SGD11: A Dataset for Dynamic Underwater Scuba Gesture Recognition

E. Baker Herrin[1], Zi-Hao Zhang[1], Puneeth Sarmak[2], Xiaomin Lin[23], and Jane Shin[1]

***Abstract*— We introduce Scuba Gesture Dataset (SGD11), a diver gesture recognition dataset designed for the underwater human-machine teaming framework. This dataset consists of annotated video sequences capturing standard diver gestures used in underwater communication. The videos are collected for 11 hand gestures in a controlled underwater environment. We also present preliminary benchmark results comparing three gesture recognition models, showing that SGD11 can serve as a benchmark for advancing robust gesture recognition for autonomous underwater systems. SGD11 is publicly available to foster research in underwater human-robot interaction (UHRI) and autonomous diver assistance.**

I. INTRODUCTION

With recent advancements in autonomy, underwater robots have expanded their roles into navigation, exploration, and cooperative underwater operations with human divers. Underwater, unlike in terrestrial environments, communication among human divers relies primarily on visual and non-verbal methods. While there exist datasets for human action recognition, such as Kinetics [1], they do not generalize easily to the underwater environment [2], which highlights the significant need for datasets in underwater environments.

There exist underwater datasets for diver detection (VDD-C [3]) and datasets for diver pose and gestures from CADDY [4]. Specifically in diver pose and gesture recognition, the CADDY dataset is widely utilized for training gesture detection models. CADDY has made significant contributions to underwater gesture recognition, offering static stereo image pairs of gestures from the CADDIAN sign language, some of which are synchronized with diver pose data collected via a full-body IMU suit (DiverNet [5]). It has been adopted in multiple studies that focus on gesture detection accuracy and robust perception in underwater environments.

Our work, while similar, differs from existing work in two main aspects: the hardware used for data collection and the scuba gestures included. First, in terms of hardware, CADDY uses a stereo camera to collect static stereo image pairs, whereas we collect monocular video clips. Another difference lies in how pose data is obtained; CADDY uses a suit of IMUs, while we use an underwater motion capture system. Second, in terms of gestures, CADDY focuses on the CADDIAN language developed by the authors for human-robot communication. In contrast, we emphasize common scuba hand signals used in diver-to-diver communication, adapted here for interaction with a robot.

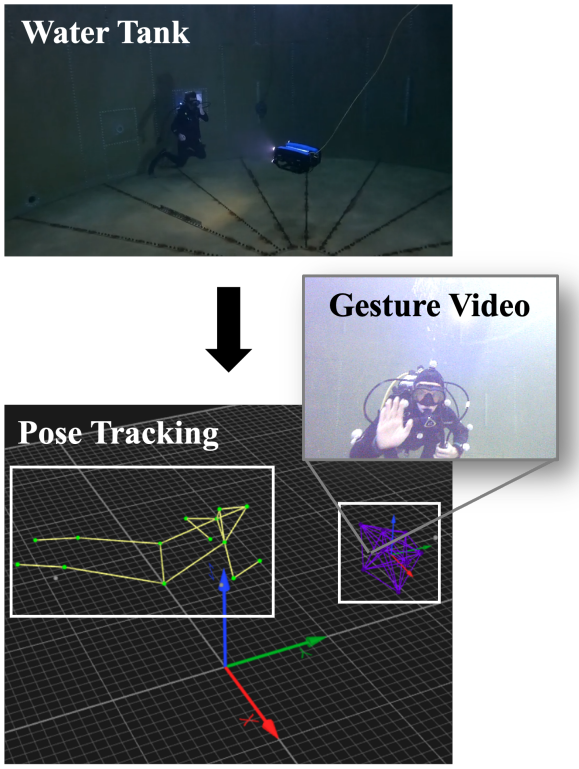

Fig. 1: The presented dataset, Scuba Gesture Dataset (SGD11), includes human divers' scuba gesture videos, human divers' poses, and the robot's pose data collected synchronously in an indoor water tank, located at the University of Florida

By prioritizing video sequences over static imagery, our work, Scuba Gesture Dataset (SGD11), seeks to provide data to address the limitations of static gesture detection. Dynamic video input has been shown to improve action understanding by enabling models to learn temporal patterns and gesture transitions [6], [7]. This shift from static image recognition to video-based analysis addresses a critical gap, enabling models to better understand gesture transitions and improve real-time underwater communication.

