Man and the Machine: Effects of AI-assisted Human Labeling on Interactive Annotation of Real-Time Video Streams

MARKO RADETA*, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, University of Belgrade, Portugal
RUBEN FREITAS, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, Portugal
CLAUDIO RODRIGUES, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, Portugal
AGUSTIN ZUNIGA, Department of Computer Science, University of Helsinki, Finland
NGOC THI NGUYEN, Department of Computer Science, University of Helsinki, Finland
HUBER FLORES, Institute of Computer Science, University of Tartu, Estonia
PETTERI NURMI, Department of Computer Science, University of Helsinki, Finland

AI-assisted interactive annotation is a powerful way to facilitate data annotation – a prerequisite for constructing robust AI models. While AI-assisted interactive annotation has been extensively studied in static settings, less is known about its usage in dynamic scenarios where the annotators operate under time and cognitive constraints, e.g., while detecting suspicious or dangerous activities from real-time surveillance feeds. Understanding how AI can assist annotators in these tasks and facilitate consistent annotation is paramount to ensure high performance for AI models trained on these data. We address this gap in interactive machine learning (IML) research, contributing an extensive investigation of the benefits, limitations, and challenges of AI-assisted annotation in dynamic application use cases. We address both the effects of AI on annotators and the effects of (AI) annotations on the performance of AI models trained on annotated data in real-time video annotations. We conduct extensive experiments that compare annotation performance at two annotator levels (expert and non-expert) and two interactive labelling techniques (with and without AI-assistance). In a controlled study with $N = 34$ annotators and a follow up study with 51 963 images and their annotation labels being input to the AI model, we demonstrate that the benefits of AI-assisted models are greatest for non-expert users and for cases where targets are only partially or briefly visible. The expert users tend to outperform or achieve similar performance as AI model. Labels combining AI and expert annotations result in the best overall performance as the AI reduces overflow and latency in the expert annotations. We derive guidelines for the use of AI-assisted human annotation in real-time dynamic use cases.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Computer vision**; • **Human-centered computing** → *Interactive systems and tools*; • **Applied computing** → *Annotation.*

Additional Key Words and Phrases: Computer Vision, Object Detection, Machine Learning, Deep Learning, Annotation, Videos, Man-Machine, Human-in-the-Loop, Intelligent User Interface, AI-assisted Interface

*Corresponding author: marko@wave-labs.org

Authors' addresses: Marko Radeta, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, University of Belgrade, Portugal; Ruben Freitas, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, Portugal; Claudio Rodrigues, Wave Labs, MARE/ARNET/ARDITI, University of Madeira, Portugal; Agustin Zuniga, Department of Computer Science, University of Helsinki, Finland; Ngoc Thi Nguyen, Department of Computer Science, University of Helsinki, Finland; Huber Flores, Institute of Computer Science, University of Tartu, Estonia; Petteri Nurmi, Department of Computer Science, University of Helsinki, Finland.

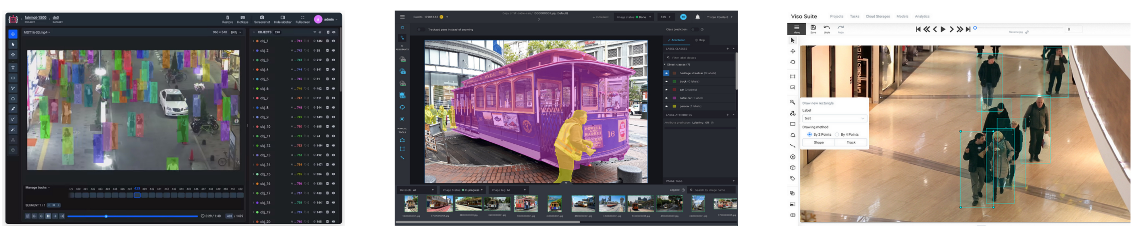
Manuscript submitted to ACM

ACM Reference Format:
Marko Radeta, Ruben Freitas, Claudio Rodrigues, Agustin Zuniga, Ngoc Thi Nguyen, Huber Flores, and Petteri Nurmi. 2018. Man and the Machine: Effects of AI-assisted Human Labeling on Interactive Annotation of Real-Time Video Streams. *ACM Trans. Interact. Intell. Syst.* 37, 4, Article 111 (August 2018), 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

High quality labeled data is a prerequisite for constructing powerful AI models. While the process of assigning labels is seemingly simple, in reality it is wrought with difficulties as the process requires significant time and resource investment and is prone to noise and errors [7, 31, 49]. *AI-assisted interactive labeling* (Figure 1) seeks to reduce the resource and cognitive demands of labeling and to improve the quality of labels by supporting the human annotation effort through interactive AI [4, 11, 54, 56]. Examples of AI techniques include visualizations that highlight patterns in the data [3, 55] and suggestions of the most likely labels [12, 47].

Evaluations of AI-assisted interactive labeling techniques have shown that, at best, the AI support can significantly decrease the time of labeling while also improving the quality of data [12, 62]. Furthermore, if the AI assisted annotations cover data that are infrequent or otherwise difficult [19], this often provides the best improvements for the final AI models that are trained on the data. While these benefits are promising, there are two main limitations to existing research. First, they have focused exclusively on static tasks where the human annotators can invest time to scrutinize and revise their annotations without examining how AI-assists in dynamic real-time tasks where the annotation is carried out in parallel to a real-world task. Second, thus far limited understand exists about the effects of using AI-assisted annotations to train AI models.



(a) (b) (c)

Fig. 1. Examples of existing AI-assisted video annotation interfaces: (a) Supervisely Video Labeling Tool; (b) CloudFactory's Accelerated Annotation; (c) Computer Vision Annotation Tool (CVAT).

The present paper contributes by systematically assessing the benefits and limitations of AI-assisted interactive labeling in real-time labeling tasks. We design a simple AI-assisted labeling interface and conduct extensive experiments that compare annotation performance between expert and non-expert users with and without interactive AI-assistance. To understand the impact of the labeling performance, we also separately investigate how annotations by these different groups impact the ML models that are trained from the data. We conduct our study considering a benign application scenario, marine biodiversity estimation, which serves as a representative example of tasks that require real-time annotation capability. Besides offering a challenging real-world task, the data that needs to be labeled also contains significant variations as it covers diverse water conditions, background details, fields of view, and so forth. This allows us to obtain better insights both into the performance of human annotators and the AI models that are trained with such data. We compare expert and non-expert annotations, considering both AI-assisted interactive labeling and interactive annotations without any AI support. In total $N = 34$ annotators participate in our evaluations.

The results of our evaluations show that the benefits of AI-assisted models are greatest for non-expert users and for cases where targets are only partially or briefly visible. Indeed, expert users tend to outperform – or at least achieve similar performance – as AI models whereas the use of AI can bring non-expert users close to expert levels. The main challenge for real-time feeds is to accurately determine the start and end points of the periods where objects are visible and AI assistance can result in annotations overflowing the actual time that an object is visible. We also conduct a follow up study where we analyse how different annotations affect the performance of AI models that are trained from the annotated data. Labels combining AI and expert annotations result in best overall performance as the AI reduces overflow and latency in the expert annotations, providing the most consistent annotations and thus making it easier for the AI to learn a robust and general model. Besides presenting the results of our studies, we derive guidelines for the use of AI-assisted human annotation in real-time use and discuss what the limitations mean for AI models that use the data.

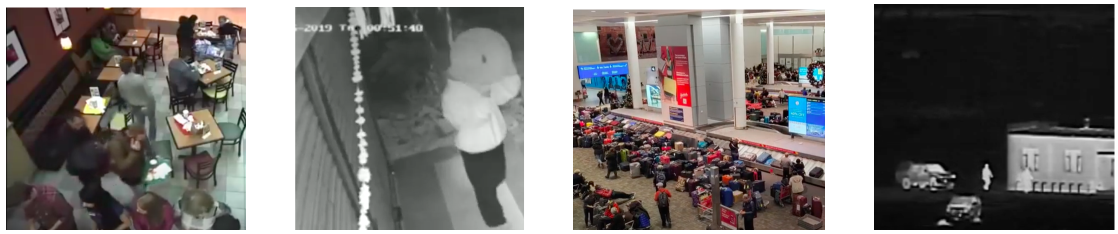
Summary of Contributions. This paper enhances the current state of the art in interactive intelligent systems by:

- Extensive assessment of the benefits, limitations and challenges of AI-assisted annotation in dynamic application use cases considering both domain experts and non-experts.
- Novel insights into the performance of AI-assisted annotation in dynamic application use cases. For example, we demonstrate that non-experts achieve highest benefits, scenes with occlusions are most impacted by AI-annotations, and that the main challenge is to identify event start and end points accurately.
- Follow-up assessment of the impacts of AI-annotated labels on AI-models trained on the data. The results show that expert annotations combined with AI-annotations achieve best AI model performance, even if the benefits of AI are small for experts.
- Based on our results, we derive guidelines and best practices for interactive AI annotation in dynamic use cases.

2 TOWARDS REAL-TIME ANNOTATION

The focus of our work is on real-time annotation of continuous video feeds which differs from the predominantly static annotation scenarios considered in existing research [25]. Indeed, even if most research explores continuous video, they do not consider constraints posed by real-time nature of the annotation. This means that existing works largely focus on scenarios where annotators can pause or adjust the speed of the video and perform operations on the video feed (e.g., pan, zoom or tag) [24, 46, 50, 69]. In these kind of tasks, the annotators can split their time and cognitive capacity between the video feed and the interactive AI, instead of the different views competing for the same cognitive resources of the annotators. Static annotation also differs from real-time annotation in that users can freely pick the label to assign, whereas real-time annotation requires users to rapidly select the right label, making it necessary to consider only a small number of labels.

