

Semantic-guided Representation Learning for Multi-Label Recognition

Ruhui Zhang^{1,2}, Hezhe Qiao^{3*}, Pengcheng Xu¹, Mingsheng Shang², Lin Chen^{2*}

¹Chongqing University of Posts and Telecommunication, China

²Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, China

³Singapore Management University, Singapore

{zhangruhui, shangms, chenlin} @cigit.ac.cn, hezheqiao@gmail.com, pxu204303@gmail.com

Abstract—Multi-label Recognition (MLR) involves assigning multiple labels to each data instance in an image, offering advantages over single-label classification in complex scenarios. However, it faces the challenge of annotating all relevant categories, often leading to uncertain annotations, such as unseen or incomplete labels. Recent Vision and Language Pre-training (VLP) based methods have made significant progress in tackling zero-shot MLR tasks by leveraging rich vision-language correlations. However, the correlation between multi-label semantics has not been fully explored, and the learned visual features often lack essential semantic information. To overcome these limitations, we introduce a Semantic-guided Representation Learning approach (SigRL) that enables the model to learn effective visual and textual representations, thereby improving the downstream alignment of visual images and categories. Specifically, we first introduce a graph-based multi-label correlation module (GMC) to facilitate information exchange between labels, enriching the semantic representation across the multi-label texts. Next, we propose a Semantic Visual Feature Reconstruction module (SVFR) to enhance the semantic information in the visual representation by integrating the learned textual representation during reconstruction. Finally, we optimize the image-text matching capability of the VLP model using both local and global features to achieve zero-shot MLR. Comprehensive experiments are conducted on several MLR benchmarks, encompassing both zero-shot MLR (with unseen labels) and single positive multi-label learning (with limited labels), demonstrating the superior performance of our approach compared to state-of-the-art methods. The code is available at <https://github.com/MVL-Lab/SigRL>.

Index Terms—Multi-label Recognition, Zero-shot recognition, Vision-Language Pre-training Models, Feature Reconstruction

I. INTRODUCTION

Multi-label recognition (MLR), which assigns multiple labels to each data instance in the image, has found widespread application in areas such as image tagging, document categorization, and medical diagnosis. Despite its critical role and increasing prominence in real-world applications, MLR faces significant challenges in practical scenarios. These challenges include: (1) the exhaustive annotation of image datasets with a large number of categories, which is prohibitively labor-intensive; and (2) the possibility of novel categories (unseen labels) emerging during testing, even with meticulous annotation efforts.

*Lin Chen and Hezhe Qiao are the corresponding authors. This work is supported by the Chongqing Key Project of Technological Innovation and Application (CSTB2023TIAD-STX0015 and CSTB2023TIAD-STX0031).

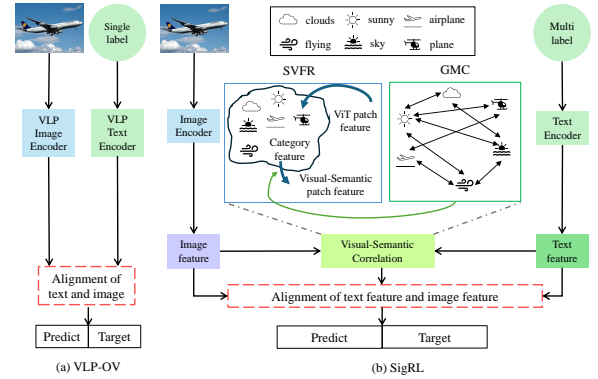


Fig. 1. (a) Previous VLP-based methods like VLP-OV [3] solely focus on modeling and aligning the visual and textual inputs [4]. (b) SigRL leverages GMC to model relationships among multi-label text and uses SVFR to reconstruct semantic visual features, thus improving multimodal representation learning.

The high cost of labeling can be mitigated by adopting more practical solutions, such as partial-label learning, where training instances are annotated with only a subset of their true labels. To address one of the typical settings, Single Positive Multi-label Learning (SPML), various methods have been proposed, including pseudo-labeling techniques [1] and novel loss functions [2]. However, these approaches face significant limitations in generalizing to unseen labels, particularly when tackling the challenge of recognizing novel categories.

Recently, Vision and Language Pre-training (VLP) models have demonstrated exceptional performance in multimodal tasks, offering a promising approach for transferring linguistic knowledge from base categories to support open vocabulary (OV) tasks through prompt learning. By pre-training on large-scale image-text pairs [4], VLP models effectively align and fuse cross-modal information by mapping representations of both image and language modalities into a shared feature space. However, most VLP-OV models [3] primarily focus on single-label classification and have received limited attention in the context of MLR with unseen labels, also known as zero-shot MLR, as illustrated in Fig. 1 (a). To address this issue, He et al. [5] proposed the multi-modal knowledge (MKT) approach for zero-shot MLR, using knowledge distillation to

align patch features with multi-labels. Xing et al. [6] leverage CLIP [4] to generate robust positive and negative pseudo-labels for zero-shot MLR.

