2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Semantically-Aware Diver Activity Recognition Framework for Effective Underwater Multi-Human-Robot Collaboration

Sadman Sakib Enan[1] and Junaed Sattar[2]

***Abstract*— Effective multi-human-robot collaboration is essential for expanding human-led operations in the challenging and high-risk underwater environment. For autonomous underwater vehicles (AUVs) to become true teammates, they must be able to comprehend their surroundings and recognize a diver's activities to offer assistance and ensure safety. Towards this goal, we introduce DAR-Net, a novel transformer-based framework that analyzes complex underwater scenes to classify diver activities. Our contribution lies in a semantically guided learning formulation that couples transformer-based temporal reasoning with pixel-level scene supervision. This multi-loss training strategy explicitly aligns global activity recognition with local human–robot interaction semantics, which is particularly critical in low-visibility underwater conditions. To address the significant challenge of data scarcity in this domain, we present the first-ever Underwater Diver Activity (UDA) dataset, a foundational resource containing over $2,600$ annotated images with pixel-level masks. Through rigorous experimental evaluations in a controlled environment, we demonstrate that DAR-Net achieves promising accuracy in recognizing six distinct diver activities, outperforming state-of-the-art models. While this dataset provides a crucial baseline, our work serves as a pioneering step, laying the groundwork for future research and facilitating the development of more intelligent, collaborative underwater robotic systems.**

I. Introduction

The use of Autonomous Underwater Vehicles (AUVs) has seen significant expansion across a broad range of tasks, such as environmental monitoring [1], mapping [2], submarine cable and wreckage inspection [3], and search-and-rescue operations [4]. This growth is primarily attributed to advancements in on-board computational power and enhanced Underwater Human-Robot Interaction (UHRI) capabilities, which enable AUVs to interact effectively with human divers without assistance from a topside operator (*e.g.*, [5]). Recent research [6] allows AUVs to determine diver attentiveness and plan trajectories for interaction if needed, making them more powerful for collaborative missions with human divers. For this collaboration to be truly effective, an AUV must be more than just a tool; it needs to be an intelligent partner capable of understanding its surroundings and its human teammates (*i.e.*, *dive buddies*). A critical component of this is the ability to recognize a diver's current activity and the significance of that activity. For instance, during a sensitive task like a rescue mission, the AUV must not disrupt the diver's focus, allowing them to concentrate fully on their responsibilities without unnecessary interference. Therefore, equipping an AUV with the capability to recognize a diver's activity is not merely an enhancement–it is a necessity for informed decision-making and safe, efficient collaboration.

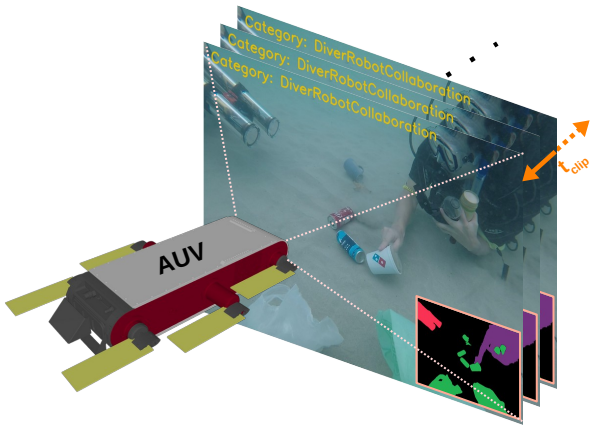


Fig. 1. Demonstration of the proposed diver activity recognition framework, making inference on real-world underwater scenarios involving multiple divers and robots. The proposed method is able to learn highly discriminative spatio-temporal features from underwater video clips by focusing on important elements within the scene (as shown in the inset), such as divers, robots, and objects of interest, and their interactions with each other.

While significant progress has been made in terrestrial activity recognition (*e.g.*, [7], [8]), the underwater environment presents unique and formidable challenges. Data collection is difficult due to high pressure, low visibility, extreme temperatures, and unpredictable currents [9], leading to a near-complete absence of large-scale underwater diver activity datasets, hampering the design of deep-learned recognition methods.

To address this critical gap, we introduce the first-ever *Underwater Diver Activity (UDA)* dataset, which contains over 2600 semantically segmented images with annotations for divers, robots, and objects of interest in a multi-human-robot collaborative setting. We also propose a novel, end-to-end framework named *DAR-Net* (*Diver Activity Recognition Network*) to analyze and understand diver activities. This transformer-based architecture is designed to extract highly discriminative spatio-temporal features from underwater scenes. Crucially, our approach uses supervision from scene semantics to train the model to focus on the most important elements – the divers, robots, and objects – instead of irrelevant background noise (*e.g.*, [10]). This enables the AUV to take proactive, informed actions, such as joining a team to remove trash after recognizing their collaborative activity (see Fig. 1). Through comprehensive experimental

*This work was supported in part by the National Science Foundation Grant IIS-#2220956. The authors were with the department of Computer Science & Engineering and the Minnesota Robotics Institute, University of Minnesota, MN, USA at the time this work was conducted. Email: [1]sadmansakib.enan@gmail.com, [2]junaed@umn.edu

evaluations, we demonstrate that our framework achieves a promising accuracy and outperforms existing state-of-the-art activity recognition models.

