Pattern Recognition 140 (2023) 109519

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

MEP-3M: A large-scale multi-modal E-commerce product dataset

Fan Liu[a,b,1,*], Delong Chen[a,1], Xiaoyu Du[c], Ruizhuo Gao[a], Feng Xu[a]

[a] Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, China
[b] Science and Technology on Underwater Vehicle Technology Laboratory, Harbin, China
[c] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

ARTICLE INFO

Article history:
Received 11 October 2021
Revised 15 February 2023
Accepted 11 March 2023
Available online 16 March 2023

Keywords:
Dataset
E-commerce product classification
Fine-grained learning
Hierarchical classification
Automatic Checkout

ABSTRACT

The product categories are vital for the E-commerce platforms due to the core applications on automatic product category assignment, personalized product recommendations, etc. In this paper, we construct a large-scale **M**ulti-modal **E**-commerce **P**roducts classification dataset MEP-3M, which is large-scale, hierarchical-categorized, multi-modal, fine-grained, and long-tailed. Statistically, MEP-3M consists of over 3 million products, thus achieves the largest data scale in comparison to the existing E-commerce product datasets. The products in MEP-3M are represented in three modalities: image, textual description, and OCR text, and labeled with tree-like labels. The third level labels are extremely fine-grained. In addition, we exploit four novel practical tasks on this dataset, Product classification, Hierarchical Product Classification, Fine-grained Product Classification, and Product Representation Learning. For each task, we present some image-only, text-only, and multi-modal baseline performances for further researches. The MEP-3M dataset will be released at https://github.com/ChenDelong1999/MEP-3M.

© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

The recent rise of deep learning can be traced back to the creation of ImageNet dataset [14] and the revival of deep Convolutional Neural Network (CNN) [6,28,32,33,64]. Since then, the combination of increasingly complex neural network architectures and increasingly large datasets fundamentally revolutionized Computer Vision (CV) and Natural Language Processing (NLP) fields. In recent years, the research communities are gradually moving from these single-modal tasks to multi-modal tasks [5,11]. Large-scale multi-modal datasets, especially vision-language datasets (*e.g.* Flickr30K [60], Multi30K [16], MS-COCO [3], SBU Captions [37], WIT [45]), have been constructed and presented. These datasets enable researchers to develop multi-modal models, which learn to utilize the complementary information across different modalities and bring the opportunity to combine the advancements across different fields to further improve the model performance.

Another recent hot topic in deep learning field is fine-grained learning, which aims to discover the subtle differences between different sub-categories, such as birds [22], dogs [46], cars [58], and castles [2]. A lot of fine-grained datasets are created to promote the development of this domain, such as iNaturalist [20], Products-10k [7], and iMaterialist Fashion [17]. Notably, lots of E-commerce-related datasets have been proposed. A possible reason is the construction of this type of dataset can rely on the predefined hierarchical categorization information (*e.g.*, Stock Keeping Unit, SKU).

However, recent E-commerce datasets only focus on one aspect from multi-modal or fine-grained without integrating them together. Moreover, many E-commerce product datasets remain non-public [10,13,19,31,48,62]. In this paper, we construct a large **M**ulti-modal **E**-commerce **P**roducts classification dataset named MEP-3M, which provides multi-modal and fine-grained data. It is collected from several Chinese large E-commerce platforms and consists of over 3 million image-text pairs of products and 599 classes. Since different E-commerce platforms have different product class labeling schemes, we design a text similarity-based label alignment scheme to automatically merge the multi-source data. For each product, the corresponding image and the product title text are collected. Moreover, since a large amount of E-commerce product contains text information in image, we also extract OCR text and provide it as another modality. As demonstrated in Fig. 1, MEP-3M consists of the second largest number of products, even compared with the single-modal E-commerce product datasets.

Here we briefly summarize the key characteristics of our MEP-3M: **Large-scale**: MEP-3M dataset consists of over 3 million product samples in total. Each sample consists of an image-text pair, resulting in 3,012,959 images and 156,069,329 characters. The en-

* Corresponding author.
E-mail addresses: fanliu@hhu.edu.cn (F. Liu), chendelong@hhu.edu.cn (D. Chen), duxy@njust.edu.cn (X. Du), gao1074871898@126.com (R. Gao), xufeng@hhu.edu.cn (F. Xu).
[1] Equal Contribution

https://doi.org/10.1016/j.patcog.2023.109519
0031-3203/© 2023 Elsevier Ltd. All rights reserved.

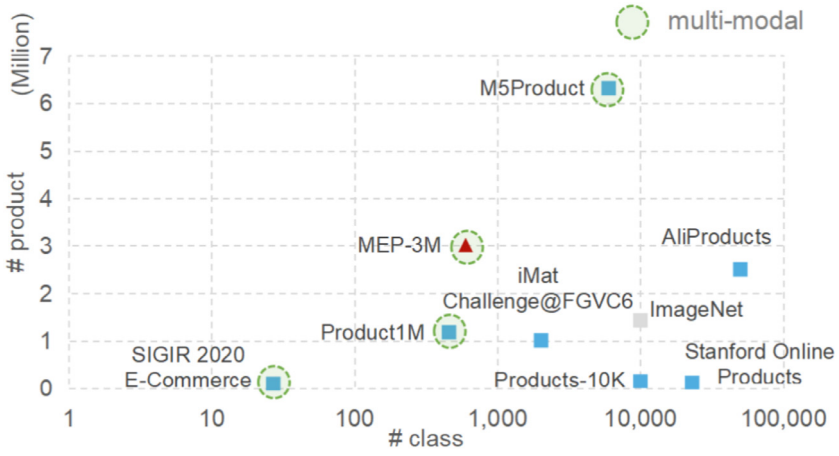


Fig. 1. The comparison between our presented dataset and existing public E-commerce product dataset.

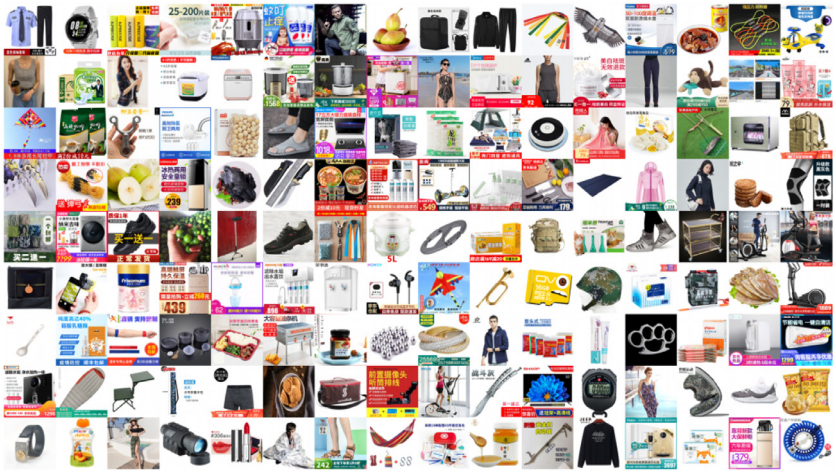


Fig. 2. Images randomly selected from the MEP-3M dataset. Our dataset covers a wide range of E-commerce products.

tire dataset takes approximately 76GB of storage. **Hierarchical-categorized**: Three levels of the label are given. There are 14 classes (first level), 599 sub-classes (second level), and 13 sub-classes have further subsub-classes (third level). **Multi-modal**: Each product has both image and Chinese label and title. Some image samples and the text cloud of the titles are given in Fig. 2 and Fig. 3. **Fine-grained**: Among a total of 599 sub-classes, many samples are visually similar but belong to different sub-classes. **Long-tailed**: MEP-3M is highly imbalanced. Some sub-classes in the dataset have more than 90k samples, while some classes have around 30 samples.

We note that the preliminary version of MEP-3M dataset has been published in the IJCAI 2021 Workshop on Long-Tailed Distribution Learning[2]. The dataset presented in this paper is an improved and extended version. In the following, we summarize the main contribution of this paper:

- We constructed a large-scale multi-modal E-commerce product dataset MEP-3M. The data are collected and merged from several E-commerce platform. MEP-3M has three modalities (*i.e.*, image, text, OCR) and three levels of labels. We present baseline results of product classification on these three modalities and three levels.
- We investigated the potential of hierarchical classification on MEP-3M and present three baselines to utilize the information from both coarse and fine labels. We pointed out that weakly-supervised hierarchical classification and multi-modal hierarchical classification are promising research direction on MEP-3M.
- We presented MEP-*meats*, MEP-*accessories*, MEP-*jewelries*, and MEP-*outdoors* as four fine-grained subsets of MEP-3M. Their uniqueness includes novel meta classes, novel image domains, and multi-modal data. Baselines for these subsets are also provided.
- We presented another special subsets named MEP-*for-RPC* and demonstrate that pre-training on it can effectively improve the models for Automatic Checkout (ACO) task.

The rest of this paper is organized as follows. We first review and compare existing E-commerce product datasets in Section 2. We describe the collection and construction process of MEP-3M in Section 3. In Section 4, Section 5, Section 6 and Section 7, we present four novel practical tasks on MEP-3M: Product classification, Hierarchical Product Classification, Fine-grained Product Classification, and Product Representation Learning. For each section, we first introduce the task and corresponding data, then present our baseline solutions and results. Section 8 concludes the paper.

