sensors

MDPI

Article

An Underwater Human–Robot Interaction Using a Visual–Textual Model for Autonomous Underwater Vehicles

Yongji Zhang [1], Yu Jiang [1,2,*], Hong Qi [1,2], Minghao Zhao [1], Yuehang Wang [1], Kai Wang [1] and Fenglin Wei [1]

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2] State Key Lab of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China
* Correspondence: jiangyu2011@jlu.edu.cn

Abstract: The marine environment presents a unique set of challenges for human–robot interaction. Communicating with gestures is a common way for interacting between the diver and autonomous underwater vehicles (AUVs). However, underwater gesture recognition is a challenging visual task for AUVs due to light refraction and wavelength color attenuation issues. Current gesture recognition methods classify the whole image directly or locate the hand position first and then classify the hand features. Among these purely visual approaches, textual information is largely ignored. This paper proposes a visual–textual model for underwater hand gesture recognition (VT-UHGR). The VT-UHGR model encodes the underwater diver's image as visual features, the category text as textual features, and generates visual–textual features through multimodal interactions. We guide AUVs to use image–text matching for learning and inference. The proposed method achieves better performance than most existing purely visual methods on the dataset CADDY, demonstrating the effectiveness of using textual patterns for underwater gesture recognition.

Keywords: autonomous underwater vehicle; underwater human–robot interaction; gesture recognition; visual–textual association

Citation: Zhang, Y.; Jiang, Y.; Qi, H.; Zhao, M.; Wang, Y.; Wang, K.; Wei, F. An Underwater Human–Robot Interaction Using a Visual–Textual Model for Autonomous Underwater Vehicles. *Sensors* **2023**, *23*, 197. https://doi.org/10.3390/s23010197

Academic Editor: Carina Soledad González

Received: 12 November 2022
Revised: 14 December 2022
Accepted: 19 December 2022
Published: 24 December 2022

1. Introduction

Autonomous underwater vehicles (AUVs) can effectively assist divers in complex operations and mitigate risks [1–5] in applications such as marine science, archaeology, and the maintenance of marine infrastructure. AUVs rely on acoustic, inertial, and visual sensors for intelligent decision-making. Figure 1 demonstrates the underwater human–robot interaction (U-HRI) platform, which is built in the CADDY [6,7] project. Specifically, the communication from the diver to the machine is recorded by the AUVs' cameras with a backend algorithm that analyzes its semantic information. Communication from the AUVs to the diver is signaled by the AUVs' flashing lights or through the screens to show more complex information feedback to the diver. The sonar keeps the diver and the AUVs at a distance by detecting the relative positions of the divers. However, due to the low-light underwater environment [8], equipment conditions [2], and restricted human action, U-HRI is much more complicated than land operations. Divers wear special equipment, such as rebreathers, which makes it very difficult to achieve voice communication [1]. Hand gestures are the standard way of communication between divers due to their clear semantics and simplicity of operation. Similarly, gesture recognition is widely used in underwater human–robot interactions.

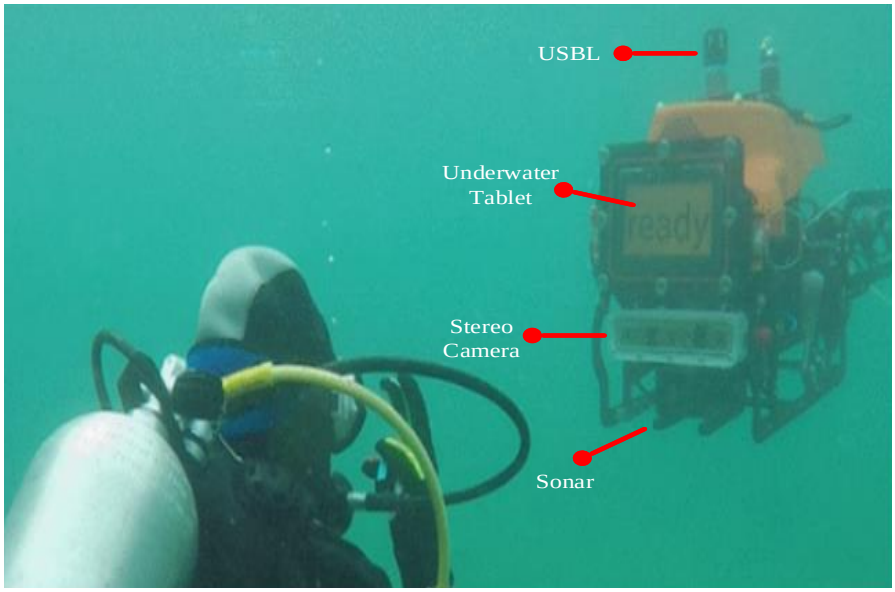

Figure 1. The CADDY system for assistance in diver missions as an example of U-HRI [6,7].

Due to the distractions of the underwater environment [8–10] and the equipment (the overall color of the diving suit is very close to that of the background) [1], current underwater gesture recognition methods used by AUVs usually use a two-stage network architecture [11]. Although the current two-stage network structure has a better accuracy rate, these methods are limited by the efficiency and accuracy of both the detector and classifier components. On the other hand, these methods rely on presegmented images of a dataset under benign conditions. Inspired by recent vision–text-related work [12–14], we realized that there was an intrinsic relationship between images and texts. We tried to consider this problem in terms of the way humans learn. Humans learn by matching the appearance of objects to their names rather than directly matching images to labels. Therefore, we applied visual–text multimodal learning in this task, using text as supervised information for a limited number of underwater images.

