TudoBonus: A Dataset for household appliance recognition to improve circular economy process

Michael Cruz
Yes Technology
São Paulo, Brazil
michael.cruz@yes.technology

Flavio Arthur Oliveira Santos
Yes Technology
São Paulo, Brazil
flavio.santos@yes.technology

Fabio Buiatti
Yes Technology
São Paulo, Brazil
fabio.buiati@yes.technology

Babak Rezaei
Yes Technology
São Paulo, Brazil
babak.rezaei@yes.technology

Abstract

In this paper, we present 'TudoBonus,' a pioneering multi-modal benchmark dataset specifically designed to address the challenges of vision recognition tasks in the context of the circular economy. Comprising over 109,000 diverse data points, including product images, SKU codes, brands, types, departments, object bounding boxes, and object descriptions, 'TudoBonus' offers a resource to drive innovation in sustainable, circular practices.To assess the complexity of 'TudoBonus,' we performed baseline experiments using pretrained models such as ResNet, ViT, DINOv1-v2, and visual-language models like CLIP and ALIGN. Our results show that even state-of-the-art methods struggle to perform classical tasks effectively with this dataset, revealing a gap between current models and the complex demands of the circular economy environment.These findings indicate that there is significant room for improvement, and we believe the diversity and quality of the 'TudoBonus' dataset will open up new avenues for research, encouraging the development of more advanced models tailored to the challenges of computer vision and automation in the circular economy, contibuting to more sustainable industries and a green economy.

1 Introduction

The circular economy has become a key topic of discussion in many government plans. For instance, Germany and the EU Commission have placed the establishment of circular economy structures at the center of the political agenda [1]. This commitment is integral to reducing strategic dependence on raw materials and ensuring intergenerational justice. These issues have gained heightened socio-political significance following the 2021 decision by the German Federal Constitutional Court and the current geopolitical crisis in Europe. Additionally, enhancing the competitiveness of circular economy companies and their processes offers significant potential for climate protection, impacting the entire industrial value chain and preserving natural resources.

In recent years, AI approaches have unblocked possibilities for practical solutions and increased industries' competitiveness. The automotive industry, for example, is a highly coordinated sector with extensive AI applications, such as self-driving vehicles, voice recognition, and more [1, 2]. In the entertainment industry, AI is already used to generate scripts, create characters, and even

[1] $https://environment.ec.europa.eu/strategy/circular-economy-action-plan_en$

Latinx in AI @ NeurIPS-24 (LXAI 2024).

assist in dubbing [3]. On the other hand, the circular economy is still in its early stages. AI-based solutions can hold significant potential for contributing to this sector, reducing costs across the entire remanufacturing value chain and benefiting all participants involved.

Although sophisticated and robust AI solutions have improved many fields, they rely on large datasets. The robustness is possible because of the abundance of web data and established benchmarks such as COCO [2], ImageNet [3], and GLUE [4]. Pre-training on these datasets has led to significant advancements in algorithms and applications across various tasks. However, recent computer vision methods, described in Section 4, struggle to perform classical tasks like classification and detection when dealing with household appliance objects, which are primary in the value chain of circular economy industries. A possible explanation for the timid results is the fine-grained characteristics, as shown in Figure 1. Unlike coarse-grained classification, identifying subtle differences among house appliance objects has proven challenging.

Figure 1: Fine-grained example of home appliance objects.

Given the points discussed, we propose and release a new image dataset called TudoBonus[5]. This novel multiview and multimodal dataset is specifically for computer vision tasks. The TudoBonus dataset provides more than one image from each object and metadata information. W features are annotated during the remanufacturing process, and a bounding box and production description are also added. For SKU labels, for example, an average of X to Y images are available, all of which are quality-controlled and human-annotated, as detailed in Section 3. In its first version, TudoBonus will feature 109,000 high-quality images.

The remainder of this work is structured as follows: We first present a selection of related works in Section 2. Section 3 provides a detailed description of the dataset and the labeling process. In Section 4, we present baseline results for classification and localization tasks and finally, Section 5 provides the conclusion. We aim to demonstrate that TudoBonus is a valuable resource for visual recognition tasks.

