3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset

Malte Pedersen*, Joakim Bruslund Haurum*, Stefan Hein Bengtson, Thomas B. Moeslund
Visual Analysis of People (VAP) Laboratory, Aalborg University, Denmark
mape@create.aau.dk, joha@create.aau.dk, shbe@create.aau.dk, tbm@create.aau.dk

Abstract

In this work we present a novel publicly available stereo based 3D RGB dataset for multi-object zebrafish tracking, called 3D-ZeF. Zebrafish is an increasingly popular model organism used for studying neurological disorders, drug addiction, and more. Behavioral analysis is often a critical part of such research. However, visual similarity, occlusion, and erratic movement of the zebrafish makes robust 3D tracking a challenging and unsolved problem.

The proposed dataset consists of eight sequences with a duration between 15-120 seconds and 1-10 free moving zebrafish. The videos have been annotated with a total of 86,400 points and bounding boxes. Furthermore, we present a complexity score and a novel open-source modular baseline system for 3D tracking of zebrafish. The performance of the system is measured with respect to two detectors: a naive approach and a Faster R-CNN based fish head detector. The system reaches a MOTA of up to 77.6%. Links to the code and dataset is available at the project page `http://vap.aau.dk/3d-zef`

1. Introduction

Over the past decades, the use of zebrafish (*Danio rerio*) as an animal model has increased significantly due to its applicability within large-scale genetic screening [1, 2]. The zebrafish has been used as a model for studying human neurological disorders, drug addiction, social anxiety disorders, and more [3, 4, 5, 6, 7, 8]. Locomotion and behavioral analysis are often critical parts of neuroscientific and biological research, which have traditionally been conducted manually [9, 10, 11]. However, manual inspection is subjective and limited to small-scale experiments. Therefore, tracking systems are getting increasingly popular due to their efficiency and objectivity. The majority of the solutions has been developed for terrestrial animals or fish in shallow water, and most studies have been based on 2D observations in scientific [12, 13, 14, 15, 16, 17, 18] and com-
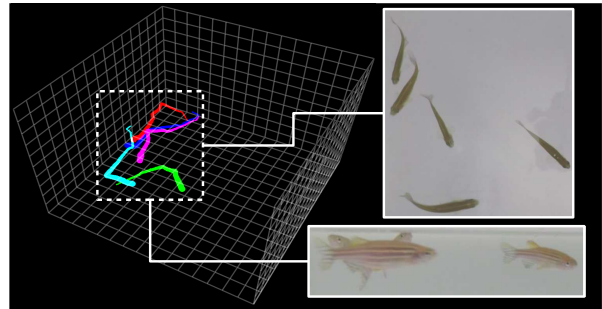

Figure 1: An example that illustrates the difference between the two perspectives. The 3D trajectories are estimated based on the head point annotations.

mercial systems [19, 20, 21, 22]. However, observations in a single plane cannot capture all the relevant phenotypes of fish [23, 24, 25]. Estimating the 3D trajectories of multiple zebrafish accurately is difficult due to their erratic movement, visual similarity, and social behavior [26], see Figure 1. This may be one of the reasons why no commercial solution has been developed yet. Only few groups in the scientific community have addressed the problem, focusing mainly on stereo vision [27, 28, 29, 30, 31] and monocular stereo using mirrors [32, 33]. However, no labeled datasets have been made publicly available within the field, which makes a fair comparison between the applied methods difficult. This ultimately hinders significant developments in the field as we have seen in other computer vision fields with common datasets. Therefore, our contributions are

- a publicly available RGB 3D video dataset of zebrafish with 86,400 bounding box and point annotations.

- an open-source modular baseline system.

A large part of 3D multi-object tracking methods are developed for LiDAR-based traffic datasets [34, 35, 36, 37, 38] or RGB-D tracking [39, 40]. However, to the best of our knowledge, there exists no publicly available annotated RGB stereo dataset with erratic moving and similarly looking subjects like the one we propose.
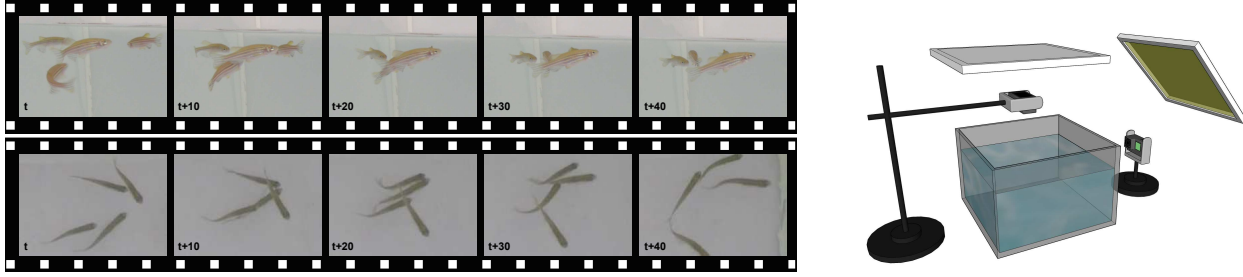
*Equal contribution

Figure 2: Five frames from two different occlusion scenarios. The upper frames are from the front-view and the lower frames are from the top-view. An illustration of the experimental setup is shown to the right.

2. Related Work

Multi-Object Tracking (MOT). Reliably tracking multiple objects is widely regarded as incredibly difficult. The interest in solving MOT has been steadily increasing since 2015 with the release of the MOT [41, 42, 43], UA-DETRAC [44, 45], and KITTI [34, 35] challenges. Within the MOT challenges, the current focus is on either aiming to solve the association problem using deep learning [46], using techniques such as intersection-over-union based tracking [47], or disregarding tracking-specific models and utilizing the improvements within object detections [48].

Zebrafish Tracking. Vision-based tracking systems developed for studying animal behavior have traditionally been based on 2D [18, 49, 50, 51, 52, 53, 54] due to simplicity and because the movement of most terrestrial animals can be approximated to a single plane. The majority of research in zebrafish tracking has followed this path by only allowing the fish to move in shallow water and assuming that motion happens in a 2D plane.

A 2D animal tracker, called idTracker presented by Perez-Escudero *et al.* in 2014 [49], uses thresholding to segment blobs and is able to distinguish between individual zebrafish based on intensity and contrast maps. In 2019, Romero-Ferrero *et al.* presented an updated version of idTracker, called idtracker.ai [18], which is the current state-of-the-art 2D tracker system based on convolutional neural networks (CNN) for handling occlusions and identifying individuals. The subjects are observed with a camera positioned above a tank with a water depth of 2.5 cm and the distance between camera and subjects is, therefore, approximately the same at all times. As stated by the authors, this simplifies the task compared to a real 3D tracking scenario.

