Bringing Multimodality to Amazon Visual Search System

Xinliang Zhu*
xlzhu@amazon.com
Amazon.com
Palo Alto, CA, USA

Sheng-Wei Huang*
shengweh@amazon.com
Amazon.com
Palo Alto, CA, USA

Han Ding*
handing@amazon.com
Amazon.com
Santa Clara, CA, USA

Jinyu Yang
viyjy@amazon.com
Amazon.com
Palo Alto, CA, USA

Kelvin Chen
kelchen@amazon.com
Amazon.com
New York, NY, USA

Tao Zhou
taozho@amazon.com
Amazon.com
Palo Alto, CA, USA

Tal Neiman
taneiman@amazon.com
Amazon.com
New York, NY, USA

Ouye Xie
ouyexie@amazon.com
Amazon.com
Seattle, WA, USA

Son Tran
sontran@amazon.com
Amazon.com
Palo Alto, CA, USA

Benjamin Yao
benjamy@amazon.com
Amazon.com
Seattle, WA, USA

Douglas Gray
douggray@amazon.com
Amazon.com
Palo Alto, CA, USA

Anuj Bindal
anbindal@a9.com
Amazon.com
Palo Alto, CA, USA

Arnab Dhua
adhua@amazon.com
Amazon.com
Palo Alto, CA, USA

ABSTRACT

Image to image matching has been well studied in the computer vision community. Previous studies mainly focus on training a deep metric learning model matching visual patterns between the query image and gallery images. In this study, we show that pure image-to-image matching suffers from false positives caused by matching to local visual patterns. To alleviate this issue, we propose to leverage recent advances in vision-language pretraining research. Specifically, we introduce additional image-text alignment losses into deep metric learning, which serve as constraints to the image-to-image matching loss. With additional alignments between the text (e.g., product title) and image pairs, the model can learn concepts from both modalities explicitly, which avoids matching low-level visual features. We progressively develop two variants, a 3-tower and a 4-tower model, where the latter takes one more short text query input. Through extensive experiments, we show that this change leads to a substantial improvement to the image to image matching problem. We further leveraged this model for multimodal search, which takes both image and reformulation text queries to improve search quality. Both offline and online experiments show strong improvements on the main metrics. Specifically, we see 4.95% relative improvement on image matching click through rate with the 3-tower model and 1.13% further improvement from the 4-tower model.

CCS CONCEPTS

• **Information systems** → **Novelty in information retrieval**; **Top-k retrieval in databases**; *Clustering and classification*; *Retrieval effectiveness*; **Image search**.

KEYWORDS

Image Retrieval, Deep Metric Learning, Vision Language Model, Multimodal Search

ACM Reference Format:
Xinliang Zhu, Sheng-Wei Huang, Han Ding, Jinyu Yang, Kelvin Chen, Tao Zhou, Tal Neiman, Ouye Xie, Son Tran, Benjamin Yao, Douglas Gray, Anuj Bindal, and Arnab Dhua. 2024. Bringing Multimodality to Amazon Visual Search System. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona,Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3637528.3671640

1 INTRODUCTION

Image matching is a well-studied problem in Computer Vision with a wide range of applications. In this work, we focus on the task of matching a visual query to items in a catalog, where queries are typically lifestyle images with cluttered background (i.e., images in

*Contributed equally to this research

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671640

Figure 1: Street-to-shop problem: we use a lifestyle query image (left) to match product images (right) with simple or white background. Note the domain shift between the query and product images.

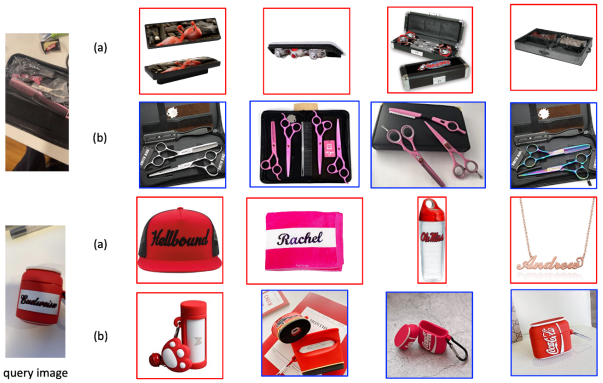

Figure 2: Comparing image match results using traditional pure image-to-image loss versus the proposed method in this paper. Left: query images. Right: retrieved results. Row (a) shows the results from a traditional method and row (b) shows the results from a model trained with the proposed training paradigm. Red boxes mean wrong results. Blue boxes represent exact/similar matches to the query image. In the first query image, a customer tries to find a hair cut set. In the second query image, a customer tries to find an airpod case.

the wild) and catalog consists of item images with simple or white backgrounds. This task, illustrated in Fig. 1, and referred to as the street-to-shop problem in previous works [10, 12, 24], is applicable in various real world settings, such as social networks [3, 48], visual search engines [48], and e-commerce websites [10, 43, 51].

When developing a street-to-shop matching system for a customer-facing visual search engine, it is essential to take into account two key aspects: the quality of the training data and the compatibility and scalability of the algorithm. Training data usually consist of image pairs (e.g., a clean product image and a corresponding lifestyle image depicting the same product). Learning algorithms utilize techniques in distance metric learning to embed images in high dimensional space such that street and shop images of the same object are close to each other while images of different objects are far away from each other. Pair based losses such as triplet loss [31], proxy based losses like proxy anchor [17], and classification based losses like multi-similarity [42] are examples of commonly used training objectives in deep metric learning (DML). Among them, paired losses are easy to scale up [26].

