# Retail-786k: a Large-Scale Dataset for Visual Entity Matching

**Bianca Lamm**                                          BIANCA.LAMM@DE.MARKANT.COM
*Markant Services International GmbH*
*Offenburg, Germany*


**Janis Keuper**                                          KEUPER@IMLA.AI
*Institute for Machine Learning and Analytics (IMLA)*
*Offenburg University, Germany*

## Abstract

Entity Matching (EM) defines the task of learning to group objects by transferring semantic concepts from example groups (=entities) to unseen data. Despite the general availability of image data in the context of many EM-problems, most currently available EM-algorithms solely rely on (textual) meta data.

In this paper, we introduce the first publicly available large-scale dataset for "visual entity matching", based on a production level use case in the retail domain. Using scanned advertisement leaflets, collected over several years from different European retailers, we provide a total of ∼786k manually annotated, high resolution product images containing ∼18k different individual retail products which are grouped into ∼3k entities. The annotation of these product entities is based on a price comparison task, where each entity forms an equivalence class of comparable products.

Following on a first baseline evaluation, we show that the proposed "visual entity matching" constitutes a novel learning problem which can not sufficiently be solved using standard image based classification and retrieval algorithms. Instead, novel approaches which allow to transfer example based visual equivalent classes to new data are needed to address the proposed problem. The aim of this paper is to provide a benchmark for such algorithms.

Information about the dataset, evaluation code and download instructions are provided under the website: `https://www.retail-786k.org/`.

**Keywords:**  entity matching, images, long-tail, retail, leaflets, products

## 1 Introduction

*Entity Matching* (EM) Barlaug and Gulla (2021), also known as *Record Linkage* or *Fuzzy Matching* in literature, describes the task of identifying semantically grouped objects across diverse data sources. One of the main applications of EM arises in the context of automated product price monitoring, which aims at the comparability of product prices between different retailers and exchangeable products. For example, a common setting is to match large numbers of products from different online shops against a user defined semantic product entity which is grouping products by certain properties. Most of the available product matching approaches are currently solely based on textual information extracted from product descriptions (e.g., in

Figure 1: In the context of retail products, the term "visual entity matching" refers to the task of linking individual product images from diverse sources to a semantic product grouping. Here all images show different products from the same entity which is defined by the fact that single images are used as "placeholders" by retailers to promote all products of the entity. For higher resolution version, refer to Figure 15 in the appendix.

online shops) and available meta data from product data bases. However, in many practical cases, additional product images could be utilized to improve the semantic matching beyond fixed meta data categories.

## 1.1 Definitions

Before elaborating into the properties of visual entities, some basic definitions are needed:

**Products.** Every retail item on sale is uniquely identified by the internationally standardized *Global Trade Item Number* (GTIN).[1] This is a 14-digit number which differs for any possible variation of an item, e.g., different package designs or packaging levels. We follow the definition that a product is an item with a unique GTIN. Figure 9 in the appendix shows three different products with their GTIN. Different products might be very similar or even visually indistinguishable and the application of image classification methods on products often results in *fine-grained image classification* (FC) Wei et al. (2021) problems.

**Entities** are then defined as semantic groupings of products, creating an equivalence class over products induced by some external tasks, e.g. price comparison. In the context of the proposed dataset, an entity represents a group of different GTINs, i.e., contains multiple products which are used by different retailers as equivalent placeholders to promote all of the different products. For example, the three different products in Figure 9 belong to the same entity. A subset of the samples of this entity is illustrated in Figure 1.

**Visual Product Entity Matching.** Extending product matching by the extraction and semantic comparison of image features is a straight forward extension of existing approaches, which receives growing attention. From a computer vision perspective, visual EM is closely related to image retrieval Chen et al. (2022) and fine-grained, long-tail image classification. However, the way that many product entities are defined for applications like price monitoring, demand algorithms to handle different similarity measures and visual intra-entity variances.

---

1. https://www.gtin.info/

Figure 1 illustrates an example for a visual EM instance with a high visual intra-entity variance. In the context of the practical retail task, entities are "things that should be compared" (mostly in price). The actual task is to learn these entities by examples, forming a novel learning problem.

## 1.2 A Dataset for Visual Entity Matching.

Despite the popularity and practical importance of EM, there are currently only a few publicly available datasets which provide sufficient image data to further drive research and development in this area. Therefore, we propose *Retail-786k*, a large open dataset with over 786k manually annotated images of more than 18.8k products in 3,298 different entities. This dataset has been created from large collections of product advertisements in printed leaflets issued by about 130 different European retail chains over a period of two years. It contains high resolution images of typical retail products. The manual entity annotations map the semantic grouping of a commercial price monitoring application, where entities represent groups of products between which prices should be compared. The promotion for all flavors of a yoghurt (strawberry, vanilla, cherry, or raspberry) in one advertisement differs among retailers because different retailers are depicting different flavors as product placeholder in their advertisements. Such different advertisements are illustrated in Figure 1. In contrast to classification, which is identifying predefined discrete subdivisions of classes, the "visual entity matching" task includes a strong transfer learning problem: given only abstract samples of entities as sets of "things that should belong together", the objective is to learn this metric and transfer it to unseen objects. Hence, the challenges in EM do not come from high variance within entities and image backgrounds, but from a combination of fine-grained classification (very similar entities) and retrieval (unknown matching metric). Both tasks, image classification and image retrieval, do not solve the visual EM problem like our baseline experiments show (see Section 4).

**The Aim of the Dataset:** the data of our dataset represents a highly relevant practical task in retail. Retailers have a significant interest about the product pricing to be competitive. Before comparing prices the promotions of the products have to be predicted to belong to the same group and entity, respectively. For this task our dataset was created. The EM task differs significantly from standard classification problems and thus provides an interesting new CV problem. We provide the first large-scale dataset to address these research questions and with our dataset the automatization of the product price monitoring can be pushed.

## 2 Related Work

We focus our literature review on publicly available and annotated datasets that are utilized in the retail context. Here, most image datasets are used to solve in the context of fine-grained classification (FC) task. Regarding entity matching (EM), there are only a few datasets containing images. The images used in the datasets can be roughly grouped into two categories: staged "studio" images recorded in controlled environments and "in the wild" images. The Table 1 and the Table 2 in the appendix give a comprehensive overview over all works discussed in the following paragraphs and additional datasets.

**Online Shopping Products with Images.** The *Retail Product Categorisation Dataset* Elayanithottathil and Keuper (2021) covers about 48k products in 21 categories. The images

of the products have been recorded in controlled environments. The *iMaterialist* competition at FGVC5 iMa (b) created a fine-grained visual categorization dataset of furniture. The competition used staged "studio" images.

**Product Images in the Wild.** Most images of the previous described datasets are highly optimized studio productions. For more realistic real-life scenarios, the *Grozi-120* dataset provides 676 in-studio product images from websites and 4,973 images taken from videos of retail items recorded inside a grocery store Merler et al. (2007). It contains 120 grocery products. An extension of this dataset is *Grozi-3.2k* George and Floerkemeier (2014), offering more products and higher-quality images. *Products-10k* Bai et al. (2020) contains about 10k product classes for about 150k in-shop photos and customer images with human annotations.

**Shelf Images.** Peng et al. (2020); Karlinsky et al. (2017); Goldman et al. (2019) provide datasets containing images of supermarket shelves. Consequently, the datasets consist of "in the wild" images. The *Retail-121* dataset Karlinsky et al. (2017) consists of 121 fine-grained retail product categories. The *RP2K* dataset Peng et al. (2020) provides a combination of shelf images (about 10k) and product images (about 380k).

**Datasets for Product Entity Matching.** While there are several available large datasets for product entity matching, e.g., Köpcke et al. (2010) and Das et al., most of them are solely relying on textual or tabular data. Only a few public sources are providing additional image data: like Plummer et al. (2015) or Wilke and Rahm (2021). The *Flickr30K Entities* dataset contains textual image descriptions for "in the wild" images that do not have any relation to retail products. The dataset from Wilke and Rahm is based on image URLs provided by the *WDC Product Categorization Goldstandard* WDC dataset. The images are grouped into the staged "studio" category. The *Stanford Online Products* datasetOh Song et al. (2016) is built of product photos from the online e-commerce website *eBay.com*. It contains 120,053 images in total and 22,634 classes.