Our SGD11 dataset builds on and extends existing methodologies by introducing a dynamic video-based dataset. It aims to complement and advance research in autonomous diver assistance and broader UHRI applications, offering a new benchmark for evaluating gesture recognition models in underwater environments. In this paper, our contributions

[1]Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611, USA. {eherrin, zhangzihao, jane.shin}@ufl.edu

[2]Maryland Robotics Center, University of Maryland, College Park, MD 20742, USA. {xlin01, akondamu}@umd.edu

[3]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. xlin52@jhu.edu

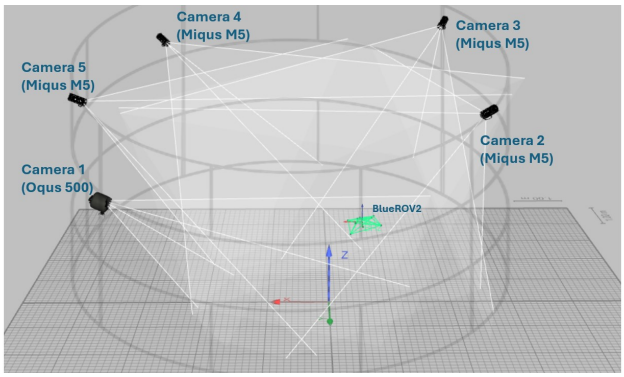

Fig. 2: Camera layout for the PERC water tank Qualisys motion capture system. One Oqus camera is positioned on the lower ring of the tank, while four Miqus cameras are evenly mounted around the upper ring. The field of view for each camera is indicated by white lines.

include (1) action dataset for both static and dynamic underwater scuba gestures; (2) preliminary benchmark results of action recognition models on different gesture types; (3) discussion on challenges in applying the dataset to the autonomy system development.

II. Dataset Collection

The dataset collection has been conducted in a controlled underwater environment at the University of Florida, using a 15-foot deep, 26-foot diameter (62,282-gallon) indoor water tank, as shown in Figure 1. This facility provides a stable and enclosed setting to ensure the consistency and reliability of the captured data. This water tank is equipped with five underwater motion capture cameras from Qualisys, which can provide the ground truth of human pose. The camera layout is depicted in 2. Four of the cameras are Miqus M5s, which have a frame rate of 180 fps, 4MP display, and resolution of 2048 × 2048 in normal mode. The Oqus 500 camera is an older model with the same specifications. The Qualisys QTM software is backwards compatible and is capable of syncing both old and new cameras. Unlike IMU-based pose estimates, Qualisys can provide up to precision on the order of millimeters for diver pose.

SGD11 captures temporal information by providing 880 annotated monocular video sequences of 11 commonly used SCUBA gestures (ranging from 75 to 85 clips per class). SGD11 further provides a subset of these sequences synchronized with diver body pose measurements recorded via the underwater motion capture system, with the diver in both the horizontal and vertical orientation (lower in Figure 1). The recordings in the subset are the same duration as the previous recordings in SDG11 (2-second clips). Motion capture data is provided as JSON files, including the available rigid bodies of both the diver and BlueROV2. Rigid bodies are taken as the center of the set of available marker measurements. All consistently tracked markers for each gesture example are included and labeled according to their location on the diver pose (e.g., HEAD, KNEE, ...). The dataset consists of

TABLE I: Standardized set of diver hand gestures used in SDG11. Each gesture is visually represented with its corresponding label.

eleven carefully selected gestures that are commonly used in the scuba community, chosen for their relevance in robotic collaboration scenarios. These gestures, which are detailed in Table I, provide a foundation for training and evaluating underwater action recognition systems.

All videos in SDG11 are taken from (*i*) a Raspberry Pi Camera mounted inside the BlueROV2 and (*ii*) a GoPro video camera. These recordings consist of short underwater hand gesture videos, captured under controlled conditions to provide high-quality training data for action recognition models. All videos in the dataset are collected from a viewpoint facing the diver, reflecting the assumption that the robot will reorient itself for optimal gesture detection. The subjects were two trained scientific divers. The human pose data is collected using a set of custom underwater markers manufactured by APRILab, mounted on a 5mm wetsuit with Velcro as in Figure 3.