We make the following contributions in this paper:

1) We propose an end-to-end transformer-based deep network called DAR-Net to analyze and classify different diver activities in underwater multi-human-robot collaborative scenes. The training pipeline includes a multi-loss objective function that prioritizes important regions to learn from, instead of learning from the whole image.
2) Additionally, we present UDA, the first-ever Underwater Diver Activity dataset that includes 2640 annotated underwater multi-human-robot collaborative scenes divided into 6 diver activity categories. The data were collected from several closed-water robot trials and contain pixel-level annotations for divers, robots, and objects of interest.
3) Furthermore, we conduct both quantitative and qualitative experiments, demonstrating that the proposed framework derives significant benefits from incorporating additional supervision from semantic labels. This enables the model to learn highly discriminative spatio-temporal features essential for recognizing different diver activities.

II. Related Work

While significant advancements have been made in autonomous robotics for various underwater tasks, the field of UHRI remains an emerging and challenging domain. A critical bottleneck to realizing truly effective collaboration is the lack of a robust system for diver activity recognition. While research in terrestrial activity recognition is extensive, the unique challenges of the underwater environment – such as low visibility, unpredictable lighting, and the complex interactions of divers with their surroundings – have left this area largely unexplored.

Human Activity Recognition (HAR) is an active research area within computer vision and robotics, with research efforts extending over several decades [11], [12], [13]. Sensor-based HAR has been a cornerstone in activity recognition research, owing to its ubiquity and ease of data collection [14]. Early approaches (*e.g.*, [15]) utilized wearable sensors and traditional machine learning models like Hidden Markov Models (HMMs) and Support Vector Machines (SVMs). Advancements in deep learning have led to more sophisticated models (*e.g.*, [16]), such as Convolutional Neural Networks (CNNs), that learn features directly from raw sensor data.

Vision-based HAR (*e.g.*, [17]) has also gained significant traction due to the proliferation of cameras. Early work focused on handcrafted features (*e.g.*, [18], [19]), while more recent deep learning techniques (*e.g.*, [20]) have revolutionized the field. Researchers have explored architectures like two-stream CNNs that process both spatial and temporal information from video frames, achieving state-of-the-art (SOTA) performance in action recognition tasks [21], [22]. Early approaches typically involved handcrafted feature extraction from video sequences [23], [24], followed by classification using methods, such as SVMs or decision trees. With the advent of deep learning, however, researchers have developed end-to-end architectures for video-based HAR (*e.g.*, [25], [26]) Hara *et al.* [27] have also proposed to use 3D kernels in CNNs to extract spatio-temporal features from videos for activity recognition. Although 3D CNNs are typically susceptible to overfitting due to their large number of parameters, the use of large-scale activity datasets (*e.g.*, UCF-101 [28], Sports-1M [29], ActivityNet [30], Kinetics [31]) for training has mitigated the issue. In contrast, Wu *et al.* [32] have showed that augmenting 3D CNNs with a long-term feature bank can yield SOTA results in activity recognition task. Furthermore, advancements in Large Language Models (LLMs) have helped create robust HAR methods [33], [34].

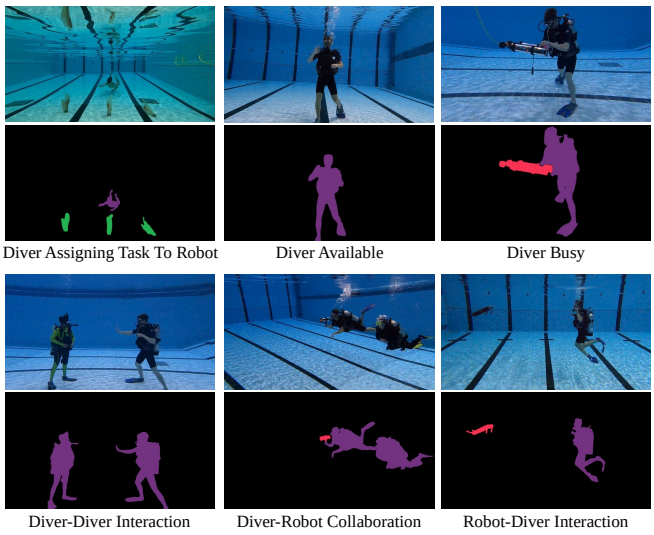


Fig. 2. A few sample images and their semantic labels from the proposed UDA dataset. In our dataset, divers, robots, and objects of interest are annotated with purple, red, and green colors, respectively.

On the contrary, HAR techniques for underwater domain have received considerably less attention. While there have been a few research endeavors addressing issues, such as diver motion prediction in the context of UHRI [35], [36], [37], non-human motion prediction [38], [39], and monitoring divers to ensure safety [40], [41], none of these specifically tackle the problem of diver activity recognition. This is primarily due to the lack of large-scale human activity recognition datasets tailored specifically for underwater environments. There are a few diver dataset available in the literature (*e.g.*, [42]), however, they do not include data of divers actively engaged in different activities or tasks, specifically collaborating with AUVs. To this end, we propose to formulate the first-ever UDA dataset involving multiple divers and AUVs, so that it can be used to supervise the learning and validation of diver activity recognition frameworks.