Figure 2 illustrates challenges in performing the annotations in real-time using detection of suspicious activity as an example. The events of interest are rare, the time window to detect them is short, and the detection process is prone to errors due to distractions and dullness [53]. As illustrated in Figure 2, the annotator may also struggle to identify the suspicious activity due to lack of context, difficulty in understanding the events that are happening in the scene, lack of focus in the picture, or lack of precision due to obstacles, environmental variations or other factors. Finding these events after the event has occurred means the negative consequences of the actions have already occurred and even this task is difficult as examining large amounts of video is resource intensive. The video material is also continually increasing,



(a) (b) (c) (d)

Fig. 2. Challenges for AI in identifying objects of interest in video streams: (a) lack of context - a thief in action or a friendly person? The AI needs to have a memory for understanding the exact sequence of happenings; (b) lack of identifying unpredictable behaviours - an object moving towards the door may not be understood as a human performing the hazardous behaviour; (c) lack of focus - focusing on people or luggage; (d) lack of precision - discriminating persons in front of the lights using thermal imagery.

requiring lots of resources to stay up-to-date with the events that would be relevant for labeling. The detection of suspicious activity is but one example of the application domains that require real-time annotation. Other examples include remote operated search and rescue operations [60], remote surgery [5], and maritime monitoring carried out on board ships or using remotely accessed video streams [36]. Understanding how AI can assist human labeling in these kinds of everyday tasks is essential for understanding the quality of data these tasks provide – as well as the potential limitations they pose on the AI models developed from such data.

3 ANNOTATION PIPELINE

We study real-time annotation using labeling of marine species from continuous video streams as a representative example of tasks that require real-time capability. Currently, this task is carried out by dedicated watchers who record sightings made aboard vessels and these records are used to establish counts of marine species, an indicator of marine biodiversity. The work in this paper is part of a longer-term project that seeks to support this process, including using AI and interactive technologies. We investigate annotation performance by conducting experiments that use video clips of marine species collected from marine excursions and openly available images. The video footage was labeled by two of the authors who also labeled the additional images. The two authors performing the annotations have long-term expertise in working with marine species and thus are qualified to analyze the footage. From the labeled footage, we extract individual frames and combined them with the image data. The AI model used for annotation was trained using the image data, and the annotation experiments were conducted on the video footage. The overall pipeline is illustrated in Figure 3. Below we describe the used datasets, creation of the base AI model to support the AI-assisted interface and additionally performed video analysis.

3.1 Dataset Collection

Obtained Imagery. To train the model, we collected imagery using online search. We downloaded 10 010 images of the aforementioned 5 marine species (Table 1) seen from diverse field of views (e.g. top-down, profile, semi-profile, etc), with mixed scenes (underwater, surface, aerial), etc. All obtained imagery was taken from 3 different online sources including: Open Images Dataset (OID) [32, 59], Kaggle and bulk image collection from Google Image Search.

Obtained Footage. To test the model, we collected video footage (Figure 5). As the primary source of data we considered proprietary real-time video footage obtained from four dolphin-watching trips. To ensure generality and increase

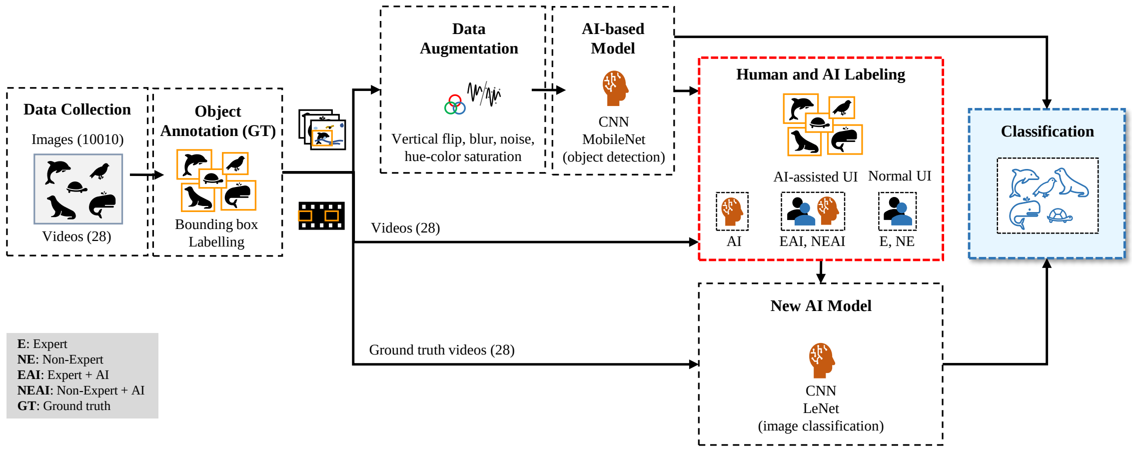


Fig. 3. Man and the Machine Pipeline - Human and AI collaborative object annotation and image classification pipeline.

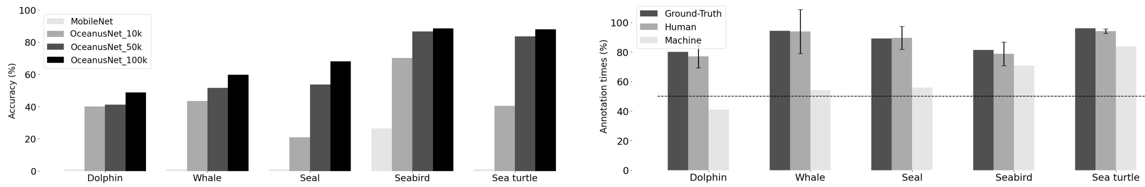
variations in the footage, we further augmented this data with video footage available on YouTube. The video footage from the dolphin watching trips (16 videos, 30.85 minutes, 720p resolution) contain different dolphin species and were recorded with a mobile camera being held horizontally from the sea-vessel front deck during the trip, facing the sea surface towards spotted dolphins. The YouTube videos (12 videos, 26.9 minutes, 1080p resolution) contain other marine species (commonly seen on such trips): whales, seabirds, turtles and seals, and were uploaded by users after they have taken trips abroad vessels. All samples were recorded with 30*FPS*. Detailed video timings are shown in Table 2. Figure 5 shows the screenshots of the footage and corresponding keywords depicting the scene type, with Figures 5 (a)-(l) correspond to YouTube footage and the remaining 16 screenshots (Figures 5 (m)-(ab)) are examples of data collected during the dolphin watching trips (indicated with § symbol).

Table 1. Total obtained images from online and personal footage including number of annotations as bounding boxes and labels prepared for model training using MobileNet. Grayed areas indicate difficult videos.

#	Objects	No. of Images, (%)		No. of Images, (%) (after augmentation)		No. of Objects, (%) (after cleaning)		No. of Objects, (%) (after augmentation)	
1	Sea Turtles	1787	.18	3887	.20	2552	.13	4983	.13
2	Seabirds	1660	.17	3217	.16	4185	.20	8101	.20
3	Dolphins	2217	.22	4310	.22	4487	.22	8718	.22
4	Whales	1966	.20	3805	.19	2132	.11	4123	.11
5	Seals	2380	.23	4634	.23	6501	.33	12678	.33
	Total	**10010**		**19853**		**19857**		**38603**	

3.2 Base AI Model Procedures

Object Annotation. Each image was further subject to object annotation by manually placing rectangle bounding boxes around the spotted objects of interests (in this case, species), performed by the two authors of the study. From all images, 19 857 annotated objects were obtained, presented in the Table 1. Together with rectangle objects, 5 main classes were used as labels for each species: (i) *whales* - baleen whales (e.g. Blue Whale); (ii) *dolphins* - toothed cetaceans (e.g. Bottlenose dolphin); (iii) *sea turtles* - e.g. loggerhead turtle; (iv) *seabirds* - e.g. seagull; and (v) *seals* - e.g. monk seal. The total amount of labels was identical to the total amount of objects (19 857) having one object per image. All



(a) Training epoch benchmark

(b) Human vs Machine vs GT

Fig. 4. From left to right: (a) Evolution of accuracy with more iteration steps indicates that all objects of interest reached an adequate accuracy after 100k epochs; (b) Comparison between annotation times of ground truth annotators (GT), human annotators and AI, relative to total cumulative video length indicating the human annotators to be overall better in recognizing the objects of interest across all videos. The vertical line indicates threshold 50% accuracy.

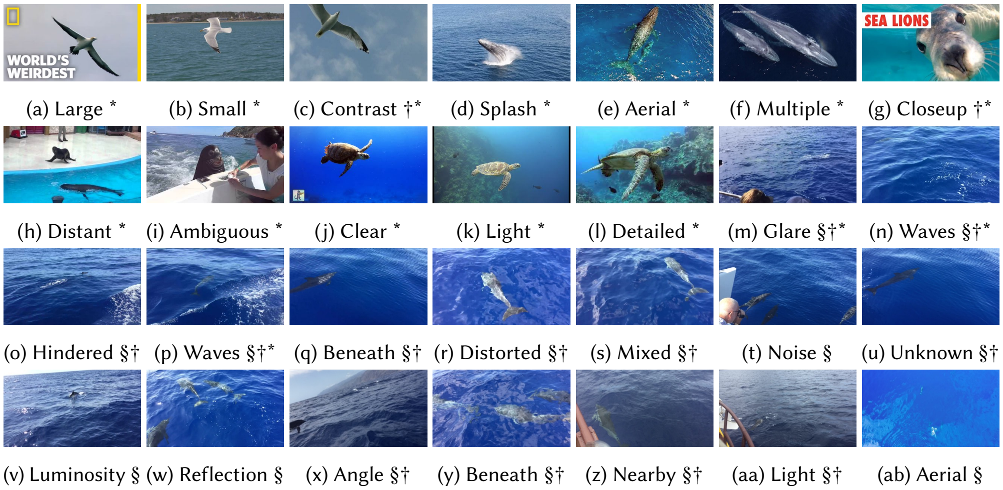
obtained imagery was downsampled to match the minimum image resolution of the image sample (800×600). Note that these measurements are solely used for training the AI model and they are not part of the annotation experiments.

Dataset Augmentation. To enhance the accuracy of image classification [38] and increase model robustness, i.e., mitigating ambiguous pixel representations such as water reflection that lead to false positives by the AI [58], we applied data augmentation techniques on the training images: (i) *Vertical flip*, (ii) *Hue-Color Saturation*; (iii) *Blur*; and (iv) *Noise*. With this step, training image sample increased from 10 010 to 19 853 images, while the annotation objects and labels increased from 19 857 to 38 603 respectively (see Table 1 for details).

Model Training. We trained a Convolutional Neural Network (CNN) model to recognize marine species. As state-of-the-art trained models do not support fully detecting marine species (e.g., YOLO5 dataset may recognize fish [23] or specific types of birds but not separate between marine species), we created a custom model (hereinafter, OceanusNet) that is based on Single Shot Detection (SSD) and MobileNetV2 architecture. Similar approach was used when discriminating marine litter underwater [44]. Model was trained on all sample imagery (19 857) using all object annotations (38 603). Default MobileNet hyperparameters were used: accuracy threshold of .5, batch size of 24, learning rate of 0.004, ReLU6 activation function, and without dropout layers. Number of different epochs was used to select the top model, comparing 1k, 10k, 50k and 100k iterations. To boost model performance, the model was further quantized [48, 70] by converting 32-bit floating points to 8-bit integers as this does not result in significant loss of accuracy while it reduces the bandwidth and the memory storage.

