Unifying Visual and Semantic Feature Spaces with Diffusion Models for Enhanced Cross-Modal Alignment

Yuze Zheng, Zixuan Li, Xiangxian Li, Jinxing Liu, Yuqing Wang, Xiangxu Meng, and Lei Meng(✉)

School of Software, Shandong University, Jinan, China
lmeng@sdu.edu.cn

Abstract. Image classification models often demonstrate unstable performance in real-world applications due to variations in image information, driven by differing visual perspectives of subject objects and lighting discrepancies. To mitigate these challenges, existing studies commonly incorporate additional modal information matching the visual data to regularize the model's learning process, enabling the extraction of high-quality visual features from complex image regions. Specifically, in the realm of multimodal learning, cross-modal alignment is recognized as an effective strategy, harmonizing different modal information by learning a domain-consistent latent feature space for visual and semantic features. However, this approach may face limitations due to the heterogeneity between multimodal information, such as differences in feature distribution and structure. To address this issue, we introduce a Multimodal Alignment and Reconstruction Network (MARNet), designed to enhance the model's resistance to visual noise. Importantly, MARNet includes a cross-modal diffusion reconstruction module for smoothly and stably blending information across different domains. Experiments conducted on two benchmark datasets, Vireo-Food172 and Ingredient-101, demonstrate that MARNet effectively improves the quality of image information extracted by the model. It is a plug-and-play framework that can be rapidly integrated into various image classification frameworks, boosting model performance.

Keywords: Image classification · Cross-modal alignment · Diffusion model

1 Introduction

Visual classification is a critical task in the field of computer vision [3,15,28,35]. However, the quality of visual images is susceptible to various factors, including but not limited to, interference from non-main elements and changes in lighting angles, leading to inconsistent performance in image classification [18,33,34].

M. Wand et al. (Eds.): ICANN 2024, LNCS 15021, pp. 110–125, 2024.
https://doi.org/10.1007/978-3-031-72347-6_8

With the rapid development of social media platforms, a vast amount of textual information related to visual images has emerged. These pieces of information present a complex relationship of mutual dependence and complementarity, which can compensate for the shortcomings of single-modal information. However, the potential complementarity between images and texts is often limited due to the fundamental differences between these two forms of information, thus affecting the effectiveness of multimodal learning [24,27]. Therefore, effectively integrating and utilizing cross-modal data information becomes key to enhancing the performance of multimodal learning.

In recent years, researchers in multimodal learning have commonly adopted cross-modal representation alignment strategies to reduce the heterogeneity between different modal information. These strategies can be broadly divided into two categories: those based on distance metrics [12,13,16] and those based on contrastive learning [11,32,36]. Distance-based alignment methods mainly constrain the spatial distance between different sources of information, such as category center distance or decision space distance, to effectively mitigate the problem of distance differences between modal information in the information space. In contrast, contrastive learning-based alignment methods divide multimodal information into positive and negative samples and enhance the similarity between positive samples while separating them from negative samples by comparing sample similarities. This approach strengthens the distinction and interactivity of information in the representation space. However, both methods tend to focus on the distance between representations while aligning cross-modal representations as much as possible, neglecting the significant distribution differences between different modal representations, which is a challenge that needs to be addressed in multimodal learning.

To deeply address the challenges in cross-modal representation alignment, this study first thoroughly analyzes the common algorithmic frameworks within the two categories of alignment methods, assessing their strengths and limitations. Based on this analysis, we introduce an innovative **m**ultimodal **a**lignment and **r**econstruction **net**work, named MARNet. As illustrated in Fig. 1, MARNet, through a contrastive learning-based alignment strategy, effectively resolves the issue of representation confusion in the visual space while enhancing the separation of samples within the same category space. Unlike previous alignment methods, we designed a cross-modal diffusion reconstruction module to complement the deficiencies of traditional alignment methods. By introducing a diffusion model with guided conditions, we achieved deep interaction between visual and textual information, significantly optimizing the distribution of visual information and enhancing the model's perception of the visual information's core areas. Through this representation fusion strategy, we enabled two independent modules to complement each other, thereby achieving the goal of enhancing visual representation and improving model robustness.