III. UDA Dataset

The lack of publicly available, large-scale datasets for underwater human-robot activity has long been a major

impediment to research in this field. To address this critical gap, we have curated the UDA dataset. This dataset is the first of its kind, meticulously compiled from real-world human-robot collaborative trials conducted in closed-water environments and specifically designed to support the development of diver activity recognition frameworks. We collect the data as video clips having resolutions of 1920×1080 pixels, using GoPros [43]. All data are captured in a closed-water pool environment with controlled lighting and visibility. The scenes include between one and three divers, one to two robots, and one to three task-specific objects. Each clip is approximately three seconds long and captures diver poses arising naturally from real task execution. These categories were chosen based on our extensive research and engagement with aquatic professionals to represent key interactions between multiple divers and robots. We chose these six categories as representative activities to demonstrate the efficacy of our proposed framework for the classification task. The activities are defined as follows:

1) *Diver Assigning Task to Robot:* A diver gives instructions to a robot, often without the robot being in the scene.
2) *Diver Available:* A diver is present but not actively engaged, indicating an opportunity for interaction.
3) *Diver Busy:* A diver is occupied with a task and is not available for interaction, even if a robot is in the scene.
4) *Diver-Diver Interaction:* Two divers are communicating with each other.
5) *Diver-Robot Collaboration:* Divers and robots are actively working together on a shared task.
6) *Robot-Diver Interaction:* A diver and robot are engaged in non-verbal communication, separate from active task execution.

Each image in the dataset has been carefully annotated at the pixel level to capture the scene's semantics. Using the Segmentation Anything Model (SAM) [44] as a base, we created detailed segmentation masks for *divers*, *robots*, and *objects of interest*, which were manually verified and corrected to ensure accurate boundaries. These pixel-level annotations are crucial, as they allow our framework to learn from and focus on the most relevant elements within a scene.

Unlike existing underwater datasets that primarily capture isolated diver presence or motion, UDA focuses on task-driven, multi-human–robot interactions. Each activity category includes visually diverse instances with varying numbers of divers, robot proximity, interaction patterns, and object configurations, providing meaningful intra-class variability for activity recognition.

The UDA dataset is a cornerstone of our research, enabling the training of robust deep-learning models for a task previously considered infeasible. We make this dataset publicly available at `https://irvlab.cs.umn.edu/uda`, to accelerate future research in the critical domain of underwater human-robot collaboration.

IV. Diver Action Recognition Framework

Our proposed DAR-Net (Diver Activity Recognition Network) is an end-to-end framework designed to process underwater video clips and extract robust spatio-temporal features to recognize various diver activities. The core of our approach lies in a unique multi-loss objective function that minimizes both a classification loss and a segmentation loss, enabling the model to learn from both global and local context simultaneously.

A. Feature Extraction

To effectively extract rich, discriminative features from the underwater environment, we use a ResNeXt-101 [45] network as the backbone of our model. This architecture is renowned for its high modularity and ability to capture complex features, and has shown good performance in underwater applications (*e.g.*, [39]). ResNeXt employs a "*split-transform-aggregate*" strategy, which creates a homogeneous, multi-branch structure that is highly efficient for learning. By increasing the network's cardinality (the size of its transformations), we achieve a significant improvement in performance over traditional approaches that simply increase depth or width.

Following recent advancements in video action recognition [33], [34], we further enhance our feature representation by incorporating both positional and classification encodings. These encodings, initialized with a normal distribution $\mathcal{N}(0, 0.02^2)$, ensure that each feature location retains its positional information, which is crucial for accurately classifying activities. This enriched feature representation is then passed to two distinct branches for further processing (see Fig. 3): the *Classification Branch* and the *Segmentation Branch*, which we describe below.

B. Objective Function Formulation

The core of DAR-Net's learning process is our multi-loss objective function, which allows the model to analyze and determine diver activities in a robust manner. We design two separate processing branches for the spatio-temporal features:

1) *The Classification Branch*: Based on a transformer [33] architecture, this branch is designed to focus on the most important temporal regions of the features, effectively capturing the global context of the diver's actions. The final output of this branch is supervised by the class labels and trained with a standard cross-entropy loss function.
2) *The Segmentation Branch*: This branch, built on an encoder-decoder architecture, is used to leverage spatial scene semantics, thereby capturing the local context. This branch is supervised by the semantic labels and trained with a binary cross-entropy loss function.

This hybrid learning strategy, guided by both global classification and local semantic information, facilitates the learning of robust features for the classification task. Given x_{n,y_n} as the unnormalized logit for the diver activity category y_n, where n refers to the n-th sample from the minibatch,

