacts to egg predators. Fish can prey on eggs in the nest when the male is away (common predators are female corkwing wrasse, goldsinny wrasse, and rock cook (i.e., *Centrolabrus exoletus*)), and when he returns, he chases off the intruders by swimming fast towards them and follows them out of the nest.
- *Chase (C)* events occur when the nesting male rapidly accelerates to chase other fish away from the nest. The difference from EP is that the male is present at the nest before the chase.
- *Goldsinny Chase (GC)* occurs when a goldsinny chases off other fish, typically corkwing sneaker males. There have been multiple observations of a goldsinny male persisting close to the nest, seemingly aiding the corkwing nesting male in defending the nest. In exchange, the goldsinny male is permitted to court goldsinny females near the nest. The data in this category pertains to the response of the male wrasse during GC events.

Nest maintenance events:

- *Foraging Maintenance (FM)* refers to instances where the nesting male corkwing wrasse performs maintenance on the nest by bringing in new nest material, i.e., algae, and incorporating it into the nest.
- *Non-Foraging Maintenance (NFM)* involves nest maintenance through the recycling or repositioning of nest material. Unlike foraging maintenance, in NFM, the male does not collect new algae.

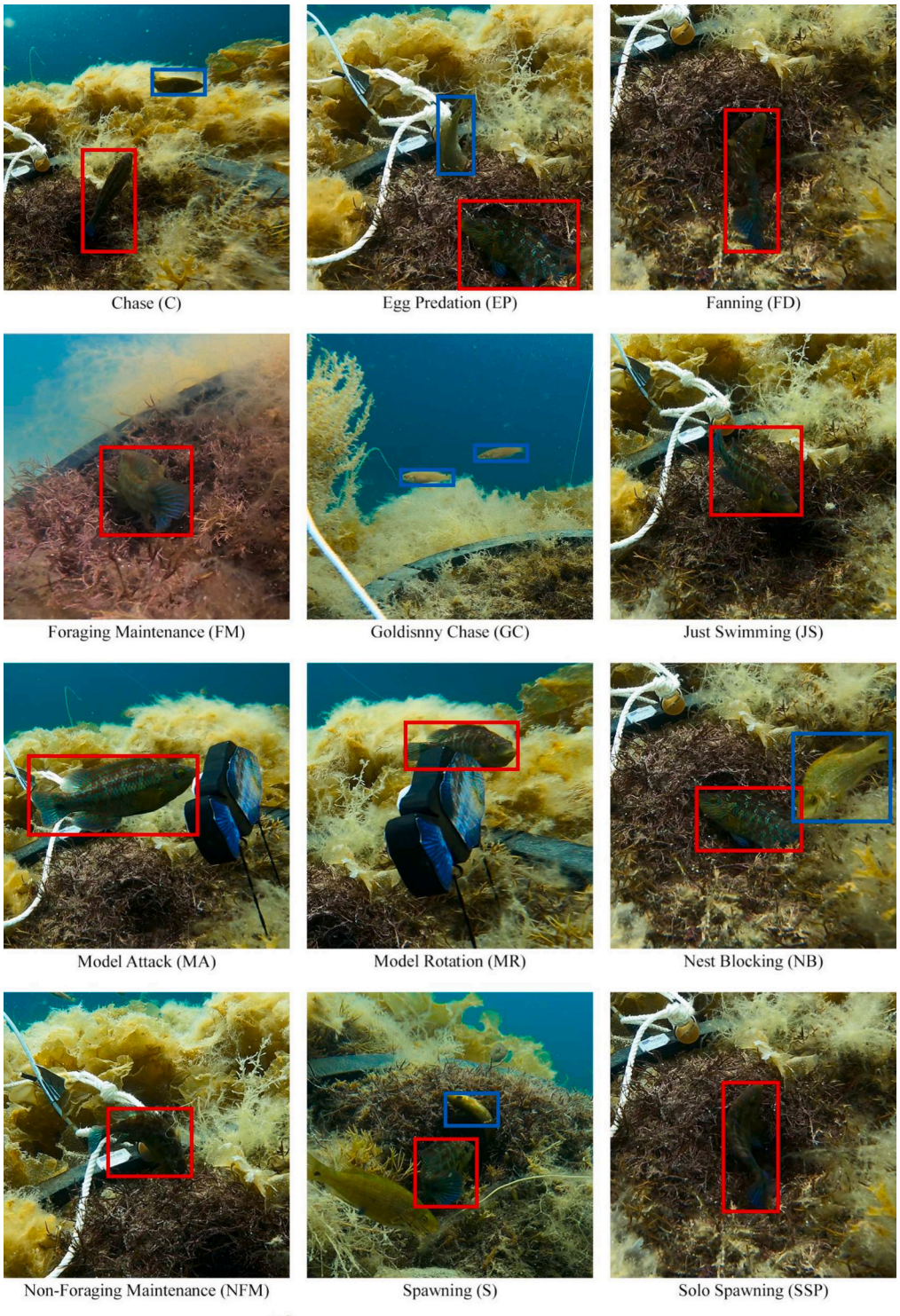


Fig. 1. Example frames from the dataset associated with each event. Red bounding boxes represent a male corkwing wrasse, while blue bounding boxes indicate a goldsinny. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Miscellaneous events:

- *Just Swimming (JS)* is considered as the default non-specific behavior, where the male corkwing simply swims around without engaging in any particular other behavior described above.

Experimental object response:

- *Model Attack (MA)* refers to the events occurring when the nesting male corkwing wrasse attacks a wooden model placed close by the nest. The wooden model imitates another nesting male corkwing wrasse to provoke an aggressive response. The model was introduced at the nest to quantify the consistent individual differences in agonistic behaviors between individual parental males, as part of a separate study. Using experimental objects such as models to quantify behavioral responses is a technique often used in the field of ecology as can be seen in (Rowland, 1999). Successful annotation of

experimentally provoked behaviors as well as those that passively occur would increase the versatility of the proposed model and therefore were included as a part of the training and testing splits.
- *Model Rotation (MR)* occurs when the nesting male corkwing wrasse rotates around the wooden model.

The final set comprises 16,937 trajectories, each with a length of 13 (refer to Section 4 for the rationale behind choosing a length of 13). The number of trajectories for each event category is presented in ascending order in Table 1. As one can observe, this distribution of event trajectories heavily favors certain classes, resulting in an imbalanced dataset. Specifically, the majority of the trajectories belong to the *Fanning (FD)* class, followed by other breeding season-related events such as *Non-Foraging Maintenance (NFM)*. As expected, *Just Swimming (JS)* also occurs frequently. Some classes have only a few samples, such as *Model Rotation (MR)*, *Goldsinny Chase (GC)*, and *Egg Predation (EP)*. The trajectories are available for download from https://github.com/NoeCanovi/Fish_Behaviors_Generative_Models.

4. Methodology & Implementation Details

The proposed approach consists of two stages. In the initial stage, unsupervised pre-training occurs in which features are acquired without reliance on the ground-truth labels associated with the task. Subsequently, in the second stage, recognition takes place, entailing the training of a classifier using the acquired features, followed by inference.

By utilizing trajectories represented as $Tra = (Tra_1, Tra_2, \ldots, Tra_N)$, where each detection Tra_i comprises three coordinates: x_i and y_i, denoting the center of the bounding box of the tracked fish, and z_i, indicating the ratio between the width and height of the bounding box, as inputs, we harness the reconstruction capability of diffusion models for unsupervised pre-training. By doing so, we aim to attain effective feature representations that are potentially less noisy. The learned features are subsequently utilized as input to an MLP for fish event recognition. It is essential to highlight that after the unsupervised training of the diffusion model, it remains frozen and detached from the event recognition component. An overview of the proposed method is depicted in Fig. 2. Below, we describe each component in detail, along with the implementation details, such as the encoder-decoder architectures and values of the hyperparameters used.

4.1. Fundamentals on diffusion models

Diffusion Models acquire complex data representations through a process of sequentially introducing noise and subsequently denoising the data. During the forward phase, Gaussian noise with standard deviation σ is iteratively added to a data point x_T, sampled from a distribution $p_{data}(x)$ with standard deviation σ_{data}, for each iteration $t \in [0, T]$ (T was taken as nine, in this study). The values of σ dictate the speed at which the noise is introduced such that higher values result in a faster conversion of the data into random noise. The distribution of the noised data is denoted as $p(x, \sigma)$, when the noise levels are $\sigma_0 = \sigma_{max} > \sigma_1 > \ldots > \sigma_{T-1} > \sigma_T = 0$. Consequently, we can sample a point $x_0 \sim N(0, \sigma_{max}\mathbf{I})$. Since the sum of Gaussian distributions remains Gaussian, we can directly compute the noisy version of a data point at a specific iteration t, without needing to calculate all previous versions. Thus, during training, a data point from a random step t can be sampled without iterating over all potential noisy versions of the same data point.

Table 1
The number of trajectories in the dataset for each event category.

Event	MR	GC	EP	SSP	S	FM
# of trajectory	4	8	13	23	31	48
Event	MA	C	NB	NFM	JS	FD
# of trajectory	67	174	198	895	1019	14,457

In the backward process, a denoising function $D_\theta(x; \sigma)$, implemented as a neural network, learns to predict the noise added to each data point. The diffusion model is then trained with Denoising Score Matching (Hyvarinen and Dayan, 2005), minimizing the expected L_2 denoising error (also called Mean Square Error (MSE)) for samples drawn from p_{data} for every σ:

$$E_{x\sim p_{data}} E_{\epsilon\sim\mathcal{N}(0,\sigma\mathbf{I})} \| D_\theta(x + \epsilon; \sigma) - x \|_2^2. \quad (1)$$

The score function used in the reverse process is:

$$\nabla log p(x; \sigma) = (D_\theta(x; \sigma) - x)/\sigma^2. \quad (2)$$

4.2. Our diffusion model

The diffusion model utilized in this study is largely based on the design outlined in (Tur et al., 2023a), which is a variant of *k-diffusion* presented in (Karras et al., 2022). The design of our model is given in Table 2.