Despite the large number of diverse datasets discussed above, only a small fraction is suitable for the development and evaluation of visual EM-algorithms, as most only provide annotations for (fine-grained) image classification. The few which provide appropriate settings, are mainly based on textual data. Further, the datasets that actually provide images data, offer much smaller numbers of images and classes than our proposed dataset.

## 3 Dataset Description

### 3.1 Data Sources

The starting point of the dataset creation was a large collection of full page leaflet images from about 130 random well-known European retailers between the years 2016 and 2022 provided by the company Markant Services International GmbH. These source images in JPG format either have been obtained by high resolution scans of physically collected leaflets, or have been extracted from publicly available digital leaflets. Large retail chains often consists of different subsidiaries. Such companies potentially use the same product images in their leaflets. We grouped the subsidiaries of a retail chain to avoid the poisoning of the test sets of each entity. The mainly promoted products in leaflets are food and beverages. But household goods, cosmetics, pet foods, or (small) electric devices are also available. Figure 10 in the appendix gives three representative samples of leaflet pages from different retailers.

### 3.2 Data Creation

Multiple steps were necessary to create our dataset. The first step was the manual segmentation of the leaflet page into so-called product information boxes. Each box must consist of the product image, the price, and the product description. Some additional information like logos, price tags, quality seals, or other icons can be included inside a box. Such step was done by a team of 23 persons over about two years. This procedure is integrated in a productive process of a price monitoring tool. After the segmentation process, each product information box is cropped from the leaflet page. Such cropped images form the base of our dataset. The last step of the dataset creation is the actual grouping of the images into entities: each entity represents a group of products with a set of according GTINs. The original leaflet pages are not included in our dataset. Also, neither text information like price, brand, nor description of the promotion are included in the dataset. Furthermore, text detection or recognition annotations of the information boxes on the leaflet pages are also not stored.

### 3.3 Dataset in Numbers

The dataset consists of 786,179 images labeled with 3,298 different entities. The images are split into sub-sets of 748,715 training images and 37,464 test images by the following constraints: I) each entity in the training set must have at least 10 samples, II) each entity in the test set must have at least 3 samples, III) in order to prevent data pollution, images from one retailer can only exist in the training or in the test set for each considered entity, and IV) the minimum resolution of the images is 82 pixels in width and 129 pixels in height while the longer edge is always fixed to 512. This dataset has a size of 65.6 GiB. We also provide the same dataset with lower resolution (with a long edge size of 256), which results in a total size of 22.6 GiB. Figure 5 in the appendix illustrates the resulting unbalanced distribution of the images in training and test set for all entities descending sorted by the training entities. When splitting images into training and test set, we guarantee that no images from the same retailers or subsidiaries exist in both split sets for each entity in order to avoid test set poisoning.

### 3.4 Image Properties and Variance

The images of the products within an entity can be similar and diverse at the same time. We observe high intra-entity variances which differ from typical natural images dominantly used in the computer vision literature: I) Each retailer has its leaflet design. Thus, the background of the images can be very different, varying from a uniform colored bright white, or dark black background over natural background images to more abstract textures. II) Products of an entity differ usually in flavor. The advertisement of such a variance of products also appears in the number and/or the choice of the product placeholders of the promotion image. III) The imaging perspectives, i.e., the point of view, of a product image underlie strong variations. If there are multiple products presented in an image, the perspectives of each might differ, too. Also, the alignment of the product can be different. IV) The color distribution of products often varies strongly between different leaflets. Nevertheless, images from different entities

can also have strong similarities. There is a typically very high number of entities with very low extra-entity variations, i.e., many different entities might be visually very similar.

For instance, the images presented in full resolution in Figure 26, Figure 27, and Figure 28 in the appendix show the training sets of three different entities. All entities represent the product "apple sauce". The brands of the products as well as the packaging sizes differ for each entity. One special feature for the product "apple sauce" is the different given measure units, g or ml, in the advertisement. Figure 2 shows an example for an intra-entity variance along with two samples of a low extra-entity variance. The images inside the dashed rectangle form an entity. The promotions for this entity vary strongly. However, an image of this entity has a strong similarity to images of other entities. Such an occurrence is illustrated by the images in the solid rectangle, circle, and polygon.



Figure 2: Inside the dashed rectangle, samples from a representative entity show the typical intra-entity variations. In contrast to this, there are samples having a strong visually similarity although they belong to different entities. These samples are framed by the solid rectangle, circle, and polygon.

Figure 6 in the appendix shows a further example of samples from different entities although there is a strong visual similarity. The difference between these entities are the package size of the products.

## 4 Baseline Results

We conduct two baseline experiments: First, we treat the proposed EM problem as (fine-grained) classification task. In the second, more realistic dynamic case, we evaluate an image retrieval approach. As discussed, both approaches are theoretically not well suited to actually solve the EM problem and are merely provided to show that standard computer vision algorithms fail to provide the results needed to atomize practically highly relevant EM tasks like price comparison.

### 4.1 EM as Classification Problem

We take our dataset as basis for a static classification problem, i.e., the entities are fixed. We chose three different state-of-the-art image classification models: *ResNet50* He et al. (2016), a *Vision Transformer* model Dosovitskiy et al. (2020), and *ConvNeXt* Liu et al. (2022). The specific configuration settings are described in Section D.1. The results are shown in Table 3 in the appendix. Overall, the *ConvNeXt* model attains the best test set accuracy of 0.855.

### 4.2 EM as Retrieval Problem

For most real applications of visual entity matching, the entities will rapidly change over time as typical retailers are exchanging about 40% of their products during a single year. Hence, the formulation of visual EM as image retrieval problem is more realistic than as

classification task. For our baseline experiment we used the *ROADMAP* approach from Ramzi et al. (2021), which originally has been evaluated on the *Stanford Online Products* (SOP) datasetOh Song et al. (2016), to obtain the 10 most similar images from the test set to a given test image. The settings for our experiments are described in Section D.2. Musgrave et al. (2020) introduces the metric *mAP@R*. A further explanation of this metric and a comparison of used dataset in Ramzi et al. (2021) and our dataset is described in Section D.2. On the test set of our dataset, we achieved a *mAP@R* score of 72.23%. Further, the *R@10* score is 56.34%. Figure 3 shows the image matching results of two different query images, where the left most image of the sub-figures is the query image. The ten most similar images to the query image are predicted. The green-framed images belong to the same entity as the query image. The red-framed images belong to another entity. The left figure of Figure 3 shows a query image with seven of ten positive image matching. The false matched samples have different package size than the query image. However, all package bags of the products have a yellow border at the top. Also, two of the three false matched images have the same price than the query image. The right figure of Figure 3 shows also negative matched results. Some of these images represent another product but from the same brand. It is more difficult to differ matches when the difference is the package size. Two of the false matched images have a bigger package size than the query image. More query examples are provided in the appendix, see Figure 31, Figure 32, and Figure 33.



(a) Package size

(b) Package size and/or product itself

Figure 3: Illustration of retrieval samples showing the difficulty of matching products. The green-framed images belong to the same entity as the query image. The red-framed images belong to another entity. The sub-captions describe the difference between the query image and the false matched images. The images are presented in full resolution in Figure 29 and Figure 30 in the appendix.

### 4.3 Error Analysis

In this section, we go deeper into the error analysis of the two approaches, image classification and image retrieval, respectively.

**Image Classification.** First, we analyze the best classification model *ConvNeXt*. This model achieves a precision and recall score of 0.861 and 0.846, respectively. As mentioned in Section 4.1, the test accuracy is 0.855 and the F1-score reaches 0.832. The Precision, Recall, and F1-score metrics were calculated by the macro average, i.e. for each label the metric

is calculated and the label imbalance of the dataset is not taken into account. Section D.3 describes further investigations regarding misclassifications of images from different entities.