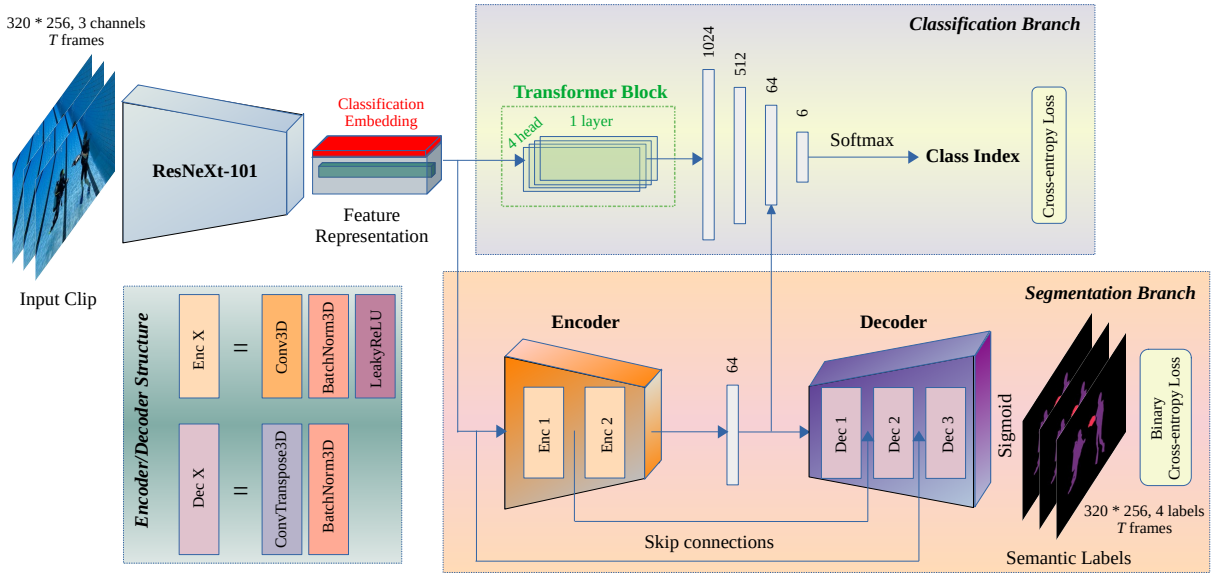
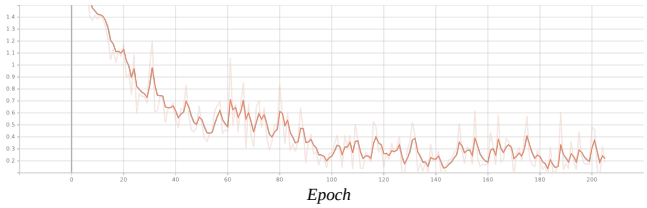
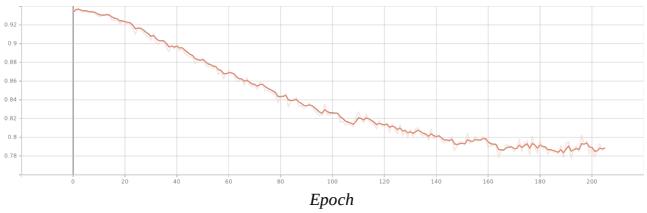


Fig. 3. An overview of the network architecture of DAR-Net. It takes an underwater diver activity video clip as input and extracts highly discriminative spatio-temporal features by incorporating additional supervision from scene semantics. Intermediary skip connections are used to avoid overfitting.



(a) Classification Loss.



(b) Segmentation Loss.

Fig. 4. The training performance of DAR-Net. Note the convergence in both the classification and segmentation validation loss graphs.

we define the classification cross-entropy loss $\mathcal{L}_{class}$ and the segmentation binary cross-entropy loss $\mathcal{L}_{seg}$ functions as follows.

$$\mathcal{L}_{class} = -\sum_{n=1}^{N} \log \frac{\exp(x_{n,y_n})}{\sum_{i=1}^{\tau} \exp(x_{n,i})} y_n$$

$$\mathcal{L}_{seg} = -\sum_{n=1}^{N} [y_n \log x_{n,y_n} + (1 - y_n) \log(1 - x_{n,y_n})]$$

where N is the batch size, and $\tau = 6$ is the number of activity categories.

The overall training is performed by minimizing a combined multi-loss objective function:

$$\mathcal{L} = \alpha \mathcal{L}_{class} + \beta \mathcal{L}_{seg} \quad (1)$$

The weights α and β are set as trainable parameters, allowing the model to dynamically adjust the importance of each loss during training.

C. Implementation Details

We implemented our framework using the PyTorch library [46]. Although the transformer backbone follows from our prior work [39], the primary contribution of this work is the integrated injection and joint optimization of semantic supervision with activity classification. For the segmentation branch, we employed an encoder-decoder architecture to learn from the scene semantics and fed the encoded information into the classification branch. Intermediary skip connections were used within the segmentation branch to mitigate the vanishing gradient problem, further enhancing the model's performance and stability. The individual elements of each encoder and decoder block is shown in Fig. 3. The α and β values in the multi-loss objective function (Eq. 1) were set as trainable parameters. We employed various data augmentation techniques [47], including random cropping, image distortion, and flipping, to increase the robustness of the model. The model was trained on the UDA dataset using an $80/20$ split for training and validation, respectively, for 200 epochs on an Nvidia RTX6000 Ada Generation GPU; validation losses were observed to converge during training (Fig. 4). We used a batch size of 4, a learning rate of 10^{-5}, and the ADAMW optimizer [48] with a momentum of 0.9. The input data were resized to a spatial resolution of 320×256 pixels and processed in 64-frame video chunks, which represent approximately 3 seconds of video.

V. Experimental Evaluations

This section outlines the process for evaluating the performance of the proposed DAR-Net framework against several SOTA models for diver activity recognition.

TABLE I

DIVER ACTIVITY RECOGNITION PERFORMANCE* ON THE TEST SET. PRECISION, RECALL, AND F1-SCORE ARE COMPUTED AS WEIGHTED AVERAGES. THE VALUES ARE IN PERCENTAGES.