These algorithms can exactly match certain visual structures well. However, since they are trained on instance-based pairs, they often ignore high level semantic relationships, and as a result, suffer from several practical shortcomings. Fig. 2 shows the output of one of such algorithm on two example queries. In the first case, the algorithm tends to match irrelevant but dominant features such as background, while placing less weight on the main object. In the second case, matching to local features leads to irrelevant and inconsistent results. The problem is further aggravated when exact matches cannot be found and the outputs are often degraded in an unpredictable and ungraceful manner. Object localization [30], segmentation [14] or weighted salient maps [40] can be used to reduce irrelevant background. However, they often lead to cumbersome models, and increase complexity, or suffer from their own accuracy problems. One reason for these limitations is that their training data consists of instance-based pairs, which biases the algorithm toward low-level visual feature matching. When the training data has annotated category-level labels, the model might be able to capture higher categorical concepts. There exist datasets with such annotations in the literature such as CUB [38] and iNaturalist [37]. However, they typically require expensive annotations while having limited coverage, for example, with respect to number of classes. It is challenging to acquire large and comprehensive annotations for categories or attributes especially for complex and diverse man-made objects such as products in an online marketplace.

In this work we propose to utilize text information associated with images to learn high level concepts for matching. Multimodal signals are available at large scale in public datasets such as LAION [32], YFCC100M [35], and CC12M [4]. They contain millions to billions of image-text pairs. E-commerce sites and social networks can have up to billions of image-text pairs. Unlike category or attribute labels, the accompanied text is typically unstructured (free text) and complex. Those text can be image captions, web alternative text, text tags or product titles and descriptions with informative content. In particular, for shop products, the titles often contain key information such as category, various attributes (e.g. material, style, shape, appearance) and fine-grained features. Many of them have corresponding visual meaning which we aim to mine for high level semantic matching.

At a high level, our approach is as follows. We align text and image representation into a common embedding space so that matching can be carried out interchangeably across modalities. We design two variants (3-tower and 4-tower models) following the proposed paradigm. In the 3-tower model, we seek to perform alignment in three ways. The first is between query image and catalog image. This is primarily to encourage visual similarity in the output, similar to previous works in image matching [1, 12, 24, 34, 39, 42]. The second is between catalog image and its associated text. Its purpose is to unify the embedding space and establish a correspondence between matching visual and language concepts. It is similar to various works in multimodal learning, especially in vision-language

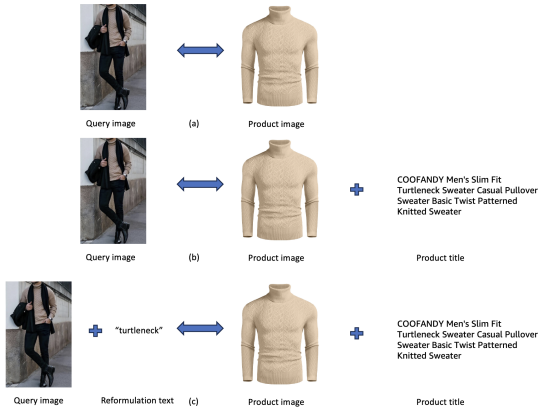

Figure 3: Comparison between existing image match, proposed multimodal image match and multimodal search: (a) existing street-to-shop image matching is query image to product image matching; (b) we propose to use multimodal signals of the products for matching; (c) multimodal search where query image and reformulation text are used to perform the match.

pre-training [16, 18, 19, 29]. The third is between the query image and product text. It is similar to the second alignment, but is a harder task. It helps by reducing the domain gap between lifestyle image, typically with distracting background, and clean text in the catalog. In the 4-tower design, we introduce one more set of alignments between the lifestyle image, catalog image, product title, and a short text query. The short text query is more information dense and cleaner compared to the product title, which further boosts performance. All of these cross modality alignments are carried out using contrastive learning. We name the new method as multimodal image matching (MIM) model. See Fig. 3 for the comparison between traditional image match and proposed MIM and section 3 for further details.

At query time, we match the query embedding to a fused signal from catalog items. We chose a simple fusion which averages the embeddings of catalog text and image. From online experiments, we saw 4.95% relative improvement for the 3-tower model, and a 1.13% further improvement for the 4-tower model as measured by click through rate. See section 4 for further details.

As a by-product of the proposed MIM models, we leverged MIM models for multimodal search, where both query side and catalog side are multimodal. Given the alignment of image and text in MIM models, arithmetic operations can be performed in the latent embedding space when processing both image and text data. Through a combination of offline and online experiments, we demonstrate that employing a straightforward weighted sum of image and text embeddings for query and catalog inputs effectively enhances street-to-shop performance. In practice, users would provide some refinement text (e.g., the type of product they are interested in, its brand, etc.) as the reformulation text into the Multimodal Search system (see Fig. 3 (c) for an example). Although the term "multimodal search" shares the same name with the engine proposed in [33], ours is designed to deal with more scenarios (e.g. reinforcing an attribute, altering an attribute) for more categories (not limited to fashion in [33]).

In summary, our main contributions of this work are:

(1) We introduce vision language alignment to the street-to-shop retrieval problem, where both 3-tower and 4-tower models are designed to boost the match performance. To the best of our knowledge, this is the first work in the direction of image retrieval using multimodal signals, especially for the street-to-shop retrieval problem.
(2) We develop a multimodal search system utilizing the proposed MIM models, which stands among the first multimodal search systems in the industry.
(3) Via extensive experiments, we show the effectiveness of our methods from two perspectives: scaling of the model size and scaling of the training dataset size.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes our approach in details. Section 4 reports our experimental results. Section 5 concludes the paper with a discussion about the potential applications of our approach to multimodal search.

