Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining

Xunlin Zhan[1†], Yangxin Wu[1†], Xiao Dong[1], Yunchao Wei[2], Minlong Lu[3], Yichi Zhang[3], Hang Xu[4], and Xiaodan Liang[1⋆]

[1]Sun Yat-sen University, [2]Beijing Jiaotong University, [3]Alibaba Group, [4]Huawei Noah's Ark Lab

{*zhanxlin, wuyx29*}*@mail2.sysu.edu.cn*, {*dx.icandoit,wychao1987,chromexbjxh,xdliang328*}*@gmail.com, ymlml@zju.edu.cn, yichi.zyc@alibaba-inc.com*

Abstract

Nowadays, customer's demands for E-commerce are more diversified, which introduces more complications to the product retrieval industry. Previous methods are either subject to single-modal input or perform supervised image-level product retrieval, thus fail to accommodate real-life scenarios where enormous weakly annotated multi-modal data are present. In this paper, we investigate a more realistic setting that aims to perform weakly-supervised multi-modal instance-level product retrieval among fine-grained product categories. To promote the study of this challenging task, we contribute Product1M, one of the largest multi-modal cosmetic datasets for real-world instance-level retrieval. Notably, Product1M contains over 1 million image-caption pairs and consists of two sample types, i.e., single-product and multi-product samples, which encompass a wide variety of cosmetics brands. In addition to the great diversity, Product1M enjoys several appealing characteristics including fine-grained categories, complex combinations, and fuzzy correspondence that well mimic the real-world scenes. Moreover, we propose a novel model named Cross-modal contrAstive Product Transformer for instance-level prodUct REtrieval (CAPTURE), that excels in capturing the potential synergy between multi-modal inputs via a hybrid-stream transformer in a self-supervised manner. CAPTURE generates discriminative instance features via masked multi-modal learning as well as cross-modal contrastive pretraining and it outperforms several SOTA cross-modal baselines. Extensive ablation studies well demonstrate the effectiveness and the generalization capacity of our model. Dataset and codes are available at https://github.com/zhanxlin/Product1M.

1. Introduction

(a) Image-level Retrieval

(b) Multi-modal Instance-level Retrieval

Figure 1. Our proposed task performs **instance-level** retrieval among **multi-modal** data.

The past two decades have witnessed the high enrichment of the commodity types and the diversification of online customer's demand in E-commerce. On the one hand, online merchandise has increasingly diversified categories and a large proportion of them are exhibited as a product portfolio where multiple instances of different products exist in one image. On the other hand, online customers or merchants may want to retrieve the single product in a portfolio for price comparison [42] or online commodity recommendation [34]. Furthermore, with the ever-accelerating accumulation of heterogeneous data generated by multimedia, it remains a problem how an algorithm can handle large-scale and weakly annotated data [45] to perform multi-modal retrieval.

In this paper, we explore a realistic problem: *how to perform instance-level*[1] *fine-grained product retrieval given the*

† Equal contribution. ⋆ Corresponding Author.

[1]Instance-level product retrieval refers to the retrieval of all single products existed in a product portfolio image.

Dataset	#samples	#categories	#instances	#obj/img	weak supervision	multi-modal	instance-level retrieval
RPC checkout [47]	30,000	200	367,935	12.26			
Twitter100k [17]	100,000	-	-	-	✓	✓	
INRIA-Websearch [22]	71,478	353	-	-	✓	✓	
Dress Retrieval [7]	20,200	-	∼20,200	∼1.0	✓	✓	
Product1M(Ours)	**1,182,083**	**458**	*92,200*	*2.83*	✓	✓	✓

Table 1. Comparisons between different datasets. '-' indicates inapplicable. The #instances and #obj/img of Product1M are in italics since there are no instance labels for the *train* set and we only count the instances in the *val* and *test* set. Product1M is one of the largest multi-modal datasets as well as the first dataset specifically tailored for real-world instance-level retrieval scenarios.

large-scale weakly annotated multi-modal data? We compare different paradigms of retrieval in Figure 1. As can be seen, image-level retrieval tends to return trivial results since it does not distinguish different instances, while multi-modal instance-level retrieval is more favorable for searching for various kinds of products among multi-modal data. Despite the generality and the practical value of this problem, it is not well studied due to the lack of real-world datasets and a clear problem definition. In the literature of product retrieval, intra-modal [32, 1, 31, 30] and cross-modal retrieval [43, 12, 48, 4, 44, 8] take as input single-modal information, e.g., an image or a piece of text, and performs matching search between separate data points. Unfortunately, such retrieval schemes significantly restrict their use in many scenarios where multi-modal information exists in both the queries and targets. More importantly, previous works focus on the relatively simple case, i.e., image-level [2] retrieval for single-product images [24, 13] and the instance-level nature of retrieval is unexplored.

To bridge this gap and advance the related research, we collect a large-scale dataset, named Product1M, proposed for multi-modal instance-level retrieval. Product1M contains over 1 million image-caption pairs and consists of two types of samples, i.e., single-product and multi-product samples. Each single-product sample belongs to a fine-grained category and the inter-category difference is subtle. The multi-product samples are of great diversity, resulting in complex combinations and fuzzy correspondence that well mimic the real-world scenarios. To the best of our knowledge, Product1M is one of the largest multi-modal datasets as well as the first dataset specifically tailored for real-world multi-modal instance-level retrieval scenarios.