Method	Accuracy	Precision	Recall	F1-Score
3DResNet [27]	53.33	59.84	53.33	53.31
R(2+1)D [49]	60.00	70.63	60.00	56.83
SlowFast [22]	56.67	65.00	56.67	57.67
LateTemporal [34]	66.67	71.25	66.67	65.60
RRCommNet [39]	60.00	67.64	60.00	61.05
Ours	**73.33**	**76.90**	**73.33**	**72.17**

*The performance is evaluated based on the classification accuracy.

A. Evaluation Process and Metrics

To ensure a fair and rigorous evaluation, we created a test set of 30 video clips that were distinct from the training and validation data. This test set includes five video clips for each of the six diver activity categories. To maintain fairness, all SOTA baselines were fully retrained on the UDA dataset using their recommended hyperparameters.

Each video clip in the test set, representing approximately 3 seconds of footage, was processed in chunks of 64 frames. These clips, in a four-dimensional RGB tensor format $(64, 3, h_{im}, w_{im})$ where h_{im} and w_{im} are height and width of the image, respectively, were fed into the trained model to produce classification scores, $\mathbf{x}_{pred} = [x^1_{pred}, \ldots, x^6_{pred}]^T$. A softmax function was then applied to convert these scores into probability scores:

$$P(x^i_{pred}) = \frac{\exp(x^i_{pred})}{\sum_{j=1}^{\tau} \exp(x^j_{pred})}$$

where i refers to the i-th category index and can have values in the range $[1, \tau]$. Finally, the predicted category is found by selecting the index with the maximum probability score.

The performance of each model was measured using four key metrics:

1) *Accuracy:* The overall correctness of the model's predictions, calculated as $\frac{TP+TN}{TP+TN+FP+FN}$. Here, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.
2) *Precision:* Measures the model's ability to avoid false positive predictions, calculated as $\frac{TP}{TP+FP}$. This is particularly useful when the cost of false positives is high.
3) *Recall:* Measures the model's ability to correctly identify all positive samples, calculated as $\frac{TP}{TP+FN}$. Recall is particularly useful when the cost of false negatives is high.
4) *F1-Score:* Combines precision and recall into a single metric, providing an overall measure of effectiveness for the classification task. It is calculated as $\frac{2 \times Precision \times Recall}{Precision + Recall}$. This metric can help assess the model's overall effectiveness in the classification task.

B. Results

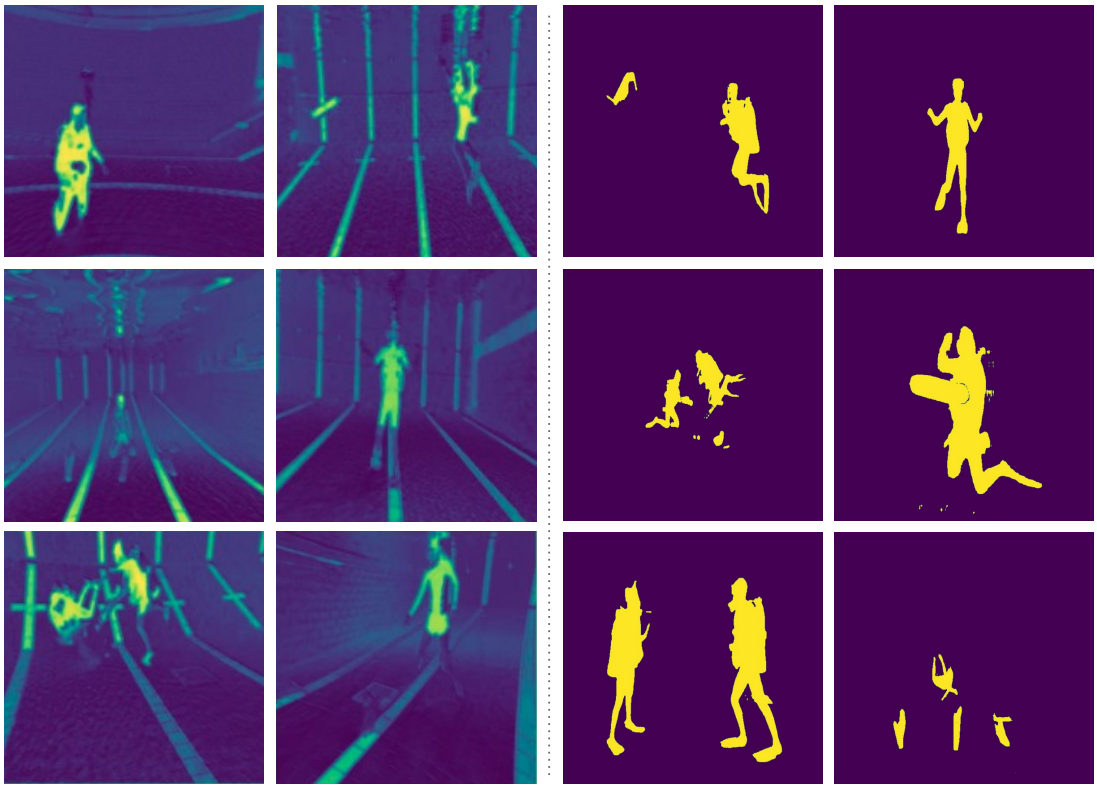
Our experimental results demonstrate that DAR-Net consistently outperforms the SOTA models. As shown in Table I, our framework achieves a classification accuracy of 73.33%, which is notably higher than the other models tested. This is significant, as activity recognition from video is a complex task, and the SOTA models struggled to accurately classify diver activities. The Late Temporal model was the only one that achieves a comparable classification accuracy of 66.67%, likely due to its use of transformer blocks and the self-attention mechanism [50]. We note that some SOTA baselines, originally designed for larger datasets, may not fully converge under limited data availability, further underscoring the benefit of semantically guided supervision in data-scarce underwater applications.