However, these approaches fail to thoroughly explore the correlation between multi-labels, and the learned visual features lack essential semantic information. While some studies [7] utilize Graph Neural Networks (GNNs) with predefined, static graph structures to model inter-label dependencies, these methods are not specifically tailored for zero-shot MLR, resulting in suboptimal performance.

Therefore, we propose a semantic-guided representation learning approach (SigRL). SigRL explores the relationships between multiple labels and enriches semantic information by integrating text features into visual representations during reconstruction, as shown in Fig. 1(b). This general framework explicitly models the interaction between semantic image features and multiple labels, enhancing text-image alignment within VLP models.

Our contributions can be summarized as follows:

- We propose a Semantic-guided Representation Learning approach (SigRL) to effectively learn the representations of visual images and text, which can benefit the downstream alignment between visual images and categories.
- We further introduce a graph-based Multi-label correlation (GMC) and Semantic Visual Feature Reconstruction (SVFR) to implement SigRL. GMC constructs a graph structure to capture the intricate semantic relationships among labels, while SVFR integrates semantic features into the visual representation during reconstruction to learn effective visual representations.
- Extensive experiments on three widely used large-scale datasets demonstrate that our proposed SigRL achieves state-of-the-art performance on both zero-shot MLR and SPML tasks.

II. METHOD

A. Problem Definition

In this study, we primarily focus on two types of MLR with uncertain annotations: zero-shot MLR and SPML tasks. The training dataset consists of N samples, $\mathcal{X} = \{\mathbf{X}_k, \mathbf{Y}_k\}_{k=1}^N$ where \mathbf{X}_k denotes the k -th image and \mathbf{Y}_k is a set of labels containing C classes. Each label $y_k^i \in \mathbf{Y}_k, i \in [1, C]$ can be 1 (positive), -1 (negative), or 0 (unknown).

In zero-shot MLR, the dataset is divided into two classes: base classes \mathbf{Y}^B and novel classes \mathbf{Y}^U . $\mathbf{Y}_k \in \mathbf{Y}^B$ represents the known label for the training image, while \mathbf{Y}^U denotes the set of unseen labels that are not present in the training data. For VLP model, such as CLIP, which has a strong vocabulary \mathbf{L} , this vocabulary does not strictly cover \mathbf{Y}^B and \mathbf{Y}^U , and may also include predefined words outside of the new categories. Model training is not solely focused on \mathbf{Y}^B but rather on \mathbf{L} . Zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) are two primary tasks within zero-shot MLR. ZSL focuses on learning a classifier $f_{zsl} : \mathcal{X} \rightarrow \mathbf{Y}^U$. GZSL extends this by requiring

the model to classify both seen and unseen categories, *i.e.*, $f_{gzsl} : \mathcal{X} \rightarrow \mathbf{Y}^B \cup \mathbf{Y}^U$.

In SPML task, \mathbf{Y}_k associated with an image contains only one positive label. Specifically, for each y_k^j in \mathbf{Y}_k , $y_k^j \in \{0, 1\}^C$ and $\sum_{j=1}^C \mathbb{I}_{[y_k^j=1]} = 1$, where $\mathbb{I}_{[\cdot]}$ denotes the indicator function, signifying that only one label in \mathbf{Y}_k is positive. Although the training data is highly sparse, the test task requires predicting multiple labels. In this work, we randomly select and retain a single positive label for each training image, treating all other labels as unknown. This selection is performed once per dataset. The validation and test sets remain fully labeled.

B. The Overall Framework

The overall structure of the SigRL model is illustrated in Fig. 2. SigRL comprises two main modules: a Graph-based Multi-label Correlation module (GMC) and a Visual Feature Reconstruction module (VFR). Specifically, a VLP image encoder is employed to extract features from both images and constructed text. The patch features are obtained by distilling knowledge from the VLP encoder. In the GMC, SigRL captures label relationships by constructing a topology where labels can exchange information. Meanwhile, the VFR integrates the corresponding semantic information into the visual representation by reconstructing the patch embeddings. Finally, SigRL aligns the visual image and labels based on a similarity score.

