Multilateral Semantic Relations Modeling for Image Text Retrieval

Zheng Wang[13][*], Zhenwei Gao[1], Kangshuai Guo[1], Yang Yang[1][†], Xiaoming Wang[1], Heng Tao Shen[12]

University of Electronic Science and Technology of China[1], Peng Cheng Lab[2]

Institute of Electronic and Information Engineering of UESTC in Guangdong[3]

Abstract

*Image-text retrieval is a fundamental task to bridge vision and language by exploiting various strategies to fine-grained alignment between regions and words. This is still tough mainly because of one-to-many correspondence, where a set of matches from another modality can be accessed by a random query. While existing solutions to this problem including multi-point mapping, probabilistic distribution, and geometric embedding have made promising progress, one-to-many correspondence is still under-explored. In this work, we develop a **M**ultilateral **S**emantic **R**elations **M**odeling (termed **MSRM**) for image-text retrieval to capture the one-to-many correspondence between multiple samples and a given query via hypergraph modeling. Specifically, a given query is first mapped as a probabilistic embedding to learn its true semantic distribution based on Mahalanobis distance. Then each candidate instance in a mini-batch is regarded as a hypergraph node with its mean semantics while a Gaussian query is modeled as a hyperedge to capture the semantic correlations beyond the pair between candidate points and the query. Comprehensive experimental results on two widely used datasets demonstrate that our MSRM method can outperform state-of-the-art methods in the settlement of multiple matches while still maintaining the comparable performance of instance-level matching.*

1. Introduction

Image and text are two important information carriers to help human and intelligent agents to better understand the real world. Many explorations [9, 18, 35] have been conducted in the computer vision as well as natural language processing domains to bridge these two modalities [16]. As a fundamental yet challenging topic in this research, image-
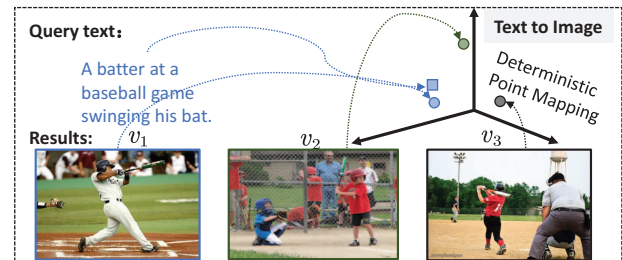

Figure 1. Examples of one-to-many correspondence caused by the inherent nature of different modalities. The existing point-to-point mapping can not capture the semantic richness of data.

text retrieval can benefit other vision-language tasks [11] in two ways, *e.g.* images search for a given sentence and the retrieval of descriptions for an image query, and spread to a variety of applications, such as person search [48], sketch-based image retrieval [33], and food recipe retrieval [52].

Due to the power of deep metric learning [30, 31] in visual embedding augmentation, its core idea is intuitively extended into image-text retrieval to consider the domain differences. The naive strategy [3, 8, 38] is based on triplet loss to learn distinctive representations at the global level only with the help of positive pair and a hard negative one. However, such random sampling cannot effectively select informative pairs, which causes a slow convergence and poor performance [43]. Thus, several re-weighting methods [1, 4, 42, 43] are proposed to address this issue by assigning different weights to positive and negative pairs. Moreover, a flat vector is difficult to infer the complex relationships among many objects existing in a visual scene [16]. Hence, advanced methods formulate various attention mechanisms [2, 15, 20, 40, 50, 51] to distinguish important features from those negligible or irrelevant ones based on Top-K region features obtained from the pre-trained Faster R-CNN [29].

Actually, the prevailing image-text retrieval approaches are instance-based, which only focus on the match of the ground-truth sample. Despite their amazing success, image-text retrieval is still very difficult because of the one-to-many correspondence [6] where a set of candidates can be

---

obtained. This phenomenon is partially caused by the inherent heterogeneity between vision and language. In detail, an image can cover all objects in a given scene yet lacks context reasoning like text [23], while a textual description may only describe a part scene of interest based on the subjective choices [6]. As illuminated by Figure 1, in the case of image retrieval under the textual description of '*A batter at a baseball game swinging his bat*', the ground-truth image v_1 can be retrieved with effort but other instances with sufficient similarity like v_2 and v_3 are possibly discarded. A similar phenomenon also exists in another case of descriptions search for a given image. The essential cause of multiple matches is the point-to-point mapping strategy adopted by the instance-level approaches. That is, they only struggle to capture the one-to-one correspondence based on the ground-truth pairs in the semantic space. Undoubtedly, such a plain strategy suffers from insufficient representation in one-to-many correspondence.

Recently, several works attempt to learn more distinctive representations by cross-modal integration [23, 45] and progressive relevance learning [22, 24]. However, they still adopt point-to-point mapping and can not address the issue of multiple matches. Based on the hedged instance embedding [25] and the box embedding [17, 37], Probabilistic Cross-Modal Embedding (PCME) [6] and Point-to-Rectangle Matching (P2RM) [41] are successively developed to learn richer representations based on semantic uncertainty capture. Motivated by them [6, 41], this work introduces a novel Multilateral Semantic Relations Modeling (**MSRM**) method to capture the one-to-many correspondence between a given query and candidates in image-text retrieval. Concretely, our work mainly includes two parts: semantic distribution learning for a query and multilateral relations modeling for retrieval. The first part maps a given query as a probabilistic embedding to learn its true semantic distribution based on Mahalanobis distance. Then each candidate instance in a mini-batch is regarded as a hypergraph node with its mean semantics while a Gaussian query is modeled as a hyperedge. Afterwards, the second part leverages the hyperedge convolution operation to capture the beyond pairwise semantic correlations between candidate points and the query.