In addition to the constructed dataset, we also propose a novel self-supervised training framework that extracts representative instance-level features from large-scale weakly annotated data. Specifically, we first train a multi-product detector from pseudo-labels by incorporating a simple yet effective data augmentation scheme. Then, CAPTURE is proposed to capture the potential synergy of images and texts via several pretext tasks. We showcase that some prevailing cross-modal pretraining methods [27, 25, 6, 38] might be flawed under the multi-instance setting due to the design defects in the network architecture or the inappropriate pretext task. In contrast, CAPTURE utilizes a hybrid-stream architecture that encodes data of different modalities separately and fuses them in a unified way, which is experimentally shown to be beneficial for our proposed task. Moreover, we introduce the cross-modal contrastive loss to enforce CAPTURE to reach alignment between image and texts, which avoids the mismatch issue incurred by the inappropriate pretext task.

Crucially, CAPTURE surpasses the SOTA cross-modal baselines in terms of all main metrics by a large margin. We further conduct extensive ablation experiments to demonstrate the generalization capacity of CAPTURE and explore several critical factors of our proposed task. We hope the proposed Product1M, CAPTURE, and solid baselines can help advance future research on real-world retrieval.

[2]Image-level product retrieval refers to recognizing a specific product instance in a single-product image.

2. Related Work

Intra- and Cross-Modal Retrieval. Intra-modal retrieval [32, 1] has been extensively studied in the keyword-based web document retrieval [11], content-based image retrieval [29], and product recommendation [19, 20]. Cross-modal retrieval [43, 12, 48, 4, 44, 8] emerges as a promising avenue for efficient indexing and searching among large-scale data with different modalities, and is widely used in search engines [2, 14], E-commerce [18, 7], to name a few. However, these approaches [30, 26, 7, 47, 46] are typically subject to single modal inputs, which makes them hard to apply to many real-world scenarios where multi-modal information exists in both the queries and targets.

WSOD: Weakly Supervised Object Detection. WSOD [39, 36, 50] reduces its excessive reliance on fine-grained labels by learning from cheaper or freely-available data. PCL [39] iteratively generates proposal clusters to facilitate the learning of instance classifiers. Pseudo labels generated from image labels [36] and unstructured textual descriptions like captions [50] are also beneficial for boosting the performance of WSOD. However, WSOD typically relies on a fixed-size collection of predefined classes and is not readily applicable to our proposed task where class labels are not available and categories can be updated dynamically.

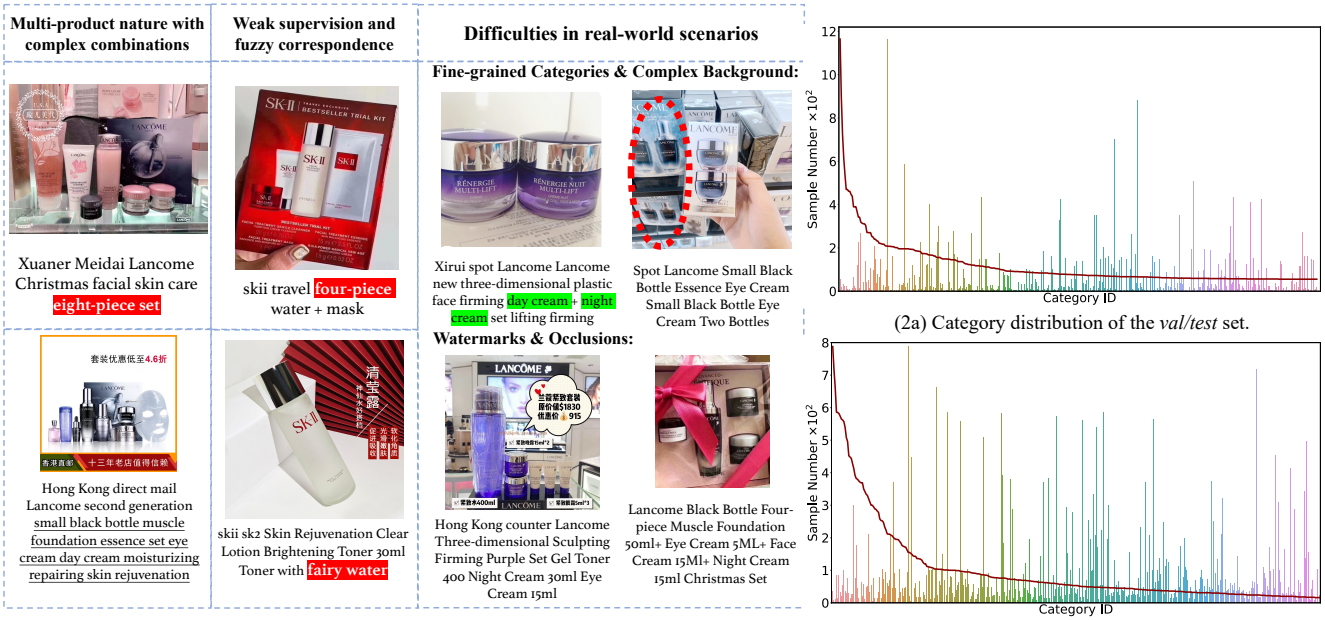


Figure 2. Characteristics and statistics of Product1M: (1a) Complex combinations of single-product; (1b) Weak supervision and fuzzy correspondence; (1c) Difficulties in real-world scenarios; (2) Long-tailed category distribution of Product1M. The line displays the sample number of each category in decreasing order. Product1M contains a wide variety of categories and the long-tailed class distribution aligns well with real-world scenarios.