2 RELATED WORK

2.1 Deep Metric Learning

Deep metric learning is vital to modern image matching systems [3, 43, 48, 51]. It aims to learn an embedding space, where similar objects are projected close to each other while distinct objects are kept away. The deep embedding models were trained with different types of losses, including paired losses [13, 15, 31], classification-based losses [1, 27, 47], and proxy-anchor based losses [17, 44]. To boost the performance of the deep embedding model, some researchers also designed multitask learning methods [2], and tried to train models with more data [2]. In this paper, we use a type of paired loss (contrastive loss) since it is scalable to a large number of classes (millions) compared to classification-based and proxy-anchor based losses.

2.2 Vision-language Pretraining

Recent advancements in vision-language pre-training (VLP) have significantly enhanced our ability to align image and text concepts within a shared latent space. This alignment facilitates a range of applications including visual question answering (VQA), image captioning, cross-modal retrieval, etc. Early works in this direction aim to capture visual-language interaction with a multimodal transformer encoder [20, 21, 25, 50]. These works usually require pre-extracted image and text features and rely on object detectors to align image and text concepts.

Recent works such as CLIP [29] and ALIGN [16] demonstrated that pre-training dual-encoder with contrastive objectives on web-scale image-text datasets leads to impressive downstream performance (cross-modal retrieval, zero-shot classification etc.). Following the success of CLIP, many works have focused on scaling up training [6] or improving computation efficiency [11, 22] to further improve downstream performance.

Research has also been performed on combining dual-encoder architectures with existing learning paradigms. In ALBEF [19], the

authors proposed to fuse image and text embeddings with a multi-modal encoder and utilize masked language modeling loss. The authors of LiT [49] combined text encoders with fixed vision encoders that were pre-trained on large-scale image annotation datasets. Authors of Florence [46] proposed a unified contrastive objective which incorporates both contrastive and cross-entropy losses for dual-encoder training. In DeCLIP [23], the authors combined contrastive loss with SimSiam loss [5] and masked language modeling loss in training dual-encoder models, where they observed superior scaling behavior compared to vanilla CLIP model.

There have also been attempts at tackling various vision-language tasks in one single framework by combining generative objectives with contrastive objectives. In [36], the authors studied the scaling property of training a vision encoder with only captioning loss. The trained vision encoder was then combined with a text encoder in LiT [49] fashion to enable zero-shot evaluations. Authors of CoCA [45] proposed to jointly train a dual-encoder and a multi-modal text decoder with both contrastive and captioning losses, which led to state of the art (SOTA) downstream performance. In BEiT3 [41], the author proposed to train a multi-way transformer with both masked image modeling and masked language modeling.

In this paper, we focus on improving a dual-encoder architecture by training it to align multiple input domains. Specifically, we provide insights on the effectiveness of cross-aligning street-to-shop data in the multimodal context.

3 METHOD

Considering the multi-modal and multi-entity nature of Amazon data, we develop two models for the street-to-shop style image match problem. The basic version is a 3-tower model (MIM-3-tower, Fig. 4) based on CLIP [29], which is trained on the triples of {query image, catalog image, product text}. The advanced version is a 4-tower model (MIM-4-tower, Fig. 4), which adds one more text arm to accommodate a new short query text entity in the training data. Compared to the common uses of vision language models (e.g., for VQA, cross-model retrieval), the two models proposed in this paper are demonstrated for the first time at Amazon scale (billions of products) for street-to-shop style product search.

3.1 3-tower Architecture

In this section, we explain the MIM-3-tower model in detail. As shown in Fig. 4, it consists of three encoders: an image encoder $e_q(\cdot)$ for query images, an image encoder $e_c(\cdot)$ for catalog images, and a text encoder $e_t(\cdot)$ for product text. The weights between the two image encoders are shared to save GPU memory during training. Compared to the original CLIP model [29], we add two new contrastive losses to align input pairs of 1) query image and catalog image and 2) query image and product text. As with CLIP, we have multiple choices of the vision encoder size (from ViT-B/16 [9] to ViT-g/14). For the text encoder, we use a 12-layer BERT_{base} [8].

We train the MIM-3-tower model with two groups of objectives: image-image contrastive learning (IIC) on the image encoder, and image-text contrastive learning (ITC) on the image and text encoders. IIC is designed for matching the query images and catalog images as shown in Fig. 4. Specifically, we adopt InfoNCE loss

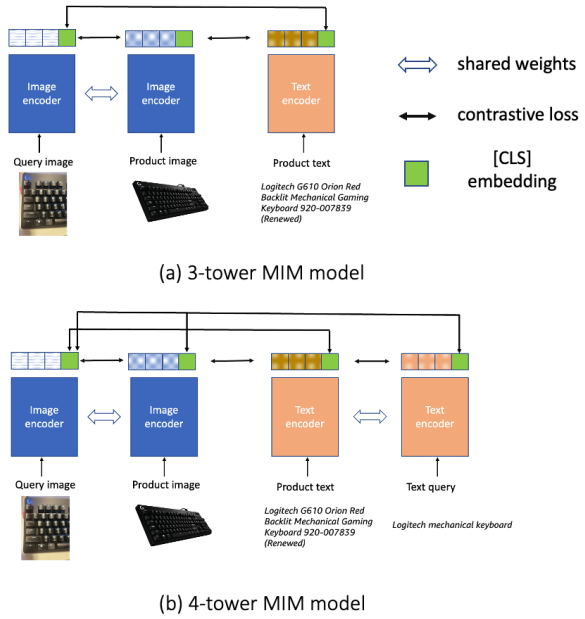

Figure 4: MIM diagram. We develop two variants (3-tower and 4-tower) of the MIM model. In the 3-tower architecture, we have two image encoders and one text encoder. The two image encoders are for query image and product image respectively, and the text encoder is for processing product text. The weights between two image encoders are shared. In the 4-tower architecture, we add one extra text encoder to process short text queries. The two text encoder weights are shared in the 4-tower architecture.