In practical image-text multimodal classification tasks, we conducted extensive experimental validation of MARNet, especially on two single-label datasets involving dishes and ingredients, Vireo-Food172 and NUS-WIDE. Compared

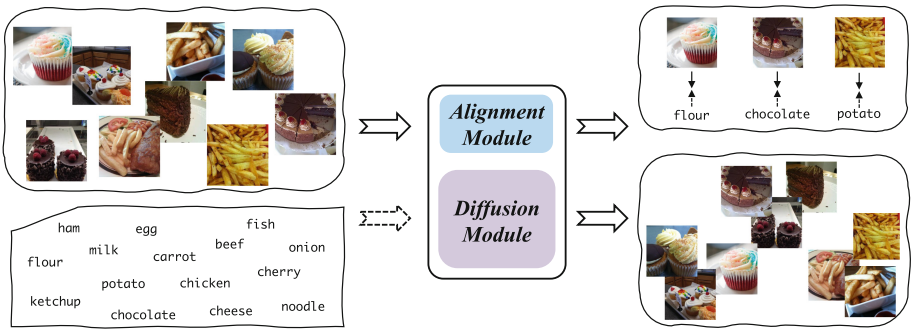

Fig. 1. The proposed MARNet schematic diagram. The network mainly comprises alignment and diffusion modules, wherein the alignment module matches and aligns image and text information, and the diffusion module reconstructs the distribution of image information.

to previous alignment frameworks, MARNet, as a model-agnostic algorithmic framework, significantly enhanced the quality of visual representation and improved the framework's performance in downstream tasks through its unique embedding matching alignment module and cross-modal diffusion reconstruction module. Through case analysis, we further demonstrated how the cross-modal diffusion reconstruction module, by incorporating textual information, significantly improved the distribution of visual information in the representation space and strengthened the model's ability to recognize visual subjects, showcasing MARNet's strong potential and practical value in the field of multimodal learning. The innovative contributions of this chapter are mainly reflected in:

1. We propose an innovative multimodal alignment and reconstruction network, named MARNet. Under the fine guidance of textual information, this network significantly improves the quality of visual information and markedly enhanced the model's decision-making ability in the visual domain. MARNet is designed with flexibility, allowing easy integration with current mainstream two-channel models to strengthen the capability of feature representation, demonstrating outstanding versatility and reusability.
2. We introduce a cross-modal diffusion reconstruction module. This module utilizes a diffusion model to smoothly unfold multimodal data along the time axis, correcting visual modal information through deep interaction, effectively optimizing the aggregation distribution of similar visual information. This module not only enhances the model's ability to process visual information but also deepens its understanding of multimodal data.
3. We explore the respective advantages and limitations of existing cross-modal alignment methods and diffusion model reconstruction representations, providing referential conclusions for future research.

2 Related Work

2.1 Cross-Modal Alignment

In common cross-modal learning scenarios, there is a clear distribution difference in the representation space among different modal data, and representations of the same category from different modalities are disorganized. Cross-modal representation alignment is needed to mitigate differences between cross-modal representations. Leading-edge cross-modal alignment methods can be divided into two paradigms: distance metric-based and contrastive learning-based methods.

In distance metric-based alignment methods, Lee *et al.* [13] project the decision information from different modalities into a spherical space, and optimize the distance using the Wasserstein metric. Li *et al.* [16] propose centroid alignment, which explicitly pulls the distance of modal corresponding class closer by calculating the centroids of clusters, while also incorporating decision information for implicit alignment. Kang *et al.* [12] adopt a clustering alignment method, optimizing the distance within and between category clusters by constructing an intra-cluster sample matrix, achieving class-aware alignment.

In contrastive learning-based alignment methods, Jiang *et al.* [11] calculate the cosine similarity of visual and textual representations among positive samples, enhancing the similarity between modalities. Xie *et al.* [36] integrate attention mechanisms into alignment methods on top of traditional global representation alignment, aligning internal information of representations at a finer granularity. Wang *et al.* [32] improving upon the InfoNCE [26] through contrastive learning, maximize the similarity of positive image-text pairs in a common representation space while minimizing the negative impact of other sample pairs.

2.2 Diffusion Models for Representation Learning

The diffusion model is inspired by non-equilibrium thermodynamics [30]. Ho *et al.* [9] treat the diffusion process as a Markov chain by progressively adding random noise to the data. They train neural networks to learn the diffusion process, enabling them to denoise images corrupted with Gaussian noise.

Currently, diffusion models are mostly applied to generative tasks [17,31]. In cross-modal diffusion models, there are commonly two approaches. One is using classifier-free guidance [10], where text is used as a condition to guide image generation with noise. The other is simultaneously adding noise from multiple modalities into the network for multi-modal generation [22].