However, as most aquatic species move in three dimensions, trajectories in 3D are required to thoroughly describe their behavior [55, 56]. The most frequently used acquisition method when dealing with studies of animal behavior in 3D is stereo vision [28, 30, 31, 56, 57, 58, 59, 60, 61]. 3D tracking of zebrafish has been focused mainly on single subjects or small groups, as occlusion is a big hindrance for maintaining correct IDs due to their shoaling behavior [26].

Furthermore, the visual appearance of the fish can change dramatically depending on the position and posture, which makes re-identification more complex compared to 2D.

The Track3D module from the commercial EthoVision XT [19] is popular for tracking zebrafish in 3D, but is limited to a single individual [56, 61]. An early semi-automatic 3D tracking system was developed by Viscido *et al.* [58] to investigate the relationship between individual members of fish schools. Initial 2D tracks were generated by a nearest neighbor algorithm followed by a step allowing the user to adjust and correct the proposed 2D trajectories, and subsequently triangulated to reconstruct the 3D trajectories.

Qian *et al.* have worked extensively with tracking of zebrafish and have developed a 2D tracking system with a top-view camera using an augmented fast marching method (AFMM) [62] and the determinant of the Hessian [15]. This was expanded to 3D tracking by extending the setup with a side-view camera. AFMM was utilized to generate a feature point based fish representation in each view followed by 2D tracklet construction based on motion constraints. 3D tracks were then constructed by associating the 2D tracklets with side-view detections using epipolar and motion consistency constraints [29]. Liu *et al.* [63] extended this method to better handle occlusions based on a set of heuristic methods and the epipolar constraint. A third camera was added in [31], and the feature point representation method was extended.

Cheng *et al.* [28] utilized a similar three-camera setup, applying an iterative unsupervised learning method to train a CNN-based classifier to distinguish between the individual fish from a camera placed above the water tank. The classifier was trained on the head region of the fish during periods when all fish were visible at the same time. By iteratively retraining the classifier, they were able to generate 2D tracks from the top-view and reconstruct the 3D tracklets based on detections from the two other side-view cameras under epipolar and motion constraints.

Wang *et al.* [30] also utilized a three-camera setup, using a Gaussian Mixture Model, a Gabor filter and an SVM-based method to detect the fish heads in the top- and side-

	Trn2	Trn5	Val2	Val5	Tst1	Tst2	Tst5	Tst10	Total
Length	120 s	15 s	30 s	15 s	15 s	15 s	15 s	15 s	240 s
Frames	14,400	1,800	3,600	1,800	1,800	1,800	1,800	1,800	28,800
BBs	28,800	9,000	7,200	9,000	1,800	3,600	9,000	18,000	86,400
Points	28,800	9,000	7,200	9,000	1,800	3,600	9,000	18,000	86,400
OC	1.82 / 1.42	3.60 / 2.93	0.93 / 0.47	2.67 / 3.80	0.00 / 0.00	0.67 / 0.67	3.07 / 2.93	4.40 / 6.53	
OL	0.41 / 0.51	0.56 / 0.64	0.22 / 0.63	0.25 / 0.66	0.00 / 0.00	0.10 / 0.38	0.25 / 0.36	0.28 / 0.35	
TBO	0.69 / 0.89	1.00 / 1.21	1.79 / 3.20	1.64 / 0.73	15.00 / 15.00	2.41 / 2.18	1.38 / 1.28	1.86 / 1.40	
IBO	0.29 / 0.26	0.28 / 0.28	0.24 / 0.35	0.22 / 0.34	0.00 / 0.00	0.19 / 0.19	0.25 / 0.23	0.26 / 0.24	
Ψ	**0.26**	**0.50**	**0.03**	**0.63**	**0.00**	**0.01**	**0.16**	**0.28**	

Table 1: Overview of the proposed dataset. OC, OL, TBO, and IBO are listed for the top- and front-view, respectively, and the number of fish is denoted in the sequence name. OC: average amount of occlusions per second, OL: average occlusion length in seconds, TBO: average amount of seconds between occlusions, IBO: intersection between occlusions, Ψ: complexity measure based on OC, OL, TBO and IBO (see Equation (2)).

views, respectively. The top-view detections are associated into 2D tracklets based on a cross-correlation method and by applying a Kalman filter; near linear movement is achieved by a frame rate of 100 FPS. The 2D tracklets are then constructed into 3D tracklets by associating the side-view detections under epipolar and motion constraints. In [64], Wang *et al.* proposed to model the top-view movement of the zebrafish through long short-term memory networks, which were used to improve the motion constraints in a new iteration of their 3D system [65]. Lastly, Wang *et al.* used a CNN for re-identification of zebrafish heads from the top-view [66], although this has yet to be incorporated into a 3D tracking setup. None of the methods are able to track multiple zebrafish in 3D for more than a few seconds without ID swaps; this is still a difficult and unsolved problem.

Datasets. As in other MOT challenges, there is a mutual agreement that occlusion is what makes 3D tracking of zebrafish difficult. Nonetheless, only Wang *et al.* [65] describe their recordings based on occlusion frequency; however, they do not define how it is measured. Qian *et al.* [31] indicate their complexity based on the amount of fish, but only four occlusion events occur during their 15 seconds demo video with ten fish. For comparison, there are 66 occlusion events in our 15 seconds sequence with ten fish.

3. Proposed Dataset

The proposed 3D zebrafish dataset, 3D-ZeF, has been recorded from a top- and front-view perspective. This approach was taken to minimize events of total occlusion typical for side-by-side binocular setups. An example of the visual variation between the views is shown in Figure 2 together with an illustration of the experimental setup.

3.1. Experimental Setup

The setup used to record the proposed dataset was built entirely from off-the-shelf hardware, whereas previous methods have used specialized camera equipment. An illustration of the setup is shown in Figure 2. The two light pan-

els are IKEA FLOALT of size 30×30 cm with a luminous flux of 670 lumen and a color temperature of 4000K. The test tank is a standard glass aquarium of size $30 \times 30 \times 30$ cm with a water depth of 15 cm. The top and front cameras are GoPro Hero 5 and GoPro Hero 7, respectively. All the videos are recorded with a resolution of 2704×1520, 60 FPS, 1/60 s shutter speed, 400 ISO, and a linear field of view. However, the fish tank does not take up the entire image, therefore, the effective region of interest is approximately 1200×1200 and 1800×900 for the top- and front-view, respectively. Diffusion fabric was placed in front of the top light in order to reduce the amount of glare in the top-view. Semi-transparent plastic was attached to three out of four of the window panes in order to reduce reflections. Furthermore, the front camera was placed orthogonally to the water level, which reduced reflections from the water surface. Lastly, the pair-wise recordings have been manually synchronized using a flashing LED, which results in a worst case temporal shift of $\frac{1}{2 \cdot \text{FPS}}$.