[28] as our contrastive learning objective, which uses categorical cross-entropy loss to identify the positive sample amongst a set of negative samples. Given a query-catalog image pair (Q, C), we define $s(Q, C) = f_q(e_q(Q))^T f_c(e_c(C))$, where $f_q(\cdot)$ and $f_c(\cdot)$ are two linear project layers that map image embeddings to low-dimensional space. For each (Q, C) pair in the training batch, we calculate the softmax-normalized query-to-catalog and catalog-to-query similarity as follows:

$$p^{q2c}(Q) = \frac{exp(s(Q,C)/\tau)}{\sum_{m=1}^{M} exp(s(Q,\tilde{C}_m)/\tau)},$$
$$p^{c2q}(C) = \frac{exp(s(C,Q)/\tau)}{\sum_{m=1}^{M} exp(s(C,\tilde{Q}_m)/\tau)}, \quad (1)$$

where τ is a learnable temperature parameter, $\tilde{C}_m$ and $\tilde{Q}_m$ are category images and query images gathered from all GPUs, respectively. Here, $M = B \times N_g$, where B is the training batch size and N_g is number of GPUs. The rationale of introducing $\tilde{C}_m$ and $\tilde{Q}_m$ is that the success of contrastive learning heavily depends on the number of negative samples, which cannot be easily achieved by in-batch negatives. Denoting $y^{q2c}(Q)$ and $y^{c2q}(C)$ the ground-truth one-hot similarity, with 1 assigned to the positive pairs and 0 to the negative

pairs. The image-image contrastive loss is:

$$\mathcal{L}_{iic} = \frac{1}{2}\mathbb{E}_{(Q,C)\sim D}[H(y^{q2c}(Q), p^{q2c}(Q)) + H(y^{c2q}(C), p^{c2q}(C))], \tag{2}$$

where $H(y, p)$ is the cross-entropy between y and p. Similarly, we can construct the ITC losses between the query image and text, and the catalog image and text. We refer to the CLIP [29] paper for the details for the ITC loss. The full training objectives of MIM-3-tower is:

$$\mathcal{L} = \mathcal{L}_{iic} + \mathcal{L}_{itc}, \tag{3}$$

where L_{itc} represents ITC loss. It is worth noting that the ITC loss is the sum of two parts: query image to product text and catalog image to product text. From equation 3, The IIC is actually a way in deep metric learning to train an image embedding model. The ITC is from the vision language pre-training field, which can be treated as a constraint for the image to image loss (i.e. IIC) in this work. The new way of applying IIC makes a big difference to large scale street-to-shop image match as it can significantly reduce irrelevant matches which are returned due to partial visual pattern matches (e.g., zebra pattern pants vs. zebra pattern mouse pads).

3.2 4-tower Architecture

Encouraged by the success of 3-tower architecture, we experimented with a 4-tower architecture by adding query text as an additional alignment target, illustrated in Fig. 4. Query text is usually search strings containing key attributes pertaining to corresponding products. Compared to text in the Amazon catalog (title, product description, etc), query text strings are shorter in length but have higher information density. Query text strings are also composed of words used by customers and therefore reside in a different language domain than the Amazon catalog.

Our intuition behind training with query text strings was twofold: 1) bridging the language domain gap between shoppers and the Amazon catalog can enhance multimodal search capability of the model. 2) training with text that contains high information density can guide the model toward capturing more essential concepts.

The 4-tower architecture consists of 4 encoder towers (query image, query text, catalog image and product text). We designed the encoders with the same modality to share weights, similar to the 3-tower architecture. During training, the model was supervised with 6 contrastive losses, including every pair of the 4 input types.

3.3 Multimodal Search

Bringing ITC loss to the shop-to-street retrieval problem not only reduces defects for image match but also brings a new opportunity to develop a system we refer to as Multimodal Search (MMS). The definition of Multimodal Search is that given a multimodal query (e.g., an image and a reformulation text), the system returns relevant products meeting the combined intention of the multimodal query. Mathematically, it can be represented as $\{Q_1, Q_2, ..., Q_m\} \rightarrow \{P_1, P_2, ..., P_n\}$, where m and n are the total numbers of modalities in the query and product side, and Q_m, P_n represent a single modality from the query side and product side respectively.

In this paper, we have two modalities (i.e., image and text) for both the query and catalog sides. To make the MMS system work and make use of the popular approximate nearest neighbor search infrastructure in industry, we need a way to convert both multimodal query and catalog into a dense feature vector. There are multiple ways (e.g. training a dedicated neural network to fuse the embeddings) to do the conversion. However, we find the simple weighted sum of image embedding and text embedding works well in practice since the key features from the image and text are aligned in the latent space during training. The formula is as following:

$$X_m = w * X_{image} + (1 - w) * X_{text}, \tag{4}$$

where X_m is the fused embedding, X_{image} is the image embedding and X_{text} is the text embedding. We use different weight w for query and catalog embeddings. With MMS, users can refine pure image search results (e.g., want a specific brand product) or correct image search results by providing more hints/details (e.g., color, product type).