**Image Retrieval.** For image retrieval the metric *Recall@K* is used to evaluate the model. The definition of this metric is "For a single query the *Recall@K* is 1 if a positive instance is in the K nearest neighbors, and 0 otherwise. The Recall@K is then averaged on all the queries."Ramzi et al. (2021). For our test dataset, we achieve *Recall@1* score of 44.85%, *Recall@10* score of 56.34%, and *mAP@R* score of 72.23%. The scores show obviously that the used approach fails for some cases on our dataset. Figure 8 in the appendix shows an example of image retrieval where the most predictions belong to another entity. The top left image beside the vertical line is red framed, i.e. the nearest image to the query image belongs to an entity where single sparkling wine bottles with a content of 0.2l is promoted.

## 5 Discussion and Outlook

With *Retail-786k*, we present the first large-scale visual entity matching dataset that has the potential to become a widely accepted benchmark for retail related image classification and matching tasks. First baseline evaluations show that the EM task that can not sufficiently solved by standard computer vision algorithms like image classification or image retrieval as the best F1-score of the three classification models achieve barely 83.2%. The more realistic image retrieval algorithm even accomplishes only a *R@10*-score of 56%.

**Limitations.** Despite our best efforts to create and release a dataset which holds up to the highest standards, there are some known limitations:

I) Manual miss-annotations: during the dataset creation process, we eliminated many images which have been labeled wrongly. However, there is no guarantee that we did not miss more and less obvious of these cases.

II) Missing textual data: many existing EM algorithms use text information, like price, price reduction, and product descriptions which is implicitly contained in our images as well. To this end, we have not yet extracted this information.

III) Entity labeling: the entities in our dataset have been defined based on the very specific application of price monitoring in printed advertisements. Other EM tasks might define different groupings of the same products. Hence, it is not clear how good results on our dataset will transfer to other problems.

**Outlook.** Even though the presented version of our dataset provides a good basis for the development of new approaches towards visual product entity matching, the underlying data has even more potential for further improvements: the extraction of textual information and price tags could further enhance the database toward multi-modal solutions.

## Broader Impact Statement

The intentions of the proposed dataset includes to provide a benchmark for novel (visual) EM-algorithms. As in any automation task, the development of strong EM-solutions could have negative impact on the employment of people who are currently performing this task manually. On the positive side, i.e. access to powerful price comparison tools will allow customers (especially those with low income) to purchase groceries at the best price.

**Reproducibility Statement**

Information about the dataset and the downloading steps are provided under the website: `https://www.retail-786k.org/`. The code necessary to replicate the experiments discussed in this paper has been made publicly available on the GitHub website: `https://github.com/retail786k/retail786k-dataset`. Also, a comprehensive description of the dataset is described in Section Appendix B Datasheets for Datasets.

**References**

Wdc product categorization goldstandard. `http://webdatacommons.org/structureddata/2014-12/products/gs.html`.

imaterialist challenge at fgvc 2017. `https://www.kaggle.com/competitions/imaterialist-challenge-FGVC2017/overview`, a.

imaterialist challenge (furniture) at fgvc5. `https://www.kaggle.com/c/imaterialist-challenge-furniture-2018`, b.

Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020.

Nils Barlaug and Jon Atle Gulla. Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–37, 2021.

Syed Talha Bukhari, Abdul Wahab Amin, Muhammad Abdullah Naveed, and Muhammad Rzi Abbas. Arc: A vision-based automatic retail checkout system. *arXiv preprint arXiv:2104.02832*, 2021.

Yuanqiang Cai, Longyin Wen, Libo Zhang, Dawei Du, and Weiqiang Wang. Rethinking object detection in retail stores. In *The 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, 2021.

Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan, and Yinghui Xu. Weakly supervised learning with side information for noisy labeled images. In *The European Conference on Computer Vision (ECCV)*, August 2020. Also available as `https://retailvisionworkshop.github.io/recognition_challenge_2020/`.

Sanjib Das, AnHai Doan, Paul Suganthan G. C., Chaitanya Gokhale, Pradap Konda, Yash Govind, and Derek Paulsen. The magellan data repository. `https://sites.google.com/site/anhaidgroup/useful-stuff/the-magellan-data-repository`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Febin Sebastian Elayanithottathil and Janis Keuper. A retail product categorisation dataset. *arXiv preprint arXiv:2103.13864*, 2021.

Patrick Follmann, Tobias Bottger, Philipp Hartinger, Rebecca Konig, and Markus Ulrich. Mvtec d2s: densely segmented supermarket dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 569–585, 2018.

Klaus Fuchs, Tobias Grundmann, and Elgar Fleisch. Towards identification of packaged products via computer vision: Convolutional neural networks for object detection and image classification in retail environments. In *Proceedings of the 9th International Conference on the Internet of Things*, pages 1–8, 2019.

Klaus Fuchs, Tobias Grundmann, Mirella Haldimann, and Elgar Fleisch. Holoselecta dataset: 10'035 gtin-labelled product instances in vending machines for object detection of packaged products in retail environments. *Data in Brief*, 32:106280, 2020. Also available as `https://data.mendeley.com/datasets/gz39ggf35n/1`.

Marian George and Christian Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *European Conference on Computer Vision*, pages 440–455. Springer, 2014.

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5236, 2019.

Yu Hao, Yanwei Fu, and Yu-Gang Jiang. Take goods from shelves: a dataset for class-incremental object detection. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 271–278, 2019.

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.

J. Kittler J. Burianek, A. Ahmadyfard. Soil-47: Surrey object image library. `http://www.ee.surrey.ac.uk/CVSSP/demos/colour/soil47/`.

Philipp Jund, Nichola Abdo, Andreas Eitel, and Wolfram Burgard. The freiburg groceries dataset. *arXiv preprint arXiv:1611.05799*, 2016.

Leonid Karlinsky, Joseph Shtok, Yochay Tzur, and Asaf Tzadok. Fine-grained recognition of thousands of object categories with single-example training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2017.

Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3(1-2):484–493, 2010. Also available as `https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution/`.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Michele Merler, Carolina Galleguillos, and Serge Belongie. Recognizing groceries in situ using in vitro training data. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019. Also available as `https://nijianmo.github.io/amazon/index.html`.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

Jingtian Peng, Chang Xiao, and Yifan Li. Rp2k: A large-scale retail product dataset for fine-grained image classification. *arXiv preprint arXiv:2006.12634*, 2020.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

Elias Ramzi, Nicolas Thome, Clément Rambour, Nicolas Audebert, and Xavier Bitot. Robust and decomposable average precision for image retrieval. *Advances in Neural Information Processing Systems*, 34:23569–23581, 2021.

Jose Rivera-Rubio, Saad Idrees, Ioannis Alexiou, Lucas Hadjilucas, and Anil A Bharath. Small hand-held object recognition test (short). In *IEEE Winter Conference on Applications of Computer Vision*, pages 524–531. IEEE, 2014.

Anderson Rocha, Daniel C Hauagge, Jacques Wainer, and Siome Goldenstein. Automatic fruit and vegetable classification from images. *Computers and Electronics in Agriculture*, 70(1):96–104, 2010.

Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 3–11, 2016. Also available as `https://www.epfl.ch/labs/mmspg/downloads/food-image-datasets/`.

Venkatesh Umaashankar, Aditi Prakash, et al. Atlas: a dataset and benchmark for e-commerce clothing product categorization. *arXiv preprint arXiv:1908.08984*, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Gül Varol and Rıdvan Salih Kuzu. Toward retail product recognition on grocery shelves. In *Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*, volume 9443, pages 46–52. SPIE, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, Lingqiao Liu, and Jian Yang. Rpc: a large-scale and fine-grained retail product checkout dataset. `https://www.kaggle.com/datasets/diyer22/retail-product-checkout-dataset`, 2022.

Moritz Wilke and Erhard Rahm. Towards multi-modal entity resolution for product matching. In *GvDB*, 2021.

## Appendix A. Dataset Information

**Dataset Description.** An comprehensive explanation of our dataset is described in "Datasheets for Datasets", Section B.

**License.** Our dataset is licensed under the *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International* license.

**Hosting.** The dataset is uploaded to the data repository `https://zenodo.org/record/7970567` and the DOI of the dataset is 10.5281/zenodo.7970566.

**Support and Maintenance.** The supporting, hosting and maintenance plan is described in Section B.7.

**Download and Reproducibility.** Information about the dataset and code as well as the downloading steps are provided under the website: `https://www.retail-786k.org/`.