[2] https://ltdl-ijcai21.github.io

Fig. 3. The text cloud (after jieba word segmentation) of product title in the MEP-3M dataset. The text size corresponds to the appearance frequency.

2. Related work

Product classification is a critical issue for an E-commerce platform since it can significantly improve the accuracy and reduce the workload of manual product category assignments. Since the product title usually aims at delivering the product information to users accurately and comprehensively as possible, text-based product classification has drawn more attention in the past years. In contrast, image data is generally harder to collect than text information, but its effectiveness is well demonstrated by a recent study [62]. Therefore, in this section, we review and compare our presented MEP-3M dataset with several E-commerce product datasets, and mainly focus on image-based ones.

In the past several years, different methods have been proposed to improve the performance of product classification, and many product datasets are collected and constructed, but unfortunately, they remained non-public [10,13,19,31,48,62]. On the other hand, there is also some public product dataset that only focuses on a limited subset of products (such as iMaterialist Fashion [17]), but classification models on this type of dataset are not applicable for general E-commerce platforms. Meanwhile, there are also some retail groceries datasets such as RPC dataset [52], but they differ from E-commerce datasets fundamentally since they are created for training automatic checkout systems. In the following, we briefly review the existing public E-commerce product datasets that aim at general products categories.

- **Stanford Online Products**[3] [43] is a E-commerce product dataset collected by a group from Stanford University using the web crawling API of eBay.com. Duplicate and irrelevant images in the dataset are filtered out. Each product in this dataset has approximately 5.3 images.
- **iMat Challenge@FGVC6**[4] is the dataset of iMaterialist Challenge on Product Recognition at FGVC6, CVPR 2019, provided by Malong Technologies and FGVC workshop. This dataset has a total number of 2,019 product categories, which are organized into a hierarchical structure with four levels.
- **SIGIR 2020 E-commerce**[5] [1] refers to the dataset used by SIGIR 2020 eCom Rakuten Data Challenge. It is a multi-modal dataset, where each sample consists of the image, the title, and the description of a product. Text information is in French.
- **AliProducts**[6] [12] is a large-scale fine-grained SKU-level E-commerce product dataset without human-labelling. It also contains side information, such as hierarchical relationships between classes.
- **Products-10K**[7] [7] is a large-scale product recognition dataset covering 10k fine-grained SKU-level products from JD.com. It contains both in-shop photos and customer images. All samples are manually checked to reduce noise.
- **Product1M**[8] [63] is a multi-modal cosmetic dataset for real-world instance-level retrieval. It consists of both single-product and multi-product samples, and each sample of the dataset contains an image-caption pair. The samples in the Product1M dataset encompass a wide variety of cosmetics brands. Some appealing characteristics of this dataset including well mimic the real-world scenes, including fine-grained categories, complex combinations, and fuzzy correspondence.
- **M5Product**[9] [15] is a large-scale multi-modal pre-training dataset with coarse and fine-grained annotation. It contains 6 million multi-modal samples and 5,000 properties with 24 million values. It is also annotated with 6,000 classes and the dataset also presents five modalities of data, including image, text, table, video and audio.

A detailed comparison of the MEP-3M dataset and the existing public E-commerce datasets is shown in Table 1. Importantly, among the above datasets, only four datasets are multi-modal, which are SIGIR 2020 E-commerce dataset, Product1M dataset, M5Product dataset and MEP-3M dataset. Compared to SIGIR 2020 E-commerce dataset, our MEP-3M dataset has much more samples and more categories. Moreover, the text of the SIGIR 2020 E-commerce dataset is in French, while our dataset is in Chinese. Since China has been the world's largest online retail market, the MEP-3M dataset may have more potential application value. Though the texts of the Product1M dataset and M5Product dataset are also in Chinese, the two datasets are quite different from our MEP-3M dataset. For example, the sample domain of the Product1M dataset is limited to cosmetics only, while the MEP-3M dataset includes a total of 599 general categories. Additionally, MEP-3M has more than three million samples, making it approximately three times larger than Product1M. Although the M5Product dataset is larger in scale, MEP-3M has several

[3] https://github.com/rksltnl/Deep-Metric-Learning-CVPR16
[4] https://www.kaggle.com/c/imaterialist-product-2019/data
[5] https://sigir-ecom.github.io/ecom2020/data-task.html
[6] https://tianchi.aliyun.com/competition/entrance/231780
[7] https://www.kaggle.com/c/products-10k
[8] https://github.com/zhanxlin/Product1M
[9] https://xiaodongsuper.github.io/M5Product_dataset/

Table 1
Comparison with existing E-commerce datasets.

Dataset	Year	#class	#image	Modality
Stanford	2016	23K	0.120M	image
iMat FGVC6	2019	2K	1.012M	image
SIGIR 2020	2020	27	0.098M	image, text (French)
AliProducts	2020	50K	2.500M	image
Products-10K	2020	10K	0.150M	image
Product1M	2021	458	1.182M	image, text (Chinese)
M5Product	2022	6K	6.313M	image, text (Chinese), table, video and audio
MEP-3M	**2021**	**599**	**3.012M**	**image, text (Chinese), OCR**

unique characteristics and advantages. For instance, the class labels in MEP-3M are organized in a hierarchical tree structure, while M5Product only has a single level of categories. Additionally, MEP-3M includes OCR as an additional modality, which is not included in the M5Product dataset. Furthermore, a MEP-*for-RPC* subset has been derived from MEP-3M to benefit downstream retail product checkout models, and we also demonstrated this potential. Finally, rather than being sourced from a single platform like M5Product, MEP-3M contains samples collected from multiple e-commerce platforms, which increases the diversity of MEP-3M samples. Moreover, the label of the E-commerce platform can be used as an auxiliary source of supervision signal. For example, it is possible to use the platform labels to learn domain-adapted product representation.

3. MEP-3M dataset

3.1. Data collection

The collected images are stored in `.jpg` or `.png` file format. We find that a large proportion of images contain texts. Therefore, besides image and text, we also extract the OCR text as another complementary modality. A text detection model from [4] and a text recognition model from [41] for OCR extraction. We extract both Simplified Chinese characters and English characters. The extracted texts are concatenated into a single line of text. From the extracted OCR texts, We find that some of them provide information that describe features of the product or have strong relationship with the product category, as shown in the top of Fig. 4. But as the bottom of Fig. 4, some other OCR text are irrelevant to the product, such as promotional information.

In the following, we give an example of an item in MEP-3M dataset annotation file.

```
{
    'class_id': '5',
    'class_name': '食品 / 酒水 / 生鲜 / 特产',
    'sub_class_id': '523',
    'sub_class_name': '水果',
    'subsub_class_id': '640',
    'subsub_class_name': '苹果',
    'img_path': 'Images/523/3.jpg',
    'img_resolution': (220, 220, 3),
    'title': '【第 2 件 9.8 ， 2 件共发带箱 10 斤】脆甜冰糖心红富士苹果 5 斤
       鲜果时令大果新鲜水果陕西洛川一整箱非烟台 5 斤装（净重 5 斤）',
    'OCR': 'FLASE'
}
```

The `class_id` denotes the first level of class label, ranging from 1 to 14. The `sub_class_id` is the second level of class label, ranging from 1 to 599. The `subsub_class_id` corresponds to the third level index, which ranges from 600 to 688. Construction of these three level of labels will be described in Section 3.2. The rest fields `img_path`, `img_resolution`, `title`, and `OCR` provide the information of product image, title text, and OCR text. For the image without OCR, the `OCR` field is set to `'FLASE'`.

3.2. Multi-source label alignment

This section introduce our proposed method of aligning multi-source labels. The label alignment is based on the analysis of the collected first-level labels (denote as 'class') and second-level labels ('sub-class'). To take the different granularity across different E-commerce platforms into account, we also create third-level labels ('subsub-class') for some of the sub-classes.

For first level, due to the number of classes are relatively small (all the platforms have less than 20 first level classes), we manually align them across different platforms. Classes with similar meanings are merged to a single class, whose new `class_name` is designated to cover the meaning of both sides. Meanwhile, unique classes are preserved as separated classes. Finally, there are a total of 14 different first level classes. A number is assigned to each class as its `class_id`. The `class_id` and the corresponding `class_name` of all the 14 first level classes are shown in Table. 2.

The number of sub-classes is far more than the first level, making manual alignment impossible. Therefore, we design an automated alignment approach based on quantitative text analysis. Specifically, the goal of the alignment is to figure out the sub-class pairs that are semantically similar across different E-commerce platforms. We assume these sub-class pairs have the following

Fig. 4. Examples of OCR information in MEP-3M dataset. **Top**: informative OCR texts. **Bottom**: less informative OCR texts.

Table 2
The Numbers of samples and sub-classes of the 14 classes.

class_id	class_name (CN)	class_name (EN)	#sample	#sub_class
1	手机/运营商/数码	Mobile phones/Digital devices	122,312	21
2	家用电器	Home appliance	240,779	51
3	电脑办公	Computers/office	85,699	17
4	家居/家具/家装/家纺/厨具	Home/decoration/kitchen	534,460	123
5	食品/酒水/生鲜/特产	Foods/drinks	411,046	53
6	美妆/个护清洁/宠物	Health care/makeup	139,049	30
7	母婴/玩具/童装	Baby care/ toys/clothes	337,425	73
8	运动/户外	Sports/outdoors	346,451	54
9	男装/女装/内衣/鞋靴	Clothes/shoes	536,842	110
10	箱包/钟表/珠宝	Luggage/jewelries	86,648	13
11	艺术/礼品鲜花/农资绿植	Art/flowers/plants	46,316	14
12	汽车/电摩/汽车用品	Cars service	66,963	21
13	图书	Books	23,208	5
14	医药保健/计生情趣	Pharmaceuticals	35,761	14

three characteristics: 1) they belong to the same first level class, 2) their names share a certain degree of similarity, 3) their title contents have similar features on term frequency. For the second and the third characteristics, we respectively calculate label similarity S_{label} and content similarity $S_{content}$ as metrics.