Model Inference. To validate the performance, 15 obtained video clips depicted with asterisks in Figure 5 were used. We selected these videos as their average time was around 3 minutes per video clip having balanced three video clips per each class (dolphins, whales, seals, seabirds, and sea turtles). For evaluating the AI performance, we computed cumulative times that the objects of interests were spotted by the AI model inside of the video clip. We propose to use this metric to simulate the real-time video feed, as traditional object detection metrics such as intersection of union [68] may be too laborious and may require more annotation time.

All 5 classes of marine species have been successfully identified by the OceanusNet in both types of footage (collected from the internet and collected from the field trip). As expected, the more iterations the model had, the higher was the average accuracy (Figure 4a). The highest average accuracy for identifying all 5 marine species was with the 100k iterations model, reaching 70.58%. The accuracy for individual classes were: sea turtles (87.87%), whales (59.69%), dolphins (48.75%), seabirds (88.54%), and seals (68.04%). These performance numbers are in line with state-of-the-art object recognition results for sea mammals, even if the models usually have focused on single class classification:



(a) Large * (b) Small * (c) Contrast †* (d) Splash * (e) Aerial * (f) Multiple * (g) Closeup †*
(h) Distant * (i) Ambiguous * (j) Clear * (k) Light * (l) Detailed * (m) Glare §†* (n) Waves §†*
(o) Hindered §† (p) Waves §†* (q) Beneath §† (r) Distorted §† (s) Mixed §† (t) Noise § (u) Unknown §†
(v) Luminosity § (w) Reflection § (x) Angle §† (y) Beneath §† (z) Nearby §† (aa) Light §† (ab) Aerial §

Fig. 5. Obtained sample footage screenshots with observed video and object characteristics. Symbol nomenclature: (§) proprietary obtained footage from dolphin-watching trips; (†) difficult video; (*) video clips used in first study.

whales [1, 8], dolphins [37], sea turtles [6, 9]. Training footage for the sea turtles contained very clear shots of the turtles, reaching highest confidence (96.52%) as turtles in images were occupying larger portion of the screen and were not against complex backgrounds. Seals had more ambiguity which resulted from ambient conditions, such as water reflections, affecting the accuracy. Contrary, as expected traditional MobileNet model was unable to detect any animals except seabirds (YOLO5 constraint).

Obtained results highlight that the use of existing trained models may not be suitable for recognizing specific objects of interests, however suggesting state of the art architectures as adequate. This further motivates the need of the proposed OceanusNet model and confirms the robustness of transfer learning [41]. Still, using these metrics, we notice the difficulties in detecting whales and dolphins as such stem from the inclusion of aerial footage, which makes hard to distinguish the two from frames where the animals are only partially visible. Nevertheless, as stated, the performance of our model is in line with state-of-the-art results for the individual animal classes, even if there is room for further improvements by considering more varied training data, adding processing techniques that eliminate reflections and other effects, and considering further data augmentations.

3.3 Video Procedures

Video Annotation. We next used annotations by two of the authors (non-expert users who have many years of knowledge of working in the field and have sufficient knowledge of correctly identifying the marine species focused in this work) as the gold standard labels. The researchers annotated all 28 video clips (Figure 5) by looking at the video, and counting the duration that the species were inside of the frame. Each video has been individually labelled[1] with the species name by inspecting each video with pause, rewind, and forward functions. All disagreements between the two annotators were carried out verbally until a consensus was achieved. This was performed to ensure high consistency in the annotations and ensure the annotations of other groups can be analyzed in detail. In cases where the baseline

[1]Throughout the paper, we use term "annotation" to indicate the labeling, i.e. if single frame has an object of interest in it, the whole image frame belongs to one class.

annotations had disagreements, a third researcher was invited to inspect the same frame until a consensus was found. These generally corresponded to situations where the object were partially visible or subject to long interpretation, e.g., a whale that jumps can leave a splash on sea surface and there is ambiguity on whether the animal can be seen in subsequent frames or not.

Video Properties. Asides annotations, all videos were analyzed to by categorizing them into different properties (Table 2). Properties are presented as follows: (type) - the type of the scene (aerial, surface or underwater); (FOV) - the camera field of view or perspective, being from boat, from diver or drone; (quality) - video recording quality, being good or bad, based on observable pixels in video; (visibility) - size of the object of interest compared to the video frame size; (recording) - being edited or raw, from amateur or professional sources; (pace) - the rate of change between scenes as slow, fast or extreme fast (fast+); and (hard) - indicating the video clip overall difficulty which is further calculated using annotator performance. These properties are used in the analysis and will be elaborated later in the paper.

Table 2. Video clips parameters used in the study. Asterisk (*) indicates video clips used in first study.

#	Object of Interest	Duration (m)	Frames	Difficult	Type	FOV	Quality	Pace	Recording	Visibility
1	Birds *	1.47	2625		aerial	from boat	good	slow	amateur	>50%
2	Birds *	0.78	1396		aerial	from boat	good	slow	amateur	>50%
3	Birds *	2.45	4391	✓	mixed	mixed	good	**fast+**	prof.	
4	Whales *	3.28	5920		surface	from UAV	good	slow	prof.	>50%
5	Whales *	1.50	2289		mixed	mixed (UAV, underwater)	good	slow	prof.	>50%
6	Whales *	4.82	7219		surface	from UAV	good	slow	prof.	>50%
7	Seals *	1.02	5736	✓	mixed	mixed	good	**mixed**	mixed	
8	Seals *	1.13	5108		surface	audience	good	slow	amateur	>50%
9	Seals *	1.08	6068		surface	from boat	good	slow	amateur	>50%
10	Turtles *	3.20	1830		underwater	diver recording	good	slow+	prof.	>50%
11	Turtles *	2.85	2008		underwater	diver recording	good	slow+	prof.	>50%
12	Turtles *	3.38	1627		underwater	diver recording	good	slow+	prof.	>50%
13	Dolphins *	0.57	991	✓	surface	from boat	good	**fast**	amateur	
14	Dolphins *	0.38	663	✓	surface	from boat	good	**fast**	amateur	
15	Dolphins	0.23	406	✓	surface	from boat	good	**fast**	amateur	
16	Dolphins *	0.40	702	✓	surface	from boat	good	**fast**	amateur	
17	Dolphins	0.67	1184	✓	surface	from boat	good	**fast**	amateur	
18	Dolphins	0.90	1589	✓	surface	from boat	good	**fast**	amateur	
19	Dolphins	0.15	268	✓	surface	from boat	good	**fast**	amateur	
20	Dolphins	5.55	9982		surface	from boat	good	slow	amateur	
21	Dolphins	0.67	1185	✓	surface	from boat	good	**fast**	amateur	
22	Dolphins	3.95	7090		surface	from boat	bad	slow	amateur	
23	Dolphins	6.97	12533		surface	from boat	bad	slow	amateur	
24	Dolphins	4.32	6193	✓	surface	from boat	good	**fast**	amateur	
25	Dolphins	4.00	5741	✓	surface	from boat	good	**fast**	amateur	
26	Dolphins	1.08	1538	✓	surface	from boat	good	**fast**	amateur	
27	Dolphins	1.02	1440	✓	surface	from boat	good	**fast**	amateur	
28	Dolphins	4.32	6191		surface	from UAV	good	slow	prof.	

4 EFFECTS OF AI ON HUMAN ANNOTATIONS

We first compare the effects of AI on human annotations by conducting a study where non-experts and experts annotate the videos with or without AI-assistance. We refer to these four groups as experts (E), non-experts (NE), experts+AI (EAI), and non-experts+AI (NEAI). We next detail our experiment, analysis, and the key findings.

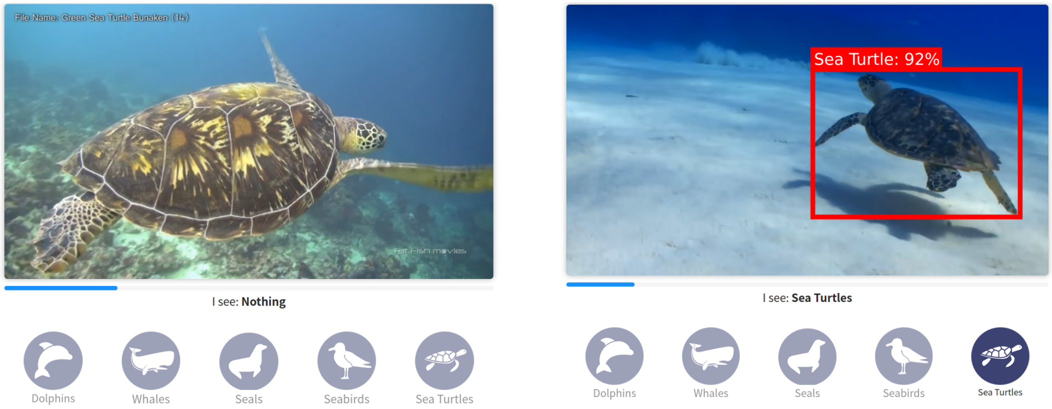
4.1 Experimental Setup

Annotation Interface. The core idea in AI-assisted labeling (or annotation) is to take advantage of AI techniques for reducing the cost and labour-effort needed for collecting high quality labeled data. To assess the benefits and limitations of AI-assisted interactive annotation in real-time labeling tasks, we designed an AI-assisted annotation interface that is motivated by mixed-initiative interaction [14, 22] and interface design for video annotation [43]. In line with the core principles of mixed-initiative interfaces, the main control is with the user performing the annotations, and the AI is merely a support tool that offers suggestions. The interface used for annotation has been designed to enforce real-time annotation by not allowing any frame-per-frame video analysis, such as pausing, rewinding, fast forwarding, speeding-up or slowing-down the video feed. The interface, shown in Figure 6, was designed as web-based and to be used on portable devices. The interface assumes video clips to be played in sequence (following the same order as in Table 1) where once the first video clip "stream" ends, the interface awaits for annotator confirmation to proceed with the next video. Annotations are made on separate buttons (5 buttons, one for each animal type); see Figure 6a. Restricting the number of options is necessary to ensure the annotators can choose the right option in real-time and this is in line with other real-time annotation tasks, as described in Section 2. Once the button is pressed, the annotation remains active until the button is pressed again. Annotation can be used with or without AI-assistance. When the AI-assistance is enabled, the AI benchmark model described in the previous section was used to perform real-time inference and to visualize the results as bounding boxes and confidence scores overlaid on top of the video whenever the AI detects an animal in the video; see Figure 6b. The motivation is to direct the user's attention to a sub-region that has the highest likelihood of containing relevant information, thus minimizing cognitive overload. This strategy is also supported by the literature where similar strategies have been shown to be effective at improving labeling efficiency for other content types, such as document labels [11]. To give visual feedback to the annotators, the interface displays the text "I see: «specie(s)»" while the button is in the ON state. Although real-time video streams typically do not contain indicators of video length, we included a vertical bar matching the video clip duration in the interface to give the test subjects an indication of task length and to ensure they do not get frustrated with the task length as this could decrease their performance.

