DOI: https://doi.org/10.15276/hait.08.2025.30
UDC 378.147:004.6:331.5

Unsupervised Re-identification architecture based on segmented tracklets for animal behavior analysis

Natalia P. Volkova[1)]
ORCID: https://orcid.org/0000-0003-3175-2179; volkova.n.p@op.edu.ua. Scopus Author ID: 36104775700
Maksym A. Shvandt[1)]
ORCID: https://orcid.org/0000-0002-4580-3961; maxim.shvandt@gmail.com
[1)] Odesa National Polytechnic University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

In this paper, we present Mask-TAUDL, an advanced unsupervised re-identification architecture that combines instance segmentation, unsupervised deep learning, and tracklet association for detailed analysis of object behavior in long-term recordings. It combines the Mask R-CNN dual-stream detector/segmenter with dual ResNet-18 backbones and the unsupervised deep learning module based on tracklet association (TAUDL). Mask R-CNN provides accurate object localization and binary masks from which we construct tracklets with improved segmentation. The two ResNet-18 streams use these masks to extract appearance and motion-sensitive features at the tracklet level, which are combined into a common feature descriptor. The TAUDL module operates directly on the masked tracklet features and co-trains discriminative embeddings and cross-session associations without manual labeling. The proposed Mask-TAUDL architecture trains a model so that features of a single individual remain close in embedding space over time, while providing a clear separation of features between different individuals. Integrating pure masked regions with temporally aggregated features helps suppress spurious variations caused by shadows, reflections, or overlapping objects. Long-term animal re-identification is challenging due to frequent overlaps, appearance drift, and subtle visual differences between individuals, and most existing solutions rely on large annotated datasets, which limits their applicability in real-world laboratory settings. The Mask-TAUDL architecture overcomes these limitations by explicitly modeling temporally consistent, mask-refined tracks and training embeddings that preserve identity in a fully unsupervised manner. Mask-TAUDL is designed for animal behavior studies, namely small laboratory species such as mice and fish observed in closed or semi-structured arenas, where reliable long-term identity tracking is essential for quantitative behavioral analysis, longitudinal experiments, and high-throughput screening.

Keywords: Identification; segmentation; image processing; unsupervised re-identification; deep learning; TAUDL; Mask R-CNN; tracklets; architecture; animal behavior tracking

***For citation*: Volkova N. P., Shvandt M. A. "Unsupervised Re-identification architecture based on segmented tracklets for animal behavior analysis". *Herald of Advanced Information Technolog*. 2025; Vol.8 No.4: 476–487. DOI: https://doi.org/10.15276/hait.08.2025.30**

INTRODUCTION

Computer vision is now a core technology for automating a wide range of visual workflows. Many computer vision pipelines initially perform object detection, followed by temporal tracking to enable subsequent quantitative analysis. Such steps are used in a variety of industries, including surveillance, military, and transportation systems [1], [2], as well as in the life sciences, particularly in animal behavior studies in the laboratory and field [1], [3], [4], [5], [6]. In this field, research is particularly active, focusing on the detailed analysis of animal behavior, with fish and mice serving as common model organisms for studying locomotion, social interaction, and responses to environmental perturbations [1], [3], [4], [5], [6].

In such experimental studies, multiple animals are often in the same field of view, often overlapping or partially covering each other, making it necessary to distinguish individual instances within a single frame. To address this problem, current approaches combine object detection with instance-level segmentation to obtain accurate pixel masks for each animal and reduce the influence of complex or dynamic backgrounds [7], [8]. Based on these approaches, multifunctional animal tools such as DeepLabCut, SLEAP, and AlphaTracker have been developed, which integrate pose estimation, instance association, and behavioral analysis to obtain advanced kinematic and interactive measures of animal movement and interaction from video data [4], [5], [6].

However, when the goal of the study is to track specific individuals over time, precise detection and segmentation are not sufficient. This is important because many behavioral paradigms aimed at studying animal movement, interaction, or response require the sequential tracking and assignment of an identity to each animal across frames, trials, and recording sessions. The process of re-establishing the identity of individuals (re-identification (Re-ID)) is crucial for reconstructing continuous trajectories, assessing individual activity levels, characterizing

social interactions, and quantifying longitudinal changes in animal behavior. Re-ID in animal datasets is particularly challenging: individuals are often visually similar, undergo substantial non-rigid deformations, and experience frequent occlusions and crossings; lighting, reflections, and turbidity further complicate underwater recordings of fish, while rapid postural changes and self-occlusions are common in mice. Without robust re-identification, tracking systems suffer identity switches, fragmented tracks, and biased metrics (for example, artificially increased speed due to broken trajectories or incorrect interaction counts). Embedding-based re-ID models, trained to map image crops into a discriminative feature space, alleviate these issues by maintaining identity through appearance changes. When coupled with accurate segmentation masks, these models can focus on animal pixels while suppressing background and noise, thereby improving identity preservation in demanding multi-animal tracking scenarios [2], [4], [5], [6].

STATEMENT OF THE PROBLEM

Mice and fish are widely used model organisms in both ecotoxicology and ethology. Much of this research relies on monitoring animal behavior in confined or controlled conditions during the initial stages of the study. For rodents, a typical example of such controlled conditions is a perforated test box containing several holes. A camera is rigidly mounted above this box, and the animals are continuously recorded for extended periods of time (approximately 30 minutes to several hours). During these sessions, investigators quantify how often each animal moves between holes and how many times it investigates or peers into a hole, which serves as an indicator of exploratory activity and territory assessment (Fig.1a, b, c).