2 Related works

This section presents datasets related to TudoBonus. We grouped the two main categories: 1) SKU Datasets and 2) Fine-grained image classification. As TudoBonus dataset contributes to the research on these two tasks.

2.1 SKU Datasets

SKU recognition is a task that combines image processing and machine learning techniques. The process typically involves capturing images of products, using Optical Character Recognition (OCR) to detect text details, and employing Convolutional Neural Networks (CNNs) to classify and recognize the SKUs based on visual features. So, a large dataset of SKUs is crucial for improving image classification accuracy by providing diverse examples for training models. This diversity helps to improve feature extraction and generalization, leading to more precise and reliable identification of various products. [4] developed a manually labeled dataset called "Products-10K", which includes 10,000 commonly bought SKU-level products from JD.com. The research also presents multiple methods aimed at improving the detailed recognition of products. In [5] the authors introduce RP2K,

[2] https://cocodataset.org/#home

[3] https://imagenet.org/

[4] https://gluebenchmark.com/

[5] `www.willbereleased.com.br`

a large-scale retail product data set with more than 500,000 images of retail products on shelves belonging to more than 2000 different products (SKUs).

The SKU110K dataset [6] is a large-scale dataset specifically designed to evaluate object detection models, particularly in retail and store environments. It contains 11,762 images with more than 1.7 million annotated bounding boxes captured in densely packed scenarios, featuring various store shelves and products. The images originate from numerous supermarket locations and exhibit a range of scales, angles of view, lighting, and noise intensities.

2.2 Fine-grained image classification

Fine-grained image classification is a specialized area of computer vision that focuses on identifying subtle differences between similar categories within a broader class. Unlike coarse-grained classification, which deals with distinguishable categories (e.g., recognizing cats versus dogs), fine-grained classification might involve distinguishing between different species of birds, car models, or types of household appliances. This task is challenging due to the high intra-class and low inter-class variations, requiring sophisticated models and techniques to achieve high accuracy.

[7] presents a systematic survey of fine-grained image analysis, providing a comprehensive review of main techniques based on deep learning, including commonly accepted problem definitions, benchmark datasets, different families of fine-grained methods, along with covering domains specific applications. In [8], the authors present a new technique called Mutual-Channel Loss, aimed at improving classification accuracy by utilizing channel-wise information. This method prioritizes the role of each channel in an image, thus boosting the model's capability to discern subtle variations between similar objects. The effectiveness of this approach is validated through a series of experiments, which exhibit notable enhancements in classification outcomes compared to conventional techniques.

In this survey [9], the authors examine four categories of deep learning methods for fine-grained image classification: the general convolutional neural networks, approaches based on part detection, techniques involving ensembles of networks, and those utilizing visual attention. Regarding datasets, BRCars [10] is a dataset designed for fine-grained classification and various computer vision tasks. It comprises 300,325 images from 52,000 car advertisements, encompassing 427 distinct car models. These images were taken under varied angles, lighting, and environmental settings.

3 The TudoBonus dataset

In this section, we provide details about the dataset capture and labeling process. Additionally, we present a descriptive analysis of the data and outline the approach to generating English descriptions and bounding boxes for each product.

3.1 Data Capturing and labeling

Here, we describe the process of generating the TudoBonus image dataset. Figure 2 depicts the process step by step. First, the house appliance product arrives in a remanufacturing facility. Then, after unpacking the objects, they go through a pipeline process in the remanufacturing chain. In this process, a worker is responsible for evaluating and recording specific product details. To do this, the worker uses a mobile app to log information such as the brand, product type, color, SKU, department, and all relevant features. After that, another worker carefully moves the home appliance product to a designated spot in the station to capture RGB images.

Figure 3 shows an example of the products from the TudoBonus dataset. As shown, the objects vary in size and visual features. Based on the object type, the worker adjusts them to different angles and captures multiple images. There is no fixed angle for positioning the objects, and the number of images taken ranges from 2 to 10. After capturing the images, they are uploaded to the TudoBonus site [6]. It is worth to mention that, a natural language description of each object and bounding box information is also part of the dataset, as described in Subsections 3.2 and 3.4.