Participants. In total, 34 participants were recruited for the study. This comprised of participants with domain expertise (E) and those without (NE). As the experts, we consider 14 marine ecologists all of whom had prior field experience in recognizing aforementioned marine species. Additional participants with no domain expertise (NE) were 20 CS students, who did not have knowledge of those species. Both the expert and non-expert groups were split into even-sized subgroups with one using AI-assistance and the other not. Participants of the controlled groups (E, NE) use the non-AI annotation interface[2] (see Figure 6a) whereas those of the treatment groups (EAI, NEAI) use our developed AI annotation interface (see Figure 6b). We run the study as between-groups design as the tasks were the same and knowledge of the video contents would affect performance significantly. As ground truth, we consider the annotations from the two separate researchers, as described in Section 3. As an additional baseline for comparison, we consider the annotations made solely by the benchmark AI model trained with 100k iterations; see Section 3.

Procedure. The user study was conducted in blocks of 75 minutes per each human annotator (groups E, EAI, NE, and NEAI), during 4 consecutive days. Each participant annotated 28 video clips (all videos from Figure 5), totaling 62.07 minutes. To simulate real-time video streaming, the playback of a video clip (30 FPS) cannot be paused once its start playing. Participants can have a break between video clips. All annotators were distributed into separate rooms and were given a laptop computer with the annotation web-based interface. Participants performed annotation using a computer mouse to select any of the 5 objects of interest. The research author showcased the annotation procedure to each annotator individually using the first video after which participants started the annotation starting from the beginning on that same video. At the end of the experiment, participants were invited to share their thoughts and provide open feedback. Participants were asked to turn off their mobile devices prior to the study.

[2]https://wave-labs.org

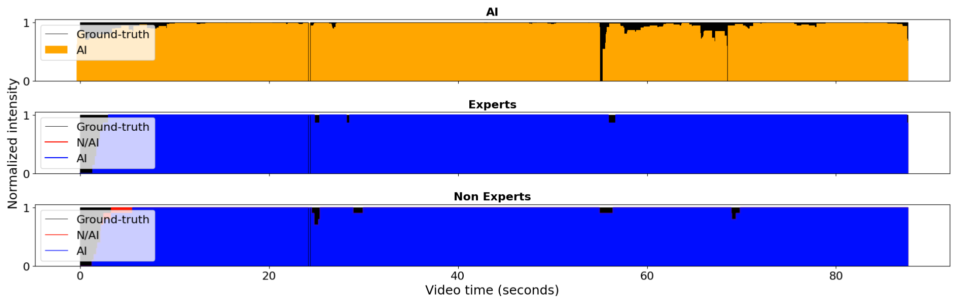


(a) Annotation UI (b) AI-assisted annotation UI

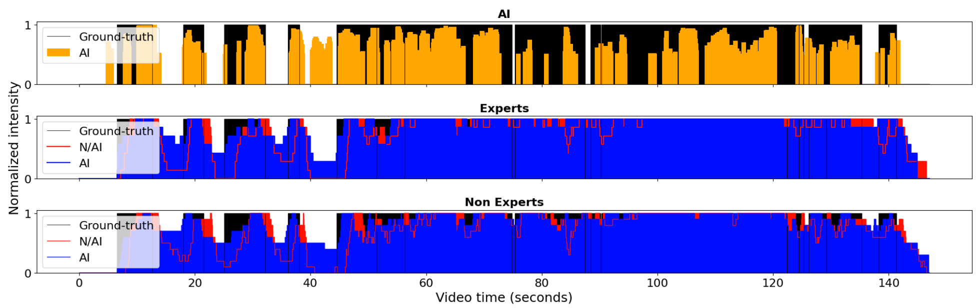
Fig. 6. Designed AI-assisted interface for real-time Video stream annotation.

Annotation Timelines as Evaluation Metric. We assess performance using annotation events which corresponds to situations where at least one of the buttons was in active state. We derive an annotation timeline from the events by considering the presence or absence of events in frames as a binary variable. When multiple species are spotted in a specific frame, we consider only the selection of the main object of interest (species) as a single event. Likewise in the case of AI, all multiple occurrences in video frame are treated as a single annotation event. Since we consider either one or multiple annotations per frame as single annotation event, we stack and normalize all annotation timelines across one group, allowing us to empirically compare different groups. An example is shown in Figure 7. The maximum on the vertical axis (i.e., 100%) is reached when all individuals from the same group annotate an at that specific video frame at least one object. GT annotations were always kept at the maximum (100%) as they were derived as a consensus. For AI annotations the ordinate value represents the confidence score, where the minimum threshold (50% or .5) was used as suggested by the literature [63]. Figures are depicted with the ground truth (GT) in black color, while the model predictions (AI) is in yellow color. The AI-assisted groups are seen in red color and non-AI assisted groups are seen in blue color. Different signal patterns may be observed. For instance, observing the formed areas, Figure 7a depicts an overlap among all 6 groups, suggesting an easy video. Conversely, Figure 7b showcases greater differences in areas among groups, indicating that the video may be considered as difficult.

Data Inquiry. To further compare obtained timelines between different sample groups, we calculate the annotations per each video clip frame, per each annotator group (E, EAI, NE or NEAI). For each video clip, each user annotated frame was multiplied with the baseline frame (GT, AI, or another user group). Multiplication with AI signal indicates the impact of AI, the multiplication with GT indicates the annotation performance, and multiplication with another user group indicates correlation between annotations of both groups. We further sum and normalize these product frames, obtaining the correlation coefficient. This metric is known as the normalized cross-correlation (NCC) or Pearson Correlation Coefficient, where the product between two groups is a coefficient between 0 and 1. The higher the NCC value, the higher is the agreement between the two sample annotation groups. NCC per each video clip and per each



(a) Easy video timeline



(b) Difficult video timeline

Fig. 7. Example of annotation timelines depicting timestamps for each sample group and for single videos. The maximum on the vertical axis (i.e., 100%) is reached when all individuals from the same group annotate at that specific video frame at least one object. GT annotations were always kept at the maximum (100%) as they were derived as a consensus. Color legend: (AI) model results (yellow), (GT) ground truth by research authors (black), (E) experts or (NE) non-experts (red) and (EAI) experts with AI or (NEAI) non-experts with AI (blue).

sample group is computed using next expression:

$$NCC_{(x,y)} = \frac{\sum_{n=1}^{N} x_{(n)} * y_{(n)}}{\sqrt{\sum_{n=1}^{N} x_{(n)}^2 * \sum_{n=1}^{N} y_{(n)}^2}}, \tag{1}$$

where x and y are two different sample groups, and n and N are current and total amount of frames in a given footage, respectively. Based on the NCC scores, we group videos into two categories: hard (NCC $<$ .95 correlation) and easy (NCC $\geq$ .95 correlation) videos. Videos with NCC score of 1.0 (i.e., 100% correlation) are consider to be the easy as there is full agreement between the two annotator groups. NCCs are further compared with aforementioned video clip parameters, such as field of view, video quality, visibility, video pace; see Table 2.

4.2 Experimental Findings

We first briefly provide the obtained findings through video parameter analysis where we detail about the most representative parameters. Next, we compare the in-between group agreement, observing whether there were any differences among the user group annotators. Afterwards, we check whether there was an influence of the AI to the annotator groups. Finally, we check to which extent was the AI-assisted interface providing the influence (whether positive or negative) on all sample annotator groups.

Video parameters analysis. We start with aggregating all NCCs in Table 3 (columns D-I), and observing how they relate to the proposed categorization of video parameters. Considering all NCCs with the aforementioned threshold (.95), we define the easy and difficult videos for annotators in column C (presented previously in Table 2). Regarding the parameter "field of view" (FOV), a mix of multiple options (being with video scenes changing from a boat to a diver, to a drone) was found to harm the correlation between groups, resulting in hard videos. Videos exclusively taken from Unmanned Aerial Vehicles (UAVs) or from solely diving footage resulted in "easy" videos. Similarly, as seen with parameter "type", a combination of multiple typologies of images (being surface, aerial or underwater) in the same video clip also resulted in worse correlations. Only underwater- or aerial- only videos were perceived as being easy videos. Parameter "recording" seems to follow the parameter "FOV" where amateur videos were mostly taken from the boat, or as in surface as parameter "type". Additional inspection of low NCC correlations happens when seeing from surface the underwater species, which was also the case in our previous study. In such cases, sea glare or splashes hinder the visibility of species, which is also noticeable in the AI timeline. Parameter "quality" of the footage did not affect the correlation as not enough lower-quality videos were used throughout the study. Regarding the parameter "visibility", all clips with objects occupying appx. 50% of the screen resulted in being easy videos. The "pace" parameter of the video was shown to be the variable of the most impact which is presented as lowest correlation scores in Table 3 seen as column "hard". This showcases a linear agreement that if the video scenes are with the rapid change (fast, or extremely fast abbreviated as "fast+" in the "pace" column), then it results in being a hard video. Conversely, slow pace videos were with high correlations (NCC $\geq$ .95).

Table 3. Normalized cross correlation (NCC) coefficients and T-test scores with p-values for in-between groups.