In summary, our contributions can be concluded as:

- We introduce an interpretable method named Multilateral Semantic Relations Modeling to better resolve the one-to-many correspondence for image-text retrieval.

- We propose the Semantic Distribution Learning module to extract the true semantics of a query based on Mahalanobis distance, which can infer more accurate multiple matches.

- We leverage the hyperedge convolution to model the high-order correlations between a Gaussian query and candidates for further improving the accuracy.

2. Related Work

Image-Text Instance Retrieval is the dominant solution to bridge the semantic discrepancy between image and text in a common space by focusing mainly on the ground-truth match. The efforts can also be classified into two categories from the perspective of the feature granularity: coarse-grained matching [3, 8, 12, 14, 39, 46] which directly compares the similarity of image and text based on their global features; fine-grained matching [1, 15, 20, 43, 50] which aggregates similarity scores of all region-word as the overall similarity of image-text pair.

Coarse-grained matching usually deploys a two-branch deep architecture to extract global features of image and text, then various techniques such as metric learning [8], generative models [12], consensus concepts [38], discrete-continuous policy gradient [46] and generalized pooling operator [3] are employed to narrow the distance between semantic similar samples while holding the dissimilar ones father away based on a ranking loss. However, SCAN [15] and VSRN [16] argue that a flat vector can not capture the richness of an image which thus leads to poor performance of fine-grained matching. Therefore, many following works devote efforts to explore better local alignments by filtering important relevance from irrelevant correlation with various elegant strategies, which include neighbor information gathering based on graph [16, 21], vector-based similarity graph reasoning [7, 49], different re-weighting methods for the weight of negative pairs [1, 4, 42, 43], various attention mechanisms [15, 20, 50, 51].

While great effectiveness in the retrieval of the most similar sample, the prevailing instance-based methods suffer from poor performance in the case of one-to-many correspondence. We argue that such limitations are partially attributable to their point-to-point mapping strategy, which only pays attention to the absolute matched or unmatched sample yet ignores many samples with acceptable similarity. Hence, we should take the one-to-many correspondence into account. Otherwise wrong conclusions may be made for the inappropriate predictions.

One-to-Many Correspondence recently attracts great attention for its aggravating the difficulty of image-text retrieval. Polysemous Visual-Semantic Embedding (PVSE) [34] pioneers M-points mapping via multi-head self-attention to address the issue of polysemy in cross-modal matching. But the increase of the predefined parameter M can not work well on multiple matches. Then, Zhou et al. [24] pay attention to the bipolar relevance between queries and candidates where a uniform margin is set to all non-positive samples and thus overlooks various semantic
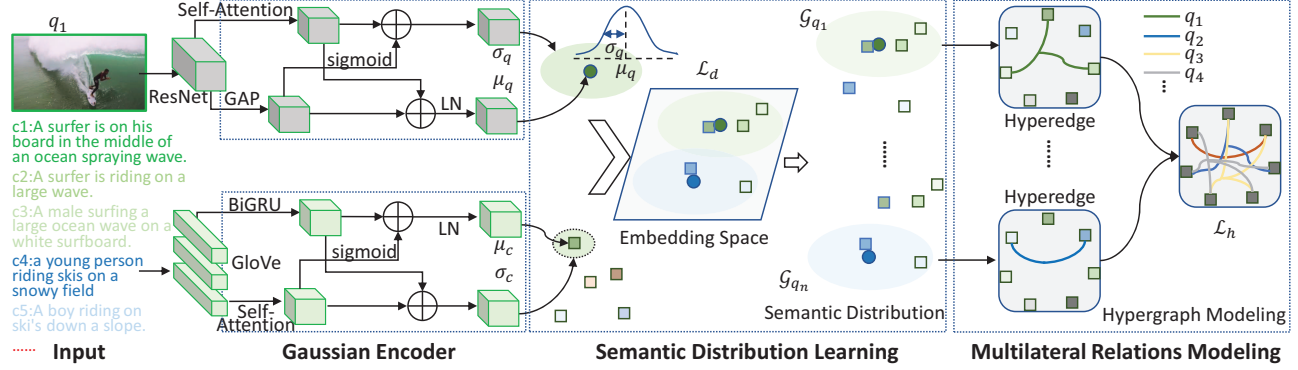
Figure 2. The overview of our proposed method. Gaussian Encoder maps queries into probabilistic embeddings, Semantic Distribution Learning attempts to extract the true semantic distribution of queries, and the last part treats each query as a hyperedge and then leverages hypergraph to model the multilateral semantic relations for one-to-many correspondence. For brevity, the overview is illustrated by image-to-text retrieval which indicates the mean semantic of each text is regarded as a candidate while a query image is a probabilistic embedding.

proximity between image and text. However, their solution of ladder loss is not too practical as it has multiple manually pre-set semantic margins and is still essentially based on one-to-one relations. On the strength of PVSE [34] and inspired by hedged instance embedding [25], Probabilistic Cross-Modal Embedding (PCME) [6] first introduces the semantic uncertainty estimation based on Gaussian distribution into image-text retrieval as a promising solution of one-to-many correspondence albeit its unsatisfactory performance. The latest work P2RM [41] leverages the volume of a rectangle query to contain many candidate points which is actually another uncertainty estimation based on hyper-cube embedding [17]. Nevertheless, the P2RM method suffers from the difficulty of optimization for sparse gradients caused by hard edge [17].

