Causal Inference over Visual-Semantic-Aligned Graph for Image Classification

Lei Meng[1,2],Xiangxian Li[1,3],Xiaoshuo Yan[1]*,Haokai Ma[1],Zhuang Qi[1],Wei Wu[1], Xiangxu Meng[1]

[1]School of Software, Shandong University, Jinan, China
[2]Shandong Research Institute of Industrial Technology, Jinan, China
[3]School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, China
{lmeng, xiangxianli, mxx}@sdu.edu.cn, {yanxiaoshuo, mahaokai, z_qi, wu_wei}@mail.sdu.edu.cn

Abstract

Incorporating tagging information to regularize the representation learning of images usually leads to improved performance in image classification by aligning the visual features with the textual ones of higher discriminative power. Existing methods typically follow the predictive approach, which uses tags as the semantic labels for visual input to make predictions. However, they typically face the problem of handling the heterogeneity between modalities. In order to learn accurate visual-semantic mapping, this paper presents a visual-semantic causal association modeling framework termed VSCNet. It aligns visual regions with tags, uses a pre-learned hierarchy of visual and semantic exemplars to refine tag predictions and constructs an augmented heterogeneous graph to perform causal intervention. Specifically, the **fine-grained visual-semantic alignment (FVA)** module adaptively locates the semantic-intensive regions corresponding to tags. The **heterogeneous association refinement (HAR)** module associates the visual regions, semantic elements and pre-learned visual prototypes in a heterogeneous graph to filter the error predictions and enrich the information. The **causal inference with graphical masking (CIM)** module applies self-learned masks to discover the causal nodes and edges in the heterogeneous graph to address the spurious association, forming robust causal representations. Experimental results from two benchmarking datasets show that VSCNet effectively builds the visual-semantic associations from images and leads to better performance than the state-of-the-art methods with enriched predictive information.

Introduction

In the field of image classification, achieving consistent representations for each class is challenging due to the diversity of visual factors such as background and lighting conditions in the data (Wang et al. 2021). Prior studies have primarily focused on enhancing the representational learning capabilities of visual models by either refining the architectural design of models (Dosovitskiy et al. 2021) or improving the strategic augmentation of data (Yun et al. 2019). Despite these efforts, a performance bar still persists for these vision-based techniques. To alleviate the problem in visual modality, recent methods have incorporated semantic information, like tags, to guide the learning of visual representations since semantic elements tend to maintain stability across different environments. However, the heterogeneity between visual and semantic modality limits the potential gains from cross-modal enhancement (Meng et al. 2019).

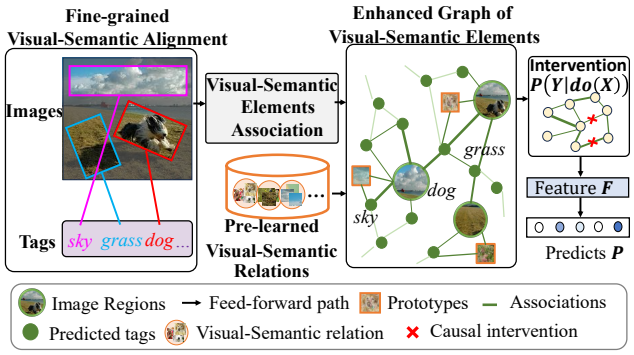

Figure 1: The illustration of the proposed VSCNet, which constructs fine-grained visual-semantic graphs to associate image regions, predicted tags, and visual prototypes, and find causal associations using graphical causal inference.

To this end, recent works mainly adapt semantic tags as privileged information (Vapnik and Vashist 2009) and form three types of approaches to alleviate the heterogeneous problem. The first involves implicit alignment, where tags are treated as semantic labels, applying tagging classification as the auxiliary task of image classification (Xiao et al. 2023; Chen et al. 2021; Jiang et al. 2019; Huo et al. 2024). The second is explicit alignment, which focuses on constraining the similarity between image features and tagging features (Xu, Zhu, and Clifton 2023; Li et al. 2020, 2021; Ye et al. 2023; Li et al. 2024b). The third approach involves fine-grained alignment methods, which locate visual regions and map them to a cross-modal space to apply fine-grained constraints or achieve feature fusion (Min et al. 2019; Chen et al. 2021; Wang et al. 2022; Luo et al. 2023; Wu et al. 2024; Wang et al. 2024a). However, existing methods face challenges in managing heterogeneity at the global level, there are inherent inaccuracies in the mapping from fine-grained visual regions to semantic elements, which leads to errors that propagate from tag predictions to class predictions.

*Corresponding Author

To address the above problems(Xu et al. 2024), this paper proposes a fine-grained visual-semantic causal association modeling network termed VSCNet. Different from previous cross-modal methods, VSCNet associates image regions, tags, and visual prototypes in a heterogeneous graph for representation learning, and it discovers causal associations in the graph for robust image classification. To achieve these goals, VSCNet needs to handle two main challenges: the first is to align the visual and semantic factors accurately, and the second is to handle the data-driven bias in visual-semantic alignment. The overall pipeline is illustrated in Figure 1, VSCNet adaptively locates semantic-intensive regions and aligns these regions with corresponding tags. To achieve more precise alignment, it refines visual-semantic associations using the pre-learned hierarchy of visual and semantic exemplars, thereby constructing an augmented heterogeneous graph. Then, VSCNet employs causal inference with graphical masking to eliminate spurious correlations and discover causal associations related to class within the graph. Finally, it forms causal information as enhanced features to boost image classification performance.

Experiments are conducted on the two common datasets, including performance comparison, ablation study of VSCNet's key components, case studies, and error analysis to assess the efficacy of visual-semantic associations. The results confirm that VSCNet can accurately identify the visual regions corresponding to semantic tags and significantly enhances the robustness of image classification. In summary, the main contributions of this paper are:

- This paper proposes a fine-grained visual-semantic associations modeling network, termed VSCNet, which enhances visual representation learning by using adaptive locating and knowledge-guided filtering to form visual and semantic factors in a heterogeneous graph.
- This paper addresses the challenge of statistical correlations interfering in semantic-guided image classification and proposes a mask-based causal inference module to discover causal factors in heterogeneous graphs.
- The experiments demonstrate that the proposed dynamic location method effectively associates fine-grained visual and semantic information. Furthermore, the hierarchical knowledge successfully models visual patterns of tags, laying a solid foundation for future research in this area.