The superior performance of DAR-Net is further validated by its high average precision (76.90%), recall, and F1-score. This indicates that our framework is highly effective at accurately classifying relevant activities while maintaining low rates of both false positives and false negatives.

Furthermore, we conduct an ablation study that isolates the effect of semantic supervision by comparing training with and without segmentation-based supervision under otherwise identical settings. In Fig. 5, we visualize attention maps from the activity recognition model, highlighting the image regions prioritized during training. Fig. 5a illustrates the attention maps obtained from the model trained without semantic supervision. From the figure, it is evident that the model focuses unnecessarily on image regions irrelevant to determining the underlying diver activity category, such as lane markings on the pool bottom and sides. This could contribute to the lower accuracy recorded when scene semantics were not considered during training. In contrast, upon integrating scene semantics into the training process of our proposed framework, the model directs attention solely to crucial image regions, such as divers, robots, and objects of interest, as shown in Fig. 5b. With DAR-Net's training supervised by semantic labels, the intermediary attention maps naturally prioritize the segmented regions, leading to superior accuracy in identifying diver activity categories. The attention maps in Fig. 5b (segmentation masks, in this case) are generated by applying binary thresholding to the output of the sigmoid function from the segmentation branch.

We also show a confusion matrix for the diver activity recognition task (see Fig. 6) by recording the predictions made by DAR-Net on our test set consisting of 30 video clips of underwater multi-human-robot collaborative scenarios. As illustrated in the figure, DAR-Net accurately classifies the majority of different diver activities. However, it encounters challenges in identifying certain activities, most notably *Diver Busy* and *Robot-Diver Interaction*. Upon closer examination of video clips from these categories, we observe that both often involve a single diver co-located with a robot and limited explicit gesturing, resulting in shared visual primitives. In several failure cases, the distinction depends on subtle temporal cues (*e.g.*, tool manipulation vs. communicative posture), which are weakly represented in short clips and further degraded by turbidity and lighting artifacts.

To further analyze this behavior, we compute the per-



(a) Without the supervision from semantic labels. (b) With the supervision from semantic labels.

Fig. 5. The effect of using semantic labels during the training of DAR-Net. The inclusion of scene semantics directs the model's focus towards relevant regions in the image for feature learning. In contrast, traditional activity recognition models frequently prioritize irrelevant areas, such as pool lane markings during training. Lighter colors indicate higher attention values. The attention maps on the right are generated by applying binary thresholding to the output of the sigmoid function from the segmentation branch.

category precision, recall, and F1-score and visualize the results in Fig. 7. A consistent trend emerges, showing lower performance for the aforementioned two categories compared to others. Specifically, for the *Diver Busy* category, the model struggles to detect positive samples, leading to low recall, while also producing false positives, resulting in reduced precision. In contrast, the *Robot-Diver Interaction* category is detected less consistently; however, when identified, the predictions are reliable, as reflected by high precision and low recall. For the remaining activity categories, DAR-Net demonstrates comparatively strong and balanced performance. This observed discrepancy highlights the need for further investigation into temporally disentangling these visually similar activities to reduce misclassification.

C. Limitations and Mitigation

While the proposed work serves as a foundational step toward enabling effective multi-human-robot collaboration in the challenging underwater environment, it is important to acknowledge the inherent limitations of this research. A key limitation of this study is the size and scope of the Underwater Diver Activity (UDA) dataset. While we have meticulously curated the first-of-its-kind dataset with over $2,600$ annotated images, this size is considered small for a transformer-based network, which typically benefits from larger and more diverse data volumes to demonstrate strong generalization capacity. Underwater data collection is a uniquely difficult process due to safety concerns, logistical complexities, and the challenge of obtaining appropriate Institutional Review Board (IRB) approvals for human trials. This makes creating a large-scale dataset, comparable to those in terrestrial robotics, an arduous task. Furthermore, our experiments were conducted in a closed-water environment. This focused scope, while necessary for a controlled study, may not fully represent the variability and unpredictability of real-world, open-water conditions. To address these limitations, future work will focus on three key areas:

1) *Dataset Expansion*: We will explore advanced data augmentation techniques and synthetic data generation to virtually expand the UDA dataset. This will improve the model's robustness and help it generalize to new scenarios beyond the scope of our initial data collection. We will also collaborate closely with our institutional ethics board to create protocols for approved open-water data collection trials to increase dataset diversity.
2) *Model Robustness*: We will conduct more extensive experiments to show the efficacy of our proposed method. A deeper analysis will be performed on existing categories (and any novel ones we introduce) to identify and decouple the visual cues that lead to confusions among them.