In terms of network structures used in diffusion models, U-Net architecture is commonly employed in the image domain for noise prediction, with intra-layer changes in image channels [9]. Additionally, some studies have utilized MLP structures for diffusion in user-item interactions without channel dimensions [20,21], focusing on simpler feature transformations.

Regarding image classification tasks based on diffusion models, Li *et al.* [14] introduced a method to evaluate diffusion models as zero-shot classifiers. Clark *et al.* [4] used density estimation calculated by a large-scale text-to-image generation model for zero-shot classification.

3 Method

In this section, we elaborate on how our proposed framework(MARNet), aligns image-text sample pairs in the representation space through **embedding matching alignment module(EMA)**, and how it mitigates the distribution differences existing in cross-modal information via **cross-modal diffusion reconstruction module(CDR)**, ultimately enhancing the interaction between cross-modal data. The overall architecture diagram of MARNet is shown in Fig. 2.

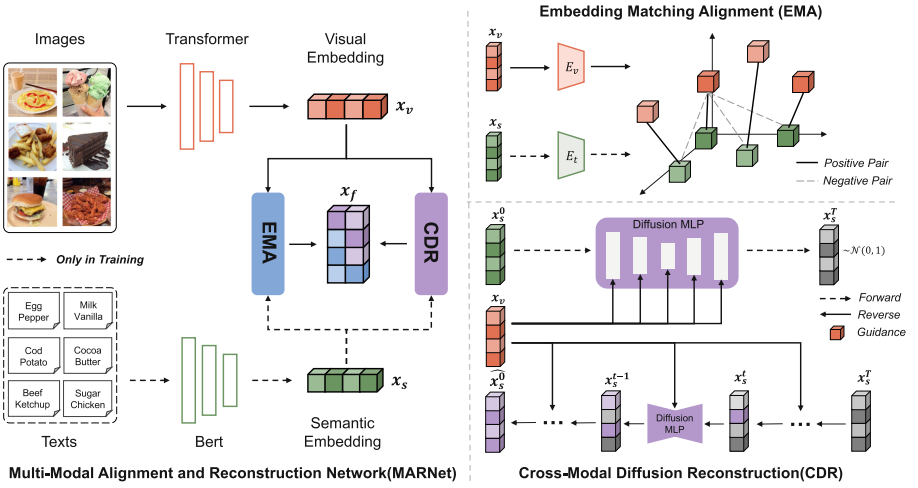

Fig. 2. The framework diagram of MARNet. The input to MARNet is image-text data pairs, which are processed by neural networks for vision and text to obtain x_v and x_s, respectively. Modules EMA and CDR handle the multi-modal representations and output representations x_{EMA} and x_{CDR}, which are fused in the end.

3.1 Overall Approach

Our goal is to learn more multi-dimensional and rich visual representations from paired image-text data through privileged information learning, in order to improve the classification performance of Multi-Modal Alignment and Reconstruction Network(MARNet). More specifically, we treat the precious and scarce ingredient data as privileged information to guide the representation of image data, that is, the training samples consist of $\mathcal{N}$ pairs of image-text data $\mathcal{S}_N = \{(p_1, i_1), (p_2, i_2), ..., (p_n, i_n)\}$, while the test samples only contain $\mathcal{M}$ pieces of photo data $\mathcal{S}_M = \{p_1, p_2, ..., p_m\}$. We use a visual encoder $\mathcal{F}_v$ to extract the representations of image data $\mathcal{R}_v = \{x_v^1, x_v^2, ..., x_v^n\}$,where $x_v = \mathcal{F}_v(p)$, and similarly for text data, $\mathcal{R}_s = \{x_s^1, x_s^2, ..., x_s^n\}$,where $x_s = \mathcal{F}_s(i)$. We take the visual

representations x_v and semantic representations x_s as inputs for the subsequent two modules. In the embedding matching alignment module, we finely align the cross-modal representations through contrastive matching learning and generate representations x_{EMA}. In the cross-modal diffusion reconstruction module, we adopt an improved diffusion model to stably and smoothly infiltrate the visual representations x_v into the semantic representations x_s and sample to generate representations x_{CDR} from Gaussian noise N_G. Finally, we fuse the output representations of the modules as x_f and predict the final classification results.