Related Work

This paper explores modeling the associations between visual and semantic information. Current works primarily achieve this through visual-semantic alignment(Zheng et al. 2024), which can be categorized into three types.

Explicit Alignment by Minimizing Multimodal Feature Discrepancies Explicit alignment methods reduce discrepancies between modalities by explicitly minimizing differences in features or predictions. These methods usually achieve feature alignment by designing distance metric functions (Chen et al. 2023; Chen and Ngo 2016; Jaritz et al. 2022; Aslam et al. 2024) or calculating cross-modal feature distribution differences under specific assumptions (Li et al. 2020; Yao et al. 2022; Ahn et al. 2024). Recently, researchers further address modality heterogeneity by decoupling and aligning features (Meng et al. 2019; Ye et al. 2023; Yang et al. 2022).

Implicit Alignment via Multimodal Information Interacting Implicit alignment methods treat the alignment process as an intermediate or implicit step without directly measuring modal discrepancies. Approaches (Jiang et al. 2019; Xu et al. 2020; George and Marcel 2021) often integrate features or predictions from different modalities to form a unified representation. Additionally, researchers have explored using attention mechanisms (Jiang and Ye 2023) or knowledge distillation (Huo et al. 2024) to model relationships between different modalities. While these methods utilize inter-modal information exchange effectively, they still face challenges related to accurate cross-modal mapping.

Fine-grained Alignment by Modeling Visual and Semantic Associations Works of fine-grained alignment seek to identify the corresponding image regions for tags and then apply implicit alignment for each region to capture finer cross-modal relationships (Xu et al. 2018). However, these methods usually adapt detection models to localize regions (Li et al. 2022; Messina et al. 2021), making it limited in image classification settings. Consequently, researchers are exploring bounding box-free approaches, including dividing feature maps into grids and using max pooling to aggregate grid information (Chen et al. 2021), pre-defining dictionary and learning region features through the combination of feature map and dictionary (Wang et al. 2022), and applying attention maps to create region masks (Luo et al. 2023). There are also methods that perform alignment based on tokens(Li et al. 2024a; Zhao et al. 2024) by contrastive learning (Wu et al. 2024) or designing cross-modal interaction modules (Xie et al. 2022). Different from these methods, VSCNet leverages a pre-learned hierarchy of visual and semantic exemplars along with causal intervention methods to enhance the association between visual regions and semantic factors.

Problem Formulation

Cross-modal enhanced image classification aims to improve visual representation learning with the help of information from another modality. The cross-modal learning dataset typically consists of N images $\mathcal{V} = \{\mathbf{v}_i | i = 1, ..., N\}$, their corresponding semantic tags $\mathcal{T} = \{\mathbf{t}_i | i = 1, ..., N\}$, and labels $Y = \{y_j | j = 1, ..., S\}$ for S classes.

Conventional feature alignment methods for image classification extract visual features $\mathbf{F}_v$ and semantic feature $\mathbf{F}_t$ during the training phase. They then apply alignment functions, such as KL-divergence, to make the distribution of visual features more similar to that of semantic features, forming aligned visual features $\mathbf{F}_{v \to t}$. Finally, these features are fused to achieve the class prediction: $\mathbf{F}_v \otimes \mathbf{F}_{v \to t} \to \mathbf{P}_f$.

Different from previous approaches(Wang et al. 2023; Liu, Li, and Lin 2023), VSCNet derives fine-grained aligned representations $\mathbf{F}_{v \to t}$ and predicted tag $\hat{\mathbf{t}}$. Then, it constructs an augmented heterogeneous graph $\mathcal{G}_h$ using visual regions $\hat{\mathbf{r}}$, corresponding tag predictions $\hat{\mathbf{t}}$ and a pre-learned hierar-

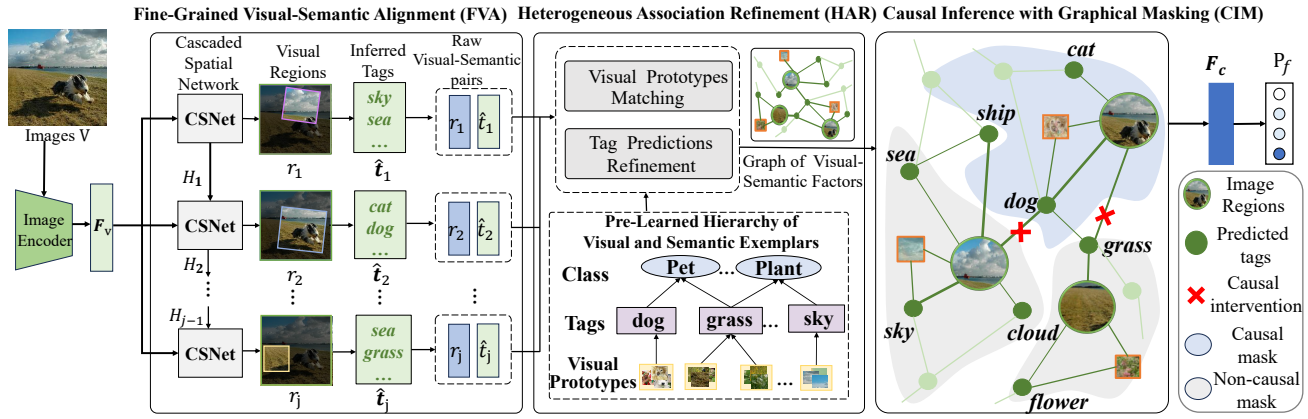

Figure 2: Illustration of the proposed VSCNet. VSCNet adaptively locates semantic-intensive regions and associates region features, predicted tags, and pre-learned visual prototypes as a graph in the HAR module. To find the causal association related to class in the graph, VSCNet applies graphical masking to extract causal information and construct causal features.

chy of visual-semantic exemplars, and forms the graph representations $\mathbf{F}_h$. Finally, causal inference is applied to filter non-causal associations, resulting in a causal graph $\mathcal{G}_c$ and forming causal representations $\mathbf{F}_c$. VSCNet learns the pipeline $\mathbf{F}_{v\to t} \to \mathbf{F}_h \to \mathbf{F}_c \otimes \mathbf{F}_v \to \mathbf{P}_f$ of three stages:

(1) **Fine-grained visual-semantic alignment**: VSCNet adaptively locates semantic-intensive visual regions $\mathbf{r}$ within images, and getting predicted tags $\hat{\mathbf{t}}$ of these regions.

