electronics

MDPI

Article

An Underwater Multi-Label Classification Algorithm Based on a Bilayer Graph Convolution Learning Network with Constrained Codec

Yun Li [1], Su Wang [2,*], Jiawei Mo [1,*] and Xin Wei [1]

[1] School of Information Science and Engineering, Liuzhou Institute of Technology, Liuzhou 545000, China; liyun@guat.edu.cn (Y.L.); 2213393053@st.gxu.edu.cn (X.W.)
[2] Yangzhou Branch, China Mobile Communications Group Jiangsu Co., Ltd., Yangzhou 225000, China
* Correspondence: 2312392109@st.gxu.edu.cn (S.W.); 2113391043@st.gxu.edu.cn (J.M.)

Abstract: Within the domain of multi-label classification for micro-videos, utilizing terrestrial datasets as a foundation, researchers have embarked on profound endeavors yielding extraordinary accomplishments. The research into multi-label classification based on underwater micro-video datasets is still in the preliminary stage. There are some challenges: the severe color distortion and visual blurring in underwater visual imaging due to water molecular scattering and absorption, the difficulty in acquiring underwater short video datasets, the sparsity of underwater short video modality features, and the formidable task of achieving high-precision underwater multi-label classification. To address these issues, a bilayer graph convolution learning network based on constrained codec (BGCLN) is established in this paper. Specifically, modality-common representation is constructed to complete the representation of common information and specific information based on the constrained codec network. Then, the attention-driven double-layer graph convolutional network module is designed to mine the correlation information between labels and enhance the modality representation. Finally, the combined modality representation fusion and multi-label classification module are used to obtain the category classifier prediction. In the underwater video multi-label classification dataset (UVMCD), the effectiveness and high classification accuracy of the proposed BGCLN have been proved by numerous experiments.

Keywords: underwater; graph convolution; multi-label classification

Citation: Li, Y.; Wang, S.; Mo, J.; Wei, X. An Underwater Multi-Label Classification Algorithm Based on a Bilayer Graph Convolution Learning Network with Constrained Codec. *Electronics* **2024**, *13*, 3134. https://doi.org/10.3390/electronics13163134

Academic Editors: Silvia Liberata Ullo and Ping-Feng Pai

Received: 23 June 2024
Revised: 3 August 2024
Accepted: 3 August 2024
Published: 7 August 2024

1. Introduction

In recent years, micro-video, as a new media form of user-generated content, is rapidly becoming one of the mainstream trends of social media, with its short, real, and instant-sharing characteristics. Micro-videos are rich in content and concise, serving as a combination of various modalities such as vision, audio, and text, and thus contain a vast amount of information. Therefore, it is of great significance to make full use of their multi-modal information for data mining and intelligent analysis. Making full use of multi-modal information focuses on mining the consistency and complementarity among multi-modal information to enhance the information representation. For example, when a micro-video shows dolphins swimming in the ocean, the audio of the micro-video is usually accompanied by the sound of the dolphins, but it is difficult to get information about the sea through the sound. In the above example, "dolphin" is information expressed by visual and acoustic modalities, representing consistency, while "ocean" is information unique to the visual modality, representing complementarity. Therefore, fully utilizing the consistency and complementarity of multi-modal information to enhance information representation is an essential step in the classification task.

The research directions for multi-label classification of micro-videos mainly include the following: (1) Micro-Video Scene Recognition. Nie et al. [1] proposed a deep migration model to accomplish the task of micro-video scene category estimation. This model

introduces external acoustic knowledge to compensate for relatively low-quality audio modality, thus enhancing the semantic representation of micro-videos. (2) Micro-Video Event Detection. Chang et al. [2] defined the concept of semantic salience to evaluate the correlation of each video segment with the event of interest. They prioritized video segments based on saliency scores and leveraged the constructed semantic ranking information to improve the model's discriminative ability in event analysis tasks. (3) Micro-Video Prevalence Prediction. Chen et al. [3] proposed a direct multi-modal learning model that seeks an optimal latent common subspace among different modalities to alleviate the information insufficiency issue arising from the short duration of micro-videos, thereby facilitating better representation. (4) Micro-Video Recommendation. Wei et al. [4] designed a multi-modal graph convolutional network framework. By constructing a binary graph of user micro-videos for each modality, the authors used the interaction behavior of users and micro-videos to guide the representation learning of each modality, further capturing users' fine-grained preferences in different modalities.

It can be seen that the scientific research into the multi-label classification of micro-videos has made achievements. However, the related work in the field of micro-videos is only limited to the consistency or complementarity between multi-modalities, while considering the consistency, complementarity, and multi-modal characterization. Furthermore, while these methods demonstrate promising performance in terrestrial video datasets, underwater environments pose significant challenges, such as color distortion, blurriness, and occlusion, which severely degrade video quality and consequently lead to a decline in detection and classification accuracy [5]. Moreover, the underwater dataset is difficult to obtain, resulting in sparsely available modality information, and it is difficult to jointly mine the multi-modal complementarity and consistency enhancement information characterization. Consequently, exploring multimodal learning approaches for underwater micro-videos that can comprehensively utilize both the consistency and complementarity of different modalities, fostering mutual reinforcement among modal information while mitigating redundancies, holds significant importance for scientific research on marine imagery.

For the multi-modal multi-label classification task of underwater micro-videos, this work makes three contributions:

(1) An underwater video multi-label classification dataset (UVMCD) is constructed, containing 3841 underwater videos covering 19 underwater categories. There were eight video classification methods used to benchmark the availability of the dataset.

(2) In the original modality features, the common information between the modalities and the specific information within the modalities are intertwined, and the redundancy between them will even contaminate the extracted representation. Therefore, it is necessary to explore the multi-modal representation learning methods that can separate these two parts of the information from the original information and minimize the redundancy.

(3) For multi-label learning, it is inevitable to consider the correlation between label categories. It is noted that there may be locality in the correlation between labels, whereby different instance groups share different label correlations rather than being globally applicable. Therefore, methods that can learn label correlations based on global and local adaptations need to be explored.

2. Related Work

Multi-label technology has been developed for many years. In the paradigm of multi-label classification, each object is associated with multiple labels simultaneously; the task of multi-label classification methods is to learn a function that can predict the corresponding label set of an input instance. Multi-label classification techniques have been widely used in various real-life scenarios such as medical diagnosis [6], bio-informatics [7], user analysis [8], and autonomous driving [9,10].