Data collection was conducted through a custom ROS2 pipeline that streamed video from the BlueROV2's onboard camera over a ROS topic, while simultaneously controlling the vehicle's lights to communicate with the diver. Short light flashes indicated transitions between gestures within a set, and a long flash signaled a change in gesture class. This allowed the diver to follow a predefined sequence autonomously, without surface communication. A second researcher monitored the live video stream on the topside computer to ensure video quality and gesture consistency during the recording session.

The data collection process for SGD11 took place over the course of one month, during three separate underwater diving sessions conducted in the University of Florida water tank.

Fig. 3: Diver equipped with a 19-marker custom set (30mm and 40mm diameter markers) for use with the Qualisys motion capture system. Example of a marker depicted in the bottom right of the Figure.

The first two sessions, which account for the majority of the dataset, focused on collecting monocular video clips of all 11 gestures using our custom data collection code and lasted approximately 30 minutes each. The third session, dedicated to recording the synchronized motion capture data using the Qualisys system, required a more involved setup and lasted approximately 90 minutes. During this session, each gesture was performed by the divers in both horizontal and vertical orientations to capture a wider range of realistic diver poses for downstream analysis and model evaluation.

All gesture annotations were performed by members of the research team, including the two scientific divers who recorded the dataset and are familiar with standard scuba hand signals. As trained AAUS-certified scientific divers, they possess domain knowledge relevant to underwater gesture recognition. This ensured high accuracy and consistency in labeling across the dataset. The scientific divers also trained several undergraduate researchers on how to properly annotate diver pose for the section of the dataset containing synchronized camera and diver pose data. Annotations were reviewed by the trained undergraduate students, in addition to our collaborators at Cornell.

III. Preliminary Benchmark for Underwater Scuba Gesture Recognition

Since this is the first video-based diver gesture recognition dataset, we present a preliminary benchmark test for three different gesture recognition models: a Multi-Layer Perceptron (MLP), a Long Short-Term Memory (LSTM) network, and a Spatial-Temporal Transformer (ST-TR). Each model is assessed using keypoints extracted by the MediaPipe Hands framework [8], focusing on the recognition of both static and dynamic gestures within the SGD11 dataset. While recent methods in gesture recognition often incorporate CNN-based feature extraction (e.g., I3D, C3D), our preliminary benchmark focuses on evaluating baseline models using pose keypoints rather than raw video frames. As such, we do not include CNN architectures in this initial benchmark. Future work will incorporate state-of-the-art video-based models, including spatiotemporal CNNs and hybrid CNN-Transformer frameworks, for direct comparison.

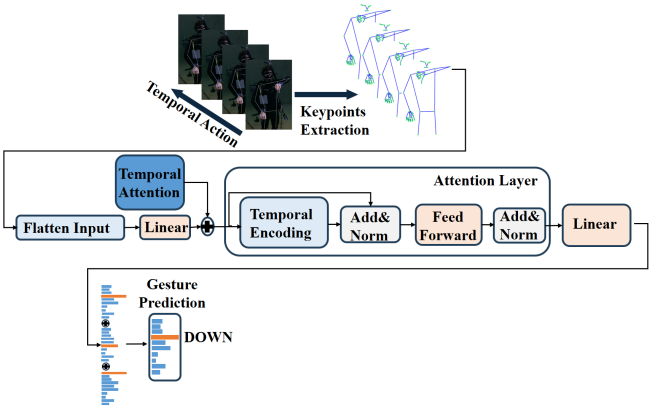

Fig. 4: ST-TR architecture for the dynamic gesture recognition

For our ST-TR model in Figure 4, the model first projects input features into an embedding space using a linear layer, followed by the addition of positional encodings to retain temporal ordering. A single-layer Transformer-encoder, with multi-head self-attention and feedforward sub-networks, captures spatial-temporal dependencies across the gesture sequence. After encoding, global average pooling is applied across the sequence dimension to aggregate the learned features, which are then passed through a dropout layer for

TABLE II: Comparison of precision, recall, and F1-score for dynamic gestures across three gesture recognition models. All tested using keypoints from MediaPipe Hands [8]. Best scores in bold.