$$x_{EMA} = EMA(\mathcal{F}_v(p), \mathcal{F}_s(i)) \tag{1}$$

$$x_{CDR} = CDR(\mathcal{F}_v(p), \mathcal{F}_s(i)) \tag{2}$$

$$x_f = fusion(x_{EMA}, x_{CDR}) \tag{3}$$

$$\hat{C} = Classifier(x_f) \tag{4}$$

3.2 Embedding Matching Alignment

Based on the positive and negative sample matching alignment method of contrastive learning, we adopt an instance-wise Alignment(ITA) approach [32]. This alignment method is an improvement based on InfoNCE [26], which calculates the matching similarity ($Sim(x_v^i, x_s^i)$) of image-text representations in feature space within a batch as a constraint to align cross-domain information. When enhancing the similarity of a set of image-text representation pairs using positive and negative sample matching methods, it also reduces the matching degree of the visual representation x_v^i with other semantic representations x_s^j, where $i \neq j$. Definition of cosine similarity is as follows:

$$Sim(x_v^i, x_s^i) = \frac{x_v^i \cdot x_s^i}{||x_v^i||||x_s^i||} \tag{5}$$

where $x_v^i \cdot x_s^i$ is the dot product of vectors, and $||x_v^i||||x_s^i||$ is the product of the modulus of vectors.

We design two encoders E_v and E_s, each consisting of a linear layer, and use an activation function $g(x)$ (*i.e.*, LeakyReLU). The encoder E_v is used to map visual representations x_v, while the encoder E_s is used to map semantic representations x_s, both to the same feature space $\mathbb{R}^d$.

$$g(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{otherwise} \end{cases} \tag{6}$$

$$x_{v'} = g(E_v(x_v)), x_{v'} = \left[x_{v'}^1, x_{v'}^2, \vdots, x_{v'}^d\right] \tag{7}$$

$$x_{s'} = g(E_s(x_s)), x_{s'} = \left[x_{s'}^1, x_{s'}^2, \vdots, x_{s'}^d\right] \tag{8}$$

where α is set to the default value of 0.01.

We use the mutual matching cosine similarity between visual representation $x_v^{'}$ and semantic representation $x_s^{'}$ as the alignment constraint $\mathcal{L}_{ITA}$.

$$\mathcal{L}_{v2s} = -\log \frac{\exp\left(Sim\left(x_{v'}^{i}, x_{s'}^{i}\right)/\tau\right)}{\sum_{b=1}^{B}\exp\left(Sim\left(x_{v'}^{i}, x_{s'}^{b}\right)/\tau\right)} \tag{9}$$

$$\mathcal{L}_{s2v} = -\log \frac{\exp\left(Sim\left(x_{s'}^{i}, x_{v'}^{i}\right)/\tau\right)}{\sum_{b=1}^{B}\exp\left(Sim\left(x_{s'}^{i}, x_{v'}^{b}\right)/\tau\right)} \tag{10}$$

$$\mathcal{L}_{ITA} = \mathcal{L}_{v2s} + \mathcal{L}_{s2v} \tag{11}$$

where B is batch size, τ is a temperature factor, which is initialized as 0.07. The definition of cosine similarity is given in Eq. (5).

Building upon cross-modal matching alignment, to enhance the model's performance in downstream visual classification tasks, we introduce a constraint cross entropy $\mathcal{L}_{CE}$ that combines image prediction results $\hat{y}$ with real labels y and weights it with the matching alignment similarity constraint $\mathcal{L}_{ITA}$ for training the EMA module.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{12}$$

$$\mathcal{L}_{EMA} = \alpha_1 \cdot \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{ITA} \tag{13}$$

where α_1 and β represent constraint weights.

3.3 Cross-Modal Diffusion Reconstruction

In this module, we further interact the visual representations x_v and semantic representations x_s based on the diffusion model. Through the diffusion model, we alleviate the impact of background noise in the visual representation x_v and, with the assistance of semantic representation x_s, generate more robust visual object representations x_r. In the following sections, we will first introduce the background of diffusion models and then describe the cross-modal reconstruction process based on diffusion model.

Background of Diffusion Models. The denoising diffusion probabilistic model mainly consists of two processes: a forward process q with diffusion and noise addition, and a reverse process p with reconstruction and denoising. In the forward process q, Gaussian noise is gradually added to the original training data x over T time steps, following a Markov process:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{14}$$

$$q\left(x_t \mid x_0\right) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right) \tag{15}$$

where $x_0 \sim q(x)$, $\mathcal{N}(.)$ means a Gaussian distribution, β_t determines the noise schedule. $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, utilize the above formula to sample the noisy sample x_t at any step t from x_0.