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Training Data. We collected 100 million images of 23 million Amazon products in the daily life context (in-the-wild) to create a training dataset for 3-tower models. Each in-the-wild image has corresponding product and product metadata, including catalog image, title, brand, product description, etc. In total 100M triples of in-the-wild images, catalog image and product text were constructed. We refer to this dataset as 100M triples dataset. The dataset is multi-lingual, covering products from 17 countries. For detailed statistics refer to Table 1.

To create training dataset for 4-tower models, we utilized the subset of 100M triples where query text strings are available. The subset consists of 56 million in-the-wild images and 13 million products. As each product is associated with large amount of query text strings, we expanded this subset by sampling up to 9 query strings per product. With this, we constructed a 400M quadruples dataset consisting of query images, query text, catalog images, and product text. Details can be found in Table 1.

4.1.2 Evaluation Data. We designed our evaluation protocol to focus on recall performance of both multimodal retrieval and image to multimodal retrieval. We designed these two tasks to reflect how our model is utilized to serve customers, where multimodal search is powered by multimodal retrieval, and image matching is augmented by image to multimodal retrieval. To that end, we created an evaluation dataset with a query set and a index set. The query set consists of 8090 unique products, 46,096 in-the-wild product query images and a total of 46,096 image-text pairs. Query strings were composed with product attributes excluding fields included in the index. On the index side, we sourced 1M distractor products from the Amazon catalog in addition to that of the query dataset. The index consists of catalog images and product titles, enabling us to replicate how products are retrieved for customers in reality. For details of the dataset please refer to Table 2.

4.1.3 Comparison Methods. For both evaluation tasks (i.e., multimodal retrieval and image-multimodal retrieval), the indices were

Name	#Products	#Query Images	#Samples
100M triples	23M	100M	100M
400M quadruples	13M	57M	400M

Table 1: Training dataset details for 3-tower and 4-tower model training.

Set Name	#Unique Products	#Image-Text Pairs
Query	8090	46,096
Index	1,008,090	1,008,090

Table 2: Statistics of evaluation dataset.

set up as multimodal, with each entry represented by the average of catalog image and product title embeddings.

For multimodal retrieval, queries were defined as weighted sum of query image and query text embeddings. For image to multimodal retrieval, on the other hand, query image embeddings were utilized as the query vectors. During evaluation, image and text embeddings were first extracted for both query and index sets. The catalog image and product title embeddings were then utilized to construct a K-nearest neighbor index. With the index ready, we iterated through the query set to retrieve top-10 products from the index and measured recall at 1st, 5th and 10th position. Such iteration is performed multiple times in a grid-search fashion to find the best image-text weight pair for the model. After the grid search, the best recall performance of multimodal retrieval (non-zero text-weight) and performance of image to multimodal retrieval (zero text-weight) were recorded.

4.1.4 Evaluation Metrics. We benchmarked our methods by evaluating recall performance at 1st, 5th and 10th position. We define correctness as retrieving the ground-truth product of the corresponding query image-text pair.

4.2 Training Procedure

4.2.1 3-tower Model Training. We trained our 3-tower model with 128 NVIDIA A100 Tensor Core GPUs, with a total batch size of 32,768. We set learning rate as 5e-4 with 600 steps of warm up and set weight decay as 0.2. We used a 1B ViT-g/14 as our vision encoder and a 354M BERT as our text encoder, which were pretrained on 2B Amazon catalog image and title pairs. We trained the model with 16 epochs on the 100M dataset which is 1.6B samples seen in total. Images are randomly cropped and resized to 224^2 without keeping the aspect ratio and text are padded to 77 tokens.

4.2.2 4-tower Model Training. We followed the same training setup as the 3-tower model training in terms of number of GPUs, batch size and warm up process. Considering that this is a fine-tuning process, we reduced the learning rate to 7.5e-5 when training with 4-tower models. We trained the model for 240k iterations, effectively observing 800M quadruples. During 4-tower model training, we padded the images to square and resized to width of 224 for keeping image aspect ratios. We also increased the maximum number of text tokens to 128. The best model was obtained by fine-tuning the best 3-tower model with 4-tower losses.

4.2.3 Computational Cost. We see 73ms and 21ms (at 90 percentile) inference latency for extracting features with the image encoder and text encoder respectively on a G5.xlarge instance from AWS.

4.3 Method Benchmarking

To measure the effectiveness of our model, we compared our 3-tower and 4-tower alignment training scheme with public metric learning paradigms, including methods that operate on both unimodal and multimodal input sources. For comparison with unimodal models, we compared 3-tower model training with a ViT trained with pure Contrastive Loss [7]. For VL pre-training, we compared our method with the publicly available CLIP model and the CLIP model trained from scratch on Amazon catalog data. In the interest of fairness, we compared different methods with the same image backbones.

4.4 Comparing 3-tower and 4-tower Architectrue

With the vision encoder set to ViT-B/16, text encoder set to BERT with 64M parameters and training dataset set to 400M quadruples, the 4-tower model outperforms the 3-tower model in image to multimodal retrieval (see Table 3, row 9 and 10) and multimodal retrieval (Table 3, row 18 and 19). We also observed improved scaling behavior with the 4-tower model with regards to the amount of compute, see figure 6. With the same vision backbone, a 4-tower model trained with 400M quadruples still outperforms a 3-tower model trained with 100M triples on both image to multimodal retrieval (Table 3, row 8 and 10) and multimodal retrieval (Table 3, row 17 and 19). This demonstrates the effectiveness of the 4-tower training strategy considering the fact that 400M quadruples dataset contains only 56% of products and query images of the 100M triples dataset.