Overall, the explorations of one-to-many correspondence remain insufficient. Based on previous achievements, our query is first mapped as probabilistic embedding to learn the true semantic distribution based on Mahalanobis distance. Then, the query is modeled as a hyperedge to capture the multi-way semantic relevance to many candidate points. Benefiting from the true semantic distribution and the powerful representation ability, we can effectively resolve one-to-many correspondence in image-text retrieval.

3. The Proposed Approach

The overview of our MSRM method is illustrated in Figure 2 and it mainly contains three components: Gaussian encoder, semantic distribution learning, and multilateral relations modeling. Next, we will elaborate on the construction and function of each component. Note that existing solutions of one-to-many correspondence [6, 34, 41] are all based on global features by learning dual encoders E_v, E_t for image and text respectively, we hence follow this paradigm.

3.1. Feature Extraction

Visual features can be obtained with the output before the global average pooling layer of ResNet [13]: $z_v = CNN(v) : v \rightarrow \mathbb{R}^{h \times w \times d_v}$, where $CNN(\cdot)$ can be replaced by any other backbone.

Textual features is constructed as an array of word embedding for a given description t by the pre-trained GloVe [27]: $z_t = GloVe(t) \in \mathbb{R}^{L \times d_t}$, where GloVe can also be other embedding models and L is the length of t.

3.2. Gaussian Encoder

Inspired by hedge instance embedding [25] and PCME [6], we also present a learnable Gaussian distribution for each query that aims to capture multilateral semantic correspondence with a true feature distribution. That is, we also project image and text into probabilistic embedding with two learned Gaussian encoders which share the same architecture with previous remedies [6, 41] to multiple matches and are based on multi-head self-attention [36].

Image Encoder E_v maps a given image v into a latent Gaussian embedding $\mathcal{G}_v$ including mean semantic μ_v extraction and uncertainty estimation for variance σ_v based on z_v, formulated as:

$$\mu_v = \text{LN}(\text{GAP}(z_v) + sg(W_v^l \text{Attn}(z_v))), \quad (1)$$

$$\sigma_v = ReLU(W_v^g(\text{GAP}(z_v) + \text{Attn}(z_v))), \quad (2)$$

where $\text{GAP}(\cdot)$, $\text{LN}(\cdot)$, $\text{Attn}(\cdot)$, and $sg(\cdot)$ indicate global average pooling, layer normalization, self-attention [36], and sigmoid function, respectively. $W_v^l, W_v^g \in \mathbb{R}^{d_v \times d}$ define learnable matrices for affine transformations. Note that the $sg(\cdot)$ and $\text{LN}(\cdot)$ operations are discarded for their restriction on the capture of semantic variance, as mentioned in PCME [6] and P2RM [41].

Text Encoder E_t can be constructed in a similar way, and the only difference is that $\text{GAP}(\cdot)$ for an image is changed

2832

to a Bi-directional Gated Recurrent Unit (BiGRU) [5] for text context learning.

Afterwards, we can represent an image v and a description t in the way of normal distributions with mean semantics and diagonal covariance matrices of variance:

$$
\begin{aligned}
\mathcal{G}_v &\sim \quad N(E_v^\mu(z_v), \mathrm{diag}(E_v^\sigma(z_v))) \sim N(\mu_v, \sigma_c), \\
\mathcal{G}_t &\sim \quad N(E_t^\mu(z_t), \mathrm{diag}(E_t^\sigma(z_t))) \sim N(\mu_t, \sigma_t),
\end{aligned}
\tag{3}
$$

where $\mu_v, \mu_t, \sigma_v, \sigma_t \in \mathbb{R}^d$.

Note that although our MSRM method maps both image and text into probabilistic embedding, we actually employ the intact embedding of a given query while only the mean semantic μ of all candidates to perform one-to-many (distribution-to-points) correspondence .

3.3. Semantic Distribution Learning

This component aims to learn true semantic distributions for each query based on probabilistic embedding, which can better capture the variation of semantics, such as the semantic ambiguity including synonymy, hyponymy, and multi-view existed in descriptions [47] or the inherent uncertainty of visual data [26].

Specifically, the distance metric between a given query q and a candidate c is based on the squared Mahalanobis distance rather than the traditional Euclidean distance, and can be formulated as

$$
d_m^2(q, c) = (u_c - \mu_q)^\top \sigma_q^{-1}(u_c - \mu_q).
\tag{4}
$$

Consequently, we develop a probability P_{qc} to evaluate the association between a query q and a candidate c:

$$
P_{qc} = \frac{exp(-\tau d_m^2(q,c))}{\sum_{c\in\mathcal{B}} exp(-\tau d_m^2(q,c))},
\tag{5}
$$

where $\mathcal{B}$ defines the set of a mini-batch, and $\tau > 0$ is a learnable temperature scaling factor. This idea is reasonable to only consider the variance of query q, because one-to-many correspondence obtains multiple candidate answers, which is caused by the incomplete semantic descriptions of the query. Additionally, it is essentially different from previous works [6, 26, 47] which learn probabilistic embedding for various visual applications but their similarity is calculated by Euclidean distance of the sampling from the learned distributions based on Monte-Carlo estimation.