This paper proposes an effective visual–textual framework for underwater gesture recognition. We improve the accuracy of gesture recognition by using text semantic information. Our multimodal framework consists of two independent unimodal encoders for images and labels and a multimodal interaction module. This framework extracts visual features from the underwater images and textual features from the labels to infer the meaning of the diver's gestures after multimodal interaction. As shown in Figure 2, the training process matches the image and the label feature pairs as closely as possible, while keeping specific image features away from other mismatched labels. As a result, visual–textual learning brings AUVs closer to human behavioral habits. In the inference phase, the proposed model is a visual–textual matching problem rather than a 1-N majority vote method.

Furthermore, the lack of samples [6,7] makes it challenging to construct visual–textual learning resources. In contrast, the Internet contains many image resources rich in textual markup. Inspired by recent work [15–18], by using pretrained models as a solution, we redefined the task to bootstrap the visual–textual task and fine-tuned the entire model for the underwater dataset. Our contributions can be summarized as follows:

- Underwater gesture recognition is constructed as a multimodal problem, and the U-HRI performance of an AUV is optimized by introducing a text modality to fully explore the feature associations between images and text.
- We propose a new underwater visual–textual gesture recognition model (VT-UHGR).

- Extensive experiments on the public benchmark dataset CADDY [6,7] show that the proposed visual–textual approach achieves superior performance over the visual-only approach.

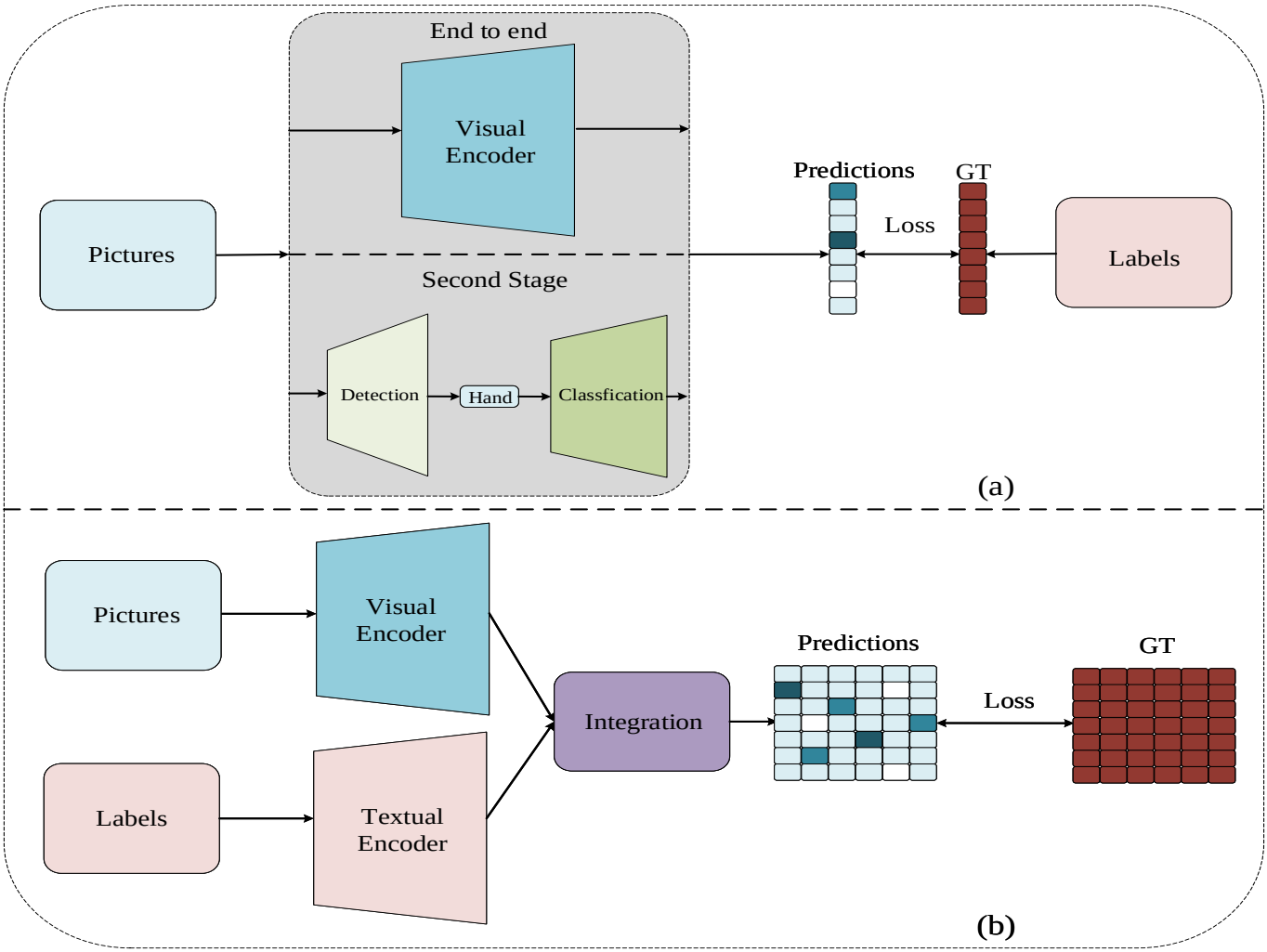

Figure 2. The current traditiona l framework (**a**) and our multimodal framework (**b**).