3.2. Dataset Construction

A total of eight sequences were recorded and divided into a training, validation, and test split. Each sequence consists of a pair of temporally aligned top- and front-view videos and the specifications of the three splits are shown in Table 1. In order to avoid data leakage, each split contains a unique set of fish. The training and validation set of fish were from the same cohort, whereas the fish in the test split were from a younger cohort. Therefore, the test set differs from the training and validation set, as the fish are smaller and behave socially different. This represent a real-life scenario where different cohorts need to be tracked, which has not generally been addressed within the field.

The zebrafish were manually bounding box and point annotated with consistent identity tags through all frames. The bounding boxes were tightly fitted to the visible parts of the zebrafish and the point annotations were centered on the head. If a set of fish touched, an occlusion tag was set for all

involved bounding boxes. During occlusions, the bounding box was fitted to the visible parts of the fish and not where it was expected to be due to the extreme flexibility of the zebrafish. The pair-wise point annotations from the two views were triangulated into 3D positions using the method proposed by Pedersen *et al.* [67]. The fish head was approximated during occlusions to ensure continuous 3D tracks.

It should be noted that the data was recorded in RGB. Zebrafish can change their body pigmentation based on their environment, stress level, and more [23]. The changes in coloration can be important in behavioral studies and may even be valuable in solving the 3D tracking problem.

3.3. Dataset Complexity

Intuitively, a higher number of fish creates a more difficult tracking problem. However, this is only true to some extent as the main complexity factor is the number and level of occlusions, which depends on a combination of the social activity and amount of space rather than the number of individuals. Therefore, we have defined a range of metrics based on occlusion events to describe the complexity of the proposed sequences. An occlusion event is defined by a set of consecutive frames, where a fish is part of an occlusion. The events are measured from the perspective of the fish; if two fish are part of an occlusion it counts as two events.

The number of occlusion events indicates how often a fish is part of an occlusion, but, few long occlusions can be just as problematic as many short. The length of the occlusions and time between them are, therefore, important to keep in mind when evaluating the complexity of a recording. Due to our definition of occlusion events there are cases where fish are part of occlusions with only minor parts of their bodies. Therefore, the intersection between occlusions is measured as an indication of the general intersection level. The metrics that we provide as basis for the complexity level of our recordings are defined here:

Occlusion Count (OC): the average number of occlusion events per second.

Occlusion Length (OL): the average time in seconds of all occlusion events.

Time Between Occlusions (TBO): the average time in seconds between occlusion events.

Intersection Between Occlusions (IBO): a measure of how large a part of the fish that is part of an occlusion event. The intersection in a frame, f, for fish i is given by

$$\text{IBO}_{i,f} = \frac{1}{|\text{bb}_i|} \sum_{j=1}^{n_{\text{occ}}} \text{bb}_i \cap \text{bb}_j, \quad \text{for } j \neq i, \qquad (1)$$

where n_{occ} is the number of fish in an occlusion event, and bb_j is the set of pixel coordinates in the bounding box of fish j. IBO is measured across all bounding boxes with an occlusion tag in a given frame, even for subjects that are not
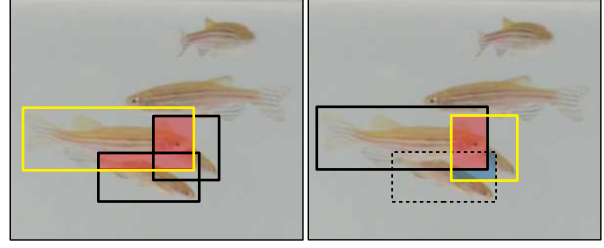

Figure 3: IBO seen from the perspective of two different individuals in the same frame. The targets are marked in yellow, the red area shows the intersection with a subject that is part of the same occlusion as the target, and the blue area shows the intersection with a subject that is not part of the same occlusion as the target.

part of the same occlusion. Two examples are presented in Figure 3, where the $\text{IBO}_{i,f}$ is calculated from the perspective of the targets enclosed in yellow. The blue area in the second example, represents the intersection with a subject that is not part of the same occlusion as the target. Additionally, the annotated bounding boxes enclose only the visible parts of the subjects. Thus, the actual intersection between the subjects can be higher if a large part of a fish is hidden. Nonetheless, the assumption is that a high IBO is an expression of heavy occlusion and vice versa. The IBO measure presented in Table 1 is an average between all fish in all frames. A single complexity measure is calculated per sequence, by combining the four proposed metrics by

$$\Psi = \frac{1}{n} \sum_{v}^{\{\text{T,F}\}} \frac{\text{OC}_v \, \text{OL}_v \, \text{IBO}_v}{\text{TBO}_v}, \qquad (2)$$

where n is the number of camera views and subscript T and F denote the top- and front-view, respectively. If a recording has no occlusions the complexity measure, Ψ, is zero; otherwise, the measure is in the interval $]0, \infty[$, where a larger value indicates a higher complexity.

4. Method

The pipeline of the proposed 3D tracker follows a modular tracking-reconstruction approach, where subjects are detected and tracked in each view before being triangulated and associated across views. This allows us to use the temporal information of the tracklets in the two views in the 3D association step in opposition to a reconstruction-tracking approach, where detections are triangulated before tracks are generated.

4.1. Object Detection in 2D

A consistent 2D point is needed in each view in order to create 3D trajectories. As the head is the only rigid part of

the body the tracking point is chosen to be located between the eyes of the fish. We present two simple methods to find the head-point of the fish: a naive approach, that does not require training, and a CNN based approach.

Naive: A background image, bg, is initially estimated for each view by taking the median of N_{bg} images sampled uniformly across the videos. Subsequently, the background is subtracted by calculating the absolute difference image, $fg = |im - bg|$. To locate the head of a fish in the top-view, the fg is binarized using the intermodes bimodal threshold algorithm [68]. The skeletonization approach of Zhang and Suen [69] is applied, and the endpoints are analyzed to locate the head of the fish. In the front-view the fg is binarized through the use of a histogram entropy thresholding method because the appearance of the fish cannot be approximated as bimodal. The head point is estimated as being either the center of the blob or one of the middle edge points of the blob along the minor axis of the detected bounding box. All three points are evaluated during the 3D reconstruction step, and the two points with the highest reprojection errors are discarded.

FRCNN-H: A Faster R-CNN [70] model has been trained for each view. The bounding boxes have been extracted from all the head-point annotations in the training sequences in order to train a head-detector model for each view. The bounding boxes have static diameters of 25 and 50 pixels for the top-, and front-view, respectively. The head-points are determined as the center of the detected bounding boxes which have a minimum confidence of c.

See the supplementary material for more detailed information on the detectors.