Other controlled conditions are used in the study of fish behavior. In our study, the focus is on the behavior of gobies. In this case, the closed test arena is a square aquarium filled with natural seawater to approximate the controlled conditions to real-world environmental conditions. Oxygen is supplied through tubes, creating a gently aerated environment. A stationary camera is installed above the aquarium and records the behavior of the gobies for extended periods (from several hours to a whole day). For practical reasons, continuous recordings are broken into 30-minute video clips to facilitate data storage management and further processing (Fig.2a, b, c).

The aquarium can contain up to 10 individual fish simultaneously. Among a wide range of possible behavioral characteristics, the current study focuses on the following indicators:

– the total number of movements (changes of position) for each individual;

– the aggregate number of movements across the entire group;

– the number of attacks or agonistic encounters between pairs of fish, defined as a directed movement of subject A toward subject B that leads to B rapidly moving away in another direction;

– preservation of a consistent identity label for each individual throughout the entire recording period. This last requirement is especially critical when two or three different bullhead species or morphs are placed in the same tank.

a

b

c

Fig. 1. **Mice/rats behaviour study:**
***a* – test environment (box);**
***b* – screenshot of a video with a lab rat;**
***c* – screenshot of a video with lab mice**
Source: **compiled by the authors**

a

b

c

Fig. 2. **Fish (gobies) behaviour study:**
***a* – test environment (aquarium);**
***b, c* – screenshots from videos with gobies**
Source: **compiled by the authors**

Currently, behavioral assessment is performed manually, by reviewing and annotating video sequences. This approach is extremely time-consuming and prone to subjective errors, as all behavioral patterns are detected and classified solely visually. Consequently, there is a strong need for an automated algorithm capable of detecting, tracking, and analyzing the fish trajectories and interactions with minimal human intervention.Several factors, however, make this task nontrivial. First, although the position of an individual fish cannot change dramatically between two or three consecutive frames, its direction of motion can vary abruptly, leading to sudden turns that complicate tracking. Second, despite the overall stability of the experimental setup, the background in the aquarium is far from constant. Food is introduced before and during the trials, and waste products from the gobies accumulate on the tank bottom. Both food particles and waste form mobile clumps that are transported by water currents generated by swimming fish and by the oxygen flow from the pipes. These clumps can reach substantial size and often share similar coloration with some of the gobies, making it difficult to localize fish when they overlap spatially with such debris. This also rules out simple strategies such as color-based tracking. All of these aspects must be carefully considered when designing a robust, fully automated algorithm for long-term behavior analysis in this environment.

RELATED WORKS

Object re-identification (Re-ID) has become a key component of modern multi-object tracking (MOT) systems. In tracking-by-detection pipelines, detectors provide per-frame bounding boxes, while Re-ID models generate appearance embeddings that support association of detections over time and, in some cases, across multiple cameras. As scenes become more crowded and objects visually similar, simple frame-level matching is often not sufficient, and many recent methods exploit tracklets, short temporal sequences of detections, to integrate richer appearance, motion, and temporal information. A number of works have proposed approaches focused on tracklet or video-level re-identification, as they are particularly relevant for constructing long, identity-consistent trajectories in a variety of fields, from human tracking to vehicle monitoring and animal behavior analysis.

There are five main approaches that comprise modern object re-identification (Re-ID) methods. These approaches differ in the way they encode identity, the use of temporal information, and the ability to cope with inter-camera variations.

The first approach is based on different levels of tracks and hierarchical/global association, which replaces frame-by-layer mapping with sequence-centric identity inference. In monocular tracking, short tracks (tracklets) are first formed, after which

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

deep models evaluate their similarity by combining appearance features with spatiotemporal motion and continuity characteristics. This allows for long and occlusion-resistant trajectories, albeit at increased computational cost [9]. Further work develops this idea, improving the integration of motion and appearance for objects with complex kinematics and extending the approach to multi-view and multi-camera systems, where identity preservation benefits from tracklet-oriented formulations [10], [11]. When tracking multiple objects in sports video, the global track association (GTA) method is used. Given a set of tracks from a base tracker, the method builds a graph in which edges reflect appearance similarity and spatiotemporal appropriateness. Further, global optimization allows detecting segments with mixed identities and re-associating those belonging to the same object. GTA significantly reduces identity switching and improves metrics (IDF1, HOTA), so it has become a standard component in modern sports benchmarks [12], [13], [14].

A graph formulation for multi-camera systems is also close in nature, where tracklets act as nodes and edges are defined by appearance similarity and spatiotemporal constraints, including motion patterns between cameras. Graph optimization through clustering or mapping provides consistent identity representation in a network of non-overlapping cameras; modern modifications are adapted for scenarios with strict latency constraints [15], [16].

In a more sophisticated version, the style of the tracklets is agreed at the tracklet level: the models transform the tracklets into a common visual domain, normalizing the differences in color and illumination between cameras while preserving temporal continuity. In parallel, part-oriented features are aggregated, which increases the robustness to partial overlaps and pose changes. This approach effectively smooths the heterogeneity of the camera network, although it increases the computational cost and creates a dependence on precise localization of features [17]. This idea is based on lighter variants, where the methods of transferring styles directly in the feature space, rather than at the image level, making the module easily pluggable into existing Re-ID pipelines [18]. Both variants emphasize that handling appearance variations between cameras is critical for reliable multi-camera tracking.

A third approach is based on reinforced attention and embedded learning to improve discrimination under occlusion and overlap conditions. An attention module integrated into the metric-learning Re-ID branch enhances identity-determining regions (e.g., clothing textures) while suppressing the background, reducing identity switching in dense scenes, while still relying on supervised identity labels and careful triplet sampling [19]. Further developments extend this idea by integrating multiple attention modules (heads) and multi-branch architectures, allowing for the integration of motion and appearance cues and stabilizing matching in complex environments [20]. Other variants combine attention with specialized motion modeling for small or fast objects where visual evidence is weak or noisy, demonstrating significant improvements in the robustness of Re-ID under partial visibility conditions [21].

