Probabilistic Compositional Embeddings for Multimodal Image Retrieval

Andrei Neculai[1], Yanbei Chen[1], Zeynep Akata[1,2,3]

[1]University of Tübingen   [2]MPI for Informatics   [3]MPI for Intelligent Systems

neculai.andrei0@gmail.com, {yanbei.chen, zeynep.akata}@uni-tuebingen.de

Abstract

Existing works in image retrieval often consider retrieving images with one or two query inputs, which do not generalize to multiple queries. In this work, we investigate a more challenging scenario for composing multiple multimodal queries in image retrieval. Given an arbitrary number of query images and (or) texts, our goal is to retrieve target images containing the semantic concepts specified in multiple multimodal queries. To learn an informative embedding that can flexibly encode the semantics of various queries, we propose a novel multimodal probabilistic composer (MPC). Specifically, we model input images and texts as probabilistic embeddings, which can be further composed by a probabilistic composition rule to facilitate image retrieval with multiple multimodal queries. We propose a new benchmark based on the MS-COCO dataset and evaluate our model on various setups that compose multiple images and (or) text queries for multimodal image retrieval. Without bells and whistles, we show that our probabilistic model formulation significantly outperforms existing related methods on multimodal image retrieval while generalizing well to query with different amounts of inputs given in arbitrary visual and (or) textual modalities. Code is here: https://github.com/andreineculai/MPC.

1. Introduction

Image retrieval aims to learn an embedding that encodes the semantics of an input query for retrieving the most relevant target images in the database. Existing works in image retrieval often consider taking one single image [19] or a text sentence [16] as query, which limit their applicability in coping with multiple queries given in arbitrary modalities. In this work, we investigate a more challenging scenario in multimodal image retrieval, where the queries may consist of an arbitrary number of queries in arbitrary visual and textual modalities (e.g. images and texts).

As Figure 1 shows, given an arbitrary number of image and (or) text queries, our goal is to retrieve the images that contain all the semantic concepts specified in the queries. Inspired by the recent advances in compositional
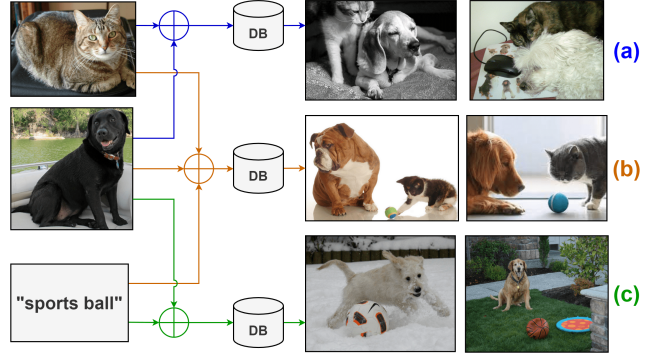

Figure 1. We consider a new compositional learning paradigm for multimodal image retrieval. Given an *arbitrary* number of multimodal queries (e.g. images and texts), the model is trained to learn a feature embedding for retrieving images that contain the composite set of semantic concepts specified in the queries, where the input could be given in *arbitrary* combinations such as (a) image + image, (b) image + image + text, (c) image + text, etc.

learning for visual recognition [38, 42, 53], we tackle this problem by learning a compositional embedding to flexibly encapsulates the multiple semantic concepts specified in the multimodal queries, and to be used for retrieving the more relevant images. For instance, when giving two query images capturing "*cat*" and "*dog*", and one text query stating "*sports ball*", we aim to learn a compositional embedding that represents all three queries to retrieve images that contain "*cat*", "*dog*" and "*sports ball*" (see Figure 1 (b)).

A line of recent works [7, 8, 24, 61, 64] also explore compositional learning upon image and text. However, all these existing works consider evaluating and modelling upon only *two* queries given in a fixed combination of image and text. Moreover, these works mainly consider using text to specify the modifications in input image for retrieving similar images, which does not apply to the case when the retrieved images are expected to represent a composite set of multiple semantic concepts. Therefore, to further investigate image retrieval using an *arbitrary* number of queries (e.g. more than two queries) given in *arbitrary* visual and (or) textual modalities, we establish a new evaluation benchmark for multimodal image retrieval. Our new benchmark

is built with the MS-COCO dataset [34] and offers a more advanced and challenging testbed to facilitate research on image retrieval. We propose this new benchmark with a comprehensive evaluation of multiple state-of-the-art multimodal methods and propose a novel Multimodal Probabilistic Composer (**MPC**) that learns an informative probabilistic compositional embedding to flexibly encode an arbitrary number of queries. Specifically, rather than fusing embeddings from different modalities through a fixed set of learnable parameters, we propose to parameterize each input as a probabilistic embedding following a multivariate Gaussian, and compose embeddings by deriving a composite multivariate Gaussian based on a probabilistic composition rule.

Our model formulation offers two unique properties to learn a compositional embedding for multimodal image retrieval. First, our probabilistic composer allows to compose embeddings of a flexible amount of queries in arbitrary modalities. Second, its probabilistic nature allows to encode semantics as well as ambiguities of a given input, thus well capturing the polysemantic information in text queries, e.g. a text query "*dog*" may refer to a variety of dog breeds that differ visually. These properties well faciliate better performance in multimodal image retrieval.

Our **contribution** is three-fold:

- We establish a new multimodal image retrieval benchmark using the MS-COCO dataset to investigate image retrieval using an arbitrary number of queries in arbitrary modalities. We evaluate a variety of settings including (1) using different combinations of input modalities, and (2) using various number of queries.

- We propose a Multimodal Probabilistic Composer (MPC), which features a new probabilistic rule to compose probabilistic emebddings and a new probabilistic similarity metric to compare probabilistic embeddings, which together lead to its superior model performance in composing multimodal queries for image retrieval.

- We show that our model outperforms existing multimodal fusion methods significantly for mulitmodal image retrieval. To further analyze our model design rationale, we also conduct an indepth experimental analysis.

2. Related Work

Multimodal image retrieval refers to image retrieval with multimodal queries. In contrast to standard image retrieval that takes in one image as query, multimodal image retrieval digests multiple inputs in different modalities such as images, spatial layout [37], sketch [18, 67, 69], texts in the attribute formats [1, 22, 29, 47, 68, 70] or natural language sentences [7, 8, 20, 21, 24, 64]. Unlike these existing works that consider the other non-visual modalities as feedback to guide image retrieval, we focus on learning the compositions of multiple semantic concepts specified in the query.