The label similarity S_{label} measures how far the two sub-class names coincide with each other, it is defined as:

$$S_{label} = 2.0 \times M/T \tag{1}$$

, where T indicates the total number of characters in both sub-class names, and M indicates the number of matches. Note that this is 1.0 if the sub-class names are identical, and 0.0 if they have nothing in common.

The content similarity $S_{content}$ is the cosine distance between term-frequency features extract from the title text content of two different sub-classes, it is defined as:

$$S_{content} = \frac{x_1 \cdot x_2}{\|x_1\| \times \|x_2\|} \tag{2}$$

, where x_1 and x_2 are the term-frequency feature vector of title text content. Each element in x_1 and x_2 counts the number of occurrences of a certain term.

The S_{label} are calculated by using python difflib package[10], while the $S_{content}$ is based on python simtext package[11]. In order to improve computational efficiency of $S_{content}$, we use the first 22000 characters of a sub-class product titles, corresponding to approximately 450 products. We iterate over all the sub-class pairs that belongs to the same classes, and filter them according to the following criterion:

$$S_{label} \geq 0.50 \text{ AND } S_{content} > 0.75 \tag{3}$$

The hyper-parameter of 0.50 is chosen by empirical hyper-parameter tuning. As for $S_{content}$, we first retrieve all sub-class pairs that have a $S_{label} = 1.00$ (i.e., with identical class names), and calculate the $S_{content}$ between these sub-class pairs. The averaged $S_{content}$ is 0.75, and we use this value as the threshold of $S_{content}$.

[10] https://docs.python.org/3/library/difflib.html
[11] https://pypi.org/project/simtext

Table 3
Examples of second-level label alignment.

sub-class name	sub-class name	S_{label}	$S_{content}$	new class name (CN)	new class name (EN)
儿童餐具	儿童餐具	1.00	0.929	儿童餐具	Children's tableware
孕妈奶粉	孕妈奶粉	1.00	0.898	孕妈奶粉	Milk powder during pregnancy
空调	空调	1.00	0.870	空调	Air conditioner
骑行装备	骑行装备	1.00	0.768	骑行装备	Cycling equipments
洗澡用具	洗澡用具	1.00	0.705	洗澡用具	Bath supplies
婴幼奶粉	婴幼儿奶粉	0.89	0.960	婴幼儿奶粉	Infant milk powder
婴儿湿巾	湿巾	0.67	0.907	湿巾	Wipes
咖啡/奶茶	咖啡	0.57	0.854	咖啡/奶茶	Coffee/milk tea
饮料	饮料饮品	0.67	0.849	饮料饮品	Beverages
办公文具	办公文仪	0.75	0.756	办公文仪	Office stationery

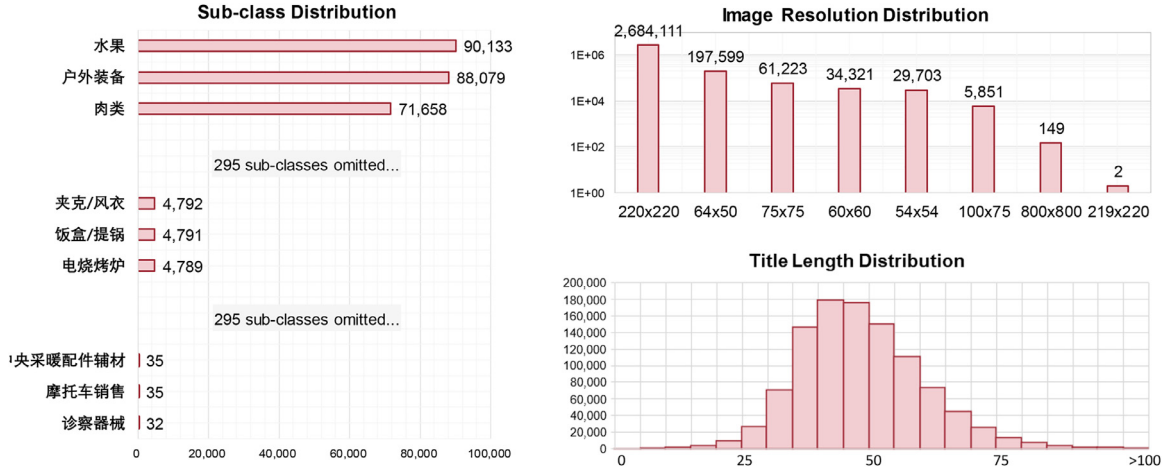


Fig. 5. The distribution of sub-classes, image resolution, and title length in MEP-3M dataset.

New names are manually assigned for those sub-class pairs that $S_{label} \neq 1.00$. Some examples of the results are listed in Table. 3.

Some platforms have further finer-grained categories. Therefore, beyond the class and the sub-class labels, we create finer-grained subsub-class labels for a total of 13 sub-classes:

'箱包' (bags), '饰品' (accessories), '手机配件' (mobile phone accessories), '男装' (men's clothing), '女装' (women's clothing), '内衣' (underwear), '户外装备' (outdoor equipment), '水果' (fruit), '肉类' (meat), '冲调饮品' (toned drinks), '南北干货' (dry foods), '纸尿裤' (diapers), and '奶瓶奶嘴' (bottle nipples).

For the sub-class that does not have finer-grained subsub-classes, the `subsub_class_id` and `subsub_class_name` are set to `'FLASE'`.

3.3. Dataset statistics

Most images are in a 220×220 resolution, and the others are in 64×50, 75×75, 60×60, 54×54, 100×75, 800×800 and 219×220 resolution. A total of 2,908,596 (96.53%) of the images are in `.jpg` format, while the other 104,363 (3.46%) images are in `.png` format.

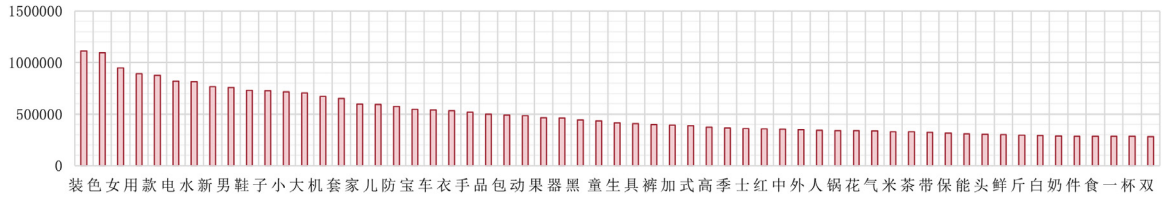
The text of `title` is in simplified Chinese. The length of `title` ranges from 2 characters to more than 100 characters. The average length of it is 49. In total, the dataset consists of 156,069,329 characters in title overall. OCR text is detected in a total of 1,404,146 images (46.6% of all products). Fig. 6 shows the statistics of character frequency in title and OCR text.

Each level of labels are provided in both Chinese and English. The entire dataset takes around 76 GB storage. See Fig 5 for the distribution of sub-classes, image resolution, and title length in MEP-3M dataset.

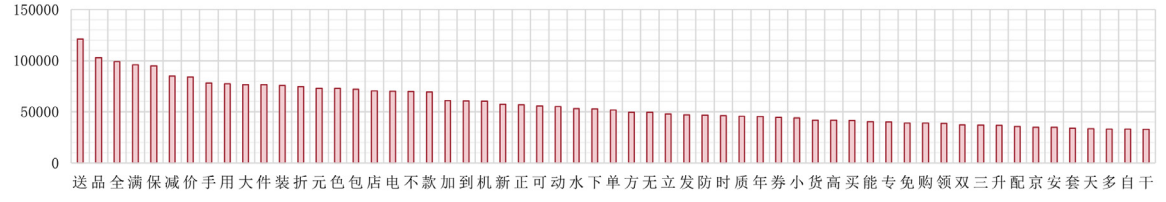
4. Product classification

4.1. Data and task

Product classification is the most straightforward task on MEP-3M dataset. The product classification task can be done on all three levels of label in MEP-3M. We defined the second level (599-way) as the predominant setting. Since MEP-3M provides multi-modal data, the model can also be trained with different modalities of data or their combination. Specifically, we defined image-only setting, text-only setting, and multi-modal setting, where the model can respectively access image data only, text data only, or both of them. The three levels of labels and three settings of data modality result in nine classification tasks. In the following, we first conduct comparative experiments on the predominant task (the sec-



(a) Character frequency of title text



(b) Character frequency of OCR text

Fig. 6. Character frequency analysis of text from title and OCR. Comparatively, text in the title is more informative, while text OCR contains more promotional-related information.

Table 4
The classification accuracies of different baseline methods on MEP-3M.