C. Graph-based Multi-label Correlation Modeling

The pre-trained VLP model CoOp [8] serves as the VLP text encoder and the corresponding label embedding \mathbf{h}^i of the image is obtained by

$$\mathbf{h}^i = \text{CoOp}(\mathbf{Y}_i). \quad (1)$$

To exploit the potential relations among labels, we build a fully connected graph for each image where each node represents the label and the attribute of each node is the learned embedding \mathbf{h}^i . Then we applied the Graph Attention Network (GAT) [9] to dynamically adjust the influence of neighboring labels through an attention mechanism in information exchange. GAT assign a score $e(v_i, v_j)$ for the edge which connects node i and node j , indicating the importance of the characteristics of the node v_j to the node v_i .

$$e(v_i, v_j) = \text{LeakyReLU}(\alpha^T [\mathbf{W}\mathbf{h}^i, \mathbf{W}\mathbf{h}^j]), \quad (2)$$

where $v_j \in \mathcal{N}(v_i)$ indicates the neighboring node of node v_i in the graph. \mathbf{W} denotes the trainable matrix mapping the attribute to the latent space, and α is the learnable attention vector. Then the normalization is performed on the attention score $e(v_i, v_j)$ to obtain the attention coefficient by:

$$\beta_{ij} = \frac{\exp(e(v_i, v_j))}{\sum_{v_k \in \mathcal{N}(v_i)} \exp(e(v_i, v_k))}. \quad (3)$$

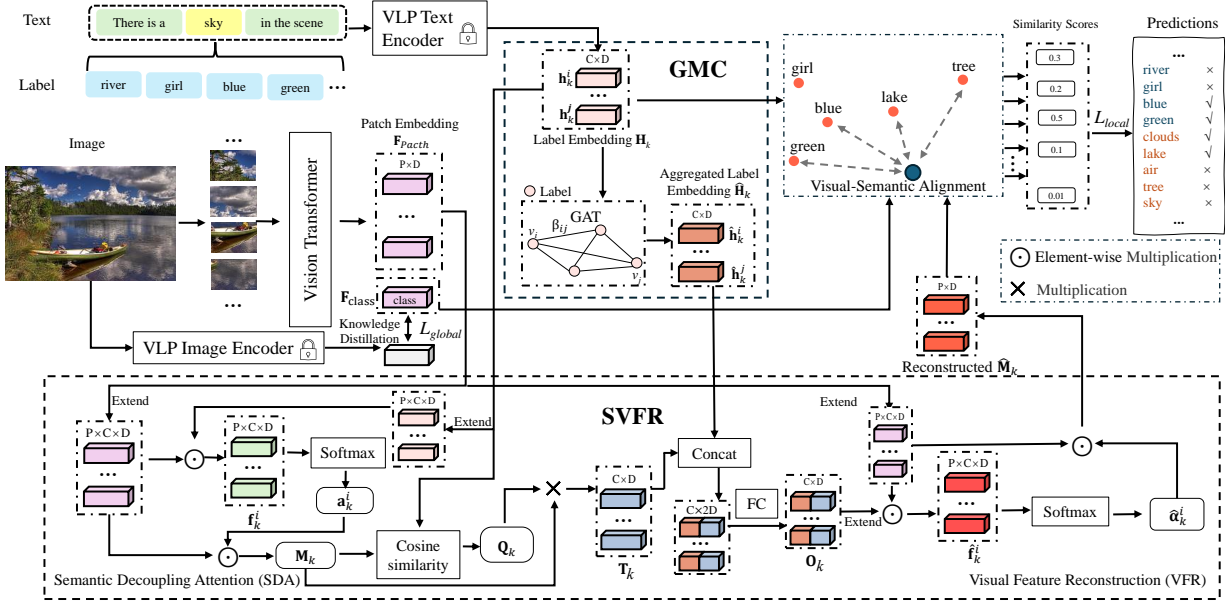


Fig. 2. The overview of SigRL. The VLP Image Encoder and VLP Text Encoder are employed to extract features from the image and label, respectively. The proposed GMC is then utilized to construct a fully connected graph. The SVFR consists of two components: SDA and VFR. SDA decouples visual features and integrates semantic information, while VFR reconstructs visual features by aligning spatial information encoded within ViT patch features.

Finally, the normalized attention coefficient β_{ij} is used to compute the new representation of each node.

$$\hat{\mathbf{h}}^i = \sigma \left(\sum_{v_j \in \mathcal{N}(v_i)} \beta_{ij} \mathbf{W} \mathbf{h}^j \right), \quad (4)$$

where σ denotes the nonlinear activate function. The aggregated label embedding $\hat{\mathbf{H}}_k = [\hat{\mathbf{h}}_k^1, \dots, \hat{\mathbf{h}}_k^C]$ with rich semantic information will be integrated into the visual representation, where C is the number of classes.