Cross-Modal Self-Supervised Learning. Existing Vision-language pre-trained models typically use a multi-layer Transformer [41] architecture such as BERT [9] to learn image-text semantic alignment on multi-modal data. Single-stream models [25, 37, 6] encode the combined multi-modal features in a unified architecture while other two-stream models [27, 38] instead utilize different encoders for inputs of different modalities. These methods are not tailored for instance-level retrieval and we showcase that they might be flawed due to the design defects in the network architecture and the inappropriate pretext tasks.

3. Instance-Level Retrieval on Product1M

3.1. Task Definition

A product sample (I, C) is an image-text pair where I is the product image and C is the caption. Given the *gallery* set of single-product samples $\mathcal{S} = \{\mathcal{S}_i | \mathcal{S}_i = (I^i_{\mathcal{S}}, C^i_{\mathcal{S}})\}$ and the set of multi-product samples $\mathcal{P} = \{\mathcal{P}_i | \mathcal{P}_i = (I^i_{\mathcal{P}}, C^i_{\mathcal{P}})\}$, the task is to retrieve and rank the single-products that appear in the query sample $\mathcal{P}_i$, i.e., to predict a list $RETR^i = [id^i_1, id^i_2, \cdots, id^i_k, \cdots, id^i_N] \quad \forall \mathcal{P}_i \in \mathcal{P}$, where id^i_k corresponds to a specific single-product sample in $\mathcal{S}$.

3.2. Dataset Statistics

We collect a large number of product samples of 49 brands from E-commerce websites. These image-text samples are then manually divided into the single-product and multi-product groups according to the corresponding product information. Product1M is split into the *train*, *val*, *test*, and *gallery* set. The *train* set contains 1,132,830 samples including both the single-product and multi-product samples, while there are only multi-product samples in the *val* and *test* set, which contain 2,673 and 6,547 samples respectively. The *gallery* set has 40,033 single-product samples for 458 categories, 392 of which appear in the *val* and *test* set and the remaining ones act as interference items for validating the robustness of a retrieval algorithm. The samples in the *gallery*, *val*, and *test* sets are annotated with class labels for the purpose of evaluation, i.e., they are not involved in the training process, and the samples in the *train* set are not annotated. The statistics of Product1M are shown in Table 1 and Figure 2. More visualizations of Product1M and comparisons with related datasets can be found in the Supplementary Materials.

3.3. Dataset Characteristics

Multi-product nature and complex combinations: The multi-product images are ubiquitous on E-commerce websites and serve as the query images of instance-level product retrieval. As is shown in Figure 2(1a), products can be organized in abundant forms and layouts and the number of instances can be large. The excessive amount and great diversity of fine-grained single-product samples give rise to the complex combinations in different portfolio images.

Weak supervision and fuzzy correspondence: We consider using data of two common modalities, i.e., images and texts, for retrieval. Unlike other datasets with clean class la-

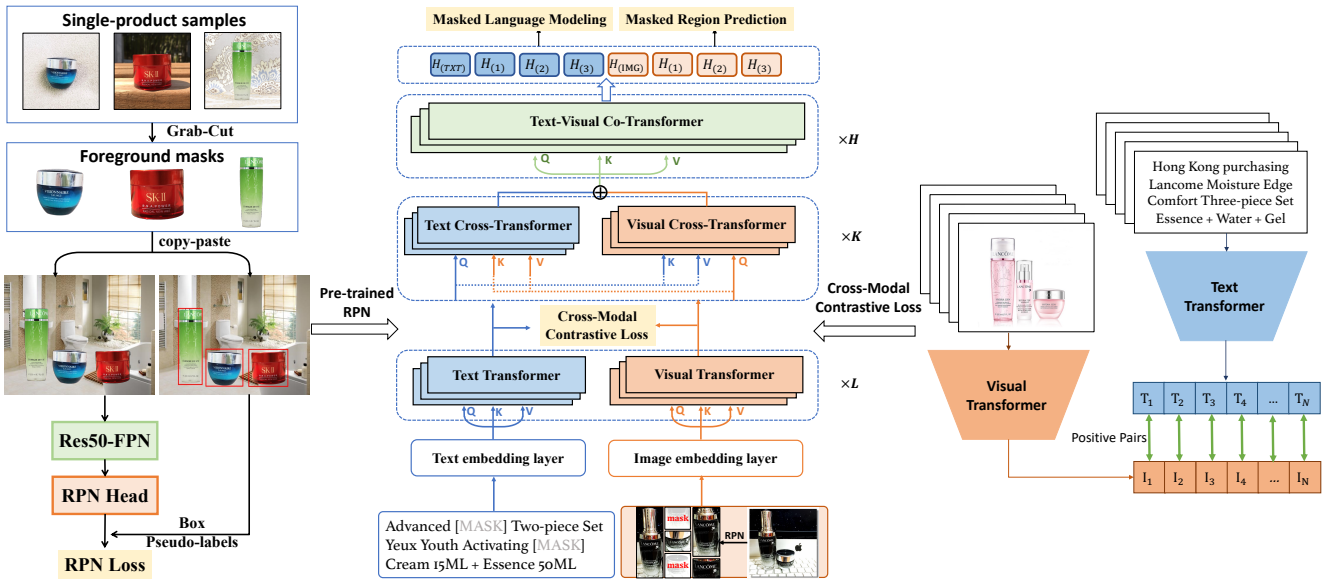


Figure 3. An overview of our instance-level retrieval pipeline. (a) Pretrain an RPN based on pseudo-labels generated by a copy-and-paste data augmentation scheme. (b) Utilize CAPTURE to capture the potential synergy across modality via a hybrid-stream architecture and several pretext tasks. (c) Construct positive pairs of matched image-text samples for cross-modal contrastive learning. Best viewed in color.

bels, the supervision from commodity captions is weak and often uninformative. We show different types of challenging samples in Figure 2(1b). Some samples contain abbreviations, i.e., a shortened form of several products, in their captions. However, the abbreviation like 'eight-piece set' does not contain any specific information about products. The second type of sample carries irrelevant information, where the commodities described in the title may not appear in the image or vice versa. The wide distribution of fuzzy correspondence between images and titles makes it even more challenging for instance-level retrieval.