(2) **Heterogeneous association refinement**: VSCNet introduces a pre-learned hierarchy of visual and semantic exemplars VSE $= \{\mathbf{p} \to \mathbf{t} \to y\}$ to construct an augmented fine-grained visual-semantic graph $\mathcal{G}_h = \{(\mathbf{r}, \hat{\mathbf{t}}_f, \mathbf{p}), \mathcal{A}_h\}$. $\mathbf{p}$ is the visual prototype obtained by clustering visual regions $\mathbf{r}$, while $\hat{\mathbf{t}}_f$ is the tags prediction $\hat{\mathbf{t}}$ refined by VSE.

(3) **Causal inference and multi-view fusion**: Through the graphical causal inference, the graph is further refined as a causal graph $\mathcal{G}_c$, and forming causal features $\mathbf{F}_c$. Regional causal features $\mathbf{F}_c$ are then integrated with global features $\mathbf{F}_v$ for outputting visual predictions $\mathbf{F}_v \otimes \mathbf{F}_c \to \mathbf{P}_f$.

Approach

Cross-modal enhanced image classification struggles with modality heterogeneity, resulting in inaccurate associations between visual and semantic information, which lead to misclassifications. To address these problems, VSCNet first adaptively locates semantic-intensive regions and aligns them with corresponding tags in the fine-grained visual-semantic alignment (FVA) module. Secondly, VSCNet leverages a pre-learned hierarchy of visual and semantic exemplars to construct and refine the associations between visual regions and predicted tags in the heterogeneous association refinement (HAR) module. Finally, VSCNet discovers causal associations in the causal inference with graphical masking (CIM) module by applying a soft mask to identify causal factors for classification, as shown in Figure 2.

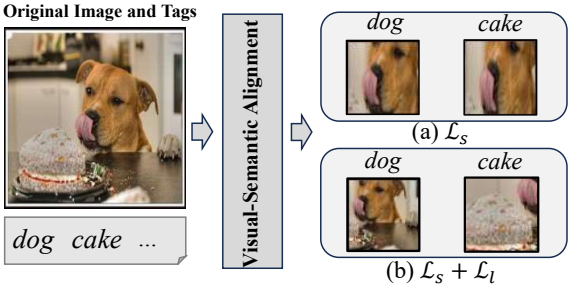

Figure 3: The illustration of visual-semantic alignment under different constraints. The $\mathcal{L}_l$ loss encourages the model to explore diverse semantic-intensive regions.

Fine-Grained Visual-Semantic Alignment (FVA)

To explore the fine-grained associations between visual and semantic factors(Guan et al. 2023), the initial challenge is to locate corresponding semantic tags within the image. The FVA module introduces the Cascaded Spatial Net (CSNet) to achieve more accurate alignment.

Given an image $\mathbf{v}$, VSCNet first extracts the global feature $\mathbf{F}_v$ using a visual encoder. Then CSNet extracts features $\{\mathbf{r}_1, ..., \mathbf{r}_J\}$ of various semantic-intensive regions and sequentially predicts the corresponding tags $\{\hat{\mathbf{t}}_1, ..., \hat{\mathbf{t}}_J\}$:

$$\mathbf{r}_j, \hat{\mathbf{t}}_j, \mathbf{H}_j = CSNet(\mathbf{F}_v, \mathbf{H}_{j-1}), \tag{1}$$

where j ranges from 1 to J, $\mathbf{H}_j$ is a learnable hidden vector and ensures the distinctiveness of region localization, and $\mathbf{H}_0$ is initialed with zero.

CSNet uses a grid generator $\mathcal{T}(.)$ to sample regions $\mathbf{r}_j$ from the whole image, using a learnable affine matrix θ, global features $\mathbf{F}_v$, and hidden vectors $\mathbf{H}_{j-1}$ from last state:

$$\mathbf{r}_j = \mathcal{T}_\theta(\mathbf{F}_v, \mathbf{H}_{j-1}) = \mathcal{T}\left(\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}, \mathbf{F}_v, \mathbf{H}_{j-1}\right), \tag{2}$$

where θ_{11} and θ_{22} control the scale, θ_{12} and θ_{21} control the

rotation, and θ_{13} and θ_{23} control the translation.

The ground-truth labels y and tags $\{\mathbf{t}_1, ..., \mathbf{t}_J\}$ of images are introduced to constrain the locating:

$$\mathcal{L}_s = -(\sum_{i=1}^{S} y_i \log(\hat{y}_i) + \sum_{j=1}^{J} \sum_{i=1}^{M} t_{ji} \log(\hat{t}_{ji})), \quad (3)$$

where $\hat{\mathbf{t}}_{ji}$ is the tag predictions of image region of $\mathbf{r}_j$, M is the total number of tags in the dataset.

As shown in Figure 3, an adaptive locating loss has been proposed to encourage CSNet to explore more semantic-intensive regions, which is formulated as:

$$\mathcal{L}_l = \sum_{j}^{J} (exp(-|\theta_{13}^{(j)} - \bar{\theta}_{13}| - |\theta_{23}^{(j)} - \bar{\theta}_{23}|) + |\theta_{12}^{(j)}| + |\theta_{21}^{(j)}|), \quad (4)$$

where $\theta^{(j)}$ is the affine matrix of the j-th region, and $\bar{\theta}^{(j)}$ is the average of affine matrix.

Heterogeneous Association Refinement (HAR)

To identify more precise correspondences between visual regions and predicted tags, VSCNet first constructs a pre-learned hierarchy of visual and semantic exemplars and matches the visual prototypes with the visual regions; secondly, it refines tag predictions and constructs an augmented heterogeneous graph after visual prototypes matching.

Visual Prototype Matching The heterogeneity of the modalities leads to incorrect tag predictions, resulting in the omission of precise visual-semantic associations. Analysis of visual-semantic predictions reveals that correct tags often appear among the top results. Therefore, VSCNet proposes a pre-learned hierarchy of visual and semantic exemplars.