Then a soft version of contrastive loss is employed as

$$
\mathcal{L}_d^{qc} = \begin{cases} - log(P_{qc}), & \text{if } q, c \text{ is a match} \\ - log(1 - P_{qc}), & \text{otherwise.} \end{cases}
\tag{6}
$$

Consider the direction of retrieval, the loss of distribution learning can be defined as

$$
\mathcal{L}_d = \mathcal{L}_{d,v2t} + \mathcal{L}_{d,t2v},
\tag{7}
$$

where $\mathcal{L}_{d,v2t} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\mathcal{L}_d^{v_i t_j}$ and $\mathcal{L}_{d,t2v}$ can be derived in a similar way.

3.4. Multilateral Relations Modeling.

Each pair in instance-based retrieval only provides bipolar supervision which pulls positive pairs closer while pushing negative pairs distant [8, 15, 21, 50]. However, this strategy causes poor performance on the multiplicity of semantic retrieval [6, 41]. Contrastively, we aim at learning a semantic distribution of a query to model the associations with multiple candidates that share similar semantics.

Formally, we construct a non-bipolar semantic correlation $\mathcal{S}(q)$ for each query based on the mini-batch $\mathcal{B}$. With the learned prototypical distribution, $\mathcal{S}(q)$ can be formalized as a semantic relation matrix $S \in [0,1]^{N_v \times N_t}$, whose entry is defined by

$$
S_{qc} = \begin{cases} 1, & \text{if } q, c \text{ is a match,} \\ e^{\alpha d_m^2(q,c)}, & \text{otherwise,} \end{cases}
\tag{8}
$$

where α is a learnable parameter to control the ratio of non-positive samples.

Since our prototypical distribution is assumed to capture the true distribution of semantics existing in a given query, these different weights are considered as the relevance of candidate answers to this query. Now, the main concern is how to model the multilateral relations rather than the pairwise relationships captured by the existing powerful graph network [16, 21].

Due to the ability of complex connections modeling, we are the first to introduce hypergraph [10] into one-to-many correspondence of image-text retrieval. Unlike binary graphs only modeling pairwise relations, the hypergraph can formulate higher-order relations by enclosing multiple nodes $v \in \mathcal{V}$ within a hyperedge $e \in \mathcal{E}$ to construct incidence matrix H (Eq.9) and then learning distinctive representations via truncated Chebyshev polynomials. Specifically, each entry of the incidence matrix H for a hypergraph G can be defined as:

$$
h(v, e) = \begin{cases} 1, \text{if } v \in e, \\ 0, \text{if } v \notin e. \end{cases}
\tag{9}
$$

Please refer to HGNN [10] for detailed explanations of hypergraph. Despite the establishment of association with multiple nodes, the hyperedge is still constructed based on bipolar similarity, presented as its definition of Eq(9). Therefore, we combine the non-bipolar semantic correlation (Eq(8)) based on Mahalanobis distance between plausible instances and the query into hypergraph to further improve the hyperedge convolution for tackling the multiplicity. That is, Eq(8) can be treated as a weighted version of the incidence matrix H, where each hyperedge connects other nodes with soft incidence weight in [0,1] computed by the squared Mahalanobis distance.

With the constructed hypergraph, the features of candidates in a mini-batch are updated via l layers of hyperedge

convolution:

$$U_c^{l+1} = sg(D_v^{-1/2} H D_e^{-1} H^\top D_v^{-1/2} U_c^l \Theta^l), \qquad (10)$$

where $U_c^l \in \mathbb{R}^{B \times d}$ denotes the input features of l layer and is constructed by the mean semantics of candidates, $sg(\cdot)$ is an activation with sigmoid function, $\Theta \in \mathbb{R}^{d \times d}$ is a learnable parameter, and D_e, D_v indicate the degrees of hyperedge and vertex, respectively. Please refer to HGNN [10] for detailed definitions and derivations.

We take the output of the last layer as our final representation of each candidate. Afterwards, the similarity between a given query and a candidate is computed with their mean semantics based on the cosine function:

$$P'_{qc} = \frac{exp(-\lambda cos(\mu_q, \mu_c))}{\sum_{c \in \mathcal{B}} exp(-\lambda cos(\mu_q, \mu_c))}. \qquad (11)$$

Finally, we can derive the loss of hypergraph modeling as $\mathcal{L}_h = \mathcal{L}_{h,v2t} + \mathcal{L}_{h,t2v}$ according to the soft contrastive loss displayed in Eq.(6) and Eq.(7). The final objective loss is the weighted combination of the two proposed losses.

$$\mathcal{L} = \gamma \mathcal{L}_d + \mathcal{L}_h, \qquad (12)$$

where γ is a hyper-parameter to control the contributions of different items.

4. Experiments

4.1. Dataset and Evaluation Metrics

Dataset. Following the recent solutions [6, 41] of one-to-many correspondence, the incomplete and non-exhaustive MS-COCO [19] including 123,287 images and a smaller yet cleaner CUB [44] with totally $11,788$ images for 200 fine-grained categories of birds are taken as benchmarks to evaluate our effectiveness. Specifically, we clone the split information of MSCOCO widely adopted by instance-level methods [3, 15, 50], where $113,287$ images and $5,000$ instances are split into the training and validation set, respectively. Each image is labeled with 5 ground-truth sentences in MSCOCO while 10 captions per image are generated by [28] for CUB. The evaluation setting for MSCOCO includes 1K test and 5K test, which are consistent with instance-based methods [7, 42, 50]. The split for CUB is the same as PCME [6] and P2RM [41], including 1) all classes training; 2) zero-shot learning which picks out 50 classes as unseen for evaluation.