Consistency with real-world scenarios: We show some challenging samples in Figure 2(1c). They can have a complex background with irrelevant objects, amorphous watermarks, or significant clutter covering the product information. Some products of different categories can have almost the same appearance except that the words on the packing are slightly different, e.g., *day cream* vs *night cream*. As is shown in Figure 2(2a,2b), the long-tailed distribution of Product1M aligns well with real-world scenarios.

4. Methodology

As is depicted in Figure 3, our framework consists of an augmentation-based detector and a self-supervised multi-modal transformer. In this section, we first elaborate the training process of RPN and the architectural design of CAPTURE in Section 4.1 and Section 4.2. Then we describe two kinds of pretext tasks that enables the self-supervised learning of CAPTURE in Section 4.3 and Section 4.4. Finally, we illustrate the inference process for instance-level retrieval in Section 4.5.

4.1. Training RPN for Multi-Product Detection

Retrieval based simply on the image-level features will lead to an undesirable condition where the retrieval results are overwhelmed by the dominated product in an image. Thus it is crucial to distinguish different products and extract proposal-wise features in a multi-product image. While many pre-trained detectors are available, they are infeasible to directly apply to multi-product detection due to the distribution difference between datasets. Thus we utilize a simple yet effective data augmentation scheme to train a Region Proposal Network (RPN) [35] based solely on the single-product images as shown in Figure 3(a). We first use GrabCut [28] to obtain the foreground masks of single-product images. With real-world background images from Places365 [51], a copy-and-paste augmentation [10] is applied to these foreground masks and background images to generate synthesized images. In this way, we are able to train a well-performing multi-product detector. Given the detected regions of RPN, we utilize RoIAlign [15] to obtain instance-wise features, which are then fed into CAPTURE for further cross-modal learning. More visualizations and details about the synthesized images and the training of RPN can be found in the Supplementary Materials.

4.2. Architectural Design of CAPTURE

After training the RPN, we can generate high-quality proposals for different products in an image. Different from the prevalent single-stream or two-stream transformer architectures, we propose CAPTURE that combines these two architectures into a unified one by stacking three types of layers for semantic alignment and joint learning of multi-

modal inputs. Details are shown in Figure 3(b). To be specific, the Text/Visual Transformer takes as input the embeddings of the texts or image and is responsible for intra-modal feature learning. The Text/Visual Cross-Transformer aims to capture and model the inter-modal relations between texts and image by exchanging key-value pairs in the multi-headed attention mechanism. After that, the features of texts and image are concatenated and serve as the query, key, and value inputs to the Co-Transformer for joint learning of multi-modal features. These three types of transformer are stacked L, K, and H times respectively. We verify the effectiveness of our architectural design in Table 4.

4.3. CAPTURE by Masked Multi-Modal Learning

We utilize several pretext tasks to enable the self-supervised learning of CAPTURE. For modality-wise feature learning, we adopt two masked multi-modal modeling tasks, i.e., Masked Language Modeling task (MLM) and Masked Region Prediction task (MRP), following the standard BERT [9] and VisualBERT [25]. Concretely, for MLM and MRP, approximately 15% of texts and proposal inputs are masked out and the remaining inputs are used to reconstruct the masked information. The MLM is handled as in BERT [9]. For the MRP, the model directly regresses the masked features, which is supervised by the features extracted by the pretrained RPN with a MSELoss. As for inter-modal relation modeling, Image-Text Matching task (ITM) is widely adopted in many previous methods [25, 6, 27, 38]. Typically, the model is asked to predict whether the text is the corresponding description of an image, which is formulated as a binary classification task. To generate negative samples, either the image or caption is randomly substituted. We argue that ITM could be problematic to the fine-grained understanding of an image-text sample at the instance-level. We hypothesize the deterioration stems from the unmatched image and caption pairs after substitution, which results in the inconsistency between detected regions and text. We further experimentally validate this claim in Table 3.

4.4. CAPTURE by Cross-Modal Contrastive Loss

Aside from intra-modal feature learning, CAPTURE is expected to generate coherent representations of multi-modal inputs and learn the correspondence between them. To this end, we resort to inter-modality contrastive learning [5, 33] to reach alignment between image and text. For a minibatch of N image-text samples, there are $2N$ data points in total. We treat the corresponding image-text pairs as N positive pairs, and the other $2(N-1)$ unmatched pairs are regarded as negative ones. Formally, given an image-text pair (x_i, x_j) and their encoded features $(\tilde{x}_i, \tilde{x}_j)$, the cross-modal contrastive loss for this positive pair is computed as:

$$\mathcal{L}(x_i, x_j) = -\log \frac{\exp\left(\text{sim}\left(\tilde{x}_i, \tilde{x}_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(\tilde{x}_i, \tilde{x}_k\right)/\tau\right)}, \tag{1}$$

where $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ computes the cosine similarity of $(\boldsymbol{u}, \boldsymbol{v})$ pairs, τ denotes the temperature parameter, $\mathbb{1}_{[k \neq i]}$ is a binary indicator function that returns 1 iff $k \neq i$. This form of contrastive loss encourages the encoded features of positive pairs from different modality to be similar while discriminates those of negative ones. We find it beneficial to inject this supervision at the Text/Visual Transformer and further discussion about the effect of cross-modal contrastive loss can be found in Section 5.3.

4.5. Inference for Instance-Level Retrieval

For both the single- and multi-product samples, the proposal-wise features extracted via the pre-trained RPN and the captions are used as input to CAPTURE. During inference, the Co-Transformer layer outputs H_{IMG} and H_{TXT} as the overall representations of visual and linguistic inputs, respectively. These two vectors are multiplied together to derive the joint representations of an instance. Furthermore, since Text/Visual Transformer is supervised with cross-modal contrastive loss, we find it beneficial to concatenate the features of this layer for retrieval. The resulting features then serve as the input of our retrieval algorithm. After computing the cosine similarity matrix between an instance and the samples in the $gallery$ set, we retrieve the corresponding single-product samples with the highest similarities for each query.

5. Experiments

5.1. Implementation Details

We attach RPN to a ResNet-50 [16] backbone pre-trained on ImageNet and follow the training schedule in [35]. We use the BERT [9] to initialize the linguistic transformer of our CAPTURE. The number of the Text/Visual Transformer, Text/Visual Cross-Transformer, and Co-Transformer is set to $L = 4, K = 4$, and $H = 4$, respectively, which adds up to 12 transformer layers. We set the hidden state size of CAPTURE and other baselines to 768 for a fair comparison. We separately attach a 512-d fully connected layer after Co-Transformer and Text/Visual Transformer for masked multi-modal learning and cross-modal contrastive learning. The concatenation of the features from these two layers results in a 1024-d feature vector for retrieval, which is also the same for other baselines. The maximum sequence length for the sentence is set to 36. We train CAPTURE with a total batch size of 128 for 10 epochs on 4 RTX 2080 GPUs. We use Adam [21] optimizer with an initial learning rate of 1e-4 and a linear learning rate decay schedule is adopted. Temperature parameter

Method	mAP@10	mAP@50	mAP@100	mAR@10	mAR@50	mAR@100	Prec@10	Prec@50	Prec@100
Image-based	40.35	36.77	34.76	17.20	15.86	15.45	32.80	30.54	29.97
Text-based	61.56	59.38	58.42	23.65	22.04	20.13	56.15	57.47	57.45
ViLBERT [27]	70.11	68.19	68.29	29.05	25.54	25.02	64.64	66.35	66.60
LXMERT [38]	71.37	67.83	66.73	29.83	23.15	23.89	65.97	64.79	64.77
CLIP* [33]	70.25	69.28	67.30	29.45	25.61	25.61	67.77	68.00	68.38
VL-BERT [37]	72.01	68.22	67.79	29.15	25.59	26.16	65.25	66.92	66.64
VisualBERT[25]	72.27	69.60	68.28	31.69	26.31	26.83	67.31	66.48	66.62
UNITER [6]	74.69	71.02	70.93	29.47	25.82	26.20	70.11	69.15	68.95
CAPTURE (Ours)	**79.36**	**74.79**	**74.63**	**34.69**	**30.04**	**30.08**	**73.97**	**72.12**	**73.86**

Table 2. Comparison with different intra- and cross-modal self-supervised baselines.

τ is set to 0.07. At inference, CAPTURE takes as input the texts and proposal-wise features to generate instance features. For a fair comparison with other baselines, we adopt the same training procedure and evaluation protocol in all experiments unless otherwise stated and we use the same augmentation-based RPN for the baselines in Table 2. More details can be found in the Supplementary Materials.

Evaluation Metrics. We adopt Precision ($\mathrm{Prec@}N$), mean Average Precision ($\mathrm{mAP@}N$) and mean Average Recall ($\mathrm{mAR@}N$) as our evaluation metrics, among which $\mathrm{Prec@}N$ and $\mathrm{mAP@}N$ are widely used in the retrieval literature [49, 3]. Since exhaustively retrieve every single product is unnecessary and impractical in many scenarios, we report mAP, mAR, and Prec for $N = 10, 50, 100$. The details of evaluation metrics can be found in the Supplementary Materials.

5.2. Weakly-Supervised Instance-Level Retrieval

We compare CAPTURE with several intra- and cross-modal baselines and the results are shown in Table 2.

Intra-Modal Schemes. We compare our method to two intra-modal schemes including *Image-based* and *Text-based* schemes. For image-based retrieval, we stack the Visual Transformer layer described in Section 4.2 and adopt the same image input and pretext task, i.e., Masked Region Prediction as CAPTURE. For text-based retrieval, we stack the Text Transformer layer and use only the text input and Masked Language Modeling pretext task. We further double the depth of these two models to 24 layers to keep the same amount of parameters as CAPTURE. It turns out that these two schemes are lagging far behind since they are subject to data of single modality, which suggests that the modeling of the relations between multi-modal data is indispensable. We provide more experiment results to validate this point in Section 5.4.