Compositional learning provides an important functionality in an artificial intelligent machine [30, 31]. The goal of compositionality is to learn representations or data distributions that encapsulate multiple semantic concepts [39, 40, 42, 43, 51, 53, 56, 58, 61, 63]. One line of works explore compositionality in zero-shot recognition [38, 42, 53], where the model is trained to compose states and objects (e.g. "*ripe*" and "*apple*") and used is to recognize their compositions. Another line of works introduce compositionality for image generation, which aim to synthesize new images that contain composite visual primitives [3, 45, 48, 56, 60] or composite context and semantic content [10, 33, 59, 66]. More recently, compositionality is explored in various tasks such as image retrieval [8, 24, 61], video recognition [9, 39], and visual relationship detection [25, 51]. In this work, we study a new compositional learning paradigm for multimodal image retrieval. Unlike the relevant works in image retrieval that consider composing image and text for modifying image content [8, 24, 61], we establish a new benchmark to investigate compositions upon an *arbitrary* number of semantics concepts given in *arbitrary* modalities. Moreover, we propose a novel probabilistic model formulation to derive probabilistic compositional embeddings that jointly capture semantics and ambiguities in input data to learn more informative compositional embeddings.

Multimodal fusion is widely adopted as an effective learning scheme to fuse complementary and (or) compositional information from multimodal input data for building a more powerful model. A large body of works exploit multimodal fusion to enhance video understanding, such as utilizing audio and visual data [9, 26, 36, 44, 62], or using text and visual data [17, 55] to learn video representations. A branch of recent works leverage multimodal fusion to fuse RGB images with depth or LiDAR input to enhance representation learning in autonomous driving [52, 65]. Another volume of works use multimodal fusion to integrate images and texts for solving multimodal image retrieval [7, 8, 24, 61, 64] or visual question answering (VQA) [2, 57]. Unlike existing works that fuse a fixed combinations of modalities, we explore multimodal fusion upon an arbitrary composite set of semantic concepts specified in different data modalities.

3. Multimodal Probabilistic Composer

Given a composite set of k input samples – where each input specifies a semantic concept (e.g. "*dog*", "*sport ball*"), our goal is to learn a compositional embedding for retrieving the corresponding target images that contain the set of specified semantic concepts. Each input can be given in visual or textual modality, and passed through modality-specific encoder to obtain its embedding. Accordingly, the composition of different embeddings should be modality-agnostic, which means our model by default is trained to combine an arbitrary set of samples in arbitrary modalities.
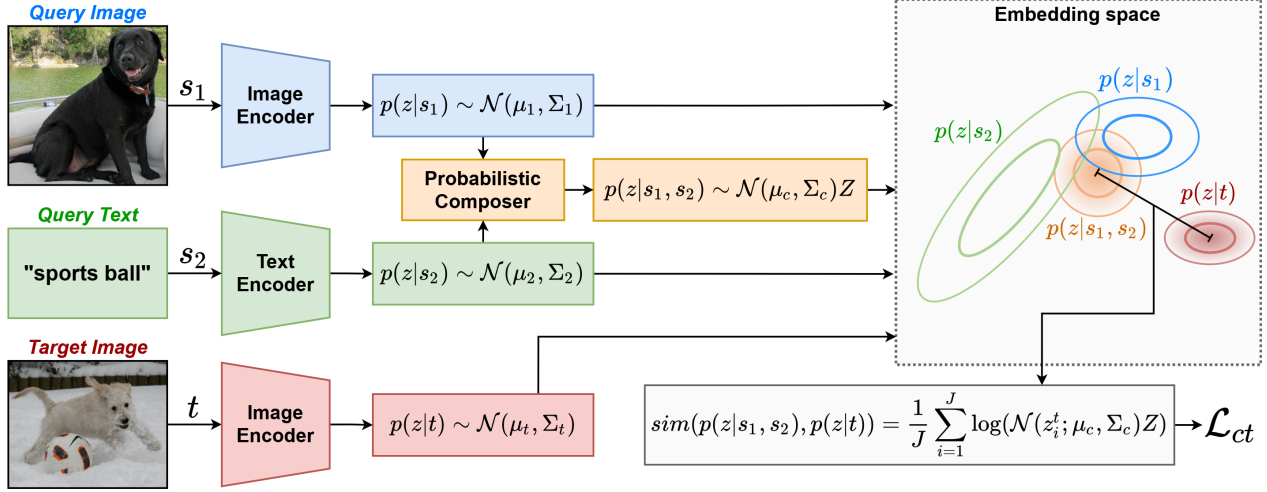
Figure 2. Model overview. We show how Multimodal Probabilistic Composer (MPC) learns from *two* queries for *visual simplicity*. Here, given the query image s_1 and text s_2, MPC first learns two modality-specific probabilistic embeddings (Section 3.1), and then composes the two embeddings through a probabilistic composer to obtain a probabilistic compositional embedding (Section 3.2) which is further aligned with the probabilistic embedding of the target image t by minimizing a probabilistic distance metric (Section 3.3). Note: MPC can digest *more than two* queries by applying the probabilistic composer on a set of probabilistic embeddings (detailed in Section 3.2).

Figure 2 gives an overview of our proposed model: Multimodal Probabilistic Composer (MPC). For simplicity, we visualize the case where MPC takes in two queries (a "*dog*" image and a "*sport ball*" text snippet) to retrieve a target image that contains "*dog*" and "*sport ball*". The image and text are passed through modality-specific encoders to learn probabilistic embeddings (Section 3.1) and then composed to derive a compositional embedding (Section 3.2). MPC is optimized to align the compositional embedding and the target image embedding by minimizing a probabilistic contrastive loss (Section 3.3). While Figure 2 shows the scenario with two inputs, MPC is designed with the flexibility to process an arbitrary set of samples in arbitrary modalities thanks to its probabilistic nature.

3.1. Modality-Specific Probabilistic Embeddings

In this section, we first describe how different modalities (i.e. image and text) are modeled and then detail how probabilistic embeddings are learned to represent each input.

Image encoder. We use a ResNet [23] backbone f_{ResNet}, with an additional linear projection layer f_{img}, as our image encoder to learn the image embeddings. Given an input image (referred as s_1), we pass it through f_{ResNet} to obtain the feature map ϕ_{img} and compute its feature encoding as: $z_{img} = f_{img}(\phi_{img})$, where $z_{img} \in \mathbb{R}^D$.