2. Related Work

2.1. Hand Gesture Recognition

The current mainstream deep-learning-based gesture recognition methods include appearance-based recognition [19,20], motion-based recognition [21,22], skeleton-based recognition [23,24], depth-based recognition [25,26], 3D-model-based recognition [27–30], etc. Chen et al. [19] proposed a two-level approach for real-time gesture classification, with the lower-level approach based on Haar-like features and AdaBoost learning algorithm and the upper-level approach based on a contextual syntactic analysis. Saha et al. [20] relied on the choice of an active difference signature-based feature descriptor and used the HMM model to improve the feature-based methods' performance significantly. Simonyan et al. [21] proposed a spatiotemporal dual-flow structure for recognition using single-frame and front–back optical flow images. Zhu et al. [22] used computational implicit motion information to reduce the high computational cost associated with computing optical flow in motion-based methods. Devineau [23] used parallel convolution to process the position sequences of hand-skeletal joints and still achieved advanced performance using only hand-skeletal data

for recognition. Cai et al. [28] proposed a weakly supervised method, adapting from a fully annotated synthetic dataset to a weakly labeled real-world dataset with a depth regularizer.

2.2. Underwater Gesture Recognition

Due to the influence of the underwater environment [8–10,31,32], underwater images suffer from noise interference, refraction effect, wavelength color attenuation, and other problems. Therefore, it is challenging to accurately identify diver images captured by AUVs in complex underwater environments. Current image-based deep learning methods do not directly yield good performance in underwater environments. In [33], pretrained versions of several classical network models were trained for underwater environments using a migration learning approach, including AlexNet [34], VggNet [35] ResNet [36], and GoogleNet [37]. In [38], an attempt was made to solve this problem from detection to classification using Faster R-CNN [39], deformable Faster R-CNN [11], and other efficient detectors, and a comprehensive analysis of the impact of underwater data was developed. Zhao [40] et al. deployed an underwater real-time FPGA system that carried a lightweight convolutional neural network for underwater image recognition.

2.3. Vision–Text Multimodality

Recent work on semantic–textual information has attracted much attention, such as pretraining [15,16], visual–textual retrieval [12–14], action recognition [17,41,42], attribute recognition [18], etc. CLIP [13] learned transferable visual models from natural language supervision and achieved surprising results on classification tasks. ActionClip [17] verified the effectiveness and scalability of text supervision in video-based action recognition. VTB [18] introduced text supervision to the multiclassification task of pedestrian attribute recognition and significantly improved the results. Fine-tuning on a specific dataset significantly improves the model's performance [15,16].

3. Method

The current gesture recognition methods used by AUVs [21–30] are essentially the classical 1-N majority vote problem that maps the corresponding labels to numerical categories without using the rich semantic information in the labels. Instead, we model the gesture recognition task used by AUVs through an image–text multimodal representation. Specifically, we use semantic information from the text to supervise the classification of images. We take advantage of the improved cognitive approach of the AUV, thus optimizing the accuracy of gesture recognition. As shown in Figure 2, (a) is the current traditional framework and (b) is our multimodal framework (b). (a) uses an end-to-end approach or a two-stage approach to map labels into numbers or one-hot vectors, while (b) tries to pull the semantic information of the label text and the corresponding image representation close to each other.

As shown in Figure 3, we propose a visual–textual baseline as an underwater gesture recognition model (VT-UHGR), which consists of three modules, including visual feature extraction (VFE), textual feature extraction (TFE), and multimodal interaction. The visual feature extraction module generates visual features from the input diver's image. The textual feature extraction module encodes the input label into the corresponding textual features. Furthermore, the multimodal interaction module projects features of two modalities to an identical high-dimensional semantic space and generate visual–textual features using a transformer encoder. Gesture classification predictions are generated from the corresponding visual–textual features by an independent feed-forward network (FFN). Benefiting from the representational power of the transformer, VT-UHGR fully explores intra- and cross-modal correlations. In addition, the location and modality type embeddings are used to maintain spatial and modal information. In subsequent subsections, we describe each module and the employed pretrained methods.