The pre-learned hierarchy of visual and semantic exemplars is constructed by clustering the regions $\mathbf{r}$ from the training set. And stipulate regions $\mathbf{r}$ from the same tag $\hat{\mathbf{t}}$ in same class y can be clustered:

$$\mathbf{p} = Cluster(\mathbf{r}, \hat{\mathbf{t}}, y), \quad (5)$$

where $\mathbf{p}$ is the visual prototype of the corresponding tag, which is represented as the center of the cluster.

The pre-learned hierarchy of visual and semantic exemplars VSE $= \{\mathbf{p} \rightarrow \hat{\mathbf{t}} \rightarrow y\}$ records the visual prototypes for tags in classes. Then, VSCNet matches the region $\mathbf{r}$ with prototypes $\mathbf{p}$ in VSE. The matched prototypes $\mathbf{p}_r$ are connecting to tagging distribution $\mathbf{t}_r$ and class predictions $\mathbf{P}_r$ from the VSE , which is illustrated as:

$$\{\mathbf{p}_r, \mathbf{t}_r, \mathbf{P}_r\} = \phi_K(\mathbf{r}), \quad (6)$$

where $\phi_K(.)$ is the operation of calculating the similarity between regions and prototypes for matching.

Tag Prediction Refinement VSCNet refines the tag predictions of visual regions after matching the visual prototypes. Then, VSCNet constructs an augmented heterogeneous graph composed of visual regions, refined tag predictions, and visual prototypes.

VSCNet refines visual-semantic associations by re-weighting previous tag predictions based on the tagging distribution $\mathbf{t}_r$ of the matched visual prototypes $\mathbf{p}_r$:

$$\hat{\mathbf{t}}_f = (1 - \alpha_f) Softmax(\hat{\mathbf{t}}) + \alpha_f Softmax(\mathbf{t}_r), \quad (7)$$

where α_f is the weight of semantic filtering.

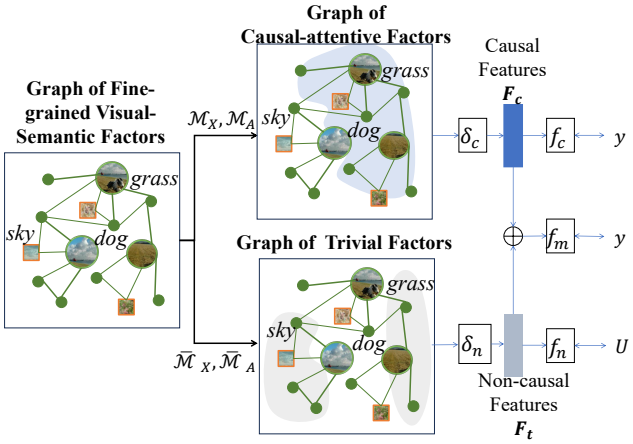

Figure 4: VSCNet uses graphical masking to achieve causal inference on the heterogeneous graph.

Then, VSCNet constructs an augmented fine-grained visual-semantic association graph to represent the relations between visual and semantic factors. The graph $\mathcal{G}_a$ contains nodes of visual regions $\mathbf{r}$ and predicted tags $\hat{\mathbf{t}}_f$:

$$\mathcal{G}_a = \{(\mathbf{r}, \hat{\mathbf{t}}_f), \mathcal{A}_a\}, \quad (8)$$

where A_a is the adjacent matrix. Edges exist between regions and corresponding tags, tags predicted from the same region, and visual regions predicting the same tag.

To enrich the information of heterogeneous graph, prototypes $\mathbf{p}$ from the VSE are integrated as a new type of node into the heterogeneous graph $\mathcal{G}_h$:

$$\mathcal{G}_h = \{(\mathbf{r}, \hat{\mathbf{t}}_f, \mathbf{p}), \mathcal{A}_h\}, \quad (9)$$

where $\mathcal{A}_h$ is the adjacent matrix. Edges exist between region nodes and their most closely matching prototype nodes, visual prototype nodes and semantic tag nodes.

Causal Inference with Graphical Masking (CIM)

After knowledge-based refining, the noise in visual-semantic mapping in graph $\mathcal{G}_h$ is alleviated. However, its efficacy in eradicating spurious correlations between nodes is limited because it relies on statistical information. To address this problem, the causal inference with graphical masking (CIM) module applies causal inference to extract causal factors about classification from the heterogeneous graph, thereby enhancing the robustness of predictions.

Graphical Factors Disentangling As shown in Figure 4, to disentangle the causal and non-causal aspects of graph $\mathcal{G}_h$, self-learning node mask $\mathcal{M}_X$ and edge mask $\mathcal{M}_A$ are utilized to differentiate the causal and non-causal aspects of features (nodes) and associations (edges) on the graph $\mathcal{G}_h$:

$$\mathcal{G}_c = \{\mathcal{A} \odot \mathcal{M}_A, \mathbf{X} \odot \mathcal{M}_X\}, \quad (10)$$

$$\mathcal{G}_n = \{\mathcal{A} \odot \overline{\mathcal{M}}_A, \mathbf{X} \odot \overline{\mathcal{M}}_X\}, \quad (11)$$

where $\mathcal{G}_c$ is the graph of causal-attentive factors, $\mathcal{G}_n$ is the graph of non-causal factors. $\overline{\mathcal{M}}_X = \mathbf{1} - \mathcal{M}_X$, $\overline{\mathcal{M}}_A =$

$\mathbf{1} - \mathcal{M}_A$, standing for complementary masks. The node features $\mathbf{X}$ are a concatenation of the visual region node feature, semantic tag node feature, and prototype node feature.

Building on the causal aspects $\mathcal{G}_c$ and non-causal aspects $\mathcal{G}_n$ derived from the heterogeneous graph $\mathcal{G}_h$, VSCNet can obtain the causal feature $\mathbf{F}_c = \delta_c(\mathcal{G}_c)$ and non-causal feature $\mathbf{F}_n = \delta_n(\mathcal{G}_n)$ respectively, where δ_c and δ_n are GCN layers designed to extract causal and non-causal features.