A fourth approach is based on motion-aware sequence modeling, explicitly encoding temporal dynamics. A canonical example is the two-stream AMOC network, which combines two branches: one branch processes RGB appearance and the other processes optical flow motion via spatiotemporal convolutions and temporal aggregation (e.g., recurrent units); this yields embeddings that are sensitive to both “how the object looks” and “how it moves,” improving resolution in the presence of gait or pose variations and short occlusions [22]. Although such models are computationally more difficult compared to other newer lightweight or transform architectures, AMOC’s understanding of the temporal accumulation of motion information over tracklets has influenced modern video-based and unsupervised approaches that combine motion modeling with tracklet clustering and spatiotemporal correlations [23].

The fifth approach is based on unsupervised re-identification, which eliminates the dependence on identification labels. Methods of this approach include unsupervised training methods based on the TAUDL principle (Tracklet Association Unsupervised Deep Learning). Training occurs on automatically generated tracklets, combining the tracklet distinctions for each camera (surrogate labels for intra-camera distinctions) with the aim of inter-camera association, which aligns the embedding of probable tracklets with the same identity between cameras; joint optimization of these losses, based on the aggregation of frame features in tracklets, ensures the scalability of Re-ID training to large deployments, but remains sensitive to tracklet fragmentation and mixed identities [24]. Further research strengthens this paradigm by gradually improving pseudo-labels with cumulative motion context and spatiotemporal tracklet correlation [23], and using contrastive targets and dynamic tracklet

clustering to detect label-free embedding patterns [25]. Related efforts further address the issue of robustness in unsupervised mode, emphasizing robustness to noise in video Re-ID [26]. These offline methods are complemented by self-supervised online formulations: for example, OUTrack updates its Re-ID network during tracking, using temporal coherence and occlusion estimation as self-control, allowing adaptation to the scene but risking drift in the case of incorrect early associations [27]. Later studies apply self-supervision to online joint detection and embedding: embeddings learn from temporal and cyclic consistency only from the tracking history [28], while JDE-style detectors incorporate unlabeled Re-ID branches that adapt via tracking feedback, achieving strong trade-offs between speed and accuracy [29]. Overall, the unsupervised and self-supervised learning approach addresses the cost and inflexibility of labeled datasets while meeting the realities of long-term, dynamic environments.

Several interrelated patterns emerge from these approaches. First, sequence-level modeling—whether hierarchical tracklet aggregation, graph optimization, or temporal aggregation—consistently enhances identity continuity and occlusion robustness, which is especially important for crowded scenes and multi-camera systems [9], [10], [11], [12], [13], [14], [15], [16], [22]. Second, the quality of local features computed at the frame-by-frame level still matters: attention mechanisms and camera domain adaptation reduce noise from redundant surroundings and domain shift, which directly improves tracklet-level association [17], [18], [19], [20], [21]. Third, unsupervised and self-supervised targets reveal the scalability and adaptability of the scene by allowing systems to learn from the very trajectories they construct—offline via pseudo-supervision focused on tracklets [23], [24], [25], [26] and online via feedback from the tracker [27], [28], [29]. Modern systems are increasingly hybrid, combining global sequential analysis (for long-term consistency), attention and adaptation (for frame discrimination), motion modeling (for dynamics), and unsupervised learning (for scaling and deployment). Such combinations are particularly effective in domains with frequent exits/re-entries and strong structural constraints (sports and traffic), where global track association and graph formulations are a natural complement to adaptive, unsupervised embeddings [12], [13], [14], [15], [16].

From the above review, it is clear that current approaches to MOT and Re-ID have made significant progress, but they remain focused mainly on people and vehicles and do not take into account the specifics of fish movement and morphology. This necessitates the need for an adapted, improved marker-free approach to analyze fish behavior based on segmented tracklets.

THE AIM OF WORK AND RESEARCH METHODOLOGY

Within the framework of the conducted experimental studies, a generalized pipeline for processing video data and behavioral analysis of animals was formed, which includes the following stages.

1. Data input. Presentation of video recordings of the experiment and, if available, a reference image of the environment.

2. Performing video preprocessing.

Correction of lighting, shadows and noise; formation of additional frame representations (grayscale, shadow correction, color filtering, etc.) that enhance the contrast of objects relative to the background.

3. Formation of difference frame representations. The image subtraction operation is used to isolate moving objects.

4. Performing segmentation.

5. Performing segment filtering and binarization. Threshold processing of segments to eliminate background fragments and applying morphological operations to remove residual artifacts.

6. Localization of objects in the frame.

7. Refined detection and re-identification. Using a neural network detector to refine object boundaries, solve merging and intersection problems, and perform markerless re-identification (re-ID).

The aim of this work is to develop an architecture for unsupervised re-identification adapted for the analysis of animal behavior in limited or controlled experimental conditions, which is implemented within the seventh stage of the pipeline.

The developed architecture should provide accurate and long-term animal tracking with identity preservation in conditions of overlap, lighting changes, and background noise. To achieve this goal, the following tasks were formulated:

– to construct a dual-backbone Mask R-CNN detector/segmenter that employs two parallel ResNet-18 branches (“twin-ResNet18”) fused via squeeze-and-excitation-based concatenation, thereby increasing robustness to illumination changes,

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

reflections, and low-contrast textures typical of enclosed arenas;

– to design an unsupervised re-identification (re-ID) module that produces 256-D L2-normalized embeddings from mask-weighted ROI features, capturing morphology- and texture-sensitive representations while attenuating background interference;

– to embed TAUDL-style unsupervised learning with Per-Camera Tracklet Discrimination (PCTD) and Cross-Camera Tracklet Association (CCTA), enabling label-free, long-term identity preservation and reliable reconstruction of behavioral trajectories.