			Paired t-Test p values						NCC Groups vs AI				NCC Groups vs GT				NCC In-between group agreement						NCC
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
#	Object	Difficult	NE-NEAI	NE-E	NE-EAI	NEAI-E	NEAI-EAI	E-EAI	E-AI	EAI-AI	NE-AI	NEAI-AI	E-GT	EAI-GT	NE-GT	NEAI-GT	NE-NEAI	NE-E	NE-EAI	NEAI-E	NEAI-EAI	E-EAI	AI-GT
1	Birds		.03	.45	.64	.02	.22	.39	.93	.93	.93	.94	.99	.99	.99	.99	1	1	1	1	1	1	.93
2	Birds		.49	.93	.26	.48	.89	.28	.70	.71	.70	.71	.98	.98	.98	.98	1	1	1	1	1	1	.69
3	Birds	x	.11	.40	.02	.40	.52	.09	.54	.57	.54	.58	.90	.93	.90	.93	.96	.99	.96	.96	.99	.97	.56
4	Whales		.45	.31	.96	.31	.55	.55	.93	.92	.92	.92	.99	.99	.99	.99	1	1	1	1	1	1	.93
5	Whales		.98	.05	.15	.06	.15	.07	.85	.85	.85	.85	.97	.96	.96	.96	.99	1	.99	.99	.99	.99	.85
6	Whales		.37	.42	.51	.92	.22	.22	.56	.56	.55	.55	.97	.96	.97	.96	1	.99	.99	.99	.99	.99	.54
7	Seals	x	.46	.47	.02	.87	.25	.14	.76	.75	.77	.76	.95	.94	.95	.94	.99	1	.99	.99	.99	.98	.76
8	Seals		.29	.96	.46	.37	.80	.53	.29	.29	.29	.30	.99	.99	.99	.99	1	1	1	1	1	1	.29
9	Seals		.22	.69	.32	.41	.89	.47	.70	.70	.70	.69	.99	.99	.99	.99	1	1	1	1	1	1	.70
10	Turtles		.35	.61	.23	.76	.13	.02	.91	.91	.91	.91	.96	.96	.96	.96	1	1	1	.99	1	.99	.92
11	Turtles		.03	.82	.16	.16	.82	.21	.93	.93	.93	.92	.99	.99	.99	.99	1	1	1	1	1	1	.94
12	Turtles		.06	.68	.14	.14	.31	.29	.89	.88	.89	.88	.98	.98	.98	.98	1	1	1	1	1	1	.90
13	Dolphins	x	.49	.71	.74	.46	.79	.50	.00	.00	.00	.46	.07	.18	.09	.35	.40	.34	.69	.56	.65	.50	.00
14	Dolphins	x	.21	.50	.23	.47	.95	.55	.09	.21	.10	.17	.21	.27	.13	.21	.76	.85	.82	.85	.93	.83	.08
15	Dolphins	x	.99	.13	.28	.12	.27	.69	.00	.00	.00	.00	.68	.63	.69	.67	.97	.98	.96	.96	.95	.97	.00
16	Dolphins	x	.33	.03	.06	.46	.58	.84	.22	.23	.20	.23	.92	.93	.92	.93	.98	.97	.98	.99	.99	.99	.24
17	Dolphins	x	.17	.52	.21	.20	.32	.12	.36	.53	.54	.59	.61	.55	.56	.40	.79	.98	.93	.78	.80	.92	.70
18	Dolphins	x	.25	.82	.19	.36	.95	.14	.79	.77	.77	.76	.88	.86	.86	.88	.98	.98	.99	.98	.99	.98	.80
19	Dolphins	x	.07	.67	.14	.15	.23	.57	.40	.46	.35	.45	.61	.54	.53	.57	.95	.95	.99	.94	.96	.95	.73
20	Dolphins		.16	.19	.12	.88	.46	.28	.54	.54	.53	.56	.97	.96	.96	.96	.98	.99	.98	.99	.99	.99	.54
21	Dolphins	x	.14	.05	.01	.89	.73	.61	.34	.32	.13	.37	.36	.38	.19	.27	.59	.66	.71	.95	.80	.85	.71
22	Dolphins		.90	.95	.07	.90	.44	.16	.62	.60	.62	.63	.97	.95	.96	.96	.99	.99	.99	.99	.99	.98	.60
23	Dolphins		.10	.43	.67	.18	.12	.43	.67	.66	.66	.68	.99	.99	.99	.98	.99	1	1	.99	.99	.99	.65
24	Dolphins	x	.71	.44	.42	.93	.41	.28	.55	.58	.56	.56	.95	.93	.94	.92	.98	.99	.99	.96	.98	.97	.56
25	Dolphins	x	.78	.53	.01	.95	.14	.04	.59	.62	.61	.63	.94	.93	.94	.93	.98	.98	.98	.96	.98	.96	.61
26	Dolphins	x	.30	.47	.85	.14	.32	.46	.34	.35	.34	.37	.96	.95	.96	.93	.98	.99	.99	.96	.98	.98	.38
27	Dolphins	x	.37	.03	.08	.91	.95	.68	.05	.16	.12	.22	.76	.86	.75	.82	.78	.88	.81	.78	.84	.79	.11
28	Dolphins		.47	.99	.03	.56	.28	.35	.03	.03	.01	.14	.97	.96	.97	.96	.99	.99	.99	.98	.99	.98	.00

In-between Groups Agreement. To understand if applied methodology caused any impact on different user groups, paired two-tailed t-tests were computed for all 6 combinations (NE-NEAI, NE-E, NE-EAI, NEAI-E, NEAI-EAI, E-EAI), presented in Table 3 (columns D-I). Next color coding was used to depict the obtained p-values: extremely significant differences (dark red, $p < .0083$ with Dunn-Bonferroni correction), statistical significance (red, $.0083 < p < .05$), no statistical differences (dark green, $.05 < p < .5$), mild agreement (light green, $.5 < p < .95$) and high agreement (white, $.95 < p < 1.0$). Then, we performed a count of aforementioned color coding cells, thus indicating tendencies of disparity between all the combinations. We identified a relevant change in the annotations of non-experts, due to existing disagreement from NE with E and EAI, while NEAI indicates agreement with same groups (E and EAI). This suggests that annotations from non-experts were somehow affected, which led their results to switch from disagreeing to agreeing with experts' annotations. Further analysis found mixed agreement in combinations NE-NEAI and E-EAI, indicating adjustment of annotations to our system.

Correlation with AI. We next analyze the impact of AI on the sample annotators by computing the NCCs using the AI as the baseline (Table 3, columns J-M). Averaged NCCs are computed and shown for both easy and hard videos per group, as seen in image 9a. The correlation of all groups with AI indicates that both non-experts (NE) and experts (E) were affected by the AI. T-test for paired two-sample towards means indicated extremely relevant statistical differences between easy and hard videos among all groups ($p < .005$). On average, for all videos, non-experts (NE) increased an agreement with AI by 5% (going from average .52 to .57 NCC). Experts increased from .52 to .54 on average NCC score, showing a 2% effect. However, looking only at hard videos, experts increased agreement with AI by 4% (going from .36 to .40 NCC) while non-experts increased agreement by 8% (going from .36 to .44). In contrast, the easy videos portray similar agreement with or without the AI, deviating only 1% for non-experts (going from .68 to .69) while creating no significant changes to experts. In addition, when comparing in-between groups agreement and observing summary Figure 8, some statistical differences are visible ($p < .05$). Combinations of different interfaces, meaning that one group is with the AI-assisted interface and the other has the normal interface (NE-NEAI, NE-EAI, E-NEAI, and E-EAI) paired t-test for means indicated statistical difference for all pairs ($p < .05$), as seen if summary Figure 8 with full lines. Conversely, performed paired t-test for means between groups using the same interface (groups on the same side of the figure with combinations in dashed lines), demonstrate mixed levels of agreement for combinations E-NE and EAI-NEAI. It was not possible to identify if the performance of AI affected the performance on users using the AI-assisted interface, presented in Table 3, column X. Thus, results indicate that a difference between the annotations of groups exists, and confirming the AI-assisted annotation interface as the source of these differences. In the following, we will analyze to what extent. In overall, the non-experts (NE) annotators were mostly affected by the AI-assisted interface. We estimate that even if AI classifications were incorrect, they managed to improve the attention span of some users, i.e. by providing visual feedback with bounding boxes and confidence levels (both low and high). This is also in line with participants' comments after the end of experiment: "AI was so wrong", "AI identified boats as whales", and "AI sees people as seals". This indicates that the AI-assisted interface managed to captivate the non-expert's attention towards detecting the objects of interest in footage and therefore improve their engagement.

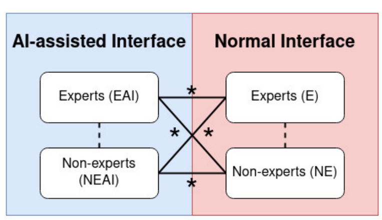
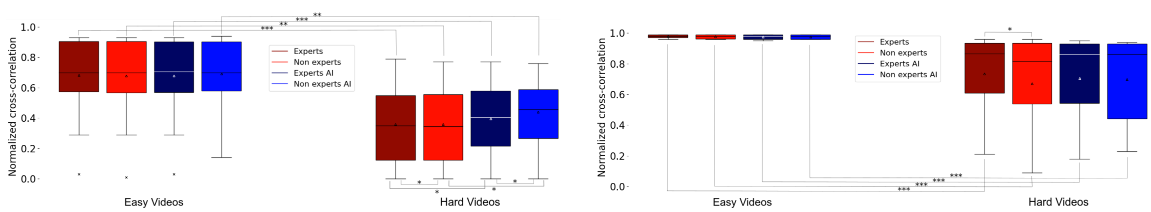


Fig. 8. Obtained statistical significance between groups and used interfaces. Full lines indicate statistical significance, while dashed lines indicate no statistical significance. Asterisk (*) indicates significant statistical significance with $p < .05$.

Correlation with GT. Next, we validate the NCCs of each annotator group against the ground-truth (GT) as a baseline to portray the overall annotation performance of all annotators (Table 3, columns N-Q). When observing differences between hard and easy videos for all groups, paired two-tailed t-test towards means indicates extremely relevant statistical differences ($p < .005$). Results are seen in Figure 9b, which depict easy videos with a grand average value of .98, while hard videos grand average obtained was .70. All videos demonstrate differences between results of groups using the AI-assisted interface. In all videos, experts saw a decrease of performance by 2% (decreasing from .86 to .84) and non-experts an improvement of 1% (going from .83 to .84). For easy videos only, the low standard deviation is observed in each annotator group, indicating almost perfect agreement (all groups with .98 NCC). However, meaningful differences occurred as a consequence of hard videos, where experts dropped by 4% (going from .74 to .70 NCC) and non-experts improved by 3% (going from .67 to .70 NCC). Findings indicate that the AI affected both groups, where the non-experts (NE) were affected positively, while experts (E) were affected negatively. Regarding t-tests towards the coefficient correlation between groups, only single statistical difference was identified with $p =< .05$ between NE vs E for hard videos (Figure 9b). These findings suggest that initially, non-experts' annotations were significantly different when compared to experts. However, while using the AI-assisted interface both groups achieved a bigger agreement, thus removing the inherited statistical difference. Therefore, we find non-experts' annotations with an AI-assisted interface to be similar with experts without the AI-assistance.