In reverse process, the data is reconstructed by the model. The optimization objective of the model is to maximize likelihood estimation $p_\theta(x_0)$ of the true data distribution, where θ represents the parameters learned by a neural network.

$$\mu_\theta\left(x_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t\right)\right) \tag{16}$$

$$p_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sigma_\theta\left(x_t, t\right)\right) \tag{17}$$

In the case of conditional guided generation, we have a data pair $(x_0, y_0) \sim (x, y)$. Similar to the above formula, we can derive:

$$\mu_\theta\left(x_t, t, y\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t, y\right)\right) \tag{18}$$

Cross-Modal Reconstruction. We employ a diffusion model to reconstruct the representations extracted by the base model across modalities. Firstly, we design an multi-layer perceptron(MLP) consisting of four linear layers and activation functions for predicting $\hat{x}_s^0 = X_\theta(x_s^t, t, x_v)$ during the reverse process.

In the forward process of the diffusion model, we treat the semantic representation x_s as the input to the diffusion model while using the visual representation x_v as a guiding condition. We smoothly interact the cross-modal representation information by gradually injecting noise. The diffusion model is to minimize the distant between $\hat{x}_s^0$ and x_s^0.

In this process, we construct the representation generation $\hat{x}_s^0$ constraint by calculating mean squared error (MSE) between the reconstructed semantic features $\hat{x}_s^0$ and the original input text features x_s^0. To enhance the performance of the generated representation on downstream tasks, similar to the Embedding Matching Align module, we introduce the cross-entropy constraint to assist in the training of the Cross-Modal Diffusion Recon module:

$$\mathcal{L}_{MSE} = \left\| X_\theta(x_s^t, t, x_v) - x_s^0 \right\|_2^2 \tag{19}$$

$$\mathcal{L}_{CDR} = \alpha_2 \cdot \mathcal{L}_{CE} + \gamma \cdot \mathcal{L}_{MSE} \tag{20}$$

where α_2 and γ represent constraint weights, the $\mathcal{L}_{CE}$ has been given in Eq. (12).

Subsequently, during the reverse process, we initialize random Gaussian noise as the model input. According to Bayes theorem, $p_\theta(x_t - 1|x_t)$ can be calculated according to the following definition:

$$\hat{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{21}$$

$$\mu_\theta\left(x_s^t, t, x_v\right) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} X_\theta\left(x_s^t, t, x_v\right) + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_s^t \tag{22}$$

$$p_\theta\left(x_{t-1}|x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_s^t, t, x_v\right), \hat{\beta}_t I\right) \tag{23}$$

Similarly, guided by the visual representation x_v, we generate the representations x_{CDR} across modalities.

3.4 Multi-modal Embedding Fusion

In the final phase, we combine the representations, x_{EMA} and x_{CDR}, outputted by the previous two modules to realize the complementation and enhancement of information across modalities.

Specifically, we utilize a range of techniques for the fusion of representations, encompassing direct concatenation, addition, multiplication, SUM, and Harmonic(HM) [25]. Based on the integrated representation, we conduct classification tasks, serving as MARNet's final output.

$$\hat{C} = Classifier(x_{EMA} \oplus x_{CDR}) \tag{24}$$

where $\oplus$ represents the process of representation fusion, and $\hat{C}$ denotes the final prediction result.

4 Experiment

4.1 Experiment Settings

Datasets. We conducted various experiments on the task of image classification using the following two datasets:

Vireofood-172 [2]: A single-label classification dataset containing 110,241 dish images across 172 categories, including 353 textual descriptions, averages three texts per image. Following the settings in the original paper, we divided the dataset into 66,071 images for training and 33,154 images for testing.

Ingredient-101 [1]: A single-label classification dataset comprising 93,425 dish images from the Food-101 dataset, featuring 446 common ingredients across 101 categories, averaging 9 ingredients per dish. According to the original paper's settings, we utilized a training set consisting of 68,175 data pairs and a testing set comprising 25,250 data pairs.

Performance Metrics. Since both datasets we used are single-label datasets, we employed accuracy rate as the performance evaluation metric:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{25}$$

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

4.2 Performance Analysis

To verify the effectiveness of our proposed MARNet in enhancing image classification performance and model robustness, we conducted experiments divided into two categories: basic visual classification and cross-modal alignment that incorporates textual information. In the visual network, we selected common structures such as Vision Transformer (ViT) and residual neural networks (ResNet). In the alignment network, we tried different alignment methods based on ViT-B/16 and BERT models, including distance measurement and similarity comparison. Table 1 shows the experimental results.

Table 1. The performance results of state-of-the-art visual neural networks and alignment networks on the datasets. **Acc-1/5** refers to Accuracy Top1/5.