One of the main objectives of traditional multi-label classification methods is to expand the migration of mature single-label classification algorithms to the multi-label field. According to different expansion ideas, this can be divided into two categories: the problem

conversion method and the algorithm conversion method. The basic idea of the problem conversion method is to transform the multi-label classification task into one of multi-label or more single-label classification tasks. The binary association method (binary relevance, BR) [11] in a multi-label classification task is transformed into multiple independent single-label binary classification tasks, and finally, the output of each binary classifier is aggregated to obtain the final multi-label prediction results. The basic idea of the algorithm adaptive method is to improve and extend the existing single-label classification algorithm so that it can process multi-label data, and then complete the task of multi-label classification. The algorithm models often used for the extension include k-Nearest Neighbors (KNN), decision trees, support vector machines (SVMs), neural networks, etc. Traditional methods often overlook or adopt rudimentary approaches to account for label correlations.

In recent years, deep learning technology has made great progress, and it has made breakthroughs in many application scenarios, such as a new model based on a deep neural network (DNN) first proposed by Yeh et al. [12]. The standard relevant autoencoder (Canonical Correlated Autoencoder, C2AE) integrates typical correlation analysis (Canonical Correlation Analysis, CCA) to derive deep latent spatial joint features and label embedding to better associate features and label domain data to improve classification performance. Fei et al. [13] proposed a latent sentiment memory network (LSMN) tailored for the multi-label sentiment classification of texts. This network is capable of learning the distribution of latent sentiments, without relying on external knowledge, and effectively integrating them into the classification network.

Multi-modal representation learning aims to represent and extract effective semantic information, and the heterogeneous differences between data of different modes are a major challenge in constructing multi-modal representations. It is usually divided into two categories of methods: joint representation and coordinated representation. The joint representation projects the multi-modal data into a common representation space. Rajagopalan et al. [14] designed a multi-view LSTM network for multimodal action recognition and image captioning tasks, which explicitly learns the changes in specific views and cross-view interactions over time or structured outputs. In contrast, coordinated representation methods process each modality independently. Fan et al. [15] combined CCA with a generative adversarial network (GAN) to propose a deep adversarial CCA model, which can simultaneously learn representations of multi-view data while possessing the ability to generate authentic multi-view samples.

Multi-modal fusion is one of the most studied directions in the field of multi-modal learning. The stage of modality fusion is divided into early fusion (feature-level fusion), late fusion (decision-level fusion), and hybrid fusion between the two. Srivastava et al. [16] proposed a multi-modal data generation model based on the deep Boltzmann machine (DBM) to generate a feature representation of missing modes and combine cross-modal features to create fusion features to complete the classification and information-retrieval tasks. The conventional multi-modal fusion method mainly uses kernel learning [17], graphical model [18], and CCA [19]. However, due to the powerful flexibility of deep learning models, the traditional classification method is no longer suitable for deep learning multi-modal fusion-based methods. Deep automatic encoder (DAE) aims to encode the input data for meaningful compression; it consists of two parts, an encoder and a decoder. The encoder converts high-dimensional input data mapping to low-dimensional space to get the potential space representation, and the decoder decodes the potential space representation to reconstruct the original input data. Shen et al. [20] proposed the focus multi-modal DAE model for the extraction of multi-modal social media content (such as text, images, and micro-video, etc.), data learning cross-modal potential representation, and using the attention mechanism integration with variable weight user global and context music preferences, converting the social media content data into the music recommendation task. Guo et al. [21] put forward the standardization of the attention mechanism and geometric perception of improvement, the self-attention mechanism to parameterize, and the latter extending the attention mechanism to explicitly consider the relative geometric

relationship of the input object; the video description, machine translation, and visual quiz task verified the generality of the improvement. Azad et al. [22] proposed a multi-label video classification model, using a self-attention mechanism to capture spatiotemporal attention in continuous video frames, to improve existing methods that consider the spatial information of only a single frame in underwater hull video inspection. Sun et al. [23] introduced a single-channel multi-target underwater acoustic signal recognition method based on deep learning, aiming to address two subproblems, identifying the unique and repetitive categories of multiple targets within a specified class. Leveraging the multimodal information embedded in videos, algorithms for video multi-label classification based on multimodal fusion were proposed. Le et al. [24] transmitted multimodal information through a unified transformer architecture to learn joint multimodal representations for multi-label video sentiment recognition. Cai et al. [25] proposed a multi-modal movie-type classification framework, which makes full use of the information complementarity in the multi-modalities and improves the classification performance. For multi-modal learning, multi-modal data describe the same concept object from different levels and perspectives, which can often complement each other. However, there are heterogeneous differences between modality data from different information sources, which hinders the direct information interaction between modalities and masks the intrinsic strong correlation and semantic consistency between them. In addition, when each mode interacts, the noise information in a certain mode data may pollute other modes with the interaction process and lead to a decrease in model performance. Therefore, how to identify the sparsity of underwater information, how to reduce the heterogeneity between modalities, and mine the intrinsic consistent and complementary semantic information of the data of different modes, while removing redundant information and filtering noise, are the main challenges for researchers.

3. The Algorithm Model

The overall framework of a bilayer graph convolution learning network based on a constrained codec (BGCLN) is shown in Figure 1, which can be roughly divided into the following three parts:

(1) Modality-specific representation and modality-common representation learning modules: The modality-common representation learns through adversarial training; the orthogonal constraint separates the common information and specific information of the modality features to reduce the redundancy between the learned representations. The reconstruction constraint preserves the effective information in the original modality features as much as possible.

(2) Attention-driven double-layer graph convolution network module: A two-layer graph convolutional network (GCN) correlates mined label information between the global and local perspectives and introduces the attention mechanism in the second GCN mining sample with characteristic and label category dimensions to enhance the modality representation.

(3) Modality representation fusion and multi-label classification module: Take the weighted fusion of the enhanced modality-common representation and each modality-specific representation as the final micro-video representation, and the fusion weight is adaptively learned by the model. The resulting representation is then input into the classifier to obtain the category prediction score.

Assume the existence of a micro-video collection, $\chi = \{x^1, x^2, \ldots, x^N\}$, as the training data, which comprise a total of N micro-video samples. For the $i(i = 1, 2, \ldots, N)$-th micro-video sample in this collection, pre-extracted multi-modal features can be used to represent it as $\left\{ x_m^i \in R^{D_m} \middle| i = 1, 2, \ldots, N \right\}$, $m \in \{v, t, a\}$, where D_m denotes the dimensionality of the multi-modal features, and m serves as a modality indicator, with values v, t, and a representing the visual, trajectory, and acoustic modalities, respectively. Additionally, the true label of x^i can be described by a binary label vector, $y_i \in \{0, 1\}^C$, where 1 indicates the

presence of the label in x^i, and 0 indicates its absence. The total number of label categories is C. Without the loss of generality, x_m is used hereinafter to represent the multi-modal features corresponding to any given micro-video sample in the collection.