(a) Annotator groups with AI (easy against hard videos)

(b) Annotator groups with GT (easy against hard videos)

Fig. 9. Correlation with AI and GT as two separate baselines for both hard and easy videos. Asterisks indicate next p-value significance: (*) $p < .05$, (**) $p < .005$, and (***) $p < .0005$.

In-between Groups Agreement. Lastly, we computed the grand average NCCs between all combination pairs of all 4 annotator sample groups (Table 3, columns R-W): (i) NE-NEAI; (ii) NE-E; (iii) NE-EAI; (iv) NEAI-E; (v) NEAI-EAI; and (vi) E-EAI. The biggest agreement resulted in combination (v), with a grand average of NCC .96, supporting previous results. Moreover, combination (i) had the smallest grand average NCC .93. This suggests that the non-experts (NE) were the group that was the most affected by the AI-assisted interface. Additionally, the comparison between experts and non-experts with and without AI, saw a slight increase of correlation on hard videos by .1 NCC when using the AI. Therefore, the non-experts saw a bigger agreement with experts when interacting with the AI-assisted interface. As described previously, significant differences occurred from the comparison of groups solely on hard videos, due to easy videos varying .003 between all groups with a maximum NCC of .996, and a minimum NCC of .993. Hard videos had a minimum NCC of .863 and a maximum of .917, with a total variance of .054.

Results Summary. The main results are summarized as follows: (i) AI-assisted interface improved the performance of non-experts, making them annotate as experts; (ii) AI-assisted interface worsen the performance of experts, due to inherited habits and distraction; (iii) Video's pace was the most discriminant video parameter being proportional to the NCC correlation threshold ($\leq$ 95 for easy videos); and (iv) Easy videos presented the highest levels of agreement among participants, while hard videos were mostly responsible for any significant differences between groups' annotation scores.

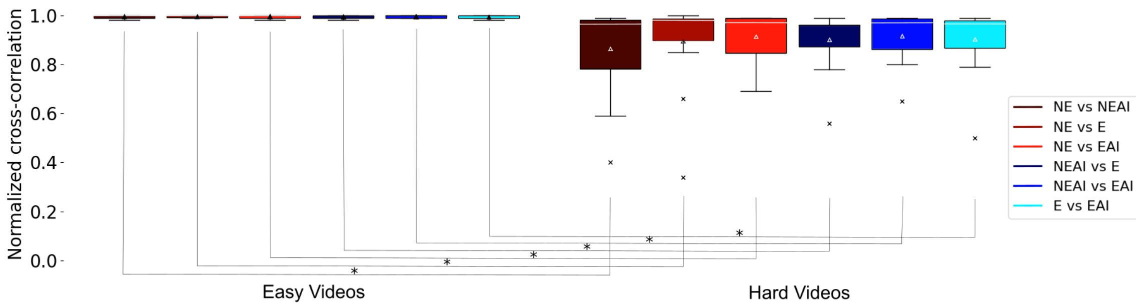


Fig. 10. Correlation between groups for hard and easy videos using ground truth (GT) as a baseline. NE: non-experts; NEAI: non-experts with AI; E: experts; and EAI: experts with AI. Asterisks indicate next statistical significance: (*) < .05, (**) $p < .005$, and (***) $p < .0005$.

5 EFFECTS OF ANNOTATIONS ON AI PERFORMANCE

The results of the previous section demonstrated that AI-assistance can have significant impact on human performance but this is moderated by human expertise. We next demonstrate that AI-annotation not only affects the accuracy of labels but also affects the performance of AI models that are trained on the annotated data. To assess this, we compare how the labels derived with or without AI-assistance impact AI models that are trained with such data – the ultimate goal of any annotation process.

5.1 Experimental Setup

From the 28 videos, we focus on the annotation frames obtained from Experts (E), Non-Experts (NE), Experts with AI (EAI), and Non-Experts with AI (NEAI). For each annotator group we train a multi-class image classification model based on traditional LeNet architecture [33] using 6 classes ("whale", "dolphin", "seal", "seabird", "turtle" and "none"). Note that the reference AI model was designed to perform object classification and present confidence scores in images, whereas we are simply interested in detecting whether the frame contains an animal or not. As this differs from the task of the reference model, we use a simpler model that has better discriminating power in classification tasks. Object detectors can also be converted into classifiers by applying a threshold on the confidence scores or by comparing the intersections of unions between detected objects and ground truth. However, this is sensitive to the threshold that is used for comparison, and hence a dedicated classifier is likely to achieve better performance. For each frame in a video for annotation class we consider the majority of annotators in a single frame (e.g. if 3 out of 5 annotators have seen an object of interest inside of the same frame, we consider that frame with one of the species classes). Conversely, if the majority of annotators agree that there are no objects of interest in the video, we assign it a class "none". Five models were trained (the four user groups and also ground truth) with average 100 epochs, reaching early stopping if there has been no improvement in validation accuracy for 10 consecutive epochs. We used standard hyperparameters and model choices: the loss function was categorical cross-entropy, the optimizer was Stochastic Gradient Descent (SGD), and learning rate was set to .001. From all videos, each image frame has been downsampled to $300 \times 300px$ resolution and was used as an input into the neural network. Used batch size was 256. The default aspect ratio was preserved by filling in with white color the remainder vertical spacing. Due to resource constraints, we downsample the input dataset imagery by selecting every second frame from each video (i.e., the input frame rate is 15 fps). The total amount of images and labels were 51963. All images were further split into three sets training (80%), validation

Table 4. Confusion matrices of AI models trained with annotated frames using randomized whole videos (inference on testing set).

E - Accuracy = .828

Actual \ Predicted	0	1	2	3	4	5
0	**255**	42	466	15	63	28
1	20	**758**	2	7	14	0
2	168	1	**1967**	2	5	3
3	2	2	0	**244**	3	3
4	15	2	7	3	**316**	11
5	2	0	1	0	0	**740**

EAI Accuracy = **.867**

Actual \ Predicted	0	1	2	3	4	5
0	**183**	36	378	10	33	29
1	6	**786**	2	3	10	1
2	106	5	**2197**	0	4	0
3	3	3	1	**240**	4	8
4	20	3	7	0	**334**	12
5	0	2	0	0	2	**739**

NE Accuracy = .842

Actual \ Predicted	0	1	2	3	4	5
0	**223**	41	472	18	55	28
1	5	**784**	3	1	8	2
2	80	1	**2086**	1	4	2
3	3	5	3	**229**	4	9
4	26	10	13	7	**292**	6
5	2	2	5	0	0	**737**

NEAI Accuracy = .841

Actual \ Predicted	0	1	2	3	4	5
0	**175**	60	480	17	33	31
1	5	**769**	5	2	5	0
2	73	1	**2128**	1	1	4
3	4	9	3	**214**	12	8
4	12	11	23	0	**326**	13
5	2	3	1	1	2	**733**

GT Accuracy = .807

Actual \ Predicted	0	1	2	3	4	5
0	**138**	59	562	10	45	49
1	5	**748**	8	12	13	1
2	68	8	**2088**	1	5	8
3	4	3	17	**219**	8	9
4	16	16	29	3	**259**	28
5	3	0	7	0	0	**718**

AI Accuracy = .784

Actual \ Predicted	0	1	2	3	4	5
0	**3497**	19	8	45	33	92
1	125	**14**	0	9	0	0
2	555	0	**2**	0	0	3
3	50	0	0	**149**	0	1
4	65	0	0	0	**124**	0
5	109	0	0	0	0	**267**

(10%), and testing (10%), comprising 41598, 5198, and 5167 images respectively. This dataset is comparable to datasets that are commonly used to evaluate object detection models (see the discussion) and is sufficient for showcasing the potential and pitfalls of annotations. Results of model inference for 6 AI models based on their annotations including training and validating performance are depicted in Table 4.

5.2 Experimental Findings

Model trained with randomized annotations from experts with AI (EAI) obtained the highest performance (.867 accuracy), followed by the experts without AI. For non-experts, AI improved the quality of annotations but did not improve the performance of AI models that are trained from the data. The performance of the AI without any human annotations was the lowest, indicating human annotations are necessary. Using ground truth to train the AI model actually decreased overall performance compared to experts or non-experts. Inspection of the confusion matrices indicates that this is due to the model trained with ground truth data missing the animals in many cases even if it can accurately distinguish between them. Overall, the results thus suggest that AI can improve the consistency in expert labels and this combination is best for training AI models. For non-experts, annotation quality goes up but this does not necessarily translate into better AI accuracy as there can be more variation in the annotations.

6 DESIGN GUIDELINES AND DISCUSSION

Based on findings from the our experiments, in the following we discuss lessons learned to derive some guidelines for delivering successful AI-assisted annotations of video streams. In addition, naturally there is also room for improvement and further work and below we discuss some of these points.

Assessing Annotator's Expertise. Infusing AI assistance into annotation interface draws annotators' attention the detected objects of interest, helping them reduce decision space. As demonstrated in our experiments, AI-assisted annotation interface helps non-experts to annotate better but reduces experts' annotation performance. These findings highlight applying AI intelligence in real-time annotation of video streams is not "one-size-fit-all" but that it is necessary to assess annotators' expertise level before recommending an AI-assisted annotation interface to annotators. At the same time, the AI benefited both groups – raising the level of non-experts while reducing the risk of sleeping for expert users. Thus, the nature of the AI assistance should be tailored according to user expertise levels.

Annotation Input Modality. Real-time annotations are also affected by latency and cognitive bandwidth of the users. Indeed, for fast-paced videos we observed that some users struggled to keep pace with the video and this resulted in incorrect labels due to delayed reactions. Minimizing this effect requires optimized interface designs for the interface so that annotators can quickly choose the correct labels without it demanding significant amounts of cognitive resources.