Method	Model	Vireo-Food172		Ingredient-101	
		Acc-1	Acc-5	Acc-1	**Acc-5**
Visual Classification	ResNet-18	77.3	93.2	78.4	93.8
	ResNet-50	81.6	95.0	82.0	94.9
	VGG-19	81.2	95.1	81.4	94.3
	WRN	82.3	95.5	82.9	95.4
	WISeR	82.8	96.5	83.2	95.8
	RepVGG	83.5	96.3	83.6	96.5
	RepMLPNet	83.3	96.2	83.8	96.5
	ViT-B/16	85.4	97.3	88.3	97.6
	ViT-B/32	84.6	97.2	87.7	97.6
	Swin-T	86.5	97.5	88.6	98.1
Cross-modal Alignment	SWD	87.6	97.9	88.6	97.7
	SSAN	87.1	97.7	88.5	97.7
	CDD	86.0	97.0	88.4	97.6
	SDM	87.6	97.7	88.7	97.7
	TEAM	87.6	97.8	88.7	97.8
	ITA	87.8	97.9	88.8	97.8
	MARNet	**88.1**	**98.0**	**89.0**	**97.9**

In experiments on visual networks, we can draw the following conclusions:

– In deep convolutional neural networks, ResNet-50 [8] has more convolutional layers and a deeper network structure compared to ResNet-18 [8], resulting in significant performance improvement. WRN [37] and WISeR [23] further enhance the performance of ResNet-50 by increasing the network's width and introducing feature attention mechanisms, respectively. RepVGG [6] and RepMLPNet [5] significantly enhance the performance of the base model VGG [29] through structural reparameterization.

– Vision Transformer [7] divides the image into small patch blocks and establishes a global understanding of the image through self-attention mechanisms, achieving optimal performance. Compared to ViT-B/32 with larger patch size, ViT-B/16 can better capture the details in image information, thus achieving better performance. Swin-Transformer [19] introduces hierarchical attention to enhance the quality of visual representations, further improving the model's performance.

In alignment network experiments, the following conclusions can be drawn:

– In alignment methods based on distance metrics, Slice WD [13] focuses on multi-modal output spaces and performs well when semantic output is good. Simultaneous Semantic Alignment Network [16] improves visual representation performance by attracting the centroids of representation clusters in latent space. Contrastive Domain Discrepancy [12] uses clustering to make the clusters more compact internally while repelling each other between clusters. However, this method is susceptible to changes in initial representation quality and cluster centroid information, resulting in slightly inferior performance compared to other methods.
– In methods based on contrastive learning, representation pairs are divided into positive and negative samples, enhancing model attention and surpassing distance-based methods. Similarity Distribution Matching [11], Instance-wise Cross-modal Alignment [32], and Token Embeddings Alignment [36] all use cosine similarity to determine sample matching degree. TEAM focuses more on positive pairs through attention, with performance heavily reliant on text quality. ITA simulates unseen negative samples within a mini-batch, increasing the distinction between positive and negative samples. Benefiting from high-quality text information in the dataset, contrastive learning-based alignment methods achieve strong and similar effects.
– Contrastive learning-based alignment methods aim to enhance the similarity between representation pairs in a shared latent space by distinguishing positive and negative sample pairs. Simultaneously, these methods enforce repulsion between matching image-text representation pairs and other non-matching pairs (i.e., negative samples). This alignment approach generally outperforms distance-based methods, which often overlook the negative impact caused by mismatched representation pairs within the same cluster.
– Building upon contrastive learning-based matching alignment, MARNet utilizes a diffusion model guided by visual representations to generate textual representations, effectively extracting crucial textual information from images. Furthermore, we strategically fuse these representations to deepen the interaction between image and text information. As a result, our innovatively proposed network achieves significantly improved performance.

4.3 Ablation Study

To validate the effectiveness of modules in MARNet, we conducted ablative experiments using ViT model as baseline. The results are shown in Table 2.

– Despite the potential noise interference in the image information, the baseline model demonstrates decent performance due to the detailed perception capability of ViT and the assistance of attention mechanisms.
– After incorporating the EMA module and introducing high-quality textual information and alignment through contrastive matching, the model's performance improved significantly. However, the presence of residual noise and interfering factors in the visual information limits the effect of alignment.
– By incorporating the CDR module, we facilitate profound interaction between visual and textual representations to derive representations founded on visual cues, thereby diminishing the impact of peripheral visual elements. Moreover, to mitigate the impact of the MLP component within the CDR module, we introduce a validation process that exclusively leverages the MLP-mapped features for assessing the efficacy of the CDR module.
– In the end, by integrating the EMA and CDR modules, we synergistically enhance the alignment representation and reconstruction representation, further strengthening the visual representation and model robustness.