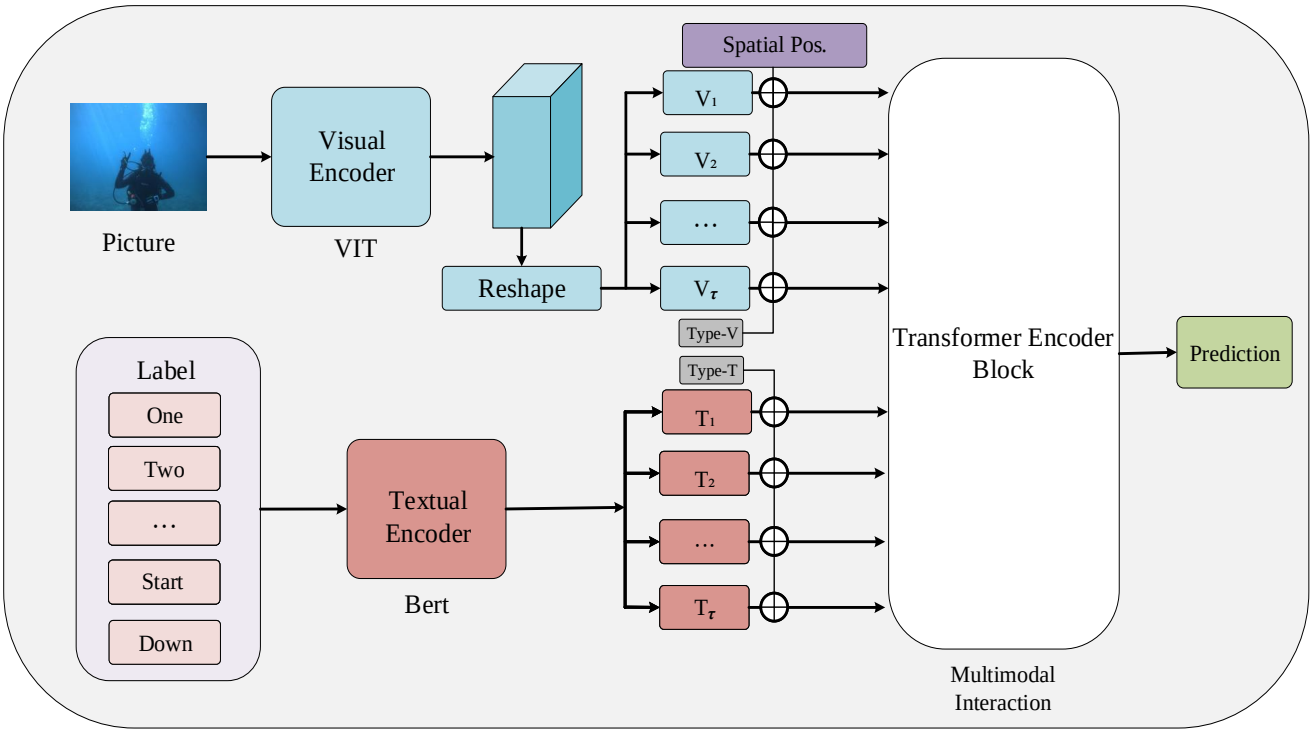

Figure 3. An overview of our proposed visual-textual baseline (VT-UHGR).

3.1. Pretraining

We fine-tuned the framework using a model that had been pretrained on a large dataset, ImageNet [43], and we adapted and recustomized the downstream task to be more like the upstream pretrained task. The traditional prompt and fine-tune approach adapts the pretrained model to the downstream classification task by attaching a new linear layer to the pretrained feature extractor. Here, we used two kinds of cues: textual and visual. We used a pretrained BERT [44] as the textual encoder and a pretrained VIT [45] as the visual encoder. Fine-tuning on a specific dataset significantly improves the model's performance [15,16]. Moreover, since we introduced additional parameters, we needed to train on these parameters. Therefore, we retrained the assembled entire framework end-to-end on the target dataset.

3.2. Transformer Block

The transformer block [45,46] is a type of multiplexed unit used in our network structure. As shown in Figure 4, a transformer block consists of multihead attention, add-and-norm, and MLP modules. Our visual feature extractor, text feature extractor, and multimodal interaction modules were all made up of transformer blocks stacked in different ways. The last layer of the transformer block in the BERT [44] network used for the textual feature extractor was the feed-forward layer.

The effectiveness of the transformer structure mainly relies on the self-attention mechanism [46]. In each block, the vector goes through the multihead self-attention module to get a weighted feature vector Z, i.e., a query feature matrix Q is used to calculate its similarity to each key feature matrix K, and then a weighted sum of all value matrices V is performed:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the dimensionality of the matrix K. In addition, to avoid the model from overly focusing on its position when encoding the information at the current position,

the transformer block uses multiple attention heads to output the encoded representation information in different subspaces, further enhancing the expressive power of the model.

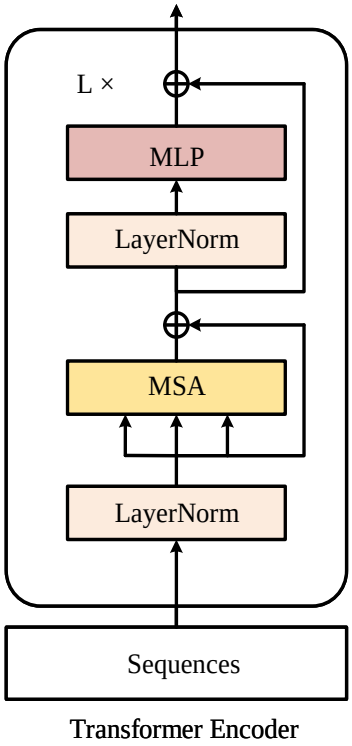

Figure 4. The overall structure of the transformer block.

3.3. Visual Feature Extraction

As shown in Figure 5, given an image I of a diver, we used the VFE module to obtain its visual features V. In a CNN, convolving the image directly in two dimensions is sufficient without a special preprocessing process. However, the transformer structure cannot process the image directly and needs to be chunked beforehand. Specifically, we chose the VIT [45] core process as the visual encoder after making patches, patch embedding, position embedding, and a transformer block encoder to get the image features vector F. That is, $F = VFE(I) \in R^{C \times H \times W}$, where C, H, and W represent the channel size, height, and width of F, respectively. To integrate with the textual features, we further plasticized F by extending it to one dimension in the spatial dimension, resulting in a set of visual feature vectors $V = [V_1, V_2, ..V_\tau] \in R^{S \times C}$, where $S = H \times W$. In summary, the features of the diver image I extracted by the visual feature extractor were defined as :

$$V = VFE(I) = Reshape(VIT(I)) \quad (2)$$

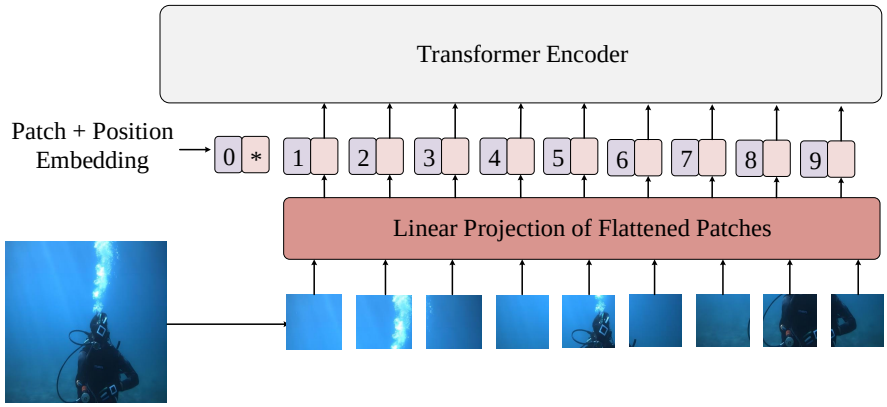

Figure 5. The overall structure of the VIT model.