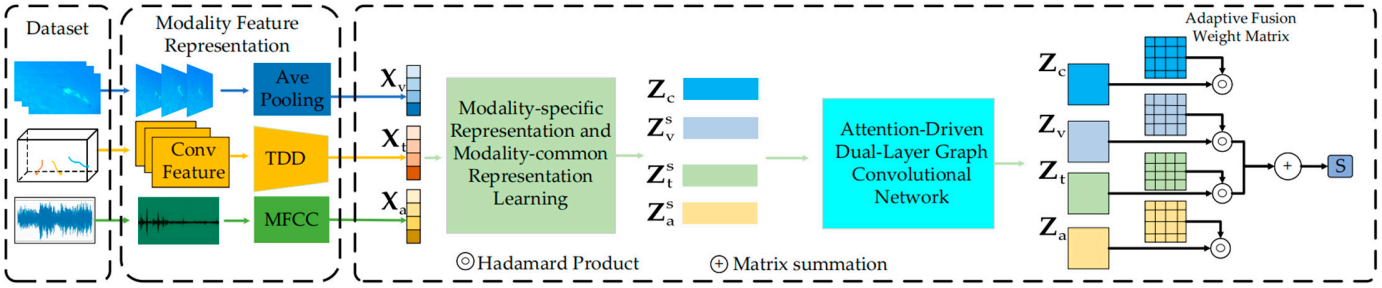


Figure 1. Framework structure diagram of bilayer graph convolution learning network based on a constrained codec.

3.1. Modality-Specific Representation and Modality-Common Representation Learning Modules

Given the consistency and complementarity among the information of different modalities, this paper proposes a codec network structure with orthogonal, adversarial similarity, and reconstruction constraints to learn the modality-common representation and the specific representation of each mode. The main body of this module consists of a modality-specific encoder, modality-common encoder, modality-common decoder, and modality discriminator.

First, the visual, trajectory, and acoustic modality features, x_m, are input into the private coding network corresponding to each modality (as shown in Figure 2) to obtain the specific representation, z_m^s, of each mode:

$$z_m^s = E_m(x_m; \theta_m) \in R^{d_m}, \tag{1}$$

where d_m is the number of dimensions represented by each modality after encoding, and the private feature $E_m(\cdot)$, the encoder of the corresponding mode θ_m, and the encoder are stacked out of multiple fully connected layers.

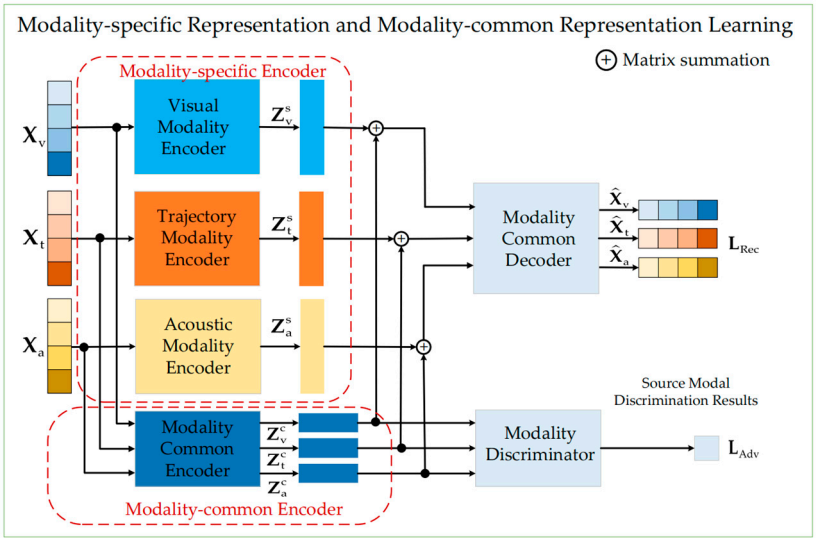


Figure 2. Framework structure diagram of modality-specific representation and modality-common representation learning modules.

At the same time, visual, trajectory, and acoustic modality features, x_m, are input into the modality-common coding network to generate common representations based on each modality:

$$z_m^c = E_c(x_m; \theta_c) \in R^{d_c}, \tag{2}$$

where the number of dimensions publicly expressed by the encoding modality d_c, representing the modality-common feature encoder $E_c(\cdot)$, with the same network structure $E_m(\cdot)$, θ_c is the learnable network parameters of the encoder. Based on the consideration of model consistency, the modality-sharing feature encoder is committed to extracting the common information between the modalities as modality-common representation, so the public representation learned from different modalities should be consistent in the model training process, which can take the public representation generated by the modality average as the final modality-common representation:

$$z^c = \frac{1}{M} \sum_{m=\{v,t,a\}} z_m^c \in R^{d_c}, \tag{3}$$

where $M = 3$, means that the modality uses the features of the three modalities as input.

3.1.1. Orthogonal Constraints

Inspired by the work [26,27], the introduction of orthogonal constraints to promote a common coding network and private coding network to explore the different aspects of input modality characteristics, separating the common and specific information between modalities, means the modality-specific representation does not contain shared information, as much as possible. The orthogonal loss is used to measure the magnitude of similarity between the common representation and the specific representation of each modality, which can be defined by the formula:

$$L_{Ort} = \left\| z_v^s (z_v^c)^T \right\|_F^2 + \left\| z_t^s (z_t^c)^T \right\|_F^2 + \left\| z_a^s (z_a^c)^T \right\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ represents the Frobenius norm. At the time $z_v^s = z_v^c$, $z_t^s = z_t^c$, $z_a^s = z_a^c$, the value of the orthogonal loss reaches the maximum, L_{Ort}. Therefore, the modality-specific representation, L_{Ort}, can be orthogonal to the modality-common representation, as much as possible, to separate the two and reduce the redundant information.

3.1.2. Adversarial Similarity Constraints

Considering the consistency between modalities, the common representations learned from different modality features should be as consistent as possible. Inspired by other work [26,28,29], we design adversarial similarity constraints based on adversarial training ideas. Specifically, the modality-common coding network is treated as a generator, $G(x_m; \theta_c)$, and a class classifier, $D(z_m^c; \theta_d)$, is introduced as modality discriminator M for its learnable network parameters, θ_d. The modality discriminant D takes the public representations generated based on different modes as input, and is committed to correctly identifying their source modalities, while the generator G aims to generate the common representations that can confuse the judgment of D. The two learn against each other in the training process, so that the common representations generated by different modalities tend to be consistent.