D. Semantic Visual Feature Reconstruction

To bridge the semantic gap given that the learned visual features often lack essential semantic information, we propose the SVFR, which leverages a multi-label semantic separation strategy to learn category-specific visual representations, thereby benefiting the downstream alignment of visual images and categories.

We denote the patches as \mathbf{V}_k , $\mathbf{V}_k = [\mathbf{F}_{class}, \mathbf{F}_{patch}]$, where \mathbf{F}_{class} and \mathbf{F}_{patch} represent the outputs of the class and patch tokens, respectively. $\mathbf{F}_{patch} \in \mathbb{R}^{P \times D}$, where P is the number of patches, D represents the output dimension of the Vision Transformer (ViT) [10]. Concurrently, from the VLP text encoder, we can obtain label embeddings $\mathbf{H}_k \in \mathbb{R}^{C \times D}$ for the k -th image by text prompt, where each row $\mathbf{h}_k^i \in \mathbb{R}^{1 \times D}$ represents the embedding for category i , along with the same dimension D as the patch features.

The SVFR framework comprises two key components: semantic decoupling attention and visual feature reconstruction. The semantic decoupling attention module disentangles the features extracted by the ViT by projecting them onto distinct

category-specific subspaces, enabling it to effectively capture the interaction between visual and semantic information.

Semantic Decoupling Attention (SDA). In SDA, we first extend the dimensions of both \mathbf{F}_{patch} and \mathbf{H}_k to $\mathbb{R}^{P \times C \times D}$. Then the semantic-guided attention mechanism is employed to adaptively weight the contribution of different spatial and semantic features using the attention score \mathbf{f}_k^i which is formulated as

$$\mathbf{f}_k^i = FC(\tanh(\mathbf{F}_{patch} \odot \mathbf{h}_k^i)), \quad (5)$$

where $FC(\cdot)$ is a fully connected layer, and \odot denotes the element-wise multiplication operation. The category-specific features \mathbf{M}_k is defined as the weighted sum of image features \mathbf{f}_k^i ,

$$\mathbf{M}_k = \sum_{i=1}^C \mathbf{a}_k^i \mathbf{F}_{patch}, \quad (6)$$

where the attention coefficients \mathbf{a}_k^i is derived from \mathbf{h}_k^i and \mathbf{F}_{patch} , which is formulated as the following

$$\mathbf{a}_k^i = softmax(FC(\tanh(\mathbf{f}_k^i))). \quad (7)$$

In addition, the category-specific attention map $\mathbf{q}_k^{ij} \in \mathbf{Q}_k^i$ representing the correlation between the label embedding \mathbf{h}_k^j and the acquired category-specific features $\mathbf{m}_k^i \in \mathbf{M}_k$ is calculated with cosine similarity by:

$$\mathbf{q}_k^{i,j} = RELU \left(\frac{(\mathbf{m}_k^i)^T \mathbf{h}_k^j}{\|\mathbf{m}_k^i\| \|\mathbf{h}_k^j\|} \right). \quad (8)$$

Then the normalization is performed on the visual feature vectors of each category to obtain \mathbf{T}_k^i representing semantic decoupled features is defined as the following

$$\mathbf{T}_k^i = \text{Norm}(\mathbf{Q}_k^i) \times \mathbf{M}_k^i. \quad (9)$$

By decoupling the patch feature based on the label embedding, the SigRL model integrates the original label information through the above operations to enrich the semantic information in the visual feature.

Visual Feature Reconstruction (VFR). Directly projecting them into $\mathbb{R}^{C \times D}$ risks the loss of this crucial spatial context by mapping the dimensions of the patch's representation into the dimension of class representation. To mitigate the loss of information, we further propose the VFR to reconstruct the representation of visual images by aligning the spatial information encoded in ViT patch features.

We first concatenate $\hat{\mathbf{H}}_k$ and \mathbf{T}_k and then transformed it through a fully connected layer to yield $\mathbf{O}_k \in \mathbb{R}^{C \times D}$. Then, $\mathbf{O}_k \in \mathbb{R}^{C \times D}$ is concatenated with the image \mathbf{F}_{patch} to generate semantic visual features as:

$$\hat{\mathbf{f}}_k^i = FC(\tanh(\mathbf{F}_{patch} \odot \mathbf{O}_k^i)) \quad (10)$$