Causal Intervention The backdoor adjustment is used to eliminate spurious correlations by integrating causal features $\mathbf{F}_c$ of the current graph with non-causal features $\mathbf{F}_n$ from other graphs. This integration aims to prevent non-causal features specific to any graph from misleadingly influencing category labels. The process can be concluded as:

$$P(Y|do(C)) = \sum_{n \in \mathcal{G}_n}^{N} P(Y|C, n)P(n), \quad (12)$$

where C is the causal feature from current sample and n is the confounding non-causal feature from $\mathcal{G}_n$.

To achieve backdoor adjustment, VSCNet makes the implicit intervention on the representation level and proposes the following loss guided by the backdoor adjustment. For causal aspects, we set the optimization objective to ground-truth label y to ensure sufficient information for classification. The objective of non-causal aspects is to capture information irrelevant to classification, ensuring these features contribute to a uniform prediction distribution U across all categories. Finally, the training loss can be formalized as:

$$\mathcal{L}_{gf} = \mathcal{L}_{KL}(f_n(\mathbf{F}_n), U) + \mathcal{L}_2(f_c(\mathbf{F}_c), y) + \mathcal{L}_2(f_m(\mathbf{F}_m), y), \quad (13)$$

where U is the uniform distribution, f_c, f_n, f_m are classifiers for causal, non-causal, intervened features and $\mathcal{L}_{KL}$, $\mathcal{L}_2$ stand for KL divergence loss and L2 norm loss.

Decision Making with Multi-view Fusion The features $\mathbf{F}_e$ from the causal graph are combined with the global features $\mathbf{F}_v$ to form enhanced causal visual representations $\mathbf{F}_e = \mathbf{F}_c \oplus \mathbf{F}_v$. Among them, $\oplus$ is implemented by linear projection in both features and follows an add operation.

Finally, predictions $\mathbf{P}_e = \phi(\mathbf{F}_e)$ is refined by the $\mathbf{P}_k$ from the hierarchy of visual and semantic exemplars, where $\phi(.)$ is a linear classifier. This can be formed as:

$$\mathbf{P}_f = Softmax\left(\mathbf{P}_e\right) \otimes Softmax\left(\mathbf{P}_k\right), \quad (14)$$

where $\otimes$ is the summation of predictions.

Training Strategies

The training of VSCNet follows two steps: The first step is training Cascaded Spatial Network. After refining tagging predictions and constructing an augmented graph, the augmented graph can be used directly in the second step for heterogeneous causal graph-based image classification. The details are as follows:

- **Step 1 (Train of Cascaded Spatial Network):** CSNet locates the association between visual regions $\mathbf{r}$ and tagging predictions $\hat{\mathbf{t}}$. The process is constrained by:

$$\mathcal{L}_{AL} = \mathcal{L}_s + \alpha_l \mathcal{L}_l, \quad (15)$$

where α_l is the weight factor.

Datasets	#Classes	#Words	#Training	#Testing
Ingredient-101	101	446	68,175	25,250
NUS-WIDE	81	1000	121,962	81,636

Table 1: Statistics of the datasets used in the experiments.

- **Step 2 (Train of causal intervention and graph fusion networks):** Freezing Cascaded Spatial Net and training causal network with graphical masking and then integrating global features $\mathbf{F}_v$. The inference of the image classification is constrained by:

$$\mathcal{L}_{ff} = \mathcal{L}_{cls}(\mathbf{P}_e, Y) + \alpha_{gf}\mathcal{L}_{gf}, \quad (16)$$

where α_{gf} is the weight factor.

Experiments

Experiment Settings

Datasets To demonstrate the generality of VSCNet in different domains, the cross-modal image classification datasets Ingredient-101 (Bolaños, Ferrà, and Radeva 2017) and NUS-WIDE (Chua et al. 2009) are used in the experiments. Table 1 shows their statistics information.

Evaluation Protocol For the Ingredient-101 dataset of single-class prediction task, we use Top-1 and -5 accuracies following previous works (Meng et al. 2019; Jiang et al. 2019) to evaluate the performance. For the NUS-WIDE dataset, we followed the original setups (Chua et al. 2009) to use Top-1 and -5 precision and recall.

Implementation Details During training, the batch size is fixed at 64 and the Adam optimizer is adapted with a learning rate chosen from 5e-5 to 5e-3. The decay rate of learning rate was chosen from 0.1 and 0.5 with a interval of 4 epochs. For VSCNet, since it is a model-agnostic framework, we investigated the ResNet18 (RN18) (He et al. 2016), ResNet50 (RN50) and ViT (Dosovitskiy et al. 2021) as base models. We use ART (Meng, Tan, and Wunsch 2015; Meng, Tan, and Xu 2013; Wang et al. 2024b; Chen et al. 2023) as our clustering algorithm, and set the clustering parameter from 0.8 to 0.9. During the training of the CSNet, we set the upper bound factor as 0.2 and the lower bound factor as 0.8, respectively. The loss weights α_l are chosen from $\{0.1, 1, 10, 100\}$. The weight of filtering α_f is selected from 0.1 to 0.5. The top number J of regions is selected from $\{5, 10, 15\}$, the number M of regions is selected from $\{3, 4, 5\}$.

Performance Comparison

We compared VSCNet with eleven state-of-the-art methods: including ResNet-18 (He et al. 2016), ResNet-50 (He et al. 2016), ViT (Dosovitskiy et al. 2021), RepVGG (Ding et al. 2021), RepMLPNet (Ding et al. 2022), VanillaNet (Chen et al. 2024), CMFL (George and Marcel 2021) and IRRA (Jiang and Ye 2023), MSMVFA (Jiang et al. 2019), SSAN (Li et al. 2020), Disentangle-VAE (Li et al. 2021), and C^{2}KD (Huo et al. 2024), CHAN (Pan, Wu, and Zhang 2023) and MGCC (Wu et al. 2024). To make a fair comparison, the hyper-parameters of all models are chosen in above section, due to the specific nature of the MGCC method, it uses ViT