4.2. 2D Tracklet Construction

As zebrafish move erratically, it is difficult to set up a stable motion model. Therefore, we use a naive tracking-by-detection approach. The tracking is done by constructing a distance matrix between the detections in a frame and the last detections of current tracklets. The matrix is solved as a global optimization problem using the Hungarian algorithm [71]. Tracklets are deliberately constructed in a conservative manner, where robustness is encouraged above length. A new detection is only assigned to a tracklet located within a minimum distance, denoted δ_T and δ_F, for the top and front view respectively. If a tracklet has not been assigned a detection within a given amount of time, τ_k, the tracklet is terminated.

The ℓ_2 distance between the head detections is used in both views for the FRCNN-H method. However, the Mahalanobis distance between the center-of-mass is used for the front-view in the Naive method. This is due to the elliptical form of the zebrafish body, which can be utilized by setting the covariance matrix of the blob as the Mahalanobis matrix; as the fish is more likely to move along the major axis
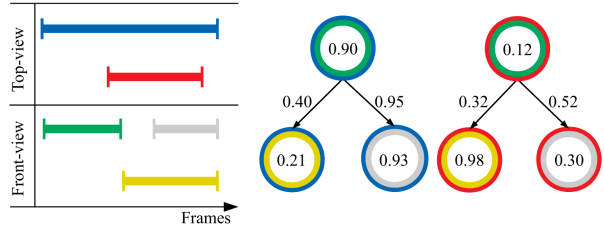

Figure 4: The colored lines represent 2D tracklets in each view, the node pairs are represented by the double-colored circles, and the edges of the DAG are shown by the arrows. The numbers represent example node and edge weights.

than along the minor axis.

4.3. 2D Tracklet Association Between Views

The 2D tracklets from each view are associated into 3D tracklets through a graph-based approach. All 2D tracklets with less than a given number of detections, α, are removed in order to filter out noisy tracklets. The 3D calibration and triangulation method from Pedersen *et al.* [67] is used.

4.3.1 Graph Construction

A directed acyclic graph (DAG) is constructed. Every node represents a 3D tracklet and consists of two 2D tracklets; one from each camera view. Each edge associates nodes, where the 3D tracklet is based on the same 2D tracklet from one of the views.

Create nodes: The graph nodes are constructed by processing each top-view tracklet and identifying all temporally intersecting front-view tracklets as given by

$$I = F_T \cap F_F, \qquad (3)$$

where F_T and F_F are the set of frames with detections in the top- and front-view tracklets, respectively, and I is the set of frames with detections in both views. If $I = \emptyset$, the node is not created.

An example is presented in Figure 4, where both the blue and red tracklets in the top-view intersects with the three tracklets in the front-view. The outer and inner circles of the six nodes represent the top- and front-view tracklets, respectively. The number inside the nodes indicates the node weight, which is calculated as follows.

For each intersecting frame in I, denoted f, the 2D tracklets are triangulated. This results in a 3D point of the zebrafish head, p_f, with a reprojection error, x_f. For the Naive method where the head is not directly detected in the front-view, the top-view 2D point is triangulated with the three estimated points to find the match resulting in the smallest reprojection error. Therefore, p_f represents the point with the smallest reprojection error. To penalize large reprojection

errors, the complimentary probability from the exponential cumulative distribution function (CDF), Φ, is utilized. The exponential CDF is chosen as it approximately models the reprojection error of the ground truth training data. The set of weights for all valid 3D points, V, can be described by the following set-builder notation

$$V = \{1 - \Phi(x_f \mid \lambda_{err}) \mid f \in I \wedge A(p_f)\}, \quad (4)$$

where λ_{err} is the reciprocal of the mean of the training data reprojection error, and A states whether p_f is within the water tank. The per-frame weights in V are combined into a single weight, W, for the entire node by

$$W = \text{median}(V)\frac{|V|}{|F_{\text{T}} \cup F_{\text{F}}|}, \quad (5)$$

and the node is added to the DAG given that $W \neq 0$. This weighting scheme considers both the reprojection error and the ratio of frames with valid 3D points compared to the set of all frames I. The median function is used instead of the mean function in order to counteract that a few extreme outliers skew the weight.

Connect nodes: The nodes in the DAG should be connected to all other nodes building on one of the same 2D tracklets, as long as the 2D tracklets in the other view do not overlap temporally, as illustrated in Figure 4. This is done by constructing the set of node pairs, P, from the set of nodes in the DAG, N. Each element of N, denoted n, consists of the 2D tracklets, t_{F} and t_{T}, the 3D tracklet, t, and the node weight, W. Nodes n_i and n_j are considered a pair if $t_{i,\text{T}} = t_{j,\text{T}}$ or $t_{i,\text{F}} = t_{j,\text{F}}$, if the 2D tracklets in the other view do not temporally overlap, and if t_i starts earlier in time than t_j. This is necessary in order to avoid assigning multiple detections to the same frame.

This can be represented by the set-builder notation

$$P = \{(n_i, n_j) \mid n_i, n_j \in N \wedge O(n_i, n_j) \wedge T(n_i, n_j)\}, \quad (6)$$

where O assesses whether t_i starts before t_j, and T ensures that the 2D tracklets in n_i and n_j do not temporally overlap, where $n = \{t_{\text{T}}, t_{\text{F}}, t, W\}$.

For each node pair in P, the weight, E, of the directed edge from n_i to n_j is based on:

- s, the speed of the fish as it moves between the last detection in t_i and the first detection in t_j.

- t_d, the temporal difference between t_i and t_j.

- W_i and W_j, the weights of the nodes.

The edge weight is calculated as the complimentary probability of the CDF of the exponential distribution, Φ. The exponential distribution is chosen as it approximately models that of the speed of the zebrafish. E is calculated by

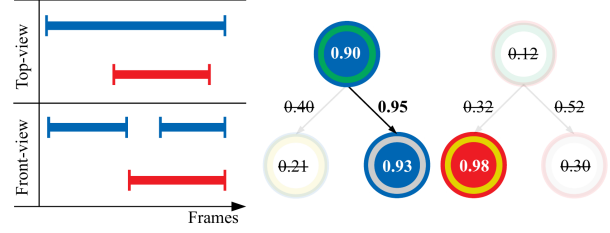$$E = (1 - \Phi(s \mid \lambda_s))e^{-\frac{t_d}{\tau_p}}(W_i + W_j), \quad (7)$$

Figure 5: Graph evaluation based on the example from Figure 4. The colored lines represent 2D tracklet pairs based on the chosen nodes in the graph; the transparent nodes are discarded.

where τ_p is an empirically chosen value, and λ_s is the reciprocal of the sum of the mean and standard deviation of the measured speed in the training data. In case a node is not present in any node pairs, the node will be assigned to the DAG, but it will have no edges. The DAG is therefore a disconnected graph.