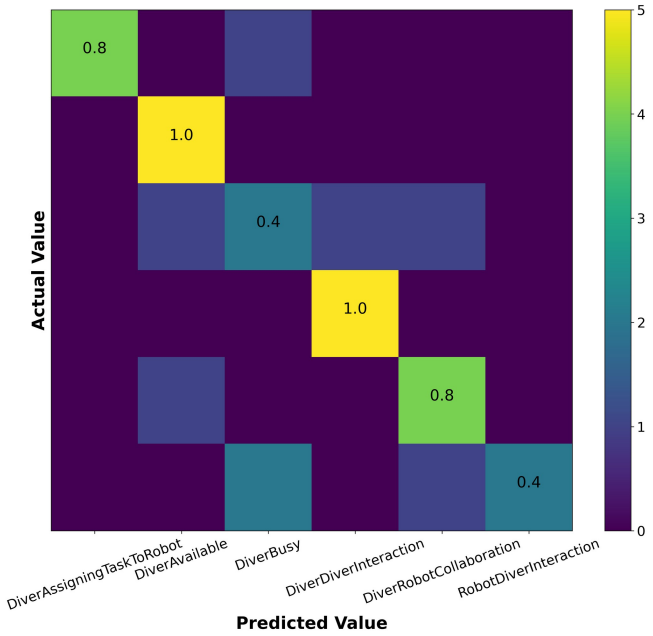


Fig. 6. Confusion matrix for diver activity recognition, computed on the 30 video clips from our test set. It highlights the robustness of the proposed framework in accurately identifying the majority of diver activity categories. The matrix entries are normalized.

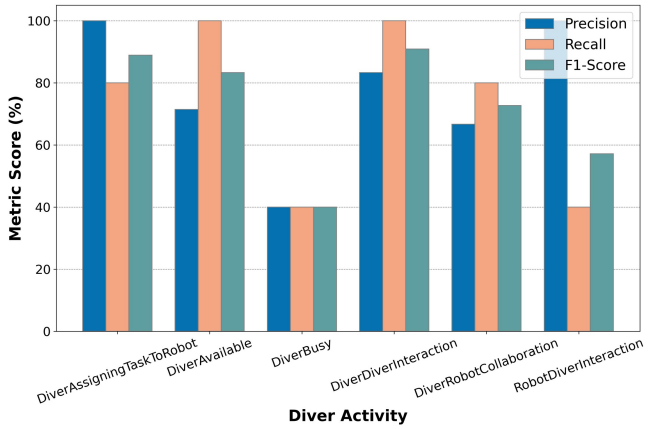


Fig. 7. Per-category Precision, Recall, and F1-Score for the proposed diver activity recognition framework. It indicates a similar trend where the model's performance is notably lower in predicting the *Diver Busy* and *Robot-Diver Interaction* categories.

3) *Community Collaboration*: By making the UDA dataset publicly available, we will invite the research community to contribute, expand, and use this resource to accelerate the development of robust and generalizable underwater perception systems.

VI. Conclusions

This work presents a significant stride in the field of underwater human-robot collaboration by introducing a novel, end-to-end framework for diver activity recognition. We have demonstrated that our model, DAR-Net, can accurately and robustly analyze diver activities from underwater scenes by learning from both global and local context. A key contribution is our proposed multi-loss objective function, which integrates supervision from scene semantics to ensure the network focuses on the most relevant elements – divers, robots, and objects of interest – and ignores irrelevant background noise. To facilitate research in this domain, we have curated and will make publicly available the first-ever Underwater Diver Activity (UDA) dataset. By providing over 2600 semantically segmented images of diverse underwater human-robot collaborative scenarios, this dataset serves as a crucial foundation for future research in this challenging domain. Through extensive experimental evaluations, DAR-Net has proven its effectiveness, consistently outperforming existing state-of-the-art models. While our framework shows promising results, particularly in its ability to focus on salient scene elements, future work will focus on expanding both the research scope and the UDA dataset, following the paths suggested in Sec. V-C. We believe this framework and dataset will pave the way for more intelligent, proactive, and safe autonomous underwater vehicles in the future.

References

[1] Y. Girdhar, N. McGuire, L. Cai, S. Jamieson, S. McCammon, B. Claus, J. E. S. Soucie, J. E. Todd, and T. A. Mooney, "CUREE: A Curious Underwater Robot for Ecosystem Exploration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 411–11 417.

[2] W. Wang, B. Joshi, N. Burgdorfer, K. Batsosc, A. Q. Lid, P. Mordohaia, and I. Rekleitisb, "Real-Time Dense 3D Mapping of Underwater Environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5184–5191.

[3] M. Lickliter-Mundon and K. B. Leverenz, "Monitoring Underwater Aircraft Sites in Lake Washington," in *Strides Towards Standard Methodologies in Aeronautical Archaeology*. Springer, 2023, pp. 211–237.

[4] J. Wu, C. Song, J. Ma, J. Wu, and G. Han, "Reinforcement Learning and Particle Swarm Optimization Supporting Real-Time Rescue Assignments for Multiple Autonomous Underwater Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6807–6820, 2022.

[5] M. J. Islam, M. Ho, and J. Sattar, "Understanding Human Motion and Gestures for Underwater Human-Robot Collaboration," *Journal of Field Robotics*, vol. 36, no. 5, pp. 851–873, 2019.