Methods	Algorithms	Reference	Ingredient-101		NUS-WIDE			
			Acc@1	Acc@5	P@1	P@5	R@1	R@5
Visual Modality	ResNet18	CVPR'16	0.784	0.938	0.785	0.391	0.439	0.846
	ResNet50	CVPR'16	0.820	0.949	0.786	0.391	0.440	0.846
	RepVGG	CVPR'21	0.836	0.965	0.797	0.394	0.448	0.856
	Repmlpnet	CVPR'22	0.838	0.965	0.801	0.405	0.453	0.877
	VanillaNet	NeurIPS'24	0.845	0.960	0.801	0.395	0.456	0.857
	ViT	ICLR'21	0.854	0.967	0.796	0.395	0.445	0.855
Implicit Alignment	MSMVFA	CVPR'19	0.841	0.965	0.798	0.395	0.443	0.855
	CMFL	CVPR'21	0.810	0.956	0.809	0.402	0.456	0.871
	IRRA	CVPR'23	0.849	0.967	0.797	0.393	0.447	0.852
Explicit Alignment	SSAN	MM'20	0.838	0.963	0.803	0.396	0.450	0.855
	D-VAE	AAAI'21	0.830	0.958	0.811	0.402	0.459	0.872
	C^2KD	CVPR'24	0.832	0.962	0.815	0.400	0.458	0.865
Fine-grained Alignment	CHAN	CVPR'23	0.846	0.970	0.810	0.403	0.461	0.871
	$MGCC_{ViT}$	AAAI'24	0.871	**0.975**	0.821	0.412	0.468	0.894
	$VSCNet_{RN18}$	Ours	0.829	0.957	0.812	0.405	0.465	0.879
	$VSCNet_{RN50}$	Ours	0.865	0.967	0.822	0.409	0.470	0.887
	$VSCNet_{ViT}$	Ours	**0.886**	0.973	**0.826**	**0.413**	**0.474**	**0.894**

Table 2: Performance comparison of algorithms. Metrics are Top-1/Top-5 Accuracy (Acc), Precision (P), and Recall (R).

as the backbone, while all other alignment methods use pre-trained ResNet-50 as the backbone. The following observations are drawn from Table 2:

- **VSCNet achieved significant improvements based on different backbones**. It demonstrates the effects of modeling fine-grained associations for image classification and shows the backbone agnostic character of VSCNet.
- **VSCNet generally achieved better performance than other algorithms in both datasets.** It is reasonable since VSCNet is able to generate a "visual-semantic-class" hierarchy to utilize the predicted visual-semantic associations to accurately infer the image classes.
- **ViT and VanillaNet have significantly improved the performance compared to other visual backbones**. Benefiting from the discriminative power of transformer and activation function, ViT and VanillaNet outperformed other methods on both datasets up to 8%.
- **The explicit alignment methods usually obtain better performance than implicit alignment**. It demonstrates that learning visual features can benefit more from explicit feature alignment than implicit regularization.
- **In fine-grained alignment methods, regional positioning and information aggregation affect its performance**. MGCC shows high top-5 performance, but poor information aggregation downgrades the top-1 performance compared to VSCNet.

Ablation Study

In this section, we further studied the working mechanisms of different modules of VSCNet, as shown in Table 3. The following findings could be observed:

- **The noise in visual-semantic inference downgrades the performance gains.** Since visual-semantic mapping incurs predictive uncertainties, using solely the predicted tags to predict classes (+L) may not bring improvement, even when combined with the heterogeneous graph (+G).

Models	Ingredient-101		NUS-WIDE	
	Acc@1	Acc@5	P@1	R@1
Base	0.854	0.967	0.796	0.445
+L	0.846	0.912	0.777	0.436
+L+G	0.858	0.932	0.778	0.437
+L+K	0.862	0.934	0.781	0.439
+L+K+G	0.869	0.959	0.822	0.469
+L+K+G+C	0.871	0.960	0.824	0.472
+L+K+G+C+F	**0.886**	**0.973**	**0.827**	**0.473**

Table 3: Ablation study of VSCNet with ViT backbone.

- **With the pre-learned hierarchical visual-semantic exemplars (+K), filtered tags achieve better classification performance.** The results of (+L+K) are comparable to the base model on Ingredient-101, demonstrating the effectiveness of HAR for handling noise. While it shows limited improvement on NUS-WIDE with noisy semantics, but the refined associations are beneficial when combine with heterogeneous graph (+L+K+G).
- **Fusion of information through heterogeneous graphs and causal inference (+C) in graphs significantly improve model performance.** Causal intervention effectively adjusts the associating weights between nodes, making the relationships in heterogeneous graphs more accurate. Further multi-view fusion (+F) of information also improves the classification effect.

Case Study

Effect of Cascaded Spatial Net in Visual-Semantic Alignment This section further studies the quality of visual-semantic associations learned by the CSNet. As shown in Figure 5, regions can well correspond to semantic tags. Obviously, in case (a) and case (c), they are just local scaling of the original image. Although the regions for "cloud" and "city" have a large rotation and stretching in cases (b) and

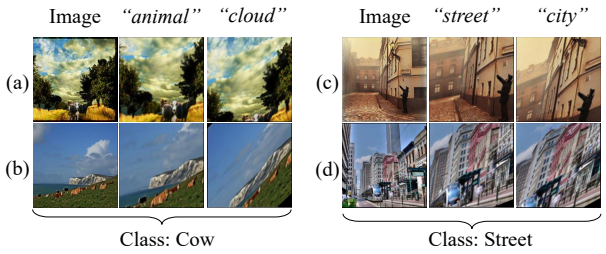

Figure 5: The effect of visual semantic predictions on located visual regions with different semantic tags.

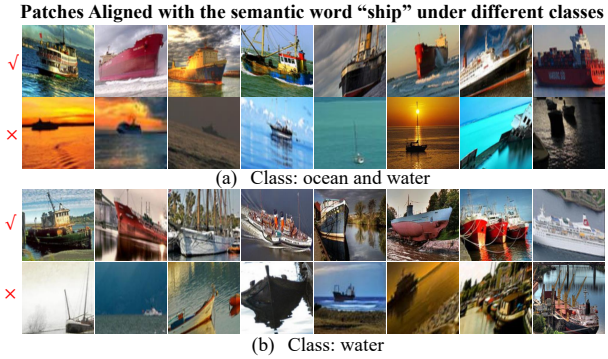

Figure 6: Showcase of the retained and discarded clusters of our generated visual-semantic hierarchy.