We propose an unsupervised re-ID framework for animal behavior analysis that integrates a customized dual-backbone Mask R-CNN detector/segmenter (Fig. 3) [30] with tracklet-association-based unsupervised deep learning in the spirit of TAUDL [17]. The intended use case is identity-consistent tracking of animals in enclosed habitats (e.g., aquaria, test boxes), where variable lighting, reflections, and partial occlusions frequently disrupt appearance stability. The method tightly couples instance-level segmentation with a re-ID branch trained on automatically generated tracklets, obviating the need for manual identity annotations while maintaining long-duration trajectories required for downstream behavioral readouts such as total distance traveled, interaction metrics, or zone occupancy patterns.

The basic architecture of Mask R-CNN contains one ResNet-18/34/50 pipeline. Figure 3 shows the corresponding convolution blocks C2-C5 of the standard ResNet-18 network. At the level of the C2 block, low-level features with higher spatial resolution are calculated, at the C3 level, medium-level features are calculated, at the C4 and C5 levels - the deepest features with the highest semantic abstraction and the smallest spatial map. These blocks are identical in their principle of operation to those that are part of the standard ResNet-50 that is included in the basic Mask R-CNN [9], [30].

The proposed architecture contains two synchronized ResNet-18 pipelines (“twin-ResNet18”) (Fig. 4), where levels C2_a, C3_a, C4_a, C5_a denote the corresponding convolutional blocks for the first ResNet-18 network (branch A), and C2_b, C3_b, C4_b, C5_b denote the convolutional blocks for the second ResNet-18 (branch B). The main branch A processes the RGB input signal, while the auxiliary branch B is designed to obtain additional derivative channels (for example, images with illumination correction or noise reduction, images with background subtraction or infrared modalities). The red color shows the feature fusion process between branches A and B at the corresponding block levels from branch B to branch A. The fusion mechanism (concatenation of compression and excitation) allows the network to selectively emphasize the most informative features from both streams, improving robustness to aquarium glare, water turbidity, and weak contrast in fur or skin patterns. For each detected image instance, mask-weighted ROI features are extracted and passed through a lightweight re-identification head, which includes global averaging, batch normalization, and a fully connected layer, to obtain a 256-dimensional L2-normalized embedding. Mask-based fusion allows for the reduction of background noise, creating descriptors that more accurately encode the shape and texture of the animal.

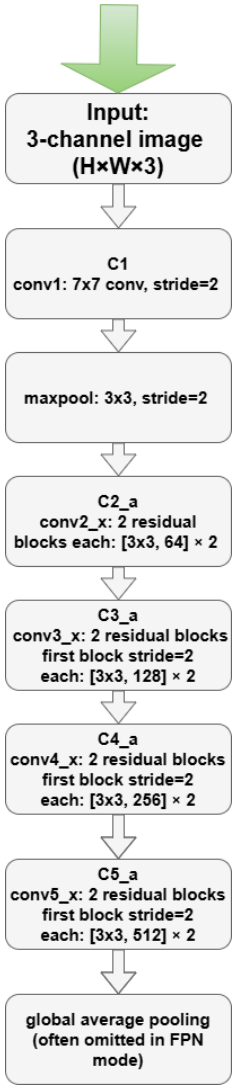

Fig. 3. **General diagram of the basic ResNet18 network**
***Source*: compiled by the authors**