For binary tasks this is reasonably simple whereas for multi-class tasks this is non-trivial as any shortcuts easily demand cognitive resources (especially memory) which risks errors in the annotations. At the same time this result also enforces the need to keep the number of classes low and sufficiently distinctive to ensure the annotators' cognitive resources are not overwhelmed by the selection of right label to apply.

Engagement and Task Duration. The non-expert group had two outlier users that consistently left the annotations on regardless of there being AI assistance or not (so-called sleepers). Findings from the post-task survey indicated that participant engagement was a critical factor in the attention they paid on the videos and hence ensuring sufficient engagement is essential for accurate labels. Non-experts indicated that they felt fatigued faster than experts, and hence there is a need to incorporate techniques that help maintain engagement over a longer period of time. One option is to rely on gamification techniques as these can potentially help to preserve engagement, as has been shown, e.g., for web-based interfaces [10].

Annotating in Context. The experiments simulated in-situ settings by playing back the video feed without options to pause the content. In some applications, such as detecting suspicious activity from surveillance feeds, this is sufficient but in other tasks there are further aspects to consider. For example, in our target domain, biodiversity annotation, the annotations need to be made on-board sea vessels – or at least on content captured from sea vessels. Waves and swaying of the vessel would affect the annotation process and make it harder to perform the annotations accurately. Potential ways to overcome this is to use a remote interface, e.g., a kiosk or a remote computer, that receives a real-time feed of the imagery captured by the vessel instead of having the annotator onboard the vessel.

Real-time AI. Incorporating the AI model as part of the annotation process requires the inference process to be able to keep up with the pace of the input data as otherwise the AI induces latency onto the process and can result in additional errors. This either requires sufficient resources and heavy optimization of the AI models or adjusting the rate of the input data. The AI model used in our experiments and on our hardware supports around $10 - 15$ frames per second. While more than sufficient for our target application of biodiversity estimation, this would result in noticeable delays in the video stream and potentially decrease engagement. Thus, in real-time applications, it is necessary to find optimal balance between AI performance and the final user interface.

Over-reliance and Ethics. We demonstrated that, when used correctly, AI-assistance can provide significant benefits for annotations, and improve the AI models that are trained on the labeled data. Conversely, there are also risks in adopting AI-assisted labeling. For example, people can become over-reliant on AI assistance, which can decrease label quality and degrade the performance of the AI models trained on the labeled data. Mitigating over-reliance is currently an active research area and there are solutions that could be adopted to facilitate overcoming this issue. For example, brief explanations can reduce over-reliance [57], but at the same time integrating explanations into dynamic real-world scenes is challenging as they can further increase the annotator's cognitive overhead. There are also differences across individuals on how easily they become reliant on AI performance and thus it is important to be aware and analyze potential effects of over-reliance. Adopting AI-assistance should also account for potential ethical issues. For example, when annotating data with human subjects, such as detecting suspicious activity, it is important to ensure the annotations are fair and free of subjective biases. Following the principle of transparency, users of annotated data should also be able to obtain information about the annotation process, including the people that performed the annotations and their background. These issues, however, are not unique to AI-assisted annotation but hold generally for any AI-based systems [29].

Room for Improvement. As with any research, naturally there is room for further work. Our experiments focused on a single target domain (marine biodiversity) and other domains may have different characteristics that affect annotation performance. The experiments on the effect of annotations on AI performance could also be repeated with larger datasets and other AI models. The dataset we considered is comparable to those that are used to test image recognition models. For example, MobileNet was tested using the CIFAR-10 dataset which has 60 000 samples [51] whereas our experiments considered 51 963 images (19 853 images were used for training the baseline model). As AI models are becoming increasingly popular, it is likely that AI-assistance would be adopted using commonly available architectures and default parameters. While other AI models might perform differently, we considered standard architectures (MobileNet and LeNet) that are commonly used as off-the-shelf tools for object recognition and thus our setup represents issues that a non-expert would face when adopting AI-annotation support. Architectures used in this study are relatively shallow compared to most recent state-of-the-art (e.g. ResNets), which makes them less likely to overfit. Taken together, our experiments should provide a sufficient foundation for analyzing AI annotation (in dynamic real-time application use cases). Nevertheless, we fully acknowledge the need for future work on benchmarking more complex (neural network) architectures, and to further explore the extent that AI may be improved by the annotators. Finally, the experiments considered scenes where at most a single object would be visible at a time. Real-world scenes may be more complex and have multiple targets that need to be detected simultaneously. In such scenarios, users might rely more on AI-annotation support to reduce cognitive demands of tracking and identifying targets. These are but some examples of future directions for our work.

7 RELATED WORK

Human-in-the-loop machine learning (HITL-ML) broadly encompasses: (i) Active Learning (AL) where the system remains in control, (ii) Interactive Machine learning (IML) with closer interaction between users and system and (iii) Machine Teaching (MT) where the domain experts have a control over the learning process [40]. Another categorization of HITL-ML has been proposed to be focused on: (i) improving model performance, (ii) improving model through intervention and (iii) system independent HITL [65]. Recent HITL-ML minimizes human queries which are typically required to train complex models [20]. Focused on IML and on understanding what is the effect of the AI on video annotators and vice versa, our research is inspired by previous works on AI-assisted labeling and studies that compare human domain expertise and AI performance. Below we briefly summarize relevant works in these fields.

AI-assisted labeling. Solutions for AI-assisted labeling can be categorized into model-based and interactive solutions with our research falling into the latter category. Model-based solutions attempt to identify candidate patterns that would be useful for labeling and to query a human annotator to label these patterns. The resulting labels can then be propagated to other data points that are sufficiently similar. Examples of model-based solutions include active learning [42, 67], semi-supervised learning [16], and few-shot learning [66]. Interactive solutions, in turn, focus on offering interactive feedback that can assist in assigning labels. Examples of interactive solutions include interactive visualizations of patterns [3, 55], the most likely labels [12], or the identification of new labels [14]. At best, AI-assisted labeling can increase human accuracy and decrease the time that is required for labeling [12, 14, 21, 56]. However, there are also concerns that excessive use of AI can result in over-reliance on the AI and result in decreased quality of data [2].Assigning the main responsibility on the human annotators can also be problematic. For example, if the set of labels is not limited, this can result in the labels diverging and the quality of the labels decreasing due to differences between the annotators [14]. Existing studies on the effects of AI-assisted labeling have focused on tasks where annotators can scrutinize and revise their annotations and the studies have largely focused on tasks where different cases can be easily distinguished. Our research contributes insights into the benefits and disadvantages of AI assisted annotations in complex real-world domains where the cognitive demands of the annotation compete with the AI support and where the distinctions between different categories are ambiguous.

Man vs. the Machine. Recent advance of machine learning (ML) user interfaces are seen in augmenting the human performance when performing different tasks [34, 64], involving human memory [39], assisting navigation of drivers [35], and aiding impaired senses of people [26]. Although human classification performance can be increased with the support of ML [45], ML solely is not yet comparable with human performance in complex situations, as such approaches are prone to errors and low precision in a wide variety of cases [30]. For instance, deep learning has been reported to achieve high classification accuracy with high resolution images, however its performance drops significantly when low resolution or blurred images are used as an input [13]. Other causes of errors include classification of objects from different angles and with diverse shapes [52], distances [15] and in distinct environmental/adversarial conditions [18]. The resulting classification errors can be severe and their exact cause can be challenging to understand thoroughly [30]. Although laborious, human observation remains to be the preferred means of classification for tasks that are complex or performance critical. Despite significant strides in AI and machine learning, humans continue to outperform the automated processes [17, 27]. Examples of such tasks include recognizing relevant data from noisy images [18, 28], e.g., CAPTCHA codes, and completing missing information [61]. More human-in-the-loop studies should be performed in involving human in video annotation.

8 SUMMARY AND CONCLUSION

Annotation of real-time video feeds is a difficult task where the annotators must divide their attention between the feed and the annotation interface. We studied the role AI-assistance has on annotation performance in real-time settings, and reversely how the differences in human labels affect performance of AI models. We compared two user groups, those with domain expertise and those without, in two conditions: with and without AI assistance. We found expert annotators generally having the highest performance. For non-experts AI significantly improves annotation performance and helps them to reach close to expert levels. When the annotated data is used to train AI models, expert users supported by AI have highest performance whereas for non-expert no improvements in AI performance can be observed. This largely stems from the consistency of the non-expert annotators having higher variation even when supported by AI whereas expert annotators tend to have more consistent agreements and disagreements. Based on our results, we discussed design considerations for interactive annotation of real-time streams and highlighted some open research issues. Taken together, our work offers new insights into designing AI-assisted annotation interfaces for real-time tasks and provides knowledge of how annotation characteristics influence the performance of AI models.

ACKNOWLEDGMENTS

The research was supported by the Foundation for Science and Technology (FCT) projects: (i) INTERWHALE - Advancing Interactive Technology for Responsible Whale-Watching (grant agreement: PTDC/CCI-COM/0450/2020), (ii) MARE - The Marine and Environmental Sciences Centre (grant agreement: UIDB/04292/2020), (iii) ARNET - Aquatic Research Network (grant agreement: LA/P/0069/2020), and (iv) PhD scholarship (grant agreement: 2022.09961.BD). It has been also financed by the EU Horizon Europe project CLIMAREST: Coastal Climate Resilience and Marine Restoration Tools for the Arctic Atlantic basin (grant agreement: 101093865), the Academy of Finland (grant number: 339614), the European

Social Fund via "ICT programme" measure, Estonian Center of Excellence in ICT Research (TK148 EXCITE), and the Nokia Foundation (grant number: 20220138). The authors thank the participants of our studies and the anonymous reviewers for their insightful comments.

REFERENCES

[1] Voncarlos M. Araújo, Ankita Shukla, Clément Chion, Sébastien Gambs, and Robert Michaud. 2022. Machine-Learning Approach for Automatic Detection of Wild Beluga Whales from Hand-Held Camera Pictures. *Sensors* 22, 11 (2022), 4107. https://doi.org/10.3390/S22114107

[2] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[3] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. 2017. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 298–308.

[4] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. 2018. VIAL: a unified process for visual interactive labeling. *The Visual Computer* 34, 9 (2018), 1189–1207.

[5] Riccardo Bertolo, Andrew Hung, Francesco Porpiglia, Pierluigi Bove, Mary Schleicher, and Prokar Dasgupta. 2020. Systematic review of augmented reality in urological interventions: the evidences of an impact on surgical outcomes are yet to come. *World journal of urology* 38 (2020), 2167–2176.