4.3.2 Graph Evaluation

The final 3D tracklets are extracted from the constructed DAG; this is done by recursively finding the longest path in the graph and storing the set of nodes as a single 3D tracklet. The longest path is the path throughout the DAG, which gives the highest value when summing all nodes and edge weights in the path, see Figure 5. After extraction of a path, the used nodes, and all other nodes using the same 2D tracklets, are removed from the DAG. This process is repeated until the DAG is empty. In case a 2D tracklet in the 3D tracklet is missing a detection, the 3D position cannot be assigned, but the known information of the 2D tracklet is kept. For the Naive method, the head position of the front-view 2D tracklet is determined by assigning the estimated point, which minimizes the ℓ_2 distance to the head positions in the consecutive frame.

4.4. 3D Tracklet Association

The final 3D tracks are constructed from the 3D tracklets in a greedy manner. A set of tracklets equal to the amount of fish present, N_{fish}, is used as initial *main tracklets*. The remaining tracklets, denoted *gallery tracklets*, are assigned one by one to a single main tracklet, until no more tracklets can be assigned.

4.4.1 Initial Tracklet Selection

The set of N_{fish} in the main tracks is selected by finding the stable tracklets that are temporally concurrent in time and span long time intervals. For each tracklet, the set of other temporally concurrent tracklets is considered. In this set, all possible combinations of size N_{fish} are investigated. If
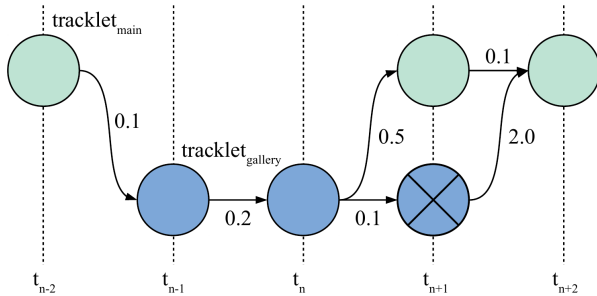
Figure 6: Example of internal spatio-temporal DAG, with the spatial distance between detections in the tracklets. The shortest path is found when switching from tracklet$_{\text{gallery}}$ to tracklet$_{\text{main}}$ in frame t_{n+1}.

all tracklets in the set overlap temporally, the set is saved as a valid tracklet set. The valid tracklet set with the highest median temporal overlap is used to construct N_{fish} full 3D tracks. This is done by using the greedy association scheme described in the following sections. No 3D tracks are created if no valid combination of size N_{fish} is identified.

4.4.2 Greedy Association

A greedy association algorithm is used when each gallery tracklet is associated with a single main tracklet. The greedy part of the algorithm concerns the way that gallery tracklets are chosen; all gallery tracks are ranked in ascending order by the shortest temporal distance to any main tracklet. If the gallery tracklet overlaps temporally with all main tracklets, it is relegated to the end of the list. When the gallery tracklet has been associated with a main track, the remaining gallery tracks are re-ranked, and the process repeated. In this way, the main tracklets are "grown" into full tracks. The gallery tracklet assignment is based on minimizing the cost of assignment. The cost is based on a set of distance measures, which are determined from two cases.

In the first case at least one main tracklet does not temporally overlap with the gallery tracklet. In this case, the association process is based on the spatio-temporal distances between the gallery tracklet and main tracklets. All temporally overlapping main tracklets are not considered.

In the second case the gallery tracklet overlaps temporally with all main tracklets. As the spatio-temporal distances between the main and gallery tracklet is no longer measurable, a different set of distance values are used: The internal spatio-temporal distances, the amount of intersecting frames, *i.e.* frames with a detection in both the main and gallery tracklets, and the ratio of intersecting frames compared to the total amount of detections in the gallery tracklet. The internal spatio-temporal distances are determined through the construction of a DAG, where each node is a detection in a frame, and the edge weights are the spatial

distances between the temporally previous nodes. The final path is the one minimizing the spatial distance traveled. An example of a graph is shown in Figure 6. The distances are calculated as the mean of the values when the graph switches from a detection in the gallery tracklet to the main tracklet and vice versa.

Association: The distance measures are consolidated into a single assignment decision through a global cost scheme. Each distance value is normalized across valid main tracklets into the range $[0; 1]$ and sum to 1. The final cost of assigning the gallery tracklet to a main tracklet, is obtained by calculating the mean of the normalized distance values. The gallery tracklet is associated with the main tracklet with the smallest cost, unless all main tracklet costs are located within a small margin, β, of each other, in which case the gallery tracklet is discarded. β directly enforces a margin of confidence in the assignment, in order to not assign a gallery traklet based on inconclusive cost values.

5. Evaluation

The metrics used in the MOT challenges [41, 42, 43] and the Mean Time Between Failures (MTBF) proposed by Carr and Collins [72] are utilized to measure the performance of the system on the proposed dataset. The MOT challenge metrics consist of the CLEAR MOT metrics [73], the mostly tracked/lost metrics [74], and the identification-based metrics [75].

The final 3D tracks are evaluated based on a subset of the MOT challenge metrics and the monotonic MTBF metric. The primary metric used is the multiple object tracking (MOTA) metric. The detected and ground truth tracklets are compared using the detected and annotated head points. A detection is only associated with a ground truth tracklet if it is within a distance of 0.5 cm. The performance of the system is evaluated with two different detection modules: Naive, and FRCNN-H. The results are compared with a hypothetical tracker, called Oracle, which tracks perfectly at all times except during occlusions. This provides an upper bound on the performance if occlusions are not handled in any way. The full set of metrics, system parameters, and results can be found in the supplementary material.

Results for all sequences compared to data complexity is shown in Figure 7, and metrics for the test sequences are shown in Table 2. It is clear that the FRCNN-H outperforms the Naive method on the training and validation splits; it even outperforms the Oracle tracker in three out of four cases. This is likely due to the method being able to detect some of the fish heads during occlusions. However, the superior performance is only seen on the two splits where the fish are from the same cohort. On the test set the FRCNN-H fails to generalize, whereas the Naive method still manages to track the fish.
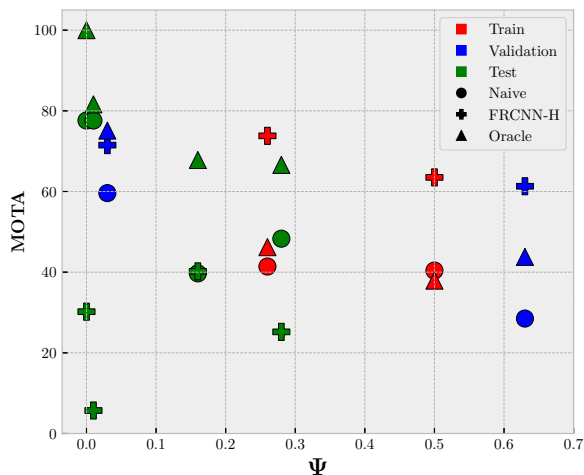
Figure 7: MOTA compared to the dataset complexity, Ψ, for all sequences in the dataset.

	Method	MOTA ↑	MT ↑	ML ↓	ID Sw. ↓	Frag. ↓	MTBF$_m$ ↑
Tst1	Naive	77.6%	1	0	0	28	12.5
	FRCNN-H	30.2%	0	0	0	15	8.212
	Oracle	100.0%	1	0	0	0	900
Tst2	Naive	77.6%	1	0	0	44	15.856
	FRCNN-H	5.7%	0	2	2	17	2.641
	Oracle	81.6%	2	0	0	25	27.396
Tst5	Naive	39.7%	0	0	7	185	6.249
	FRCNN-H	40.2%	0	0	7	115	7.577
	Oracle	67.8%	1	0	0	50	28.112
Tst10	Naive	48.3%	0	0	11	268	9.075
	FRCNN-H	25.2%	0	3	32	225	4.904
	Oracle	66.6%	1	10	0	119	23.105

Table 2: Evaluation of 3D tracks on test split. The arrows indicate whether higher or lower values are better. MOTA: Multiple Object Tracking Accuracy, MT: Mostly tracked, ML: Mostly lost, ID Sw.: Number of identity swaps, Frag.: Number of fragments, MTBF$_m$: Monotonic MTBF.

It should be noted that the poor performance of the Naive method on Tst1, is suspected to be due many short tracks from erratic movement, which the pipeline with the used parameter settings does not handle well.

5.1. Comparison with Other Methods

It has not been possible to make a fair comparison with the other 3D zebrafish tracking methods mentioned in Section 2. Previous systems have been analyzed in terms of ID swaps, fragments, precision, and recall for the generated 2D and 3D tracks. However, there is no exact description of how these metrics are calculated. The evaluation protocol is further limited by not including a statement on the maximum allowed distance between estimated and ground truth tracks leading to uncertainty on the accuracy of the metrics.

Furthermore, the evaluated sequences are not described in terms of complexity, even though occlusion is repeatedly stated as a major hindrance in 3D zebrafish tracking. The only common complexity indication of the datasets is the number of fish, even though it is not representative. An example of this is the tracking demo video of Qian *et al.* [62] with ten fish and only four occlusion events during 15 seconds. Wang *et al.* [30] describes their dataset on basis of an occlusion probability but do not explain how it is measured.

There are currently no publicly available annotated data and the previous systems are evaluated on seemingly simplified cases of the problem. Furthermore, the used evaluation protocols are lacking details in such a manner that it is not possible to determine under which conditions the metrics have been calculated. This, along with inaccessible codebases, severely limits the reproducibility of the results,

and makes it impossible to ensure identical evaluation procedures. Therefore, it simply does not make sense to compare the proposed system to the other methods under the current circumstances.

6. Conclusion

Zebrafish is an increasingly popular animal model and behavioral analysis plays a major role in neuroscientific and biological research. However, it is tedious and subjective to manually describe the complex 3D motion of zebrafish. Therefore, 3D zebrafish tracking systems are critically needed to conduct accurate experiments on a grand scale. The significant development experienced in other fields of MOT has not yet translated to 3D zebrafish tracking. The main reason being that no dataset has been made publicly available with ground truth annotations. Therefore, we present the first publicly available RGB 3D zebrafish tracking dataset called 3D-ZeF.

3D-ZeF consists of eight stereo sequences with highly social and similarly looking subjects demonstrating complex and erratic motion patterns in three dimensions that are not seen in common MOT challenges. A complexity measure based on the level of occlusions has been provided for each sequence to make them comparable to future related datasets. The proposed dataset is annotated with 86,400 bounding boxes and points; the latter used for estimating ground truth 3D tracks based on the head position of the fish. Different cohorts of zebrafish are used in the training, validation, and test splits to avoid data leakage; a problem that has never been addressed within the field.

The proposed Naive method scores a MOTA between 25% and 80% across the entire dataset, which correlates well with the complexity measure of the recordings. The open-source modular based system provides a baseline and stepping stone for further development within the field of 3D zebrafish tracking and understanding.

References

[1] J. S. Eisen, "Zebrafish make a big splash," *Cell*, vol. 87, pp. 969–977, Dec. 1996.

[2] P. Haffter, M. Granato, M. Brand, M. Mullins, M. Hammerschmidt, D. Kane, J. Odenthal, F. van Eeden, Y. Jiang, C. Heisenberg, R. Kelsh, M. Furutani-Seiki, E. Vogelsang, D. Beuchle, U. Schach, C. Fabian, and C. Nusslein-Volhard, "The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio," *Development*, vol. 123, no. 1, pp. 1–36, 1996.

[3] A. D. Collier, K. M. Khan, E. M. Caramillo, R. S. Mohn, and D. J. Echevarria, "Zebrafish and conditioned place preference: A translational model of drug reward," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 55, pp. 16–25, Dec. 2014.

[4] C.-Y. Lin, C.-Y. Chiang, and H.-J. Tsai, "Zebrafish and Medaka: new model organisms for modern biomedical research," *Journal of Biomedical Science*, vol. 23, Jan. 2016.

[5] M. C. Soares, S. C. Cardoso, T. d. S. Carvalho, and C. Maximino, "Using model fish to study the biological mechanisms of cooperative behaviour: A future for translational research concerning social anxiety disorders?," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 82, pp. 205–215, Mar. 2018.

[6] A. V. Kalueff, A. M. Stewart, and R. Gerlai, "Zebrafish as an emerging model for studying complex brain disorders," *Trends in Pharmacological Sciences*, vol. 35, pp. 63–75, Feb. 2014.

[7] D. A. Meshalkina, M. N. Kizlyk, E. V. Kysil, A. D. Collier, D. J. Echevarria, M. S. Abreu, L. J. G. Barcellos, C. Song, J. E. Warnick, E. J. Kyzar, and A. V. Kalueff, "Zebrafish models of autism spectrum disorder," *Experimental Neurology*, vol. 299, pp. 207–216, Jan. 2018.

[8] K. M. Khan, A. D. Collier, D. A. Meshalkina, E. V. Kysil, S. L. Khatsko, T. Kolesnikova, Y. Y. Morzherin, J. E. Warnick, A. V. Kalueff, and D. J. Echevarria, "Zebrafish models in neuropsychopharmacology and CNS drug discovery," *British Journal of Pharmacology*, vol. 174, no. 13, pp. 1925–1944, 2017.

[9] L. Li and J. E. Dowling, "A dominant form of inherited retinal degeneration caused by a non-photoreceptor cell-specific mutation," *Proceedings of the National Academy of Sciences*, vol. 94, pp. 11645–11650, Oct. 1997.

[10] U. K. Muller, "Swimming of larval zebrafish: ontogeny of body waves and implications for locomotory development," *Journal of Experimental Biology*, vol. 207, pp. 853–868, Feb. 2004.

[11] M. B. McElligott and D. M. O'Malley, "Prey tracking by larval zebrafish: Axial kinematics and visual control," *Brain, Behavior and Evolution*, vol. 66, no. 3, pp. 177–196, 2005.

[12] E. Fontaine, D. Lentink, S. Kranenbarg, U. K. Müller, J. L. van Leeuwen, A. H. Barr, and J. W. Burdick, "Automated visual tracking for studying the ontogeny of zebrafish swimming," *Journal of Experimental Biology*, vol. 211, no. 8, pp. 1305–1316, 2008.

[13] B. Risse, D. Berh, N. Otto, C. Klämbt, and X. Jiang, "FIM-Track: An open source tracking and locomotion analysis software for small animals," *PLOS Computational Biology*, vol. 13, no. 5, pp. 1–15, 2017.

[14] S. Ohayon, O. Avni, A. L. Taylor, P. Perona, and S. E. R. Egnor, "Automated multi-day tracking of marked mice for the analysis of social behaviour," *Journal of Neuroscience Methods*, vol. 219, no. 1, pp. 10 – 19, 2013.

[15] Z.-M. Qian, X. E. Cheng, and Y. Q. Chen, "Automatically detect and track multiple fish swimming in shallow water with frequent occlusion," *PLOS ONE*, vol. 9, no. 9, pp. 1–12, 2014.

[16] G. d. O. Feijó, V. A. Sangalli, I. N. L. d. Silva, and M. S. Pinho, "An algorithm to track laboratory zebrafish shoals," *Computers in Biology and Medicine*, vol. 96, pp. 79 – 90, 2018.

[17] J. E. Franco-Restrepo, D. A. Forero, and R. A. Vargas, "A review of freely available, open-source software for the automated analysis of the behavior of adult zebrafish," *Zebrafish*, vol. 16, June 2019.

[18] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. H. Heras, and G. G. d. Polavieja, "idtracker.ai: tracking all individuals in small or large collectives of unmarked animals," *Nature Methods*, vol. 16, pp. 179–182, Jan. 2019.

[19] L. P. J. J. Noldus, A. J. Spink, and R. A. J. Tegelenbosch, "EthoVision: A versatile video tracking system for automation of behavioral experiments," *Behavior Research Methods, Instruments, & Computers*, vol. 33, pp. 398–414, Aug. 2001.

[20] Loligo Systems, "LoliTrack v.4." https://www.loligosystems.com/lolitrack-v-4.

[21] ViewPoint, "ZebraLab." http://www.viewpoint.fr/en/p/software/zebralab.

[22] TSE Systems, "VideoMot2 - versatile video tracking system." https://www.tse-systems.com/product-details/videomot.

[23] A. V. Kalueff, M. Gebhardt, A. M. Stewart, J. M. Cachat, M. Brimmer, J. S. Chawla, C. Craddock, E. J. Kyzar, A. Roth, S. Landsman, S. Gaikwad, K. Robinson, E. Baatrup, K. Tierney, A. Shamchuk, W. Norton, N. Miller, T. Nicolson, O. Braubach, C. P. Gilman, J. Pittman, D. B. Rosemberg, R. Gerlai, D. Echevarria, E. Lamb, S. C. F. Neuhauss, W. Weng, L. Bally-Cuif, H. Schneider, and t. Z. Neuros, "Towards a comprehensive catalog of zebrafish behavior 1.0 and beyond," *Zebrafish*, vol. 10, pp. 70–86, Mar. 2013.

[24] J. M. Cachat, P. R. Canavello, S. I. Elkhayat, B. K. Bartels, P. C. Hart, M. F. Elegante, E. C. Beeson, A. L. Laffoon, W. A. Haymore, D. H. Tien, A. K. Tien, S. Mohnot, and A. V. Kalueff, "Video-aided analysis of zebrafish locomotion and anxiety-related behavioral responses," in *Zebrafish Neurobehavioral Protocols* (A. V. Kalueff and J. M. Cachat, eds.), pp. 1–14, Totowa, NJ: Humana Press, 2011.

[25] S. Macrì, D. Neri, T. Ruberto, V. Mwaffo, S. Butail, and M. Porfiri, "Three-dimensional scoring of zebrafish behavior unveils biological phenomena hidden by two-dimensional analyses," *Scientific Reports*, vol. 7, May 2017.

[26] N. Miller and R. Gerlai, "Quantification of shoaling behaviour in zebrafish (Danio rerio)," *Behavioural Brain Research*, vol. 184, pp. 157–166, Dec. 2007.

[27] H. AlZu'bi, W. Al-Nuaimy, J. Buckley, L. Sneddon, and Iain Young, "Real-time 3D fish tracking and behaviour analysis," in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp. 1–5, Nov. 2015.

[28] X. E. Cheng, S. S. Du, H. Y. Li, J. F. Hu, and M. L. Chen, "Obtaining three-dimensional trajectory of multiple fish in water tank via video tracking," *Multimedia Tools and Applications*, vol. 77, pp. 24499–24519, Feb. 2018.

[29] Z. Qian, M. Shi, M. Wang, and T. Cun, "Skeleton-based 3D tracking of multiple fish from two orthogonal views," in *Communications in Computer and Information Science*, pp. 25–36, Springer Singapore, 2017.

[30] S. H. Wang, X. Liu, J. Zhao, Y. Liu, and Y. Q. Chen, "3D tracking swimming fish school using a master view tracking first strategy," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2016.

[31] Z.-M. Qian and Y. Q. Chen, "Feature point based 3D tracking of multiple fish from multi-view images," *PLOS ONE*, vol. 12, no. 6, pp. 1–18, 2017.

[32] G. Audira, B. Sampurna, S. Juniardi, S.-T. Liang, Y.-H. Lai, and C.-D. Hsiao, "A simple setup to perform 3D locomotion tracking in zebrafish by using a single camera," *Inventions*, vol. 3, p. 11, Feb. 2018.

[33] G. Xiao, W.-K. Fan, J.-F. Mao, Z.-B. Cheng, D.-H. Zhong, and Y. Li, "Research of the fish tracking method with occlusion based on monocular stereo vision," in *2016 International Conference on Information System and Artificial Intelligence (ISAI)*, IEEE, June 2016.

[34] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," in *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[35] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: multi-object tracking and segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: a multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.

[37] "Waymo open dataset: An autonomous driving dataset," 2019. https://www.waymo.com/open.

[38] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3D tracking and forecasting with rich maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757, 2019.

[39] S. Song and J. Xiao, "Tracking revisited using RGBD camera: Unified benchmark and baselines," in *Proceedings of the IEEE international conference on computer vision*, pp. 233–240, 2013.