Text encoder. To encode text information, we use GloVe word embeddings [49] to encode each word (denoted as f_{GloVe}) and train a bidirectional GRU [11] (denoted as f_{txt}) to learn the text embeddings. Given a text snippet (referred as s_2), we obtained the word embeddings: $\phi_{txt} = f_{GloVe}(s_2)$. We pass ϕ_{txt} through the GRU to obtain its

feature encoding: $z_{txt} = f_{txt}(\phi_{txt})$, where $z_{txt} \in \mathbb{R}^D$.

Probabilistic embeddings. Our model is motivated with the flexibility to take a composite set of k queries in arbitrary modalities. With this design rationale in mind, we propose to model each embedding as a multivariate Gaussian probability density function (PDF), such that the compositions of different embeddings can be achieved by composing different Gaussian PDFs through a parametric probabilistic rule, i.e. the product of k Gaussian PDFs [5]. Below, we detail how each embedding is modeled as a multivariate Gaussian, similar to recent probabilistic embeddings works [12, 46], and present how we derive compostional embeddings through our probabilistic composer in Section 3.2.

Given any embedding and feature map (denoted as z_m and ϕ_m) from either the image encoder or the text encoder, we model it as a Gaussian with mean μ_m and diagonal variance matrix Σ_m. For computational reasons, we work with only the diagonal of the covariance matrix Σ_m, referred as σ_m^2. As seen in [12], μ_m and σ_m^2 are computed as follows:

$$\begin{aligned} \mu_m &= \text{LN}((z_m + s(\text{fc}(\text{attn}(\phi_m))))) \\ \log(\sigma_m^2) &= z_m + \text{fc}(\text{attn}(\phi_m)) \end{aligned} \tag{1}$$

where LN denotes a LayerNorm [4] operation, fc denotes a linear layer, $s(\cdot)$ is a sigmoid activation function and attn denotes a self-attention module [35]. The model outputs the logarithm of the variance instead of the actual variance for numerical stability. In essence, the Gaussian PDF (Eq. 1) offers a more informative and expressive estimation about the embedding, capturing not only the semantics of an input but also its ambiguities. By design, imposing a probabilistic

distribution on a feature embedding has the following merits. First, it is essential in practice to encode both semantics and ambiguities of an input – e.g. encoding a query of "*cat*" needs to represent its semantic content as a cat but also need to represent its ambiguities to reflect different possible variations as not all cats visually look the same. Second, composing a set of k probabilistic embeddings can be easily derived by composing their Gaussian PDFs through a parametric probabilistic formulation, as detailed next.

3.2. Probabilistic Composer

Probabilistic composer. Let $S=\{s_1, \ldots s_k\}$ be the set of k samples in the query. The goal of the composition is to find a probability distribution $p(z|S)$ that unifies all the individual probability distributions $\{p(z|s_i) \sim \mathcal{N}(\mu_i, \Sigma_i)\}_{1,k}$. As aforementioned, composing a set of k probabilistic embeddings can be achieved based on a parametric probabilistic rule which derives the product of k Gaussian PDFs [5]. We refer this process as our probabilistic composer and explain the case where $k = 2$ (i.e. two queries) below. Formally, the product of two Gaussian PDFs can be written as:

$$\mathcal{N}(z;\mu_1,\Sigma_1)\mathcal{N}(z;\mu_2,\Sigma_2) = \mathcal{N}(z;\mu_c,\Sigma_c)Z \quad (2)$$

where $\mathcal{N}(z;\mu_1,\Sigma_1), \mathcal{N}(z;\mu_2,\Sigma_2)$ are the two Gaussian PDFs for the set of two queries s_1, s_2. $\mathcal{N}(z;\mu_c,\Sigma_c)$ is the new composite Gaussian with mean μ_c and variance Σ_c, and Z is a normalization constant, defined as:

$$\begin{aligned} \Sigma_c &= (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} \\ \mu_c &= \Sigma_c(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) \\ Z &= \mathcal{N}(\mu_1;\mu_2,\Sigma_1 + \Sigma_2) \end{aligned} \quad (3)$$

While Eq. 2, Eq. 3 show the special case of $k = 2$, it can be easily generalized to the case where $k > 2$ by deriving the product of k Gaussian PDFs sequentially – i.e. multiplying the k_{th} Gaussian PDF with the product of all the $(k-1)$ Gaussian PDFs, which results in a composite new Gaussian with a normalization constant Z written as follows.

$$Z = \prod_{i=1}^{k-1} Z_{i,i+1} \quad (4)$$

where $Z_{i,j}$ is the normalization constant of the product between the Gaussian PDFs of input i and input j. Thanks to the inductive bias induced by our probabilistic composer, our model can generalize to compositions of a flexible set of queries even without training. In our experiments, we show that despite not containing any learned parameters, our probabilistic composer outperforms other composition methods that learn fusion layers to compose embeddings.

3.3. Model Optimization

Similar to standard objectives in non-probabilistic metric learning such as triplet loss and contrastive loss, our training objective is imposed to pull the distribution of the compositional embedding and the target image distribution closer, while pushing away the distributions of negative pairs. To achieve this aim, we first need to define a probabilistic similarity function between two probability distributions.

Probabilistic similarity. To quantify the similarity between two probabilistic distribution, Monte-Carlo estimation can be adopted [12], which draws a number of J data points $\{z_i^x\}_{1,J}$, $\{z_i^y\}_{1,J}$ from two distributions $p(z|x), p(z|y)$, resulting in J^2 pairs of feature vectors that can be used to compute the similarity between two distributions as:

$$sim(p(z|x), p(z|y)) = \frac{1}{J^2} \sum_{i=1}^{J} \sum_{j=1}^{J} \kappa(z_i^x, z_j^y) \quad (5)$$

where $\kappa(\cdot, \cdot)$ refers to a standard similarity metric between feature vectors, e.g. cosine similarity. Although Eq. 5 can measure the similarity between two probabilistic embeddings that follow Gaussian distributions, it could not be directly applied in our scenario given that our composite Gaussian is scaled by a constant Z (Eq. 4); while sampling data points from $\mathcal{N}(z;\mu_c,\Sigma_c)$ is equivalent to sampling from $\mathcal{N}(z;\mu_c,\Sigma_c)Z$.