Evaluation Metrics. First, we adopt the traditional **Recall@K** as our protocols to evaluate the performance of ground-truth match. However, R@K may be not good at semantic ranking especially only with bipolar relevance provided by datasets, which may cause an unsatisfactory performance [6]. Hence, we also evaluate our model under the protocols of Plausible Match Recall Precision (**PMRP**)

and Recall-Precision (**R-P**) for MSCOCO and CUB respectively, which are designed to focus more on the 'precision' for semantics, not just 'recall' for the most similar instance [6, 41]. R-P is calculated by the ratio of positive instances in the top-$|\Omega(query)|$ retrieved results while PMRP is designed to retrieve items with almost similar concepts to the query as MSCOCO is non-exhaustive annotated. Specifically, at most $\zeta \in \{0, 1, 2\}$ positions difference for total 80 labels are taken into account. Please refer to PCME [6] for detailed definitions of the above metrics.

4.2. Implementation Details

As aforementioned, we share the main architecture with PCME [6], therefore most of the implementation details also stay the same to PCME for fair comparisons. Concretely, we pre-train ResNet [13] on ImageNet [32] and GloVe [27] with 2.2M vocabulary to extract our initial features for vision and language. The other unlisted settings including backbone, dimension ($d = 1024$ for MS-COCO and 512 for CUB), batch size $\mathcal{B}$ (64 for CUB and 128 for MSCOCO), AdamP optimizer, and so on are consistent with PCME [6]. The value of γ is set to 0.8, the initial values of learnable temperatures λ, τ are set to 5 and 10, and layers l of HGNN equals to 2, respectively.

4.3. Performance Comparison

As aforementioned, PVSE [34], PCME [6], and P2RM [41] are employed to solve the issue of one-to-many correspondence by multi-point mapping, probabilistic embedding, and geometric representation. Therefore, we mainly compare our performance on multiple matches with them to demonstrate the effectiveness of our proposed MSRM method.

Additionally, several instance-based methods which focus on the ground-truth match, including VSE++ [8] which mined hardest negative, VSRN [16] based on graph reasoning for better alignment, and AOQ [4] which draws the idea of metric learning to make positive pairs closer, are also selected as our competitors.

Results on CUB. We report the performance of multiple matches under different metrics for image-text retrieval in the CUB [44] dataset with two settings. Table 1 is for the setting of 'all classes training' and Table 2 is for 'zero-shot learning' setting. Deep analyze these comprehensive comparisons between our MSRM method and other competitors, we can make some important conclusions:

(1) The increase of M value can indeed improve the R@K performance of PVSE [34] while suffering an opposite trend on R-P performance, which to a certain degree confirms that R@K metric pays more attention to the most similar instance in the retrieval yet ignores the phenomenon of multiple matches caused by semantic variation of image or text. It also indicates that this multi-point mapping

Table 1. Performance comparisons in the CUB dataset [44] for 'all classes training' setting. The 'μ' indicates that only the mean value of probabilistic embedding was used to evaluate.

Methods	Venue	Image-to-text		Text-to-image	
		R-P	R@1	R-P	R@1
VSE0 [8]	BMVC'18	22.40	44.20	22.60	32.70
PVSE M=1 [34]		22.30	40.90	20.50	31.70
PVSE M=2 [34]	CVPR'19	19.70	47.30	21.20	28.00
PVSE M=4 [34]		18.40	47.80	19.90	34.40
PCME μ [6]		24.70	46.40	25.60	35.50
PCME [6]	CVPR'21	26.30	46.90	26.80	35.20
P2RM [41]	MM'22	26.88	49.11	27.93	37.30
MSRM	**Ours**	**27.91**	**51.03**	**28.82**	**39.20**

which is a deterministic point mapping in nature cannot learn the true semantic distribution. By introducing probabilistic fitting and rectangle embedding, PCME [6] and P2RM [41] can effectively alleviate the one-to-many correspondence, which is mainly due to the introduction of uncertainty estimation to capture the rich semantic space of different modalities. However, they also suffer from various limitations and one-to-many correspondence is still under exploration, which calls for more effective solutions.

Table 2. Performance comparisons among various methods under zero-shot setting for CUB [44].

Methods	Venue	Image-to-text		Text-to-image	
		R-P	R@1	R-P	R@1
VSE0 [8]	BMVC'18	22.35	44.19	22.57	32.71
PVSE M=1 [34]		22.34	40.88	20.51	31.71
PVSE M=2 [34]	CVPR'19	19.67	47.29	21.16	27.98
PVSE M=4 [34]		18.38	47.76	19.94	34.39
PCME μ [6]		24.70	46.38	25.64	35.50
PCME [6]	CVPR'21	26.28	46.92	26.77	35.22
P2RM	MM'22	26.75	47.12	27.87	37.16
MSRM	**Ours**	**27.92**	**50.53**	**28.43**	**37.54**

(2) Differently, we learn the true semantic distribution of queries based on Mahalanobis distance and probabilistic embedding, and then capture the complex relevance between image and text via hyperedge convolution. Based on that, we can capture the semantic variation from representation learning and meanwhile better model the one-to-many correspondence from high-order correlation exploration. Consequently, our MSRM method can comprehensively surpass the existing PCME and P2RM approaches

under all evaluation metrics and test settings. Compared with PCME, which also utilizes probabilistic embedding to capture semantic uncertainty, we are ahead with a margin of at least **0.56%** occurring in the image-to-text retrieval scene, and by a maximum improvement up to **1.90%**. Even compared with the latest SOTA, we can still exceed P2RM [41] with a satisfactory margin. For example, the maximum relative lead is **1.17%**. Various performance improvements experimentally demonstrate the effectiveness of our approach in tackling the one-to-many correspondence.