3.4. Textual Feature Extraction

As shown in Figure 6, given a set of labels $\lambda = [\gamma_1, \gamma_2, \ldots \gamma_\tau]$, we used the TFE module to extract the corresponding textual features T. All original texts (e.g., "boat", "on", etc.) were first assembled into statements (e.g., "This picture represents 'boat"', "The action of the diver is 'on"' etc.) and then encoded them to text features by the natural language method [44,47]. We referred to some natural language processing methods to transform the text of the feature labels to match them in the vector space. Specifically, for the label γ, the word vector was first obtained by an embedding, which contained the word vector token embedding, the segment embedding to distinguish the tag, and the position embedding to encode the position. In addition, the last layer of the transformer block used in BERT was the feed-forward layer. The textual encoder was not involved in the training process. It only used the pretrained model to output the corresponding textual features. Therefore, our textual module did not burden the training considerably.

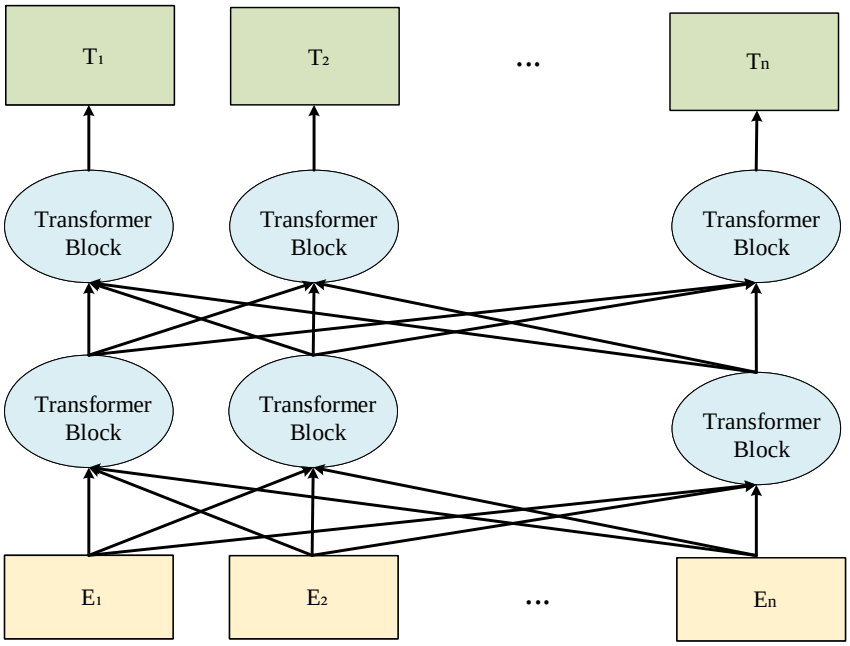

Figure 6. The overall structure of the BERT model.

We finally obtained the textual features T by a pretrained BERT [44] consisting of two-way transformer blocks, i.e., $T = [t_1, t_2, \ldots t_\tau] = BERT(Embedding[\gamma_1, \gamma_2, \ldots \gamma_\tau])$. In summary, the textual extractor extracted the label λ features defined as :

$$T = TFE(\lambda) = BERT(Embedding(\lambda)) \tag{3}$$

3.5. Multimodal Interaction

Multimodal interaction is essential for the exchange of information between multiple modalities. An intuitive approach to modal fusion is integrating different modalities' features by simple operations, such as weighting or cascading. In this paper, we used the transformer encoder for a deep cross-modal fusion. As input, the encoder received a sequence of features from two modalities, textual and visual. The information from both modalities interacted in the encoder through the self-attention mechanism and eventually output multimodal features.

As shown in Figure 3, our model extracts visual features V and textual features T from two feature extractors that interact in a multimodal interaction module to finally generate visual–textual features Z for gesture recognition. We used the transformer encoder [45] as a multimodal interaction module to allow unrestricted interaction between visual–textual sequences. On the other hand, it allowed the framework to model the correlations within and across modalities deeply. We first mapped V and T to the same higher-dimensional space D for sequence concatenation, as follows :

$$V_\varsigma = \left[\Phi(V_1) + e_1^{pos}, \Phi(V_2) + e_2^{pos}, \ldots \Phi(V_\tau) + e_s^{pos}\right] \tag{4}$$

$$T_\varsigma = [\varphi(t_1), \varphi(t_2), \ldots \varphi(t_\tau)], \tag{5}$$

where $\phi(.)$ and $\varphi(.)$ are both linear 1×1 convolution layers, $E^{pos} = \left[e_1^{pos}, e_2^{pos}, \ldots e_s^{pos}\right] \in R^{M\times D}$ is the learnable location embedding, which needs to add the spatial information of the image as visual features to the extracted features. The final obtained visual–textual feature pair Z_0 was as follows :

$$Z^0 = \left[V_\varsigma + e_v^{type}; T_\varsigma + e_t^{type}\right] \tag{6}$$

The learnable modal type embedding e_v^{type}, e_t^{type} was added with the corresponding sequences to maintain the different modal information. Next, the initial visual–textual vectors were learned by a transformer encoder. The transformer encoder consisted of a stack of L transformer blocks, each of which included a multiheaded self-attention (MSA) layer and a multilayer perceptron (MLP) layer, as well as normalization layers (LN); details are shown in Figure 4. The final gesture classification results were generated by feature comparisons extracted from pretrained text labels. Specifically, we derived the corresponding predicted values by an independent feed-forward network (FFN) containing linear layers. The network used the same text features for the same set of different classified diver images.