As seen in Table 2, the model includes a σ-dependent skip connection, which allows the network to perform differently based on the noise magnitude. The denoising network D_θ mentioned in Section 4.1 is in terms of F_θ, which is the effective network to train, c_{skip} modulating the skip connection, c_{in} and c_{out} scaling input and output magnitude, and c_{noise} scales, which is formulated as:

$$D_\theta(x; \sigma) = c_{skip}(\sigma)x + c_{out}(\sigma)F_\theta(c_{in}(\sigma)x\,; c_{noise}(\sigma)) \quad (3)$$

Our F_θ is an MLP with an encoder-decoder structure such that the input is a set of noised data, composed of three channels x, y, and z, the layers of the encoder are progressively reduced ({1024, 512, 256}), while the number of channels is increased. Conversely, in the decoder layers, the length ({256, 512, 1024}) and the channels gradually return to their original sizes.

The learning rate scheduler and Exponential Moving Average (EMA) of our model are taken as the default values of *k-diffusion* (Karras et al., 2022). In both the encoder and decoder parts of the model, the time step σ is integrated through transformation via Fourier embedding and FiLM layers (Perez et al., 2018). Hence, the overall diffusion process for an input of the network X and its reconstructed counterpart X_r consists of *a)* noise sampling: $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, *b)* diffusion input corruption: $X_t = X + \epsilon^* \sigma$, and *c)* reconstruction of the data with k-diffusion: $X_r = \text{sampling}(D(X_t, \sigma))$.

4.3. Unsupervised pre-training with our diffusion model

During training, the diffusion model is provided with a dataset comprising trajectories of fixed length. Determining the appropriate length of these trajectories is a non-trivial task, as it can significantly impact the performance of the final model. To address this, we initially plotted the histogram of the original trajectory lengths and extracted statistics including the minimum, maximum, average, and median lengths. The analysis revealed that the majority of trajectories have a length of 13. Consequently, we established a standard trajectory length of 13 for our model. After segmenting the trajectories into fixed lengths of 13, we discard segments shorter than 6 detections. For segments with 6 to 12 detections, we replicate their data points until they reach the required length of 13. Finally, we scale each coordinate of the fish detection within the range of [−1,1].

Given that our model aims to learn to reconstruct these trajectories without relying on labels, the entire dataset is utilized following the representation learning literature such as (D'incà et al., 2023; Paoletti et al., 2021a; Paoletti et al., 2022b). The optimization of the network's parameters is achieved using the Adam optimizer (Kingma and Ba, 2015), coupled with an inverse decay learning rate scheduler. This scheduler initializes the learning rate at a default minimum value, which is zero. Then, it progressively increases the learning rate until it reaches

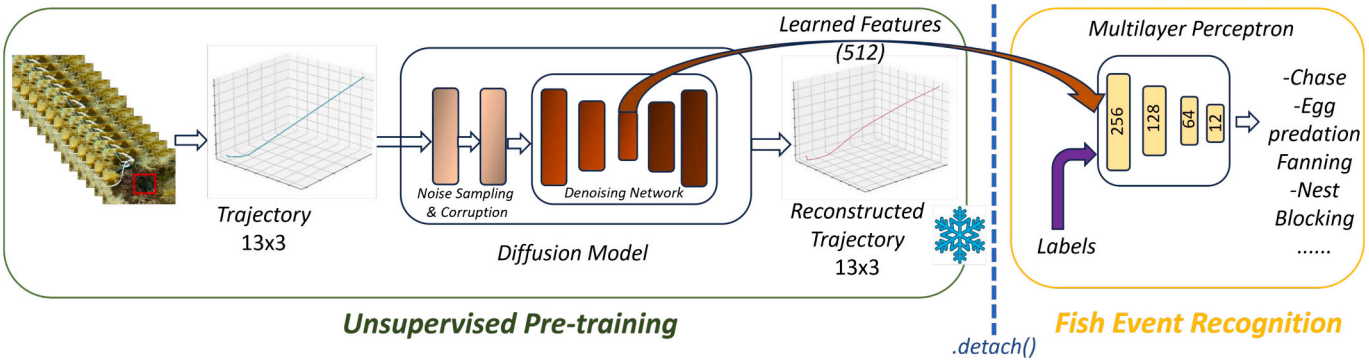


Fig. 2. The proposed method is divided into two stages. The first stage, known as *unsupervised pre-training* (feature learning phase), utilizes fish trajectories containing three dimensions over a fixed time frame: the center of the bounding box of the tracked fish, and the ratio between the width and height of the bounding box. Leveraging the diffusion model, we aim to generate a robust feature space. Once the feature learning process is completed, the learned features are extracted from the encoder of the diffusion model for use in the second stage, referred to as *fish event recognition*. During the training and inference of the second stage, the diffusion model remains frozen and detached. In this stage, the learned features act as input to a classifier (i.e., Multilayer Perceptron), enhancing its ability to distinguish between fish events.

Table 2
Design choices of our *k-diffusion*.

Sampling	
ODE solver	LMS
Time steps	$\sigma_{max}^{\frac{1}{\rho}} + \frac{i}{T-1}\left(\sigma_{min}^{\frac{1}{\rho}} - \sigma_{max}^{\frac{1}{\rho}}\right)\Big)^{\rho}$
Network and preconditioning	
Architecture of F_θ	MLP (see text)
Skip scaling $c_{skip}(\sigma)$	$\frac{\sigma_{data}^2}{\sigma^2+\sigma_{data}^2}$
Output scaling $c_{out}(\sigma)$	$\frac{\sigma\cdot\sigma_{data}}{\sqrt{\sigma^2+\sigma_{data}^2}}$
Input scaling $c_{in}(\sigma)$	$1/\sqrt{\sigma^2+\sigma_{data}^2}$
Noise scaling $c_{noise}(\sigma)$	$\frac{1}{4}ln(\sigma)$
Training	
Noise distribution	$ln(\sigma)\sim\mathcal{N}(P_{mean},P_{std}^2)$
Loss weighting	$\frac{(\sigma^2+\sigma_{data}^2)}{(\sigma\cdot\sigma_{data})^2}$

its maximum value at the last epoch of the training.

Noise plays an important role in the diffusion process, thus its distribution and parameters are to be set depending on the task and the dataset. Here, the noise is sampled from a log-normal distribution with parameters P_{mean} and P_{std}, representing the distribution's mean and standard deviation. These parameters are connected with the maximum and minimum σ values (σ_{max} and σ_{min}), through the following formula:

$$\sigma_{max}, \sigma_{min} = e^{P_{mean} \pm 5P_{std}}. \tag{4}$$

This relationship, as elucidated in (Karras et al., 2022; Tur et al., 2023a), is particularly valuable as it reduces the parameter search to two values instead of four.

Various combinations of P_{mean} and P_{std} values, along with typical hyperparameters of neural networks such as learning rate, batch size, and weight decay, were tested and explored. These values were assigned using random generators to ensure a more comprehensive evaluation of the model's performance. The parameter ranges considered are as follows: learning rate = [0.00001, 0.001], batch size = [64, 8192], weight decay = [0, 0.59], $P_{mean} = [-4, 1.8]$, and $P_{std} = [0.5, 1.68]$.

4.4. Recognition with a multilayer perceptron

After training the proposed diffusion model in an unsupervised pre-training fashion, where the data labels are not utilized, the generative model is frozen and detached and exclusively employed to extract features for both training and testing data, following the representation learning literature (D'incà et al., 2023; Franceschini et al., 2022; Koromilas and Giannakopoulos, 2021; Paoletti et al., 2021d; Paoletti et al., 2022a). These features are then used to train an MLP.

The MLP employed consists of four fully connected layers with the size of 256, 128, 64, and 12, each comprising a linear layer followed by PRelu as non-linearity. Only the third layer differs, as it is provided with Batch Normalization (Ioffe and Szegedy, 2015). During optimization, Batch Normalization was also applied to other layers as well, but the network performance was not as good. The network gets the input of size 512 consisting of learned features extracted from the diffusion model and the last layer has 12 neurons. The maximum neuron value from this layer is then used for predicting the event class, as each neuron corresponds to one of the events. The values of the 12 neurons are then compared to the real event ground truth through a loss, which serves to update the model parameters and perform the training.

During the training of the MLP, Adam optimizer (Kingma and Ba, 2015) is used, as well as the scheduler which adjusts the learning rate when the network reaches a plateau. For the training, various parameter values and different regularization techniques were explored, with the option of using either *Cross-Entropy* or *Focal Loss* (Lin et al., 2017). In detail, the learning rate was set to 0.00001 and 0.0001, batch sizes of 8, 16, 32, and 64 were tested, and weight decay was varied between 0.0001 and 0.001 with a dropout of 0.1. Additionally, Focal Loss (Lin et al., 2017) introduce additional parameters to investigate: α and γ. The former serves as a balancing factor, either as a fixed value for all classes or as the inverse of each class frequency. The latter regulates the impact of the scaling factor; specifically, when set to zero, Focal Loss is equivalent to Cross-Entropy Loss. We experimented with α set as the inverse of class frequency when γ took values of 0.5, 1, 2, 3, and 5. As demonstrated in several studies and ecological informatics (Xie et al., 2021), focal loss effectively handles imbalanced data compared to cross-entropy. This motivates our adaptation.

5. Experimental Analysis & Results

The methodologies used for comparing our proposed method are given in Section 5.1, and we present the corresponding results in Section 5.2. As the evaluation metrics we follow (Beyan and Fisher, 2015; Haixiang et al., 2017; Luque et al., 2019), showing that F1-score and the geometric mean of true positives and true negatives (denoted as G-mean) are suitable evaluation metrics for imbalanced data classification. We report the F1-score and G-mean results for Macro (i.e., calculating

metrics for each class, and finding their unweighted mean, meaning that the class label imbalance is not considered) and Weighted averages (i.e., calculating metrics for each class, and finding their average, weighted by the number of true instances for each label, thus considering the class label imbalance) in Eqs. 5–10. We also present the individual class accuracies to demonstrate the detection performance of the proposed method for each event class.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{Macro F1} = \frac{1}{n} \sum_{i=1}^{n} F1_i \tag{6}$$

$$\text{Weighted F1} = \frac{\sum_{i=1}^{n} \text{TrueInstances}_i \cdot F1_i}{\sum_{i=1}^{n} \text{TrueInstances}_i} \tag{7}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{8}$$

$$\text{Macro G-mean} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\text{Sensitivity}_i \times \text{Specificity}_i} \tag{9}$$

$$\text{Weighted G-mean} = \frac{\sum_{i=1}^{n} \text{TrueInstances}_i \cdot \sqrt{\text{Sensitivity}_i \times \text{Specificity}_i}}{\sum_{i=1}^{n} \text{TrueInstances}_i} \tag{10}$$

where Precision $= TP/(TP + FP)$, Recall $= TP/(TP + FN)$, Sensitivity $= TP/(TP + FN)$, Specificity $= TN/(TN + FP)$, $F1_i$ is the F1-score for class i, TrueInstances_i is the total number of true instances $(TP + FN)$ for class i, $\text{Sensitivity}_i = TP_i/(TP_i + FN_i)$ is the sensitivity for class i, $\text{Specificity}_i = TN_i/(TN_i + FP_i)$ is the specificity for class i, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, and n is the number of classes.