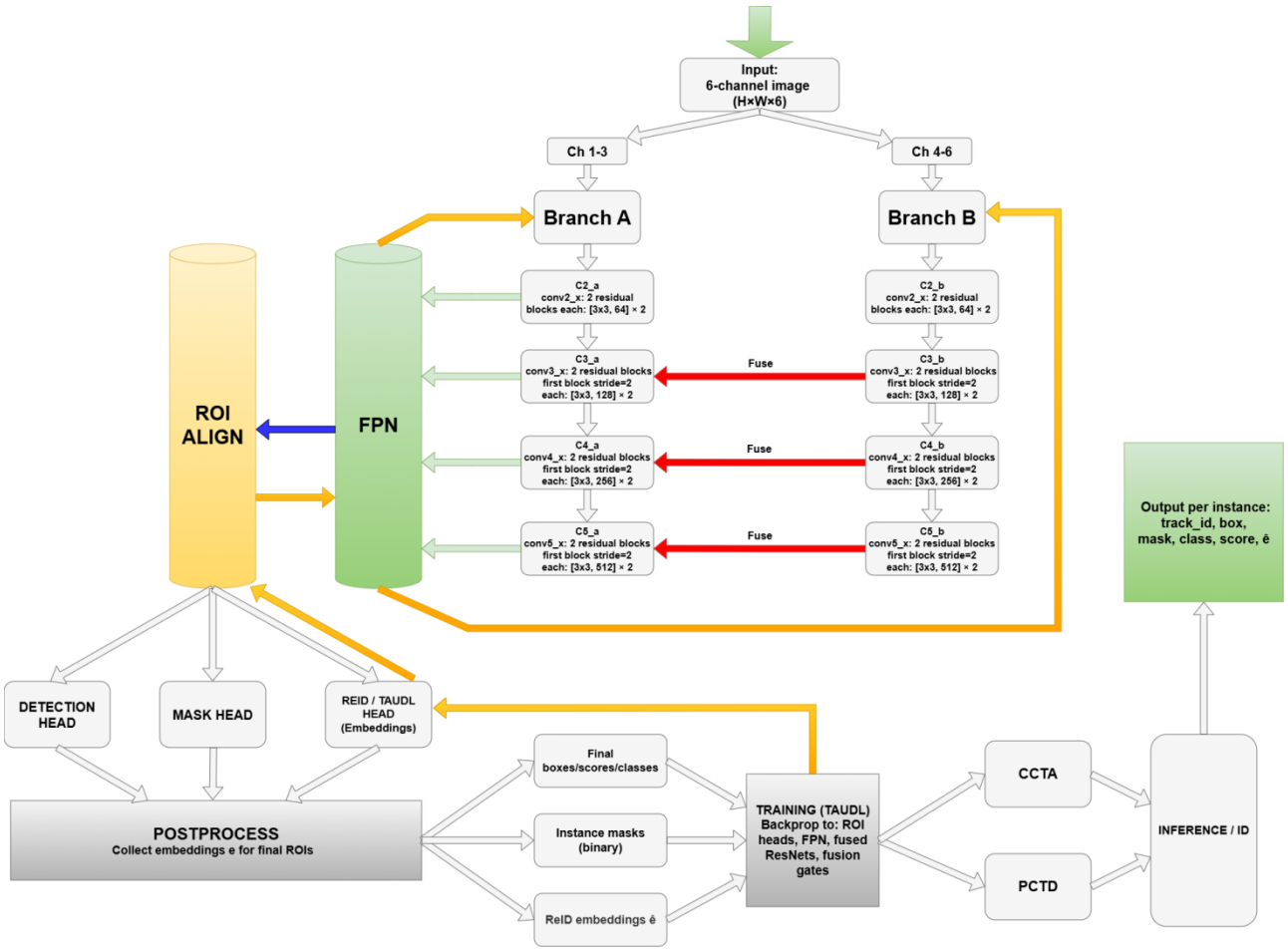

Fig. 4. **General conceptual diagram of the unsupervised re-identification architecture (R-CNN + TAUDL Mask)**

Source: **compiled by the authors**

Unsupervised learning of representations is performed according to the TAUDL framework [17], combining two complementary goals. The first, per-camera track discrimination (PCTD) treats each extracted tracklet in a given session (or "chamber") as a pseudo-class and trains a classifier to distinguish between these tracks, applying regularization to reduce the effects of fragmentation and noisy associations. The second, cross-chamber track association (CCTA), periodically identifies pairs of mutual nearest-neighbor tracks across different sessions or data acquisition conditions and approximates their embedding using contrast losses. In this context, "chambers" are interpreted as different recordings of the same environment with altered lighting, water conditions, or experimental parameters. This reinterpretation allows CCTA to learn invariance to interfering noise, even in nominally single-chamber laboratories, using multiple sessions of the same arena. Both loss functions are optimized jointly, with a short warm-up phase during which PCTD learns to stabilize the embedding space before activating cross-session associations.

The general architecture of unsupervised re-identification is shown in Fig. 4. In addition to the RoIAlign and Detection Head modules, the model includes the ReID module. It operates on the following components:

- fully connected layer creating ReID embedding vector:

$$e_r = W^{\text{reid}} h_r + b^{\text{reid}} \in \mathrm{R}^D,$$

where D is the embedding size;

- *L2*-normalization to obtain the final embedding:

$$f_r = \frac{e_r}{\|e_r\|^2}.$$

Each region of interest (RoI) in a training package has an associated tracklet ID:

$$y_r^{\text{trk}} \in \{0,...,T-1\} \cup \{-1\},$$

where -1 is used to indicate RoIs that TAUDL should ignore (e.g. background or non-matching). These identifiers are aligned with the regions of interest (RoIs) generated by the Detection Target Layer, so TAUDL losses are only calculated for valid foreground RoIs with a given tracklet identifier. Also, the loss for this module (Association Loss/TAUDL-Style Association Loss (heavy triplet variant in batch form)) is calculated as follows. Let R be the set of RoIs with valid track identifiers (non-negative). For each RoI $r \in R$, embedding $f_r \in \mathrm{R}^D$, tracklet y_r^{trk} a pairwise squared Euclidean distance matrix is calculated:

$$d_{rs} = \| f_r - f_s \|_2^2 .$$

For each anchor *r*, one defines:

- positive set $P(r) = \{s \in R : y_s^{\text{trk}} = y_r^{\text{trk}}, s \neq r\}$,
- negative set $N(r) = \{s \in R : y_s^{\text{trk}} \neq y_r^{\text{trk}}\}$.

If *P(r)* and *N(r)* are not empty, the most complex positive distance is $d_r^+ = \max\limits_{s \in P(r)} d_{rs}$ and negative one is $d_r^- = \min\limits_{s \in N(r)} d_{rs}$.

Thus the TAUDL association loss for each anchor is:

$$l_r = \max\left(0, m + d_r^+ - d_r^-\right),$$

where *m* is the margin hyperparameter. The final loss TAUDL is the average value for all valid anchors:

$$L_{\text{taudl}} = \frac{1}{|\Re|} \sum_{r \in \Re} l_r ,$$

where $\Re$ is a set of RoIs with at least one positive and one negative counterpart in the packet. This packet-complex triplet loss is one of the key elements of TAUDL: for each tracklet identity, the most complex positive and negative examples in the current mini-packet are used to guide metric learning, encouraging small distances between embeddings of the same tracklet and large distances between different tracklets, with margin *m*.