3.6. Loss Function

Given a set of attribute annotations and a training dataset, we use the cross-entropy method, widely used for multiclassification tasks. Our loss function was formulated as follows:

$$L = -\frac{1}{N}\sum_{i}^{n}\sum_{j}^{M}(y_{ij}log(p_{ij})) \tag{7}$$

In the inference phase, all text labels were encoded as text features by the TFE module, and the corresponding textual features were fed to the multimodal interaction module for prediction. The proposed visual–textual method was trained by optimizing L end-to-end, except for the TFE module, which was not involved in model training. Since our training

goal was to make the feature vectors of an image–text pair as similar as possible, rather than a pair as similar as possible, here, the similarity was calculated using the inner vector product. In this way, our loss function was computed by N positive samples and $N^2 - N$ negative samples.

4. Results

4.1. Datasets

BUDDY-AUV collected the CADDY [6,7] data used in this paper, and the University of Zagreb designed it. It is equipped with navigation sensors, a Doppler velocity logger (DVL), an ultrashort baseline (USBL), and perception sensors, including a multibeam sonar and a stereo camera in the underwater housing. Additional underwater housing enabling two-way human–machine interaction is also included.

The underwater robot captured the gesture image dataset in eight underwater scenes and the dataset contains over 10,000 images. For different scenes, each image contains the image number, shooting scene, gesture name meaning, gesture number, left-hand position, right-hand position, semantics, and other information. As shown in Figure 7, the AUV intelligently senses divers' gestures in the underwater environment. The dataset contains common underwater communication gestures such as up dive, carry, digit, stay, and boat.

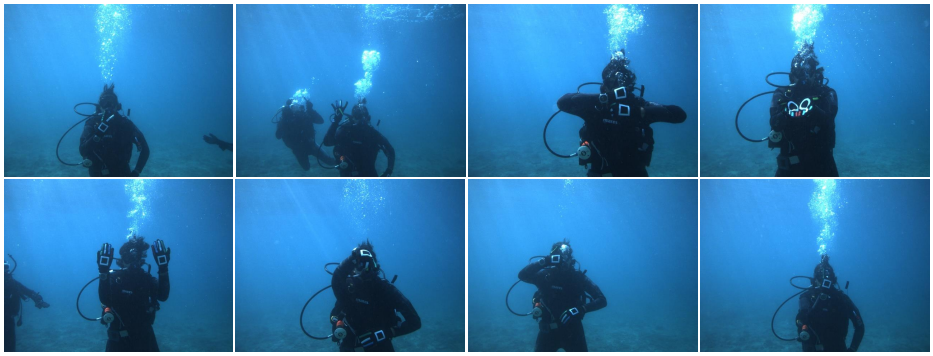

Figure 7. Example images from the CADDY [6,7] dataset.

We evaluated the performance of the model using accuracy, i.e., the number of samples correctly predicted as a percentage of the total. The CADDY dataset has a relatively homogeneous sample distribution, and using accuracy gave a more intuitive picture of the model's performance level.

In addition, to test the performance of our model in a more general scenario, we additionally chose to supplement it with the SCUBANet dataset [48], which was recorded by the Milton unmanned underwater vehicle (UUV). As shown in Figure 8, it contains two distinct scenarios of cold and warm water. Since the SCUBANet dataset is primarily designed to detect the body parts of divers, we filtered the dataset by explicit gesture semantics.

Figure 8. Example images of the SCUBANet [48] dataset.

4.2. Implementation Details

We implemented VIT [45] as our visual coder, which was pretrained on ImageNet [43]. Correspondingly, we used a pretrained BERT [44] as a text encoder. The cell structure in the transformer encoder for multimodal interaction was the same as that of the transformer block in VIT. The dimensionality of the visual–textual features was set to 256. The images used in the training process were enhanced with random level flipping and random cropping. We used the Adam optimizer to optimize our model with $\beta_1 = 0.9, \beta_2 = 0.999$. We used a warm-up strategy to increase the learning rate linearly from 0 to an initial learning rate of 1×10^{-3} in the first ten epochs and decreased the learning rate by a factor of 0.1 when the number of iterations increased. The batch size was set to 32.

4.3. Comparison with the State of the Art

We compared our method with state-of-the-art end-to-end methods for underwater gesture recognition, as shown in Table 1. We compared the model's accuracy, the number of parameters, and the average time of inference for diver gesture recognition with different methods on the same test set. The data for the comparison methods were mainly from [33]. Experiments showed that our method significantly outperformed the end-to-end methods alone in terms of accuracy. Since the sample distribution of the CADDY dataset was relative, using the accuracy rate could reflect the performance level of the model intuitively. In addition, due to the larger number of transformer structural parameters, our inference time in the same environment was slightly longer.

Table 1. Quantitative comparison of end-to-end methods on the CADDY dataset.