Inspired by [30], the gradient reversal layer (GRL) is used to achieve this adversarial learning by redefining the backward function of the module. The network introducing the GRL remains consistent with the original network during forward propagation, but during backpropagation, the gradient is multiplied by a negative constant to reverse the gradient direction. Therefore, the gradient direction of the learnable network parameters θ_c, of the GRL inversion generator G, is used to form a confrontation between the discriminator and the generator.

The formula for the antagonistic similarity loss, L_{Adv}, is as follows:

$$L_{Adv} = \sum_{i=1}^{N} \sum_{m \in \{v,t,a\}} l_i^m \log P_i^m(z_m^c), \tag{5}$$

where the source modality, $l_i^m \in \{0,1\}$, of the current public m representation z_m^c is described, the common representation is generated by the currently indicated modality, which is the probability, $P_i^m(z_m^c)$, that the common representation comes from the indicated m modality predicted by the mode discriminator D. In essence, the generator G is dedicated to minimization, L_{Adv} the mode D, and the discriminant is dedicated to maximizing L_{Adv}, forming a confrontation between the two.

The antagonistic similarity loss L_{Adv} can measure the difference between the public representations generated by different modalities. When the value L_{Adv} is small, the modality discriminator can better judge the source modality of the representation. Therefore, the similarity of the public representations obtained by different modalities is improved by minimizing L_{Adv}.

3.1.3. Refactoring Constraints

To ensure the integrity and validity of each piece of modality information during the encoding process, reconstruction constraints are introduced. Specifically, the specific representation generated by the same modality and the common representation are first input into the modality-common decoder network to obtain the reconstruction vector, $\hat{x}_m$, of the mode feature:

$$\hat{x}_m = D_s((z_m^s + z_m^c); \theta_s) \in R^{D_m}, \tag{6}$$

where $D_s(\cdot)$ represents the modality-common decoder, the learnable network parameters θ_s. Without a loss of generality, the reconstructed vector $\hat{x}_m$ should be as similar as possible to the original modality features, x_m, to ensure that the encoded representation retains the effective information in the original modality features to the greatest extent. The mean squared error with constant proportions was used. To measure the difference between x_m and $\hat{x}_m$, the calculation formula of reconstruction loss is as follows: L_{Rec}

$$L_{Rec} = \sum_{m \in \{v,t,a\}} \frac{1}{k}(\|x_m - \hat{x}_m\|_2)^2 - \frac{1}{k^2}([x_m - \hat{x}_m] \times \mathbf{1}_k)^2, \tag{7}$$

where $\|\cdot\|_2$ is the L2 normalization, k is the number x_m of elements contained, and $\mathbf{1}_k$ is a one vector of length k. The model is minimized, L_{Rec}, to ensure the integrity of the modality feature information during the encoding process.

3.2. Attention-Driven Double-Layer Graph Convolutional Network Module

Considering the globality and locality of label category correlation, this paper designs attention-driven two-layer convolutional networks to adaptively learn the correlation matrix to mine the dependencies between labels from the global perspective and sample-specific local perspective, respectively, for enhanced modality representation. In addition, the attention mechanism is introduced in the GCN to mine the correlation structure in the feature dimension and the label category dimension of specific samples and further enhance the modality representation. Otherwise, residual connections are added between the two-layer graph convolutions to prevent network degradation.

Since both the modality-common representation z^c and the modality-specific representation z_m^s need to be processed through the bilayer GCN without a loss of generality, the reference modality-common representation $z \in R^d$ and the modality-specific representation d are used as the number of dimensions of the modality representation.

3.2.1. Static Graph Convolutional Network

The first layer of an attention-driven bilayer graph convolutional network is a static graph convolutional network, its structure is presented in Figure 3. The modality repre-

sentation **Z** is first extended by a row to obtain the initial category representation matrix $\mathbf{V} \in R^{C\times d}$, where the jth$(j = 1, 2, \ldots, C)$ row of the matrix represents the class representation of the sample specific to the jth label. Accordingly, a global-based static graph is constructed, which has C nodes. The initial category representation matrix **V** is the node feature matrix of the graph, and its correlation matrix **V** characterization is based on the label dependence of the whole training dataset. Enter the static GCN to get the intermediate output **H**, which can be described by the formula:

$$\mathbf{H} = \text{LeakyReLU}(\mathbf{A}_s\mathbf{V}\mathbf{W}_s) \in R^{C\times d_1}, \tag{8}$$

where, LeakyReLU$(\cdot)$ is the nonlinear activation function, $\mathbf{A}_s \in R^{C\times C}$ is the correlation matrix of the static GCN, and $\mathbf{W}_s \in R^{d\times d_1}$ is the state update matrix of the static GCN, characterizing the linear transformation from dimension d to dimension d_1. Both are randomly initialized and updated by gradient descent during training. $\mathbf{A}_s$ is shared by all the training samples in the dataset and therefore able to capture global label correlation information.

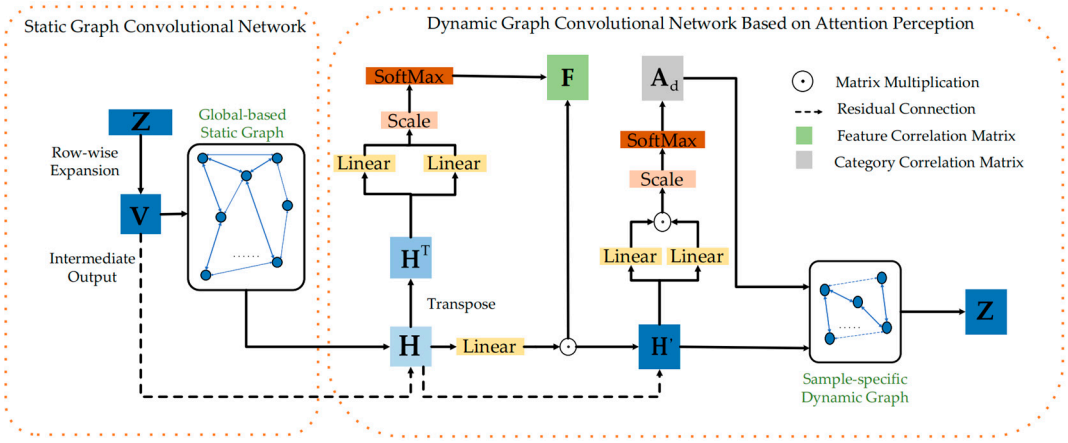


Figure 3. Framework structure diagram of attention-driven double-layer graph convolutional network module.