We also compared 3-tower and 4-tower model training by using ViT-g/14 as vision encoder and BERT with 354M parameters as text encoder. Compared to a 3-tower model trained on 100M triples, the 4-tower model trained with 400M quadruples achieved superior image to multimodal retrieval performance (Table 3, row 12, 13) despite being trained on a smaller dataset. The 4-tower model also outperformed the 3-tower model in multimodal retrieval (Table 3, row 21, 22).

To fully leverage the increased query image and product diversity in 100M triples dataset, we obtained our best 4-tower model by fine-tuning a 100M triples pre-trained 3-tower model with 4-tower training setup. As shown in Table 3, row 14 and 23, the fine-tuned model outperforms every other method in both retrieval evaluation tasks.

4.5 Image Only versus Multi-modal Contrastive Learning

In order to understand the impact of incorporating text data in contrastive training, we conducted fine-tuning on the ViT-B/16 CLIP model. This fine-tuning involved using an image-only contrastive loss for one set and the 3-tower image-text contrastive losses for

Row	Method Name	Vision Encoder	Text Encoder	Initialization	Training Set	Recall@1	Recall@5	Recall@10
				Image to Image Retrieval				
1	Contrastive	ViT-B/16	n/a	random	100M triples	0.18	0.27	0.31
2	3-Tower Model	ViT-B/16	BERT(64M)	random CLIP	100M triples	0.19	0.30	0.35
3	3-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	100M triples	**0.30**	**0.45**	**0.51**
4	CLIP (OpenAI)	ViT-B/16	BERT(63M)	n/a	n/a	0.06	0.10	0.12
5	Amazon CLIP	ViT-B/16	BERT(64M)	n/a	Amazon Catalog	0.04	0.07	0.09
				Image to Multimodal Retrieval				
6	CLIP (OpenAI)	ViT-B/16	BERT(63M)	n/a	n/a	0.07	0.13	0.15
7	Amazon CLIP	ViT-B/16	BERT(64M)	n/a	Amazon Catalog	0.07	0.13	0.15
8	3-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	100M triples	0.37	0.51	0.55
9	3-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	400M quadruples	0.34	0.52	0.58
10	4-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	400M quadruples	**0.38**	**0.54**	**0.58**
11	Amazon CLIP	ViT-g/14	BERT(354M)	n/a	Amazon Catalog	0.15	0.24	0.27
12	3-Tower Model	ViT-g/14	BERT(354M)	Amazon CLIP	100M triples	0.49	0.65	0.70
13	4-Tower Model	ViT-g/14	BERT(354M)	Amazon CLIP	400M quadruples	0.49	0.67	0.72
14	4-Tower Model	ViT-g/14	BERT(354M)	3-Tower Model	400M quadruples	**0.54**	**0.74**	**0.79**
				Multimodal Retrieval (Multimodal Search)				
15	CLIP (OpenAI)	ViT-B/16	n/a	n/a	n/a	0.09	0.15	0.18
16	Amazon CLIP	ViT-B/16	BERT(64M)	n/a	Amazon Catalog	0.10	0.18	0.21
17	3-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	100M triples	0.38	0.52	0.57
18	3-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	400M quadruples	0.37	0.55	0.61
19	4-Tower Model	ViT-B/16	BERT(64M)	Amazon CLIP	400M quadruples	**0.57**	**0.74**	**0.79**
20	Amazon CLIP	ViT-g/14	BERT(354M)	n/a	Amazon Catalog	0.25	0.41	0.47
21	3-Tower Model	ViT-g/14	BERT(354M)	Amazon CLIP	100M triples	0.53	0.68	0.72
22	4-Tower Model	ViT-g/14	BERT(354M)	Amazon CLIP	400M quadruples	0.61	0.78	0.82
23	4-Tower Model	ViT-g/14	BERT(354M)	3-Tower Model	400M quadruples	**0.64**	**0.82**	**0.86**

Table 3: Benchmarking of different retrieval sub-tasks. For image to image retrieval, leveraging vision-language pre-training improved recall at all positions (row 2). Positive scaling behavior was observed of the 3-tower and the 4-tower model w.r.t. model size, where utilizing ViT-g/14 and BERT(354M) significantly improves recall performance. Results also indicate that 4-tower model outperforms competing methods on both image to multimodal retrieval and multimodal retrieval. Initializing the 4-tower model with 3-tower model weights (pre-trained on 100M triples) led to the best performing model (row 11, 20).

the other. The results, presented in Table 3, reveal performance improvements when 3-tower training is applied. Specifically, rows 1 and 2 of the table show that fine-tuning the ViT-B/16-based CLIP model with text data results in increases of 1% in Recall@1, 3% in Recall@5, and 4% in Recall@10.

4.6 Impact of Multimodal Index

Our experiments show that incorporating text data in contrastive training improves model performance. This led to the hypothesis that integrating text information in the index would similarly enhance image retrieval performance. To test this hypothesis, we conducted an evaluation comparing recall metrics using an image-only index versus a multimodal index. This evaluation was performed on both VL (VL) models and multimodal models. The results, as depicted in Table 3, show notable improvements when text data is added during index construction. Specifically, for the out-of-the-box CLIP model (row 4, 6) and Amazon-CLIP model (row 5, 7), the addition of text data to the index resulted in increases of 1.2%, 2.5%, and 3.1%, and 3.4%, 5.2%, and 5.9% in recall@1, recall@5, and recall@10, respectively. Furthermore, for 3-tower model, the inclusion of text in index building showed even more improvements (row 3, 8) of 7% recall@1, 6% recall@5 and 4% recall@10. Fig. 5 shows one example of the comparison.