(3) The advantages of our method can again be verified in the horizontal comparison between the two settings of all-class training and zero-shot learning. That is partly because CUB is relatively clean and our power of learned true semantic distribution. The clean indicates that captions and images are basically describing the same class, false positives and false negatives are unlikely to happen [6].

Results on MS-COCO. Unlike CUB [44], MSCOCO is a large-scale yet incomplete especially bipolar relevance dataset, which exacerbates the difficulty of one-to-many correspondence. Similarly, we report the results of PMRP metrics for multiple matches and widely used Recall@K for ground-truth retrieval on MS-COCO in Table 3. The following can be drawn from this table:

1) The performance difference between only mean semantic and the intact probabilistic embedding of PCME is very small with only about **0.1%**. This phenomenon shows that PCME does not capture semantic uncertainty well on MSCOCO. By contrast, our approach solves the one-to-many problem by learning true semantic distributions which can effectively represent the semantic richness of different modalities. Particularly, whether it is under 1K testing or 5K testing and regardless of the retrieval direction, our performance on multiple matches can beat PCME by a minimum of **1.33%**.

2) The performance gap between two kinds of evaluation metrics of VSRN [16] and AOQ [4] further indicates that R@K focusing only on ground-truth match (the most similar instance) may lead to poor performance in semantic retrieval. Even though fine-grained alignment based on graph structure and the power of metric learning is applied to image-text retrieval, seeking a way to address both instance-based and semantic retrieval is a great challenge.

3) Different from P2RM [41] which mainly solves the one-to-many issue by introducing rectangle embedding to contain many candidate points with similar semantics, we resolve both one-to-one and one-to-many matching by learning a true semantic distribution and modeling multilateral connections with hypergraphs. Specifically, although P2RM is ahead of PCME in PMRP, its R@K performance is inferior to the latter by a maximum margin of **2.2%**. However, our MSRM method is superior to the point-to-rectangle matching strategy with varying degrees in the one-

Table 3. The performance comparison of different metrics in MS-COCO between our MSRM method and several latest methods under 1K and 5K test. Note that † indicates the results re-produced by P2RM [41] with the released code of PCME [6].

| Methods | Dimension | 1K Test | | | | 5K Test | | | |
| | | Image-to-text | | Text-to-image | | Image-to-text | | Text-to-image | |
		PMRP	R@1	PMRP	R@1	PMRP	R@1	PMRP	R@1
VSE++ (BMVC'18) [8]	1024	-	64.60	-	52.00	-	41.30	-	30.30
PVSE M=1 (CVPR'19) [34]	1024	40.30	66.70	41.90	53.50	29.30	41.70	30.10	30.60
PVSE M=2 (CVPR'19) [34]	1024 × 2	42.80	69.20	43.70	55.20	31.80	45.20	32.00	32.40
VSRN (ICCV'19) [16]	2048	41.20	76.20	42.40	62.80	29.70	53.00	29.90	40.50
VSRN +AOQ (ECCV'20) [4]	2048 × 2	44.70	**77.50**	45.60	**63.50**	33.00	**55.10**	33.50	**41.10**
PCME μ only (CVPR'21) [6]	1024	45.00	68.00	45.90	54.60	34.00	43.50	34.30	31.70
PCME (CVPR'21) [6]	1024 × 2	45.10	68.80	46.00	54.60	34.10	44.20	34.40	31.90
PCME (CVPR'21)†	1024 × 2	45.10	65.90	46.00	53.30	34.10	41.70	34.40	31.20
P2RM (ACM MM'22) [41]	1024 × 2	45.90	66.60	46.42	54.22	35.52	42.12	35.11	31.50
MSRM (Ours)	1024 × 2	**46.43**	68.85	**47.35**	56.12	**35.62**	44.32	**35.81**	33.40

to-many evaluation and continues to outperform PCME in the ground truth matching, which is sufficient to demonstrate the effectiveness of our method even on a more difficult dataset MSCOCO. Note that our performance on R@K is still inferior to VSRN [16] and AOQ [4], which is clearly caused by our selection of global-level alignment.

Table 4. Ablation study on each loss in our method.

| Methods | I2T | | T2I | |
	R-P	R@1	R-P	R@1
MSRM	**27.9**	**51.0**	**28.8**	**39.2**
μ only	26.8	50.2	27.6	38.7
w/o $\mathcal{L}_d$	27.2	49.2	27.8	38.1
w/o $\mathcal{L}_h$	25.8	45.7	26.2	35.2

Table 5. Analysis on non-binary relevance matrix.

Variants of $\mathcal{S}$	I2T	T2I
$e^{\alpha d_m^2(q,c)} = 0$	27.1	28.1
$\alpha = 1$	27.7	28.7
$\alpha = 2$	**27.9**	**28.8**
$\alpha = 3$	27.8	28.8
$\alpha = 5$	27.8	28.7

4.4. Further Analysis

To demonstrate the effectiveness of our proposed method, we further make comprehensive ablation studies in CUB dataset with 'all classes training' setting. Note that this is the default configuration unless stated otherwise.
Ablation Study. We ablate either loss of our MSRM method (without $\mathcal{L}_d$ or without $\mathcal{L}_h$) to examine the effectiveness of different components. The varying degrees of performance degradation shown in Table 4 prove that our two components can both help the model to solve the multiplicity of semantic matching. We further bypass the variance of query embedding, the expected performance fall indicates that we have indeed learned the true semantic distri-

bution of queries to facilitate multiple matches.
Analysis of Non-bipolar Correlation. We first set the value of α as 0 to validate the advantage of non-bipolar semantic correlation in tackling multiple matches. The first line of Table 5 validates the argument that the performance of bi-polar setting is indeed inferior to that of our proposed method. Additionally, we also investigated the impact of different α on the multiplicity. Table 5 shows that the varying of α within a certain range have little effect on multiple matches, and we set to 2 for best performance.
Parameter Sensitivity. We further explore the effect of weight γ, batch size $\mathcal{B}$, and hypergraph layers l on our performance. Specifically, we set a numerical range [0.5,2.0] for γ to dynamically reflect our performance under different evaluation metrics. The figure 4 shows that the performance of different evaluators increases monotonically and then decreases, and the best performance is achieved at 0.8, which is the final value of our γ.

Additionally, the impact of l on our method is presented in Figure 5. Too many layers of hypergraph may cause overfitting while a few layers are insufficient to explore the one-to-many relations between vision and language. Therefore our final layers are configured to 2, same as HGNN [10].

As our MSRM method attempts to capture high-order correlations between a query and candidates in a minibatch, we further investigated the impact of the mini-batch size. Figure 6 displays the results of our MSRM method with $\mathcal{B} \in \{32, 64, 96, 128\}$. As we can see, our performance of one-to-many correspondence is not very sensitive to the mini-batch size $\mathcal{B}$ and is optimal with a small value. It also reveals that our MSRM method has learned the true semantic distribution of a query, which facilitates the mul-
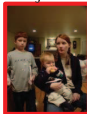
Query	Our MSRM	PCME
	1. A couple of men are loading a truck with glass. $\zeta=0$ 2. Many men work together to put objects in a truck $\zeta=0$ 3. A man bending into the back of a truck on a street $\zeta=1$ 4. A couple are approaching a man sitting down outside of a small shop. $\zeta=3$ 5. A man reaches in the back of a truck . $\zeta=0$ 6. A truck with a bunch of people in back of it. $\zeta=1$	1. A couple of men are loading a truck with glass. $\zeta=0$ 2. A man bending into the back of a truck on a street. $\zeta=1$ 3. A man reaches in the back of a truck. $\zeta=0$ 4. A couple are approaching a man sitting down outside of a small shop $\zeta=3$ 5. A man leaning over the back of a truck in front of buildings. $\zeta=3$ 6. Some people trying to load an item onto a motorcycle. $\zeta=3$
Two children play while eating in a restaurant.		

Figure 3. Visualization of one-to-many correspondence examples, and ζ indicates the semantic concept differences from GT (green) under the evaluation of PMRP. The red samples are not positive labeled in dataset but share similar semantics with that of query.
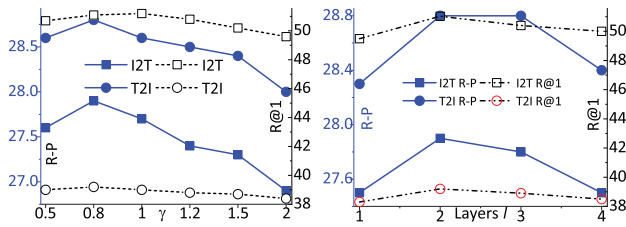

Figure 4. Performance effect of different γ.

Figure 5. Impact of different layers l for hypergraph.

tiple matches between images and texts. In the end, we set it as $\mathcal{B} = 64$ for optimal performance and fair comparisons to PCME [6] in CUB dataset.
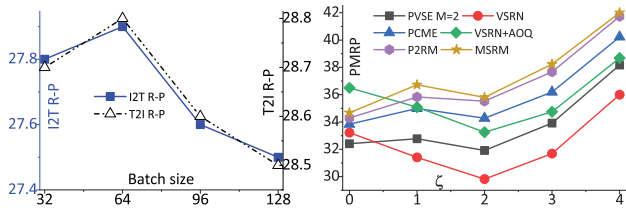

Figure 6. Exploration on mini-batch size $\mathcal{B}$.

Figure 7. PMRP performance vs. ζ varying.

Performance vs ζ Varying. We investigate the effect of ζ varying on the PMRP performance in MS-COCO with six methods, illustrated by Figure. 7. Our MSRM method can achieve better one-to-many correspondence than PCME [6] and P2RM [41] at all different ζ values. The figure 7 can also demonstrate that the traditional instance-based meth-

ods such as VSRN [16] and AOQ [4] are better at matching the most similar instance yet can not capture the semantic richness to match those plausible samples ($\zeta > 0$).
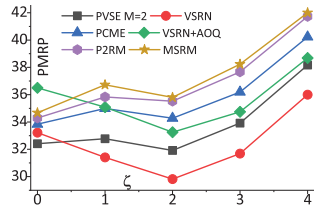
Visualization. To further demonstrate our validity, we visualize several examples of top-6 retrieval results in MSCOCO between our MSRM method and the PCME, shown in Figure 3. The ground truths marked in dataset are in green while the other retrieval samples are in red. From Figure 3, we find that compared to PCME, our MSRM model can perform better both in ground-truth retrieval and one-to-many match. We ascribe all the advances to the adoption of semantic distribution learning and multilateral relations modeling to the settlement of multiple matches.