Then, the vectors \mathbf{O}_k and \mathbf{F}_{patch} are processed with an attentional function (*i.e.*, a fully connected operation). Subsequently, a softmax function is applied to normalize the resulting values and obtain the attention coefficients $\hat{\mathbf{a}}_k^i$ as follows

$$\hat{\mathbf{a}}_k^i = \text{softmax}(\hat{\mathbf{f}}_k^i). \quad (11)$$

Finally, the reconstructed representation with rich semantic information $\hat{\mathbf{M}}_k \in \mathbb{R}^{P \times D}$ associated with the patches are obtained using the weighted average pooling of patches, which is formulated as the following:

$$\hat{\mathbf{M}}_k = \sum_P \hat{\mathbf{a}}_k^i \mathbf{F}_{patch}. \quad (12)$$

E. Alignment of Visual Image and Category

By fusing the semantic and visual features, the predicted scores s_i for each category are computed based on the $\mathbf{F}_{patch} = \{\hat{\mathbf{M}}_k^1, \hat{\mathbf{M}}_k^2, \dots, \hat{\mathbf{M}}_k^P\}$ and \mathbf{F}_{class} , which is formulated as

$$s_i = \langle \mathbf{h}^i, \mathbf{F}_{class} \rangle + \text{Top-K} \left(\left\{ \langle \mathbf{h}^i, \hat{\mathbf{M}}_k^1 \rangle, \dots, \langle \mathbf{h}^i, \hat{\mathbf{M}}_k^P \rangle \right\} \right), \quad (13)$$

where $\mathbf{h}^i \in \mathbb{R}^{1 \times D}$ denotes the original label embedding for each category, $\langle \cdot \rangle$ denotes inner product, and Top-K (\cdot) refers to *top-k* mean pooling.

To align the visual image and category, we employ a total loss function L_{total} consisting of L_{global} and L_{local}

$$L_{total} = L_{local} + L_{global}, \quad (14)$$

where the L_{local} is defined based on a ranking loss to enhance the representational capacity of local features extracted by the ViT,

$$L_{local} = \sum_k \sum_{p \in \mathbf{Y}_k^i, n \notin \mathbf{Y}_k^i} \max(1 + s_k^n - s_k^p, 0), \quad (15)$$

where $\mathbf{Y}_k^i \in \mathbf{Y}^B$ denotes the ground truth labels of base classes for k -th image, and s_k^n and s_k^p represent the scores of negative label n and positive label p , respectively. L_{global} is a knowledge distillation loss function defined based on the \mathbf{F}_{class} as the following

$$L_{global} = \|\text{CLIP}(\mathbf{X}_k) - \mathbf{F}_{class}\|_1, \quad (16)$$

where \mathbf{X}_k denotes the feature of k -th image, $\text{CLIP}(\cdot)$ denotes the output of the pre-trained CLIP image encoder, and $\|\cdot\|_1$ represents the L_1 loss, known as the mean absolute error.

III. EXPERIMENTS

A. Experiments Setup

Datasets. We evaluate the proposed SigRL on three widely used datasets. The NUS-WIDE dataset [11], containing 925 labels, is employed for the zero-shot MLR task. In this setting, 925 labels are considered seen, while the remaining 81 labels are unseen. The Microsoft COCO (COCO) [12] includes 80 object categories, with 82,801 training and 40,504 validation images. VOC (VOC) [13] covers 20 categories, offering 5,011 training/validation and 4,952 test images. Following previous studies [5], [14], we evaluated models using mean average precision (mAP), F1, Precision (P), and Recall (R) for the Top-K predictions.

Implementation Details. The ViT-B/16 model pre-trained on ImageNet-1K is used as the ViT backbone of the SigRL model. For each image with the resolution of 224×224 , ViT-B/16 splits it into 16×16 patches. Regarding hyperparameter settings, k in *top-k* pooling is set to 16 defined in the SigRL, and the ViT feature dimension D is set to 512. The optimizer employs AdamW with a base learning rate of $1e-4$, a minimum learning rate of $1e-7$, and a weight decay of 0.05.

B. Main Results

In the experiments, we performed two separate tasks: zero-shot MLR and SPML. In the zero-shot MLR comparison, we compared the SigRL model with several state-of-the-art methods published in recent years. This comparison included a diverse set of techniques, including Vision-Language Pre-training (VLP)-based models, *e.g.*, CLIP model with ranking loss (CLIP-FT), MKT, Gen-MLZSL models [21], (ML)²P_Encoder [22] and SCP [23].