Model	**Buddy Up**			**Follow Me**			**Left**		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
MLP	96.3%	–	–	**97.8%**	–	–	52.0%	–	–
LSTM	58.3%	25.0%	30.0%	8.3%	33.0%	24.0%	0.0%	0.0%	0.0%
ST-TR	**100.0%**	**83.0%**	**91.0%**	91.7%	**92.0%**	**96.0%**	**58.3%**	**17.0%**	**25.0%**

Model	**Right**			**You**		
	Prec.	Rec.	F1	Prec.	Rec.	F1
MLP	0.0%	–	–	31.0%	–	–
LSTM	33.3%	**92.0%**	27.0%	8.3%	0.0%	0.0%
ST-TR	**83.3%**	92.0%	**29.0%**	**66.7%**	**33.0%**	**38.0%**

regularization before being classified via a fully connected output layer.

The ST-TR model was trained on a dataset consisting of $C = 11$ action classes, each with 60 videos spanning 2 seconds at 30 FPS, resulting in 660 video samples. Each video provides a sequence of $T = 60$ frames, representing temporal joint configurations over time.

For dynamic gestures, the gesture recognition performance is compared using the precision scores in Table II. There are two dynamic gestures considered: *Buddy Up*, *Follow Me* and three static gestures: *Left*, *Right*, and *You*. The ST-TR model demonstrated better performance, particularly for the dynamic gestures. It achieved a perfect 100% precision on the *Buddy Up* gesture, outperforming the MLP model by 3.7%. Moreover, the ST-TR model consistently outperformed baseline models for dynamic gestures, except for *Follow Me*, where the MLP model had an advantage.

For static gestures such as *Ascend* and *Descend*, the MLP model achieved 100% accuracy, while the ST-TR model showed reduced performance, achieving only 75% and 33%, respectively. This suggests that frame-level classifiers like MLP are particularly effective for recognizing minimal motion gestures, whereas the ST-TR model's strength lies in capturing temporal dependencies critical for dynamic gestures.

IV. Discussion

We release this dataset to facilitate the development of gesture recognition models for diver-robot collaboration, such as in cave diving scenarios where robots may understand divers' intentions and assist divers by responding to predefined gesture commands. In addition to gesture labels, SDG11 provides synchronized diver pose annotations, enabling the training of image-based pose estimation models that are robust to varied diver orientations, including horizontal poses commonly observed in underwater settings. Furthermore, by including both camera and diver position data, the dataset supports research on monocular depth estimation methods adapted to the unique challenges of the underwater domain.

Beyond dynamic gesture recognition, SDG11 also supports static gesture analysis and image-based classification. Keyframes extracted from gesture sequences naturally capture intra-class variation across different executions of the same gesture. These static frames can be used to augment training data or adapt existing recognition models for underwater deployment. In this way, SDG11 serves as a bridge between traditional image-based classification methods and emerging approaches to video-based understanding in underwater human-robot interaction.

Looking ahead, we plan to expand SDG11 in both scale and scope to better support the development of robust underwater gesture recognition systems. The current release includes a limited set of annotated video clips, with only a subset containing synchronized motion capture data. Increasing the volume of recorded gestures, particularly those involving complex or subtle motions, such as *Buddy Up*, will improve model generalization across real-world conditions. Additionally, enhancing the fidelity of motion capture and marker tracking will increase the accuracy of pose annotations, which are essential for training and evaluating pose-aware models.

To broaden its applicability, SDG11 can also be integrated with other underwater datasets such as CADDY. While SDG11 primarily targets dynamic gesture recognition, individual keyframes can be extracted to support static gesture analysis. In this context, CADDY's stereo image pairs could be paired with SDG11-derived frames to augment training data for static recognition models. This combined approach leverages complementary strengths – SDG11's temporal annotations and CADDY's spatial cues – to enable more comprehensive evaluation and training pipelines across diverse underwater environments.