Model	Data	Top-1	Top-5	AP	F1-score
VGG-19	image-only	76.36%	91.77%	0.3966	0.7275
Inception-v3	image-only	79.48%	94.27%	0.6326	0.7493
LSTM	text-only	89.13%	98.33%	0.9200	0.8796
Bi-LSTM	text-only	90.68%	98.70%	0.9309	0.8931
VGG+BiLSTM	multi-modal	**91.07%**	**98.84%**	**0.9354**	0.8540
TFN	multi-modal	90.70%	98.74%	0.9289	0.8899
LMF	multi-modal	89.22%	98.19%	0.8924	**0.9125**

ond level) to demonstrate the effectiveness of multi-modal data, then compare the difficulty of the three levels.

4.2. Baseline results

For the classification on the second level, we test both single modal models (*i.e.*, LSTM for text-only, VGG-19 [42] and Inception-V3 [47] for image only) and multi-modal models (Low-rank Multi-modal Fusion (LMF) [34] and Tensor Fusion Network (TFN) [61]). For LSTM-based text-only model, we first remove meaningless characters from texts with regular expressions and then implement Chinese word segmentation. Word2vec model from the gensim toolkit is used to obtain word embedding. The representation is further passed to the LSTM or BiLSTM model for classification. We divide the full dataset randomly into training and test set at a ratio of 8:2. The model is implemented by TensorFlow, using an Intel i5-9400F CPU and NVIDIA TITAN RTX GPU. All experiments are trained with Adam optimizer, and the initial learning rate is set to 1e-3 and decreases every 2 epochs at a rate of 0.5. The batch size is 64.

The testing accuracies, average precision score (AP) and F1-score of baseline models are shown in Table 4. We can see that the multi-modal methods (VGG+BiLSTM, LMF [34] and TFN [61]) achieved better results than single-modal methods, which demonstrates the advantage of multi-modal product classification over single-modal-based methods. The results also show that the text-based classification is much easier than the image-based one and reveal the potential of utilizing text information as weak annotations to improve the vision models.

Table 5
The classification accuracies of different level of class labels on MEP-3M.

Model	First-level	Second-level	Third-level
ResNet	88.81%	71.37%	71.15%
BERT	98.76%	94.29%	94.10%

Then we compare the classification tasks on different label level. For image-only setting, we use a ResNet-50 as feature extractor. For text-only setting, we use a pre-trained BERT model. For each of them, we use two 4096-d fully connected layers as the classifier. The results are shown in Table 5. It can be seen that first level classification is much easier than the second and the third level. Comparing the results of ResNet and BERT, we find a large performance gap between them. It indicates that image-only classification task is much harder than text-based ones. It is partially due to the fact that the images of MEP-3M are fine-grained. We select the nearest neighbors of several images to demonstrate this point. The clustering is done by calculating the pixel-wise distance. As shown in Fig. 8, many images are visually similar but belong to different classes.

Except the fine-grained nature of MEP-3M, the lone-tailed distribution may also brings additional challenge. To verify this point, we evaluate the per-class accuracy of the ResNet trained on third-level labels, and analyze the relationship between the number of each class's training sample and its corresponding testing accuracy. The results are shown in Fig. 8. We find that with the number of training sample decrease, the lower bound of per-class classification accuracy drop sharply. The lowest accuracy approaches 14.29%. On the opposite direction, we find that 6000+ training samples could yield satisfying per-class accuracy.

5. Hierarchical product classification

5.1. Data and task

Hierarchical classification [25,51], or coarse-and-fine learning [9], is an active research area in the machine learning field [49,57]. It shares many similarities with fine-grained learning, such as they

Fig. 7. Two groups of visually similar images that belongs to different classes.

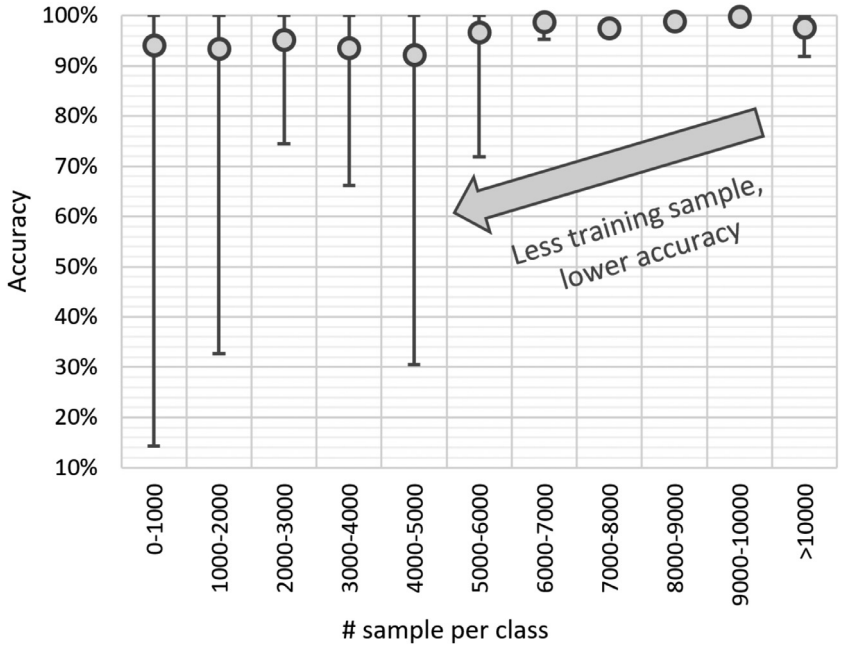


Fig. 8. The relationship between the number of sample per class and the per-class accuracy.

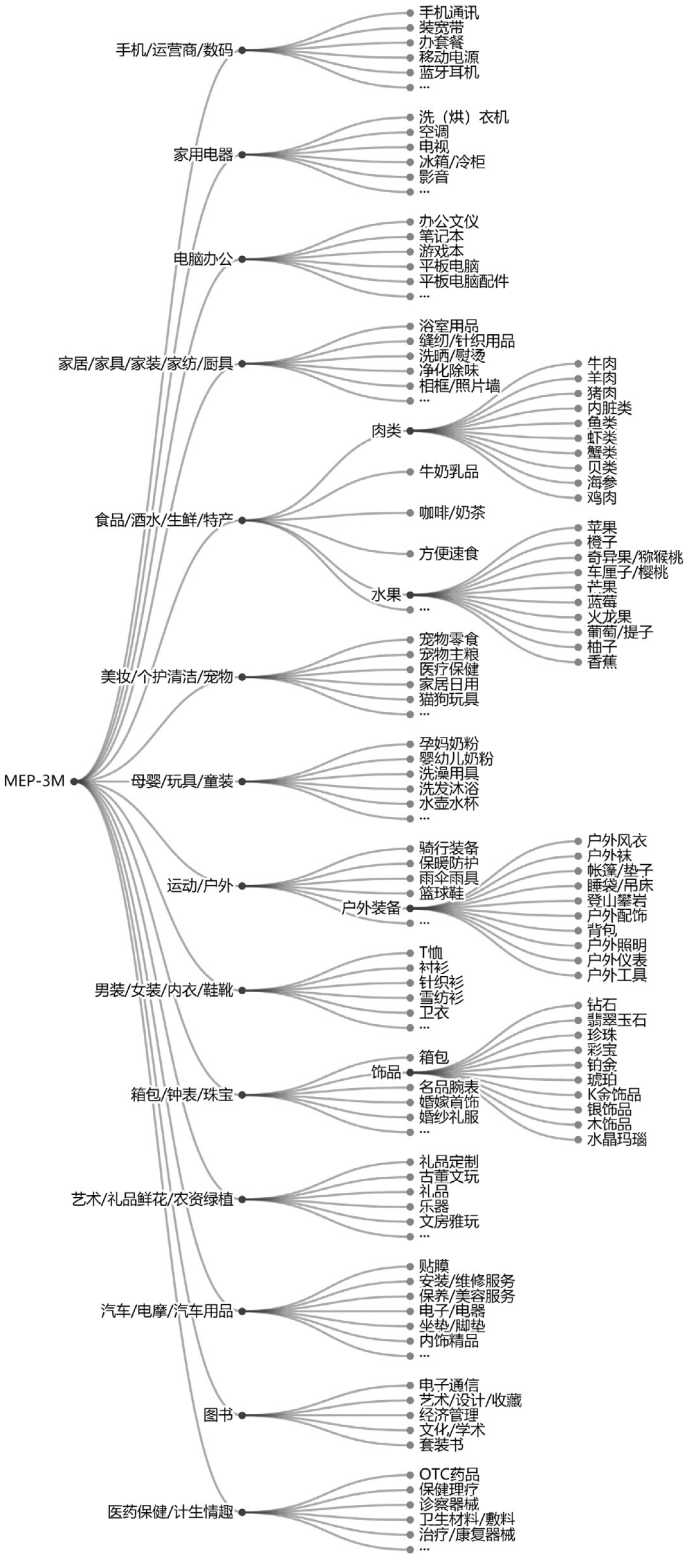


Fig. 9. Illustration of the hierarchical structure of MEP-3M. Note that this figure only shows about 10% of the entire MEP-3M.