[40] A. Lukezic, U. Kart, J. Kapyla, A. Durmush, J.-K. Kamarainen, J. Matas, and M. Kristan, "Cdtb: A color and depth visual object tracking dataset and benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10013–10022, 2019.

[41] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942.

[42] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[43] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "CVPR19 tracking and detection challenge: How crowded can it get?," *arXiv:1906.04567 [cs]*, June 2019. arXiv: 1906.04567.

[44] S. Lyu, M. Chang, D. Du, L. Wen, H. Qi, Y. Li, Y. Wei, L. Ke, T. Hu, M. Del Coco, P. Carcagnì, D. Anisimov, E. Bochinski, F. Galasso, F. Bunyak, G. Han, H. Ye, H. Wang, K. Palaniappan, K. Ozcan, L. Wang, L. Wang, M. Lauer, N. Watcharapinchai, N. Song, N. M. Al-Shakarji, S. Wang, S. Amin, S. Rujikietgumjorn, T. Khanova, T. Sikora, T. Kutschbach, V. Eiselein, W. Tian, X. Xue, X. Yu, Y. Lu, Y. Zheng, Y. Huang, and Y. Zhang, "UA-DETRAC 2017: Report of AVSS2017 IWT4S challenge on advanced traffic monitoring," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–7, Aug 2017.

[45] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020.

[46] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[47] E. Bochinski, T. Senst, and T. Sikora, "Extending IOU based multi-object tracking by visual information," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, Nov 2018.

[48] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[49] A. Perez-Escudero, J. Vicente-Page, R. C. Hinz, S. Arganda, and G. G. de Polavieja, "idTracker: tracking individuals in a group by automatic identification of unmarked animals," *Nature Methods*, vol. 11, pp. 743 – 748, 2014.

[50] V. H. Sridhar, D. G. Roche, and S. Gingins, "Tracktor: Image-based automated tracking of animal movement and behaviour," *Methods in Ecology and Evolution*, vol. 10, pp. 815–820, Mar. 2019.

[51] H. J. Mönck, A. Jörg, T. v. Falkenhausen, J. Tanke, B. Wild, D. Dormagen, J. Piotrowski, C. Winklmayr, D. Bierbach, and T. Landgraf, "BioTracker: an open-source computer vision framework for visual animal tracking," *CoRR*, vol. abs/1803.07985, 2018.

[52] A. Rodriguez, H. Zhang, J. Klaminder, T. Brodin, P. L. Andersson, and M. Andersson, "ToxTrac: a fast and robust software for tracking organisms," *Methods in Ecology and Evolution*, vol. 9, pp. 460–464, Sept. 2017.

[53] A. M. T. Harmer and D. B. Thomas, "pathtrackr: An r package for video tracking and analysing animal movement," *Methods in Ecology and Evolution*, May 2019.

[54] X. Liu, P. R. Zhu, Y. Liu, and J. W. Zhao, "Tracking full-body motion of multiple fish with midline subspace constrained multicue optimization," *Scientific Programming*, vol. 2019, pp. 1–7, June 2019.

[55] L. Zhu and W. Weng, "Catadioptric stereo-vision system for the real-time monitoring of 3D behavior in aquatic animals," *Physiology & Behavior*, vol. 91, no. 1, pp. 106 – 119, 2007.

[56] J. Cachat, A. Stewart, E. Utterback, P. Hart, S. Gaikwad, K. Wong, E. Kyzar, N. Wu, and A. V. Kalueff, "Three-dimensional neurophenotyping of adult zebrafish behavior," *PLOS ONE*, vol. 6, no. 3, pp. 1–14, 2011.

[57] X. E. Cheng, S. H. Wang, and Y. Q. Chen, "3D tracking targets via kinematic model weighted particle filter," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2016.

[58] S. V. Viscido, J. K. Parrish, and D. Grünbaum, "Individual behavior and emergent properties of fish schools: a comparison of observation and theory," *Marine Ecology Progress Series*, vol. 273, pp. 239–249, 2004.

[59] A. D. Straw, K. Branson, T. R. Neumann, and M. H. Dickinson, "Multi-camera real-time three-dimensional tracking of multiple flying animals," *Journal of The Royal Society Interface*, vol. 8, pp. 395–409, July 2010.

[60] K. Müller, J. Schlemper, L. Kuhnert, and K. D. Kuhnert, "Calibration and 3D ground truth data generation with orthogonal camera-setup and refraction compensation for aquaria in real-time," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3, pp. 626–634, Jan. 2014.

[61] A. M. Stewart, F. Grieco, R. A. J. Tegelenbosch, E. J. Kyzar, M. Nguyen, A. Kaluyeva, C. Song, L. P. J. J. Noldus, and A. V. Kalueff, "A novel 3D method of locomotor analysis in adult zebrafish: Implications for automated detection of CNS drug-evoked phenotypes," *Journal of Neuroscience Methods*, vol. 255, pp. 66–74, Nov. 2015.

[62] Z.-M. Qian, S. H. Wang, X. E. Cheng, and Y. Q. Chen, "An effective and robust method for tracking multiple fish in video image based on fish head detection," *BMC Bioinformatics*, vol. 17, p. 251, June 2016.

[63] X. Liu, Y. Yue, M. Shi, and Z.-M. Qian, "3-D video tracking of multiple fish in a water tank," *IEEE Access*, vol. 7, pp. 145049–145059, 2019.

[64] S. H. Wang, X. E. Cheng, and Y. Q. Chen, "Tracking undulatory body motion of multiple fish based on midline dynamics modeling," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2016.

[65] S. H. Wang, J. Zhao, X. Liu, Z. Qian, Y. Liu, and Y. Q. Chen, "3D tracking swimming fish school with learned kinematic model using LSTM network," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1068–1072, March 2017.

[66] S. H. Wang, J. W. Zhao, and Y. Q. Chen, "Robust tracking of fish schools using CNN for head identification," *Multimedia Tools and Applications*, pp. 1–19, 2016.

[67] M. Pedersen, S. Hein Bengtson, R. Gade, N. Madsen, and T. B. Moeslund, "Camera calibration for underwater 3D reconstruction based on ray tracing using Snell's law," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[68] J. M. S. Prewitt and M. L. Mendelsohn, "The analysis of cell images," *Annals of the New York Academy of Sciences*, vol. 128, no. 3, pp. 1035–1053, 1966.

[69] T. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, 1984.

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[71] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[72] P. Carr and R. T. Collins, "Assessing tracking performance in complex scenarios using mean time between failures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, March 2016.

[73] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, p. 246309, May 2008.

[74] Bo Wu and R. Nevatia, "Tracking of multiple, partially occluded humans based on static body part detection," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 951–958, June 2006.

[75] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Computer Vision – ECCV 2016 Workshops* (G. Hua and H. Jégou, eds.), (Cham), pp. 17–35, Springer International Publishing, 2016.