To align the probabilistic compositional embedding (Eq. 2) – modeled as a composite Gaussian $\mathcal{N}(z;\mu_c,\Sigma_c)Z$ – and the target image embedding – modeled as a multivariate Gaussian $\mathcal{N}(z;\mu_t,\Sigma_t)$), we propose a new probabilistic similarity function to measure the similarity between the composite Gaussian $p(z|S)$ and the target image distribution $p(z|t)$ using the following procedure. First, we sample J data points $\{z_i^t\}_{1,J}$ from the target distribution $\mathcal{N}(\mu_t,\Sigma_t)$. Then, we assign these data points to the composite distribution $\mathcal{N}(z;\mu_c,\Sigma_c)Z$ to compute the probabilistic similarity scores. The similarity function is formally defined as the logarithm of these probabilistic scores:

$$sim(p(z|S), p(z|t)) = \frac{1}{J} \sum_{i=1}^{J} \log(\mathcal{N}(z_i^t;\mu_c,\Sigma_c)Z) \quad (6)$$

where we compute the overall similarity with logarithm to ensure numerical stability given that Z may contain small numerical values. In essence, Eq. 6 provides a tractable probabilistic formulation to measure the similarity between two probabilistic embeddings that follow the Gaussian distributions $\mathcal{N}(z;\mu_c,\Sigma_c)Z$ and $\mathcal{N}(z;\mu_t,\Sigma_t)$. Eq. 6 is also computationally more efficient than Eq. 5, given that their computational complexity is $O(J)$ vs $O(J^2)$.

Learning objective. In similar spirit as the cross entropy loss and contrastive loss [6], we define our loss function as:

$$\mathcal{L}_{ct} = \frac{1}{B} \sum_{i=1}^{B} -\log \frac{\exp(sim(p(z|S_i), p(z|t_i)))}{\sum_{j=1}^{B} \exp(sim(p(z|S_i), p(z|t_j)))} \quad (7)$$

where B denotes the batch size. $sim(p(z|S_i), p(z|t_i))$ is the similarity between probabilistic distributions of two positive pairs. The loss $\mathcal{L}_{ct}$ is computed across all positive pairs. To ensure the training stability and prevent the learned variance σ_m^2 (Eq. 1) from collapsing to zero or exploding to very high values, we add a ℓ_2 regularization term on the logarithm of the variance, as defined below:

$$\mathcal{L}_{\ell_2} = \frac{1}{B|S|} \sum_{i=1}^{B} \sum_{j=1}^{S} (\log(\sigma_{i,j}^2))^2 \tag{8}$$

where $|S|$ is the number of queries being composed. $\sigma_{i,j}^2$ is the variance of the input j in the i_{th} pair of queries in the batch. The final loss is $\mathcal{L} = \mathcal{L}_{ct} + \lambda_{\ell_2} \mathcal{L}_{\ell_2}$.

Model deployment. At test time, we compute the compositional probabilistic embeddings of the set of queries by first computing the probabilistic embeddings (Eq. 1) and then composing the embeddings through our probabilistic composer (Eq. 2, Eq. 3). For image retrieval, we compute the probabilistic embeddings for all the images in the database, and match the composite queries with all images based on the cosine similarity between the mean μ_c of a compositional probabilistic embeddings and the mean μ_t of an image embedding. The images with the top-k similarity scores in the database are the top-k retrieved items.

4. Experiments

4.1. Constructing Benchmark

Dataset. To explore compositional learning upon a flexible combination of queries, we establish a new benchmark based on the MS-COCO dataset [34], which contains bounding box annotations of 80 unique object categories. We utilize these data annotations to construct our multi-modal image retrieval benchmark. As each image in MS-COCO often contains a set of objects labeled with bounding boxes, we can easily obtain *full images* that contain a composite set of objects, as well as *cropped images* that contain a single object by applying the bounding boxes. Concretely, we use the *cropped images* with one object (e.g. *dog*, *cat* and *truck*) as the query images, and use the category label (e.g. *sport balls*, *carrot* and *stop sign*) as the query texts. It is worth noting that we consider short texts in this work; however, our model is able to handle free-form texts, given that the text encoder can encode both words and sentences. The query images and (or) query texts can be composed by composition methods to retrieve the target images that contain the composite set of specified semantic concepts. More details about the benchmark are given in the supplementary.

Benchmark setup. To build a comprehensive benchmark, we consider various evaluation setups, as detailed next.

- composing different input modalities: images only, texts only, and multimodal query with images and texts.

- seen compositions vs unseen compositions: the former considers that the composite concepts encountered at test time are seen during training, while the latter considers the compositions are new and unseen at test time.
- varying number of queries: 2, 3, 4 queries.

We construct a training/validation/test data split with around 78,000/20,000/20,000 image samples using the MS-COCO dataset. For the seen compositions, we select 1000 category tuples for training and validation and we use the same tuples for testing. In the unseen compositions scenario we pick 100 pairs for training and validation and 500 different pairs for testing (the training and testing pairs use the same categories). Due to the high number of image samples for each category, the actual number of possible compositions of images and (or) texts is extremely high. Thus, the same compositions of queries are rarely repeated during training, offering rich diversity in the constructed query-target pairs for training. Due to space limit, we give data statistics and evaluate on other datasets in supplementary.

Evaluation metrics. We use two metrics to evaluate image retrieval. First, we use recall@K (R@K), which measures the percentage of at least one correct retrieved images occurred in the top K retrieved items. Second, we use R-Precision (R-P) [41], which takes into account that there are multiple correct retrieved images for each query, i.e. R-P$= \frac{1}{N} \sum_{i=1}^{N} \frac{r}{R}$, where N is the number of test queries; for each test query, R is the number of correct target images in the gallery, and r is the number of correct retrievals in the top-R retrieved items. R-P has a high score only if the model ranks all the correct items before the incorrect ones.