The experimental results for zero-shot MLR are summarized in Table I. Notably, the VLP model CLIP-FT achieves significantly higher performance compared to traditional ZSL approaches. While MKT achieved the second-highest F1 and mAP scores on GZSL by focusing on multi-modal alignment between labels and image features. Our SigRL further leverages multi-label information for image feature reconstruction, leading to enhanced performance in zero-shot MLR. Quantitatively, SigRL outperformed MKT by 2.13 and 2.21 in Top-3 and Top-5 F1 scores for GZSL, respectively, and also achieved state-of-the-art results on the ZSL task. Meanwhile, the number of parameters in SigRL is 59.82M, which is only 4.47% more than the base model MKT.

TABLE I

STATE-OF-THE-ART COMPARISON FOR ZSL AND GZSL TASKS ON THE NUS-WIDE DATASET. THE TOP-PERFORMING SCORE IS HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST SCORE IS UNDERLINED.

Methods	Zero-Shot Learning (ZSL)							Generalized zero-Shot Learning (GZSL)						
	P	Top-3 R	F1	P	Top-5 R	F1	mAP	P	Top-3 R	F1	P	Top-5 R	F1	mAP
LabelEM (2015) [15]	15.60	25.00	19.20	13.40	35.70	19.50	7.10	15.50	6.80	9.50	13.40	9.80	11.30	2.20
Fast0Tag (2016) [16]	22.60	36.20	27.80	18.20	48.40	26.40	15.10	18.80	8.30	11.50	15.90	11.70	13.50	3.70
OAL (2018) [17]	20.90	33.50	25.80	16.20	43.20	23.60	10.40	17.90	7.90	10.90	15.60	11.50	13.20	3.70
LESAM ₁₀ (2020) [18]	25.70	41.10	31.60	19.70	52.50	28.70	19.40	23.60	10.40	14.40	19.80	14.60	16.80	5.60
BiAM (2021) [19]	-	-	33.10	-	-	30.70	26.30	-	-	16.10	-	-	19.00	9.30
SDL _{M7} (2021) [20]	24.20	41.30	30.50	18.80	53.40	27.80	25.90	27.70	<u>13.90</u>	18.50	23.00	<u>19.30</u>	22.10	12.00
CLIP-FT (2021) [4]	-	-	23.50	-	-	21.70	30.50	-	-	20.30	-	-	23.20	16.80
MKT(2023) [5]	-	-	34.10	-	-	31.10	<u>37.60</u>	-	-	<u>22.00</u>	-	-	<u>25.40</u>	<u>18.30</u>
Gen-MLZSL(2023) [21]	<u>26.60</u>	<u>42.80</u>	32.80	<u>20.10</u>	<u>53.60</u>	29.30	25.70	<u>30.90</u>	13.60	18.90	<u>26.00</u>	19.10	22.00	8.90
(ML) ² P_Encoder(2023) [22]	-	-	32.80	-	-	32.30	29.40	-	-	16.80	-	-	19.20	10.20
SCP(2024) [23]	-	-	33.80	-	-	31.00	27.20	-	-	16.10	-	-	18.80	10.10
SigRL(Ours)	27.73	44.32	34.11	21.77	58.01	<u>31.66</u>	38.46	39.43	17.38	24.13	32.60	23.95	27.61	20.85

TABLE II
COMPARISON FOR SPML TASKS ON THE COCO DATASET.

Ann. Labels	Methods	Loss	mAP	P	Top-3 R	F1	P	Top-5 R	F1
1 P.&0N.	DualCoOp [14]	-	69.20	-	-	-	-	-	-
	MKT [5]	IUN [2]	73.65	63.92	76.99	69.85	45.56	85.67	59.49
		EM_APL [24]	74.66	<u>65.85</u>	<u>78.92</u>	<u>71.80</u>	<u>46.62</u>	<u>87.32</u>	<u>60.79</u>
		EM [24]	73.72	65.06	78.05	70.97	46.16	86.65	60.23
	SigRL(Ours)	IUN [2]	74.19	65.43	77.64	70.48	45.82	86.07	59.80
		EM_APL [24]	<u>74.70</u>	65.63	78.83	71.63	46.58	<u>87.41</u>	60.77
		EM [24]	74.86	65.87	78.99	71.84	46.68	87.45	60.87

TABLE III
COMPARISON FOR SPML TASKS ON THE VOC DATASET.

