Supplementary Material: 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 Noahs Ark Lab

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

1. Related Datasets

RPC [7] contains 53,739 single-product exemplar images and 30,000 checkout images. Three different granularities of annotations are provided including category information, point-level annotations, and bounding boxes. However, there exists a gap between RPC and real-world scenarios since RPC was captured under a controlled environment and no textual information is available in this dataset. **Twitter100k** [3] is proposed for weakly supervised cross-media retrieval. It contains 100k image-text pairs crawled from Twitter. There is no constraint for the categories of images and the text is written in informal language by users. Similarly, this dataset focuses on image-level retrieval and use single-modal input as the query, which is not suitable for instance-level multi-modal retrieval.

INRIA-Websearch [4] is a cross-modal retrieval dataset that contains 71,478 image-text pairs with 353 different search queries, including actors, films, etc. The images are collected via internet search and textual information consists of text surrounding images on websites. Cross-modal retrieval algorithms are performed on this dataset to solve the text-to-image searching problem.

Dress Retrieval proposed in [1] collects noisy labeled data crawled on E-commerce website catalogs. Each fashion image is associated with some selected relevant fields filtered by predefined vocabulary. Typically, each image contains a single instance, e.g., a model wearing a dress and the background is relatively clean.

In comparison, our Product1M differs notably in four aspects from the above datasets:

- Product1M is the first dataset tailored for instance-level retrieval, which is of great potential in E-commerce industry.
- Product1M extends the canonical intra- and cross-modal retrieval to multi-modal retrieval, where images and texts exist both in the query and the target. This is a more practical setting since the multi-modal information is ubiquitous in real world scenarios.
- The weakly annotated samples of Product1M enforces the model to excavate useful features without clean labels, which endows the model with great generalization capacity to large-scale and noisy data.
- Product1M is one of the largest datasets for retrieval and the product samples are of great diversity and well aligns with E-commerce industry.

2. Single-Product Samples

We show some single-product images in Figure 1 and captions are not included for simplicity. The inter-category differences are subtle and some products of different categories have almost the same appearance except that the words on the packing are slightly different.

3. Multi-Product Samples

Figure 2 shows some multi-product images and captions are not included for simplicity. As can be seen, the combinations of single-product are complicated and noises like irrelevant objects, watermarks and complex background are widely existed.

4. RPN Training Details

We utilize a simple yet effective data augmentation scheme to train a Region Proposal Network (RPN) based

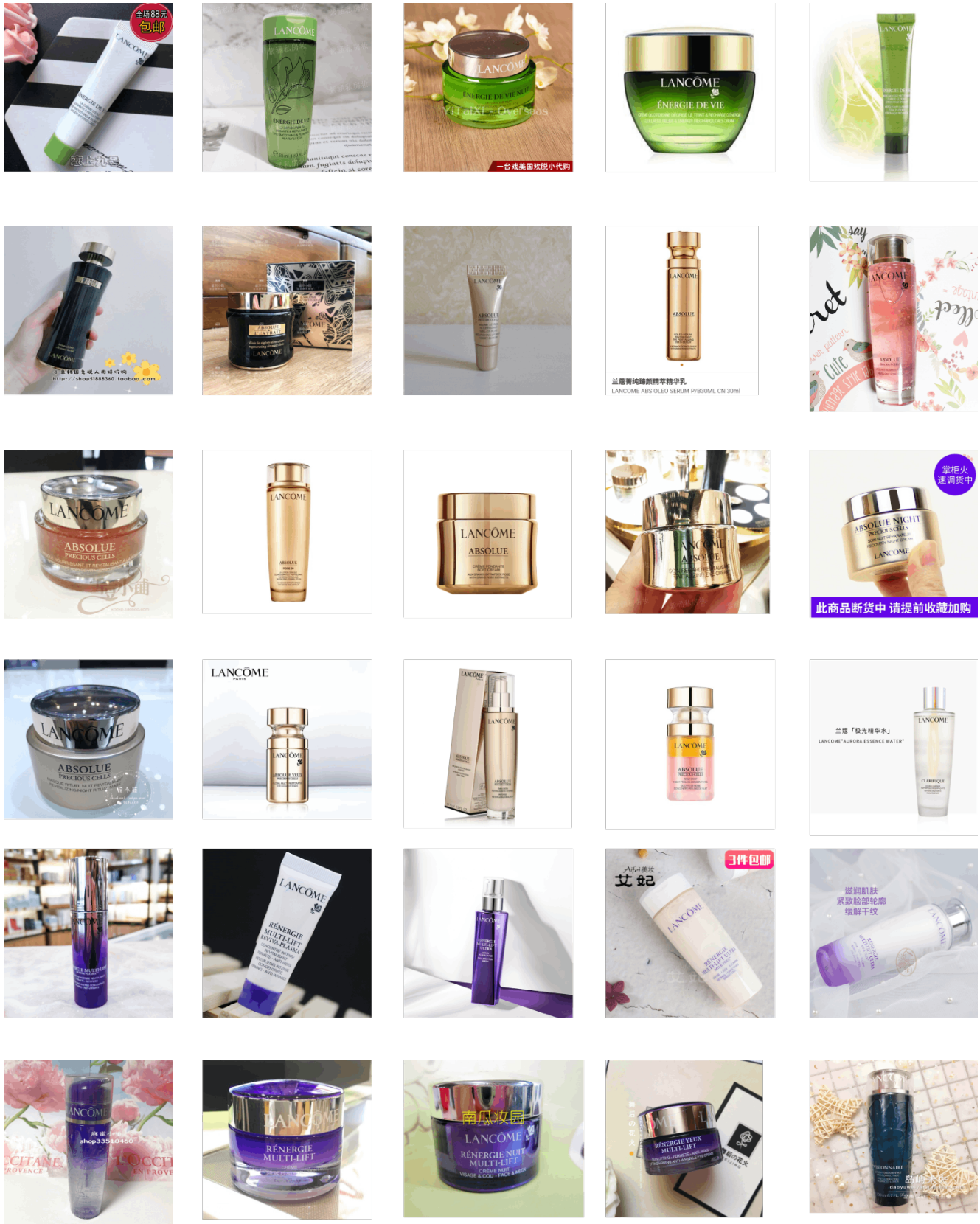

Figure 1. Single-product samples.

Figure 2. Multi-product samples.

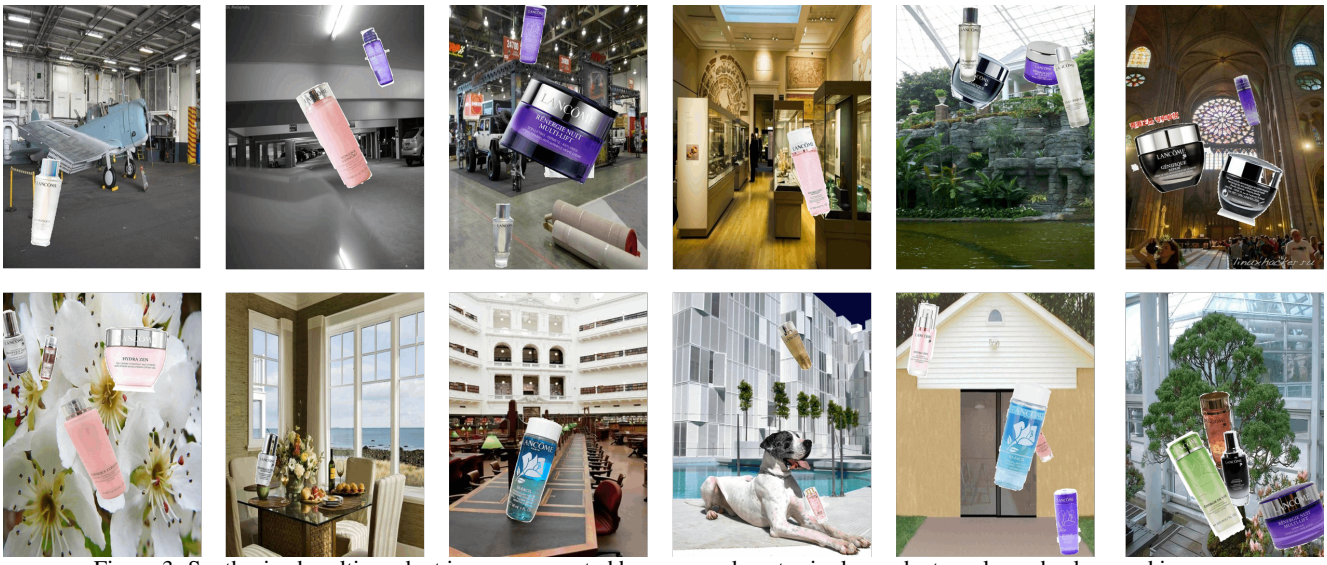

Figure 3. Synthesized multi-product images generated by copy-and-paste single-product masks on background images.

Figure 4. Visualizations of detection results of our RPN.

on the single-product images. We select about 5000 images with simple background from single-product images. Then GrabCut [6] method is adopted to extract the foreground masks of these single-product samples. These masks are used to generate the bounding box pseudo-labels and are pasted onto the background images from Place365 [8] following the copy-and-paste strategy in [2]. More than 40,000 synthetic images are generated in this step and Figure 3 showcases some synthesized images by copy-and-paste augmentation. We fix the parameters of backbone pre-trained on ImageNet and train RPN for 5 epochs on the synthesized images. During training, we adopt an SGD optimizer with a momentum of 0.9 and a weight decay of 1×10^{-4}. The mini-batch size is set to 16. The input images are resized into 11 scales ranging from 480 to 800 with a step size of 32. The detection results are visualized in Figure 4.

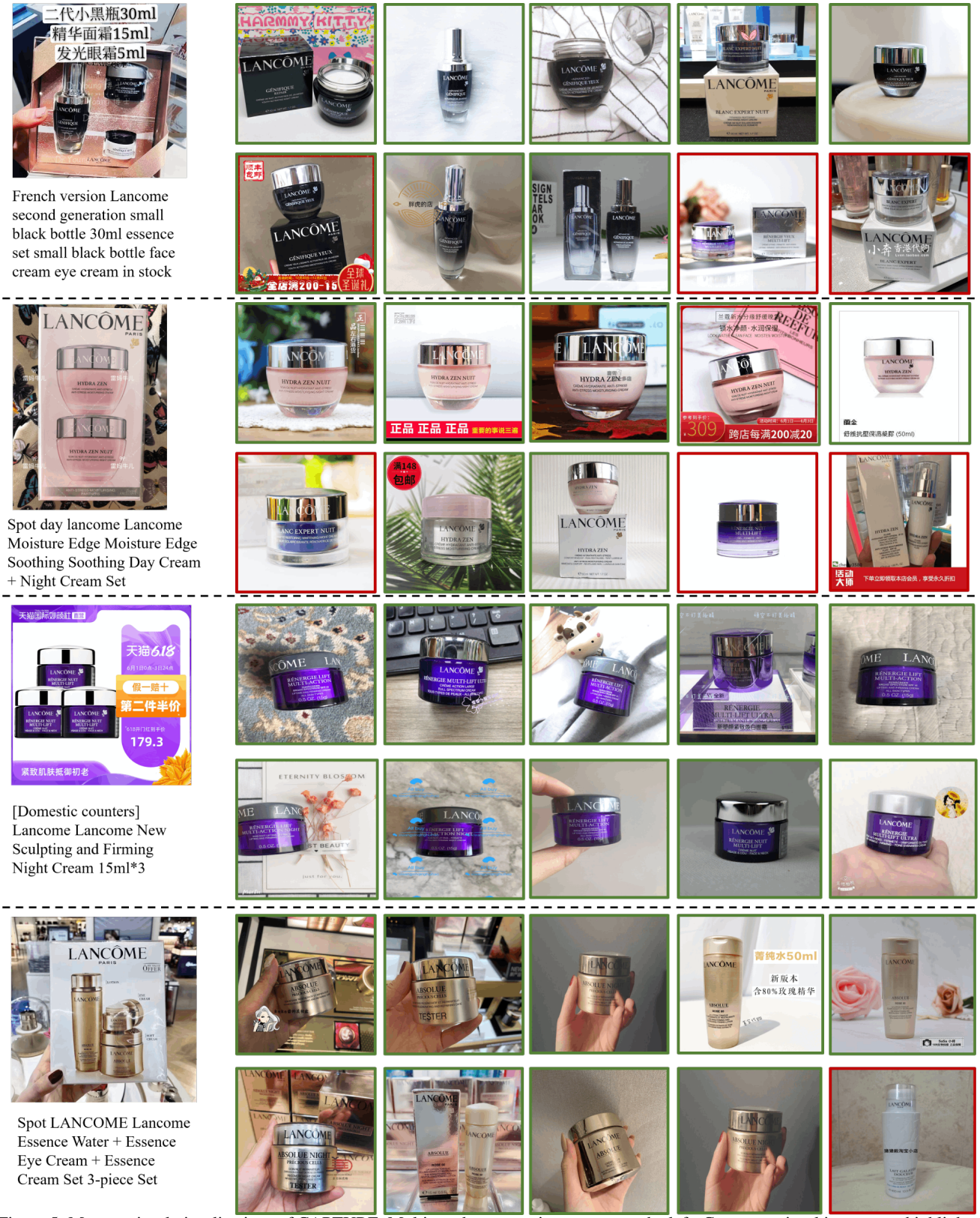

Figure 5. More retrieval visualizations of CAPTURE. Multi-product query images are on the left. Correct retrieval images are highlighted in green boxes while the incorrect ones are highlighted in red boxes. The captions of retrieved samples are omitted for simplicity.

5. Evaluation Metrics

We adopt Precision (Prec@N), mean Average Precision (mAP@N) and mean Average Recall (mAR@N) as our evaluation metrics. Prec@N evaluates the average accuracy of the top-N predictions per image and is widely used in the retrieval literature. Prec@$N(q)$ is computed as follows:

$$\text{Prec@}N(q) = \frac{1}{N}\sum_{i=1}^{N}\text{acc}_q(i), \tag{1}$$

where $\text{acc}_q(i)$ is a binary indicator function that returns 1 when the i-th prediction is correct for the q-th query and 0 otherwise. mAP@N is computed as the average AP@N per image, where AP@$N(q)$ is computed as follows:

$$\text{AP@}N(q) = \frac{1}{\min(m_q, N)}\sum_{k=1}^{N}\text{P}_q(k)\text{rel}_q(k), \tag{2}$$

where m_q is the total number of ground truth images, i.e., corresponding single-product images that appear in the q-th query image, $P_q(k)$ is the precision at rank k for the q-th query, and $\text{rel}_q(k)$ is a binary indicator function that returns 1 when the k-th prediction is correct for the q-th query and 0 otherwise. To evaluate the recall of instance-level retrieval results, we propose metric mAR@N, which can be computed as the average AR@N per image. AR@$N(q)$ is computed as follows:

$$\text{AR@}N(q) = \frac{1}{C_q}\sum_{c=1}^{C}\mathbb{1}_c(q)\min(1, \frac{RETR_c^q}{\min(\lfloor r_q^c \cdot N \rfloor, G_c)}) \tag{3}$$

where C is the total number of single-product categories in the *gallery* set, C_q equals to the number of existing categories in the q-th query, $\mathbb{1}_c(q)$ is a binary indicator function that returns 1 when class c exists in the q-th query, $RETR_c^q$ is the number of retrieved products belonging to class c for the q-th query, G^c is the number of ground truths belonging to class c in the *gallery* set, r_q^c is the instance ratio[1] of category c in the q-th query, and $\lfloor\cdot\rfloor$ is rounding operation. As per the equation, AR@$N(q)$ takes the category distribution into account, i.e., the inclusion of instance ratio is informative for evaluating both the correctness and diversity of a retrieval algorithm and guarantees that some trivial results are not overestimated[2].

6. More Retrieval Details of CAPTURE

We keep the number of input regions between 10 to 36 by selecting regions with predicted confidence higher than a threshold as in [5]. The features of region extracted by the RoIAlign are then flatten and fed into CAPTURE. The transformer blocks in CAPTURE have hidden state size of 768 with 8 attention heads. During retrieval, we compute cosine similarities between each box query with single-product samples and rank all retrieval single-product results according to their similarities. The top-N samples are returned as the retrieval results. Fig 5 shows more retrieval results by CAPTURE.

[1] For instance, for a 2A+3B query image, $r_q^A = 0.4$ and $r_q^B = 0.6$.

[2] For a 2A+3B query image and $N = 100$, AR@100(q) returns 0.51 and 1.0 for retrieval results 1A+99B and 40A+60B, respectively.

References

[1] 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. 1

[2] 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

[3] 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. 1

[4] 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. 1

[5] 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. 6

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

[7] 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. 1

[8] 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