Implementation details. We implement our model using PyTorch. For the image encoder, we use the ResNet-50 network [13] pre-trained on ImageNet. The 2048-D image embeddings are passed through a self-attention module with a hidden layer of dimension 1024 and a fc layer that maps it to a 512-D probabilistic embedding (Eq. 1). For the text encoder, we use the GloVe 300-dimensional word embeddings pre-trained on the corpus with 1.9M vocabulary [49], and train a bidirectional GRU [11] with 256 features in its hidden state. The 256-D text embeddings are passed through a self-attention module with a hidden layer of dimension 150 and a fc layer that maps it to a 512-D probabilistic embedding (Eq. 1). To compute the probabilistic similarity in Eq. 6, we use $J = 7$, same as [12]. We set the weight λ_{ℓ_2} for the regularizer $\mathcal{L}_{\ell_2}$ to 0.001. For data augmentation, we use Cutout [14], random resizing and random horizontal flips for the images and apply random word dropout with probability 0.1. For training, we use the Adam optimizer [28] and our learning rate is 2e-4. We set the learning rate of ResNet layers before the GAP layer to 1/10 the general learning rate. We train our model for 1600 epochs and decrease the learning rate by a factor of 1/10 every 600 epochs. We use the same data augmentation and training scheme for differ-

Method	images only			multimodal			texts only		
	R@5	R@10	R_P	R@5	R@10	R_P	R@5	R@10	R_P
Relationship	0.36	0.72	0.11	0.47	0.77	0.13	0.61	1.31	0.14
FiLM	0.40	0.92	0.15	0.51	1.06	0.18	0.37	1.06	0.15
MRN	19.81	28.17	4.52	24.59	34.83	5.46	33.20	42.48	6.73
TIRG	21.98	31.58	5.80	18.94	26.85	4.87	34.79	51.43	9.09
PCME + addition	15.19	21.86	3.66	20.02	29.41	4.88	28.78	39.94	6.62
MPC	**27.32**	**36.52**	**6.72**	**37.76**	**48.23**	**9.10**	**59.32**	**69.42**	**15.14**

Table 1. Evaluation of composing *two* queries for image retrieval on a *seen* composition setup.

Method	images only			multimodal			texts only		
	R@5	R@10	R_P	R@5	R@10	R_P	R@5	R@10	R_P
Relationship	0.40	0.49	0.13	0.41	1.02	0.17	0.57	1.62	0.27
FiLM	0.00	0.16	0.04	0.01	0.12	0.05	0.00	0.00	0.03
MRN	**9.89**	**16.96**	**2.21**	17.63	26.97	3.73	25.16	36.48	5.14
TIRG	6.95	11.97	1.62	4.77	8.93	1.30	2.91	5.01	0.76
PCME + addition	4.61	8.00	1.09	7.37	12.78	1.68	10.44	16.26	2.08
MPC	8.90	15.13	1.94	**19.26**	**28.03**	**4.15**	**32.52**	**43.28**	**6.75**

Table 2. Evaluation of composing *three* queries for image retrieval on a *seen* composition setup.

ent methods to ensure fair comparisons.

4.2. Comparing to the State-of-the-Art

Competitors. We compare our model MPC to a variety of state-of-the-art approaches that perform multimodal fusion. We re-implement some methods to ensure they could perform flexible compositions of varying numbers of queries in arbitrary modalities, as described next.

- Relationship [54]: A relation reasoning module. It concatenates image features and text features, followed by an MLP to obtain a compositional embedding.

- FiLM [50]: A feature-wise linear modulation module. It learns affine transformation layers to modulate the image features based on the text features.

- MRN [27]: A multimodal fusion module. It fuses the image and text features through three blocks of element-wise multiplication and residual learning.

- TIRG [61]: An multimodal composition module. It composes image and text features by concatenation, followed by a gating function and a residual connection for fusion.

- PCME + addition [12]: It uses the same probabilistic embeddings in Eq. 1. For composition, we add the two probabilistic embeddings directly, as defined below.

$$X \sim \mathcal{N}(\mu_1, \Sigma_1); Y \sim \mathcal{N}(\mu_2, \Sigma_2)$$
$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2) = \mathcal{N}(\mu_c, \Sigma_c) \quad (9)$$

<u>**Discussion.**</u> Except for PCME, the other methods above were specifically designed to combine deterministic image and text features (either for VQA or for image search with text feedback). Our method is designed to leverage probabilistic embeddings and combine them in a way in which we can capture the polysemantic nature of the data. Our composition method is also agnostic of the input modality or the number of queries which allows us to combine any

type of multimodal input. In our experiments, we reproduce the existing methods using the same ResNet backbone and pre-trained word embeddings for fair comparisons.

Quantitative results. As mentioned in Section 4.1, we test various setups, including different input modalities, seen vs unseen compositions, and varying number of inputs. We present our experimental results on these setups below.

Seen compositions with two inputs. Table 1 shows our experimental results on composing two queries for multimodal image retrieval, where the model is tested on compositions seen during training. As can be seen, when given an arbitrary combination of the input modalities (i.e. images only, multimodal queries, texts only), we can observe that our model MPC outperforms the other competitors with substantial margins. On the compositions of image inputs, our MPC obtains a R@5 of 27.32% vs 21.98% by the best competitor TIRG. On the compositions of multimodal inputs, MPC obtains a R@5 of 37.76% vs 24.59% by the best competitor MRN. On the compositions of text inputs, MPC surpasses the best competitor TIRG by a margin of 24.53% (59.32-34.79). Overall, our results show that MPC outperforms the other methods significantly, which indicates its strong generalization to digest an arbitrary combination of different input modalities for image retrieval.

Seen compositions with three inputs. Table 2 presents our experimental results on composing three queries for image retrieval, where we also test combinations of different input modalities. As shown, our MPC achieves the best performance on the compositions of multimodal and text-only inputs, obtaining R@5 of 19.26%/32.52% vs 17.63%/25.16% by the best competitor MRN. Although MPC does not surpass MRN on the compositions of image-only inputs, its performance is quite close to MRN, obtaining a R@5 of 8.90% vs 9.89% by MRN. From Table 1 and Table 2, we observe that while the other methods could perform robustly
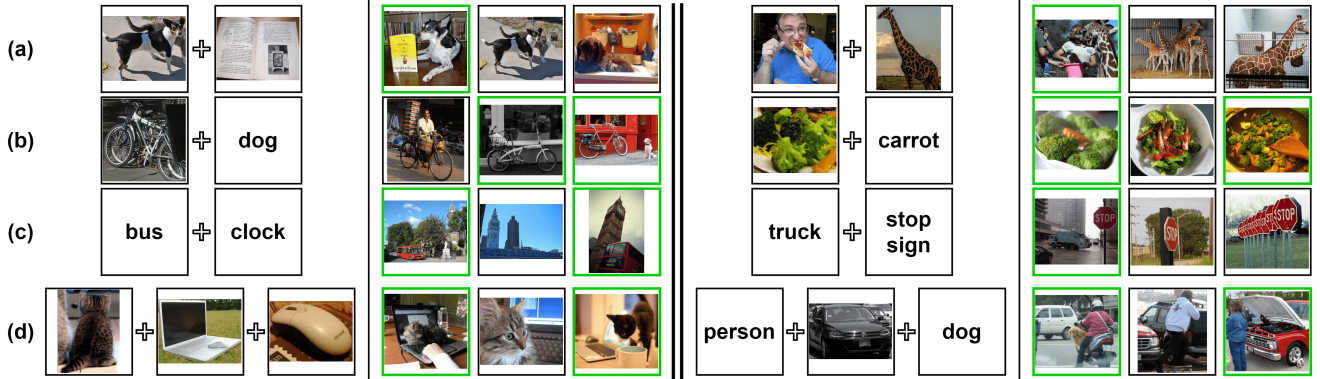
Figure 3. Qualitative results of various types of multimodal queries: (a) image + image, (b) image + text, (c) text + text, (d) image + image + image / text + image + text. For each example, we show the top 3 retrieved images and highlight the groundtruth with a green box.