3.2.2. Dynamic Graph Convolutional Network Based on Attention Perception

The second layer of the attention-driven dynamic graph convolutional network is based on attentional perception, introducing attention mechanism dynamic capture specific to the sample characteristics of the dimension correlation structure and label category dimension correlation structure. The static GCN intermediate output **H** is further enhanced to obtain the enhanced category **Z**. The structure of dynamic GCN is presented in Figure 3.

Inspired by [31], the model introduces the attention mechanism to mine the correlation structure of the sample itself and dynamically generates the category correlation matrix $\mathbf{A}_d$ and feature correlation matrix **F** for specific samples. Firstly, the attention score calculation formula based on the scaling point product is given as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right), \tag{9}$$

where softmax is the nonlinear activation function, **Q** and **K** represents the query matrix and the key matrix, respectively, and d_k is the scale factor; its value should be the same as the number of dimensions of the key matrix **K**.

The calculation formula for defining the feature correlation matrix **F** is as follows:

$$\mathbf{F} = \text{softmax}\left(\frac{(\mathbf{U}_1\mathbf{H}^T)(\mathbf{U}_2\mathbf{H}^T)^T}{\sqrt{d_{k1}}}\right) \in R^{d_1\times d_1}, \tag{10}$$

where the scaling factors $d_{k1} = d_1$, $\mathbf{U}_1 \in R^{d_1 \times d_1}$, and $\mathbf{U}_2 \in R^{d_1 \times d_1}$ transform the transposed $\mathbf{H}^T \in R^{d_1 \times C}$ of the intermediate output into a mapping matrix of the query and bond matrix. The attention score matrix calculated by Formula (10) is used as the feature correlation matrix $\mathbf{F}$ to characterize the correlation structure of the sample feature dimensions.

Then, a specific sample-based dynamic graph is constructed, whose node feature matrix is the intermediate output $\mathbf{H}'$ of the feature correlation matrix $\mathbf{F}$ enhancement, and the attention mechanism is introduced to dynamically generate its correlation matrix $\mathbf{A}_d$, to characterize the label dependence based on the specific sample. Dynamic GCN takes $\mathbf{H}'$ as input and outputs further enhanced category representation $\mathbf{Z}$, which can be formally defined as follows:

$$\begin{cases} \mathbf{Z} = \text{LeakyReLU}(\mathbf{A}_d \mathbf{H}' \mathbf{W}_d) \in R^{C \times d_2} \\ \mathbf{A}_d = \text{softmax}\left(\frac{(\mathbf{W}_1 \mathbf{H}')(\mathbf{W}_2 \mathbf{H}')^T}{\sqrt{d_{k2}}} \right) \in R^{C \times C} \\ \mathbf{H}' = (\mathbf{U}_3 \mathbf{H}) \cdot \mathbf{F}^T \in R^{C \times d_1}, \end{cases} \tag{11}$$

which is the correlation matrix, $\mathbf{A}_d \in R^{C \times C}$, of the dynamic GCN, the state update matrix $\mathbf{W}_d \in R^{d_1 \times d_2}$, and the scaling factor $d_{k2} = C$. $\mathbf{W}_1 \in R^{C \times C}$ and $\mathbf{W}_2 \in R^{C \times C}$ are the mapping matrices that transform the intermediate output $\mathbf{H}'$ into a query matrix and a key matrix. $\mathbf{A}_d$ dynamically constructs based on the current input sample, which can better capture label dependency relationships specific to the current sample. $\mathbf{H}'$ represents the intermediate output, $\mathbf{H}$, of the dynamically adjusted feature correlation matrix $\mathbf{F}$, which enhances $\mathbf{H}$ by introducing the correlation information of samples in the feature dimension by $\mathbf{F}$. $\mathbf{U}_3 \in R^{C \times C}$ is a linear transformation matrix. Finally, residual connections are added between the static GCN and the dynamic GCN to prevent network degradation.

In short, the proposed attention-driven bilayer graph convolutional network works by introducing label category correlation information and feature correlation information.

3.3. Modality Representation Fusion with Multi-Label Classification

As shown in Figure 1, the modality-specific representation z_m^s and modality-common representation z^c are input into the attention-driven bilayer convolutional network to get the corresponding enhanced category representation matrix $\mathbf{Z}_c$, $\mathbf{Z}_v$, $\mathbf{Z}_t$ and $\mathbf{Z}_a$, and then weighted to get the final category representation:

$$\mathbf{Z}' = \mathbf{W}_c \odot \mathbf{Z}_c + \mathbf{W}_v \odot \mathbf{Z}_v + \mathbf{W}_t \odot \mathbf{Z}_t + \mathbf{W}_a \odot \mathbf{Z}_a, \tag{12}$$

where the Hadamard product $\odot$, namely, the corresponding elements in the matrix $\mathbf{W}_v$, $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_c \in R^{C \times d_2}$, are multiplied separately, is the fusion weight matrix of adaptive learning, representing the contribution degree of each modality representation and the modality-common representation in the corresponding label category.

Then, the category vector specific to each label in the final category representation matrix $\mathbf{Z}' = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_C]$ is put into the corresponding binary classifier $\mathbf{z}_j (j = 1, 2, \ldots, C)$ to predict the category score $\mathbf{s} = [s_1, s_2, \ldots, s_C]$, and the prediction score of each category is obtained. Thus, the calculation formula for classification loss is as follows:

$$L_{Cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_j^i \log\left(\delta\left(s_j^i\right)\right) + \left(1 - y_j^i\right) \log\left(1 - \delta\left(s_j^i\right)\right), \tag{13}$$

where it is the sigmoid activation function $\delta(\cdot)$, in which y_j^i is the true label i of the first sample. Taking 1 as meaning that the sample has the class label j, and taking 0 as the opposite, s_j^i is the prediction result of the label of the network model.

In conclusion, the combination of Formulas (4), (5), (7), and (13) can obtain the overall loss function of the proposed BGCLN model as follows:

$$L = L_{Cls} + \alpha L_{Rec} + \beta L_{Ort} + \gamma L_{Adv}, \tag{14}$$

where α, β, and γ are the trade-off parameters balancing different loss contributions, and the pseudo-code of the training process of the proposed BGCLN model is shown in Algorithm 1.