Ann. Labels	Methods	Loss	mAP	P	Top-3 R	F1	P	Top-5 R	F1
1 P.&0N.	DualCoOp [14]	-	83.60	-	-	-	-	-	-
	MKT [5]	IUN [2]	86.37	43.25	93.86	59.47	27.30	96.57	42.57
		EM_APL [24]	84.58	44.23	94.53	60.26	27.76	97.45	43.21
		EM [24]	86.25	43.31	92.94	59.09	27.32	96.26	42.56
	SigRL(Ours)	IUN [2]	87.20	43.74	94.03	59.71	27.46	96.87	42.79
		EM_APL [24]	84.60	43.11	93.04	58.92	27.35	96.54	42.62
		EM [24]	86.45	44.50	95.28	60.67	27.90	98.01	43.44

Figure 3 illustrates zero-shot MLR predictions on NUS-WIDE for the tested models. Our proposed model accurately predicted both "building" and "glass", as well as correctly inferred the relationship between "bear" and "river", thereby correctly predicting "lake". In contrast, MKT failed to predict "glass" and misclassified "river" as "lake", highlighting the importance of multi-label semantic visual features for enhanced VLP performance.

To evaluate the performance of SigRL on SPML tasks, we compared it with other models using different SPML loss functions, including Ignores Unobserved Negatives (IUN) loss [2], entropy-maximization (EM) loss [24], and entropy-maximization asymmetric pseudo-labeling (EM_APL) [24]. The results on the COCO and VOC datasets are presented in Tables II and III, respectively. We employ three loss functions: ignores unobserved negatives (IUN) [2] loss, entropy-maximization (EM) [24] loss for SPML task, and entropy-maximization asymmetric pseudo-labeling (EM_APL) [24] loss, which were compared across different models. Notably, the experimental results of DualCoOp [14], were reported by Xing et al. [6]. Consistent with the original DualCoOp [14], we employ ResNet-101 as the visual encoder, resizing

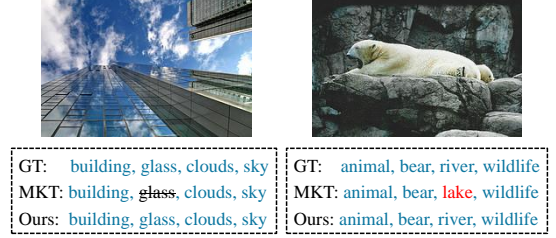


Fig. 3. The qualitative analysis of zero-shot MLR, where ground truth labels are denoted by GT, correct predictions are shown in blue, incorrect predictions are shown in red, and missed predictions are denoted by knockout.

images to 448x448 pixels. Table II presents the experimental results on the COCO dataset. The EM loss function yields the most significant improvement. Compared to MKT, our model achieves a 1.14 increase in mAP and 0.87 and 0.64 improvements in Top-3 and Top-5 F1 scores, respectively. Table III demonstrates the effectiveness of our model on the VOC dataset. Compared to MKT, which also uses the IUN loss function, our model achieves a 0.20 higher mAP and 1.88 and 0.88 higher Top-3 and Top-5 F1 scores, respectively. The observed performance gains can be attributed to SigRL's mechanism of reconstructing features with richer semantic and image information, thereby enhancing multi-modal alignment by leveraging the properties of the VLP model, even with only limited valid annotations.

C. Ablation Study

To evaluate the contribution of each module, we conducted ablation studies on the SigRL model. As shown in Table IV, the experimental results demonstrate that these modules improve the VLP model's performance on zero-shot MLR. Notably, the VFR module within the SigRL model contributed to a 0.37 improvement in mAP for the ZSL task. For GZSL, the GMC module led to a 0.08 improvement in mAP and 0.48 and 0.72 improvements in F1 for Top-3 and Top-5, respectively. Furthermore, the SVFR, which consists of the SDA and VFR modules, improved performance across all test metrics. The enhanced MLR performance in SVFR is

TABLE IV
ABLATION EXPERIMENT FOR ZSL AND GZSL TASKS ON THE NUS-WIDE.

Methods			Zero-Shot Learning (ZSL)							Generalized zero-Shot Learning (GZSL)						
GMC	SDA	VFR	Top-3			Top-5			mAP	Top-3			Top-5			mAP
P	R	F1	P	R	F1	P	R	F1		P	R	F1	P	R	F1	
×	×	×	25.67	41.04	31.59	19.92	53.08	28.97	37.60	37.33	16.46	22.85	30.89	22.70	26.17	19.98
✓	×	×	25.73	41.13	31.66	20.28	54.04	29.49	<u>38.44</u>	38.13	16.81	23.33	31.74	23.32	26.89	20.06
✓	✓	×	26.57	42.48	32.69	20.77	<u>55.35</u>	<u>30.21</u>	38.09	<u>39.05</u>	<u>17.22</u>	<u>23.90</u>	<u>32.39</u>	<u>23.80</u>	<u>27.44</u>	<u>20.62</u>
✓	✓	✓	27.73	44.32	34.11	21.77	58.01	31.66	38.46	39.43	17.38	24.13	32.60	23.95	27.61	20.85

attributed to the fusion of semantic and visual features, and the GMC’s capacity to model multi-label correlations.