**Author responsibility.** As authors, we bear all responsibility in case of any violation of rights during the collection of the data or other work, and will take appropriate action when needed, e.g., to remove data with such issues.

## Appendix B. Datasheets for Datasets

### B.1 Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The dataset was created as benchmark for "visual entity matching" tasks. Most existing Entity Matching (EM) approaches solely rely on meta data even if the data basis are products.

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
The company Markant Services International GmbH and the Institute of Machine Learning and Analytics from the Offenburg University are involved in the dataset creation. From the company a team of 23 persons worked for about two years on the annotation of the dataset. Also, Janis Keuper and Bianca Lamm from the institute participated on the dataset creation.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
The creation costs of about 1 million USD of the dataset was funded by the German company Markant Services International GmbH that provides a price monitoring service.

**Any other comments?**
None.

### B.2 Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The instances are cropped images in JPG format from German leaflets. Each image promotes one or multiple retail product(s). The instances are grouped in so-called entities being a semantic grouping of products. The entities are sequentially numbered with an ID. Two instances from different groupings are shown in Figure 4.



Figure 4: Illustration of two instances from different groupings.

**How many instances are there in total (of each type, if appropriate)?**
There are 786,179 images in total that are distributed over 3,298 entities (semantic groupings).
These images are split into a training set of 748,715 images and a test set of 37,464 images.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset contains all possible instances because no more products were considered. Therefore, there are no more cropped images from leaflets of other products.

**What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
An instance is a cropped image from a leaflet. These leaflets either have been obtained by high resolution scans of physically collected leaflets, or have been extracted from publicly available digital leaflets. At the resulted images in JPG format are unprocessed images.

**Is there a label or target associated with each instance?** If so, please provide a description.
The label is the entity ID that is given by the folder name in which the instance is stored.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
There is no text extraction of the product descriptions from each promotion image. The wording of this text is unavailable. Information about price, discount, brand, and some attributes are not available at the time of the dataset creation. But these information can be received and another version of the dataset could be published.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
An instance, a promotion image, is assigned a list of GTIN from the promoted products. Hence, an entity has a set of GTIN lists. Such GTIN lists are disjunctive for all entities. Inside an entity, an instance must have at least one equal GTIN of its GTIN-list with another instance of this entity.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.
The data splits are given by the dataset itself. There are training and test splits. No images from the same retailers or subsidiaries exist in both split sets for each entity in order to avoid test set poisoning due to the visual similarity of promotions.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Redundancies can appear in the dataset, e.g. there could exist different entities that represent the same product. This could be caused by false manual classification of GTINs for the promoted products in the instance.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not link to external resources.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

No, there are no personal data included in the dataset.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

No, such data does not exist in the dataset.

**Any other comments?**

None.

### B.3 Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The instances are derived from online available leaflets. The leaflet images were neither validated nor verified.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?** How were these mechanisms or procedures validated?

The following procedures were used to collect the data:

1. Downloading the public available leaflets of each retailer at each calendar week.

2. Uploading the digital leaflets into an internal software of the company Markant Services International GmbH. The software converts the leaflets from PDF format into JPG format.

3. The employees of the company draw rectangles onto each leaflet page to manual segment the page into so-called product information boxes. Each box must consist of the product image, the price, and the product description. Some additional information like logos, price tags, quality seals, or other icons can be included inside a box.

4. The leaflet can be cropped into the instances of the dataset with the stored rectangle coordinates from the previous step.

There was no validation of the procedures, especially for the drawing of the product information boxes.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
The dataset is not a sample from a larger set.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and howwere they compensated (e.g., how much were crowdworkers paid)?**
The company Markant Services International GmbH, who funded the dataset, involve an own team of its employees for the data collection process. The specific payment cannot be disclosed because of the confidentiality of company internal data. But the is a lower bound given by the legal minimum wage of 12.60 USD per hour in Germany (in 2022). The total amount of costs aggregate about 1m USD.

**Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The team of 23 persons worked in a period of two years to create the dataset. The basis of the dataset consists of about 68k leaflets. The release years of the considered leaflets and hence the data collection timeframe are from 2016 to 2022.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Unknown to the authors of the datasheet. But there are no participant risks. There is no personal data collection from the people who are created the dataset.

**Any other comments?**
None.

## B.4 Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.
There are some entities that group arbitrary products due to manual miss-annotations of the GTINs to these products. Because of this occurrence, cake baking mix, cream, and pudding form an entity. We eliminated many images that have been labeled wrongly to clean the dataset. However, there is no guarantee that we did not miss more and less obvious of these cases.

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
No, the "raw" data is not saved separately. The leaflets being the basis for the dataset creation are stored at the company Markant Services International GmbH internally.

**Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.
No, the cleaning of the data was done by manual screening. There is no code / software available.

**Any other comments?**
None.

## B.5 Uses

**Has the dataset been used for any tasks already?** If so, please provide a description.
The dataset itself has not been used for any tasks already. But a subset of the dataset was used company internally for the price monitoring service.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
No, there is no repository.

**What (other) tasks could the dataset be used for?**
The dataset can be also used for the image classification or the image retrieval task. Both tasks are investigated in our experiments. Unfortunately, they do not solve the visual EM problem. Furthermore, the images in the dataset can be used for the OCR task. But the Ground Truth data of the text in the images is not provided by our dataset.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or

other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
Unknown to the authors of the datasheet.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.
Unknown to the authors of the datasheet. But the license of the dataset prohibits commercial purpose and the distribution of a modified version of the dataset.

**Any other comments?**
None.

### B.6 Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
Yes, the dataset is publicly available on the internet.

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
Information about the dataset and code as well as the downloading steps are provided under the website: `https://www.retail-786k.org/`. The dataset can be also downloaded directly via the URL `https://zenodo.org/record/797056`. The DOI of the dataset is 10.5281/zenodo.7970566.

**When will the dataset be distributed?**
The dataset was released at Jun 06, 2023.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
This dataset is licensed under the *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International* license. For more information about the license see `https://creativecommons.org/licenses/by-nc-nd/4.0/`.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
Unknown to the authors of the datasheet.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
Unknown to the authors of the datasheet.

**Any other comments?**
None.

### B.7 Maintenance

**Who will be supporting/hosting/maintaining the dataset?**
Bianca Lamm and Janis Keuper from the Institute of Machine Learning and Analytics from the Offenburg University are supporting/maintaining the dataset.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
The curators of the dataset, Bianca Lamm and Janis Keuper, can be contacted via info@Retail-786k.org, bianca.lamm@hs-offenburg.de, and keuper@imla.ai.

**Is there an erratum?** If so, please provide a link or other access point.
No.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
In the future, another version of the dataset could be published with information about price, discount, brand, and so on of the products per instance.
There is no fixed date for the upload of the new version, yet. Also, there will be no fixed periods for updating the dataset.
The new versions will be updated by the authors of the datasheet. The updates to the dataset will be communicated via the dataset website: `https://www.retail-786k.org/`.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.
The dataset does not relate to people.

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
There are no older versions of the dataset, yet. If newer versions are available, then old version will still be available.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

The license of our dataset does not allow to extend/augment/build on/contribute to the dataset. Interested persons can contact the curators via info@Retail-786k.org.

**Any other comments?**
None.

## Appendix C. Additional Related Work

In addition to the datasets described in Section 2, there are further datasets of interest that are described in this section. The Table 1 and the Table 2 summarize all described datasets from this section and Section 2 in relation to the number of images and number of classes per dataset. Besides, we indicate whether the dataset is suitable for the FC or EM tasks. The last two columns define the image category: staged "studio" images recorded in controlled environments and "in the wild" images.

**Online Shopping Products with Images.** The assignment of attribute labels for product images was the task of the *iMaterialist* competition at FGVC4 iMa (a). The *Amazon Review Data* (2018) Ni et al. (2019) describes a dataset of products offered by *Amazon*. It contains about 233.1m reviews, e.g., ratings, text, helpfulness votes, or user provided product images. Additionally, the dataset consists of metadata for 15.5m products, including product images. The data is comprised in 29 categories, like *Electronics, Books, All Beauty, Home and Kitchen*, or *Pet Supplies*. The *Atlas* dataset Umaashankar et al. (2019) focuses on clothing images, which were collected from popular Indian e-commerce stores. Image descriptions and pre-defined product taxonomies are also provided in this 52 category dataset.