both require a model to recognize the subtle differences between classes. The key difference between them lies in the fact that fine-grained learning usually has only one level of labels, but hierarchical classification covers multiple levels of labels. One hierarchical model should recognize both coarse classes and finer sub-classes at the same time. One distinguished feature of our MEP-3M dataset is that it is hierarchically categorized. As shown in Fig. 9, there are three levels of the label in MEP-3M, including 14 classes (first level), 599 sub-classes (second level), and 13 sub-classes have further subsub-classes (third level). These hierarchical labels MEP-3M make it a suitable dataset for experiments of hierarchical classification algorithms. Compared to the CIFAR-100 dataset or ImageNet dataset that has been commonly used by well-known hierarchical classification models, our MEP-3M dataset provides three levels of labels, E-commerce product domain images, and more importantly, multi-modal data. Based on the characteristics of its hierarchical annotations, we point out two possible tasks on MEP-3M for hierarchical classification studies:

- **Weakly-supervised hierarchical classification.** Current hierarchical classification tasks can be divided into two types. The first is all samples are labeled with both coarse-and-fine labels, while in the second type only part of the samples has fine labels. Since coarse labels are easier to collect than fine labels, the second type is much closer to a real-world scenario. Our MEP-3M falls in this type. For example, one E-commerce platform may end up with “fruits” class, while the other may further classify the “fruits” into “apples”,“bananas”, “pitches”, etc. In this case, the samples from the first platform have only coarse labels, they can be served as weak supervision that assist finer-grained classes learning [24,29,40].
- **Multi-modal hierarchical classification.** It should be noticed that most existing hierarchical classification studies are in single-modal settings. One likely reason is that there are not many multi-modal hierarchical classification datasets. With MEP-3M, we propose a novel learning task: multi-modal hierarchical classification. Compared to traditional hierarchical classification, multi-modal setting brings additional challenge that the model should simultaneously learn different granularity *within* (*e.g.*, “fruits” v.s. “apples”, “bananas”, “pitches”) and *across* modalities (*e.g.*, a samples of “apples” class may contain text like “fresh fruit”).

5.2. Baseline results

We present three baseline approaches that aim to make full use of the hierarchical annotations of MEP-3M. Specifically, our goal is to improve classification accuracy with finer-grained information. For example, for the baseline of first-level classification on MEP-3M in Section 4 (Table 5), only the corresponding 14-way labels are available. Such results might be further improved when more information is available (*i.e.* with 599-way labels and 688-way labels). Inspired by Guo *et al.* [18], we design three baseline models. Their network structures are shown in Fig. 10. We combine them with the three baselines in Section 4 by attaching them to the features extracted by ResNet, BERT, and ResNet+BERT. The features are passed to a series of dense layers with skip connections. Three classifiers corresponding to the three levels in the category hierarchy are connected with dense layers in different orders. The network is trained with a joint loss composed of cross-entropy losses of the three classifiers. During training, the backbone network (ResNet/BERT) is fixed as in [18].

The results are shown in Table 6. We note accuracy improvements in parentheses. For most settings, the classification accuracies got improved, which shows the effectiveness of hierarchical information. As shown in the ”Average improvement” row in Table 6, all of the three hierarchical classification algorithms bring improvements on average, and the “coarse to fine” setting with +0.71% improvement is the best one by comparison. In addition, the results also illustrate that hierarchical information is more beneficial to image model (ResNet) compared to text model (BERT). The best improvement (+1.20%) in this table is introduced by the image-only coarse to fine model, and both of the remaining hierarchical clas-

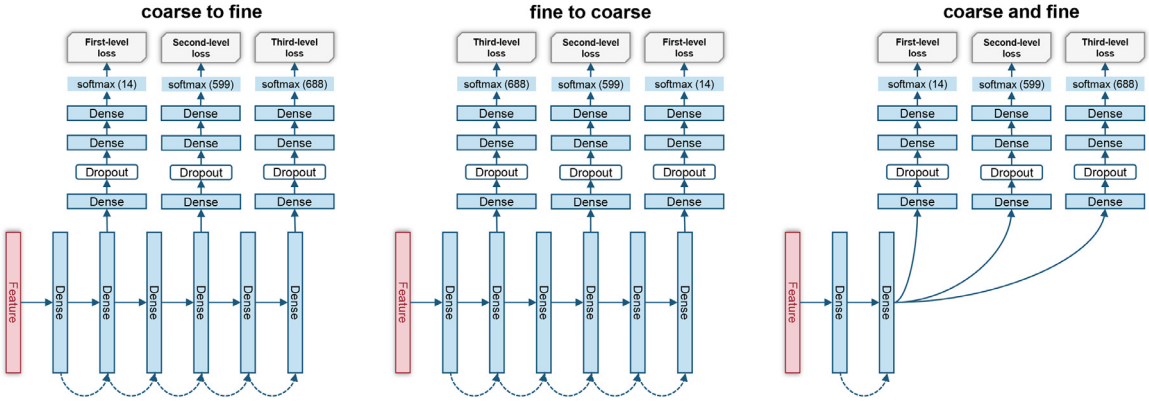


Fig. 10. Network architectures of proposed baselines for hierarchical classifications. The dotted lines represents skip connections.

Table 6
Hierarchical classification results.

Model	Baseline	coarse to fine	fine to coarse	coarse and fine
ResNet	88.81%	90.01% (+1.20%)	89.50% (+0.69%)	89.73% (+0.92%)
BERT	98.76%	98.58% (- 0.18%)	98.61% (- 0.15%)	98.61% (- 0.15%)
ResNet + BERT	97.70%	98.81% (+1.11%)	98.81% (+1.11%)	98.81% (+1.11%)
Average improvement	-	+0.71%	+0.55%	+0.63%

sification algorithms on ResNet also yield improvements (+0.69% and +0.92%). However, for BERT-based models, the hierarchical information fails to improve the performance. This may be due to the fact that text classification is generally easier than image classification and the naive baseline for text classification already performs very well at 98.76%

6. Fine-grained product classification

6.1. Data and task

In recent years, with deep learning technologies, researchers in the computer vision community have made remarkable progress in Fine-Grained Image Analysis (FGIA) tasks [38,53,65]. However, FGIA still remains an challenging and unsolved problem due to the small inter-class variation and the large intra-class variation. Our MEP-3M is collected from several E-commerce platforms. The label granularity is different across different platforms. On one hand, this fact brings additional need for label alignment when merge multi-source data into a single dataset (as described in Section 3). On the other hand, we can naturally acquire several groups of high-quality coarse-and-fine label annotations.

Based on this, we select four sub-classes and accordingly build up four fine-grained subsets from MEP-3M. The selected sub-classes (may also be named as “meta-classes” or “super-classes”) include meat, mobile phone accessory, jewelry, and outdoor equipment. The resulting subsets are respectively named as MEP-*outdoors*, MEP-*meats*, MEP-*accessories* and MEP-*jewelries*. Among them, MEP-*outdoors* is divided into 17 categories including backpack, climbing equipment, sleeping bag and so on, and MEP-*meats* contains 18 categories including chickens, ducks, fishes, etc. The MEP-*accessories* is made up of 7 categories, such as mobile phone holder, cell phone battery, earphones and so on, and MEP-*jewelries* involves 12 categories including diamond, jade, etc. In average, each category in these datasets has 4000-5000 samples. Table 7 gives an detailed comparison between presented four MEP-3M fine-grained subsets and existing fine-grained datasets. Our datasets enjoy four unique characteristics compared to existing fine-grained datasets, as listed in the following:

- **Novel meta-class.** Existing fine-grained datasets mainly cover different types of animals (birds [8,50], dogs [26]), plants (flowers [36], vegetable [23], fruits [23]), vehicles (cars [27], planes [35]). A few of existing datasets involves products, such as clothes [17], retail products [52], etc. Comparatively, our presented fine-grained subsets of MEP-3M are with very different meta-classes. To our best knowledge, there are not any existing fine-grained image datasets that specifically focus on meat, mobile phone accessory, jewelry, or outdoor equipment.
- **Novel image domain.**. The images in most existing fine-grained datasets are collected in the wild [21] or in a controlled environment [52], but the images in our fine-grained subsets of MEP-3M are E-commerce product images, which usually have cleaner backgrounds and sometimes contain OCR information. Moreover, although our datasets are not annotated with the bounding box, it should be noticed that E-commerce product images are implicitly annotated. The seller (who uploads the product image) tends to place the product in the center of the image and adjust it to a proper size. This characteristic is unique compared with most existing datasets that contains photos in the wild, and leads to the potential of applying self-supervised learning technologies
- **Multi-modality.** Recently, the combination of multi-modal learning and FGIA attracts much attention [39,44,56]. Most studies in this direction use the CUB-200-2011 dataset and Oxford Flowers dataset, in which the text modality is collected by Reed et al. [39].. Specifically, the text is annotated through Amazon Mechanical Turk (AMT) platform by “non-Master” certified workers. Our fine-grained subsets differ from theirs in two aspects. First, the text description of our datasets is in Chinese instead of English. Second, our text description is much more accurate and informative since it is given by the product seller. With such text data, the MEP-3M fine-grained subsets can also be considered as multi-modal fine-grained datasets with complementary vision-language information and high-quality text annotations.
- **High diversity.** As shown in Fig. 11, we sampled some images from our four fine-grained datasets. Many of the images have very small inter-class variance (as noted by green dotted box), and at the same time, the intra-class variance (horizontal axis) is large. In other words, these fine-grained subsets provide im-

Table 7
Comparison between our presented MEP-3M fine-grained subsets and widely used fine-grained datasets.

Dataset		Meta-class	# Images	# Categories	Modaility
Oxford Flower	[36]	Flowers	8,189	102	Image/Text (English)
CUB200-2011	[50]	Birds	11,788	200	Image/Text (English)
Stanford Dog	[26]	Dogs	20,580	120	Image
Stanford Car	[27]	Cars	16,185	196	Image
FGVC Aircraft	[35]	Aircrafts	10,000	100	Image
Birdsnap	[8]	Birds	49,829	500	Image
Fru92	[23]	Fruits	69,614	92	Image
Veg200	[23]	Vegetable	91,117	200	Image
iNat2017	[21]	Plants and Animals	859,000	5,089	Image
RPC	[52]	Retail products	83,739	200	Image
MEP-*outdoors*	Ours	Outdoor equipment	82,187	17	Image/Text (Chinese)/OCR
MEP-*meats*		Meats	68,102	18	Image/Text (Chinese)/OCR
MEP-*accessories*		Mobile phone accessories	34,168	7	Image/Text (Chinese)/OCR
MEP-*jewleries*		Jewelries	28,193	12	Image/Text (Chinese)/OCR

Table 8
The classification accuracies of different baseline methods on MEP-3M fine-grained subsets.

	MEP-*accessories*	MEP-*jewelries*	MEP-*outdoors*	MEP-*meats*
ResNet	90.95%	64.88%	69.39%	64.76%
BERT	99.27%	100%	93.44%	81.66%

ages with high diversity. In addition, in existing fine-grained learning datasets, one category normally has only around one hundred samples. In comparison, MEP-3M fine-grained subsets have up to several thousand images per category. These images carry very different patterns and result in a very high diversity, which brings challenges for visual models. Such diversity, combined with other three characteristics as mentioned before, will make these fine-grained subsets valuable for researching.

6.2. Baseline results

We also present baseline results to benchmark MEP-3M fine-grained subsets. As before, an image-only ResNet and a text-only BERT are employed to perform fine-grained classification. The results are shown in Table 8. Similar to the baseline results of MEP-3M (*i.e.*, Table 4), there is a large gap between the performance of the image-only model and the text-only model. In addition, we noticed that the baseline performance of the four subsets differs significantly from each other. There are mainly two reasons for this. Firstly, the number of categories varies across the datasets, leading to varying levels of difficulty in learning a classification model. Secondly, the samples in the MEP-3M dataset are diverse because they are collected from different platforms, and the qualities of samples in different categories and modalities may also be different. Based on the results, the four subsets can be described as: strong image and strong text (MEP-*accessories*), weak image and strong text (MEP-*jewelries*), and weak image and weak text (MEP-*outdoors* and MEP-*meats*). This classification of strong and weak is not only evident in the experimental results, but can also be seen in the dataset samples. For example, as shown in Fig. 11, the image samples in the MEP-accessories subset are generally easier to recognize because different categories have significantly different visual patterns. As a result, the ResNet-based classification model achieved a high accuracy (90.95%) on MEP-accessories. In contrast, in the other three datasets, the ResNet-based model only achieved accuracies of around 60% to 70% because the images have smaller inter-class differences (especially for the images in green dotted boxes in Fig. 11). As for the BERT-based text classification models, the accuracies in MEP-accessories and MEP-jewelries are very high (99.27% and 100%) because many product titles contain the exact class names. However, in MEP-outdoors and MEP-meats, the titles are generally shorter and contain more irrelevant information, resulting in lower accuracies.

7. Product representation learning

7.1. Data and task

In this section, we propose another special subset and a practical usage of MEP-3M. We present MEP-*for-RPC*: as a pre-training dataset for the Automatic Checkout (ACO) task. ACO is a novel computer vision task, which refers to recognize and count products from a given image. ACO has high research value since it has the potential to reduce human labor amount in the retail industry [54]. However, such task is particularly challenging since a model needs to recognize subtle differences between a large number of products. Moreover, due to the rapid updating of the products, it is desirable to perform online learning to avoid frequent re-training. In 2019, Wei *et al.* [52] introduced a high-quality dataset for the ACO problem named Retail Product Checkout (RPC) dataset. Compared to previous relevant datasets, RPC has a more clearly defined ACO task setting. It is also significantly larger and closer to real-world application scenarios. Existing solutions [30,52,55,59] on RPC dataset mainly used GAN-based data augmentation [66] to synthesize training images from single-product exemplar images. Although the number of samples can be significantly increased, the diversity of synthesized training data is limited. As a result, the robustness and generalization ability of learned feature representation of an ACO model is limited, which is also undesirable in the online learning scenario.

Therefore, we propose to use the MEP-3M dataset to improve the feature for ACO task on RPC. We select a special subset of MEP-3M named MEP-*for-RPC* to exclude unrelated samples in MEP-3M. Specifically, 26 sub-classes in MEP-3M that have semantic overlap with RPC are selected. Due to difference in granularity, there is a many-to-many correspondence between RPC dataset and MEP-*for-RPC* dataset. For example, as shown in Fig. 12, the "personal hygiene" (id=15) in RPC dataset corresponds to four different subclasses in MEP-*for-RPC*, while "candy/choclate" (id=304) in MEP-*for-RPC* dataset covers three classes in RPC dataset. Our MEP-*for-RPC* dataset has a total of 118,170 different products, which is 500 times larger than the RPC dataset. As a result, as shown in Fig. 12, the samples in MEP-*for-RPC* have significantly larger variance than RPC. Such variance is beneficial for learning robust product representation and improving ACO model.

7.2. Baseline results

To demonstrate the superiority of the pre-training on MEP-*for-RPC*, we conduct a comparative experiment simulating a real-world incremental online learning scenario. We first divide the training

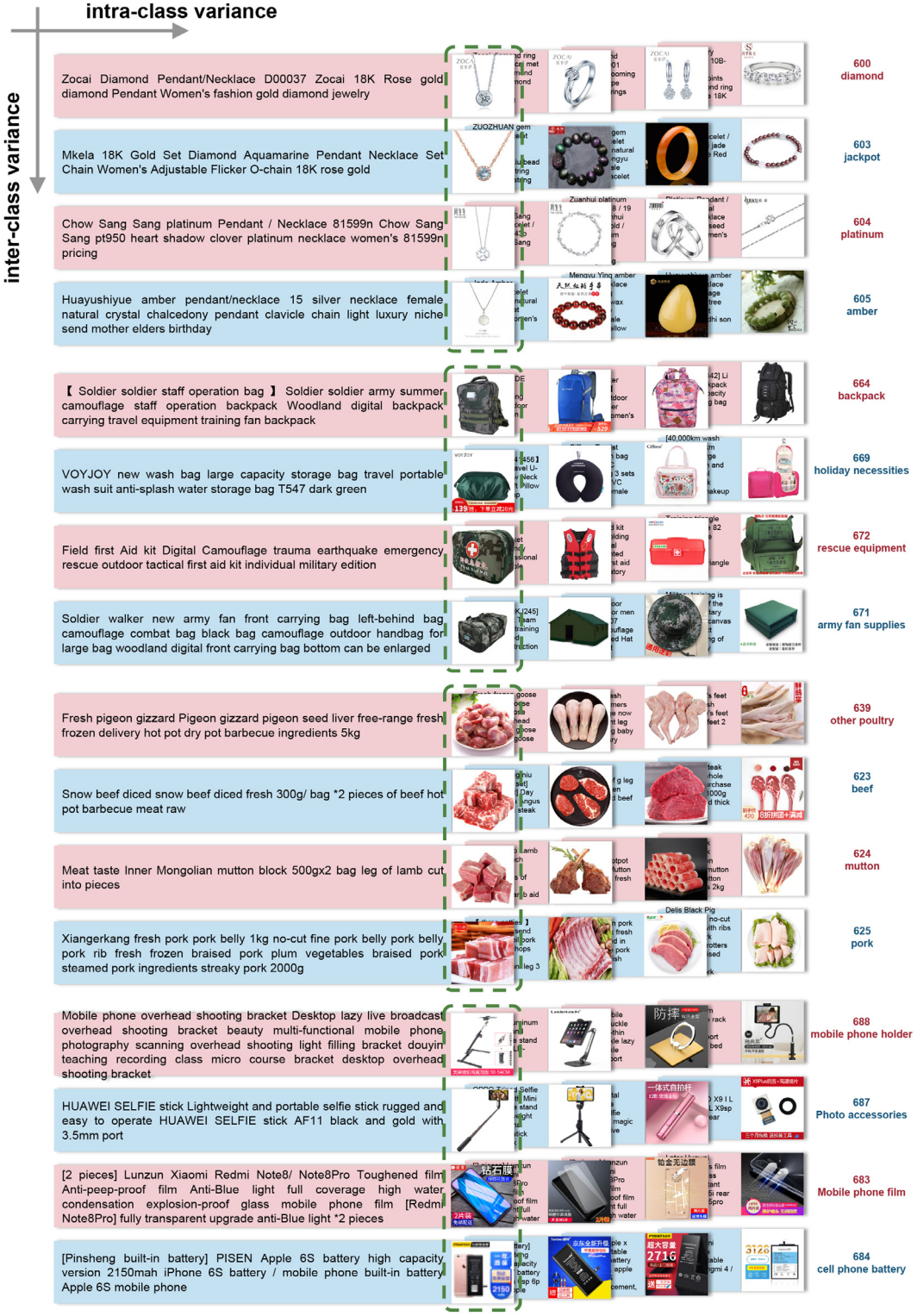


Fig. 11. Examples of presented MEP-3M fine-grained subsets. From top to bottom: MEP-*jewelries*, MEP-*outdoors*, MEP-*meats*, MEP-*accessories*. The key challenge on our datasets lies in the small inter-class variance (as noted in green dotted boxes) and large intra-class variance (horizon axis).

set of RPC into base classes and novel classes. The novel class represents the newly updated products. From each of 17 meta classes in RPC dataset we select two products as novel class, and leave the rest classes as base class. It splits a total of 200 classes in RPC into 166 base classes and 34 novel classes. We randomly select 10 images for each class as training data. We conduct experiment to validate the effectiveness of pre-training on MEP-*for-RPC*. A random initialized ResNet is firstly trained with 166 base classes only, and then transferred to classify 200 base + novel classes. The comparison of with pre-training has the same transferring pipeline, except the ResNet is initialized with weights pre-trained on MEP-*for-RPC*. The results are shown in Table 9. We present the final accuracy of all classes (166 base classes + 34 novel classes) and the accuracy of only the novel class. It can be seen that the represen-

Fig. 12. Class correspondence between RPC dataset and MEP-*for-RPC* dataset.

Table 9
Comparison of MEP-*for-RPC* pre-training

Pre-training on MEP-*for-RPC*	Accuracy-all	Accuracy-novel
×	59.63%	52.05%
√	64.40%	64.01%

tation learned from MEP-*for-RPC* pre-training effectively improves the performance of retail product classification. Moreover, the accuracy of novel class is lower than overall accuracy without pre-training, while the accuracy-all and accuracy-novel are almost the same under the pre-training setting. It indicates that the feature representation learned from MEP-*for-RPC* is more robust than the feature learned from RPC only.

8. Conclusion

This paper introduced MEP-3M, a large-scale multi-modal E-commerce product dataset, which is unique in terms of its large-scale, multi-modality, hierarchical and fine-grained categorization, and long-tailed distribution. One of the key strengths of MEP-3M is its alignment with recent progress in the field of vision-language research, which makes it a valuable resource for researchers to explore the challenges and opportunities of multi-modal learning in the context of E-commerce products. For instance, with its large-scale data, MEP-3M can serve as a powerful pre-training dataset for E-commerce vision-language foundation models.

However, the MEP-3M dataset also has several limitations that need to be considered. First, although the dataset covers 599 fine-grained product categories, it may still not include all the possible product categories in the E-commerce domain. The reason is that the products of E-commerce platforms are constantly updating, making it difficult to keep all the categories up-to-date. This can limit the generalizability of the models trained on MEP-3M to real-world E-commerce scenarios. Second, the data quality of MEP-3M may also affect the performance of the models. For example, some images or textual descriptions may contain noise, which can negatively impact the learning process.

Despite these limitations of MEP-3M, we believe that it has the potential to advance the field of vision-language research, and make a significant impact on E-commerce related research. In the future, we plan to continuously expand the dataset and improve its quality by label denoising and image enhancement. Additionally, we will further extend the MEP-3M dataset to more applications, such as cross-modal retrieval and product clustering, to meet the growing needs of the research community. We hope to see more researchers using MEP-3M to explore new ideas and develop more advanced models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was partially supported by Key Laboratory of Information System Requirements, No: LHZZ 2021-M04, Research Fund from Science and Technology on Underwater Vehicle Technology Laboratory (2021JCJQ-SYSJJ-LB06905), Water Science and Technology Project of Jiangsu Province under grant No. 2021072, 2021063.

References

[1] H. Amoualian, SIGIR 2020 e-commerce workshop data challenge overview, 2020.

[2] C.E. Anderson, A. Farrell, B. Young, Have fun storming the castle(s)!, WACV, 2021.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: visual question answering, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 2425–2433, doi:10.1109/ICCV.2015.279.

[4] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, Character region awareness for text detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 9365–9374, doi:10.1109/CVPR.2019.00959.

[5] J. Bai, C. Liu, F. Ni, H. Wang, M. Hu, X. Guo, L. Cheng, Lat: latent translation with cycle-consistency for video-text retrieval, CoRR abs/2207.04858 (2022), doi:10.48550/arXiv.2207.04858.

[6] S. Bai, F. Zhang, P.H.S. Torr, Hypergraph convolution and hypergraph attention, Pattern Recognit. 110 (2021) 107637, doi:10.1016/j.patcog.2020.107637.

[7] Y. Bai, Y. Chen, W. Yu, L. Wang, W. Zhang, Products-10K: a large-scale product recognition dataset, CoRR abs/2008.10545 (2020). https://arxiv.org/abs/2008.10545

[8] T. Berg, J. Liu, S.W. Lee, M.L. Alexander, D.W. Jacobs, P.N. Belhumeur, Birdsnap: large-scale fine-grained visual categorization of birds, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014, pp. 2019–2026, doi:10.1109/CVPR.2014.259.

[9] G. Bukchin, E. Schwartz, K. Saenko, O. Shahar, R. Feris, R. Giryes, L. Karlinsky, Fine-grained angular contrastive learning with coarse labels, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 8730–8740. https://openaccess.thecvf.com/content/CVPR2021/html/Bukchin_Fine-Grained_Angular_Contrastive_Learning_With_Coarse_Labels_CVPR_2021_paper.html

[10] Z. Cao, S. Mu, M. Dong, Two-attribute e-commerce image classification based on a convolutional neural network, Vis. Comput. 36 (8) (2020) 1619–1634, doi:10.1007/s00371-019-01763-x.

[11] D. Chen, Z. Wu, F. Liu, Z. Yang, Y. Huang, Y. Bao, E. Zhou, Prototypical contrastive language image pretraining, CoRR abs/2206.10996 (2022), doi:10.48550/arXiv.2206.10996.

[12] L. Cheng, X. Zhou, L. Zhao, D. Li, H. Shang, Y. Zheng, P. Pan, Y. Xu, Weakly supervised learning with side information for noisy labeled images, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 306–321.

[13] J. Dai, T. Wang, S. Wang, A deep forest method for classifying e-commerce products by using title information, in: International Conference on Computing, Networking and Communications, ICNC 2020, Big Island, HI, USA, February 17-20, 2020, IEEE, 2020, pp. 1–5, doi:10.1109/ICNC47757.2020.9049751.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society, 2009, pp. 248–255, doi:10.1109/CVPR.2009.5206848.

[15] X. Dong, X. Zhan, Y. Wu, Y. Wei, M.C. Kampffmeyer, X. Wei, M. Lu, Y. Wang, X. Liang, M5product: self-harmonized contrastive learning for e-commercial multi-modal pretraining, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21220–21230, doi:10.1109/CVPR52688.2022.02057.

[16] D. Elliott, S. Frank, K. Sima'an, L. Specia, Multi30k: multilingual english-german image descriptions, in: Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany, The Association for Computer Linguistics, 2016, doi:10.18653/v1/w16-3210.

[17] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M.R. Scott, S.J. Belongie, The imaterialist fashion attribute dataset, in: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, IEEE, 2019, pp. 3113–3116, doi:10.1109/ICCVW.2019.00377.

[18] Y. Guo, Y. Liu, E.M. Bakker, Y. Guo, M.S. Lew, CNN-RNN: a large-scale hierarchical image classification framework, Multim. Tools Appl. 77 (8) (2018) 10251–10271, doi:10.1007/s11042-017-5443-x.

[19] V. Gupta, H. Karnick, A. Bansal, P. Jhala, Product classification in e-commerce using distributional semantics, in: N. Calzolari, Y. Matsumoto, R. Prasad (Eds.), COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, ACL, 2016, pp. 536–546. https://www.aclweb.org/anthology/C16-1052/

[20] G.V. Horn, O.M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S.J. Belongie, The inaturalist species classification and detection dataset, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 8769–8778, doi:10.1109/CVPR.2018.00914.

[21] G.V. Horn, O.M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S.J. Belongie, The inaturalist species classification and detection dataset, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 8769–8778, doi:10.1109/CVPR.2018.00914.

[22] G.V. Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, S.J. Belongie, Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 595–604, doi:10.1109/CVPR.2015.7298658.

[23] S. Hou, Y. Feng, Z. Wang, Vegfru: a domain-specific dataset for fine-grained visual categorization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 541–549, doi:10.1109/ICCV.2017.66.

[24] Q. Jiao, Z. Liu, L. Ye, Y. Wang, Weakly labeled fine-grained classification with hierarchy relationship of fine and coarse labels, J. Vis. Commun. Image Represent. 63 (2019), doi:10.1016/j.jvcir.2019.102584.

[25] C.N. Silla Jr, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (1-2) (2011) 31–72, doi:10.1007/s10618-010-0175-9.

[26] A. Khosla, N. Jayadevaprakash, B. Yao, F.-F. Li, Novel dataset for fine-grained image categorization: Stanford dogs, in: Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), volume 2, Citeseer, 2011.

[27] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3d object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013, IEEE Computer Society, 2013, pp. 554–561, doi:10.1109/ICCVW.2013.77.

[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: P.L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, 2012, pp. 1106–1114. https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[29] J. Lei, Z. Guo, Y. Wang, Weakly supervised image classification with coarse and fine labels, in: 14th Conference on Computer and Robot Vision, CRV 2017, Edmonton, AB, Canada, May 16-19, 2017, IEEE Computer Society, 2017, pp. 240–247, doi:10.1109/CRV.2017.21.

[30] C. Li, D. Du, L. Zhang, T. Luo, Y. Wu, Q. Tian, L. Wen, S. Lyu, Data priming network for automatic check-out, in: L. Amsaleg, B. Huet, M.A. Larson, G. Gravier, H. Hung, C. Ngo, W.T. Ooi (Eds.), Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, ACM, 2019, pp. 2152–2160, doi:10.1145/3343031.3350989.

[31] G. Li, N. Li, Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network, Electron. Commer. Res. 19 (4) (2019) 779–800, doi:10.1007/s10660-019-09334-x.

[32] Y. Li, H. Zhou, Y. Yin, J. Gao, Multi-label pattern image retrieval via attention mechanism driven graph convolutional network, in: H.T. Shen, Y. Zhuang, J.R. Smith, Y. Yang, P. César, F. Metze, B. Prabhakaran (Eds.), MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, ACM, 2021, pp. 300–308, doi:10.1145/3474085.3475695.

[33] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–21, doi:10.1109/TNNLS.2021.3084827.

[34] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, Association for Computational Linguistics, 2018, pp. 2247–2256, doi:10.18653/v1/P18-1209.

[35] S. Maji, E. Rahtu, J. Kannala, M.B. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, CoRR abs/1306.5151 (2013). http://arxiv.org/abs/1306.5151

[36] M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008, IEEE Computer Society, 2008, pp. 722–729, doi:10.1109/ICVGIP.2008.47.

[37] V. Ordonez, G. Kulkarni, T.L. Berg, Im2text: Describing images using 1 million captioned photographs, in: J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, 2011, pp. 1143–1151. https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html

[38] C. Qiu, W. Zhou, A survey of recent advances in cnn-based fine-grained visual categorization, in: 20th IEEE International Conference on Communication Technology, ICCT 2020, Nanning, China, October 28-31, 2020, IEEE, 2020, pp. 1377–1384, doi:10.1109/ICCT50939.2020.9295723.

[39] S.E. Reed, Z. Akata, H. Lee, B. Schiele, Learning deep representations of fine-grained visual descriptions, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 49–58, doi:10.1109/CVPR.2016.13.

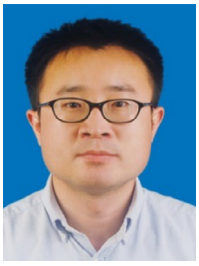
[40] M. Ristin, J. Gall, M. Guillaumin, L.V. Gool, From categories to subcategories: large-scale image classification with partial class label refinement, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 231–239, doi:10.1109/CVPR.2015.7298619.

[41] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2017) 2298–2304, doi:10.1109/TPAMI.2016.2646371.

[42] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. http://arxiv.org/abs/1409.1556

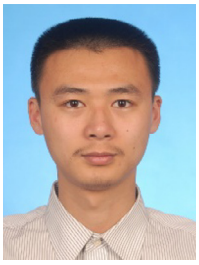
[43] H.O. Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: 2016 IEEE Conference on Computer Vision

and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 4004–4012, doi:10.1109/CVPR.2016.434.
[44] K. Song, X. Wei, X. Shu, R. Song, J. Lu, Bi-modal progressive mask attention for fine-grained recognition, IEEE Trans. Image Process. 29 (2020) 7006–7018, doi:10.1109/TIP.2020.2996736.
[45] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, WIT: wikipedia-based image text dataset for multimodal multilingual machine learning, CoRR abs/2103.01913 (2021). https://arxiv.org/abs/2103.01913
[46] M. Sun, Y. Yuan, F. Zhou, E. Ding, Multi-attention multi-class constraint for fine-grained image recognition, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI, Lecture Notes in Computer Science, volume 11220, Springer, 2018, pp. 834–850, doi:10.1007/978-3-030-01270-0_49.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2818–2826, doi:10.1109/CVPR.2016.308.
[48] Y. Tang, F. Borisyuk, S. Malreddy, Y. Li, Y. Liu, S. Kirshner, MSURU: large scale e-commerce image classification with weakly supervised search data, in: A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis (Eds.), Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, ACM, 2019, pp. 2518–2526, doi:10.1145/3292500.3330696.
[49] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, H. Jégou, Grafit: learning fine-grained image representations with coarse labels, CoRR abs/2011.12982 (2020). https://arxiv.org/abs/2011.12982
[50] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset(2011).
[51] J. Wehrmann, R. Cerri, R.C. Barros, Hierarchical multi-label classification networks, in: J.G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, Proceedings of Machine Learning Research, volume 80, PMLR, 2018, pp. 5225–5234. http://proceedings.mlr.press/v80/wehrmann18a.html
[52] X. Wei, Q. Cui, L. Yang, P. Wang, L. Liu, RPC: a large-scale retail product checkout dataset, CoRR abs/1901.07249 (2019). http://arxiv.org/abs/1901.07249
[53] X. Wei, J. Wu, Q. Cui, Deep learning for fine-grained image analysis: a survey, CoRR abs/1907.03069 (2019). http://arxiv.org/abs/1907.03069
[54] Y. Wei, S.N. Tran, S. Xu, B.H. Kang, M. Springer, Deep learning for retail product recognition: Challenges and techniques, Comput. Intell. Neurosci. 2020 (2020), doi:10.1155/2020/8875910. 8875910:1–8875910:23
[55] Y. Wei, S. Xu, S.N. Tran, B.H. Kang, Data augmentation with generative adversarial networks for grocery product image recognition, in: 16th International Conference on Control, Automation, Robotics and Vision, ICARCV 2020, Shenzhen, China, December 13-15, 2020, IEEE, 2020, pp. 963–968, doi:10.1109/ICARCV50220.2020.9305421.
[56] H. Xu, G. Qi, J. Li, M. Wang, K. Xu, H. Gao, Fine-grained image classification by visual-semantic embedding, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 1043–1049, doi:10.24963/ijcai.2018/145.
[57] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 2740–2748, doi:10.1109/ICCV.2015.314.
[58] L. Yang, P. Luo, C.C. Loy, X. Tang, A large-scale car dataset for fine-grained categorization and verification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 3973–3981, doi:10.1109/CVPR.2015.7299023.
[59] Y. Yang, L. Sheng, X. Jiang, H. Wang, D. Xu, X. Cao, Increaco: incrementally learned automatic check-out with photorealistic exemplar augmentation, in: IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021, IEEE, 2021, pp. 626–634, doi:10.1109/WACV48630.2021.00067.
[60] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229
[61] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 1103–1114, doi:10.18653/v1/d17-1115.
[62] T. Zahavy, A. Krishnan, A. Magnani, S. Mannor, Is a picture worth a thousand words? A deep multi-modal architecture for product classification in e-commerce, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 7873–7881. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16579
[63] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, X. Liang, Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 11762–11771, doi:10.1109/ICCV48922.2021.01157.
[64] H. Zhang, Y. Li, Z. Zhuang, L. Xie, Q. Tian, 3d-gat: 3d-guided adversarial transform network for person re-identification in unseen domains, Pattern Recognit. 112 (2021) 107799, doi:10.1016/j.patcog.2020.107799.
[65] B. Zhao, J. Feng, X. Wu, S. Yan, A survey on deep learning-based fine-grained object classification and semantic segmentation, Int. J. Autom. Comput. 14 (2) (2017) 119–135, doi:10.1007/s11633-017-1053-3.
[66] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2242–2251, doi:10.1109/ICCV.2017.244.



Fan Liu is currently a professor of Hohai University. He received his B.S. degree and Ph.D. degree from Nanjing University of Science and Technology (NUST) in 2009 and 2015. From September 2008 to December 2008, he studied at Ajou University in South Korea. From February 2014 to May 2014, he worked at Microsoft Research Asia. His research interests include computer vision, pattern recognition, and machine learning. Dr. Liu also serves as a reviewer of IEEE TKDE, ACM TIST, Information Sciences, Neurocomputing, Pattern Analysis and Application, and KSII Transaction on Internet and Information Systems.

Delong Chen received the B.S. degree in computer science in Hohai University, Jiangsu, China. He is currently a research intern in MEGVII Technology, and a part-time research assistant in Hohai University. His research includes computer vision and multi-modal learning.



Xiaoyu Du is currently an associate professor in Intelligent Media Analysis Group (IMAG) at School of Computer Science and Engineering, Nanjing University of Science and Technology (NJUST). He was a research fellow in the NExT++ center of National University of Singapore. He received his Ph.D. degree in computer science and technology from University of Electronic Science and Technology of China. His research interests include recommender systems, machine learning, multimedia.

Ruizhuo Gao received the B.S. degree in computer science in Hohai University, Jiangsu, China. He is currently pursuing the M.S. degree in computer science in Hohai University. His research interests include computer vision, natural language processing, and multi-modal learning.

Feng Xu is currently a professor at Hohai University. He received his Ph.D. degree from Nanjing University in 2008. He received his B.S. and M.S. degrees from Hohai University in 1998 and 2001, respectively. His research interests include cloud computing, network information security and domain software engineering etc. He has authored over 100 journal and conference papers in these areas.