IV. CONCLUSION

In this paper, we propose a novel semantic guide representation learning approach SigRL, which provides a unified framework for improving MLR under two scenarios including zero-shot MLR and SPML. In SigRL, we first introduce the GMC to exploit semantic relationships between multiple labels by incorporating a trainable graph learning structure during the representation learning of text. Second, to enrich the semantic information of the representation of visual images, we further introduce the SVFR to integrate the semantic information into the representation of visual images during the reconstruction. By doing this SigRL can learn effective representation of visual images which enhance the alignment of text-image pairs in the VLP model. Finally, the alignment of visual images and categories is achieved by optimizing the knowledge distill loss and ranking loss from global and local features respectively. We conducted experiments on the zero-shot MLR task using the NUS-WIDE dataset and the SPML task using the COCO and VOC datasets. The experimental results demonstrate the significant effectiveness of our method.

REFERENCES

- [1] Ming-Kun Xie, Jiahao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang, “Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning,” *Advances in Neural Information Processing Systems*, pp. 1–17, 2023.
- [2] Elijah Cole, Oisín Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic, “Multi-label learning from single positive labels,” in *Proceedings of the CVPR*, 2021, pp. 933–942.
- [3] Yu Du, Fangyun Wei, Ziheng Zhang, Miaojing Shi, Yue Gao, and Guoqi Li, “Learning to prompt for open-vocabulary object detection with vision-language model,” in *Proceedings of the CVPR*, 2022, pp. 14084–14093.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *ICML*. PMLR, 2021, pp. 8748–8763.
- [5] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xijun Shu, Bo Ren, and Shu-Tao Xia, “Open-vocabulary multi-label classification via multi-modal knowledge transfer,” in *Proceedings of the AAAI*, 2023, vol. 37, pp. 808–816.
- [6] Xin Xing, Zhexiong Xiong, Abby Stylianou, Srikumar Sastry, Liyu Gong, and Nathan Jacobs, “Vision-language pseudo-labels for single-positive multi-label learning,” in *Proceedings of the CVPR*, 2024, pp. 7799–7808.
- [7] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao, “Attention-driven dynamic graph convolutional network for multi-label image recognition,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* 16. Springer, 2020, pp. 649–665.
- [8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [9] Shaked Brody, Uri Alon, and Eran Yahav, “How attentive are graph attention networks?,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. 2022, OpenReview.net.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *ICLR*, 2021.
- [11] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao 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.
- [12] Tsung-Yi Lin, Michael Maire, et al., “Microsoft coco: Common objects in context,” in *ECCV*. Springer, 2014, pp. 740–755.
- [13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, pp. 303–338, 2010.
- [14] Ximeng Sun, Ping Hu, and Kate Saenko, “Dualcoop: Fast adaptation to multi-label recognition with limited annotations,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 30569–30582, 2022.
- [15] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, “Label-embedding for image classification,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.
- [16] Yang Zhang, Boqing Gong, and Mubarak Shah, “Fast zero-shot image tagging,” in *Proceedings of the CVPR*. IEEE, 2016, pp. 5985–5994.
- [17] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, “Bilinear attention networks,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 1571–1581, 2018.
- [18] Dat Huynh and Ehsan Elhamifar, “A shared multi-attention framework for multi-label zero-shot learning,” in *Proceedings of the CVPR*, 2020, pp. 8776–8786.
- [19] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah, “Discriminative region-based multi-label zero-shot learning,” in *Proceedings of the ICCV*, 2021, pp. 8731–8740.
- [20] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi Zelnik-Manor, “Semantic diversity learning for zero-shot multi-label classification,” in *Proceedings of the ICCV*, 2021, pp. 640–650.
- [21] Akshita Gupta, Sanath Narayan, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Joost Van De Weijer, “Generative multi-label zero-shot learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14611–14624, 2023.
- [22] Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, and Fushuo Huo, “²p-encoder: On exploration of channel-class correlation for multi-label zero-shot learning,” in *Proceedings of the CVPR*, 2023, pp. 23859–23868.
- [23] Zhongchen Ma, Runze Ma, Guangchen Wang, Qirong Mao, and Ming Dong, “Label correlation preserving visual-semantic joint embedding for multi-label zero-shot learning,” *Multimedia Tools and Applications*, pp. 1–18, 2024.
- [24] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng, “Acknowledging the unknown for multi-label learning with single positive labels,” in *European Conference on Computer Vision*. Springer, 2022, pp. 423–440.