Thus, the overall loss metric of the Mask R-CNN+ TAUDL model is:

$$L_{\text{total}} = L_{\text{rpn\_cls}} + L_{\text{rpn\_bbox}} + \\ + L_{\text{mrcnn\_cls}} + L_{\text{mrcnn\_bbox}} + L_{\text{mrcnn\_mask}} + \\ + \lambda_{\text{taudl}} L_{\text{taudl}},$$

where $L_{\text{rpn\_cls}}, L_{\text{rpn\_bbox}}$ are the losses of class and bounding box for RPN [9,30] respectively, and $L_{\text{mrcnn\_cls}}, L_{\text{mrcnn\_bbox}}, L_{\text{mrcnn\_mask}}$ are the losses of the class identification, object bounding box localization and segmentation mask of Mask R-CNN respectively. In the current implementation, λ is effectively 1 unless further scaling is performed within the losses. The important point is that the gradients from the TAUDL losses are propagated back through the ReID head into the same ResNet-18 pooled basis used by Mask R-CNN, thereby coordinating feature training for detection/segmentation and ReID at the tracklet level. Given the connection to the separate TAUDL model, which is to use a dual pooled ResNet-18 backbone (the dual backbone consumes 6-channel input, the global pooling and FC layer create the embedding f, and the TAUDL losses are applied at the image/tracklet level), all of this together allows the backbone to be pretrained with tracklet-oriented TAUDL objectives and then tuned together with Mask R-CNN, or vice versa.

EXPERIMENT AND RESULTS

Due to its modularity, the proposed architecture can be trained both together (if there is a pre-prepared dataset with data for training the TAUDL module) and separately. In this case, an alternate training approach is implemented: first, the detector-classifier of the MaskRCNN itself is trained, then the unsupervised re-identification module TAUDL is trained on its basis. The experiments were performed on a laptop with an Intel Core i9-13980HX processor, 64 GB of RAM and one NVidia GeForce RTX 4090 (16GB) video card. Standard versions of ResNet architectures were used for the experiments. The test dataset contained 41 test images and 8 validation images, including the corresponding tracklets. The Batch size was chosen taking into account the size of the feature extractor. Fig.5 and Fig.6 demonstrate examples of object detection and re-identification.

The mean Average Precision (mAP) and Rank-k (also known as Cumulative Matching Characteristic / top-k accuracy) metrics were used to assess the accuracy. The results (Table) demonstrate that the dual ResNet18 architecture configuration in the re-identification module demonstrates overall higher identification accuracy compared to the baseline ResNet18 and ResNet34 variants and approaches ResNet50. In addition, additional experiments have shown that the Mask R-CNN mask detector and classifier with dual ResNet18

performs well at the ResNet50 level. Due to this, the integration of the TAUDL module eliminates the need for additional feature extractors, which significantly simplifies the architecture, reduces computational costs, and reduces the overall model size. Fig. 5 and Fig. 6 show examples of object detection and re-identification.

Table. **TAUDL accuracy vs feature extractor**

Backbone	TAUDL	Batch size	Epochs	mAP	R1	R5	R10
ResNet18	Random init	10	100	0.31	0.5	0.75	0.875
ResNet34	Random init	10	100	0,3	0.375	0.625	0.875
ResNet50	Random init	8	100	0.37	0.375	0.75	0.875
Fused double ResNet18	Backbone weights ontained from MaskRCNN detector training	8	120	0.33	0.625	0.875	1.0

***Source*: compiled by the authors**

a *b*

Fig. 5. **Object detection via modified Mask R-CNN:**
***a, b* – detection examples**
***Source:* compiled by the authors**

a *b* *c* *d*

Fig. 6. **Examples of re-identification** :
***a, b, c, d* – determining the object ID**
Source: compiled by the authors

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

CONCLUSIONS

In this study, we present an improved unsupervised label-free re-identification architecture that combines instance segmentation with tracklet association learning to achieve stable identity tracking without manual labeling. The proposed architecture is implemented as a practical pipeline that combines a dual-backbone Mask R-CNN for detection and segmentation with an unsupervised re-identification module built on the TAUDL principle, which allows for continuous maintenance of the identities of fish and mice in closed experimental environments over a long period of time.

The tasks were accomplished: an efficient detector-segmenter with twin-ResNet18 and a compression and excitation mechanism was implemented, a re-identification module based on 256-dimensional L2-normalized embeddings was developed, and unsupervised learning under the TAUDL scheme using PCTD and CCTA was implemented. The use of instance masks provides clearer object separation and more informative motion features, while tracklet association learning forms robust unlabeled appearance embeddings. This reduces identity switching and track fragmentation during occlusions and pose variations. The overall architecture is modular and data-efficient, making it suitable for long-term behavioral experiments where manual annotation is too resource-intensive.

Future work will focus on large-scale benchmarking, adapting the architecture to broader conditions, further improving the accuracy of re-identification, and developing a specialized tracking algorithm built on the proposed architecture for animal behavior analysis.

AKWNOLEDGMENTS

Special thanks for the research assistance and provided test videos and images of lab animals to Faculty of Biology of Odesa I.I. Mechnikov National University.

REFERENCES

1. Spampinato, C. et al. “Understanding fish behavior during typhoon events in real-life underwater environments”. *Multimedia Tools and Applications*. 2014; 70 (1): 199–236, https://www.scopus.com/inward/record.uri?eid=2-s2.0-84901839205&partnerID=MN8TOARS. DOI: https://doi.org/10.1007/s11042-012-1101-5.

2. Wojke, N., Bewley, A. & Paulus, D. “Simple online and realtime tracking with a deep association metric”. In *2017 IEEE International Conference on Image Processing (ICIP)*. 2017. p. 3645–3649. DOI: https://doi.org/10.1109/ICIP.2017.8296962.

3. Diakogiannis, F. I., Waldner, F., Caccetta, P. & Wu, C. “ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data”. *ISPRS Journal of Photogrammetry and Remote Sensing*. 2020; 162: 94–114. DOI: https://doi.org/10.1016/j.isprsjprs.2020.01.013.

4.Lauer, J. et al. “Multi-animal pose estimation and tracking with DeepLabCut”. *bioRxiv. Advance Online Publication*. 2021, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127960932&partnerID=MN8TOARS. DOI: https://doi.org/10.1101/2021.04.30.442096.