Method	average		
	R@5	R@10	R_P
MRN	6.24	9.77	1.16
TIRG	14.34	21.28	3.58
PCME + addition	19.27	27.80	4.40
MPC	**26.11**	**37.41**	**6.05**

Table 3. Evaluation of composing *two* queries for image retrieval on a *unseen* composition setup.

Method	average		
	R@5	R@10	R_P
MPC w/o Eq. 2	15.40	24.17	3.43
MPC w/o Eq. 6	30.26	38.62	8.05
MPC	**40.54**	**50.60**	**10.02**

Table 4. Ablation study on our multimodal probabilistic composer.

well under different setups, MPC obtains the best overall model performance in varying number of inputs (2 or 3) and in different combinations of input modalities (image-only, text-only, or multimodal). Moreover, MPC surpasses others with great margins in handling text-only queries, thanks to the fact that it probabilistic nature can well capture the polysemantic semantics of words.

Unseen compositions. We further test a challenging setup with *unseen* compositions – in similar spirit to compositional zero-shot learning [53], we consider the compositions of semantic concepts encountered at test time are unseen during training. Table 3 shows the averaged results reported over all possible combinations of two input modalities (i.e. images only, texts only, multimodal queries), in comparison to three best competitors: MRN, TIRG, PCME + addition. We find that MPC outperforms other methods significantly, obtaining a R_P of 6.05 vs 4.40 by the best competitor PCME + addition. Overall, these results suggest that MPC generalizes well at test time to digest unseen compositions of semantic concepts specified in queries.

Qualitative results. Figure 3 shows the qualitative image retrieval results using a composite set of queries in different modalities. When given two inputs in arbitrary modalities (see (a), (b), (c)), our model can retrieve the images that contain the set of semantic concepts specified in the input, e.g. in the second example in (b), "*broccoli*" image and "*carrot*" text are composed to retrieve a dish with both concepts. When given three inputs (see (d)), our model can discover the images that cover the multiple semantic concepts specified in the input. Another interesting observation

is that when given text queries, we find that our model is capable of capturing their polysemantic nature, e.g. "*clock*" may refer to a wall clock, but when composed with "bus", its semantic meaning becomes a clock tower, as shown in the 3 retrieved images in the first example in (c).

4.3. Model Analysis

Generalization to more complex compositions. As motivated, one uniqueness of our work is that we consider compositions upon an arbitrary number of queries. To test the generalization to handle an increasing number of inputs without additional training, we evaluate our model and three best competitors by training on compositions of 2 inputs and testing on compositions of 2, 3, 4 inputs. The more inputs are composed, the more challenging the task is, as the model needs to retrieve images that match all query concepts at once. Figure 4 shows that under the varying number of inputs, our MPC maintains the best model performance under different combinations of input modalities. It is worth noting that the performance of non-probabilistic methods (MRN, TIRG) degrade to almost random guess ($\approx 0\%$ in R@5) when increasing the number of queries from 2 to 4. In contrast, the probabilistic methods (our MPC, PCME + addition) are able to retain a relatively reasonable model performance even when increasing the number of inputs to 3 or 4 without additional training. Overall, these results indicate that the probabilistic nature of our model helps to encode richer semantics and interactions between different semantic concepts, which strengthens its capability of composing an flexible amount of queries for image retrieval.

Ablation study. To gain insights of our model, we analyze the following two components: our probabilistic composer
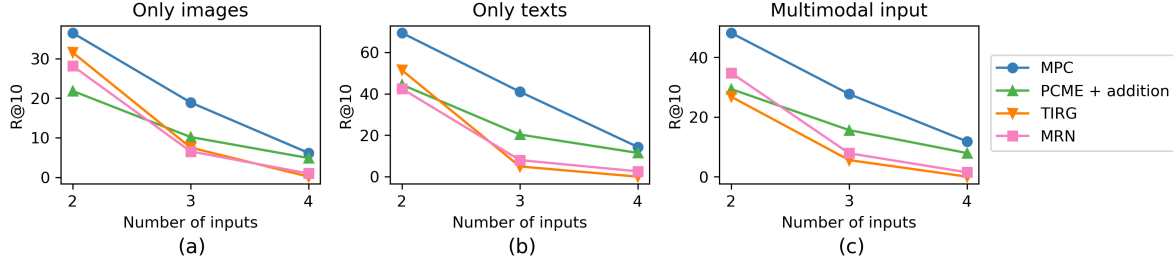
Figure 4. Evaluation of composing an increasing number of multimodal queries for image retrieval. All methods are trained to compose 2 inputs but tested to compose an increasing number of queries (i.e. 2, 3, 4). (a), (b), (c) show results on different combinations of input modalities. Note: as we increase the number of queries, the task difficulty is also increased. Numerical results are in the supplementary.

(Eq. 2) and our probabilistic similarity function (Eq. 6). First, we compare our proposed probabilistic composer (Eq. 2, Eq. 3) with a baseline "MPC w/o Eq. 2" that fuses two embeddings by MLPs: $[\mu_c, \sigma_c^2] = f_{\text{MLP}}([\mu_1, \sigma_1^2, \mu_2, \sigma_2^2])$. Second, we compare our proposed probabilistic similarity function (Eq. 6) with a baseline 'MPC w/o Eq. 2" that uses an existing similarity function (Eq. 5). Table 4 reports the averaged results over different combinations of input modalities. When comparing MPC and "MPC w/o Eq. 2", we find that the model performance degrades significantly from 40.54% to 15.40% in R@5. This suggests that our probabilistic composer works much more effectively compared to simple composition with MLPs, owing to the rich inductive bias induced by our probabilistic composer. When comparing MPC and "MPC w/o Eq. 6", we find that the model performance also decreases substantially from 40.54% to 30.26% in R@5. This indicates that our probabilistic similarity function offers a better estimated similarity as compared to the other similarity function, making our model converge well and learn a better embedding.