During the training and validation of the models, a stratified train-test split was employed. This procedure is particularly beneficial for imbalanced datasets as it maintains the proportion, in our case 80%–20%, for each class of the original dataset. This means that every class have approximately 80% of its samples in the train set and approximately 20% in the validation set. This ensures that all classes participate in both the training and validation phases of the model. However, this stratified train-test split was not applied to the dataset composed of trajectories of length 13, but rather to the original trajectories of variable length. This was done to ensure that each trajectory belongs exclusively to either the train or validation splits, but not both. As a consequence, in terms of fixed-length trajectories, each with a length of 13, we obtained 13,489 samples in training and 3448 samples in validation.

5.1. Methods employed for comparisons

We adopted the following methods to compare against the proposed method. Their implementation details are given as follows.

5.1.1. Trajectory Features of [40, 41]

As discussed in Section 2, it is notable that there is a scarcity of studies delving into the understanding of fish trajectories through the analysis of underwater videos. We adhere to the methodologies outlined in the existing works (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b), which define several hand-crafted features extracted from fish trajectories. Building upon the findings reported in (Beyan and Fisher, 2015), we incorporate these features without feature selection into Support Vector Machines (SVM), showing the best performance for imbalanced data classification. It is important to note that the hierarchical framework introduced in (Beyan and Fisher, 2013b) for detecting unusual trajectories is not directly adaptable to our study, as it assumes binary classes. The features that we adopted include Curvature Scale Space (CSS), Moment Descriptors, velocity and acceleration, turn, Centroid Distance Function (CDF), and vicinity. However, features such as loop, fish pass-by, and displacement on the location are deemed unsuitable for the dataset under examination. SVM was applied with the radial basis kernel function (RBF) with the kernel parameters set as $C = 2^i$, where $i = -1, 1, 3, \ldots, 31$, and the RBF was used with $\gamma = 2^j$, where $j = -11, -9, -7, \ldots, 11$. If it is worth mentioning, we also attempted to use linear SVM and our MLP; however, the results obtained were poorer, indicating that such classifiers are insufficient for these hand-crafted features in the dataset under consideration.

5.1.2. 1D-CNN

One-dimensional convolutional neural networks are particularly well-suited for processing sequential data (Kiranyaz et al., 2019; Kiranyaz et al., 2021; Sujatha et al., 2023; Troullinou et al., 2020), including trajectories (Hsieh et al., 2021; Zamboni et al., 2022), as well as being applied for action / event recognition (Cho and Yoon, 2018; Escottá et al., 2022; Hosseini et al., 2020; Javidani and Mahmoudi-Aznaveh, 2022; Trelinski and Kwolek, 2021) and imbalanced data classification (Alex et al., 2024; Eren, 2017; Mattioli et al., 2022; Qazi et al., 2022; Sujatha et al., 2023). Their advantages compared to LSTM and GRUs occur especially in scenarios where efficiency, scalability, and processing of fixed-length relatively lower-size data is considered (Kiranyaz et al., 2019; Kiranyaz et al., 2021). Our implementation aligns with the architectural design of a trajectory-based approach (Hsieh et al., 2021), as well as a recent paper in ecological informatics (Sujatha et al., 2023), which has input data sizes similar to ours while handling an imbalanced dataset.

The input consists of the concatenation of three channels: x, y, and z. Before concatanation zero-padding is applied to isolate the x, y, and z coordinates, preventing interference during the convolution operation as applied in (Hsieh et al., 2021). Each convolutional block includes convolution, max pooling, and batch normalization, applied sequentially. ReLu serves as the activation function. Additionally, multiple dense layers are present, with dropout utilized between them. The output dense layer employs softmax activation. The network is trained with Cross-Entropy and Focal Loss, with Adam optimizer (Kingma and Ba, 2015) in line with the proposed method. Different architectural configurations were evaluated, varying the number of filters from four, eight, 16, and 32, adjusting the filter size from three to seven, and experimenting with the number of convolutional and dense layers each ranging from two to four. The learning rate, batch size, weight decay, and dropout were set on par with the MLP of the proposed method.

5.1.3. Autoencoder

Numerous studies (D'incà et al., 2023; Koromilas and Giannakopoulos, 2021; Paoletti et al., 2021a; Paoletti et al., 2021d; Paoletti et al., 2022a) have showcased the effectiveness of autoencoders in unsupervised pre-training, where the acquired features are subsequently leveraged for training and testing classifiers. In this study, a convolutional autoencoder inspired by (Paoletti et al., 2021a; Paoletti et al., 2021d) was utilized, tailored to our dataset. The network's input comprises trajectories of length 13, encompassing three channels. With each encoder layer, the input length decreases by half while the number of channels increases. Conversely, in the decoder layers, the reverse occurs. Both encoder and decoder blocks consist of two layers, with the latent representation naturally positioned between them. The best-performing autoencoder processes inputs of size 13×3, with the encoder layer comprising layers of sizes 6×64 and 3×128, respectively. Conversely, the decoder is composed of layers sized 3×128 and 6×64. During the training of this model, the network weights are iteratively updated using the Adam optimizer (Kingma and Ba, 2015). Additionally, a scheduler adjusts the learning rate: when the loss plateaus, meaning it does not decrease over a certain number of epochs, the learning rate is reduced.

Various parameter values, including learning rate (i.e., 0.0001, 0.001, 0.01, and 0.1), batch size (i.e., 32, 64, 128, 256, and 512), and latent dimension (i.e., 128, 256, 512, 2048), have been explored and evaluated.

Once the latent representations were extracted from the Autoencoder model, an MLP was trained for the classification of fish events. Throughout the training of MLP, Adam optimizer (Kingma and Ba, 2015) with the learning rate modified by a scheduler when the network reaches a plateau was used. The MLP was trained for several epochs with various parameter configurations and regularization techniques, including both Cross-Entropy and Focal Loss, in line with the proposed method. As regularization techniques, weight decay, batch normalization (Ioffe and Szegedy, 2015), and dropout (Srivastava et al., 2014) were explored. In particular, batch normalization (Ioffe and Szegedy, 2015) was tested on each layer of the network separately and on all layers simultaneously, with the best results achieved when applied to the third layer. Dropout has been applied to the input layer, the hidden layers, and all layers, yielding various results in combination with other parameters. The parameter values explored alongside the aforementioned autoencoder include the following: learning rates of 0.00005, 0.0005, 0.0001, 0.005, 0.001, and 0.01; batch sizes of 8, 16, 32, 64, and 128; weight decay values of 0, 0.00001, 0.0001, and 0.001, with dropout set to 0.05, 0.1, 0.2, and 0.25. For the focal loss parameters, α was defined as the inverse of class frequency, 0.25, or 0.5, while γ was set to 0.25, 0.5, 0.75, 1, 1.5, and 2.

5.1.4. SMOTE [131]

Imbalanced datasets pose challenges as the dominant classes heavily influence training, while minority classes have minimal impact. Addressing the issue of imbalanced class distribution can involve undersampling the majority classes and/or oversampling the minority classes. Undersampling entails randomly removing samples from the majority classes while oversampling involves duplicating samples from the minority classes (Beyan and Fisher, 2015; Galar et al., 2011; Kubat and Matwin, 1997). A more sophisticated and widely used approach is the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). SMOTE augments the dataset by generating synthetic samples based on the feature space and combining features from neighboring samples. SMOTE has been frequently applied in ecological informatics studies, where imbalanced data is common, demonstrating its effectiveness such as in (Bourel et al., 2021; Shin et al., 2021). In this paper, we used SMOTE both for the proposed method and also for the methods used in comparisons. When SMOTE was integrated with 1D-CNN, it was applied to flattened data, which includes zero-padding. For other methods used in conjunction with SMOTE, it should be noted that the features are already one-dimensional.

5.2. Results: Comparisons with other methods

Table 3 presents the results in terms of Macro F1-score, Weighted F1-score, Macro G-mean, and Weighted G-mean. The key distinction between macro and weighted lies in their treatment of class imbalances. Macro treats all classes equally, while weighted assigns more weight to larger classes. As a result, the numerical values for weighted metrics are notably higher than those for macro. Nevertheless, the proposed method outperforms others across all metrics. Its combination with SMOTE yields the best results overall, while its combination with *Focal Loss* comes in as the runner-up.

The results of the 1D-CNN, autoencoder, and the proposed method reveal that the models trained with plain *Cross-Entropy Loss* consistently achieve a lower or comparable performance in every metric. This aligns with expectations, as both *Focal Loss* and *SMOTE* are specifically designed to address imbalanced datasets. The models exploiting *SMOTE* obtain the best performance across all metrics. On the other hand, the features from (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b) combined with SVM yield the lowest results across all combinations.

Table 3
The best results of each method. CE, FL, F1, and G-mean denote cross-entropy loss, focal loss (Lin et al., 2017), F1-score, and the geometric mean of true positives and negatives, respectively. Proposed can also refer to "Diffusion + MLP". The best of all results for each metric are given in bold while the second best is presented underlined.

Method	Macro F1	Weighted F1	Macro G-mean	Weigthed G-mean
Features (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b) w/ SVM	27.7	79.8	51.2	73.8
Features (Beyan and Fisher, 2013a; Beyan and Fisher, 2013b) w/ SVM + SMOTE	28.0	80.1	51.8	74.6
1D-CNN w/ CE	35.1	81.5	54.7	75.3
1D-CNN w/ FL	36.4	83.7	56.6	75.7
1D-CNN w/ CE + SMOTE	36.8	83.8	57.7	78.9
Autoencoder + MLP w/ CE	38.3	87.3	61.0	79.4
Autoencoder + MLP w/ FL	44.5	87.3	64.6	81.4
Autoencoder + MLP w/ CE + SMOTE	45.8	87.3	66.4	83.2
Proposed w/ CE	45.5	89.2	64.5	80.4
Proposed w/ FL	<u>47.8</u>	<u>89.9</u>	<u>66.6</u>	<u>83.0</u>
Proposed w/ CE + SMOTE	**50.6**	**90.7**	**68.1**	**83.6**

This suggests that features learned from raw trajectories (especially through unsupervised pre-training) can be preferable to hand-crafted features.

For the diffusion model, the starting point of the reverse process t has a remarkable effect on the training and the performance of the classifier (Karras et al., 2022; Tur et al., 2023a) (see also the corresponding results in Section 5.2.1). The ideal choice for t is the one that retains the most relevant information for the classifier. Herein, we found out that this value corresponds to 4 or 5. The results indicate that utilizing *SMOTE* leads to improved performance, with the best results achieved on the test set created with t equal to 4 when other parameters are set as follows: batch size equal to 16, learning rate equal to 0.00005, and weight decay set to 0.0001.

In all models, there is a noticeable discrepancy in performance among behavioral classes. Specifically, the two best-performing methods, the autoencoder-based model and the proposed method yield the following results. The autoencoder-based model classifies well *Fanning (FD)* and *Model Attack (MA)*, with class accuracies reaching 97% and 86% (or 71%) respectively. On the other hand, there are considerable challenges in recognizing certain classes, particularly *Goldsinny Chase (GC)* and *Model Rotation (MR)*, which have a class accuracy of zero. The *Focal Loss*-trained autoencoder-based model stands out as the only one with sufficiently good class accuracy on *Egg Predation (EP)*, and it also exhibits a slightly better class accuracy on *Chase (C)*. Conversely, the *Cross-Entropy Loss* and *SMOTE*-enhanced autoencoder-based model performs well on *Solo Spawning (SSP)*, and slightly outperforms in *Foraging Maintenance (FM)*, *Foraging Maintenance (NFM)*, and *Spawning (S)*.

For the proposed method, the majority class *Fanning (FD)* has a remarkably good class accuracy with the lowest value being 96%. In contrast, *Goldsinny Chase (GC)* and *Model Rotation (MR)* are consistently unrecognized by the model, resulting in a constant class accuracy of zero. These are also the classes with the lowest amount of samples, four and eight respectively. Additionally, *Chase (C)* and *Model Attack (MA)* achieve the best class accuracy with *Focal Loss*. However, this model struggles to classify *Egg Predation (EP)* and *Foraging Maintenance (FM)* with scores of zero and 10.0%, respectively. On the other hand, *Egg Predation (EP)* and *Foraging Maintenance (FM)* reach a class accuracy of 50.0% and 30.0% with the model trained with *SMOTE*. In general, *Spawning (S)* and *Solo Spawning (SSP)*'s class accuracy remains consistent across all three models with scores of 83% and 50%, *Fanning (FD)* and *Just Swimming (JS)*'s class accuracies are only subject to small variations. The reader can refer to Table 4 and Fig. 3 for the individual class

Table 4
Class accuracies obtained for the proposed method and corresponding number of samples per class. The best results out of the three settings are given in bold.

Event	Cross-Entropy Loss	Focal Loss	Cross-Entropy + *SMOTE*
C	19.0	**58.0**	42.0
EP	**50.0**	00.0	**50.0**
FD	97.0	96.0	**98.0**
FM	10.0	10.0	**30.0**
GC	00.0	00.0	00.0
JS	59.0	**62.0**	60.0
MA	50.0	**79.0**	71.0
MR	00.0	00.0	00.0
NB	64.0	**65.0**	45.0
NFM	33.0	**43.0**	**43.0**
S	**83.0**	**83.0**	**83.0**
SSP	**50.0**	**50.0**	**50.0**

accuracies as well as the confusion matrices of the proposed method.

The confusion matrices of the proposed method in the best-performing setting, i.e., Cross-Entropy Loss with *SMOTE*, reveal that several event classes were misclassified as *Just Swimming (JS)*, *Fanning (FD)*, or *Non Foraging Maintenance (NFM)*, which are among the three most dominant classes. In detail, *Chase (C)*, *Model Attack (MA)*, and *Spawning (S)* were mainly misclassified as *Just Swimming (JS)*. *Goldsinny Chase (GC)*, *Model Rotation (MR)*, and *Foraging Maintenance (FM)* were primarily misclassified as *Non Foraging Maintenance (NFM)*. *Egg Predation (EP)*, *Nest Blocking (NB)*, *Non Foraging Maintenance (NFM)*, and *Just Swimming (JS)* were mainly misclassified as *Fanning (FD)*.

5.2.1. Results: Trajectory reconstruction

While it is true that our diffusion and autoencoder models do not know which sample belongs to which class, there may be some classes that are implicitly simpler to reconstruct. Additionally, a model could be learning those samples belonging to the most popular classes more effectively. Consequently, we also conduct an analysis of the trajectory reconstruction error for the proposed diffusion model and the proposed autoencoder. This involves assessing the networks' loss (in terms of MSE) and comparing the reconstructed data with the original data to measure their similarity.

The reverse process of a diffusion model is not required to start from the maximum level of noise variance σ^2_{max}. It can start from any level of noise, with any arbitrary step $t \in [0, T]$, where a t close to zero indicates a noised x_t closer to isotropic Gaussian, i.e., $\sigma^2_{max} = \sigma^2_0$, while t closer to T means a noised x_t closer to the original data distribution. Depending on the value of t, the network can yield various loss values and reconstructed trajectories. Furthermore, as the noise is initialized through $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and different values are sampled each time, the same network can produce distinct loss values and trajectories upon multiple runs of the evaluation process. Thus, for a fair comparison between diffusion models and across time steps t, the noise is initialized once and then maintained fixed. The loss values achieved by the best-performing model for each time step t from zero to nine are 0.002152, 0.000603, 0.000306, 0.000176, 0.000099, 0.000058, 0.000033, 0.000014, 0.000004, and 0.000001, respectively. This progression demonstrates that as t increases and the noise decreases, the error tends to assume lower values, and the reconstructed trajectories more closely resemble the original ones.

Fig. 4 illustrates the reconstruction of the same trajectory with different time steps, showing the effect of t on the reconstruction capabilities of our diffusion model. In addition, Fig. 5 demonstrates trajectories belonging to 12 event classes and their reconstructed counterparts, extracted when $t = 9$. Notably, the reconstructed trajectories exhibit a high fidelity to the original ones at this specific time step.

Additionally, Table 5 provides the MSE for each event class at various time steps t. Notably, the differences between the classes are minimal and affected by the time step t. At lower values of t, classes such as *Fanning (FD)*, *Nest Blocking (NB)*, and *Egg Predation (EP)* appear to have comparatively lower MSE. However, as t increases, this distinction diminishes.

On the other hand, we observed that when using the proposed autoencoder, the *Fanning (FD)* and *Nest Blocking (NB)* events had the lowest overall MSE value, equal to 0.00008. The event *Model Attack (MA)* yielded the highest MSE value, equal to 0.00044.

6. Discussions

Cameras are increasingly being used to monitor marine fauna, but in most cases, only a fraction of the data is utilized due to the limitations of the manual processing of videos. In practice, this means that video data containing diverse and complex ecological information tend to be distilled to counts and size of a few species of interest, while behavior information is rarely collected unless it is the primary objective of the study (Goodwin et al., 2022; Weinstein, 2017). In addition to automating the detection and counting of fish (Catalán et al., 2023; Knausgård et al., 2022), developing effective and precise machine learning methods for classifying the behavior of key species can produce data that opens new avenues to deeper understanding of marine ecosystems and the behaviors that link the different components together (Aguzzi et al., 2015; Ditria et al., 2020). In the light of accelerating environmental

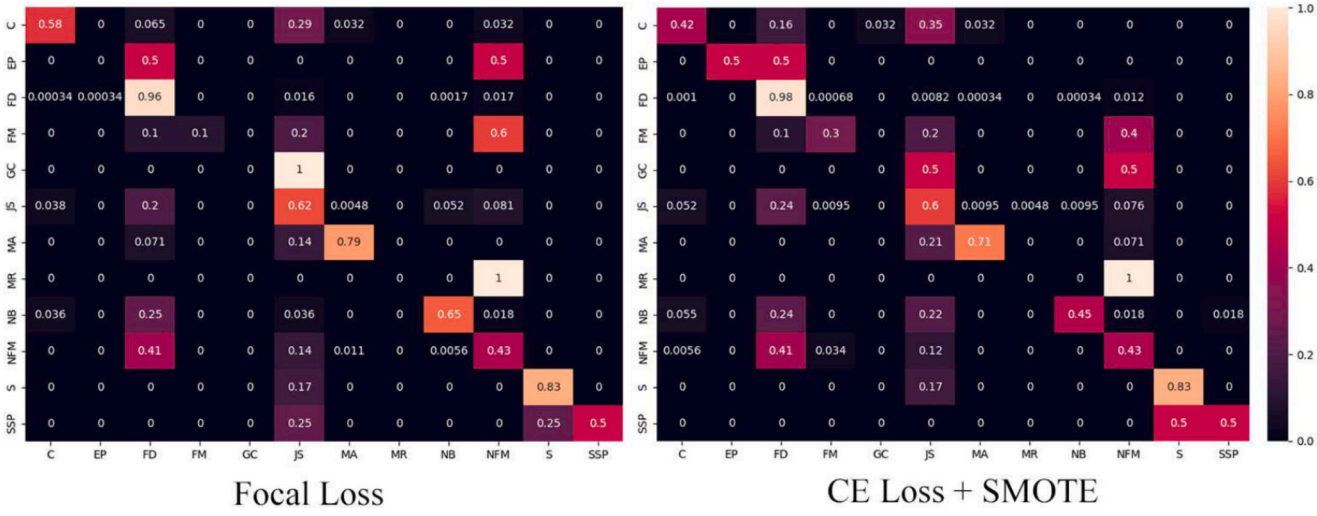


Fig. 3. Confusion Matrices of the Proposed Method with the two best performing configurations: Focal Loss and Cross-Entropy Loss with *SMOTE*.

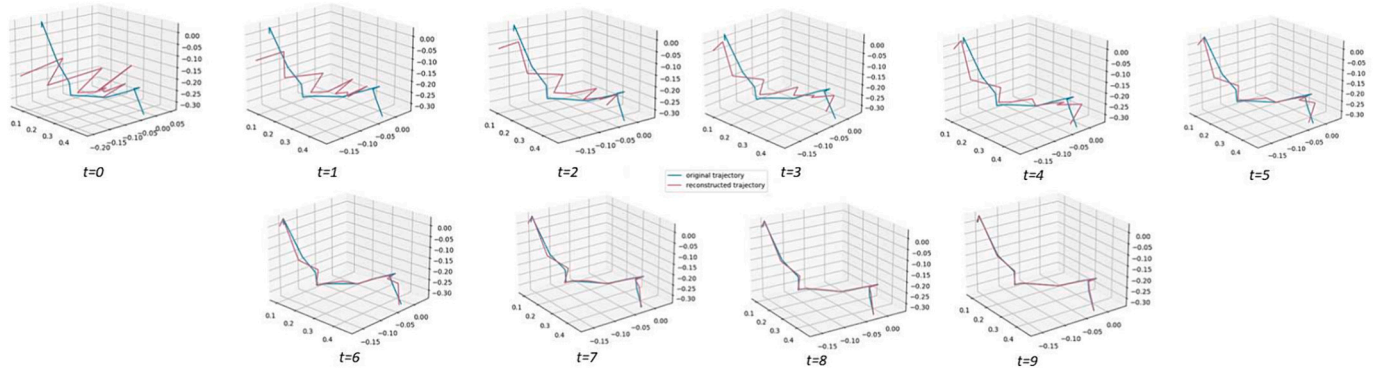


Fig. 4. Original and respective reconstructed trajectories at various time steps (t) for the event class *Model Rotation (MR)*.

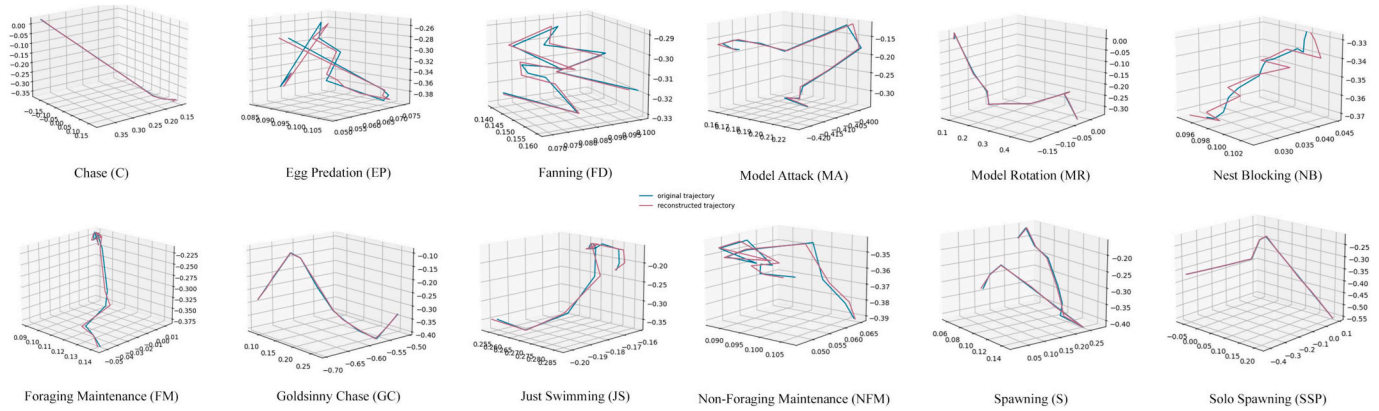


Fig. 5. Original and corresponding reconstructed trajectories, randomly selected from each class.

Table 5
MSE for each event class during the training of the proposed diffusion model.

t =	0	1	2	3	4	5	6	7	8	9
C	0.004	0.001	0.0009	0.0004	0.0001	9e-5	4e-5	1e-5	4e-6	1e-6
EP	0.001	0.0008	0.0005	0.0002	0.0001	8e-5	4e-5	1e-5	4e-6	1e-6
FD	0.001	0.0004	0.0002	0.0001	8e-5	5e-5	3e-5	1e-5	4e-6	1e-6
FM	0.002	0.001	0.0005	0.0001	0.0003	8e-5	4e-5	1e-5	4e-6	1e-6
GC	0.003	0.001	0.0006	0.0003	0.0001	8e-5	3e-5	1e-5	4e-6	1e-6
JS	0.003	0.001	0.0007	0.0003	0.0001	7e-5	3e-5	1e-5	4e-6	1e-6
MA	0.006	0.002	0.001	0.0005	0.002	0.0001	4e-5	1e-5	3e-6	9e-7
MR	0.007	0.003	0.001	0.006	0.003	0.001	5e-5	1e-5	4e-6	1e-6
NB	0.001	0.0005	0.0003	0.0001	9e-5	5e-5	3e-5	1e-5	4e-6	1e-6
NFM	0.002	0.0009	0.0004	0.0002	0.0001	7e-5	3e-5	1e-5	4e-6	1e-6
S	0.002	0.001	0.001	0.0007	0.0002	0.0001	4e-5	1e-5	4e-6	1e-6
SSP	0.002	0.001	0.001	0.0002	0.0006	0.0001	4e-5	1e-5	4e-6	1e-6

change, understanding species' behavior responses becomes even more important, as many species cannot adapt to rapid changes through evolution, but must rather rely on adapting their behavior (Hoffmann and Sgrò, 2011; Wong and Candolin, 2015) For example, North Atlantic right whales have been observed adjusting their calling behavior in response to increased background noise (Parks et al., 2011), birds have modified their flight distances in response to speed limits to reduce collision risks with vehicles (Legagneux and Ducatez, 2013), and certain marine organisms have migrated poleward, leading to new species interactions (Poloczanska et al., 2013). However, not all organisms are capable of adapting their biology to changing environments. With current threats disrupting ecosystems, it is crucial to detect and monitor changes early to understand which variables trigger specific behaviors and anticipate their effects. Given the complexity and often enormous spatiotemporal variation in such natural phenomena, the only way to monitor this behavior on large scales is through increased use of cameras and other passive sensors (e.g. hydrophones) combined with machine learning tools for automated data processing.

6.1. Broad impact of the proposed method

The method presented in this paper represents an important step towards automating the collection of behavioral data from video surveys and observatories. The traditional procedure, which lacks automation, is highly time-consuming and requires expert knowledge of species behavior, thus limiting opportunities for scaling up and expanding studies. The proposed method holds promise in addressing the challenge of analyzing data-intensive video surveys, as evidenced by its effective results, particularly for certain event classes. The data under examination is particularly challenging due to its highly imbalanced nature,

where some event classes occur very infrequently. This characteristic resembles scenarios encountered in few-shot learning, which is a prominent research topic in the machine/deep learning field. Moreover, the proposed method offers a new tool for marine biologists to monitor behavioral changes in response to environmental shifts. For instance, video monitoring can help detect the onset and duration of spawning periods, crucial for managing fisheries with temporal closures during reproductive phases (Halvorsen et al., n.d.; Halvorsen et al., 2017b).

Focusing on trajectory-based generative model pre-training, it seems the use of diffusion models shows superior results compared to the use of autoencoders. In particular, diffusion models achieve better results in *F1-score macro* and *Geometric Mean macro*, meaning they can correctly classify samples from a greater range of classes. Regardless of the method, both *SMOTE* and *Focal Loss* prove their effectiveness in addressing imbalanced scenarios, outperforming the plain *Cross-Entropy Loss* model. Notably, *SMOTE* emerges as the more efficient technique here. Overall, our method is a simple yet effective approach that can be applied to behavior analysis of other marine species or broader ecological studies. Since the feature learning part is unsupervised, it has the potential to generalize well to other event classes, including those outside the scope of the breeding season or other fish species. Additionally, the same architecture, with minor modifications to the encoder of the diffusion model and/or by using a pre-trained model to extract initial features, can be adapted to inputs in the form of video frames. This flexibility allows for future studies to discard the use of trajectories if desired.

6.2. Analysis of class performance and dataset imbalance

At the class level, disparities in performance are influenced in part by the imbalance within the dataset. Indeed, imbalanced datasets present challenges, as models tend to be biased towards the majority class, resulting in difficulties in accurately recognizing and classifying instances from the minority classes. As such, classes with consistently poor performance, such as *Goldsinny Chase (GC)* and *Model Rotation (MR)*, have low sample sizes: eight and four, respectively. The limited number of samples may hinder the network's ability to discern patterns unique to these classes, assimilating them with more prevalent classes. While *SMOTE* was employed to augment these classes, it may still be insufficient. Conversely, *Fanning (FD)* consistently achieves higher class accuracy, being the class with the highest number of samples in the whole dataset. The abundance of *Fanning (FD)* trajectories gives the model plenty of chances to learn and discover patterns to recognize the class. The class with the second-largest sample size *Just Swimming (JS)* (with 10% of that of the samples in *Fanning (FD)*), was often incorrectly predicted to be *Fanning (FD)*. This may be possibly due to the fact *Just Swimming (JS)* is the class derived from the information of the male present in the scene when it is not performing any behavior. Trajectories extracted for certain classes of events may overlap to some extent, as some segments of an event might inherently share characteristics with *Just Swimming*.

6.3. Challenges in trajectory-based behavior classification

The semantic similarities between certain behaviors, such as *Spawning (S)* and *Solo Spawning (SSP)*, as well as *Foraging Maintenance (FM)* and *Non-Foraging Maintenance (NFM)*, poses challenges for trajectory-based classification. These behaviors, while functionally distinct, have subtle differences in trajectory level that may be difficult to discern. *Spawning (S)* and *Solo Spawning (SSP)* both represent spawning behavior. In the former, the male and the female corkwing wrasse take turns in spawning, while in the latter the male fish conduct the behavior by himself. The trajectories extracted from the male fish motion may not be sufficient to determine which of the two behaviors is performed. Indeed, this is reflected in the confusion matrices, where *Solo Spawning (SSP)* trajectories are occasionally misclassified as *Spawning (S)*. If the presence of another individual could be taken into account, the two classes would probably be separated more clearly. Similarly, the distinction between *Foraging Maintenance (FM)* and *Non-Foraging Maintenance (NFM)*, both maintenance behaviors, involves subtle differences in the male fish's interactions with the nest. The trajectories alone may not provide clear indications of whether material is added to the nest (*Foraging Maintenance (FM)*) or if the fish is simply poking the nest (*Non-Foraging Maintenance (NFM)*). Between the two, *Foraging Maintenance (FM)* suffers the most as it has a lower sample size, being misclassified as *Non-Foraging Maintenance (NFM)* around half of the time. In addition, *Egg Predation (EP)* and *Goldsinny Chase (GC)* prove to be particularly challenging to classify. For instance, in the case of *Egg Predation (EP)*, males often attempt to protect the nest, resulting in trajectories overlapping with the events *Nest Blocking (NB)* or *Chasing (C)* the intruders. As the trajectory of the nesting male is the only information exploited by the model, it is inherently difficult to perform well on such classes. Addressing these cases may involve incorporating in the pipeline additional context, for example, the appearance features that exist in video clips. Such should be especially beneficial for those behaviors that are semantically similar or that involve more fish.

6.4. Practical implementation

While real-time operational capability is beneficial for scenarios like monitoring fish in aquaculture pens (Rosten et al., 2023) or tracking fish migration upstream (Magaju et al., 2023), it is not essential for our ecological study. Conducting behavior analysis on the video data collected during the field season suffices. In essence, performing offline behavior analysis provides adequate behavior classifications for marine biologists to address their research questions. Nevertheless, the proposed method is very efficient such that its inference time for one epoch is 3.5 s on a machine equipped with an Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz, 64GB of RAM, and a single NVIDIA RTX2080 GPU.

6.5. Dataset bias

As previously mentioned, the collected dataset focuses on the corkwing wrasse breeding season, capturing and carefully annotating the relevant event classes on the corkwing male nest. In addition to breeding season activities, we included the default non-specific behavior where the male corkwing simply swims. There is no artificial bias introduced in the event classes, such as by discarding certain events or decreasing their occurrences. Any potential bias could arise only if the object tracker we used fails to detect certain event classes. Although this can be addressed by fine-tuning the tracker with more underwater video data, our visual inspections performed on a subset of the trajectory data, confirm that this is not the case. Instead, the data is naturally imbalanced, with some events occurring infrequently during the timeline of video capturing.

7. Conclusions

The main research question addressed in this study was whether it is feasible to classify fish behaviors, particularly the behavior of the nesting male corkwing wrasse (*Symphodus melops*), through trajectory-based generative model pre-training. In essence, the aim was to examine individual trajectories, applying unsupervised pre-training with diffusion models as the feature learning step, and then applying a relatively shallow MLP for fish event classification. We have shown that diffusion models serve as effective pre-training models, yielding superior results concerning autoencoders, state-of-the-art fish trajectory features, and fully supervised methods. We have also assessed the efficacy of specific techniques to address imbalanced learning, such as *Focal Loss* and *SMOTE*, showing that their involvement improves the classification results.

To this point, few studies have applied deep models to analyze and

classify fish trajectories, and even fewer studies were performed in natural environments and with an elevated number of classes. This study represents the pioneering use of generative models for pre-training trajectories in a fish behavior classifier, highlighting the potential of such an approach for fish event classification. Nonetheless, further research efforts can be made to leverage the performance and generalize the pipeline to other fields.

7.1. Limitations and future work

One notable limitation is the use of a single dataset in both the proposed method and other methods employed in this study. While this dataset has its unique qualities, it is crucial to validate these models on alternative datasets when they become publicly available. Moreover, expanding the dataset is a critical next step for us, but it is important to acknowledge the significant challenges associated with collecting relevant data in a natural setting. This includes the considerable time required to observe specific events and the need for meticulous annotation. Waiting for these events to occur naturally further complicates the data collection process, highlighting the value of the collected data. Additionally, integrating appearance-based information from RGB video clips has the potential to enhance effectiveness, which we plan to address in future work.

CRediT authorship contribution statement

Noemi Canovi: Writing – review & editing, Writing – original draft, Software, Formal analysis. **Benjamin A. Ellis:** Writing – review & editing, Writing – original draft, Conceptualization. **Tonje K. Sørdalen:** Writing – review & editing, Writing – original draft, Validation, Investigation, Funding acquisition, Data curation. **Vaneeda Allken:** Writing – review & editing, Software, Formal analysis. **Kim T. Halvorsen:** Writing – review & editing, Writing – original draft, Validation, Project administration, Investigation, Funding acquisition, Data curation. **Ketil Malde:** Writing – review & editing, Writing – original draft, Validation, Software. **Cigdem Beyan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Data availability

A link to access to the trajectory data and the code of the proposed method will be injected upon acceptance of this paper.

Acknowledgment

This work is funded by the Norwegian Research Council through the project “Computer Vision to Expand Monitoring and Accelerate Assessment of Coastal Fish (CoastVision)” with project number 325862, as well as the Norwegian Institute of Marine Research, project number 15638. The dataset utilized in this paper derives from data collected as part of a University of Plymouth Research Masters undertaken in collaboration with the Norwegian Institute of Marine Research.

References

Aguzzi, J., Doya, C., Tecchio, S., De Leo, F.C., Azzurro, E., Costa, C., Sbragaglia, V., Del Río, J., Navarro, J., Ruhl, H.A., Company, J.B., Favali, P., Purser, A., Thomsen, L., Catalán, I.A., 2015. Coastal observatories for monitoring of fish behaviour and their responses to environmental changes. Rev. Fish Biol. Fish. 25 (3), 463–483. https://doi.org/10.1007/s11160-015-9387-9.

Alex, S.A., Nayahi, J.J.V., Kaddoura, S., 2024. Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification. Appl. Soft Comput. 111491.

Allken, V., Rosen, S., Handegard, N.O., Malde, K., 2021. A deep learning-based method to identify and count pelagic and mesopelagic fishes from trawl camera images. ICES J. Mar. Sci. 78 (10), 3780–3792.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features. In: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I 9. Springer, pp. 404–417.

Ben Tanfous, A., Drira, H., Ben Amor, B., 2018. Coding kendall's shape trajectories for 3d action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2840–2849.

Bergmann, P., Meinhardt, T., Leal-Taixe, L., 2019. Tracking without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 941–951.

Beyan, C., Browman, H.I., 2020. Setting the stage for the machine intelligence era in marine science. ICES J. Mar. Sci. 77 (4), 1267–1273.

Beyan, C., Fisher, R.B., 2012. A filtering mechanism for normal fish trajectories. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). IEEE.

Beyan, C., Fisher, R.B., 2013a. Detecting abnormal fish trajectories using clustered and labeled data. In: International Conference on Image Processing. IEEE.

Beyan, C., Fisher, R.B., 2013b. Detection of abnormal fish trajectories using a clustering based hierarchical classifier. In: British Machine Vision Conference (BMVC).

Beyan, C., Fisher, R.B., 2013c. Detecting abnormal fish trajectories using clustered and labeled data. In: 2013 IEEE International Conference on Image Processing. IEEE, pp. 1476–1480.

Beyan, C., Fisher, R., 2013d. Detection of abnormal fish trajectories using a clustering based hierarchical classifier. In: Proceedings of the British Machine Vision Conference. BMVA Press, pp. 1–11.

Beyan, C., Fisher, R., 2015. Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recogn. 48 (5), 1653–1672.

Beyan, C., Katsageorgiou, V.-M., Murino, V., 2017. Moving as a leader: Detecting emergent leadership in small groups using body pose. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1425–1433.

Beyan, C., Katsageorgiou, V.-M., Fisher, R.B., 2018. Extracting statistically significant behaviour from fish tracking data with and without large dataset cleaning. IET Comput. Vis. 12 (2), 162–170.

Boom, B.J., He, J., Palazzo, S., Huang, P.X., Beyan, C., Chou, H.-M., Lin, F.-P., Spampinato, C., Fisher, R.B., 2014. A research tool for long-term and continuous analysis of fish assemblage in coral-reefs using underwater camera footage. Eco. Inform. 23, 83–97.

Bourel, M., Segura, A.M., Crisci, C., López, G., Sampognaro, L., Vidal, V., Kruk, C., Piccini, C., Perera, G., 2021. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. Water Res. 202, 117450.

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., Li, S.Z., 2024. A survey on generative diffusion models. IEEE Trans. Knowl. Data Eng. 36 (7), 2814–2830. https://doi.org/10.1109/TKDE.2024.3361474.

Caravaggi, A., Banks, P.B., Burton, A.C., Finlay, C.M., Haswell, P.M., Hayward, M.W., Rowcliffe, M.J., Wood, M.D., 2017. A review of camera trapping for conservation behaviour research. Remote Sens. Ecol. Conserv. 3 (3), 109–122.

Catalán, I.A., Álvarez-Ellacuría, A., Lisani, J.L., Sánchez, J., Vizoso, G., Heinrichs-Maquilón, A.E., Hinz, H., Alós, J., Signarioli, M., Aguzzi, J., Francescangeli, M., Palmer, M., 2023. Automatic detection and classification of coastal Mediterranean fish from underwater images: good practices for robust training. Front. Mar. Sci. 10 (April), 1–11. https://doi.org/10.3389/fmars.2023.1151758.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, S., Sun, P., Song, Y., Luo, P., 2023. Diffusiondet: diffusion model for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 19830–19843.

Cho, H., Yoon, S.M., 2018. Divide and conquer-based 1d cnn human activity recognition using test data sharpening. Sensors 18 (4), 1055.

Claridge, A.W., Mifsud, G., Dawson, J., Saxon, M.J., 2004. Use of infrared digital cameras to investigate the behaviour of cryptic species. Wildl. Res. 31 (6), 645–650.

Dell, J.A., Bender, I.D., Branson, K., Couzin, G.G., de Polavieja, L.P., Noldus, U., Brose, 2014. Automated image-based tracking and its application in ecology, 29 (7), 417–428.

D'incà, M., Beyan, C., Niewiadomski, R., Barattin, S., Sebe, N., 2023. Unleashing the transferability power of unsupervised pre-training for emotion recognition in masked and unmasked facial images. IEEE Access 11, 90876–90890. https://doi.org/10.1109/ACCESS.2023.3308047.

Ditria, E.M., Lopez-Marcano, S., Sievers, M., Jinks, E.L., Brown, C.J., Connolly, R.M., 2020. Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. Front. Mar. Sci. 7, 429. https://doi.org/10.3389/fmars.2020.00429.

Ditria, E.M., Jinks, E.L., Connolly, R.M., 2021. Automating the analysis of fish grazing behaviour from videos using image classification and optical flow. Anim. Behav. 177, 31–37.

Ellis, B.A., 2023. Personality and parental care in the corkwing wrasse (*symphodus melops*). Ph.D. thesis. University of Plymouth.

Eren, L., 2017. Bearing fault detection by one-dimensional convolutional neural networks. Math. Probl. Eng. 2017, 1–9.

Erhan, D., Courville, A., Bengio, Y., Vincent, P., 2010. Why does unsupervised pre-training help deep learning?. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, pp. 201–208.

Escottá, Á.T., Beccaro, W., Ramírez, M.A., 2022. Evaluation of 1d and 2d deep convolutional neural networks for driving event recognition. Sensors 22 (11), 4226.

Estevam, V., Pedrini, H., Menotti, D., 2021. Zero-shot action recognition in videos: a survey. Neurocomputing 439, 159–175.

Fisher, R.B., Chen-Burger, Y.-H., Giordano, D., Hardman, L., Lin, F.-P., 2016. Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data, 104. Springer.
Frainer, G., Dufourq, E., Fearey, J., Dines, S., Probert, R., Elwen, S., Gridley, T., 2023. Automatic detection and taxonomic identification of dolphin vocalisations using convolutional neural networks for passive acoustic monitoring. Eco. Inform. 78, 102291.
Franceschini, R., Fini, E., Beyan, C., Conti, A., Arrigoni, F., Ricci, E., 2022. Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. In: 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, pp. 2589–2596.
Friard, O., Gamba, M., 2016. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods Ecol. Evol. 7 (11), 1325–1330.
Fundel, F., Braun, D.A., Gottwald, S., 2023. Automatic bat call classification using transformer networks. Eco. Inform. 78, 102288.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (4), 463–484.
Ge, J., Tang, S., Fan, J., Jin, C., 2023. On the provable advantage of unsupervised pretraining. The Twelfth International Conference on Learning Representations.
Goodwin, M., Halvorsen, K.T., Jiao, L., Knausgård, K.M., Martin, A.H., Moyano, M., Oomen, R.A., Rasmussen, J.H., Sørdalen, T.K., Thorbjørnsen, S.H., 2022. Unlocking the potential of deep learning for marine ecology: overview, applications, and outlook. ICES J. Mar. Sci. 79 (2), 319–336.
Gu, Z., Chen, H., Xu, Z., et al., 2024. Diffusioninst: diffusion model for instance segmentation. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, pp. 2730-2734, doi: 10.1109/ICASSP48485.2024.10447191.
Gui, L.-Y., Wang, Y.-X., Liang, X., Moura, J.M., 2018. Adversarial geometry-aware human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 786–803.
Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R., 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, 36, pp. 762–770.
Gurevitch, J., Morrison, J.A., Hedges, L.V., 2000. The interaction between competition and predation: a meta-analysis of field experiments. Am. Nat. 155 (4), 435–453.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. 73, 220–239.
Halvorsen, K.T., Sørdalen, T.K., Durif, C., Knutsen, H., Olsen, E.M., Skiftesvik, A.B., Rustand, T.E., Bjelland, R.M., Vøllestad, L.A., 2016. Male-biased sexual size dimorphism in the nest building corkwing wrasse (*symphodus melops*): implications for a size regulated fishery. ICES J. Mar. Sci. 73 (10), 2586–2594.
Halvorsen, K.T., Sørdalen, T.K., Vøllestad, L.A., Skiftesvik, A.B., Espeland, S.H., Olsen, E. M., 2017a. Sex- and size-selective harvesting of corkwing wrasse (*Symphodus melops*)—a cleaner fish used in salmonid aquaculture. ICES J. Mar. Sci. 74 (3), 660–669. https://doi.org/10.1093/icesjms/fsw221.
Halvorsen, K.T., Larsen, T., Sørdalen, T.K., Vøllestad, L.A., Knutsen, H., Olsen, E.M., 2017b. Impact of harvesting cleaner fish for salmonid aquaculture assessed from replicated coastal marine protected areas. Mar. Biol. Res. 13 (4), 359–369. https://doi.org/10.1080/17451000.2016.1262042.
Halvorsen, K.T., Sørdalen, T.K., Larsen, T., Browman, H.I., Rafoss, T., Albretsen, J., Skiftesvik, A.B., 2024. Mind the Depth: The Vertical Dimension of a Small-Scale Coastal Fishery Shapes Selection on Species, Size, and Sex in Wrasses, Marine and Coastal Fisheries, 6, pp. 404–422. https://doi.org/10.1002/mcf2.10131.
Hoffmann, A.A., Sgrò, C.M., 2011. Climate change and evolutionary adaptation. Nature 470 (7335), 479–485.
Holden, D., Saito, J., Komura, T., Joyce, T., 2015. Learning motion manifolds with convolutional autoencoders. In: SIGGRAPH Asia 2015 Technical Briefs. ACM, pp. 1–4.
Hosseini, B., Montagne, R., Hammer, B., 2020. Deep-aligned convolutional neural network for skeleton-based action recognition and segmentation. Data Sci. Eng. 5 (2), 126–139.
Hsieh, C.-H., Lo, Y.-S., Chen, J.-Y., Tang, S.-K., 2021. Air-writing recognition based on deep convolutional neural networks. IEEE Access 9, 142827–142836.
Hu, J., Zhao, D., Zhang, Y., Zhou, C., Chen, W., 2021. Real-time nondestructive fish behavior detecting in mixed polyculture system using deep-learning and low-cost devices. Expert Syst. Appl. 178, 115051.
Hyvarinen, A., Dayan, P., 2005. Estimation of non-normalized statistical models by score matching. J. Mach. Learn. Res. 6 (4).
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456.
Javidani, A., Mahmoudi-Aznaveh, A., 2022. Learning representative temporal features for action recognition. Multimed. Tools Appl. 81 (3), 3145–3163.
Karaszkiewicz, M., 2020. Reproductive biology in corkwing wrasse (*symphodus melops*). Master's thesis.
Karras, T., Aittala, M., Aila, T., Laine, S., 2022. Elucidating the design space of diffusion-based generative models. Adv. Neural Inf. Proces. Syst. 35, 26565–26577.
Katsageorgiou, V.-M., Zanotto, M., Tucci, V., Murino, V., Sona, D., 2017. Data-driven study of mouse sleep-stages using restricted boltzmann machines. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 4549–4556. https://doi.org/10.1109/IJCNN.2017.7966433.
Kindsvater, H.K., Halvorsen, K.T., Sørdalen, T.K., Alonzo, S.H., 2020. The consequences of size-selective fishing mortality for larval production and sustainable yield in species with obligate male care. Fish Fish. 21 (6), 1135–1149. https://doi.org/10.1111/faf.12491.
Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., Gabbouj, M., 2019. 1-d convolutional neural networks for signal processing applications. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8360–8364. https://doi.org/10.1109/ICASSP.2019.8682194.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations (ICLR).
Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J., 2021. 1d convolutional neural networks and applications: a survey. Mech. Syst. Signal Process. 151, 107398.
Knausgård, K.M., Wiklund, A., Sørdalen, T.K., Halvorsen, K.T., Kleiven, A.R., Jiao, L., Goodwin, M., 2022. Temperate fish detection and classification: a deep learning based approach. Appl. Intell. 52 (6), 6988–7001.
Kong, Y., Fu, Y., 2022. Human action recognition and prediction: a survey. Int. J. Comput. Vis. 130 (5), 1366–1401.
Koromilas, P., Giannakopoulos, T., 2021. Unsupervised multimodal language representations using convolutional autoencoders. arXiv 1–5.
Kubat, M., Matwin, S., 1997. Addressing the curse of imbalanced training sets: one-sided selection. In: Proc. of ICML, 97, p. 179.
Kundu, J.N., Gor, M., Uppala, P.K., Radhakrishnan, V.B., 2019. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1459–1467.
Legagneux, P., Ducatez, S., 2013. European birds adjust their flight initiation distance to road speed limits. Biol. Lett. 9 (5), 20130417.
Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S., 2018. Unsupervised Learning of View-Invariant Action Representations, Advances in Neural Information Processing Systems, 31.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988.
Lin, L., Song, S., Yang, W., Liu, J., 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2490–2498.
Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D., 2023. AudioLDM: text-to-audio generation with latent diffusion models. In: Proceedings of the 40th International Conference on Machine Learning (ICML'23), Vol. 202. JMLR. org, Article 886, 21450–21474.
Long, L., Johnson, Z.V., Li, J., Lancaster, T.J., Aljapur, V., Streelman, J.T., McGrath, P.T., 2020. Automatic classification of cichlid behaviors using 3d convolutional residual networks. Iscience 23 (10).
Lopez-Marcano, S., Jinks, E.L., Buelow, C.A., Brown, C.J., Wang, D., Kusy, B., Ditria, E. M., Connolly, R.M., 2021. Automatic detection of fish and tracking of movement for ecology. Ecol. Evol. 11 (12), 8254–8263.
Luque, A., Carrasco, A., Martín, A., de Las Heras, A., 2019. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recogn. 91, 216–231.
Magaju, D., Montgomery, J., Franklin, P., Baker, C., Friedrich, H., 2023. Machine learning based assessment of small-bodied fish tracking to evaluate spoiler baffle fish passage design. J. Environ. Manag. 325, 116507.
Måløy, H., Aamodt, A., Misimi, E., 2019. A spatio-temporal recurrent network for salmon feeding action recognition from underwater videos in aquaculture. Comput. Electron. Agric. 167, 105087.
Marjani, M., Ahmadi, S.A., Mahdianpari, M., 2023. Firepred: a hybrid multi-temporal convolutional neural network model for wildfire spread prediction. Eco. Inform. 78, 102282.
Martinez, J., Black, M.J., Romero, J., 2017. On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2891–2900.
Mattioli, F., Porcaro, C., Baldassarre, G., 2022. A 1d cnn for high accuracy classification and transfer learning in motor imagery eeg-based brain-computer interface. J. Neural Eng. 18 (6), 066053.
McIntosh, D., Marques, T.P., Albu, A.B., Rountree, R., De Leo, F., 2020. Movement tracks for the automatic detection of fish behavior in videos. arXiv 1–6.
Nie, Q., Liu, Z., Liu, Y., 2020. Unsupervised human 3d pose representation with viewpoint and pose disentanglement. In: Springer European Conference on Computer Vision (ECCV).
Palazzo, S., Spampinato, C., Beyan, C., 2012. Event detection in underwater domain by exploiting fish trajectory clustering. In: Proceedings of the 1st ACM International workshop on Multimedia Analysis for Ecological Data, pp. 31–36.
Paoletti, G., Cavazza, J., Beyan, C., Del Bue, A., 2021a. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. In: The 32nd British Machine Vision Conference (BMVC).
Paoletti, G., Cavazza, J., Beyan, C., Del Bue, A., 2021b. Unsupervised human action recognition with skeletal graph Laplacian and self-supervised viewpoints invariance. In: The 32nd British Machine Vision Conference (BMVC).
Paoletti, G., Cavazza, J., Beyan, C., Del Bue, A., 2021c. Subspace clustering for action recognition with covariance representations and temporal pruning. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 6035–6042.
Paoletti, G., Cavazza, J., Beyan, C., Del Bue, A., 2021d. Unsupervised human action recognition with skeletal graph laplacian and self-supervised viewpoints invariance. In: 32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22–25, 2021. BMVA.

Paoletti, G., Beyan, C., Del Bue, A., 2022a. Graph laplacian-improved convolutional residual autoencoder for unsupervised human action and emotion recognition. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3229478.
Paoletti, G., Beyan, C., Del Bue, A., 2022b. Graph laplacian-improved convolutional residual autoencoder for unsupervised human action and emotion recognition. IEEE Access 10, 131128–131143.
Paoletti, G., Beyan, C., Del Bue, A., 2022c. Graph laplacian-improved convolutional residual autoencoder for unsupervised human action and emotion recognition. IEEE Access 10, 131128–131143. https://doi.org/10.1109/ACCESS.2022.3229478.
Pareek, P., Thakkar, A., 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artif. Intell. Rev. 54, 2259–2322.
Parks, S.E., Johnson, M., Nowacek, D., Tyack, P.L., 2011. Individual right whales call louder in increased environmental noise. Biol. Lett. 7 (1), 33–35.
Patrick, M., Campbell, D., Asano, Y., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., Henriques, J.F., 2021. Keeping your eye on the ball: trajectory attention in video transformers. Adv. Neural Inf. Proces. Syst. 34, 12493–12506.
Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. Film: visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, 32.
Phan, N., Dou, D., Piniewski, B., Kil, D., 2016. A deep learning approach for human behavior prediction with explanations in health social networks: social restricted boltzmann machine (srbm+). Soc. Netw. Anal. Min. 6, 1–14.
Poloczanska, E.S., Brown, C.J., Sydeman, W.J., Kiessling, W., Schoeman, D.S., Moore, P. J., Brander, K., Bruno, J.F., Buckley, L.B., Burrows, M.T., et al., 2013. Global imprint of climate change on marine life. Nat. Clim. Chang. 3 (10), 919–925.
Potts, G., 1974. The colouration and its behavioural significance in the corkwing wrasse, *crenilabrus melops*. J. Mar. Biol. Assoc. U. K. 54 (4), 925–938.
Potts, G.W., 1985. The nest structure of the corkwing wrasse, *crenilabrus melops* (labridae: Teleostei). J. Mar. Biol. Assoc. U. K. 65 (2), 531–546.
Qazi, E.U.H., Almorjan, A., Zia, T., 2022. A one-dimensional convolutional neural network (1d-cnn) based deep learning system for network intrusion detection. Appl. Sci. 12 (16), 7986.
Ramsey, A.B., Sawaya, M.A., Bullington, L.S., Ramsey, P.W., 2019. Individual identification via remote video verified by dna analysis: a case study of the american black bear. Wildl. Res. 46 (4), 326–333.
Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B., 2021. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Inf. Sci. 569, 90–109.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-Cnn: Towards Real-Time Object Detection with Region Proposal Networks, Advances in Neural Information Processing Systems, 28.
Rosten, C.M., Mathiassen, J.R., Volent, Z., 2023. Acoustic environment of aquaculture net-pens varies with feeding status of Atlantic salmon (salmo Salar). Aquaculture 563, 738949.
Rowland, W.J., 1999. Studying visual cues in fish behavior: a review of ethological techniques. Environ. Biol. Fish 56, 285–305.
Saharia, C., Chan, W., Saxena, S.E.A., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. 35, 36479–36494.
Schindler, F., Steinhage, V., 2021. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. Eco. Inform. 61, 101215.
Shi, Y., Tian, Y., Wang, Y., Huang, T., 2017. Sequential deep trajectory descriptor for action recognition with three-stream cnn. IEEE Trans. Multimed. 19 (7), 1510–1520.
Shin, J., Yoon, S., Kim, Y., Kim, T., Go, B., Cha, Y., 2021. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. Eco. Inform. 61, 101202.
Shuster, S.M., Wade, M.J., 2003. Mating Systems and Strategies. Princeton University Press.
Skiftesvik, A.B., Blom, G., Agnalt, A.-L., Durif, C.M., Browman, H.I., Bjelland, R.M., Harkestad, L.S., Farestveit, E., Paulsen, O.I., Fauske, M., et al., 2014. Wrasse (labridae) as cleaner fish in salmonid aquaculture–the hardangerfjord as a case study. Mar. Biol. Res. 10 (3), 289–300.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Su, K., Liu, X., Shlizerman, E., 2020. Predict & cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9631–9640.
Sujatha, M., Jaidhar, C., Lingappa, M., 2023. 1d convolutional neural networks-based soil fertility classification and fertilizer prescription. Eco. Inform. 78, 102295.
Suryawati, E., Pardede, H.F., Zilvan, V., Ramdan, A., Krisnandi, D., Heryana, A., Yuwana, R.S., Kusumo, R.B.S., Arisal, A., Supianto, A.A., 2021. Unsupervised feature learning-based encoder and adversarial networks. J. Big Data 8, 1–17.
Trelinski, J., Kwolek, B., 2021. Embedded features for 1d cnn-based action recognition on depth maps. In: VISIGRAPP (4: VISAPP), pp. 536–543.
Troullinou, E., Tsagkatakis, G., Chavlis, S., Turi, G.F., Li, W., Losonczy, A., Tsakalides, P., Poirazi, P., 2020. Artificial neural networks in action for an automated cell-type classification of biological neural networks. IEEE Trans. Emerg. Top. Comp. Intellig. 5 (5), 755–767.
Truong, T.H., Du Nguyen, H., Mai, T.Q.A., Nguyen, H.L., Dang, T.N.M., et al., 2023. A deep learning-based approach for bee sound identification. Eco. Inform. 78, 102274.
Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E., 2023a. Exploring diffusion models for unsupervised video anomaly detection. In: IEEE International Conference on Image Processing (ICIP), pp. 2540–2544.
Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E., 2023b. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In: International Conference on Image Analysis and Processing. Springer, pp. 49–62.
Uglem, I., Rosenqvist, G., 2002. Nest building and mating in relation to male size in corkwing wrasse, *symphodus melops*. Environ. Biol. Fish 63, 17–25.
Uglem, I., Rosenqvist, G., Wasslavik, H.S., 2000. Phenotypic variation between dimorphic males in corkwing wrasse. J. Fish Biol. 57 (1), 1–14. https://doi.org/10.1006/jfbi.2000.1283.
Walker, J., Gupta, A., Hebert, M., 2015. Dense optical flow prediction from a static image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2443–2451.
Wang, H., Schmid, C., 2013. Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558.
Wang, H., Klaser, A., Schmid, C., Cheng-Lin, L., 2011. Action recognition by dense trajectories. computer vision and pattern recognition (cvpr). In: 2011 IEEE Conference on, pp. 3169–3176.
Wang, L., Qiao, Y., Tang, X., 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Wang, P., Li, W., Li, C., Hou, Y., 2018. Action recognition based on joint trajectory maps with convolutional neural networks. Knowl.-Based Syst. 158, 43–53.
Wang, G., Muhammad, A., Liu, C., Du, L., Li, D., 2021. Automatic recognition of fish behavior with a fusion of rgb and optical flow data based on deep learning. Animals 11 (10), 2774.
Weinstein, B.G., 2017. A computer vision for animal ecology. J. Anim. Ecol. 87 (3), 533–545 arXiv:0608246v3. https://doi.org/10.1111/1365-2656.12780.
Wey, T., Blumstein, D.T., Shen, W., Jordán, F., 2008. Social network analysis of animal behaviour: a promising tool for the study of sociality. Anim. Behav. 75 (2), 333–344.
Wong, B.B., Candolin, U., 2015. Behavioral responses to changing environments. Behav. Ecol. 26 (3), 665–673.
Xiang, W., Yang, H., Huang, D., Wang, Y., 2023. Denoising diffusion autoencoders are unified self-supervised learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15802–15812.
Xie, J., Hu, K., Guo, Y., Zhu, Q., Yu, J., 2021. On loss functions and cnns for improved bioacoustic signal classification. Eco. Inform. 64, 101331.
Yang, X., Wang, X., 2023. Diffusion model as representation learner. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 18938–18949.
Xu, S., Rao, H., Hu, X., Cheng, J., Hu, B., 2023. Prototypical contrast and reverse prediction: unsupervised skeleton based action recognition. IEEE Trans. Multimed. 25, 624–634. https://doi.org/10.1109/TMM.2021.3129616.
Yang, L., Liu, Y., Yu, H., Fang, X., Song, L., Li, D., Chen, Y., 2021. Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: a review. Archiv. Comp. Methods Eng. 28, 2785–2816.
Zamboni, S., Kefato, Z.T., Girdzijauskas, S., Norén, C., Dal Col, L., 2022. Pedestrian trajectory prediction with convolutional neural networks. Pattern Recogn. 121, 108252.
Zanfir, M., Leordeanu, M., Sminchisescu, C., 2013. The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection. In: Proceedings of IEEE ICCV, pp. 2752–2759.
Zhang, C., Yang, T., Weng, J., Cao, M., Wang, J., Zou, Y., 2022. Unsupervised pre-training for temporal action localization tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14031–14041.
Zhao, Y., Xiong, Y., Lin, D., 2018a. Trajectory Convolution for Action Recognition, Advances in Neural Information Processing Systems, 31.
Zhao, J., Baoa, W., Zhanga, F.E.A., 2018b. Modified motion influence map and recurrent neural network-based monitoring of the local unusual behaviors for fish school in intensive aquaculture. Aquaculture 493, 165–175.
Zhao, J., Bao, W., Zhang, F., Zhu, S., Liu, Y., Lu, H., Shen, M., Ye, Z., 2018c. Modified motion influence map and recurrent neural network-based monitoring of the local unusual behaviors for fish school in intensive aquaculture. Aquaculture 493, 165–175.
Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z., 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: AAAI Conference on Artificial Intelligence.