5. Pereira, T. D. et al. “SLEAP: A deep learning system for multi-animal pose tracking”. *Nature Methods*. 2022; 19 (4): 486–495, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128789099&partnerID=MN8TOARS. DOI: https://doi.org/10.1038/s41592-022-01426-1.

6. Chen, Z., Zhang, R., Zhang, Y. E., Wang, W., Fan, Z., Zhang, Y. & Lu, C. “AlphaTracker: A multi-animal tracking and behavioral analysis tool”. *Frontiers in Behavioral Neuroscience*. 2023: 17: 1111908. DOI: https://doi.org/10.3389/fnbeh.2023.1111908.

7. Csurka, G., Volpi, R. & Chidlovskii, B. “Semantic image segmentation: Two decades of research”. *Foundations and Trends® in Computer Graphics and Vision*. 2022; 14 (1-2): 1–162. DOI: https://doi.org/10.1561/0600000095.

8. Ramesh, C. S. & Kumar, V. V. “A review on instance segmentation using Mask R-CNN”. In *Proceedings of the International Conference on Systems, Energy & Environment (ICSEE)*. 2021. p. 1–4. DOI: https://doi.org/10.2139/ssrn.3794272.

9. Babaee, M., Athar, A. & Rigoll, G. “Multiple people tracking using hierarchical deep tracklet re-identification”. *arXiv*. 2018. DOI: https://doi.org/10.48550/arXiv.1811.04091.

10. Yang, F., Wang, Z., Wu, Y., Sakti, S. & Nakamura, S. “Tackling multiple object tracking with complicated motions: Re-designing the integration of motion and appearance”. *Image and Vision Computing*. 2022; 124: 104514. DOI: https://doi.org/doi.org/10.1016/j.imavis.2022.104514.

11. Yang, F., Wang, Z., Wu, Y., Sakti, S. & Nakamura, S. “A unified multi-view multi-person tracking framework”. *Computational Visual Media*, 2024; 10 (1): 137–160. DOI: https://doi.org/10.1007/s41095-023-0334-8.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

12. Sun, J., Wang, T., Chen, K. & Liu, Y. GTA: “Global tracklet association for multi-object tracking in sports”. In *Computer Vision – ACCV 2024 Workshops. Springer*. 2025. DOI: https://doi.org/10.1007/978-981-96-2644-1_6.

13. Jian, R.-L., Luo, M.-C., Huang, C.-W., Lee, C.-M. & Hsu, C.-C. “GTATrack: Winner solution to SoccerTrack 2025 with Deep-EIoU and global tracklet association”. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM’25). ACM*. 2025. DOI: https://doi.org/10.1145/3728423.3759416.

14. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G. & Wang, L. “SportsMOT: A large multi-object tracking dataset in multiple sports scenes”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. p. 9887–9897. DOI: https://doi.org/10.1109/ICCV51070.2023.00910.

15. Nguyen, T. T., Nguyen, H. H., Sartipi, M., & Fisichella, M. “Multi-vehicle multi-camera tracking with graph-based tracklet features”. *IEEE Transactions on Multimedia*. 2024; 26: 972–983. DOI: https://doi.org/10.1109/TMM.2023.3274369.

16. Nguyen, T. T., Nguyen, H. H., Sartipi, M. & Fisichella, M. “Real-time multi-vehicle multi-camera tracking with graph-based tracklet features”. *Transportation Research Record*. 2024; 2678 (1): 296–308. DOI: https://doi.org/10.1177/03611981231170591.

17. Dorai, Y., Kiruthiga, R. & Jayasudha, J. “Tracklet style transfer and part-level feature description for person re-identification in a camera network”. *Pattern Analysis and Applications*. 2021; 24: 875–886. DOI: https://doi.org/10.1007/s10044-021-00990-0.

18. Liu, Y., Li, M., Wang, F., Zhang, R. & Wan, J. “Feature-level camera style transfer for person re-identification”. *Applied Sciences*. 2022; 12 (14): 7286. DOI: https://doi.org/10.3390/app12147286.

19. Ahn, N., Kim, I., Jeong, S. & Ko, S.-J. “Multiple object tracking using re-identification model with attention module”. *Applied Sciences*. 2023; 13 (7): 4298. DOI: https://doi.org/10.3390/app13074298.

20. Si, H., Zhu, H., Fang, G. & Li, J. “Multi-object tracking with integrated heads and attention”. *Neurocomputing*. 2022; 510: 95–106. DOI: https://doi.org/10.1016/j.neucom.2022.09.045.

21. Ma, X., Huang, S., Li, Y. & Chen, L. “Multi-feature re-identification enhanced dual motion modeling for multi small-object tracking”. *Sensors*. 2025; 25 (18): 5732. DOI: https://doi.org/10.3390/s25185732.

22. Liu, H., Jie, Z., Karlekar, J., Zhao, Y., Chai, D., Lim, J. H. & Pranata, S. “Video-based person re-identification with accumulative motion context”. *IEEE Transactions on Circuits and Systems for Video Technology*. 2018; 28 (10): 2788–2802. DOI: https://doi.org/10.1109/TCSVT.2017.2715499.

23. Yang, J., Wang, S. & Jiang, T. “Progressive unsupervised video person re-identification with accumulative motion and tracklet spatial–temporal correlation”. *Future Generation Computer Systems*. 2023; 142: 90–100. DOI: https://doi.org/10.1016/j.future.2022.12.023.

24. Li, M., Zhu, X. & Gong, S. “Unsupervised person re-identification by deep learning tracklet association”. In *Computer Vision – ECCV 2018 (Lecture Notes in Computer Science*. 2018; 11208: 737–753, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055454377&partnerID=MN8TOARS.
DOI: https://doi.org/10.1007/978-3-030-01225-0_45.