**Product Images in the Wild.** The *AliProducts Challenge* dataset Cheng et al. (2020) consists of real-life ("in the wild") image scenarios, e.g., products in customer hand, on a table, in a shopping basket, or on a supermarket shelf. The dataset comprises about 50k different products and about 3m images in total. A dataset considering images recorded from hand-held or wearable cameras and showing hand-held grocery products is the *SHORT* dataset Rivera-Rubio et al. (2014). The training images are of high quality but form only a small set. Oppositely, the test set comprises nearly 135k smartphone-captured images of 30 products.

**Shelf Images.** 25 grocery classes with about 5k images are included in the *Freiburg Groceries Dataset* Jund et al. (2016). The dataset from Varol and Kuzu (2015) comprises more than 5k shelf images consisting of 10 tobacco brands. Karlinsky et al. (2017) provides the *GameStop* dataset that contains of video game store images. The dataset covers about 3.7k video game categories. More than 1m individual product images retrieved from 11,762 store shelf photographs are provided by the *SKU-110* dataset Goldman et al. (2019). This dataset is intended for object detection in densely packed scenes. The *Locount* dataset Cai et al. (2021), which consists of about 50k images with more than 1.9m object instances in 140 product categories, aims towards localize groups of products with the number of instances.

**Product Checkout Images.** The *RPC* dataset Wei et al. (2022) includes checkout data. It consists of 200 categories and 83,739 images in total. The images are divided into single-product images for the training set and 30k multi-product checkout images for the test set. The *ARC-100* dataset Bukhari et al. (2021) contains 31k RGB images with a size of $640 \times 480$ pixel. The images show 100 commonly found retail items in Lahore, Pakistan. The products were lied on a black, matte finish conveyor belt in various logical orientations. The dataset is made in a controlled environment.

**Products in Vending Machines.** Another way to record images of retail products is in vending machines. For the *TGFS* dataset Hao et al. (2019) 38k images from self-service vending machines with a resolution of $480 \times 640$ pixel are collected. The images show the product removal. The dataset comprises 24 fine-grained and 3 coarse classes. Fuchs et al. (2019) creates a dataset of 300 images illustrating the view on vending machines. There are 15k labeled instances of 90 products in the images. A similar build dataset as before is the *HoloSelecta* dataset Fuchs et al. (2020). It also contains shelf images of vending machines of about 295 images with 10,035 labeled instances of 109 products. The considered vending machines include food as well as beverage products. The labeled images provide bounding box coordinates and also information about the brand, product name, and the GTIN of the product.

**Data Bases with Food Images.** More datasets deal with retail products in the food domain. One dataset limits oneself to fruits and vegetables Rocha et al. (2010). This dataset consists of 2,633 images and 15 categories. Moreover, the Food-5K dataset Singla et al. (2016) contains 2.5k food and 2.5k non-food images. The food images illustrate cooked meals on plates and the non-food images show, e.g., humans, animals, flowers, or landscapes. Singla et al. (2016) provides another food image dataset called *Food-11*. This dataset has about 16k food images that are grouped in 11 categories.

**Miscellaneous.** The *MVTec D2S* dataset Follmann et al. (2018) provides semantic segmentation annotated images of retail products. The images are recorded in a laboratory environment. Another retail product dataset is *SOIL-47* J. Burianek. The characteristic feature of the product images is the 20 different captured horizontal views. The dataset consists of 47 product categories and each category covers 21 images.

Number of images in training and test set



Figure 5: Entity histograms by train and test split of the dataset.



| (a) 175g | (b) 250g | (c) 290g | (d) 4×50g |

Figure 6: Illustration of samples from different entities that have a strong visual similarity. The difference of the entities is the package size that is specified by the caption of each subfigure. For higher resolution versions, refer to Figure 22, 23, 24, and 25.

## Appendix D. Detailed Elaboration

Figure 5 illustrates the resulting unbalanced distribution of the images in training and test set for all entities descending sorted by the training entities. Figure 6 shows a further example of samples from different entities although there is a strong visual similarity. The difference between these entities are the package size of the products.

### D.1 Baseline Results on EM as Classification Problem

We chose three different state-of-the-art image classification models: *ResNet50* He et al. (2016), a *Vision Transformer* model Dosovitskiy et al. (2020) with a *vit_b_16* architecture and *ConvNeXt* Liu et al. (2022) using the base model architecture. Each model has been pretrained on *ImageNet* Deng et al. (2009) and was trained for 50 epochs of fine-tuning. For the *ResNet50* model, we used a NVIDIA GeForce RTX 3090 GPU and a batch size of 32. The

same batch size is used for the training for the *Vision Transformer* model and the *ConvNeXt* model. These models were trained on three NVIDIA A100-SXM4-40GB. All models use a SGD optimizer with a learning rate of 0.001 plus a momentum of 0.9. Furthermore, we add a decay of the learning rate of 0.1 every 7 epochs. Results are shown in Table 3. The *ResNet50* model results in the lowest test set accuracy of 0.824. An improvement is achieved by the *Vision Transformer* model with a test set accuracy of 0.834. The *ConvNeXt* model attains the best test set accuracy of 0.855. The F1-score values of the test data are also calculated for each model.

## D.2 Baseline Results on EM as Retrieval Problem

We trained on 4 NVIDIA TITAN GPUs for 43 hours, using the following settings for the training: ResNet50 as backbone model, 100 epochs, batch size of 128, Adam optimizer, and an embedding dimension of 512. The settings for the *ROADMAP* approach are the same as in the experiments of the paper Ramzi et al. (2021). Musgrave et al. (2020) introduces the metric *mAP@R*, where the authors define $R$ as "For each query[...], let $R$ be the total number of references that are the same class as the query" Musgrave et al. (2020). Hence, the value of $R$ is different for each query and is therefore not a configurable setting value. The total number of references for each entity is consequently the sum of images of this entity. The authors of Ramzi et al. (2021) investigate the three dataset SOP Oh Song et al. (2016), CUB Wah et al. (2011), and INaturalist Van Horn et al. (2018). The last called dataset contains images of animals and comprises 8,142 classes. A specialization to the animal bird is the CUB dataset. It contains 200 fine-grained classes of birds. The SOP dataset has the highest number of classes, 22,634 classes. A comparison of the three datasets and our dataset in particular to the mAP@R score and the number of classes is shown in the Table 4. On the test set of our dataset, we achieved a *mAP@R* score of 72.23%. Further, the *R@10* score is 56.34%.

## D.3 Error Analysis on Image Classification

Figure 7 shows three typical examples of misclassification. The left image of the three subfigures are example images from an entity that are misclassified to the next to it illustrated entity. The left image of the left subfigure indicates the advertisement for a box of six sparkling wine bottles. This entity is misclassified to the entity that only represents only one bottle. Not only the number of products but also the packaging size of a product defines the difference between entities. The centered subfigure illustrated such case. The entity representing the packaging size of 250g is often misclassified to the entity with the packaging size of 500g. Furthermore, the design of the packaging of products leads to misclassification. The left image of the right subfigure promotes 1l of milk. However, the product is misclassified to the product 1l of cream. Often, there are also a strong difference between the number of the training images of both entities. For deeper, manual error analysis, we created a much smaller long-tail dataset which consists of 137,416 images in total, where the train split contains 115,889 images, and the test split contains 21,527, respectively. Table 5 shows the main reasons for false predictions based on a manual evaluation.

Figure 7: Example image pairs of failure cases: The images of the entities on the left side are misclassified to the entities illustrated by the images on the right side. The images are presented in full resolution from Figure 34 to Figure 39.

## D.4 Error Analysis on Image Retrieval



Figure 8: Illustration of retrieval sample. The green-framed images belong to the same entity as the query image. The red-framed images belong to another entity. This image query illustrates the lack of the investigated image retrieval model. The most predictions and especially the "nearest" image in the embedding space to the query image belong to another entity. The image is presented in full resolution in Figure 40.