(d). In general, they can still retain the corresponding semantics. Notably, the regions generated by CSNet are easy to build visual-semantic connections for specific tags (such as "animal" and "street"). This is mainly because they have regular forms of vision. For some semantics with diversity in visual form (such as "cloud" and "city"), the resulting regions are usually getting more rotated and scaled for searching the attention region of corresponding semantic tags.

Quality Estimation of Pre-learned Visual-Semantic Exemplars As shown in Figure 6, the hierarchy of visual-semantic exemplars discards some low-quality patterns and retains the high-quality patterns. The semantic tag "ship" exhibits various behaviors in different clusters. For the class "ocean and water", the features of "ship" in the first line are clearly visible, but not obvious in the second line. The reason lies in the poor quality of the original image. The hierarchical association filtering filters the patterns shown in the second line. Obviously, for the class "water", the visual pattern in lines three and four are chaotic and unified respectively. And the patterns with no clear features are discarded and the patterns with visible features are retained.

Error Analysis

This section provides an analysis of the VSCNet in success and failure cases, where VSA is visual-semantic alignment, TPR is tag predictions refinement, VPM is visual prototype matching, CMI is causal inference with graphical masking, and MF is multi-view fusion. As observed in Figure 7 (a), the

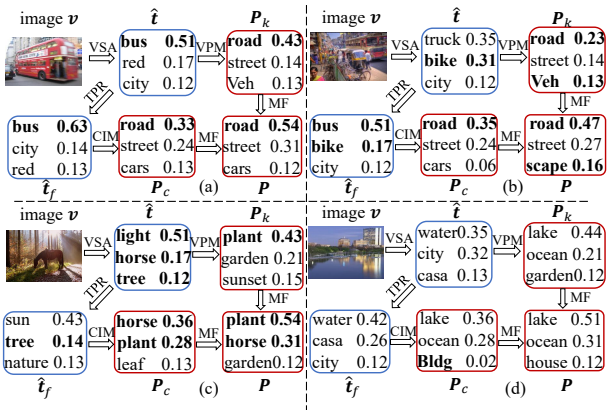

Figure 7: Error analysis of VSCNet. The ground-truth labels are red in predictions. (a) Both visual and knowledge-based predictions work for easy cases. (b) TPR refines the semantic predictions, leading to improved content-based and fused predictions. (c) Failures in TPR refinement can be alleviated by global information. (d) Small visual targets may lead to wrong predictions.

visual and knowledge-based predictions successfully predicted the correct class, and fused for robust predictions. In Figure 7 (b), for multiple semantic prediction, the predicted semantics only hits one correct tag, and the WPR refined some wrong predictions, which helps the raw and knowledge-based prediction make correct decisions. Figure 7 (c) shows the case that semantic features are similar to others not in the image, refined semantics may fail and lead to a decline in the performance of knowledge-based predictions, but the raw predictions can help to reduce the error in fusion. Figure 7 (d) depicts the case that the predicted and refined semantics make wrong predictions for small targets. The knowledge-based and raw predictions failed in image recognition, which makes the model unable to predict the correct class. They demonstrate that VSCNet can produce more robust classification results by filtering misinformation and fusing multi-channel knowledge.

Conclusion

This paper presents a novel approach termed VSCNet, to achieve fine-grained alignment in image classification, which can dynamically locate semantically meaningful visual regions without the supervision of bounding boxes. Then, pre-learned hierarchy of visual and semantic exemplars and graphical masking module are created to handle the predictive errors in cross-modal inference. Our future work has two key concerns. First, the cost of constructing VSE will increase as the scale of datasets grows, the balance of efficiency and accuracy as well as applying it to other fields like federated learning (Qi et al. 2024; Liu, Li, and Lin 2023; Meng et al. 2024) are important. Second, it is worth exploring whether the fine-grained alignment could benefit the foundation models in their training or fine-tuning.

Acknowledgements

This work is supported in part by the Shandong Province Excellent Young Scientists Fund Program (Overseas) (Grant no. 2022HWYQ-048), the TaiShan Scholars Program (Grant no. tsqn202211289)

References

Ahn, W.-J.; Yang, G.-Y.; Choi, H.-D.; and Lim, M.-T. 2024. Style Blind Domain Generalized Semantic Segmentation via Covariance Alignment and Semantic Consistence Contrastive Learning. In *CVPR*, 3616–3626.

Aslam, M. H.; Zeeshan, M. O.; Belharbi, S.; Pedersoli, M.; Koerich, A. L.; Bacon, S.; and Granger, E. 2024. Distilling privileged multimodal information for expression recognition using optimal transport. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–10. IEEE.

Bolaños, M.; Ferrà, A.; and Radeva, P. 2017. Food ingredients recognition through multi-label learning. In *ICIAP*, 394–402.

Chen, H.; Wang, Y.; Guo, J.; and Tao, D. 2024. Vanillanet: the power of minimalism in deep learning. *In NeurIPS*, 36.

Chen, J.; and Ngo, C.-W. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *MM*, 32–41.

Chen, J.; Zhu, B.; Ngo, C.-W.; Chua, T.-S.; and Jiang, Y.-G. 2021. A Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition. *IEEE Transactions on Image Processing*, 30: 1514–1526.

Chen, Z.; Qi, Z.; Cao, X.; Li, X.; Meng, X.; and Meng, L. 2023. Class-level Structural Relation Modeling and Smoothing for Visual Representation Learning. In *MM*, 2964–2972.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 1–9.

Ding, X.; Chen, H.; Zhang, X.; Han, J.; and Ding, G. 2022. Repmlpnet: Hierarchical vision mlp with re-parameterized locality. In *CVPR*, 578–587.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. Repvgg: Making vgg-style convnets great again. In *CVPR*, 13733–13742.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

George, A.; and Marcel, S. 2021. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, 7882–7891.

Guan, Q.-L.; Zheng, Y.; Meng, L.; Dong, L.-Q.; and Hao, Q. 2023. Improving the generalization of visual classification models across IoT cameras via cross-modal inference and fusion. *IEEE Internet of Things Journal*, 10(18): 15835–15846.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *CVPR*, 16006–16015.

Jaritz, M.; Vu, T.-H.; De Charette, R.; Wirbel, É.; and Pérez, P. 2022. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 1533–1544.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2787–2797.