5. Conclusion

In this work, we developed a novel method to better resolve the issue of one-to-many correspondence in image-text retrieval. Specifically, we first learned a true semantic distribution based on Mahalanobis distance for each query which can better estimate the semantic uncertainty. Then we regarded the distributions of queries as different hyper-edges and leveraged the high-order correlation ability of hypergraph to capture the multilateral relations in candidates on the basis of their own mean semantics. Benefiting from these novel designs, our MSRM method can significantly improve the performance of image-text semantic retrieval. Finally, extensive experiments with achieving new SOTA and solid ablation studies under different evaluations in two widely used benchmarks demonstrated that we can learn true semantic distributions for queries, which facilitated our effectiveness and superiority on multiple matches.

References

[1] Feiyu Chen, Jie Shao, Yonghui Zhang, Xing Xu, and Heng Tao Shen. Interclass-relativity-adaptive metric learning for cross-modal matching and beyond. *IEEE Transactions on Multimedia*, 23:3073–3084, 2021. 1, 2

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, June 2020. 1

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15784–15793, 2021. 1, 2, 5

[4] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, pages 549–565, 2020. 1, 2, 5, 6, 7, 8

[5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014. 4

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[7] H. Diao, Y. Zhang, L. Ma, and H. Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, volume 35, pages 1218–1226, 2021. 2, 5

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 1, 2, 4, 5, 6, 7

[9] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *CVPR*, pages 18009–18019, June 2022. 1

[10] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao. Hypergraph neural networks. In *AAAI*, volume 33, pages 3558–3565, 2019. 4, 5, 7

[11] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, pages 14105–14115, June 2022. 1

[12] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5

[14] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *ICCV*, pages 5753–5762, 2019. 2

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 212–228, 2018. 1, 2, 4, 5

[16] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4653–4661, 2019. 1, 2, 4, 5, 6, 7, 8

[17] Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the geometry of probabilistic box embeddings. In *ICLR*, 2019. 2, 3

[18] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, June 2022. 1

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5

[20] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM MM*, page 3–11, 2019. 1, 2

[21] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10918–10927, 2020. 2, 4

[22] Hongying Liu, Ruyi Luo, Fanhua Shang, Mantang Niu, and Yuanyuan Liu. Progressive semantic matching for video-text retrieval. In *ACM MM*, page 5083–5091, 2021. 2

[23] Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. Gradual: Graph-based dual-modal representation for image-text matching. In *WACV*, pages 2463–2472, 2022. 2

[24] Zhou Mo, Niu Zhenxing, Wang Le, Gao Zhanning, Zhang Qilin, and Hua Gang. Ladder loss for coherent visual-semantic embedding. In *AAAI*, pages 13050–13057, 2020. 2

[25] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. Modeling uncertainty with hedged instance embedding. In *ICLR*, 2019. 2, 3

[26] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *CVPR*, pages 14711–14721, June 2022. 4

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 3, 5

[28] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016. 5

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 1

[30] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning. In *CVPR*, pages 16177–16189, June 2022. 1

[31] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Non-isotropy regularization for proxy-based deep metric learning. In *CVPR*, pages 7420–7430, June 2022. 1

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg,

and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[33] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*, pages 7462–7471, June 2022. 1

[34] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019. 2, 3, 5, 6, 7

[35] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *CVPR*, pages 19437–19446, June 2022. 1

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 6000–6010, 2017. 3

[37] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *ACL*, pages 263–272, 2018. 2

[38] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 2020. 1, 2

[39] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2019. 2

[40] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *IJCAI*, page 3792–3798, 2019. 1

[41] Zheng Wang, Zhenwei Gao, Xing Xu, yadan Luo, Yang Yang, and Heng Tao Shen. Point to rectangle matching for image text retrieval. In *ACM MM*, 2022. 2, 3, 4, 5, 6, 7, 8

[42] Jiwei Wei, Xing Xu, Zheng Wang, and Guoqing Wang. Meta self-paced learning for cross-modal matching. In *ACM MM*, page 3835–3843, 2021. 1, 2, 5

[43] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *CVPR*, pages 13002–13011, 2020. 1, 2

[44] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 5, 6

[45] Dongqing Wu, Huihui Li, Yinge Tang, Lei Guo, and Hang Liu. Global-guided asymmetric attention network for image-text matching. *Neurocomputing*, 481:77–90, 2022. 2

[46] Shiyang Yan, Li Yu, and Yuan Xie. Discrete-continuous action space policy gradient-based attention for image-text matching. In *CVPR*, pages 8092–8101, 2021. 2

[47] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *CVPR*, pages 12522–12531, 2021. 4

[48] Rui Yu, Dawei Du, Rodney LaLonde, Daniel Davila, Christopher Funk, Anthony Hoogs, and Brian Clipp. Cascade transformers for end-to-end person search. In *CVPR*, pages 7267–7276, June 2022. 1

[49] H. Zhang, Z. Mao, K. Zhang, and Y. Zhang. Show your faith: Cross-modal confidence-aware network for image-text matching. In *AAAI*, volume 36, pages 3262–3270, 2022. 2

[50] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *CVPR*, pages 15661–15670, June 2022. 1, 2, 4, 5

[51] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3533–3542, 2020. 1, 2

[52] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R²gan: Cross-modal recipe retrieval with generative adversarial network. In *CVPR*, pages 11469–11478, 2019. 1