(a) 04028900007284          (b) 04028900007413          (c) 04028900007321

Figure 9: Illustration of three different products, each product has its own GTIN described in the caption. They differ from each other in the flavor but might be used independently by different retailers to promote discounts for all three of them. Hence, they form an entity in a price comparison task. For higher resolution versions, refer to Figure 16, 17, and 18.

## Appendix E. Additional Images

Example images of a training and a test set of the entity ID 28 are shown in Figure 11 and Figure 12, respectively.

A further example for strong similarity between the images of different entities are shown in Figure 13 and Figure 14.

Furthermore, more image retrieval results for different query images are generated. Figure 31, Figure 32, and Figure 33, are shown the results.

Subsequently, some figures of the paper are illustrated in full resolution.

Figure 10: Examples of leaflet pages from different European retailers which are the basis of the proposed dataset. For higher resolution versions, refer to Figure 19, 20, and 21.

Table 1: Overview of different datasets described in Section 2 and in Section C with the number of images and number of classes per dataset. Furthermore, we indicate whether the dataset is suitable for the FC or EM tasks. Also, we define the image category for the data, if possible. Images recorded in controlled environment are grouped into the staged "studio" image category (studio). Other images belong to the "in the wild" image category (wild). (*) Number of advocated images in Wilke and Rahm (2021), but the dataset could contain more images.

| | # images | # classes | FC | EM | image category studio | wild |
|---|---|---|---|---|---|---|
| **Online Shopping Products with Images** | | | | | | |
| *Amazon Review Data* (2018) Ni et al. (2019) | - | 29 | ✓ | ✗ | ✓ | ✗ |
| *Retail Product Categorisation Dataset* Elayanithottathil and Keuper (2021) | 48,000 | 21 | ✓ | ✗ | ✓ | ✗ |
| *Atlas* Umaashankar et al. (2019) | 186,150 | 52 | ✓ | ✗ | ✓ | ✗ |
| *iMaterialist* competition at FGVC4 iMa (a) | 50,461 | 381 | ✓ | ✗ | ✓ | ✗ |
| *iMaterialist* competition at FGVC5 iMa (b) | 214,028 | 128 | ✓ | ✗ | ✓ | ✗ |
| **Product Images in the Wild** | | | | | | |
| *Grozi-120* Merler et al. (2007) | 5,649 | 120 | ✓ | ✗ | ✓ | ✓ |
| *Grozi-3.2k* George and Floerkemeier (2014) | 9,030 | 80 | ✓ | ✗ | ✓ | ✓ |
| *Products-10k* Bai et al. (2020) | 150,000 | 10,000 | ✓ | ✗ | ✓ | ✓ |
| *AliProducts Challenge* Cheng et al. (2020) | 3,000,000 | 50,000 | ✓ | ✗ | ✗ | ✓ |
| *SHORT* Rivera-Rubio et al. (2014) | 134,524 | 30 | ✓ | ✗ | ✓ | ✓ |
| **Shelf Images** | | | | | | |
| Varol and Salih Varol and Kuzu (2015) | 5,000 | 10 | ✓ | ✗ | ✗ | ✓ |
| *Freiburg Groceries Dataset* Jund et al. (2016) | 5,021 | 25 | ✓ | ✗ | ✗ | ✓ |
| *Retail-121* Karlinsky et al. (2017) | 688 | 121 | ✓ | ✗ | ✗ | ✓ |
| *GameStop* Karlinsky et al. (2017) | 1,039 | 3,700 | ✓ | ✗ | ✗ | ✓ |
| *RP2K* Peng et al. (2020) | 2,388 | 394,696 | ✓ | ✗ | ✗ | ✓ |
| *SKU-110* Goldman et al. (2019) | 11,762 | 110,712 | ✓ | ✗ | ✗ | ✓ |
| *Locount* Cai et al. (2021) | 50,394 | 140 | ✓ | ✗ | ✗ | ✓ |
| **Product Checkout Images** | | | | | | |
| *RPC* Wei et al. (2022) | 83,739 | 200 | ✓ | ✗ | ✓ | ✗ |
| *ARC-100* Bukhari et al. (2021) | 31,000 | 100 | ✓ | ✗ | ✓ | ✗ |
| **Our Retail-786k dataset** | 786,179 | 3,298 | ✗ | ✓ | ✓ | ✗ |

Table 2: Continuation of the overview of different datasets described in Section 2 and in Section C with the number of images and number of classes per dataset. Furthermore, we indicate whether the dataset is suitable for the FC or EM tasks. Also, we define the image category for the data, if possible. Images recorded in controlled environment are grouped into the staged "studio" image category (studio). Other images belong to the "in the wild" image category (wild). (*) Number of advocated images in Wilke and Rahm (2021), but the dataset could contain more images.

| | | | | | image category | |
| | # images | # classes | FC | EM | studio | wild |
|---|---|---|---|---|---|---|
| **Products in Vending Machines** | | | | | | |
| *TGFS* Hao et al. (2019) | 38,000 | 24 | ✓ | ✗ | ✗ | ✓ |
| Fuchs et al. Fuchs et al. (2019) | 300 | 90 | ✓ | ✗ | ✗ | ✓ |
| *HoloSelecta* Fuchs et al. (2020) | 295 | 109 | ✓ | ✗ | ✗ | ✓ |
| **Data Bases with Food Images** | | | | | | |
| Rocha et al. Rocha et al. (2010) | 2,633 | 15 | ✓ | ✗ | ✓ | ✗ |
| *Food-5K* Singla et al. (2016) | 5,000 | 2 | ✓ | ✗ | ✗ | ✓ |
| *Food-11* Singla et al. (2016) | 16,643 | 11 | ✓ | ✗ | ✗ | ✓ |
| **Miscellaneous** | | | | | | |
| *MVTec D2S* Follmann et al. (2018) | 21,000 | 60 | ✓ | ✗ | ✓ | ✗ |
| *SOIL-47* J. Burianek | 987 | 47 | ✓ | ✗ | ✓ | ✗ |
| **Product Entity Matching** | | | | | | |
| *Benchmark datasets for entity resolution*Köpcke et al. (2010) | 0 | - | ✗ | ✓ | - | - |
| *The Magellan Data Repository* Das et al. | 0 | - | ✗ | ✓ | - | - |
| *Flickr30K Entities* Plummer et al. (2015) | 31,783 | - | ✗ | ✓ | - | - |
| Wilke and Rahm Wilke and Rahm (2021) | 1,763* | 1,763 | ✗ | ✓ | ✓ | ✗ |
| *Stanford Online Products* Oh Song et al. (2016) | 120,053 | 22,634 | ✗ | ✓ | ✓ | ✗ |
| **Our Retail-786k dataset** | 786,179 | 3,298 | ✗ | ✓ | ✓ | ✗ |

Table 3: Illustration of the results for the "Static" EM problem. The models *ResNet50, Vision Transformer* and *ConvNeXt* were examined. The accuracy and the F1-score of the test data as well as the training time in hours are given. (*) Macro average is used.

| model | F1-score* | test accuracy | training time [h] |
|---|---|---|---|
| *ConvNeXt* | 0.832 | 0.855 | 76 |
| *Vision Transformer* | 0.806 | 0.834 | 62 |
| *ResNet50* | 0.793 | 0.824 | 51 |

Table 4: Comparison of the number of classes and the *mAP@R* value between three image classification datasets and our dataset.

| dataset | number of classes | mAP@R [%] |
|---|---|---|
| CUB | 200 | 25.27±0.12 |
| INaturalist | 8,142 | 27.01±0.10 |
| SOP | 22,634 | 56.64+-0.09 |
| ours | 3,298 | 72.23 |

Table 5: Manual evaluation [in %] of false classified samples in a subset 21k images.

| different size or quantity | different training and test data | unknown | similar design | special illustration | focus on product | poor image quality | false GT label |
|---|---|---|---|---|---|---|---|
| 39.0 | 14.4 | 10.6 | 9.7 | 8.9 | 7.2 | 6.8 | 3.4 |

Figure 11: The training set of the entity ID 28 is depicted. Figure 12 shows the corresponding test set.

Figure 12: The test set of the entity ID 28 is depicted. Figure 11 shows the corresponding training set.
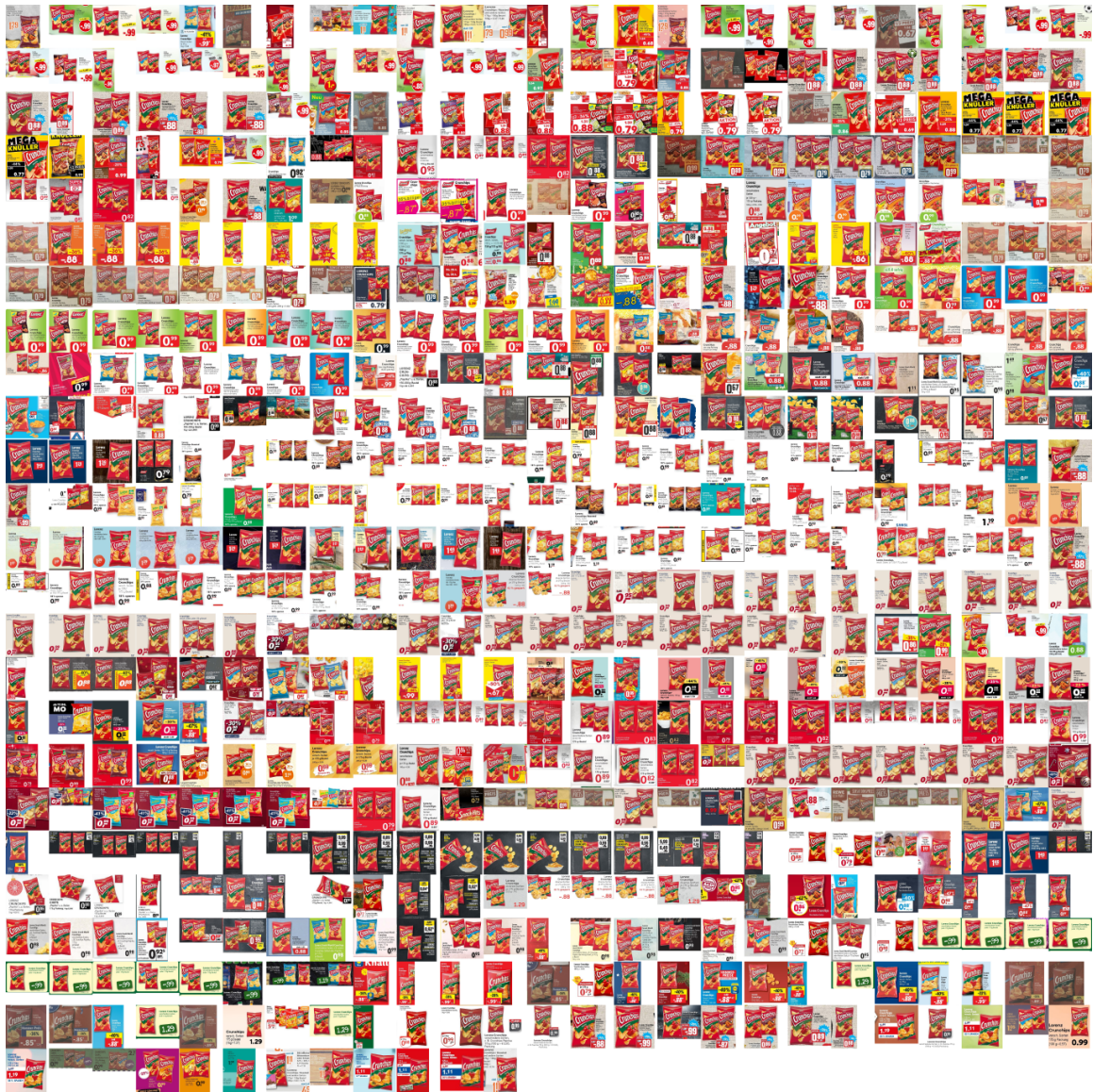
Figure 13: The training set of the entity ID 18 is depicted.
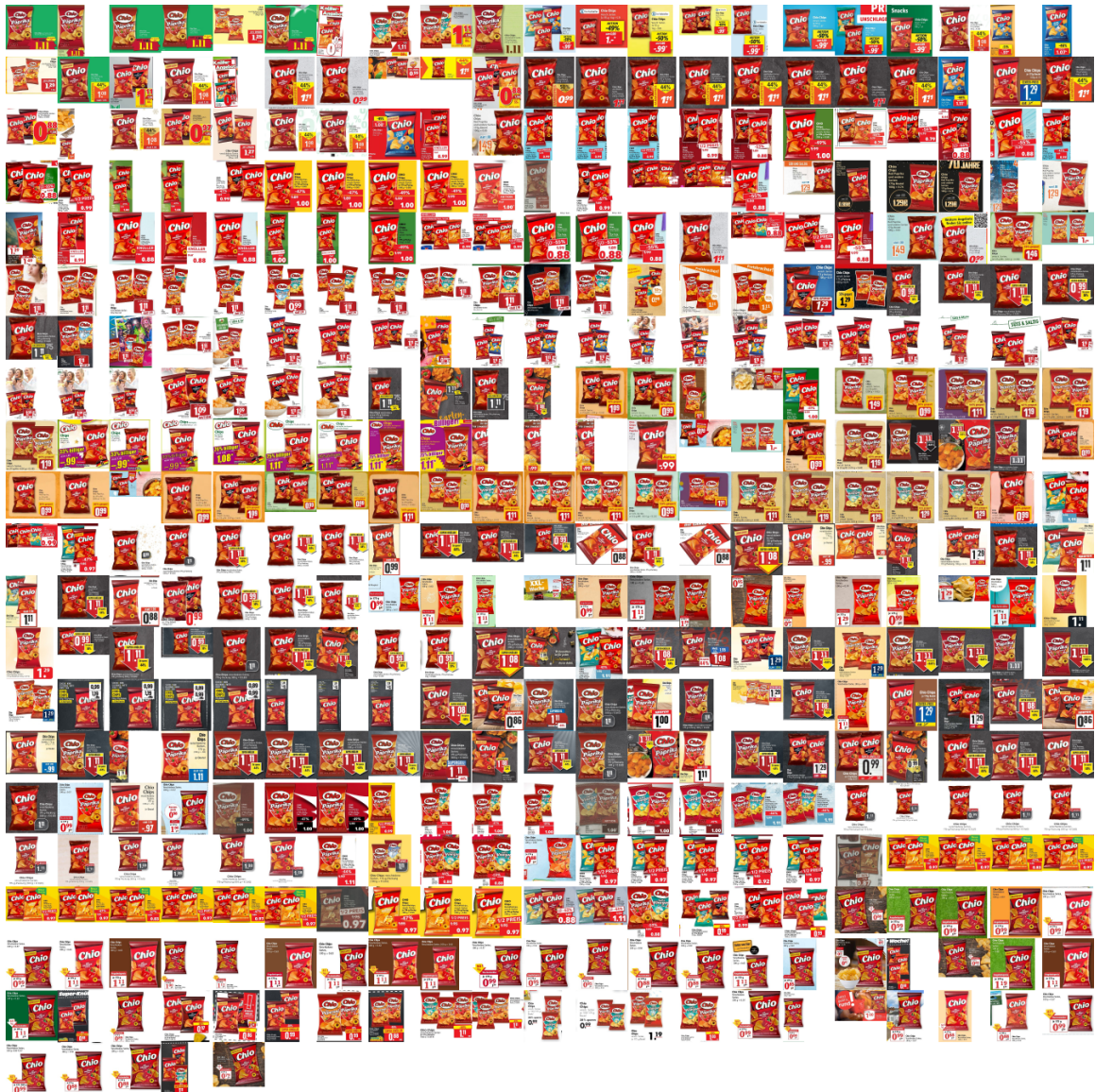
Figure 14: The training set of the entity ID 51 is depicted.

Figure 15: Illustration of the Figure 1 in full resolution.

Figure 16: Illustration of the left product image in Figure 9 in full resolution.

Figure 17: Illustration of the middle product image in Figure 9 in full resolution.

Figure 18: Illustration of the right product image in Figure 9 in full resolution.

Figure 19: Illustration of the left leaflet page in Figure 10 in full resolution.

Figure 20: Illustration of the middle leaflet page in Figure 10 in full resolution.