4.7 Online A/B Testing Results

To understand how our models affect customer experience, we performed online A/B testing with our models powering image match and multimodal search functionalities. We considered click-through rate (CTR) as the main criterion, based on the assumption that customers are more likely to click on high-quality retrieved products. We anticipated that improvements in retrieval quality due to new models will therefore be reflected in CTR trends.

As we developed the models progressively, we first conducted online testing for the 3-tower model. In this experiment, we focused on CTR improvement for image matching and monitored

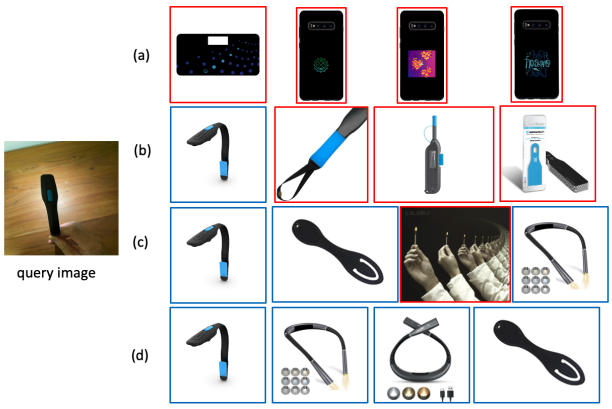

Figure 5: Comparing street-to-shop retrieval from 4 representative models: (a) - (d) show results from Row 1, 2, 12, 13 respectively. Red boxes mean wrong results. Blue boxes represent exact/similar matches to the query image. The model trained with pure image-to-image match loss tends to match partial visual patterns while models trained with additional vision-text alignment loss can find the same or similar products from the same product category. The larger the model, the better the performance. We use our evaluation index to do the comparison, containing less than 1 percent of products compared to our online index, which makes exact matches hard.

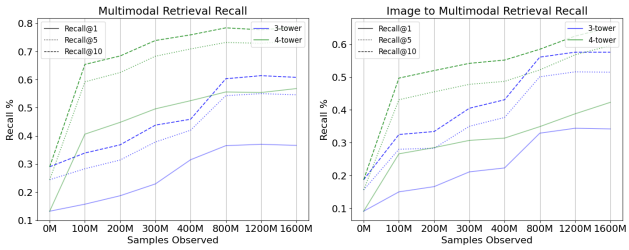

Figure 6: Comparing 3-tower and 4-tower model scaling behavior w.r.t. compute.

how customers interact with the new multimodal search functionality. We observed 4.95% relative image match CTR improvement, which demonstrated the 3-tower model to be effective in improving image matching quality. We also performed online testing for the 4-tower model, where we observed a further 1.13% image match CTR improvement and 1.35% multimodal search CTR improvement. Although the 4-tower model has less image match CTR impact compared to the 3-tower model, we believe this is because the 3-tower model had already significantly improved image matching result quality.

4.8 Ablation Study

We performed additional experiments to understand how different scaling factors affects model learning (i.e., model size, dataset size). These studies were based on 4-tower architecture, pre-trained

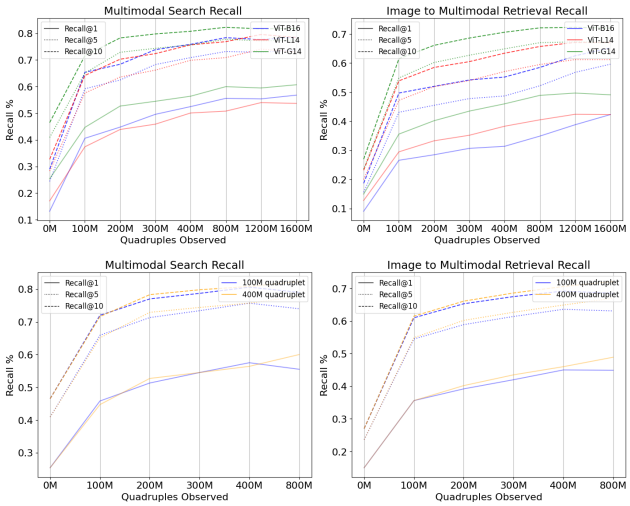

Figure 7: Model scaling behavior with different model sizes and dataset sizes. Top two images: Scaling pattern of 4-tower model with different encoder sizes. Bottom two images: Scaling pattern of 4-tower model with different training dataset sizes.

with Amazon Catalog data. All experiments were trained with 128 NVIDIA A100 TensorCore GPUs, batch size of 32k and learning rate of 7.5e-5. We utilized a 400M quadruples dataset in these experiments, following the evaluation protocol described in section 4.1.2.

4.8.1 Impact of Model Size. To observe the impact of model size, we trained a 4-tower architecture at 3 different scales, illustrated in Table 4. Each model was trained for 4 full epochs, effectively observing 1600M quadruples. Our experiment results indicated that the ViT-g/14 combined with BERT (354M) achieved the best performance on both image to multimodal retrieval and multimodal retrieval.

Vision Enc	Text Enc	Recall@1	Recall@5	Recall@10
Image to Multi-modal Retrieval				
ViT-B/16	BERT(64M)	0.42	0.60	0.65
ViT-L/14	BERT(123M)	0.42	0.62	0.67
ViT-g/14	BERT(354M)	**0.49**	**0.67**	**0.72**
Multi-modal Retrieval				
ViT-B/16	BERT(64M)	0.57	0.74	0.78
ViT-L/14	BERT(123M)	0.54	0.73	0.79
ViT-g/14	BERT(354M)	**0.61**	**0.78**	**0.82**

Table 4: Impact of model scale has on 4-tower model training. Using ViT-g/14 and BERT (354M) led to the best performance, indicating the scalability of our method.

4.8.2 Impact of Dataset Size. We also measured how dataset size impacts model performance. To that end, we subsampled a 400M

Dataset Size	Recall@1	Recall@5	Recall@10
Image to Multi-modal Retrieval			
100M quadruples	0.45	0.64	0.69
400M quadruples	**0.6**	**0.78**	**0.82**
Multi-modal Retrieval			
100M quadruples	0.49	0.67	0.72
400M quadruples	**0.57**	**0.76**	**0.81**

Table 5: Impact of dataset size on 4-tower model training. All experiments use ViT-g/14 as image encoder and BERT (354M) as text encoder. Training with 400M quadruples led to better retrieval performance across the board.

quadruples dataset and created a 100M quadruples subset. We trained a 4-tower model with the 100M quadruples subset for 8 epochs and compared it with the same model trained with the full 400M quadruples dataset. In this experiment we used ViT-g/14 as the image encoder and BERT (354M) as text encoder. We found that in early iterations the two models scaled similarity in terms of retrieval performance. However, with more compute, the model trained with 400M quadruples dataset clearly outperforms its counterpart as illustrated in Fig. 7.

5 CONCLUSION

In this paper, we propose a new algorithm named MIM for Amazon scale street-to-shop retrieval problem. MIM can improve performance on the street-to-shop retrieval problem by leveraging text information for visual semantic matching. We also develop a multimodal search service to further improve search quality and give users the flexibility to refine search results. Both offline and online experiments verify the effectiveness of MIM models and the multimodal search service. It shows that image-text alignment is beneficial to the street-to-shop problem as it can learn the product concepts well and avoid false positives due to matching low-level visual features. In the future, we plan to resolve some limitations of the proposed method such as: 1) the vision model may prioritize retrieving approximate matches over exact matches for some cases; 2) in Multimodal Search still underperforms on cases where lexical match is more important than semantic understanding (e.g., size/part numbers).

REFERENCES

[1] X. An, J. Deng, K. Yang, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu. Unicom: Universal and compact representation learning for image retrieval. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

[2] J. Beal, H.-Y. Wu, D. H. Park, A. Zhai, and D. Kislyuk. Billion-scale pretraining with vision transformers for multi-task visual representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 564–573, 2022.

[3] S. Bell, Y. Liu, S. Alsheikh, Y. Tang, E. Pizzi, M. Henning, K. Singh, O. Parkhi, and F. Borisyuk. Groknet: Unified computer vision model trunk and embeddings for commerce. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2608–2616, 2020.

[4] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[5] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[6] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning, 2023.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 2019, page 4171, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[10] M. Du, A. Ramisa, A. K. KC, S. Chanda, M. Wang, N. Rajesh, S. Li, Y. Hu, T. Zhou, N. Lakshminarayana, S. Tran, and D. Gray. Amazon shop the look: A visual search system for fashion and home. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2822–2830, 2022.

[11] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.

[12] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

[13] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[15] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015.

[16] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[17] S. Kim, D. Kim, M. Cho, and S. Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247, 2020.

[18] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[19] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[21] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[22] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He. Scaling language-image pre-training via masking, 2023.

[23] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2022.

[24] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[25] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[26] D. Manandhar, M. Bastan, and K.-H. Yap. Dynamically modulated deep metric learning for visual search. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2408–2412. IEEE, 2020.

[27] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017.

[28] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[31] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[32] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[33] I. Tautkute, T. Trzciński, A. P. Skorupa, Ł. Brocki, and K. Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019.

[34] E. W. Teh, T. DeVries, and G. W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *European Conference on Computer Vision (ECCV)*, pages 448–464. Springer, 2020.

[35] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[36] M. Tschannen, M. Kumar, A. Steiner, X. Zhai, N. Houlsby, and L. Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021.

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[39] C. Wang, W. Zheng, J. Li, J. Zhou, and J. Lu. Deep factorized metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2023.

[40] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3127–3135, 2018.

[41] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[42] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[43] F. Yang, A. Kale, Y. Bubnov, L. Stein, Q. Wang, H. Kiapour, and R. Piramuthu. Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2101–2110, 2017.

[44] Z. Yang, M. Bastan, X. Zhu, D. Gray, and D. Samaras. Hierarchical proxy-based loss for deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868, 2022.

[45] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, Aug 2022, 2022.

[46] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, and P. Zhang. Florence: A new foundation model for computer vision, 2021.

[47] A. Zhai and H.-Y. Wu. Classification is a strong baseline for deep metric learning. *British Machine Vision Conference (BMVC)*, 2019.

[48] A. Zhai, H.-Y. Wu, E. Tzeng, D. H. Park, and C. Rosenberg. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2412–2420, 2019.

[49] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[50] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[51] K. Zhao, P. Pan, Y. Zheng, Y. Zhang, C. Wang, Y. Zhang, Y. Xu, and R. Jin. Large-scale visual search with binary distributed graph at alibaba. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2567–2575, 2019.