Table 2. The chart presents the ablation experiment results of MARNet. **Acc-1/5** refers to the Top 1/5 Accuracy.

Module	Vireo-Food172		Ingredient-101	
	Acc-1	Acc-5	Acc-1	Acc-5
Baseline	85.4	97.3	88.3	97.6
+EMA	87.8	97.9	88.8	97.8
+MLP	85.5	95.4	86.4	94.8
+CDR(MLP)	86.9	92.5	88.0	90.5
+Fusion	**88.1**	**98.0**	**89.0**	**97.9**

4.4 Case Study

Reconstructed Feature Visualization. We conducted *t-SNE* visualization on features from the ViT base model and CDR module, selecting 100 samples each from five Vireo-Food172 categories, as depicted in Fig. 3. From the visualization results, it is evident that the reconstruction process based on the diffusion model significantly improved the distribution of the representations and effectively separated the confusing samples in the original representations. Due to the diffusion model generating based on random noise, a small number of unstable samples within the space. However, these minimal instances of noise have negligible impact on the performance of the model, as shown in Table 2.

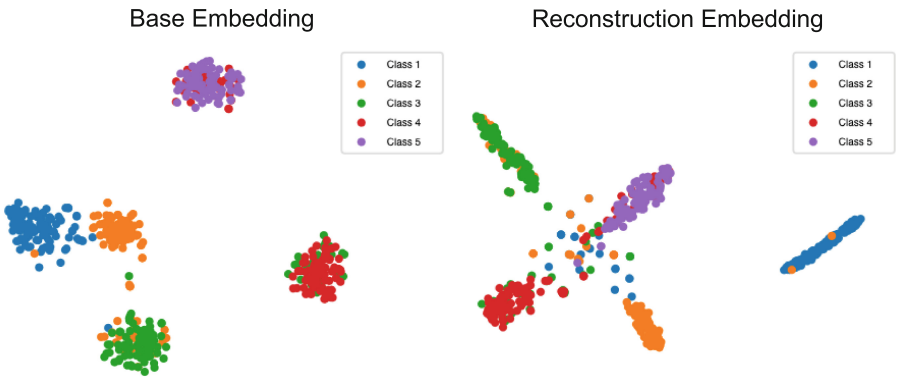

Fig. 3. Visualization of the basic representation x_v and reconstructed representation x_{CDR} using *t-SNE*. As shown in the legend, the color of dots represents the category.

Analysis of CDR Results. In ablation experiments, we can clearly observe a significant decrease in the TOP5 accuracy on the two datasets. We presented and analyzed the prediction results of base visual module and CDR module (shown in Fig. 4): the confidence predicted by base model is typically distributed among top 1–3 classes. However, the diffusion model, which incorporates semantic information, tends to be completely confident in predicting a certain class, with the confidence in other classes stemming more from the randomness of sampling process. This leads to a situation where, when the prediction of the most likely class is incorrect, base model can maintain a relatively high TOP5 accuracy, while the diffusion model struggles to improve.

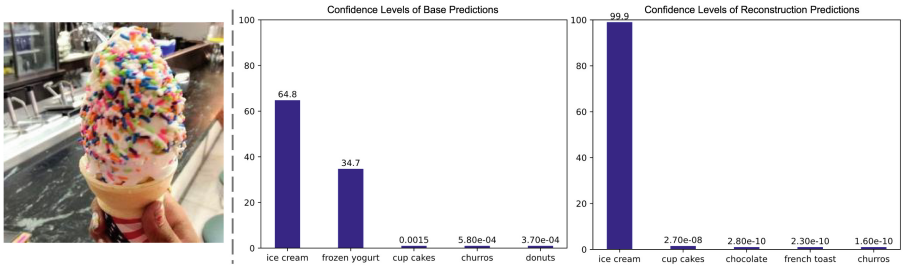

Fig. 4. Prediction results of the basic module and CDR module. The minimal confidence values are represented in scientific notation, e.g., 6.4e-1 indicates 0.64.

5 Conclusion

In this article, we tackle the issue of heterogeneity in multi-modal data by introducing the Multi-Modal Alignment and Reconstruction Network (MARNet).

This network addresses the disparities in distance and distribution within the feature space through a dual approach: embedding matching alignment(EMA) modules and cross-modal diffusion reconstruction(CDR) modules. Our experimental findings validate that MARNet significantly improves the quality of visual information and optimizes the distribution of representations.