**LEIFHEIT**

**Bügeltisch »Air Board Compact M Plus«**
Bügelfläche ca. B 120 x T 38 cm, Spezialbezug mit Thermo-Reflect-Technologie, 5 Jahre Hersteller-garantie
je

**-38%**
~~63.35~~
**38.98***

✔ Mit Elektroanschluss und Kabelhalter

Höhen-verstellbar bis ca. 98 cm

**PHILIPS**

Dampfmenge 100 g/min

**Dampfbügel-station »HI5917/20«**
Keramikbügelsohle, Pumpendruck 5 bar, Dampfstoß bis zu 180 g, Wassertank ca. 1,1 l, 2.400 W
je

📢 **-65%**
~~UVP 199.99~~
**68.23***

**Bügel-tischbezug »Air Board Thermo Reflect M«**
für Bügeltische bis ca. B 125 x T 40 cm
je

**-41%**
~~16.56~~
**9.74***

Dampfmenge 25 g/min

**PHILIPS**

**Dampfbügeleisen »GC1756/20«**
Dampfstoß bis zu 100 g, Wasser-behälter ca. 220 ml, 2.000 W
je

**nur**
**29.23***

**K-CLASSIC Doppel-treteimer**
aus Kunststoff, Eimer entnehmbar, (je ca. 11 l), ca. B 35 x H 45 x T 29 cm
je

**nur**
**9.74***

**KÄRCHER**

**Glasreiniger-Konzentrat »RM 500«**
je 500-ml-Flasche (1 l = 9.72)

**-30%**
~~UVP 6.99~~
**4.86***

Abnehmbarer Tank ca. 300 ml

Passend dazu

**Akku-Fensterreiniger »WV 1 Plus«**
Absaugdüse ca. 25 cm, Schmutzwasser-behälter ca. 100 ml, Akkuladezeit ca. 185 min, Akku-Laufzeit ca. 25 min ca. B 13 x H 28 x T 25 cm, inkl. Fensterreiniger-Konzentrat ca. 20 ml
je

**WEDO Rollhocker »Step«**
aus Kunststoff, integrierter Gummistandring, max. 150 kg, ca. Ø 29/44 x H 43 cm
je

✔ Mit 3 Gleitrollen

**-60%**
~~UVP 61.00~~
**24.36***

**3 in 1**
Boden-, Teppich- und Handdampfreiniger

**CLEANmaxx**

**Dampfbesen »1123«**
mit Knick-gelenk, Fenster-reinigeraufsatz, Bürstenaufsatz, max. 1.300 W
je

**-65%**
~~UVP 99.99~~
**34.11***

**-29%**
~~UVP 54.99~~
**38.98***

Figure 21: Illustration of the right leaflet page in Figure 10 in full resolution.

42

Figure 22: Illustration of the first image of Figure 6 in full resolution.

Figure 23: Illustration of the second image of Figure 6 in full resolution.

Figure 24: Illustration of the third image of Figure 6 in full resolution.

Figure 25: Illustration of the fourth image of Figure 6 in full resolution.

Figure 26: The training set of the entity ID 371 is depicted.

Figure 27: The training set of the entity ID 373 is depicted.

Figure 28: The training set of the entity ID 1208 is depicted.

Figure 29: Illustration of the left image of Figure 3 in full resolution.
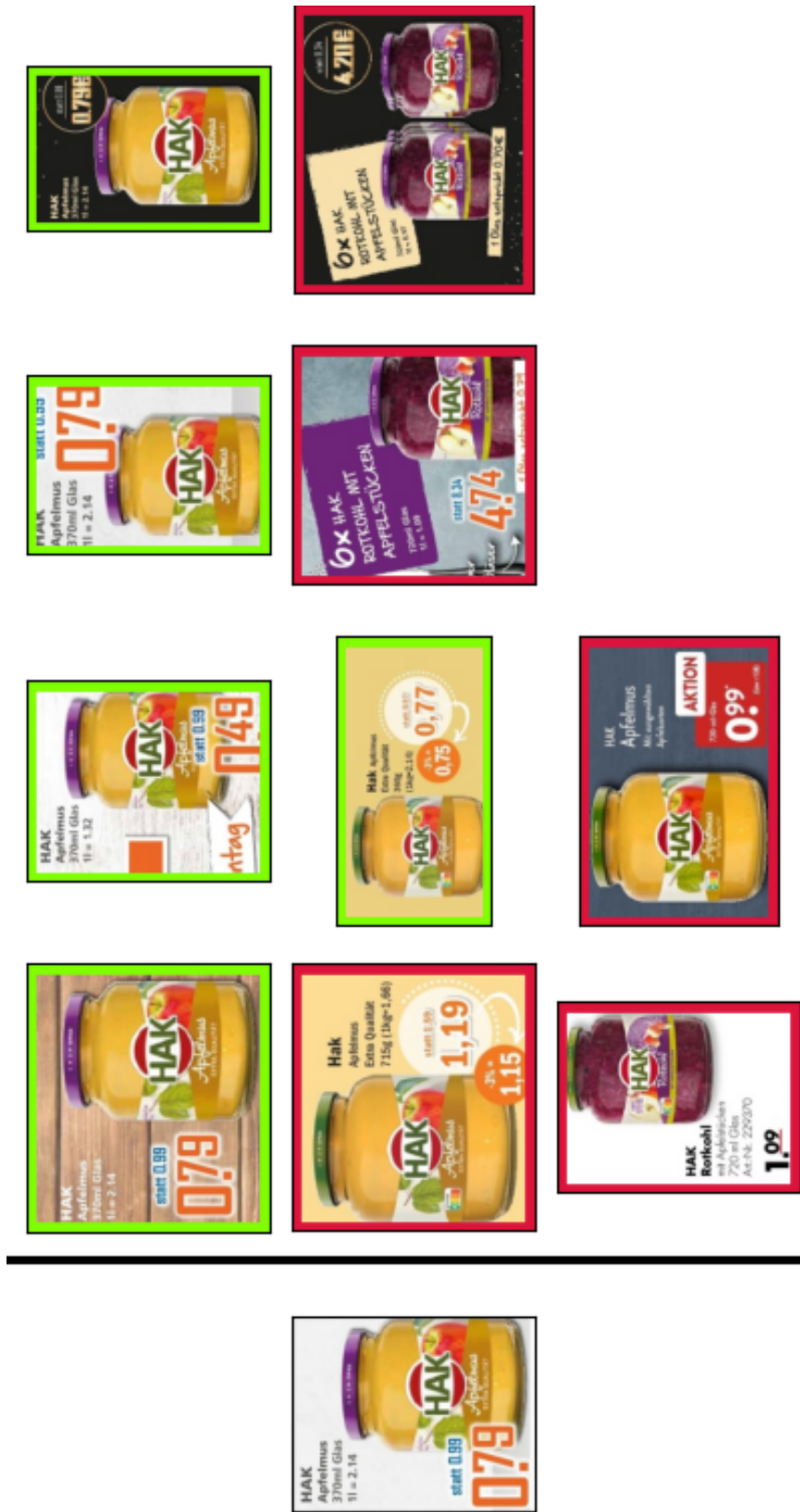
Figure 30: Illustration of the right image of Figure 3 in full resolution.

Figure 31: A further example for image retrieval with the product "apple sauce". This entity differs in the brand as the result of the right figure of Figure 3.

Figure 32: A further example for image retrieval with the product "apple sauce". This entity differs in the packaging as the result of the right figure of Figure 3 and Figure 31.

Figure 33: Illustration of an image retrieval with the product "potato chips". The query image is from the test set of the entity ID 51. The training set is shown in Figure 14.

Figure 34: Illustration of the top left image of Figure 7 in full resolution.

Figure 35: Illustration of the top centered image of Figure 7 in full resolution.

Figure 36: Illustration of the top right image of Figure 7 in full resolution.

Figure 37: Illustration of the bottom left image of Figure 7 in full resolution.

Figure 38: Illustration of the bottom centered image of Figure 7 in full resolution.

Figure 39: Illustration of the bottom right image of Figure 7 in full resolution.

Figure 40: Illustration of an image retrieval with the product "sparkling wine". The query image is from the test set of the entity ID 430.