[6] https://loja.tudobonus.com.br/

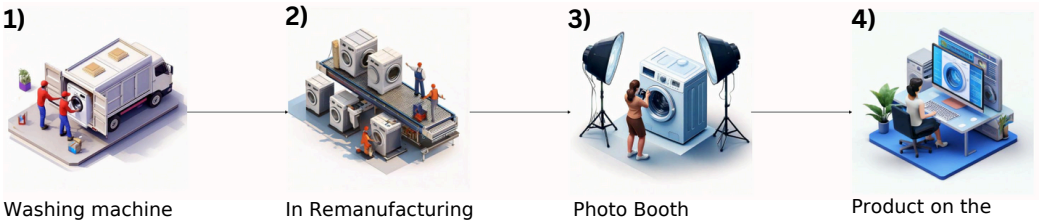

Figure 2: TudoBonus circular economy process. **1)** initially the products came to the industry and the **2)** remanufacturing process occur, then the employers **3)** take picture from the product in a photo booth and finally the **4)** product is uploaded to the tudo bonus website.

Figure 3: TudoBonus house appliance images. This grid of images shows several examples of products present in TudoBonus dataset.

3.2 Data Description

We release a new labeled data set and benchmark containing images of household appliances. The data set contains more than 109,000 entries and consists of 14 columns. The columns provide various attributes of the products, as illustrated in Table 1.

Table 1: Data columns, types and sample values

Column Name	**Data Type**	**Examples**
description	string	electrolux washer machine top load 18kg
sku	string	fx0115000001, 2117cdba106
department	categoric	cook, freeze, portable appliances
product type	categoric	oven, refrigerator, freezer, purifier
brand	categoric	brastemp, consul, panasonic
color	categoric	black, stainless steel, white, titanium
voltage	categoric	0, 127v, 127v-220v, 220v
height	float	44,2cm, 105,2cm, 76,1cm
width	float	13,2cm, 67,2cm, 38,9cm
depth	float	32,8cm, 17,22cm, 66,19cm
weight	float	9,5kg, 21,5kg, 45kg
super-class	int	0-70

Let us elaborate the main columns in detail. The column interactionid contains 23,978 unique values, representing a vast number of distinct interactions, providing a valuable opportunity for analysis and insight, potentially revealing patterns and correlations that can inform further research or applications.

The column labeled description holds crucial text information about the product, that can be used as metadata for classification tasks. tasks. The department column is a categorical variable that includes values such as cooking, freezing, portable appliances, washing, among others. The column product type has sixteen possible values such as oven, refrigerator, freezer, purifier, washing machine, etc. The

brand column has values such as Brastemp, Consul, Panasonic, Esmaltec and so on. The department column is a categorical variable listing categories like cooking, freezing, portable appliances, washing, among and others. The column product type includes sixteen distinct values, such as oven, refrigerator, freezer, purifier, washing machine, etc. and more. The brand column has values such as Brastemp, Consul, Panasonic, Esmaltec and so on.

The SKU column contains 334 unique classes, making SKU classification a challenging task due to the vast number of categories to distinguish between. To address this complexity, we have introduced a new column called "super-class", which groups similar SKUs into broader categories. This approach simplifies the classification process by reducing the number of distinct types that need to be considered while maintaining meaningful hierarchical relationships between the items. So, the super-class column implies that the SKUs were categorized into higher-level classes based on four columns (brand, color, department, and product type).

3.3 Product descriptions

In addition to the product's discrete attributes (e.g., SKU, brand, product type), the TudoBonus dataset also has the product description in Portuguese and English. Figure 4 shows a sample of the descriptions. The Portuguese descriptions are made by a human before uploading the product's image to the TudoBonus site (Step 4) of the Figure 2). This information is essential because it allows researchers to work not only with image classification models but also with multi-modal models such as visual language models [11, 12], multi-modal large language models [13], and vision foundation models [14].