Cross-Modal Schemes. We compare CAPTURE to several prevailing self-supervised cross-modal pretraining methods in Table 2, including SOTA single-stream and two-stream Vision-language models as well as a SOTA zero-shot classification model, i.e., CLIP [33]. The CLIP* baseline refers to a CLIP-like architecture that uses separate transformers to encode image and text and is trained with a contrastive objective. Notably, CAPTURE outperforms all these baselines in all three metrics for instance-level retrieval. Two-stream models, i.e., ViLBERT [27], LXMERT [38] and CLIP*, are generally worse than single-stream ones, which suggests that the fusion mode of multi-modal features is one of the critical factors. We attribute the superior performance of CAPTURE to its hybrid-stream architecture and we study the impact of different layer types in Section 5.4.

#	Masked	ITM	CTR	Concat	mAP/mAR/Prec
1	✓				72.1 / 28.9 / 72.7
2	✓			✓	71.9 / 28.5 / 72.9
3	✓	✓			70.2 / 27.1 / 70.2
4	✓		✓		73.3 / 29.1 / 73.2
5	✓		✓	✓	**74.6 / 30.1 / 73.9**

Table 3. The impact of different pretext tasks and cross-modal contrastive loss. Evaluation for $N = 100$. 'Masked' stands for two masked multi-modal pretext tasks, i.e., MLM and MRP. 'CTR' stands for cross-modal contrastive loss.

5.3. Impact of Pretext Tasks and Contrastive Loss

As shown in Table 3, ITM will hurt the accuracy of instance-level retrieval (#1 vs #3), since it gives rise to mismatch samples, which might be detrimental to the fine-grained understanding of a multi-product image. We apply the cross-modal contrastive loss at the Text/Visual Transformer layer to align the representations of image and text, which further benefits the learning of consequent layers. The inclusion of contrastive loss encourages our model to maximizes the feature similarity of positive pairs, which improves all three metrics by 1.2, 0.2, and 0.5, respectively (#1 vs #4), and we find it of little help when added to the deeper layers. Moreover, after concatenating the features from the Text/Visual Transformer with that from the Co-Transformer for retrieval, it further improves all three metrics by 1.3, 1.0, and 0.7, respectively (#4 vs #5). However, we find this concatenation operation will slightly degrade

the performance of the model without contrastive loss (#1 vs #2), which suggests that the improvement mainly comes from the contrastive loss instead of the operation itself.

Model	Config	Depth	mAP/mAR/Prec
w/o-Cross	(6,0,6)	12	73.8 / 28.2 / 71.5
w/o-Co	(6,6,0)	12	73.2 / 29.3 / 72.8
w/o-Txt/Vis	(0,6,6)	12	69.3 / 25.4 / 68.4
CAPTURE-A	(2,5,5)	12	72.8 / 29.0 / 71.3
CAPTURE-B	(5,2,5)	12	73.7 / 29.1 / 71.8
CAPTURE-C	(5,5,2)	12	73.8 / 29.5 / 72.0
CAPTURE-S	(2,2,2)	6	67.7 / 25.7 / 68.3
CAPTURE-L	(8,8,8)	24	74.7 / 30.9 / 74.2
CAPTURE	**(4,4,4)**	**12**	**74.6 / 30.1 / 73.9**

Table 4. Performance of different layer configurations. Evaluation for $N = 100$.

5.4. Impact of Layer Configuration

We investigate how the configuration of transformer layers will affect the performance of our model in Table 4. The triplet in the Config column stands for the number of Text/Visual Transformer, Cross-Transformer, and Co-Transformer layer, respectively. We first remove layers of a specific type while keeping the depth of the resulting network the same as CAPTURE's, i.e., 12 layers, for a fair comparison. The 'w/o-Cross', 'w/o-Co', and 'w/o-Txt/Vis' refer to the resulting model after removing the Cross-Transformer, Co-Transformer, and Text/Visual Transformer layers from CAPTURE. As can be seen, the performances of these three models are inferior to that of CAPTURE, which demonstrates the effectiveness of its hybrid-stream architecture. Moreover, in the second group of Table 4 (CAPTURE-A,B,C), we study the combination of three layer types in different proportions. It turns out that the (4,4,4) configuration achieves the best performance. We further explore the performance of a smaller model (CAPTURE-S) and a larger model (CAPTURE-L). As can be seen, CAPTURE with the (4,4,4) configuration achieves a better trade-off between accuracy and parameters.

5.5. Zero-Shot Instance-Level Retrieval

We argue that a retrieval-based solution generalizes better to real-world scenarios where the category set is updated continuously and large quantities of clean labels are too costly to collect. Unlike detection, our retrieval-based framework does not rely on a fixed-size collection of predefined classes or fine-grained box annotations. To emphasize this, we conduct zero-shot retrieval experiments and report the results in Table 5. We manually remove 5/10/20 brands from the *train* set and train CAPTURE on the remaining samples so that the removed categories are not disposed to our model during training. Then we evaluate CAPTURE on the classes of these unseen brands. We further compare our model with a two-stream model LXMERT and a single-stream model UNITER. As can be seen, CAPTURE achieves better performance than LXMERT and UNITER for all three metrics, which well demonstrates its generalization capacity. We also visualize the embeddings generated by CAPTURE and UNITER via t-SNE [40] in Figure 5. It turns out that the features encoded by CAPTURE are more discriminative, which thus benefits the retrieval task.