A. Limitations

The current dataset has several limitations that users should consider when applying it to underwater research:

1) The videos recorded using the onboard camera may exhibit reduced quality. This issue is due to the resolution being intentionally downsized to facilitate real-time video transmission. Consequently, users may encounter limitations when conducting tasks that require high-resolution input, such as extracting keypoints from the diver's body.
2) While the CADDY dataset provides approximately 10,000 stereo image pairs across eight different scenarios, the presented SDG11 dataset focuses on capturing both static and dynamic gestures within an indoor underwater tank. As a result, SDG11 may

lack generalizability for tasks intended for open-water environments.

3) For the marker-based dataset collected using the Qualisys system, consistent tracking is not guaranteed. This inconsistency arises from the insufficient number of markers placed on both the diver's body and the BlueROV2, limiting the accuracy and continuity of motion capture.

In this initial release, we do not evaluate model run-time performance, robustness to input noise, or perform cross-validation experiments. Our benchmark focuses on providing baseline comparisons for dynamic and static gesture classification using keypoint sequences. Future work will incorporate evaluation of inference speed, noise robustness, and more comprehensive testing across recording conditions and subjects.

V. Availability

The dataset and benchmarking code are publicly available at `https://github.com/aprilab-uf/SDG11`. The release includes monocular video clips for 11 scuba gestures, synchronized motion capture data (in JSON format) for a subset of sequences, and baseline training and evaluation code for three gesture recognition models (MLP, LSTM, and ST-TR). The dataset is intended to support research on training, benchmarking, and analysis of action recognition systems for robust underwater human-robot interaction.

VI. Conclusion and Future Work

SDG11 is the first publicly available dataset focused on dynamic underwater diver gesture recognition, providing a valuable benchmark for advancing robust AI models in support of human-machine teaming in underwater environments. By offering synchronized video and pose data, along with baseline models, SDG11 enables research in gesture recognition, pose estimation, and perception-driven interaction. Future work will focus on expanding the dataset with a broader range of gestures, enhancing the fidelity of pose annotations, and integrating the dataset into vehicle autonomy pipelines. Additionally, exploring domain adaptation techniques, such as transfer learning and fine-tuning, may improve model generalization across diverse underwater conditions. By addressing the unique challenges of underwater perception, SDG11 aims to support the development of more capable and autonomous diver-assistance systems.

ACKNOWLEDGMENT

The authors would like to thank Quang Pham, Eli Heskin, Aditya Penumarti, Andres Pulido, Zi Long He, Jia Guo, Daniele Fragiacomo, and Silvia Ferrari for their valuable contributions to the ideation, creation, and testing of our dataset.

References

[1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[2] M. A. Martija, J. I. Dumbrique, and P. Naval, "Underwater Gesture Recognition Using Classical Computer Vision and Deep Learning Techniques," *Mathematics Faculty Publications*, Mar. 2020.

[3] K. de Langis, M. Fulton, and J. Sattar, "Video diver dataset (vdd-c): 100,000 annotated images of divers underwater," April 2021, [Dataset]. [Online]. Available: https://example.com/vdd-c-dataset

[4] A. G. Chavez, A. Ranieri, D. Chiarella, E. Zereik, A. Babić, and A. Birk, "CADDY Underwater Stereo-Vision Dataset for Human–Robot Interaction (HRI) in the Context of Diver Activities," *Journal of Marine Science and Engineering*, vol. 7, no. 1, p. 16, 2019. [Online]. Available: https://doi.org/10.3390/jmse7010016

[5] G. M. Goodfellow, J. A. Neasham, I. Rendulic, D. Nad, and N. Miskovic, "Divernet — a network of inertial sensors for real time diver visualization," in *2015 IEEE Sensors Applications Symposium (SAS)*, 2015, pp. 1–6.

[6] X. Chen, G. Wang, H. Guo, C. Zhang, H. Wang, and L. Zhang, "Mfa-net: Motion feature augmented network for dynamic hand gesture recognition from skeletal data," *Sensors*, vol. 19, no. 2, p. 239, 2019.

[7] Y. Li, Z. He, X. Ye, Z. He, and K. Han, "Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition," *EURASIP Journal on Image and Video Processing*, vol. 2019, pp. 1–7, 2019.

[8] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.