Algorithm 1 Model training process—BGCLN training process

Data input:
$\{x_v, x_t, x_a\}$: visual features, trajectory features, and acoustic features of the micro-video;
y: the real category label vector of the micro-video;
$\alpha = 0.1, \beta = 0.05, \gamma = 0.05$: the term coefficient of the loss function;
1: Randomly initialize all network parameters;
2: Repeat;
3: For $i = 1, 2, L,$ epoch do;
4: Use Formula (1) to calculate the specific representation z_m^s of each mode;
5: Use Formulas (2) and (3) to calculate the public representation z^c of each mode;
6: Use Formula (6) to calculate the reconstructed vector $\hat{x}_m$;
7: Use Formula (8) to update the category to represent **V** the intermediate output **H**;
8: Use Formula (10) to calculate the feature correlation matrix **H** according to the intermediate output **F**;
9: Calculate the enhanced category representation $\mathbf{V}'$ using Formula (11);
10: Use Formula (12) to calculate the fused category representation **Z**;
11: Update all network parameters using the stochastic gradient descent method under Formula (14);
12: End for;
13: Until convergence.
Data output: all network training parameters $\theta_m, \theta_c, \theta_d, \theta_s$, etc.

4. Experimental Simulation

4.1. Dataset and Experimental Settings

To facilitate the development of tracking algorithms well-suited for underwater environments and address the lack of existing underwater visual datasets, Panetta et al. proposed the first comprehensive underwater object tracking (UOT100) benchmark dataset [32,33]. This dataset consists of 104 underwater video sequences and over 74,000 annotated frames, which are derived from both natural and artificial underwater videos, with a variety of distortions. The UOT100 dataset accessed at the following URL: https://www.kaggle.com/datasets/landrykezebou/uot100-underwater-object-tracking-dataset, accessed on 23 January 2024.

Given the scarcity of multi-label classification datasets for underwater micro-videos [22], this paper combines the UOT100 dataset and relevant underwater content from the MLSV2018 dataset; by processing and relabeling these micro-videos, an underwater micro-video multi-label classification dataset (UVMCD) was established, as shown in Figure 4. The MLSV2018 dataset accessed on 20 January 2024 at the following URL: https://github.com/tjufan/challengerai-mlsv2018. All the experiments described in this paper were conducted on the UVMCD, a large-scale multi-label classification dataset for micro-video, in order to verify BGCLN.

The multi-label classification dataset of underwater micro-videos is composed of 3841 underwater micro-videos and their corresponding audios, and each micro-video has a corresponding category label. The dataset has 19 label categories, with 1–5 labels per micro-video. The number and proportion of labels in the multi-label classification dataset of underwater micro-videos are shown in Table 1. The label distribution of the underwater micro-video multi-label classification dataset is shown in Figure 5. In our experiment, 80% of the data is used to train, and the remaining 20% is used to evaluate.

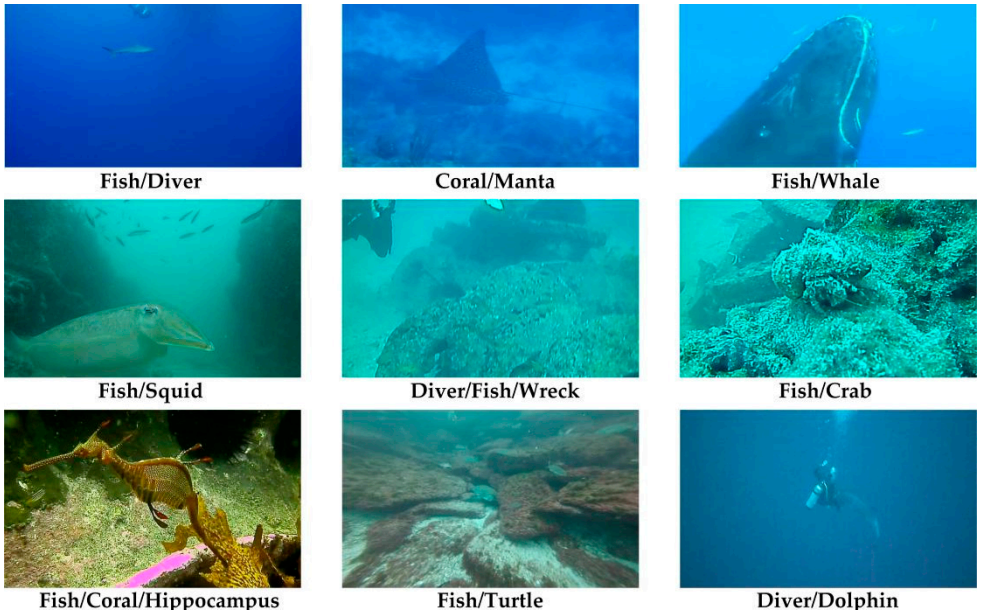


Figure 4. Sample diagram of UVMCD.

Table 1. Number and proportion of labels in multi-label classification dataset.

Category	Number	Proportion	Category	Number	Proportion
Diver	2278	59.31%	Seal	62	1.61%
Fish	1995	51.94%	Lobster	58	1.51%
Coralline	760	19.79%	Squid	56	1.46%
Manta	315	8.20%	Crab	40	1.04%
Person	256	6.66%	Medusa	39	1.02%
Wreckage	223	5.81%	Dolphin	31	0.81%
Others	173	4.5%	Whale	26	0.68%
Tortoise	167	4.35%	Sea slug	21	0.55%
Eel	143	3.72%	Seahorse	14	0.36%
Octopus	105	2.73%			