	Metric	$N = 10$	$N = 50$
5 brands	mAP@N	63.3 / 64.5 / **67.1**	60.7 / 62.2 / **64.4**
	mAR@N	23.2 / 24.7 / **25.9**	19.2 / 20.1 / **20.9**
	Prec@N	56.5 / 57.1 / **60.5**	56.5 / 57.7 / **62.4**
10 brands	mAP@N	56.8 / 58.2 / **61.5**	54.1 / 55.2 / **57.1**
	mAR@N	19.6 / 20.5 / **24.0**	16.4 /17.4 / **18.4**
	Prec@N	50.2 / 51.9 / **53.0**	51.3 / 52.7 / **54.3**
20 brands	mAP@N	42.6 / 43.3 / **44.4**	36.7 / 37.3 / **38.8**
	mAR@N	17.5 / 17.6 / **17.9**	12.9 / 13.2 / **13.5**
	Prec@N	32.2 / 32.4 / **34.0**	32.9 / 33.1 / **34.5**

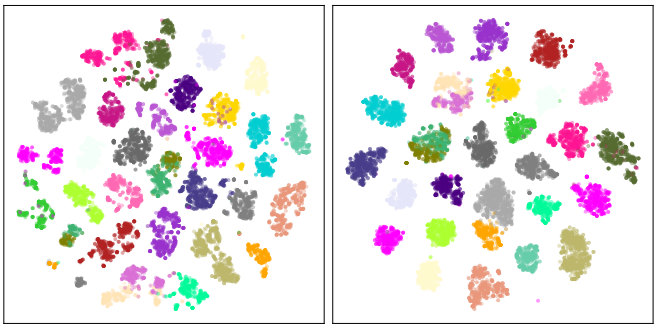
Table 5. Performance comparison of zero-shot retrieval. Organized in the order of LXMERT/UNITER/CAPTURE.

Method	mAP@100	mAR@100	Prec@100
UNITER-single	86.56	80.82	80.82
LXMERT-single	86.05	80.59	80.59
CAPTURE-single	**88.24**	**83.33**	**83.33**
CAPTURE-natural	70.36	26.46	66.53
CAPTURE-1Inst	60.03	20.43	58.42
CAPTURE	**74.63**	**30.08**	**73.86**
CAPTURE-subset	73.36	30.44	72.41
CAPTURE-gt	**77.79**	**37.40**	**77.13**

Table 6. Ablation study of single-product retrieval and the impact of detection performance on retrieval. Note that for single-product retrieval, the metric Prec@N is equivalent to mAR@N since there is only one category in an image.

5.6. Comparisons on Single-Product Retrieval

It is noteworthy that CAPTURE is applicable to both single-product and multi-product retrieval. Indeed, it excels on these two tasks and achieves better performance than other baselines in single-product retrieval. Specifically, for each single-product sample in the *gallery* set, we pick it out as a query and perform single-product retrieval among the remaining samples in the *gallery* set. We compare the performance of three models, i.e., UNITER-single, LXMERT-single and CAPTURE-single, in Table 6. As can be seen, the performance of single-product retrieval is much higher than that of multi-product retrieval since the difficulty is largely reduced when there is only one instance/entity in the image/text. Furthermore, we notice the



(a) UNITER (b) CAPTURE

Figure 4. Visualize the embeddings generated by CAPTURE and UNITER via t-SNE. Points belonging to the same category are of the same color. Best viewed in color.

performance of 'CAPTURE-single' is still better than that of two other baselines, which further demonstrates the superiority of CAPTURE.

5.7. Impact of Detection Performance on Retrieval

We conduct several experiments to explore how the performance of a detector will influence instance-level retrieval. The results are listed in Table 6. As we claim in Section 4.1, the off-the-shelf pretrained detectors are not readily applicable to our dataset due to the distribution difference between natural images and commodity images. To verify this, we replace the RPN with Faster R-CNN [35] pre-trained on Visual Genome [23] and utilize it to generate instance-wise input features of CAPTURE. The resulting model, named 'CAPTURE-natural', is inferior to CAPTURE in all three metrics. For the 'CAPTURE-1Inst' model, we feed the whole image and an image-level bounding box, which is of the same size as the image, to CAPTURE for inference. This scheme performs unsatisfactorily due to the failure of instance recognition, which suggests that the detector may become a performance bottleneck. Going a step further, to explore the upper bound of CAPTURE, we randomly select 1,338 multi-product images and manually label the bounding boxes of these images. For the 'CAPTURE-subset' model, we simply evaluate CAPTURE on this annotated subset. For the 'CAPTURE-gt' model, the ground truth boxes and their corresponding features serve as the input to CAPTURE. As can be seen, the performance gap of these two models suggests that the performance of a detector can play an essential role in instance-level retrieval. Moreover, the mAR gap between them is relatively large, which indicates that the false negatives in detection will hurt the performance of instance-level retrieval.

6. Conclusion

In this paper, we present the first effort on extending canonical intra-/cross-modal retrieval to a more generalized setting, i.e., weakly-supervised multi-modal instance-level product retrieval, which has wide application potential in the E-commerce industry. We contribute Product1M, which is one of the largest multi-modal retrieval datasets as well as the first one specifically tailored for instance-level retrieval. Aside from that, we propose a novel hybrid-stream transformer, named CAPTURE, that excels in capturing the potential synergy between data of different modalities. Moreover, we overcome the unmatched issue incurred by the inappropriate pretext task by enforcing cross-modal contrastive learning between multi-modal features. Extensive experiments demonstrate that our CAPTURE surpasses the SOTA cross-modal pretraining models in terms of all metrics by a large margin. We hope the proposed Product1M, CAPTURE, and solid baselines will spur further investigation into a more reliable and flexible retrieval system.