Jiang, S.; Min, W.; Liu, L.; and Luo, Z. 2019. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29: 265–276.

Li, B.; Shi, Y.; Yu, Q.; and Wang, J. 2024a. Unsupervised Cross-Domain Image Retrieval via Prototypical Optimal Transport. In *AAAI*, 4, 3009–3017.

Li, J.; He, X.; Wei, L.; Qian, L.; Zhu, L.; Xie, L.; Zhuang, Y.; Tian, Q.; and Tang, S. 2022. Fine-grained semantically aligned vision-language pre-training. *In NeurIPS*, 7290–7303.

Li, S.; Xie, B.; Wu, J.; Zhao, Y.; Liu, C. H.; and Ding, Z. 2020. Simultaneous semantic alignment network for heterogeneous domain adaptation. In *MM*, 3866–3874.

Li, X.; Xu, Z.; Wei, K.; and Deng, C. 2021. Generalized zero-shot learning via disentangled representation. In *AAAI*, 1966–1974.

Li, X.; Zheng, Y.; Ma, H.; Qi, Z.; Meng, X.; and Meng, L. 2024b. Cross-modal learning using privileged information for long-tailed image classification. *Computational Visual Media*, 1–12.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Luo, M.; Min, W.; Wang, Z.; Song, J.; and Jiang, S. 2023. Ingredient Prediction via Context Learning Network with Class-Adaptive Asymmetric Loss. *IEEE Transactions on Image Processing*.

Meng, L.; Chen, L.; Yang, X.; Tao, D.; Zhang, H.; Miao, C.; and Chua, T.-S. 2019. Learning using privileged information for food recognition. In *MM*, 557–565.

Meng, L.; Qi, Z.; Wu, L.; Du, X.; Li, Z.; Cui, L.; and Meng, X. 2024. Improving Global Generalization and Local Personalization for Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Meng, L.; Tan, A.-H.; and Wunsch, D. C. 2015. Adaptive scaling of cluster boundaries for large-scale social media data clustering. *IEEE transactions on neural networks and learning systems*, 27(12): 2656–2669.

Meng, L.; Tan, A.-H.; and Xu, D. 2013. Semi-supervised heterogeneous fusion for multimedia data co-clustering. *IEEE Transactions on Knowledge and Data Engineering*, 26(9): 2293–2306.

Messina, N.; Amato, G.; Esuli, A.; Falchi, F.; Gennaro, C.; and Marchand-Maillet, S. 2021. Fine-grained visual textual

alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4): 1–23.

Min, W.; Liu, L.; Luo, Z.; and Jiang, S. 2019. Ingredient-guided cascaded multi-attention network for food recognition. In *MM*, 1331–1339.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network. In *CVPR*, 19275–19284.

Qi, Z.; He, W.; Meng, X.; and Meng, L. 2024. Attentive modeling and distillation for out-of-distribution generalization of federated learning. In *ICME*, 1–6. IEEE.

Vapnik, V.; and Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6): 544–557.

Wang, R.; Qi, Z.; Meng, X.; and Meng, L. 2023. Learning to fuse residual and conditional information for video compression and reconstruction. In *ICIG*, 360–372.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal attention for unbiased visual recognition. In *CVPR*, 3091–3100.

Wang, Y.; Li, X.; Liu, Y.; Cao, X.; Meng, X.; and Meng, L. 2024a. Causal inference for out-of-distribution recognition via sample balancing. *CAAI Transactions on Intelligence Technology*.

Wang, Y.; Meng, L.; Ma, H.; Wang, Y.; Huang, H.; and Meng, X. 2024b. Modeling Event-level Causal Representation for Video Classification. In *MM*, 3936–3944.

Wang, Z.; Min, W.; Li, Z.; Kang, L.; Wei, X.; Wei, X.; and Jiang, S. 2022. Ingredient-Guided Region Discovery and Relationship Modeling for Food Category-Ingredient Prediction. *IEEE Transactions on Image Processing*.

Wu, X.; Ma, W.; Guo, D.; Zhou, T.; Zhao, S.; and Cai, Z. 2024. Text-based Occluded Person Re-identification via Multi-Granularity Contrastive Consistency Learning. In *AAAI*, 6, 6162–6170.

Xiao, L.; Wu, X.; Yang, S.; Xu, J.; Zhou, J.; and He, L. 2023. Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, 60(6): 103508.

Xie, C.-W.; Wu, J.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022. Token embeddings alignment for cross-modal retrieval. In *MM*, 4555–4563.

Xu, H.; Qi, G.; Li, J.; Wang, M.; Xu, K.; and Gao, H. 2018. Fine-grained Image Classification by Visual-Semantic Embedding. In *IJCAI*, 1043–1049.

Xu, N.; Gao, Y.; Liu, A.-A.; Tian, H.; and Zhang, Y. 2024. Multi-Modal Validation and Domain Interaction Learning for Knowledge-based Visual Question Answering. *IEEE Transactions on Knowledge and Data Engineering*.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Xu, X.; Lin, K.; Yang, Y.; Hanjalic, A.; and Shen, H. T. 2020. Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3030–3047.

Yang, X.; Wang, S.; Dong, J.; Dong, J.; Wang, M.; and Chua, T.-S. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31: 1204–1216.

Yao, S.; Kang, Q.; Zhou, M.; Rawa, M. J.; and Albeshri, A. 2022. Discriminative manifold distribution alignment for domain adaptation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2): 1183–1197.

Ye, J.; Wei, Y.; Wen, X.-C.; Ma, C.; Huang, Z.; Liu, K.; and Shan, H. 2023. Emo-DNA: Emotion Decoupling and Alignment Learning for Cross-Corpus Speech Emotion Recognition. In *MM*, 5956–5965.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.

Zhao, Z.; Liu, B.; Lu, Y.; Chu, Q.; and Yu, N. 2024. Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-to-Image Person Re-identification. In *AAAI*, 7, 7534–7542.

Zheng, Y.; Li, Z.; Li, X.; Liu, J.; Wang, Y.; Meng, X.; and Meng, L. 2024. Unifying Visual and Semantic Feature Spaces with Diffusion Models for Enhanced Cross-Modal Alignment. In *ICANN*, 110–125.