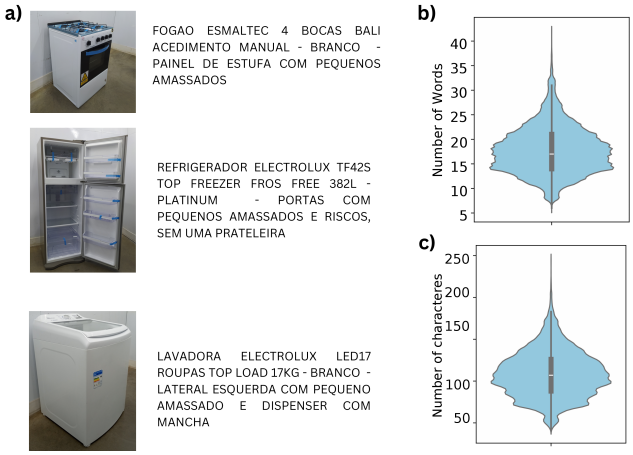

Figure 4: a) Examples of product descriptions in portuguese from the TudoBonus dataset. Each description on the right is for the product on the left. The plot in b) shows the distribution of the number of words in the description considering all the dataset, while the c) shows the distribution of the number of characters.

3.4 Product localization

One of the key components of this dataset involves product detection and the precise delineation of bounding boxes around detected objects (product localization). The dataset includes a dedicated column that provides the bounding box (bbox) coordinates for each object. This information is important because it enable multi-task model development or even multimodal models as Florence 2 [14] and PaliGemma [15]. It is worth mentioning that we use a BBoxWidget [7] to generate localization.

[7] https://github.com/gereleth/jupyter-bbox-widget

4 Baselines

This section presents baseline experiments performed with the Tudobonus dataset. The baseline experiments are important to highlight the dataset's difficulty and complexity, provide a starting point for performance comparison by other researchers, and identify challenges and opportunities. Trying to provide a reference point for other researchers, we performed experiments with object detection tasks, product classification (SKU, type, brand) based on pre-trained models (i.e. transfer learning), and zero-shot product classification with visual language models. In the following, we present each experiment independently.

4.1 Object detection and localization

Precise object detection and localization are critical in many computer vision applications, as they directly influence the accuracy and reliability of downstream tasks such as image segmentation, tracking, and scene understanding. Accurate detection enables systems to identify objects correctly, while precise localization, typically through bounding boxes, ensures that objects are accurately framed within an image. As aforementioned, the proposed dataset includes a dedicated bounding box column. Initially, bounding boxes were generated using **PaliGemma** [15] as a pre-trained model. To assess the quality of the generated bounding boxes, a sample of 300 images was randomly selected from various product categories. The selection process considered the distribution of images across different product types, aiming to create a more balanced sample dataset. These selected images were manually reviewed, and new bounding boxes were generated accordingly. To facilitate the generation of bounding boxes, an effective tool was developed using the *BBoxWidget* from the *jupyter_bbox_widget* Python library. Bounding boxes were drawn manually with careful attention to accuracy, ensuring precise localization of objects within the images. This manual approach aids in creating reliable ground truth data essential for evaluating model performance. In the search for the most suitable pre-trained model for this application, the performances of **PaliGemma** [15], **Florence-2** [14] and **YoloV9** were compared, with evaluation grounded on the manually generated bounding boxes. Three evaluation metrics were employed: *Intersection over Union (IoU)*, *Generalized Intersection over Union (GIoU)*, and *Center-to-Center Distance (CC)*, each calculated according to the following formulas:

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

$$GIoU = IoU - \frac{|C \backslash A \cup B|}{|C|} \tag{2}$$

$$CC = Euclidean\ distance\ between\ the\ two\ centers \tag{3}$$

where *A* and *B* represent the bounding boxes, and *C* is the smallest enclosing box that covers both the predicted and ground truth boxes. The evaluation results are presented in Table 2, clearly demonstrating that Florence-2 outperforms the other models in terms of accuracy across these metrics. Moreover, Figure 5 shows the same results graphically.