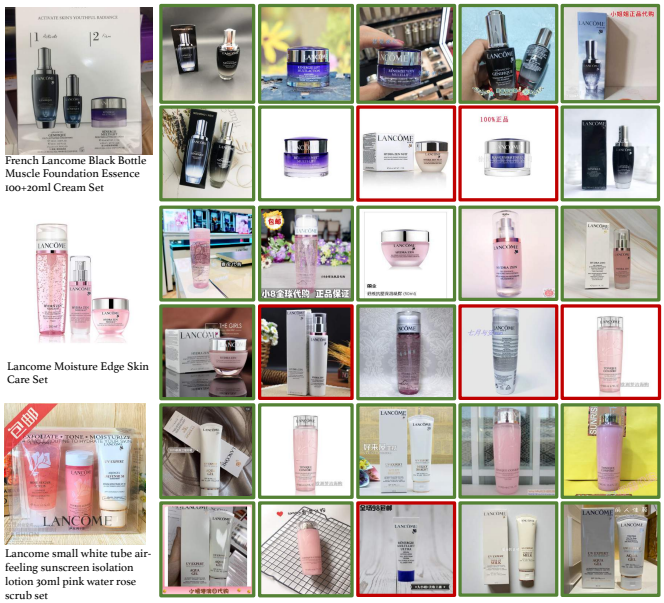


Figure 5. Visualizations of the retrieval results generated by CAPTURE. Multi-product query images are on the left. Correct/Incorrect retrieval images are highlighted in green/red boxes.

7. Acknowledgement

This work was supported in part by National Key RD Program of China under Grant No.2018AAA0100300, National Natural Science Foundation of China (NSFC) under Grant No.U19A2073 and No.61976233, Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) Grant No.2019B1515120039, Guangdong Outstanding Youth Fund (Grant No.2021B1515020061), Shenzhen Fundamental Research Program (Project No.RCYX20200714114642083, No.JCYJ20190807154211365).

References

[1] Cong Bai, Ling Huang, Xiang Pan, Jianwei Zheng, and Shengyong Chen. Optimization of deep convolutional neural network for large scale image retrieval. *Neurocomputing*, 303:60–67, 2018. 2

[2] Stefan Büttcher, Charles LA Clarke, and Gordon V Cormack. *Information retrieval: Implementing and evaluating search engines*. Mit Press, 2016. 2

[3] Deng Cai, Xiuye Gu, and Chaoqi Wang. A revisit of deep hashings for large-scale content based image retrieval. In *arXiv:1711.06016*, 2017. 6

[4] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454, 2016. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2, 3, 5, 6

[7] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2268–2274, 2017. 2

[8] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5

[10] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017. 4

[11] Faezeh Ensan and Ebrahim Bagheri. Document retrieval model through semantic linking. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 181–190, 2017. 2

[12] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014. 2

[13] Xiaoling Gu, Yongkang Wong, Lidan Shou, Pai Peng, Gang Chen, and Mohan S Kankanhalli. Multi-modal and multi-domain embedding learning for fashion retrieval and analysis. *IEEE Transactions on Multimedia*, 21(6):1524–1537, 2018. 2

[14] Donna Harman et al. Information retrieval: the early years. *Foundations and Trends® in Information Retrieval*, 13(5):425–577, 2019. 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[17] Yuting Hu, Liang Zheng, Yi Yang, and Yongfeng Huang. Twitter100k: A real-world dataset for weakly supervised cross-media retrieval. *IEEE Transactions on Multimedia*, 20(4):927–938, 2017. 2

[18] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1654–1662, 2017. 2

[19] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. Visually-aware fashion recommendation and design with generative image models. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 207–216. IEEE, 2017. 2

[20] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian McAuley. Complete the look: Scene-based complementary product recommendation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10532–10541, 2019. 2

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

[22] Josip Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web image search results using query-relative classifiers. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1094–1101, 2010. 2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 8

[24] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3066–3075, 2019. 2

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 3, 5, 6

[26] Zehang Lin, Zhenguo Yang, Feitao Huang, and Junhong Chen. Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2073–2077, 2018. 2

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations

for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2, 3, 5, 6

[28] Tang Meng, Lena Gorelick, Olga Veksler, and Yuri Boykov. Grabcut in one cut. In *IEEE International Conference on Computer Vision*, 2014. 4

[29] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2

[30] Petteri Nurmi, Eemil Lagerspetz, Wray Buntine, Patrik Floréen, and Joonas Kukkonen. Product retrieval for grocery stores. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 781–782, 2008. 2

[31] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, 2017. 2

[32] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2460–2464. IEEE, 2016. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 6

[34] Puji Rahayu, Dana Indra Sensuse, Betty Purwandari, Indra Budi, F Khalid, and N Zulkarnaim. A systematic review of recommender system for e-portfolio domain. In *Proceedings of the 5th International Conference on Information and Education Technology*, pages 21–26, 2017. 1

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4, 5, 8

[36] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative adversarial learning towards fast weakly supervised detection. In *IEEE/CVF Conference on Computer Vision Pattern Recognition*, 2018. 2

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 3, 6

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. 2, 3, 5, 6

[39] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018. 2

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[42] Sabine Vogler, Agnes Vitry, et al. Cancer drugs in 16 european countries, australia, and new zealand: a cross-country price comparison study. *The Lancet Oncology*, 17(1):39–47, 2016. 1

[43] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 2

[44] Kaiye Wang, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2010–2023, 2015. 2

[45] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016. 1

[46] Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. Effective deep learning-based multi-modal retrieval. *The VLDB Journal*, 25(1):79–101, 2016. 2

[47] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. Rpc: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019. 2

[48] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2):449–460, 2016. 2

[49] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020. 6

[50] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. 2019. 2

[51] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4