25. Wu, S., Sun, Y., Luo, J., Zhang, C. & Huang, Q. “Tracklet self-supervised learning for unsupervised person re-identification”. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020; 34 (7): 12362–12369. DOI: https://doi.org/10.1609/aaai.v34i07.6921.

26. Zang, X., Li, G., Gao, W. & Shu, X. “Exploiting robust unsupervised video person re-identification”. *IET Image Processing*. 2022; 16 (3): 729–741. DOI: https://doi.org/10.1049/ipr2.12380.

27. Liu, Q., Chen, D., Chu, Q., Yuan, L., Liu, B., Zhang, L. & Yu, N. “Online multi-object tracking with unsupervised re-identification learning and occlusion estimation”. *Neurocomputing*. 2022; 483: 333–347, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124569858&partnerID=MN8TOARS.
DOI: https://doi.org/10.1016/j.neucom.2022.01.008.

28. Li, S., Yang, L., Tan, H., Lan, L. & Lin, Y. “Self-supervised re-identification for online joint multi-object tracking”. *Knowledge and Information Systems. Advance Online Publication*. 2025. DOI: https://doi.org/10.1007/s10115-024-02237-w.

29. Erregue, I., Benyahia, Y., Mezhoud, R. & Boukhrissa, A. “YOLO11-JDE: Fast and accurate multi-object tracking with self-supervised re-ID”. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW 2025)*. 2025. p. 776–785. DOI: https://doi.org/10.1109/WACVW65960.2025.00092.

30. Volkova, N. & Shvandt, M. “Object detection network architecture for multichannel images”. In *Information Control Systems and Technologies (ICST-ODESA): Proceedings of the XIII International Scientific-Practical Conference*. 2025. p. 166–169. DOI: https://doi.org/10.36059/978-966-397-531-3.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 02.10.2025
Received after revision 28.11.2025
Accepted 05.12.2025

DOI: https://doi.org/10.15276/hait.08.2025.30
УДК 378.147:004.6:331.5

Удосконалений підхід до ідентифікації без міток на основі сегментованих треків для аналізу поведінки об'єктів

Волкова Наталія Павлівна[1)]
ORCID: https://orcid.org/ 0000-0003-3175-2179; volkova.n.p@op.edu.ua. Scopus Author ID: 36104775700
Швандт Максим Альбертович[1)]
ORCID: https://orcid.org/ 0000-0002-4580-3961; maxim.shvandt@gmail.com
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У роботі ми представляємо Mask-TAUDL, вдосконалену архітектектуру безнаглядної повторної ідентифікації, яка поєднує сегментацію екземплярів, глибоке навчання без учителя та асоціацію треклетів для детального аналізу поведінки об'єктів у довготривалих записах. Вона об’єднує двопотоковий детектор/сегментатор Mask R-CNN з подвійними магістралями ResNet-18 та модулем глибокого навчання без учителя на основі асоціації треклетів (TAUDL). Mask R-CNN забезпечує точну локалізацію об'єктів та бінарні маски, з яких ми будуємо треклети з покращеною сегментацією. Два потоки ResNet-18 використовують ці маски для вилучення ознак зовнішнього вигляду та ознак, чутливих до руху, на рівні треклетів, які об'єднуються у спільний дескриптор ознак. Модуль TAUDL працює безпосередньо з маскованими ознаками треклетів та спільно навчає дискримінативні вбудовування і міжсесійні асоціації без ручного маркування. Запропонована архітектура Mask-TAUDL навчає модель так, щоб ознаки однієї особини залишалися близькими у просторі вбудовувань упродовж часу, водночас забезпечуючи чітке розділення ознак між різними особинами. Інтеграція чистих маскованих областей із часово агрегованими ознаками допомагає пригнічувати хибні варіації, викликані тінями, відбиттями або перекриттям об’єктів. Довгострокова повторна ідентифікація тварин є складною задачею через часті перекриття, дрейф зовнішнього вигляду та незначні візуальні відмінності між особинами, а більшість існуючих рішень спираються на великі анотовані набори даних, що обмежує їхню застосовність у реальних лабораторних умовах. Архітектура Mask-TAUDL усуває ці обмеження шляхом явного моделювання часово узгоджених, уточнених за маскою треків та навчання вбудовувань, які зберігають ідентичність у повністю безнаглядному режимі. Mask-TAUDL розроблений для досліджень поведінки тварин, а саме невеликих лабораторних видах, таких як миші та риби, які спостерігаються в закритих або напівструктурованих аренах, де надійне довготривале відстеження ідентичності є важливим для кількісного аналізу поведінки, поздовжніх експериментів та високопродуктивного скринінгу.

Ключові слова: ідентифікація; сегментація; обробка зображень;;безнаглядна повторна ідентифікація; глибоке навчання; TAUDL; Mask R-CNN; треклети; архітектура; відстеження поведінки тварин

ABOUT THE AUTHORS

Natalia P. Volkova - PhD, Associate Professor, Chief of the Department of Applied Mathematics and Information Technologies, Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine.
ORCID: https://orcid.org/0000-0003-3175-2179; volkova.n.p@op.edu.ua. Scopus Author ID: 36104775700
***Research field*:** Digital Image Processing; Pattern Recognition

Волкова Наталія Павлівна – кандидат технічних наук, доцент, завідувачка кафедри Кафедра прикладної математики та інформаційних технологій. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

Maksym A. Shvandt - PhD Student, Department of Applied Mathematics and Information Technologies. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine.
ORCID: https://orcid.org/0000-0002-4580-3961; maxim.shvandt@gmail.com
***Research field*:** Computer vision, image processing

Швандт Максим Альбертович - аспірант кафедри Прикладної математики та Інформаційних Технологій, Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна