This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

Improving the Generalization of Visual Classification Models Across IoT Cameras via Cross-modal Inference and Fusion

Qing-Ling Guan, Yuze Zheng, Lei Meng, Li-Quan Dong and Qun Hao

***Abstract*—The performance of visual classification models across IoT devices is usually limited by the changes in local environments, resulted from the diverse appearances of the target objects and differences in light conditions and background scenes. To alleviate these problems, existing studies usually introduce the multimodal information to guide the learning process of the visual classification models, making the models extract the visual features from the discriminative image regions. Especially, cross-modal alignment between visual and textual features has been considered as an effective way for this task by learning a domain-consistent latent feature space for the visual and semantic features. However, this approach may suffer from the heterogeneity between multiple modalities, such as the multi-modal features and the differences in the learned feature values. To alleviate this problem, this paper first presents a comparative analysis of the functionality of various alignment strategies and their impacts on improving visual classification. Subsequently, a cross-modal inference and fusion framework (termed as CRIF) is proposed to align the heterogeneous features in both the feature distributions and values. More importantly, CRIF includes a cross-modal information enrichment module to improve the final classification and learn the mappings from the visual to the semantic space. We conduct experiments on four benchmarking datasets, i.e. the Vireo-Food172, NUS-WIDE, MSR-VTT, and ActivityNet Captions datasets. We report state-of-the-art results for basic classification tasks on the four datasets and conduct subsequent experiments on feature alignment and fusion. The experimental results verify that CRIF can effectively improve the learning ability of the visual classification models, and it is a model-agnostic framework that consistently improves the performance of state-of-the-art visual classification models.**

***Index Terms*—heterogeneous domain, image classification, feature alignment, semantic inference**

I. Introduction

Image classification is an essential task in computer vision applications, and research on it has been widely applied to the downstream visual understanding tasks, such as face recognition and intelligent robots [1], [2]. Nowadays, state-of-the-art visual classification methods mainly use the deep neural networks to fit the image data for downstream tasks and usually achieve impressive results. However, this may not happen in the IoT domain due to the complicated outside scenarios. First, in the outside environments, images need to be acquired from a variety of video cameras. This may result in the various light and color conditions of the captured images, which results in the visual diversity of the images belonging to the same class. The above reasons may lead to a decrease in the performance of traditional image classification models in practical [3]. The use of multimedia information to guide the learning of image classification typically explores the descriptive text of images which contains more discriminative information for classification [4]. It is worth mentioning that, in practice, the tagging information for camera photos is usually not available. As such, the Learning Using Privileged Information (LUPI) paradigm that only uses text in the training phase has been proposed [5]–[8]. However, due to the problem of modal heterogeneity, i.e., the different feature distribution of images and texts in the feature space, the performance gains achieved by cross-modal learning is limited.

In recent years, researchers have tried to alleviate the problem caused by heterogeneous modalities by imposing operations on the features of different modalities in the feature space, and a variety of feature alignment methods have been proposed. The existing cross-modal alignment algorithms try to shorten the distance between features by constraining the distance of cross-modal features in the latent space [9]–[11], or make the distribution of features as similar as possible by reconstructing the features [12]–[14] to improve the ability of multimodal representation learning. However, since various feature alignment methods are different in the methods of shared space mapping [6], [13], [15], [16], the selection of the distance measurement [17], [18], etc., there is still lacking a summary of how aforementioned factors affect the alignment, which further hinders the mitigation of heterogeneity in the alignment of different modalities, and also limits the performance improvement of cross-modal enhancement methods.

To address the above problems, we first conducted a comparative study of existing feature alignment methods and analyzed the effects of changing different key factors in alignment experimentally to find a more efficient alignment method. Based on the above studies, we propose a cross-modal enhanced image classification framework, termed CRIF, which introduces text information as privileged information to enhance the representation learning of visual modality. The basic idea of CRIF is shown in Fig. 1, in contrast to

Lei Meng is the corresponding author.

Qing-Ling Quan and Li-Quan Dong are with School of Optics and Photonics, Beijing Key Lab Precis Optoelect Measurement Institute, Beijing Institute of Technology, China; They are also with Yangtze River Delta Research Institute, Beijing Institute of Technology (Jiaxing), China. Email:guanql@bit.edu.cn, kylind@bit.edu.cn.

Qun Hao is with School of Optics and Photonics, Beijing Key Lab Precis Optoelect Measurement Institute, Beijing Institute of Technology, China. Email: qhao@bit.edu.cn.

Yuze Zheng is with School of Software, Shandong University, China. Email: zhengyuze@mail.sdu.edu.cn.

Lei Meng is with Shandong University and Shandong Research Institute of Industrial Technology, China. Email: lmeng@sdu.edu.cn.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

previous cross-modal alignment methods, CRIF adopts a two-level feature alignment, i.e., the distributional alignment and the content alignment, to first make the distribution of visual features similar to that in semantic modality, and then a content-level alignment learns a semantic-consistent feature representation in the shared space, which makes the visual and textual features of the same sample to be closer. Since visual noise causes error porpagation during feature alignment, CRIF further filters the visual noise by cross-modal inference, which maps the visual features to the semantic space. The final classification is performed using the fused features of the aligned and cross-modal inferred ones.

Experiments are conducted on four real-world cross-modal datasets, i.e. Vireo-Food172 [19], NUS-WIDE [20], MSR-VTT [21], ActivityNet Captions [22]. Among them, VireoFood-172 and NUS-WIDE datasets for image classification tasks contain image-text data pairs, and in particular, NUS-WIDE is a multi-classification dataset. The other two datasets, MSR-VTT and ActivityNet Captions, are used for video classification tasks. MSR-VTT contains video-description data pairs, and ActivityNet Captions contains video-audio data pairs. The experimental results show that compared with other cross-modal alignment methods, CRIF achieves stable performance improvement through the combination of two-level alignment and cross-modal inference, and CRIF is a model-independent framework that can make visual features to be closer to its semantic counterpart in the feature space than the conventional method. Through case studies, we found that cross-modal alignment can enable visual features to effectively learn semantic feature distribution, and feature alignment can effectively enhance visual representation learning under different tasks and different backbones. In summary, the contributions of this paper mainly include:

- We propose a multi-modal image classification framework, namely CRIF, which utilizes the multi-modal information to enhance the anti-interference ability of the classification model for visual noise. This alleviates the semantic gap between image contents and semantic information, and can be easily incorporated into common image classification methods to improve their accuracy.
- We propose a two-step alignment strategy to progressively alleviate the impact of feature heterogeneity on feature alignment. At the same time, privileged information can be used to enhance the quality of propagation presentation. This method leads to an incremental performance improvement over previous alignment methods, which is meaningful for multi-modal representation learning.
- We discuss a variety of cross-modal feature alignment methods and multi-modal feature fusion strategies, which can provide insights for future cross-modal and multi-modal research.

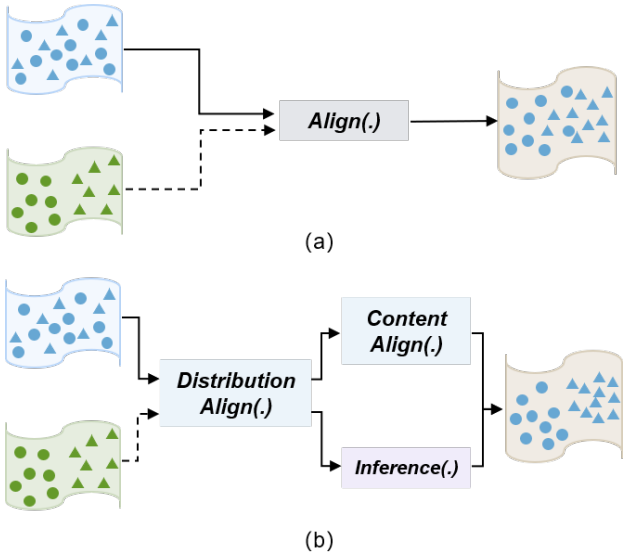

Fig. 1. Illustration of the common feature alignment method and the proposed CRIF, and the dashed line represents the text information used only in the training. In the figure, (a) represents the traditional feature alignment process; and (b) is the CRIF framework, CRIF alleviates heterogeneous by two-level feature alignment, mines hidden text information in visual modalities, and fuses above information to enhance image classification learning.

II. Related Work

A. Cross-camera Visual Classification

In recent years, with the rise of neural networks, the Internet of Thof ings (IoT) is widely combined with deep learning due to its accuracy and convenience. People use IoT devices such as IoT cameras to obtain real-world data and analyze them through deep learning models. It has achieved extensive development in medical services [23], urban governance [24], [25], video surveillance [26], and other aspects [27], [28]. For example, in wild conservation, Zualkernan [29] has developed an IoT-based system that uses deep learning and edge analytics to automatically classify large collections of images taken by specialized cameras used by ecologists in the field. And send the information to the relevant departments.

However, since IoT cameras are shot in real-time and different cameras get variant parameter settings, decoration positions, and lighting during shooting, the obtained images will be disturbed by similar factors, which will lead to deviations in the learning of the model. At the same time, the pictures taken by IoT cameras usually have complex environments and category diversity, which leads to poor performance of the model in practical applications. Therefore, in order to improve the performance of deep learning models in the face of complex data obtained by IoT cameras, people use cross-modal learning [30]–[32] to train models better than image-only learning [33]–[35]. RGSP [31] identified the same person using the daytime visible modality and nighttime thermal modality captured by IoT cameras. It makes the model more robust to color changes through the data enhancement method of Random Gray, and uses the softpooling strategies to learn more features that can be used to identify people. MAA-Net [32] uses image attention and natural language description for better person search. It uses the attention mechanism and multimodal alignment method to bridge the semantic gap between the visual mode and the text mode described by natural language. RRL-GAT [36] adopts the two graph convolution modules to effectively reduce false connections between data objects and reduce the impact of noisy connections in complex images.

B. Multi-modal Visual Classification

Visual classification has always been at the core of deep learning. With the iterative update of the neural network, the performances of visual representation learning have also been significantly improved. Traditional visual classification tasks use a number of single-modal data to train models. Many studies use convolutional neural networks to extract image features, and time series modeling is used to extract video frames for classification [37]–[40]. With the rapid development of digital media and the explosion of information, a amount of multimodal data has emerged, such as pictures and words in social media, images and captions in video, and a variety of images taken by satellites(HS, MS, LiDAR, etc.), and so on. Related studies begin to use multimodal data for visual representation learning.

Traditional multi-modal visual representation learning inputs multi-modal information into the training and testing stages of the model, and the performance of the model has been significantly improved compared with the single-modal learning method [41]–[43]. Yu [44] focuses on subspace learning and proposes label graph to efficiently utilize semi-supervised data, maintain semantic consistency between visual modes and text modes in subspace, and ensure geometric consistency of multi-modal features in subspace. In the follow-up study, Yu [45] proposed the end-to-end model DDCH, which regarded the labels with rich information as bridging modes, bridging the semantic gap between images and text and increasing the relevance of multi-modal information. COCO [46] takes images and text as input data, and uses the data potential within a modality in a self-supervised manner, while performing contrastive learning between modalities to generate high-quality visual representations. CCR-Net [47] takes two kinds of satellite images of the same area as input(HS and LiDAR), and makes the visual information cross-modal interaction by reconstructing the features, then fuses the features and classifies them.

However, this multi-modal learning method is very strict for data requirements, and the cost of screening out matching data pairs and labeling them correctly is very expensive. Therefore, a new learning method of multi-modal visual representation learning has emerged: LUPI(learning using privileged information) [5]. LUPI takes multi-modal information as input in the training phase, and the better-performing modality information is regarded as privileged information to guide the learning of the other modality information. In the testing phase, only single-modal information is used to detect the model performance. Garcia [48] improved an illusion network, which used depth pictures as privileged information to guide RGB pictures, and enabled the model to encode monocular depth features from RGB frames in the test phase by adversarial learning. In classification tasks, text information is easier to recognize due to its features [4], and related studies often use text information as privileged information to guide visual representation learning. Yan [7] adopts the active learning method, takes text features as privileged information, and makes active sample selection of visual features and text features to train the model's ability to extract visual features. Yao [8] extracts privileged information from the unlabeled corpus and denoises it, guiding the model to train multiple sub-classifiers according to the category to improve the robustness of the model.

C. Cross-Modal Alignment for Enhancing Visual Representation Learning

In this paper, we propose a cross-modal feature alignment framework using the neural network as encoder and decoder. The existing heterogeneous feature alignment models are mainly divided into two categories: the distribution alignment model and the feature alignment model. Our proposed framework belongs to the second category.

For the distribution alignment model, the current research mostly uses the generation class framework to realize the alignment, and uses GAN/VAE to learn the distribution of features respectively [49], [50], and generates the aligned features by making the distribution of heterogeneous features close to each other. Zhu built a GAN architecture with one generator and two discriminators to enhance previously extracted features [12]. The generator is used to reconstruct recipe or image features, and the discriminator is used to distinguish generated images and reconstructed image features, respectively. The overall architecture enables bidirectional image-recipe retrieval. Thomas uses three pairs of VAE to realize cross-modal distribution alignment, two pairs of VAE to reconstruct image/text features into text features, and cross-modal VAE to align the distribution between image features and text features [13]. Wan uses VAE and GAN to construct the latent space of each mode, and then maps the uniform distribution of GAN to the normal distribution of VAE through the network to achieve alignment [14].

For the feature alignment model, the current research mostly measures the distance between heterogeneous features and reduces the distance to narrow the features from different modes in hidden space. Sun extracts the features of the source domain and the target domain and then calculates the distance using the Frobenius norm between the features as a loss to minimize the second-categoriesstics between them to achieve alignment [17]. Sun improves the previous method by calculating the covariance matrix of the prediction results as the distance measure after the two modes are predicted respectively [51]. On the basis of realizing the feature alignment, Sun also avoids model overfitting to the source domain data as much as possible [51]. Li extracts features by a two-channel encoder and calculates the centroid coordinates of the source domain and the target domain [18]. By reducing the centroid distance and mining the hidden information in the source domain, Li further assists the target domain to align [18].

Based on the alignment of the image features in the source domain and the text features in the target domain, our framework further reconstructs the source domain features into the target domain features and fuses them with the aligned source domain features. This method alleviates the semantic gap between image and text and reduces the degree of internal heterogeneity between different domains.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

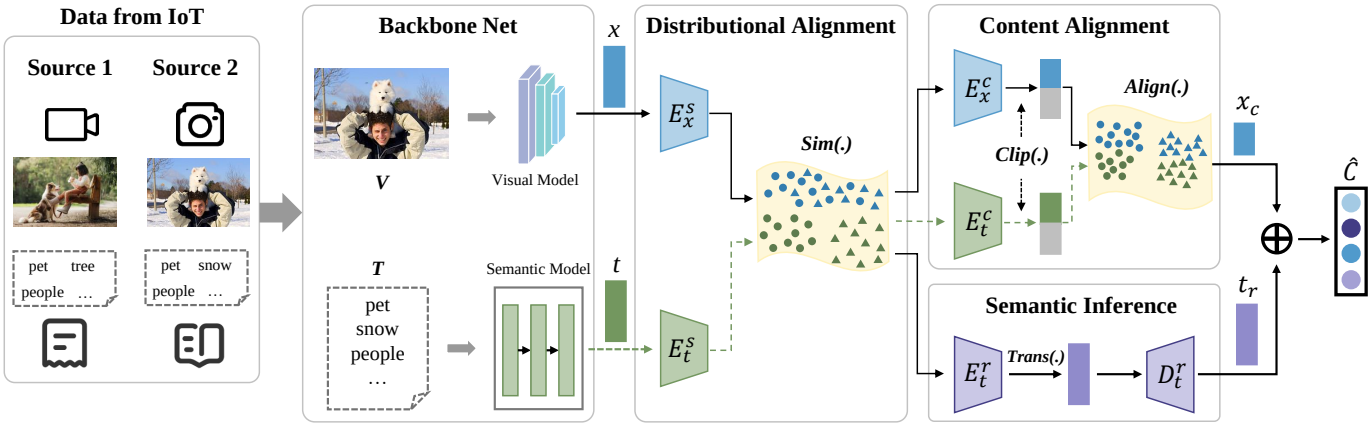

Fig. 2. Framework of CRIF. CRIF receives visual and semantic data from the IoT and extracts visual representation $\mathbf{x}$ and semantic representation $\mathbf{t}$. $Sim(.)$ represents initial heterogeneous feature alignment, $Clip(.)$ represents local feature generation, $Align(.)$ represents local heterogeneous feature alignment, $Tran(.)$ represents cross-modal mapping, and $\hat{\mathbf{C}}$ represents the final model classification result. After feature alignment and cross-modal mapping, $\mathbf{x_c}$ and $\mathbf{t_r}$ are obtained from image representation $\mathbf{x}$. CRIF fuses the generated cross-modal features for classification.

III. Method

A. Overall Framework

The proposed CRIF framework consists of three main modules: the Distributional Alignment Module, the Content Alignment Module, and the Cross-Modal Feature Fusion Module, as shown in Figure 2.

In the initial stage, backbone models extract visual features $\mathbf{x}$ and text features $\mathbf{t}$ from the visual input V and semantic input T respectively. Then the Distributional Alignment Module maps the multi-modal features $\mathbf{x}$ and $\mathbf{t}$ into the shared latent space and performs feature alignment to make their distributions similar. To further narrow the distance between modalities, the Content Alignment Module extracts part of the original features to form $\mathbf{x_c}$ and $\mathbf{t_c}$ that have the same target to represent information of classification and aligns them to alleviate the heterogeneous. In addition to filtering the noise from visual modality, the Semantic Inference Module maps visual features $\mathbf{x_a}$ into the semantic space and reconstructs them into semantic features $\mathbf{t_r}$, so as to mine the hidden semantic information in the visual modality. Finally, the aligned feature $\mathbf{x_c}$ and cross-modal inferred feature $\mathbf{t_r}$ are fused to combine advantages across modalities. The above process can be described as information flows:

$$\mathbf{x} \xrightarrow{E_x^s} \mathbf{x_a} \xrightarrow{E_x^c} \mathbf{x_c} \tag{1}$$

$$\mathbf{x} \xrightarrow{E_x^s} \mathbf{x_a} \xrightarrow{E_t^r D_t^r} \mathbf{t_r} \tag{2}$$

$$\hat{C} = f(\mathbf{x_c} \oplus \mathbf{t_r}) \tag{3}$$

where E_x^s is the mapping of shared space in distributional alignment, E_x^s is the mapping of common space for visual modality in content alignment, E_t^r is the cross-modal inferred mapping, and D_t^r is the feature decoding from inferred feature to text feature. In Equation 3, $\oplus$ is the fusion operation of features, and f represents the process by which the model makes a prediction on the fused features, $\hat{C}$ denotes the result of the model prediction.

In summary, the CRIF framework uses a two-level alignment between the visual modality and semantic modality, and performs visual cross-modal reconstruction features, which alleviate the semantic gap between vision and text, and reduce the impact of visual noise on the model, altogether improving the robustness of the model.

B. Distributional Alignment Module for Identifying Features with Similar Correlations in Visual-Semantic Domain

This module maps features from different modalities into a shared latent feature space and improves the distributional similarity, resulting in the formed aligned features being better separable in the latent space. Initially, the visual encoder $E_v(.)$ and semantic encoder $E_t(.)$ extract the original visual features $\mathbf{x}$ and semantic features $\mathbf{t}$ from the visual inputs V from inputs T respectively:

$$\mathbf{x} = E_v(V) \tag{4}$$

$$\mathbf{t} = E_t(T) \tag{5}$$

Previous methods that directly impose distance metric constraints for alignment ignore the structural differences of features from different modalities, so it is necessary to first map heterogeneous features into the same feature latent space to alleviate the inhibitory effect of different distributions between features on alignment.

To achieve this, features of both modalities are mapped to a shared feature space for further alignment:

$$\mathbf{x_a} = E_x^s(\mathbf{x}) \tag{6}$$

$$\mathbf{t_a} = E_t^s(\mathbf{t}) \tag{7}$$

where $E_x^s(.)$ and $E_t^s(.)$ are space mapping for visual and semantic features.

C. Content Alignment Module for Learning Shared Space for Heterogeneous Features with Smallest Cross-Modal Distance

This module guides the model to disentangle key features associated with the classification and achieve visual and semantic alignment. The module contains two parts of constraints: local classification prediction and feature alignment.

We use the encoder to further extract the features in the distributional alignment latent space, trying to obtain the part of the features that contribute more to the classification task, that is, the key part of the features:

$$\mathbf{x_c} = \mathrm{Clip}(\mathbf{x_a}, \rho) \tag{8}$$

$$\mathbf{t_c} = \mathrm{Clip}(\mathbf{t_a}, \rho) \tag{9}$$

where $Clip(.)$ is the operation to obtain key parts of features, and ρ is the dimension of features.

After feature extraction, we reduce the background noise part in the visual features, and also reduce the redundant information in the text features. For the obtained key features, we impose a constraint composed of Deep Coral [51] between the features of the two modalities for content alignment $Align(.)$, so as to reduce the inherent heterogeneity between modalities. The formula is as follows:

$$\mathcal{L}_{align} = \frac{1}{4d^2} \left\| \mathbf{x_c} - \mathbf{t_c} \right\|_F^2, \tag{10}$$

where $\| \cdot \|_F^2$ denotes the squared matrix Frobenius norm. $\mathcal{L}_{align}$ minimizes the difference between the characteristics of the two parts.

In addition, in order to better improve the feature extraction and selection of the model, we also introduce the classification loss in the content alignment module. For single-label and multi-label datasets, we employ cross-entropy (CE) and binary cross-entropy (BCE) as pair constraints to calculate the classification loss, respectively.

Finally, the content alignment module will output the aligned key part feature $\mathbf{x_c}$ for the next feature fusion.

D. Semantic Inference Module for Learning Visual-to-Semantic Information Mappings

This module mines the hidden text information in the visual features and reconstructs them into text features. This module mainly uses the two-part loss of feature similarity $\mathcal{L}_{Sim}$ and feature reconstruction $\mathcal{L}_{Recon}$ for constraint.

Firstly, for the original features $\mathbf{x}$ and $\mathbf{t}$ extracted by the base network, due to the difference in feature dimensions, we project them into the common latent space by linear mapping, denoted as $\mathbf{x_a}$ and $\mathbf{t_a}$. Subsequently, we align visual features $\mathbf{x_a}$ and semantic features $\mathbf{t_a}$ as privileged information by a normalized distance measure method, which reflects part of the features that match the visual modality and the text modality. We use $\mathcal{L}_{Sim}$ as a constraint, which is defined as follows:

$$\mathbf{x_a} = \mathcal{F}_x(\mathbf{x}) \tag{11}$$

$$\mathbf{t_a} = \mathcal{F}_t(\mathbf{t}) \tag{12}$$

$$\mathcal{L}_{Sim} = \left\| \mathbf{x_a} - \mathbf{t_a} \right\|_2. \tag{13}$$

where $\mathcal{F}_x$ and $\mathcal{F}_t$ are the respective feature mapping networks of visual modality and semantic modality.

Secondly, we input the visual feature $\mathbf{x_a}$ obtained after alignment into the cross-modal transfer network E_t^r, so as to transfer the original visual features to the semantic feature space, and then decode the transferred features through the semantic decoder D_t^r to obtain the reconstructed feature $\mathbf{t_r}$:

$$\mathbf{t_r} = D_t^r(E_t^r(\mathbf{x_a})) \tag{14}$$

Through cross-modal feature reconstruction work, we extract the information on visual modalities at the text level and close the semantic gap between visual and semantic modalities without destroying the distribution of semantic features. In the process of feature reconstruction in the semantic reasoning module, in order to make the prediction results of reconstructed text features $\hat{\mathbf{t}}_\mathbf{r}$ closer to the real text labels $\mathbf{y_t}$, we use the reconstruction loss $\mathcal{L}_{recon}$ defined by the cross-entropy function to constraint, and the specific formula is as follows:

$$\mathcal{L}_{recon} = -\sum_{i=1}^{N} \mathbf{y_t} \log(\hat{\mathbf{t}}_\mathbf{r}), \tag{15}$$

where N refers to the total number of samples.

Finally, the semantic inference module will generate the semantic reconstruction feature $\mathbf{t_r}$, which is also used for feature fusion.

E. Cross-Modal Feature Fusion Module for Unified Representation Learning

In the feature fusion stage, we perform the cross-modal fusion of the visual features $\mathbf{x_c}$ and the semantic features $\mathbf{t_r}$ of for recognition and classification. The generation process of the final fusion features $\mathbf{F_{mix}}$ is described as follows:

$$\mathbf{F_{mix}} = \tau\left(\mathrm{cat}\left(\tau\left(\mathbf{x_c}\right), \tau\left(\mathbf{t_r}\right)\right)\right), \tag{16}$$

where $\mathrm{cat}\left(.\right)$ refers to the splicing operation of characterization, and $\tau\left(.\right)$ refers to a fully connected layer followed by $ReLU(.)$ Activation function.

In the feature fusion module, we adopt a variety of strategies for multi-modal representation fusion. CRIF makes predictions using the fusion features, which are defined as:

$$\hat{\mathbf{C}} = f(\mathbf{x_c} \oplus \mathbf{t_r}), \tag{17}$$

where $\oplus$ can be any of the commonly used vector operators, such as plus, multiplication, max-pooling, and min-pooling, the specific splicing process is shown in Eq. 14, $f(.)$ means the final classifier of CRIF.

The feature fusion module uses the classification loss $\mathcal{L}_{mix}$ for constraint. The classification loss can be either cross entropy or binary cross entropy, depending on the task. The loss formula for the final fusion process is as follows:

$$\mathcal{L}_{mix} = \begin{cases} -\sum_1^N y \log(\phi(\mathbf{F_{mix}})) \\ -\sum_1^N \left[y \log(\psi(\mathbf{F_{mix}})) + (1-y) \log(1 - (\psi(\mathbf{F_{mix}}))) \right] \end{cases}. \tag{18}$$

We choose the first formulation when the mixture of features $\mathbf{F_{mix}}$ is used to predict single-class classification tasks, and the other when used for multi-class classification tasks, where $\phi(.)$ and $\psi(.)$ denote different two activation functions.

F. Training Strategy

The modules in the CRIF framework are trained under constraints respectively, so multi-stage training and end-to-end training can be performed.

1) Multi-stage Training: Multi-stage training is mainly divided into feature alignment training, cross-modal inference training, and feature fusion training.

In the first stage, the local extraction network of visual features is trained, and the similarity constraint of visual and semantic classification features and the loss of classification are used to send them back together.

$$\mathcal{L}_{class} = \begin{cases} \text{CE}(softmax(\hat{\mathbf{x}}), y) \\ \text{BCE}(sigmoid(\hat{\mathbf{x}}), y) \end{cases} \tag{19}$$

$$\mathcal{L}_1 = \mathcal{L}_{class} + \alpha_{sim} \cdot \mathcal{L}_{sim}. \tag{20}$$

where $\mathcal{L}_{class}$ is divided according to whether the classification task is single-class or multi-class, and $\hat{\mathbf{x}}$ represents the visual modality prediction results. α_{sim} is the weight of similarity loss, and we will show the range of α_{align} in Section.IV-B2.

In the second stage, the semantic decoder is trained, and the model reconstruction ability is improved through the alignment loss of visual features and semantic features and cross-modal reconstruction loss. The loss function is.

$$\mathcal{L}_2 = \mathcal{L}_{recon} + \alpha_{align} \cdot \mathcal{L}_{align} \tag{21}$$

where α_{align} is the weight of aligned loss, and we will show the range of α_{align} in Section.IV-B2.

The visual classification features generated by the alignment module and the semantic selection representation generated by the cross-modal semantic inference module are fused and constrained by a classification loss $\mathcal{L}_{\text{mix}}$.

2) End-to-end Training: The CRIF can also be end-to-end trained, which requires a weighted combination of the losses in multi-stage training, the formula is:

$$\mathcal{L}_{all} = \mathcal{L}_1 + \beta \cdot \mathcal{L}_2 + \gamma \cdot \mathcal{L}_{mix} \tag{22}$$

where $beta$ and $gamma$ are the weights of losses in stage 2 and feature fusion respectively, and we will show the range of α_{align} in Section IV-B2.

IV. Experiment

A. Experiments Settings

1) Datasets: We conducted experiments on four cross-modal datasets, and divided the experiments into image classification experiments and video classification experiments according to the different data types used for the task. The image classification experiments used the image-text datasets Vireo-Food172 and NUS-WIDE. The video-description dataset MSR-VTT and the video-speech dataset ActivityNet Captions are used for the video classification task. Details are presented as follows:

Vireo-Food172 [19]: a single-label classification dataset containing 110,241 images of dishes in 172 categories. The dataset contains 353 kinds of texts, and each image sample corresponds to 3 texts on average. According to the setting of the original paper, we split the data set into 66,071 and 33,154 images for training and testing respectively.

NUS-WIDE [20]: multi-label classification dataset, containing 269,648 image samples, corresponding to 81 categories. Each image sample corresponds to several texts, with a total of 1000 texts. We divided the training set and test set by referring to the original paper, and cleaned the data set. After removing the data lacking labels or texts, 203,598 image samples were left, including 121,962 samples of the training set and 81,636 samples of the test set.

MSR-VTT [21]: contains 10,000 unique YouTube video clips. Each of them is annotated with 20 different text captions, so there are 200,000 video caption pairs in total. We split the dataset into 7,010 and 2,990 videos for training and testing.

ActivityNet Captions [22]: contains 20,000 captioned videos, totaling 849 video hours, with a total of 100,000 segments, each with a unique start and end time. On average, each 20,000 video contains 3.65 temporally localized sentences, for a total of 100,000 sentences. The average length of each sentence is 13.48 words, which also shows a normal distribution. We split the dataset into 10,009 and 4,515 videos for training and testing, respectively.

2) Evaluative Measures: In experiments on the Vireo-Food172 dataset, $Accuracy$ was used for evaluation in single-label classification. In experiments on the NUS-WIDE dataset, $Precision$ and $Recall$ were used to evaluate model prediction performance in multi-label classification tasks. The three formulas are specified as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

where TP is the number of true positive samples, FP is the number of false positive samples, TN is the number of true negative samples, and FN is the number of false negative samples. For the above indicators, we calculate the average value of top-1 and top-5.

B. Experiments on Image Classification

1) Implementation details:

- We conduct experiments on two groups of backbone networks, the first group is the Visual Model Backbone:
  - **ResNet-18:** Pytorch implementation[1] for 18-layer ResNet [37].
  - **ResNet-50:** Pytorch implementation[2] for 50-layer ResNet [37].
  - **VGG19-bn:** Pytorch implementation[3] for 19-layer VGG [38] with a batch normalization layer added after the convolutional layer.

[1]https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py
[2]https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py
[3]https://github.com/pytorch/vision/blob/main/torchvision/models/vgg.py

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

TABLE I
PERFORMANCE COMPARISON OF COMMONLY USED NEURAL NETWORK MODELS ON VIREO-FOOD172 AND NUS-WIDE DATASETS WHERE THE ALIGNMENT FRAMEWORK IS BASED ON THE RESNET-50 EXPERIMENT (**ACC** REFERS TO ACCURACY, **PRE** REFERS TO PRECISION, **1** AND **5** REFER TO THE AVERAGE VALUE OF TOP-1 AND TOP-5, THE BEST PERFORMANCE OF EACH INDICATOR HAS BEEN HIGHLIGHTED IN BOLD)

Method	Model	Vireo-Food172		NUS-WIDE			
		Acc-1	Acc-5	Pre-1	Pre-5	Recall-1	Recall-5
Visual Modal Backbone	ResNet-18	0.773	0.932	0.746	0.372	0.410	0.804
	ResNet-50	0.816	0.950	0.786	0.391	0.440	**0.864**
	VGG	0.812	0.951	0.789	0.393	0.442	0.851
	WRN	0.823	0.955	0.787	0.394	0.440	0.853
	WISeR	0.828	**0.965**	0.789	0.395	0.441	0.855
	RepVGG	0.835	0.963	0.797	0.394	0.448	0.856
	RepMLPNet	0.833	0.962	0.801	0.405	0.448	0.877
	ViT	0.836	0.966	0.796	0.395	0.453	0.855
Align Framework	ATNet	0.839	0.950	0.794	0.391	0.445	0.847
	CM-VAE	0.829	0.954	0.790	0.395	0.443	0.854
	CMFL	0.831	0.958	0.809	0.405	0.456	0.871
	ViLT	0.829	0.963	0.807	**0.406**	0.458	0.880
	CRIF(Ours)	**0.841**	0.948	**0.813**	0.398	**0.458**	0.858

– **WRN50-2:** Pytorch implementation[4] for 18-layer Wide Residule Networks (WRN) [52] based on ResNet-50 with wide factor of 2.
– **WISeR:** In-house implementation for Wide-Slice Residual Networks (WISeR) [53] using WRN50-2
– **RepVGG-A2:** Pytorch implementation[5] for 22-layer RepVGG [54] with layers of each stage are 1,2,4,14,1.
– **RepMLPNet-T224:** Pytorch implementation[6] for 10-layer RepMLPNet [55] using ResNet-50 with multi-branch and re-parameterization.
– **ViT-B/16:** Pytorch implementation[7] for 12-layer Vision Transformer (ViT) [56] with patch size 16, feature dimension 768.

In the group of Align Framework, we compare with the following models:

– **ATNet:** In-house implementation for cross-modal alignment and transfer network (ATNet) [6] using ResNet-50 and LSTM for image channel and semantic channel respectively.
– **CM-VAE:** In-house implementation for Cross Modal Variational Auto-Encoder(CM-VAE) [13]. CM-VAE consists of three pairs of VAE [50], two of which are used for encoding and decoding of a single modality, and the other decodes the features of modality M1 to another modality cross-modally to achieve inter-modal feature alignment.
– **CMFL:** In-house implementation for Cross Modal Focal Loss(CMFL) [57], using improved cross-modal focal loss to focus more on indistinguishable samples.
– **ViLT:** Pytorch implementation[8] for 12-layer single-stream vision and language transformer [37], using linear projection and BERT as image and semantic channel embedding encoder.

The proposed method CRIF used Feed-forward Neural Network (FNN) and Long-Short Term Memory network (LSTM) [58] for encoding and decoding semantic features respectively. Feed-forward neural network: Its internal structure consists of 2 fully connected layers, the output of each layer is activated by ReLU, and then enters the next fully connected layer. LSTM: the corresponding cell number is generated according to the maximum word bit in the dataset, and the maximum hidden state dimension is 300. The corresponding word bit representation is calculated through the LSTM unit.

2) Hyper-Parameter Selection:

- In this experiment, we choose Adam as the optimizer of the model, where the weight decay of Adam is set to 1e-3, and the learning rate of all neural networks is set from 5e-5 to 5e-3. Every four rounds of training, the learning rate will decay to 0.1 times the original. For the weights of losses mentioned in the training strategy, we select α_{sim} and α_{align} from 0.1 to 2.0, and the value of β and γ are chosen in [0.1, 0.2, 0.5, 1.0]. We conduct experiments on NVIDIA Tesla V100 with a mini-batch input of 64 images, and each stage is trained for 13 rounds. For Vireo-Food172, a ResNet-50-based model takes $\sim$6 hours to train. For the multi-classification dataset NUS-WIDE, we set the weight of positive sample loss of BCE loss was set from 20 to 150, a ResNet-50-based model takes $\sim$10 hours to train.

3) Performance Comparison:

- In order to verify the effectiveness of the CRIF method in improving the ability of visual representation learning, we compare with two groups of baselines: visual modal backbone and align framework. In the visual modal backbone, we first compared basic visual backbones such as ResNet-18 [37], ResNet-50 [37] and VGG-19 [38], and improved WRN [52], and WISeR [53] based on ResNet-50. We also compared the backbone networks proposed in recent years, such as RepVGG [54] and RepMLPNet [55] based on reparameterization, and ViT [59] based on transformer architecture. In the align framework, we compared ATNet which uses cross-modal alignment after

[4]https://github.com/szagoruyko/wide-residual-networks
[5]https://github.com/DingXiaoH/RepVGG
[6]https://github.com/DingXiaoH/RepMLP
[7]https://github.com/asyml/vision-transformer-pytorch
[8]https://github.com/dandelin/ViLT

feature decoupling, CM-VAE [13] that achieves inter-modal information alignment by decoding features across modalities based on the vae-like architecture, CMFL [57] that achieves cross-modal information alignment through global constraints, and ViLT [56] that implicitly aligns different modalities through a single-stream Transformer network. As shown in Table I.

- Since ResNet-50 and VGG have a more complex network structure, they achieved a significant improvement (approximately the 4%) in both Vireo-Food172 and NUS-WIDE data sets compared to ResNet-18, the simplest baseline network. WRN and WISeR achieve better classification performance than ResNet-50 through the improvement of the model structure, which shows that the basic vision backbone can improve transfer ability based on structural improvement and parameter increase. Based on the reparameterization method, RepVGG and RepMLPNet convert training knowledge into the inference stage and integrate multi-branch information to improve the performance of the basic model; Compared with the above models, ViT using the Transformer architecture achieved the best visual classification performance, indicating the effectiveness of the model architecture improvement.
- ATNet based on ResNet-50 after passing the heterogeneous feature alignment module (distributional alignment and content alignment, where the content alignment module uses the $\mathcal{L}_{CORAL}$ constraint function) has significantly improved the prediction performance of image classification, which is more prominent in NUS-WIDE dataset. Due to the difference in the distribution of pre-trained data and downstream task data, there is a problem with domain adaptation when transferring a pre-trained large model to the target data set, therefore, ViLT achieved better classification performance on NUS-WIDE, which is more consistent with the distribution of pre-training data, but performed poorly on the food classification data set Vireo-Food172.
- The proposed method CRIF using the heterogeneous feature alignment module alleviates the semantic gap caused by the cross-modal difference between visual modality and text modality by imposing alignment loss constraints on features of different modes. Data in Table I shows that CRIF has a strong generalization ability on data sets in different fields, which achieved the state-of-art results both on Vireo-Food172 and NUS-WIDE.

C. Experiments on Video Classification

1) Implementation Details:

- we conduct experiments on backbones belonging to visual and semantic modalities respectively, then we show the effects of introducing feature alignment. GRU and ViLT are chosen as backbones for demonstrating the model-agnostic character of feature alignment.

TABLE II
PERFORMANCE COMPARISON OF COMMONLY USED NEURAL NETWORK MODELS ON MSR-VTT AND ACTIVITYNET CAPTIONS DATASETS WHERE THE ALIGNMENT FRAMEWORK IS BASED ON THE GRU AND VILT EXPERIMENT (**ACC** REFERS TO ACCURACY**1** AND **5** REFER TO THE AVERAGE VALUE OF TOP-1 AND TOP-5, THE BEST PERFORMANCE OF EACH INDICATOR HAS BEEN HIGHLIGHTED IN BOLD)

Methods	Model	**MSR-VTT**		**Activitynet Captions**	
		Acc-1	**Acc-5**	**Acc-1**	**Acc-5**
Visual modality Backbone	GRU	0.519	0.815	0.802	0.955
	ViLT	0.515	0.817	0.814	0.954
Semantic modality Backbone	GRU	0.526	0.828	0.224	0.425
	ViLT	0.531	0.817	0.249	0.455
Cross-modal Alignment	GRU	0.545	0.835	0.820	0.957
	ViLT	**0.560**	**0.839**	**0.835**	**0.964**

- **ViLT:** Pytorch implementation[9] for 12-layer single-stream vision and language transformer [37], using linear projection and BERT as image and semantic channel embedding encoder.
- **GRU:** In-house implementation for Gated Recurrent Unit [60] using Pytorch.

For the two multimodal video datasets MSR-VTT and Activitynet Captions visual features and semantic features of both are extracted. The visual features are extracted using a pre-trained S3D network with a feature dimension of 1024. For the MSR-VTT dataset [61], the text features of sentence descriptions in the semantics are extracted using the Google Cloud Speech to Text API network with a dimension of 768 [62], and for the Activitynet Captions dataset, the audio features in the semantics are extracted using the VGGish network pre-trained on the YouTube -8M dataset on top of the pre-trained VGGish network to extract audio features in the semantics with dimension 128 [63].

2) Hyper-Parameter Selection:

- In the experiment, we follow the feature dimension setting of the pre-trained large model ViLT and set the feature dimension to 768. The experimental model is optimized using the Adam optimizer during the training process, where the learning rate is selected from 1e-6 to 1e-3, and then for every 4 epochs of training, the learning rate of the optimizer decays by a factor of 0.1. The batch size was selected from $\{32, 64, 128, 256\}$.

3) Performance Comparison:

- In this section, we show the effect of the alignment algorithm on the MSR-VTT dataset and Activitynet Captions dataset, comparing the performance of the underlying vision model GRU backbone network on top of both GRU and ViLT models for video classification experiments based on visual information, semantic information, and multimodal feature alignment. The obtained results are shown in Table II. We have the following observations.
  - The performance of multimodal information fusion is apparently higher than that of classification with only unimodal features. Due to the complementary property between the semantic information of different

[9]https://github.com/dandelin/ViLT

TABLE III
RESULTS OF ABLATION STUDY.**ALIGN** REPRESENTS CONTENT ALIGNMENT, **INFERENCE** REPRESENTS THE CROSS-MODAL RECONSTRUCTION FEATURES OF THE SEMANTIC INFERENCE MODULE, AND **FUSION** REPRESENTS FEATURE FUSION. THE EVALUATION INDEXES ARE THE SAME AS THOSE IN TABLE 1, AND THE BEST PERFORMANCE IS MARKED IN BOLD.

Method	Vireo-Food172		NUS-WIDE			
	Acc-1	Acc-5	Pre-1	Pre-5	Recall-1	Recall-5
Baseline	0.773	0.932	0.746	0.372	0.410	0.804
+Align	0.788	**0.933**	0.776	0.382	0.433	0.825
+Inference	0.756	0.886	0.763	0.376	0.426	0.814
+Fusion	**0.803**	0.931	**0.781**	**0.385**	**0.436**	**0.831**

TABLE IV
RESULTS OF ABLATION STUDY.**ALIGN** REPRESENTS CONTENT ALIGNMENT AND **FUSION** REPRESENTS FEATURE FUSION. THE EVALUATION INDEXES ARE THE SAME AS THOSE IN TABLE 1, AND THE BEST PERFORMANCE IS MARKED IN BOLD.

Method	MSR-VTT		Activitynet Captions	
	Acc-1	Acc-5	Acc-1	Acc-5
Baseline	0.515	0.817	0.814	0.954
+Align	0.560	0.839	0.835	**0.964**
+Fusion	**0.568**	**0.840**	**0.838**	0.951

modalities, the classification enhancement of the fusion of textual and visual information is more obvious after the fusion of information.

- The overall performance of the pre-trained large model ViLT is higher than that of the GRU on the experimental dataset, which demonstrates the superiority of the pre-trained large modality on the downstream task. Meanwhile, the multimodal interaction network can further reduce the bias caused by the inconsistent distribution between different modalities, and reduce the heterogeneity between different feature frames through the collaborative learning of semantic information to visual information in the feature alignment process.
- The classification performance of the multimodal alignment algorithm appears significantly improved after applying it to different models (GRU or VILT), showing its model-independent characteristics.

D. Ablation Studies

1) Ablation study on image classification: To investigate the effectiveness of modules in CRIF, we conduct ablation experiments with ResNet-18 as the baseline model, and the results are shown in Table III. Due to the background noise in the images, the baseline networks all performed poorly on the datasets, and after heterogeneous feature alignment, with the assistance of text features, the performance of the model was partially improved. However, the performance of the model is limited due to the semantic gap between textual and visual features. After the cross-modal feature reconstruction module, the model maps visual features to the text modality space and mines the semantic information contained in the visual modality. However, due to the loss of information in the modality mapping, there is a slight drop in performance. Therefore, we fuse the semantic and visual features to make the model focus on the main parts of the image, thereby further improving the performance of the model.

2) Ablation study on video classification: To investigate the effectiveness of Align modules, we conduct ablation experiments with ViLT as the baseline model, and the results are shown in Table IV. Similar to image classification, it can be concluded that heterogeneous feature alignment can reduce the heterogeneity between different modalities, and fusion can further improve the characterization ability of the model.

E. In-depth Analysis

In order to verify the effectiveness of the method proposed in this paper on the performance improvement of the classification task of the models with the different number of model parameters, we conducted alignment-based experiments on the base model ResNet-50 and the pre-trained large model ViLT respectively. Then we use the t-SNE method to show the distribution changes of visual features caused by cross-modal alignment. Specifically, we randomly select 5 categories in the Vireo-Food172 dataset, and randomly select around 100 image samples for each category, which are not visible during the training phase. As shown in Figure 3, the distribution of visual representations extracted by ResNet-50 and ViLT changes due to cross-modal alignment.

1) Cross-modal alignment improves image discrimination: As shown in the Figure 3, image categories mixed in the original feature distribution space can be further distinguished from each other after cross-modal feature alignment. Otherwise, the discrete points outside the cluster are further concentrated to the center of the cluster, which reduces intra-cluster distance. This shows that the visual discrimination ability of the model is effectively improved by cross-modal feature alignment, explaining the effectiveness of our proposed method.

2) Pre-training improves visual representation quality: Compared with ResNet-50 whose category feature distribution is difficult to distinguish before alignment, the pre-trained large model ViLT has more visual representations concentrated near clusters, and the distance between clusters is larger, which is easy to distinguish. This shows that with the support of large-scale supervised pre-training and redundant parameters, the model can obtain higher-quality visual representation modeling capabilities. For ViLT, through cross-modal feature alignment, although the discrete points of samples outside the cluster are further gathered in the cluster center, the inter-cluster discrimination is not significantly improved.

F. Case Studies

1) Feature alignment performance: In order to verify that the alignment method proposed in this paper effectively al-

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

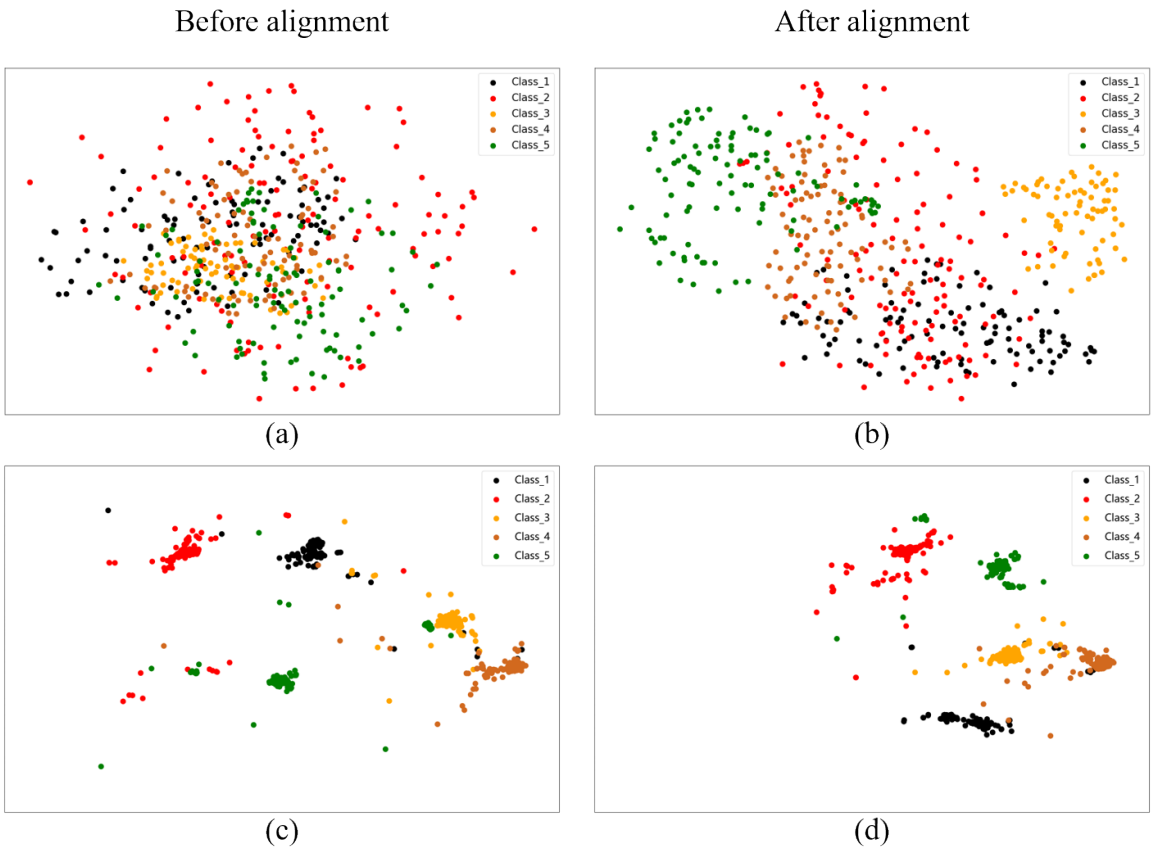

Fig. 3. The t-SNE visualization of the influence of the feature alignment method on different backbone representation learning, using two models of ResNet-50 and ViLT, randomly selected 5 categories on the VireoFood-172 dataset for experiments. Figures (a) and (b) illustrate the visualization results of resnet50, and Figures (c) and Figures (d) depict the results of ViLT.

leviates the cross-modal representation heterogeneity between visual representations and semantic representations, we conduct alignment-based experiments on visual and semantic modalities, and visualize their features by t-SNE to observe the impact of alignment operation on the distribution of visual representations and the semantic gap between the two modalities. The specific experimental process is that we randomly selected a total of 40 test samples on the Vireo-Food172 dataset, including 20 for each modality, and visualized the original features and the aligned features of these samples, the result of the visualization is shown in Figure 4.

In the initial stage(figure a), the distance between the corresponding samples of the modalities is large, and the overall distribution has obvious deviation, which is caused by modal heterogeneity. After feature alignment(figure b), the distribution of heterogeneous modal features in the latent space is closer than before, and the feature confounding of the visual modalities is alleviated. Overall, our alignment method clearly alleviates the problem of distance gap and distribution bias of cross-modal features in the latent space.

2) Analysis of feature alignment methods: In order to deeply analyze the influence of different alignment strategies and alignment functions in the heterogeneous feature alignment module on the model performance improvement, we selected ResNet-18 as the baseline algorithm and conducted experimental comparison on the Vireo-Food172 dataset. The experimental Settings were the combination of two alignment strategies $Align$ and $Clip$, where $Align$ represents the overall part of the feature for alignment and $Clip$ represents the critical part of the feature for alignment, and three alignment constraint functions ($\mathcal{L}_{L2-Norm}$, $\mathcal{L}_{KL}$, $\mathcal{L}_{CORAL}$, $\mathcal{L}_{SSAN}$) respectively, in which $\mathcal{L}_{L2-Norm}$ is a common alignment function and $\mathcal{L}_{SSAN}$ is the centroid alignment method [18]. As shown in Table V, the performance of visual features is significantly improved after alignment in a single visual modality, while text features are less affected by alignment because the original features are already performing well. After the fusion of visual and text features, it can be seen that the fusion features after alignment perform better, and the performance of the fusion features is better than the two features before fusion. Obviously, the heterogeneous feature alignment module not only ensures the fusion feature performance, but also optimizes the features of each modality.

3) Multi-modal feature fusion: We tried the prediction performance experiment after the fusion of multi-modal features, in which the fusion features were respectively from the visual features and text features after the alignment of heterogeneous features, and the reconstructed text features across the semantic inference module. The experimental results of directly concatenating the features of three different modes are shown

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2023.3265645

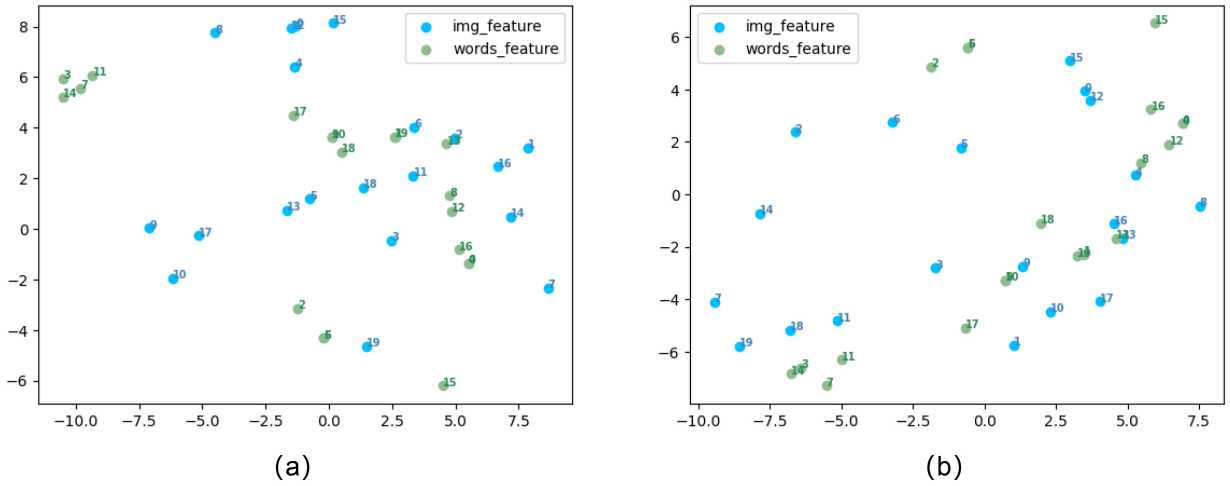

Fig. 4. The t-SNE visualization of the distribution of partial data after alignment, where Figure (a) illustrates the initial feature before alignment, Figure (b) depicts the feature after alignment, and the number indicates the category to which the feature belongs.

TABLE V
IN THE COMBINED EXPERIMENT OF DIFFERENT ALIGNMENT STRATEGIES AND CONSTRAINT FUNCTIONS ON VIREO-FOOD172, THE DATA RESULTS IN THE TABLE ARE THE AVERAGE OF TOP-1 ACCURACY. WE CHOOSE CONCATENATING AS THE METHOD OF FEATURE FUSION. THE BEST RESULTS FOR EACH FEATURE FUSION METHOD ARE HIGHLIGHTED IN BOLD.

Strategy	Function	Visual	Semantic	Inference	V+S	V+I	S+I	V+S+I
None	ResNet-18	0.773	0.975	0.756	0.975	0.776	0.975	0.977
Align	L2-Norm	0.784	0.975		0.976	0.785	0.976	0.979
	KL-D	0.778	0.976		0.979	0.790	0.977	0.983
	DeepCoral	0.787	0.977		0.979	0.800	**0.979**	0.982
	SSAN	0.778	0.975		0.979	0.794	0.977	0.981
Clip	L2-Norm	0.781	0.976		0.978	0.782	0.978	0.979
	KL-D	0.779	**0.978**		0.978	0.786	0.978	0.980
	DeepCoral	**0.788**	0.976		**0.980**	**0.803**	0.979	**0.984**
	SSAN	0.778	0.975		0.977	0.780	0.976	0.982

in Table V, we can conclude as follows.

After the experiment, it is found that after the feature fusion between the other modalities and the text modality, the performance of the fused features mainly depends on the text modality because the text modality is easier to be recognized, which is similar to the previous research findings. The cross-modal reconstruction features generated by the semantic inference module can mine the hidden information in the visual modality, and have a compensatory effect when fused with the visual modality, then the performance is improved after fusion. Thanks to the semantic inference module, some of the information extracted from the visual modality is also helpful for the text modality, but since the text modality already performs well, the improvement is very limited.

Finally, we try to concatenate the features of the three modalities, and the experimental results are similar to the splicing results of the reasoning modality and the semantic modality. The performance of the fused features is mainly dominated by the semantic features, and the visual features and the inferential features play the role of information supplement. From the results, the improvement of the three-feature fusion is also limited, which is also because the performance of the semantic modality is good enough.

V. CONCLUSION

In this paper, we summarize the effectiveness of different alignment strategies and propose a cross-modal image classification framework CRIF. The proposed CRIF effectively alleviates the modal heterogeneity problem by a two-level feature alignment method along with cross-modal inference from visual to semantic space. Experiments demonstrate that the proposed framework can enhance visual representation learning with the help of semantic privileged information and promote learning in different modalities.

In the future, we will further analyze the inner mechanism of feature alignment and study how various alignment methods affect representation learning, so as to propose more effective alignment strategies and alignment functions, and further improve the cross-modal inference proposed in this paper to encourage model mining semantic information with less noise.

ACKNOWLEDGEMENTS

This work is supported by the National Key RD Program of China (Grant no. 2021YFC3300203), the TaiShan Scholars Program (Grant no. tsqn202211289), and the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048)

REFERENCES

[1] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[2] Y. Hu, S. M. Bejarano, and G. Hoffman, "Shadowsense: Detecting human touch in a social robot using shadow image classification," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–24, 2020.

[3] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[4] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Information bottleneck learning using privileged information for visual recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1496–1505.

[5] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[6] L. Meng, L. Chen, X. Yang, D. Tao, H. Zhang, C. Miao, and T.-S. Chua, "Learning using privileged information for food recognition," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 557–565.

[7] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[8] Y. Yao, F. Shen, J. Zhang, L. Liu, Z. Tang, and L. Shao, "Extracting privileged information for enhancing classifier learning," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 436–450, 2019.

[9] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *Journal of Big Data*, vol. 4, no. 1, pp. 1–42, 2017.

[10] Y. Luo, T. Liu, Y. Wen, and D. Tao, "Online heterogeneous transfer metric learning." in *IJCAI*, 2018, pp. 2525–2531.

[11] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification," in *Twenty-fifth aaai conference on artificial intelligence*, 2011.

[12] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2gan: Cross-modal recipe retrieval with generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 477–11 486.

[13] T. Theodoridis, T. Chatzis, V. Solachidis, K. Dimitropoulos, and P. Daras, "Cross-modal variational alignment of latent spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 960–961.

[14] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 680–689.

[15] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 285–10 295.

[16] J. Duan, L. Chen, S. Tran, J. Yang, Y. Xu, B. Zeng, and T. Chilimbi, "Multi-modal alignment using representation codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 651–15 660.

[17] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[18] S. Li, B. Xie, J. Wu, Y. Zhao, C. H. Liu, and Z. Ding, "Simultaneous semantic alignment network for heterogeneous domain adaptation," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3866–3874.

[19] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 32–41.

[20] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.

[21] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[22] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[23] R. Shailendra, A. Jayapalan, S. Velayutham, A. Baladhandapani, A. Srivastava, S. Kumar Gupta, and M. Kumar, "An iot and machine learning based intelligent system for the classification of therapeutic plants," *Neural Processing Letters*, pp. 1–29, 2022.

[24] O. Ali and M. K. Ishak, "Bringing intelligence to iot edge: Machine learning based smart city image classification using microsoft azure iot and custom vision," in *Journal of Physics: Conference Series*, vol. 1529, no. 4. IOP Publishing, 2020, p. 042076.

[25] B. K. Mishra, D. Thakker, S. Mazumdar, S. Simpson, and D. Neagu, "Using deep learning for iot-enabled camera: A use case of flood monitoring," in *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE, 2019, pp. 235–240.

[26] I. Zualkernan, S. Dhou, J. Judas, A. R. Sajun, B. R. Gomez, and L. A. Hussain, "An iot system using deep learning to classify camera trap images on the edge," *Computers*, vol. 11, no. 1, p. 13, 2022.

[27] S.-K. Noh, "Recycled clothing classification system using intelligent iot and deep learning with alexnet," *Computational Intelligence and Neuroscience*, vol. 2021, 2021.

[28] K. Phasinam and T. Kassanuk, "Machine learning and internet of things (iot) for real-time image classification in smart agriculture," *ECS Transactions*, vol. 107, no. 1, p. 3305, 2022.

[29] I. A. Zualkernan, S. Dhou, J. Judas, A. R. Sajun, B. R. Gomez, L. A. Hussain, and D. Sakhnini, "Towards an iot-based deep learning architecture for camera trap image classification," in *2020 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*. IEEE, 2020, pp. 1–6.

[30] Y. Huang, X. Li, W. Wang, T. Jiang, and Q. Zhang, "Towards cross-modal forgery detection and localization on live surveillance videos," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[31] H. Liu, L. Xu, X. Tian, H. Peng, and D. Xia, "Visible-thermal person reidentification in visual internet of things with random gray data augmentation and a new pooling mechanism," *IEEE Internet of Things Journal*, 2022.

[32] Z. Ji and S. Li, "Multimodal alignment and attention-based person search via natural language description," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 147–11 156, 2020.

[33] B. Hu, K. Guo, X. Wang, J. Zhang, and D. Zhou, "Rrl-gat: Graph attention network-driven multi-label image robust representation learning," *IEEE Internet of Things Journal*, 2021.

[34] L. D. Chamain, S. Qi, and Z. Ding, "End-to-end image classification and compression with variational autoencoders," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 916–21 931, 2022.

[35] Z. Zhang and D. Li, "Hybrid cross deep network for domain adaptation and energy saving in visual internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6026–6033, 2018.

[36] B. Hu, K. Guo, X. Wang, J. Zhang, and D. Zhou, "Rrl-gat: Graph attention network-driven multilabel image robust representation learning," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9167–9178, 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[39] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[40] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[41] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 103–114, 2018.

[42] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 7276–7282.

[43] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, pp. 1325–1341, 2019.

[44] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1276–1288, 2018.

[45] E. Yu, J. Ma, J. Sun, X. Chang, H. Zhang, and A. G. Hauptmann, "Deep discrete cross-modal hashing with multiple supervision," *Neurocomputing*, vol. 486, pp. 215–224, 2022.

[46] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[47] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022.

[48] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2581–2593, 2020.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[50] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[51] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[52] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[53] Z.-Y. Ming, J. Chen, Y. Cao, C. Forde, C.-W. Ngo, and T. S. Chua, "Food photo recognition for dietary tracking: System and experiment," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 129–141.

[54] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 733–13 742.

[55] X. Ding, H. Chen, X. Zhang, J. Han, and G. Ding, "Repmlpnet: Hierarchical vision mlp with re-parameterized locality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 578–587.

[56] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[57] A. George and S. Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7882–7891.

[58] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[60] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[61] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[62] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *arXiv preprint arxiv:1907.13487*.

[63] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in *European Conference on Computer Vision (ECCV)*, 2020.