Probabilistic uncertainty for discriminating feasibility. The probabilistic nature of our model allows encoding both semantics and uncertainties for given inputs; while the *uncertainty* score of a probabilistic compositional embedding can be naturally used to indicate the *feasibility* in compositions – it is feasible to compose "*broccoli*" and "*carrot*" as they can appear together in a dish; but it is infeasible to compose "*broccoli*" and "*piano*"as they rarely appear together in real-world images. To utilize probabilistic uncertainty for discriminating feasibility, we use the Monte-Carlo estimation to derive uncertainty same as [46]. For non-probabilistic models, we use the distance between two embeddings as the uncertainty score. We construct the feasible and infeasible sets of inputs based on their existence in MS-COCO dataset, where the feasible sets may be seen or unseen during training. We test the binary discrimination of feasibility based on ROC curve [15]. Figure 5 shows the comparison between MPC and the top competitors. We find that the probabilistic models (MPC, PCME + addition) obtain much higher AUC scores compared to the non-probabilistic models (TIRG, MRN); while our MPC obtains
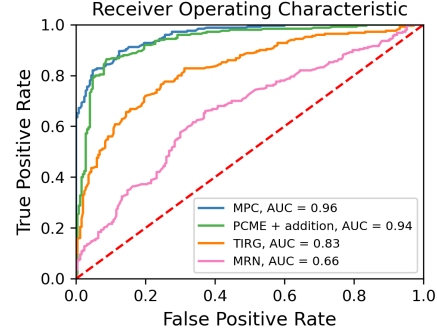

Figure 5. The ROC curves and their AUC scores (area under curve, higher is better) for discriminating feasibility in compositions.

the best AUC of 0.96 among all methods. These results suggest that the uncertainty derived from MPC serves as a superior indicator for discriminating feasibility in compositions.

5. Conclusion and Future Work

In this work, we explored a new compositional learning paradigm for multimodal image retrieval. We established a new benchmark for multimodal image retrieval based on MS-COCO dataset, which poses a new challenge of composing a flexible amount of queries given in arbitrary visual and (or) textual modalities. To tackle this challenge, we proposed a novel multimodal probabilistic composer that learns probabilistic embeddings and composes embeddings based on a parametric probabilistic rule. We formulated a probabilistic composer and a probabilistic similarity function to learn informative compositional embeddings. We demonstrated the superiority of our model compared to multiple existing methods under a diverse set of evaluation setups.

Our model has shown its efficacy; however, there are still critical research problems to be solved for the new challenge posed in this work, such as considering more complex natural language queries, and ensuring model robustness when increasing the number of inputs. Moreover, to bring a positive societal impact, we also need to consider issues such as algorithmic biases and fairness when deploying our model to real-world search engines.

Supplementary materials

A. Additional Details in Benchmark Setup

Dataset and benchmark setup. In order to construct our benchmark, we start from the 118,287 images in the training split of the MS-COCO dataset [34]. We split the images with a 4:1:1 ratio for the train/validation/test sets. Our benchmark is designed for the task of composing the concepts (which are object categories in this work) specified in the input queries for image retrieval. For this aim, we need to generate compositions of individual concepts that are well presented in a sufficient amount of images to serve for the retrieval. We choose 8:2:2 as the splitting thresholds for the train/validation/test splits. This means that a composition of concepts is valid only if it is present in at least 8 of the images in the training set, and 2 of the images in the validation and testing sets respectively. These values ensure that there is enough diversity in the training phase and also multiple correct answers (i.e. more than 1) in the evaluation phase. Using these thresholds allows us to find 1000 viable compositions for using either 2, 3 or 4 input queries. Algorithm A details the process of generating a compositions of k inputs. We will release our benchmark upon acceptance.

Training. During training, for every concept (i.e. category) in the compositional inputs, we randomly choose an image id containing that concept, and also randomly pick either the visual or textual the modality to represent it. The target image is chosen randomly from the images that contain all the k concepts specified in the inputs, but the modality is fixed to visual as we aim to solve an image retrieval task.

Evaluation. At test time, we evaluate for composing k inputs (with k fixed for an evaluation setup). The test queries are described by k concepts and we generate all the test queries randomly to obtain a roughly equal mix of modality combinations. For example, for $k = 2$, we will have around 25% cases for each modality mixture in {*image-image*, *image-text*, *text-image*, *text-text*}.

We also establish evaluation setups to test the model's ability of recognizing unseen compositions for retrieval, as well as testing the model's ability of identifying feasible/infeasible compositions. For the case of testing unseen compositions, we use $k = 2$ and generate 100 compositions of different concepts for training and 500 new compositions of different concepts for testing, where the concepts at test time are seen during training but their compositions are unseen. We choose to use only 100 compositions for the training phase but 500 unseen compositions to simulate a more challenging evaluation setup. For testing the model's ability of identifying feasible/infeasible compositions, we build the

evaluation setup upon the data generated for the seen compositions scenario for $k = 2$. On top of the 1000 compositions, we generate an additional 250 unseen compositions and 250 infeasible compositions. An infeasible composition is defined as infeasible given that it is not found in any image across all the images in the dataset.

B. Numerical Results

We present the numerical results used to generate Figure 4 in the main paper. Figure 4 shows the model performance when being trained to compose two inputs but tested on composing a varying number of inputs (i.e. 2, 3 and 4). We present these numerical results in Table A.

C. Additional Evaluation

Training with more inputs. In the main paper, we explored the scenarios where we train our model to compose two and three query inputs. However, our model formulation is not limited to a fixed number of inputs. Hence, we further explore the scenario where we train the models to compose four query inputs here. Table B contains the results for the top performing models when trained to compose four query inputs and be evaluated on a seen composition setup. As we can see, our model MPC achieves the best performance on the composition of multimodal and text-only inputs, obtaining a R@5 of 5.98%/8.46% vs 4.76%/7.40% by the best competitor MRN. Although MPC does not surpass TIRG on the compositions of image-only inputs, its performance is quite close to TIRG, obtaining a R@5 of 2.76% vs 3.40%. Another interesting observation is that while the other probabilistic model PCME+addition performs relatively well when composing two or three inputs, its performance degrades when composing four inputs. This suggests that composition with addition is weaker at preserving the information from more inputs, while our MPC model formulation with probabilistic composer can better capture the additive information from the increasingly more inputs.