[6] S. S. Enan and J. Sattar, "Visual Detection of Diver Attentiveness for Underwater Human-Robot Interaction," *arXiv preprint arXiv:2209.14447*, 2022.

[7] K. Xia, J. Huang, and H. Wang, "LSTM-CNN Architecture for Human Activity Recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[8] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. M. Snoek, "Actor-Transformers for Group Activity Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 836–845.

[9] X. Wei, H. Guo, X. Wang, X. Wang, and M. Qiu, "Reliable Data Collection Techniques in Underwater Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 404–431, 2022.

[10] M. J. Islam, R. Wang, and J. Sattar, "SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robots," in *Robotics: Science and Systems (RSS)*, NY, USA, 2022.

[11] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[12] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A Review of Human Activity Recognition Methods," *Frontiers in Robotics and AI*, vol. 2, pp. 1–28, 2015.

[13] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A Review on Video-Based Human Activity Recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[14] F. Serpush, M. B. Menhaj, B. Masoumi, B. Karasfi *et al.*, "Wearable Sensor-Based Human Activity Recognition in the Smart Healthcare System," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[15] J. Wang, R. Chen, X. Sun, M. F. She, and Y. Wu, "Recognizing Human Daily Activities From Accelerometer Signal," *Procedia Engineering*, vol. 15, pp. 1780–1786, 2011.

[16] W. Jiang and Z. Yin, "Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks," in *23rd ACM international conference on Multimedia*, 2015, pp. 1307–1310.

[17] D. Girish, V. Singh, and A. Ralescu, "Understanding Action Recognition in Still Images," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1523–1529.

[18] C.-P. Huang, C.-H. Hsieh, K.-T. Lai, and W.-Y. Huang, "Human Action Recognition Using Histogram of Oriented Gradient of Motion History Image," in *2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, 2011, pp. 353–356.

[19] H. Su, J. Zou, and W. Wang, "Human Activity Recognition Based on Silhouette Analysis Using Local Binary Patterns," in *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2013, pp. 924–929.

[20] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, "Human Activity Recognition using Binary Motion Image and Deep Learning," *Procedia Computer Science*, vol. 58, pp. 178–185, 2015, second International Symposium on Computer Vision and the Internet (VisionNet'15). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915021614

[21] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Advances in Neural Information Processing Systems*, vol. 27, p. 568–576, 2014.

[22] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[23] N. Robertson and I. Reid, "A General Method for Human Activity Recognition in Video," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 232–248, 2006, special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731420600110X

[24] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video," in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2001, pp. 123–130.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[26] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[27] K. Hara, H. Kataoka, and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 3154–3160.

[28] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv preprint arXiv:1212.0402*, 2012.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.

[30] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics Human Action Video Dataset," 2017.

[32] C.-Y. Wu, C. Feichtenhofer, H. Fan, K. He, P. Krähenbühl, and R. Girshick, "Long-Term Feature Banks for Detailed Video Understanding," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 284–293.

[33] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video Action Transformer Network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 244–253.

[34] M. E. Kalfaoglu, S. Kalkan, and A. A. Alatan, "Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 731–747.

[35] H. Hu, Z. Sun, and L. Su, "Underwater Motion and Activity Recognition using Acoustic Wireless Networks," in *2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[36] J. Yang, J. P. Wilson, and S. Gupta, "DARE: Diver Action Recognition Encoder for Underwater Human–Robot Interaction," *IEEE Access*, vol. 11, pp. 76 926–76 940, 2023.

[37] E. Delhaye, A. Bouvet, G. Nicolas, J. P. Vilas-Boas, B. Bideau, and N. Bideau, "Automatic Swimming Activity Recognition and Lap Time Assessment Based on a Single IMU: A Deep Learning Approach," *Sensors*, vol. 22, no. 15, p. 5786, 2022.

[38] H. Måløy, A. Aamodt, and E. Misimi, "A Spatio-Temporal Recurrent Network for Salmon Feeding Action Recognition From Underwater Videos in Aquaculture," *Computers and Electronics in Agriculture*, vol. 167, pp. 1–9, 2019.

[39] S. S. Enan, M. Fulton, and J. Sattar, "Robotic Detection of a Human-Comprehensible Gestural Language for Underwater Multi-Human-Robot Collaboration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3085–3092.

[40] G. M. Goodfellow, J. A. Neasham, I. Rendulić, Đ. Nađ, and N. Mišković, "DiverNet - a Network of Inertial Sensors for Real Time Diver Visualization," in *2015 IEEE Sensors Applications Symposium (SAS)*, 2015, pp. 1–6.

[41] Đ. Nađ, C. Walker, I. Kvasić, D. O. Antillon, N. Mišković, I. Anderson, and I. Lončar, "Towards Advancing Diver-Robot Interaction Capabilities," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 199–204, 2019.

[42] A. Gomez 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, 2019. [Online]. Available: https://www.mdpi.com/2077-1312/7/1/16

[43] "GoPro Hero 8," 2019, https://gopro.com.

[44] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment Anything," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[45] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:8485068

[46] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019, pp. 8024–8035.

[47] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards Good Practices for Very Deep Two-Stream ConvNets," *arXiv preprint arXiv:1507.02159*, 2015.

[48] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[49] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in *31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.