Methods	Acc	Params (M)	Times (ms)
AlexNet	0.83	61.1	0.84
ResNet	0.88	11.7	1.26
GoogleNet	0.90	6.8	1.65
VggNet	0.95	138.4	2.14
VT-UHGR (ours)	0.98	178.4	2.87

We compared our method with end-to-end underwater gesture recognition methods on the SCUBANet dataset, as shown in Table 2. All models showed a significant decrease in efficiency compared to the results on the CADDY dataset. The divers in SCUBANet did not wear markers on their hands, and problems such as complex backgrounds and motion blur in the dataset posed challenges for the recognition. In addition, due to the large image size of SCUBANet samples, the large scaling to fit the input of the model also posed difficulties for feature extraction.

Table 2. Quantitative comparison of end-to-end methods on the SCUBANet dataset.

Methods	Acc
ResNet	0.75
GoogleNet	0.78
VggNet	0.82
VT-UHGR (ours)	0.86

We compared our method with state-of-the-art two-stage underwater gesture recognition methods, as shown in Table 3. We compared the accuracy of gesture recognition for divers with different methods on the same test set. The data for the comparison methods were mainly from [38]. Experiments showed that our method had comparable performance in terms of accuracy compared to the two-stage method. Our model could achieve similar performance to detection-based methods without relying on detection labeling pairs.

Table 3. Quantitative comparison of two-stage methods on the CADDY dataset.

Methods	Acc
MD-NCMF	0.77
SSD with MobileNets	0.85
FC-CNN with ResNet-50	0.95
Deformable Faster R-CNN	0.98
VT-UHGR (ours)	0.98

The superior performance of our approach was mainly attributed to the textual features provided by the pretrained textual encoder, which helped the model learn the intrinsic feature correlation between pictures and text. On the other hand, we effectively interacted with intermodal features through the multimodal interaction module. Thus, our approach was significantly better than the visual-only based approach.

4.4. Ablation Study

In this subsection, we conducted an extensive ablation study in the CADDY dataset. As shown in Table 4, we explored the impact of different textual encoders and the absence of textual encoders on the performance of our model. Using a text feature extractor significantly improved the performance of the visual-only model, and even when using a simple one-hot encoding approach to interact with visual features, the performance was significantly better than the visual-only scheme. In addition, more detailed processing of textual features using BERT [44] could further improve the model results. Our approach treated underwater gesture recognition as a multimodal task and fully used the textual information inherent in the labels.

Table 4. Impact of different textual encoders on VT-UHGR performance on the CADDY dataset.

Methods	Textual Encoder	Acc
VIT	–	95.81
VT-UHGR	One-hot	97.13
VT-UHGR	BERT	98.32

As shown in Table 5, we tested the need for multimodal interaction. Specifically, we constructed a more straightforward visual–textual baseline using additive operations as a cross-modal fusion module instead of a transformer encoder. In addition, we tested the corresponding performance degradation without positional encoding and modal encoding, respectively.

Table 5. Ablation study of the proposed components in VT-UHGR on the CADDY dataset.

Methods	Structure	Acc
VT-UHGR	–	97.19
VT-UHGR	+ Transformer Encoder	97.78
VT-UHGR	+ E^{pos}	98.11
VT-UHGR	+ E^{type}	98.32

Experiments showed that using the transformer block as a multimodal interaction module significantly improved performance. In addition, it was necessary to provide more numerous conditions, such as positional encoding for the encoded information. On the other hand, the introduction of the learnable modal encoding also improved the performance of the final result.

5. Limitation

There is still a lack of large-scale, scenario-rich, high-resolution benchmark datasets due to the complex underwater environment and the limitations of the hardware quality of the filming equipment. In addition, the intensive use of transformer blocks in the inference process places an additional burden on the hardware side of the computational deployment. Therefore, using lightweight models to achieve more accurate recognition will make more sense in the future of edge computing.

6. Conclusions

In this paper, we formulated underwater gesture recognition as a multimodal problem and proposed a visual–textual baseline (VT-UHGR) to improve the performance of U-HRI for AUVs. We explored the correlation between visual and textual features using the transformer block. Extensive experiments on a widely used CADDY dataset demonstrated the importance of introducing textual modalities, and our proposed visual–textual baseline achieved a higher performance than the purely visual approach.

Author Contributions: Y.Z. and Y.J. designed the methodology, conducted the experiment, analyzed the results, and wrote the manuscript. F.W. and K.W. participated in the experiment and provided advice on the data analysis. M.Z., H.Q. and Y.W. were involved in part of the experiment and manuscript revision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grant 62072211, grant 51939003, and grant U20A20285.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Birk, A. A Survey of Underwater Human-Robot Interaction (U-HRI). *Curr. Robot. Rep.* **2022**, *3*, 199–211.
2. Mišković, N.; Egi, M.; Nad, D.; Pascoal, A.; Sebastiao, L.; Bibuli, M. Human-robot interaction underwater: Communication and safety requirements. In Proceedings of the 2016 IEEE Third Underwater Communications and Networking Conference (UComms), Lerici, Italy, 30 August–1 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5.
3. Sun, K.; Cui, W.; Chen, C. Review of Underwater Sensing Technologies and Applications. *Sensors* **2021**, *21*, 7849. https://doi.org/10.3390/s21237849.
4. Pan, S.; Shi, L.; Guo, S. A Kinect-Based Real-Time Compressive Tracking Prototype System for Amphibious Spherical Robots. *Sensors* **2015**, *15*, 8232–8252. https://doi.org/10.3390/s150408232.
5. Qin, R.; Zhao, X.; Zhu, W.; Yang, Q.; He, B.; Li, G.; Yan, T. Multiple Receptive Field Network (MRF-Net) for Autonomous Underwater Vehicle Fishing Net Detection Using Forward-Looking Sonar Images. *Sensors* **2021**, *21*, 1933. https://doi.org/10.3390/s21061933.
6. Chiarella, D.; Bibuli, M.; Bruzzone, G.; Caccia, M.; Ranieri, A.; Zereik, E.; Marconi, L.; Cutugno, P. A novel gesture-based language for underwater human–robot interaction. *J. Mar. Sci. Eng.* **2018**, *6*, 91.
7. Gomez Chavez, A.; Ranieri, A.; Chiarella, D.; Zereik, E.; Babić, A.; Birk, A. CADDY Underwater Stereo-Vision Dataset for Human–Robot Interaction (HRI) in the Context of Diver Activities. *J. Mar. Sci. Eng.* **2019**, *7*, 16. https://doi.org/10.3390/jmse7010016.
8. Blizard, M.A. Ocean optics: Introduction and overview. In *Ocean Optics VIII*; SPIE: Bellingham, DC, USA, 1986; Volume 637, pp. 2–17.
9. Schettini, R.; Corchs, S. Underwater image processing: State of the art of restoration and image enhancement methods. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 1–14.
10. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.* **2019**, *29*, 4376–4389.
11. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
12. Fang, H.; Xiong, P.; Xu, L.; Chen, Y. Clip2video: Mastering video-text retrieval via image clip. *arXiv* **2021**, arXiv: 2106.11097.

13. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
14. Miech, A.; Laptev, I.; Sivic, J. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv* **2018** arXiv:1804.02516.
15. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
16. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7331–7341.
17. Wang, M.; Xing, J.; Liu, Y. Actionclip: A new paradigm for video action recognition. *arXiv* **2021**, arXiv:2109.08472.
18. Cheng, X.; Jia, M.; Wang, Q.; Zhang, J. A Simple Visual-Textual Baseline for Pedestrian Attribute Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6994–7004.
19. Chen, Q.; Georganas, N.D.; Petriu, E.M. Real-time vision-based hand gesture recognition using haar-like features. In Proceedings of the 2007 IEEE Instrumentation & Measurement Technology Conference IMTC, Warsaw, Poland, 1–3 May 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–6.
20. Saha, S.; Lahiri, R.; Konar, A.; Banerjee, B.; Nagar, A.K. HMM-based gesture recognition system using Kinect sensor for improvised human-computer interaction. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2776–2783.
21. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27; pp. 568–576*.
22. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 363–378.
23. Devineau, G.; Moutarde, F.; Xi, W.; Yang, J. Deep learning for hand gesture recognition on skeletal data. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 106–113.
24. Nguyen, X.S.; Brun, L.; Lézoray, O.; Bougleux, S. A neural network based on SPD manifold learning for skeleton-based hand gesture recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 12036–12045.
25. Bakar, M.Z.A.; Samad, R.; Pebrianti, D.; Mustafa, M.; Abdullah, N.R.H. Finger application using K-Curvature method and Kinect sensor in real-time. In Proceedings of the 2015 International Symposium on Technology Management and Emerging Technologies (ISTMET), Langkawi Island, Malaysia, 25–27 August 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 218–222.
26. Wu, X.; Finnegan, D.; O'Neill, E.; Yang, Y.L. Handmap: Robust hand pose estimation via intermediate dense guidance map supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 237–253.
27. Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3d hand shape and pose estimation from a single rgb image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 10833–10842.
28. Cai, Y.; Ge, L.; Cai, J.; Yuan, J. Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 666–682.
29. Miao, Q.; Li, Y.; Ouyang, W.; Ma, Z.; Xu, X.; Shi, W.; Cao, X. Multimodal gesture recognition based on the resc3d network. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3047–3055.
30. Zhu, G.; Zhang, L.; Mei, L.; Shao, J.; Song, J.; Shen, P. Large-scale isolated gesture recognition using pyramidal 3d convolutional networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 19–24.
31. Kim, H.G.; Seo, J.; Kim, S.M. Underwater Optical-Sonar Image Fusion Systems. *Sensors* **2022**, *22*, 8445. https://doi.org/10.3390/s22218445.
32. Du, W.; Yang, Y.; Liu, L. Research on the Recognition Performance of Bionic Sensors Based on Active Electrolocation for Different Materials. *Sensors* **2020**, *20*, 4608. https://doi.org/10.3390/s20164608.
33. Yang, J.; Wilson, J.P.; Gupta, S. Diver gesture recognition using deep learning for underwater human-robot interaction. In Proceedings of the OCEANS 2019 MTS/IEEE SEATTLE, Seattle, WA, USA, 27–31 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90.
35. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

37. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
38. Chavez, A.G.; Ranieri, A.; Chiarella, D.; Birk, A. Underwater Vision-Based Gesture Recognition: A Robustness Validation for Safe Human–Robot Interaction. *IEEE Robot. Autom. Mag.* **2021**, *28*, 67–78.
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
40. Zhao, M.; Hu, C.; Wei, F.; Wang, K.; Wang, C.; Jiang, Y. Real-time underwater image recognition with FPGA embedded system for convolutional neural network. *Sensors* **2019**, *19*, 350.
41. Piergiovanni, A.; Ryoo, M. Learning multimodal representations for unseen activities. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass village, Colorado, 2–5 March 2020; pp. 517–526.
42. Alayrac, J.B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; Zisserman, A. Self-supervised multimodal versatile networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 25–37.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
44. Peng, Y.; Yan, S.; Lu, Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv* **2019**, arXiv:1906.05474.
45. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
46. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
47. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48–57.
48. Codd-Downey, R.; Jenkin, M. Finding divers with SCUBANet. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5746–5751. https://doi.org/10.1109/ICRA.2019.8793655.