Table 2: Calculated metrics for object detection and localization models

	Algorithms		
Metrics	**PaliGemma**	**Florence-2**	**Yolov9**
IoU	0.7646	0.8192	0.6351
GIoU	0.7518	0.8188	0.5018
Center-to-Center	17.8117	14.6236	57.8462

It is noteworthy that the models employed for product detection and localization in this study are pre-trained and have not undergone dataset-specific fine-tuning. Analysis of the results indicates that, although some outcomes are promising, model performance could be significantly improved through fine-tuning tailored to our dataset. As demonstrated in Figures 6 and 7, all three pre-trained models exhibited limitations in detecting certain products, underscoring the necessity of fine-tuning to enhance detection accuracy and overall model effectiveness.

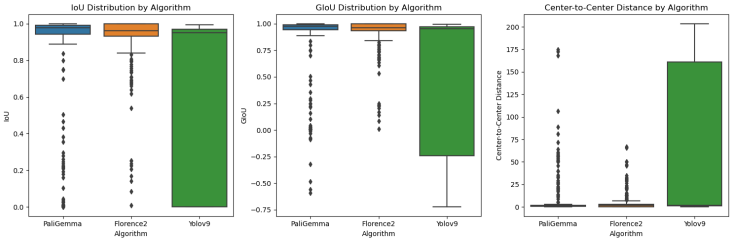

Figure 5: Comparison of IoU, GIoU and Center-to-Center metrics for different pre-trained algorithms

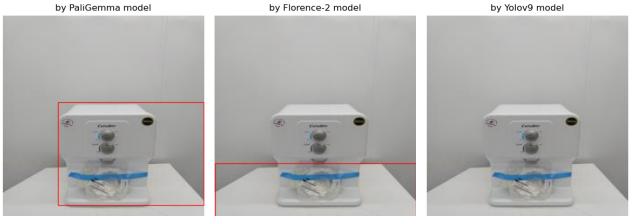

Figure 6: Sample of product detection and localization using pre-trained models

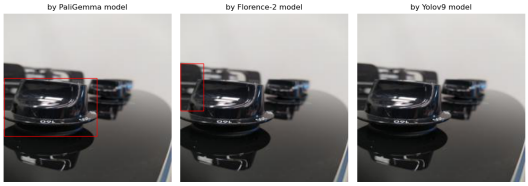

Figure 7: Sample of product detection and localization using pre-trained models

4.2 Classification based on pretrained models

A standard approach to perform image classification is using a pretrained model to extract image representation, then training a simple classifier on top of these representations. In this section, we perform this experiment using different pretrained models, such as ResNet [16], ViT [17], DINOv1 [18], DINOv2 [19], and CLIP [11]. Table 3 presents the results obtained from this experiment to the test set split. Each cell of the results is the balanced accuracy metric for the respective feature extractor and classifier model. The results shows that classify the brand, product type and department is not a challenge, as a pretrained model on the ImageNet as the ResNet achieved more than 95% on these three tasks. However, classify the SKU is a challenging task that there are room for improvement. The best SKU classifier achieved only 48% of accuracy. SKU classification from the product picture is challenging due to many products may have similar appearances, making it challenging to differentiate them based only on visual information. In addition, products can be customized with different colors, patterns, or features, which can further complicate identification. The TudoBonus dataset helps to fill this gap in the literature, providing a dataset with more than 100.000 product images.

4.3 Zero-shot classification based on VLMs

With the advent of visual language models such as CLIP [11] and ALIGN [12], the research developed zero-shot image classification with high accuracy, only prompting the VLM with a text describing the class name and an input image. Given an image x and a text d, the VLM returns the similarity score between x and d ($\phi(d_c, x)$), thus a zero-shot image classification based on VLM follows the equation 5, where each class c has a description (i.e. prompt text) d_c and the class which has the highest text score similarity with the image is defined as the predicted. The baseline prompt is "A photo of a {class name}", however we may also use the object and class descriptions as prompt, or even class attributes descriptions [20].