Moving forward, our efforts will concentrate on reducing noise interference during reconstruction phase of the diffusion model, with the overarching goal of preserving the integrity of original information to the greatest extent possible.

Acknowledgments. This work is supported in part by the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

References

1. Bolaños, M., Ferrà, A., Radeva, P.: Food ingredients recognition through multi-label learning. In: Battiato, S., Farinella, G.M., Leo, M., Gallo, G. (eds.) ICIAP 2017. LNCS, vol. 10590, pp. 394–402. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70742-6_37
2. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM Multimedia, pp. 32–41 (2016)
3. Chen, Z., Qi, Z., Cao, X., Li, X., Meng, X., Meng, L.: Class-level structural relation modeling and smoothing for visual representation learning. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 2964–2972 (2023)
4. Clark, K., Jaini, P.: Text-to-image diffusion models are zero shot classifiers. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
5. Ding, X., Chen, H., Zhang, X., Han, J., Ding, G.: Repmlpnet: hierarchical vision MLP with re-parameterized locality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 578–587 (2022)
6. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: making VGG-style convnets great again. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733–13742 (2021)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
11. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2787–2797 (2023)
12. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4893–4902 (2019)

13. Lee, C.Y., Batra, T., Baig, M.H., Ulbricht, D.: Sliced Wasserstein discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10285–10295 (2019)
14. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2206–2217 (2023)
15. Li, J., Ma, H., Li, X., Qi, Z., Meng, L., Meng, X.: Unsupervised contrastive masking for visual haze classification. In: Proceedings of the 2022 International Conference on Multimedia Retrieval, pp. 426–434 (2022)
16. Li, S., Xie, B., Wu, J., Zhao, Y., Liu, C.H., Ding, Z.: Simultaneous semantic alignment network for heterogeneous domain adaptation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 3866–3874 (2020)
17. Li, X., Meng, L., Wu, L., Li, M., Meng, X.: Dreamfont3d: personalized text-to-3D artistic font generation. In: ACM SIGGRAPH, pp. 1–11 (2024)
18. Li, X., Zheng, Y., Ma, H., Qi, Z., Meng, X., Meng, L.: Cross-modal learning using privileged information for long-tailed image classification. CVM (2024)
19. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of ICCV, pp. 10012–10022 (2021)
20. Ma, H., et al.: Plug-in diffusion model for sequential recommendation. In: Proceedings of AAAI, pp. 8886–8894 (2024)
21. Ma, H., Xie, R., Meng, L., Yang, Y., Sun, X., Kang, Z.: Seedrec: sememe-based diffusion for sequential recommendation. In: Proceedings of IJCAI, pp. 1–9 (2024)
22. Ma, H., Yang, Y., Meng, L., Xie, R., Meng, X.: Multimodal conditioned diffusion model for recommendation, pp. 1733–1740 (2024)
23. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 567–576. IEEE (2018)
24. Meng, L., et al.: Learning using privileged information for food recognition. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 557–565 (2019)
25. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual VQA: a cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12700–12710 (2021)
26. Oord, A.V.D., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
27. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8238–8247 (2022)
28. Qi, Z., Meng, L., Chen, Z., Hu, H., Lin, H., Meng, X.: Cross-silo prototypical calibration for federated learning with non-IID data. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 3099–3107 (2023)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
30. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
31. Wang, C., Wu, L., Liu, X., Li, X., Meng, L., Meng, X.: Anything to glyph: artistic font synthesis via text-to-image diffusion model. In: SIGGRAPH Asia 2023 Conference Papers, pp. 1–11 (2023)

32. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Advances in Neural Information Processing Systems, vol. 35, pp. 33536–33549 (2022)
33. Wang, Y., Li, X., Liu, Y., Cao, X., Meng, X., Meng, L.: Causal inference for out-of-distribution recognition via sample balancing. CAAI Trans. Intell. Technol. (2024)
34. Wang, Y., et al.: Meta-causal feature learning for out-of-distribution generalization. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13806, pp. 530–545. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25075-0_36
35. Wang, Y., Qi, Z., Li, X., Liu, J., Meng, X., Meng, L.: Multi-channel attentive weighting of visual frames for multimodal video classification. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2023)
36. Xie, C.W., Wu, J., Zheng, Y., Pan, P., Hua, X.S.: Token embeddings alignment for cross-modal retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4555–4563 (2022)
37. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)