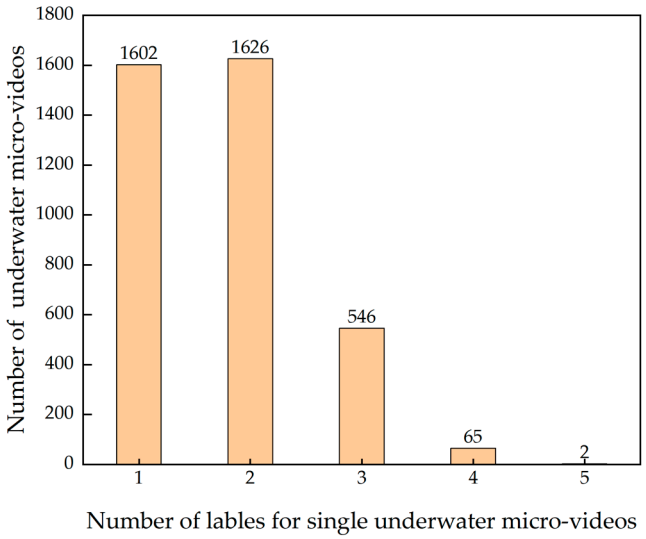


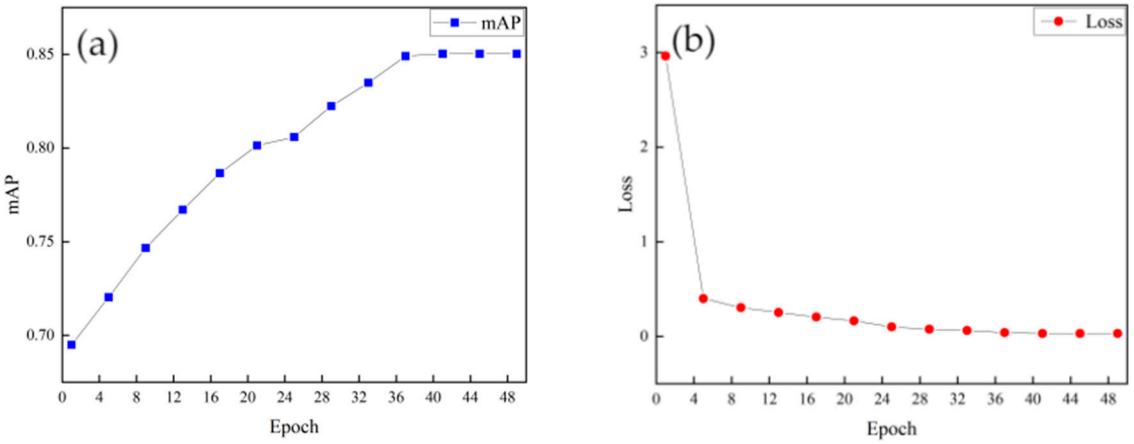
Figure 5. Label distribution of multi-label classification dataset.

4.2. Performance Evaluation

The performance of the algorithm is evaluated by five evaluation indexes [34]: mean average precision (mAP), Hamming loss, ranking loss, coverage, and one error.

4.2.1. Convergence Analysis

To analyze the convergence of the models in this chapter, the experimental results of the average precision versus the number of model iterations and the classification loss versus the number of model iterations were tested, which are shown in Figures 6a and 6b, respectively. From the figures, it can be observed that the average precision increases as the number of iterations increases and stabilizes at the optimal average precision when the number of iterations is 40. Meanwhile, the classification loss decreases as the number of iterations increases and stabilizes eventually.



(**a**) (**b**)

Figure 6. (**a**) The curve of the mean average precision changes with the epoch. (**b**) The classification loss curve changes with the epoch.

4.2.2. Ablation Experiments and Analysis

Ablation experiments, including model performance evaluation with different modes and different module configurations, were conducted. Table 2 shows the performance comparison of the different schemes on the classification results.

Table 2. Performance comparison of different schemes on classification results.

	mAP↑	One Error↓	Coverage↓	Ranking Loss↓	Hamming Loss↓
Visual modality	0.7212	0.0416	0.0906	33.9567	0.0734
Audio modality	0.7536	0.0395	0.0705	28.7821	0.0682
Visual modality+ Audio modality	0.8183	0.0301	0.0675	16.4535	0.0537
Visual modality+ Graph learning module	0.7731	0.0351	0.0786	30.1439	0.0693
Visual modality+ Graph learning module	0.8007	0.0262	0.0591	25.1644	0.0646
Ours	**0.8503**	**0.0224**	**0.0472**	**14.8036**	**0.0393**

In Table 2, ↑ indicates that a higher value of the metric corresponds to better model performance, ↓ indicates that a lower value of the metric corresponds to better model performance. It can be seen that the multi-modal fusion method outperforms the unimodal method, which indicates that the complementarity between different modalities can be effectively utilized by integrating the multi-modal fusion to promote compatibility modeling. Meanwhile, each individual modality can have a different degree of positive impact

on the model performance. The audio modality is superior to the visual modality, which means that the audio modality contains more valuable information about underwater categories than the visual modality. Whether in the unimodal or multi-modal case, the graph associative learning module plays an important role in the framework proposed in this chapter, which proves the necessity of the graph associative learning module by learning the semantic association representation between labels to better complete the multi-label classification task for underwater micro-videos.

4.2.3. Model Validity Analysis

To illustrate the validity of the methods presented in this chapter, comparisons are made with the following different types of methods and all the experiments using the same training and test sets. Table 3 compares the performance of the different methods on the UVMCD dataset.

Table 3. Contrast classification performance.

Method	mAP↑	One Error↓	Coverage↓	Ranking Loss↓	Hamming Loss↓
GoolgeNet	0.7213	0.0389	0.0917	34.4535	0.0605
C3D	0.7215	0.0391	0.0892	33.4416	0.0621
MLKNN	0.7446	0.0381	0.0744	0.3414	0.0559
GlOCAL	0.7020	0.0415	0.1188	0.4930	0.0697
SIMM	0.7258	0.0423	0.0717	0.3149	0.0782
TM3L	0.7598	0.0380	0.0501	**0.2298**	0.0401
MANET	0.8019	0.0291	0.0591	22.9874	0.0485
BGCLN	**0.8503**	**0.0224**	**0.0472**	14.8036	**0.0393**

In Table 3, it can be observed that deep representation-based methods, namely GoogleNet and C3D, typically model only a single modality and lack semantic correlation modules, and are also sensitive to lighting, color cast, and water flow disturbances in underwater micro-videos, resulting in unsatisfactory performance. In addition, the results of the four multi-label learning methods, MLKNN, GLOCAL, SIMM, and TM3L, are not satisfactory, but they perform better than the baseline methods in the coverage metric, which indicates that multi-label learning methods can better handle the correlation between labels and have lower complexity compared to multi-modal semantic enhancement methods. Finally, the multi-label classification method based on multi-modal semantic enhancement achieved relatively promising results, reflecting the positive role of multi-modal fusion semantic enhancement in multi-label classification tasks.

5. Summary

In this paper, a bilayer graph convolution learning network based on a constrained codec (BGCLN) is proposed. First, considering the consistency and complementarity of multi-modal information, learning modality-specific and modality-common representations through codec networks with constraints is constructed. Secondly, the correlation information between the labels from both global and local perspectives is mined through the attention-driven bilayer graph convolutional network, while introducing the attention mechanism to explore the correlation structure of the samples in the label dimension and feature dimensions in the graph convolution to enhance the modality representation. Finally, the enhanced public representation and each modality-specific representation are weighted, fused, and input into the classifier to complete the multi-label classification task. Based on the large-scale multi-label micro-video dataset UVMCD, a series of experiments show that the proposed BGCLN model has better achievement in parameter sensitivity analysis, module ablation analysis, modality combination analysis, and other aspects. Meanwhile, compared with other models, BGCLN has better classification performance to verify its effectiveness.

Author Contributions: Conceptualization, Y.L. and S.W.; methodology, Y.L.; software, S.W.; validation, S.W., J.M. and X.W.; formal analysis, S.W.; investigation, X.W.; resources, J.M.; data curation, X.W.; writing—original draft preparation, S.W.; writing—review and editing, Y.L. and S.W.; visualization, J.M.; supervision, Y.L.; project administration, Y.L.; funding acquisition, Y.L. 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 (No. 62360102) and the Intelligent Gateway for Data Exchange in the Lijiang River Basin, integrating the Beidou Navigation System with the Water Network (No. 20221B074250).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: Author Su Wang was employed by the company China Mobile Communications Group Jiangsu Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Nie, L.; Wang, X.; Zhang, J.; He, X.; Zhang, H.; Hong, R.; Tian, Q. Enhancing micro-video understanding by harnessing external sounds. In Proceedings of the ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1192–1200.
2. Chang, X.; Yu, Y.L.; Yang, Y. Semantic pooling for complex event analysis in untrimmed videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1617–1632. [CrossRef] [PubMed]
3. Chen, J.; Song, X.; Nie, L.; Wang, X.; Zhang, H.; Chua, T.S. Micro tells macro: Predicting the popularity of microvideos via a transductive model. In Proceedings of the ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 898–907.
4. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the ACM Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
5. Li, Y.; Sun, S.; Huang, Q.; Jing, P. Underwater Image Enhancement Network Based on Multi-channel Hybrid Attention Mechanism. *J. Electron. Inf. Technol.* **2017**, *46*, 118–128.
6. Zou, Y.; Chou, C.A. A combinatorial optimization approach for multi-label associative classification. *Knowl.-Based Syst.* **2022**, *240*, 108088. [CrossRef]
7. Pham, T.; Tao, X.; Zhang, J.; Yong, J.; Li, Y.; Xie, H. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowl.-Based Syst.* **2022**, *235*, 107662. [CrossRef]
8. Chen, Z.; Ke, H.; Xu, J.; Peng, T.; Yang, C. Multichannel Domain Adaptation Graph Convolutional Networks-Based Fault Diagnosis Method and With Its Application. *IEEE Trans. Ind. Inform.* **2023**, *19*, 7790–7800. [CrossRef]
9. Chen, L.; Zhan, W.; Tian, W.; He, Y.; Zou, Q. Deep integration: A multi-label architecture for road scene recognition. *IEEE Trans. Image Process.* **2019**, *28*, 4883–4898. [CrossRef] [PubMed]
10. Zhu, L.; Zhou, Y.; Jia, R.; Gu, W.; Luan, T.H.; Li, M. Real-Time Fault Diagnosis for EVs With Multilabel Feature Selection and Sliding Window Control. *IEEE Internet Things J.* **2022**, *9*, 18346–18359. [CrossRef]
11. Boutell, M.R.; Luo, J.; Shen, X.; Brown, C.M. Learning multi-label scene classification. *Pattern Recognit.* **2004**, *37*, 1757–1771. [CrossRef]
12. Yeh, C.K.; Wu, W.C.; Ko, W.J.; Wang, Y.C.F. Learning deep latent space for multi-label classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada, 4–9 February 2017; Volume 31, pp. 2838–2844.
13. Fei, H.; Zhang, Y.; Ren, Y.; Ji, D. Latent emotion memory for multi-label emotion classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7692–7699.
14. Rajagopalan, S.S.; Morency, L.P.; Baltrusaitis, T.; Goecke, R. Extending long short-term memory for multi-view structured learning. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 338–353.
15. Fan, W.; Ma, Y.; Xu, H.; Liu, X.; Wang, J.; Li, Q.; Tang, J. Deep adversarial canonical correlation analysis. In Proceedings of the Society for Industrial and Applied Mathematics International Conference on Data Mining, Cincinnati, OH, USA, 7–9 May 2020; pp. 352–360.
16. Srivastava, N.; Salakhutdinov, R.R. Multimodal learning with deep boltzmann machines. *Adv. Neural Inf. Process. Syst.* **2012**, *25*.
17. Gao, X.; Zhang, G.; Xiong, Y. Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel. *Measurement* **2022**, *194*, 111001. [CrossRef]
18. Iyer, G.; Chanussot, J.; Bertozzi, A.L. A graph-based approach for data fusion and segmentation of multimodal images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4419–4429. [CrossRef]
19. Chandar, S.; Khapra, M.M.; Larochelle, H.; Ravindran, B. Correlational neural networks. *Neural Comput.* **2016**, *28*, 257–285. [CrossRef] [PubMed]
20. Shen, T.; Jia, J.; Li, Y.; Wang, H.; Chen, B. Enhancing music recommendation with social media content: An attentive multimodal autoencoder approach. In Proceedings of the International Joint Conference on Neural Networks, Glasgow, UK, 19–24 July 2020; pp. 1–8.

21. Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; Lu, H. Normalized and geometry-aware self-attention network for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10327–10336.
22. Azad, M.A.; Mohammed, A.; Waszak, M.; Elvesæter, B.; Ludvigsen, M. Multi-label Video Classification for Underwater Ship Inspection. In Proceedings of the OCEANS, Limerick, Irish, 5–8 June 2023.
23. Sun, Q.; Wang, K. Underwater single-channel acoustic signal multitarget recognition using convolutional neural networks. *The J. Acoust. Soc. Am.* **2022**, *151*, 2245–2254. [CrossRef] [PubMed]
24. Le, H.D.; Lee, G.S.; Kim, S.H.; Kim, S.; Yang, H.J. Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning. *IEEE Access* **2023**, *11*, 14742–14751. [CrossRef]
25. Cai, Z.; Ding, H.; Wu, J.; Xi, Y.; Wu, X.; Cui, X. Multi-label movie genre classification based on multimodal fusion. *Multimed. Tools Appl.* **2023**, *83*, 36823–36840. [CrossRef]
26. Liu, P.; Qiu, X.; Huang, X. Adversarial Multi-task Learning for Text Classification. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1–10.
27. Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; Erhan, D. Domain separation networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 343–351.
28. Duan, Y.; Zheng, W.; Lin, X.; Lu, J.; Zhou, J. Deep adversarial metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2780–2789.
29. Chen, S.; Gong, C.; Yang, J.; Li, X.; Wei, Y.; Li, J. Adversarial metric learning. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 2021–2027.
30. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2130.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
32. Kezebou, L.; Oludare, V.; Panetta, K.; Again, S.S. Underwater Object Tracking Benchmark and Dataset. In Proceedings of the IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, MA, USA, 5–6 November 2019; pp. 1–6.
33. Panetta, K.; Kezebou, L.; Oludare, V.; Again, S.S. Comprehensive Underwater Object Tracking Benchmark Dataset and Underwater Image Enhancement with GAN. *IEEE J. Ocean. Eng.* **2022**, *47*, 59–75. [CrossRef]
34. Zhang, M.L.; Zhou, Z.H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 1819–1837. [CrossRef]