Evaluation on Fashion200k. In the main paper, we showed how our model is capable of composing queries of arbitrary sizes and modalities. In this section we want to showcase the performance of our model on the problem of image retrieval using image and attribute-based text feedback [8, 32, 61]. This is a slightly different setting than our original problem, as in this case the aim is to model the interactions between the image and text queries instead of adding them together. Fashion200k [22] is a dataset that contains around 200k images of fashion products. Each image is tagged with a set of attributes. Using these attributes, [61] generated pairs of products that differ by only one attribute. The text modifications are generated using the attribute that is different between the 2 products. Table C contains the results on the Fashion200k dataset. We limited our compar-

Algorithm A Algorithm for generating a compositions of k concepts (i.e. categories)

compositions_set ← empty set   ▷ initialize a set to store the compositions of k concepts
num_of_compositions ← 0   ▷ record the number of found compositions in dataset
target_num_of_compositions ← 1000   ▷ expected number of compositions to be found
threshold_{train/val/test} ← 8 : 2 : 2   ▷ minimal numbers of compositions in the train/val/test splits
while num_of_compositions < target_num_of_compositions **do**
    cur_composition ← sample k categories w/out replacement   ▷ get images for the current composition of k concepts
    images_{train/val/test} ← get {train/val/test} images that contain "cur_composition"
    if len(images_{train/val/test) ≥ threshold_{train/val/test} **then**
        if cur_composition not in compositions_set **then**
            compositions_set.insert(cur_composition)   ▷ add a found composition to the set
            num_of_compositions ++
        end if
    end if
end while

Input modalities	images only			multimodal			texts only		
Number of inputs	2	3	4	2	3	4	2	3	4
MRN [27]	28.17	6.55	0.97	34.84	7.88	1.49	42.48	8.01	2.57
TIRG [61]	31.58	7.52	0.16	26.85	5.60	0.03	51.43	4.94	0.00
PCME [12] + addition	21.85	10.19	4.87	29.41	15.65	7.95	44.43	20.34	11.48
MPC	**36.52**	**18.93**	**6.17**	**48.23**	**27.73**	**11.85**	**69.42**	**41.02**	**14.35**

Table A. Numerical results of model performance trained for composing 2 inputs but evaluated for composing 2, 3, and 4 inputs. Results are accompanied to Figure 4 in the paper. Metrics: R@10 (%).

Method	images only			multimodal			texts only		
	R@5	R@10	R_P	R@5	R@10	R_P	R@5	R@10	R_P
MRN [27]	2.11	5.19	0.50	4.76	8.44	1.03	7.40	11.78	1.51
TIRG	**3.40**	**6.98**	**0.63**	4.73	7.98	0.97	7.10	9.67	1.21
PCME + addition	0.00	0.16	0.01	0.13	0.19	0.31	0.00	0.00	0.00
MPC	2.76	5.52	0.59	**5.98**	**10.03**	**1.30**	**8.46**	**16.77**	**1.95**

Table B. Evaluation of composing *four* query inputs for image retrieval on a ***seen*** composition setup.

isons to methods that are capable of modeling the original task in the paper. Other methods [8, 32] have better performance, but use complex network architectures, specialized on this task, that can not be used to model arbitrary queries. Because of this, we chose to not add them to the comparison. As we can see, our model MPC achieves the best performance among the compared methods. We can also see that using simple additions of probabilistic embeddings (i.e. PCME + addition) is not capable of modeling the interactions between the image and text parts of the query, leading to poor results. These results show that our MPC also generalizes well on the other benchmark dataset Fashion200k.

Method	R@1	R@10	R@50
Relationship	13.0	40.5	62.4
Film	12.9	39.5	61.9
MRN	12.3	39.4	60.9
TIRG	14.1	42.5	63.8
PCME + addition	1.8	11.4	27.3
MPC	**14.6**	**45.4**	**66.0**

Table C. Retrieval results on the Fashion200k dataset.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 2020. 2

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[5] Paul Bromiley. Products and convolutions of gaussian probability density functions. *Tina-Vision Memo*, 2003. 3, 4

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on*

Machine Learning, 2020. 4

[7] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *European Conference on Computer Vision*, 2020. 1, 2

[8] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 9, 10

[9] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[10] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *IEEE International Conference on Computer Vision*, 2019. 2

[11] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014. 3, 5

[12] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 4, 5, 6, 10

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[14] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. 5

[15] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 2006. 8

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013. 1

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, 2020. 2

[18] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *IEEE International Conference on Computer Vision*, 2019. 2

[19] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, 2016. 1

[20] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, 2018. 2

[21] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019. 2

[22] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. 2, 9

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[24] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[25] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, 2020. 2

[26] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 2021. 2

[27] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, 2016. 6, 10

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 2

[30] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 2

[31] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 2017. 2

[32] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 802–812, June 2021. 9, 10

[33] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 5, 9

[35] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding, 2017. 3

[36] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*, 2019. 2

[37] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[38] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[39] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[40] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[41] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check, 2020. 5

[42] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[43] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *European Conference on Computer Vision*, 2018. 2

[44] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, 2021. 2

[45] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[46] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher. Modeling uncertainty with hedged instance embedding. In *International Conference on Learning Representation*, 2019. 3, 8

[47] Devi Parikh and Kristen Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, 2011. 2

[48] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 2

[49] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 3, 5

[50] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. 6

[51] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[52] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[53] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 7

[54] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, 2017. 6

[55] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *IEEE International Conference on Computer Vision*, 2019. 2

[56] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[57] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[58] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. Learning compositional representations for few-shot recognition. In *IEEE International Conference on Computer Vision*, 2019. 2

[59] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[60] Hung-Yu Tseng, Hsin-Ying Lee, Lu Jiang, Ming-Hsuan Yang, and Weilong Yang. Retrievegan: Image synthesis via differentiable patch retrieval. In *European Conference on Computer Vision*, 2020. 2

[61] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 9, 10

[62] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[63] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *IEEE International Conference on Computer Vision*, 2019. 2

[64] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[65] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal end-to-end autonomous

driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2

[66] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)*, 2012. 2

[67] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *European Conference on Computer Vision*, 2018. 2

[68] Aron Yu and Kristen Grauman. Thinking outside the pool: Active training image creation for relative attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[69] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[70] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2