Table 3: Accuracy results for feature extractor models.

Feature Extractor	Classifier	SKU	Brand	Product Type	Department
DINOv2	Linear	3.42	43.99	36.11	59.08
ViT16	Linear	3.56	42.67	39.32	73.26
DINO	Linear	4.21	39.29	38.01	59.89
ViT8	Linear	9.54	44.86	41.81	67.93
ViT8	XGBoost	9.93	51.03	49.29	72.40
ViT8	KNN	10.32	41.04	39.05	67.04
ViT16	XGBoost	10.64	51.93	51.94	79.90
DINOv2	XGBoost	11.77	55.54	47.98	70.71
DINO	XGBoost	12.24	53.01	51.44	74.87
ViT16	KNN	20.45	64.82	67.04	88.95
ResNet34	XGBoost	27.23	86.22	90.40	98.17
ResNet18	XGBoost	28.41	89.49	90.59	97.69
DINOv2	KNN	31.87	83.41	81.79	95.32
DINO	KNN	31.92	84.48	81.94	95.84
ResNet34	Linear	36.7	82.57	88.37	94.85
ResNet18	Linear	38.75	85.25	88.01	94.65
CLIP	KNN	42.69	95.14	96.23	**99.72**
ResNet34	KNN	46.33	95.57	97.41	99.70
ResNet18	KNN	**48.70**	**97.21**	**97.60**	99.67

In this experiment, we evaluate the performance of zero-shot product type classification based on two well-known VLMs, CLIP and ALIGN. As a prompt strategy, we performed the experiment with the baseline and the object description, thus evaluating if the class description has an impact on the score similarity. Table 4 presents the results from this experiment. In general, the ALIGN model achieved better results than CLIP. Besides, the object descriptions improved the ALIGN accuracy by approximately 5%. These results show that performing zero-shot product type classification based on pre-trained VLM is still challenging (even though these VLMs have been trained with hundreds of millions of images and text pairs). Therefore, the Tudobonus dataset has a text description of all products, which enables the development and research of VLMs specific to the product domain.

$$s(c, x) = \phi(d_c, x) \tag{4}$$

$$P(x) = \underset{c \in C}{\operatorname{argmax}}\, s(x, c) \tag{5}$$

Table 4: Accuracy results for zero-shot image classification.

VLM	Prompt strategy	Accuracy for Product Type
CLIP	Baseline	29.96
	Object description	27.19
ALIGN	Baseline	40.25
	Object description	**45.21**

5 Conclusion

TudoBonus dataset presents a valuable resource for advancing computer vision tasks in a circular economy context. By offering a comprehensive, multi-view, and multimodal dataset of remanufactured household appliances, TudoBonus enables the exploration and enhancement of object recognition and classification models. The dataset's detailed bounding box annotations and rich metadata provide a robust foundation for evaluating detection and localization algorithms. Initial evaluations reveal

the challenges posed by the dataset, particularly for pre-trained models that exhibit limitations in handling the fine-grained characteristics of household appliances. These findings underscore the need for dataset-specific fine-tuning to achieve optimal performance. TudoBonus has the potential to contribute significantly to both AI-driven circular economy research and industry applications, fostering new approaches to sustainable manufacturing and resource efficiency.

References

[1] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020.

[2] Chee Yang Loh, Kai Lung Boey, and Kai Sze Hong. Speech recognition interactive system for vehicle. In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pages 85–88. IEEE, 2017.

[3] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9313–9319, 2023.

[4] Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset, 2020.

[5] Jingtian Peng, Chang Xiao, and Yifan Li. Rp2k: A large-scale retail product dataset for fine-grained image classification, 2021.

[6] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*, 2019.

[7] 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*, 44(12):8927–8948, 2021.

[8] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29:4683–4695, 2020.

[9] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017.

[10] Daniel M Kuhn and Viviane P Moreira. Brcars: a dataset for fine-grained classification of car images. In *2021 34nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2021.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[13] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[14] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[15] Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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

[17] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[20] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.