[6] Trevor Beugeling and Alexandra Branzan-Albu. 2014. Computer vision-based identification of individual turtles using characteristic patterns of their plastrons. In *2014 Canadian Conference on Computer and Robot Vision*. IEEE, USA, 203–210.

[7] Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research* 11 (1999), 131–167.

[8] John Calambokidis, Jay Barlow, Kiirsten Flynn, Elana Dobson, and Gretchen H Steiger. 2017. *Update on abundance, trends, and migrations of humpback whales along the US West Coast*. Technical Report SC/A17/NP/13. International Whaling Commision.

[9] Steven JB Carter, Ian P Bell, Jessica J Miller, and Peter P Gash. 2014. Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. *Journal of experimental marine biology and ecology* 452 (2014), 105–110.

[10] Chih-Ming Chen, Ming-Chaun Li, and Tze-Chun Chen. 2020. A web-based collaborative reading annotation system with gamification mechanisms to improve reading performance. *Computers & Education* 144 (2020), 103697.

[11] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[12] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 392–401.

[13] Samuel Dodge and Lina Karam. 2017. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE, Institute of Electrical and Electronics Engineers Inc., United States, 1–7.

[14] Cristian Felix, Aritra Dasgupta, and Enrico Bertini. 2018. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 153–164.

[15] Hongbo Gao, Bo Cheng, Jianqiang Wang, Keqiang Li, Jianhui Zhao, and Deyi Li. 2018. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics* 14, 9 (2018), 4224–4231.

[16] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*. Springer, Cham, United States, 510–526.

[17] Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. 2017. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *CoRR* abs/1706.06969 (2017). arXiv:1706.06969 http://arxiv.org/abs/1706.06969

[18] Philippe Golle. 2008. Machine learning attacks against the Asirra CAPTCHA. In *Proceedings of the 15th ACM Conference on Computer and Communications Security* (Alexandria, Virginia, USA) *(CCS '08)*. Association for Computing Machinery, New York, NY, USA, 535–542. https://doi.org/10.1145/1455770.1455838

[19] Michael Damien Haberlin. 2010. *Insights into jellyfish distribution and abundance provided by a platform of opportunity.* Ph. D. Dissertation. NUI.

[20] Donald Joseph Hejna III and Dorsa Sadigh. 2023. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*. PMLR, Auckland, New Zealand, 2014–2025.

[21] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[22] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 159–166.

[23] Guanxi Huang. 2021. A Comparative Study of Underwater Marine Products Detection based on YOLOv5 and Underwater Image Enhancement. *International Core Journal of Engineering* 7, 5 (2021), 213–221.

[24] Robert L Hulsman and Jane van der Vloodt. 2015. Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity. *Patient Education and Counseling* 98, 3 (2015), 356–363.

[25] Wu-Yuin Hwang, Chin-Yu Wang, and Mike Sharples. 2007. A study of multimedia annotation of Web-based materials. *Computers & Education* 48, 4 (2007), 680–699.

[26] Md Milon Islam and Muhammad Sheikh Sadi. 2018. Path hole detection to assist the visually impaired people in navigation. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*. IEEE, United States, 268–273.

[27] Robert A Jacobs and Christopher J Bates. 2019. Comparing the visual representations and performance of humans and deep neural networks. *Current Directions in Psychological Science* 28, 1 (2019), 34–39.

[28] Hojin Jang, Devin McCormack, and Frank Tong. 2021. Noise-robust recognition of objects by humans and deep neural networks. *bioRxiv* (2021). https://doi.org/10.1101/2020.08.03.234625 arXiv:https://www.biorxiv.org/content/early/2021/06/09/2020.08.03.234625.full.pdf

[29] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[30] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. 2019. Testing Robustness Against Unforeseen Adversaries. *CoRR* abs/1908.08016 (2019). arXiv:1908.08016 http://arxiv.org/abs/1908.08016

[31] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3075–3084.

[32] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982 [cs.CV]

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[34] Joe Lemley, Shabab Bazrafkan, and Peter Corcoran. 2017. Deep Learning for Consumer Devices and Services: Pushing the limits for machine learning, artificial intelligence, and computer vision. *IEEE Consumer Electronics Magazine* 6, 2 (2017), 48–56.

[35] Shih-Chieh Lin, Chang-Hong Hsu, Walter Talamonti, Yunqi Zhang, Steve Oney, Jason Mars, and Lingjia Tang. 2018. Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 531–542.

[36] Chenguang Liu, Xiumin Chu, Wenxiang Wu, Songlong Li, Zhibo He, Mao Zheng, Haiming Zhou, and Zhixiong Li. 2022. Human–machine cooperation research for navigation of maritime autonomous surface ships: A review and consideration. *Ocean Engineering* 246 (2022), 110555. https://doi.org/10.1016/j.oceaneng.2022.110555

[37] Rosalia Maglietta, Vito Renò, Giulia Cipriano, Carmelo Fanizza, Annalisa Milella, Ettore Stella, and Roberto Carlucci. 2018. DolFin: an innovative digital platform for studying Risso's dolphins in the Northern Ionian Sea (North-eastern Central Mediterranean). *Scientific reports* 8, 1 (2018), 1–11.

[38] Wei-Lung Mao, Wei-Chun Chen, Chien-Tsung Wang, and Yu-Hao Lin. 2021. Recycling waste classification using optimized convolutional neural network. *Resources, Conservation and Recycling* 164 (2021), 105132.

[39] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics* 19, 6 (2018), 1236–1246.

[40] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[41] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[42] Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. Alto: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1158–1169.

[43] M Radeta, Z Shafieyoun, and M Maiocchi. 2014. Affective timelines towards the primary-process emotions of movie watchers: Measurements based on self-annotation and affective neuroscience. In *9th International Conference on Design and Emotion (eds J Salamanca, P Desmet, A Burbano, G Ludden, and J Maya), Bogotá, Colombia*. Universidad de los Andes, Bogota, Colombia, 679–688.

[44] Marko Radeta, Agustin Zuniga, Naser Hossein Motlagh, Mohan Liyanage, Ruben Freitas, Moustafa Youssef, Sasu Tarkoma, Huber Flores, and Petteri Nurmi. 2022. Deep learning and the oceans. *Computer* 55, 5 (2022), 39–50.

[45] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun-Cheng Chen, Vishal M Patel, Carlos D Castillo, and Rama Chellappa. 2018. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine* 35, 1 (2018), 66–83.

[46] Peter J Rich and Michael Hannafin. 2009. Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection. *Journal of teacher education* 60, 1 (2009), 52–67.

[47] Eric Saund, Jing Lin, and Prateek Sarkar. 2009. Pixlabeler: User interface for pixel-level labeling of elements in document images. In *2009 10th International Conference on Document Analysis and Recognition*. IEEE, United States, 646–650.

[48] Tao Sheng, Chen Feng, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Mickey Aleksic. 2018. A quantization-friendly separable convolution for mobilenets. In *2018 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*. IEEE, United States, 14–18.

[49] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, New York, NY, USA, 614–622.

[50] Vikash Singh, Celine Latulipe, Erin Carroll, and Danielle Lottridge. 2011. The choreographer's notebook: a video annotation system for dancers and choreographers. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*. Association for Computing Machinery, New York, NY, USA, 197–206.

[51] Debjyoti Sinha and Mohamed El-Sharkawy. 2019. Thin MobileNet: An Enhanced MobileNet Architecture. In *10th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, UEMCON 2019, New York City, NY, USA, October 10-12, 2019*. IEEE, New York, NY, USA, 280–285. https://doi.org/10.1109/UEMCON47517.2019.8993089

[52] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng. 2012. Convolutional-recursive deep learning for 3d object classification. In *Advances in neural information processing systems*. Curran Associates Inc., Red Hook, NY, USA, 656–664.

[53] Mohammad Soleymani and Martha Larson. 2010. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*. ACM, Geneva, Switzerland.

[54] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 558–569.

[55] Yuandong Tian, Wei Liu, Rong Xiao, Fang Wen, and Xiaoou Tang. 2007. A face annotation framework with partial clustering and interactive labeling. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, United States, 1–8.

[56] Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. 2021. Biological data annotation via a human-augmenting AI-based labeling system. *NPJ digital medicine* 4, 1 (2021), 1–7.

[57] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[58] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

[59] Angelo Vittorio. 2018. Toolkit to download and visualize single or multiple classes from the huge Open Images v4 dataset. https://github.com/EscVM/OIDv4_ToolKit.

[60] Sonia Waharte and Niki Trigoni. 2010. Supporting search and rescue operations with UAVs. In *2010 international conference on emerging security technologies*. IEEE, United States, 142–147.

[61] Dylan Wang, Melody Moh, and Teng-Sheng Moh. 2020. Using Deep Learning to Solve Google reCAPTCHA v2's Image Challenges. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, United States, 1–5.

[62] Isaac Wang, Pradyumna Narayana, Jesse Smith, Bruce Draper, Ross Beveridge, and Jaime Ruiz. 2018. Easel: Easy automatic segmentation event labeler. In *23rd International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 595–599.

[63] Simon Wenkel, Khaled Alhazmi, Tanel Liiv, Saud Alrshoud, and Martin Simon. 2021. Confidence score: the forgotten dimension of object detection performance evaluation. *Sensors* 21, 13 (2021), 4350.

[64] H James Wilson and Paul R Daugherty. 2018. Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review* 96, 4 (2018), 114–123.

[65] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[66] Zhongwen Xu, Linchao Zhu, and Yi Yang. 2017. Few-shot object recognition from machine-labeled web images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, United States, 1164–1172.

[67] Jie Yang et al. 2003. Automatically labeling video data using multi-class active learning. In *Proceedings Ninth IEEE international conference on computer vision*. IEEE, United States, 516–523.

[68] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 516–520.

[69] Gang Zhai, Geoffrey C Fox, Marlon Pierce, Wenjun Wu, and Hasan Bulut. 2005. eSports: collaborative and synchronous video annotation system in grid computing environment. In *Seventh IEEE International Symposium on Multimedia (ISM'05)*. IEEE, United States, 9–pp.

[70] Neta Zmora, Guy Jacob, Lev Zlotnik, Bar Elharar, and Gal Novik. 2019. Neural Network Distiller: A Python Package For DNN Compression Research. *CoRR* abs/1910.12232 (2019). arXiv:1910.12232 http://arxiv.org/abs/1910.12232

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009