ePillID Dataset: A Low-Shot Fine-Grained Benchmark for Pill Identification

Naoto Usuyama, Natalia Larios Delgado, Amanda K. Hall
Microsoft Healthcare

{naotous,nlarios,amhal}@microsoft.com

Jessica Lundin
Work done at Microsoft

lundinmachine@outlook.com

Abstract

Identifying prescription medications is a frequent task for patients and medical professionals; however, this is an error-prone task as many pills have similar appearances (e.g. white round pills), which increases the risk of medication errors. In this paper, we introduce ePillID, the largest public benchmark on pill image recognition, composed of 13k images representing 8184 appearance classes (two sides for 4092 pill types). For most of the appearance classes, there exists only one reference image, making it a challenging low-shot recognition setting. We present our experimental setup and evaluation results of various baseline models on the benchmark. The best baseline using a multi-head metric-learning approach with bilinear features performed remarkably well; however, our error analysis suggests that they still fail to distinguish particularly confusing classes.

1. Introduction

At least 1.5 million preventable adverse drug events (ADE) occur each year in the U.S. [22]. Medication errors related to ADEs can occur while writing or filling prescriptions, or even when taking or managing medications [1]. For example, pharmacists must verify thousands of pills dispensed daily in the pharmacy, a process that is largely manual and prone to error. Chronically-ill and elderly patients often separate pills from their original prescription bottles, which can lead to confusion and misadministration. Many errors are preventable; however, pills with visually similar characteristics are difficult to identify or distinguish, increasing error potential.

The pill images available for vision-based approaches fall into two categories: reference and consumer images (Figure 1). The reference images are taken with controlled lighting and backgrounds, and with professional equipment. For most of the pills, one image per side (two images per pill type) is available from the National Institutes of Health (NIH) Pillbox dataset [36]. The consumer images are taken with real-world settings including different lighting, back-

Figure 1: **Representative examples of reference images (top row) and corresponding consumer images (bottom row)** Both sides (front and back) of the two pill types (NDC:00093-7180 and NDC:50111-0441, respectively) are shown.

grounds, and equipment. Building pill-image datasets, especially for prescription medications, is costly and has additional regulatory obstacles as prescription medications require a clinician's order to be dispensed. A labeled dataset is scarce and even unlabeled images are hard to collect. This setup requires the model to learn representations from very few training examples. Existing benchmarks are either not publicly available or smaller scale, which are not suitable for developing real-world pill identification systems.

Over the years, deep learning has achieved unprecedented performance on image recognition tasks with efforts from both large-scale labeled datasets [41, 66] and modeling improvements [18, 30, 43]; however, fine-grained visual categorization (FGVC) is still a challenging task, which needs to distinguish subtle differences within visually similar categories. FGVC tasks mainly include natural categories (e.g., birds [50, 52], dogs [26] and plants [38, 56]) and man-made objects (e.g., cars [29, 58] and airplanes [35]). In this work, we target the pill identification task, which is an under-explored FGVC task yet important and frequent in healthcare settings. The distributions of pill shape and color are skewed to certain categories (Figure 2), highlighting the importance of distinguishing subtle differences such as materials, imprinted text and symbols.

The main contribution of this paper is introducing ePillID, a new pill identification benchmark with a real-world
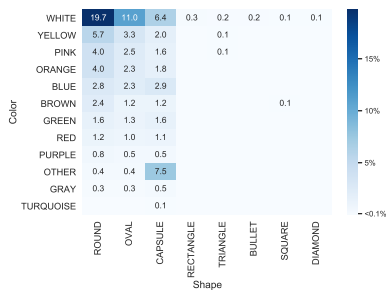
Figure 2: **The distribution of shape and color of pills.** For pills registered with multiple colors, we grouped them as OTHER. The combinations less than 0.1% are omitted.
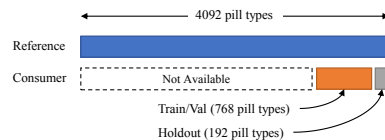

Figure 3: **The distribution of reference and consumer images on the ePillID benchmark.** The reference images are available for 4092 pill types, while the consumer images are only available for 960 pill types. The dataset includes 9804 reference images (two images, front and back, for each pill type) and 3728 consumer images.

low-shot recognition setting. Leveraging two existing NIH datasets, our benchmark is composed of 13k images representing 8184 appearance classes (two sides for 4092 pill types). This is a low-shot fine-grained challenge because (1) for most of the appearance classes there exist only one image and (2) many pills have extremely similar appearances. Furthermore, we empirically evaluate various approaches with the benchmark to serve as baselines. The baseline models include standard image classification approaches and metric learning based approaches. Finally, we present error analysis to motivate future research directions.

2. Related Work

Image-based Pill Identification: In 2016, the NIH held a pill image recognition challenge [60], and released a dataset composed of 1k pill types. The challenge winner [63] proposed a deep similarity-based approach [53] and developed a mobile-ready version using knowledge-distillation techniques [19]. Following the competition, classification-based approaches were applied, with higher recognition performance reported [12, 48, 57]. Aside from deep neural networks, various approaches with feature engineering have been proposed [6, 11, 31, 37, 57]. For instance, the Hu moment [20] was applied [11, 31], because of its rotation-invariant properties. Other methods [7, 47, 61, 62] were proposed to generate imprint features, recognizing the importance of imprints for pill identification. The past methods achieved remarkable success; however, the lack of benchmarks prevents us from developing methods for real-world low-shot settings. Wong *et al.* [57] created a 5284-image dataset of 400 pill types with a QR-like board to rectify geometric and color distortions. Yu *et al.* [61] collected 12,500 images of 2500 pill categories. Unfortunately, neither of these datasets are publicly available.

Fine-Grained Visual Categorization (FGVC): In many FGVC datasets [5, 3, 25, 9], the number of categories is not extremely large, often less than 200. Recent large-scale datasets [17, 14, 51] offer large numbers of categories with many images (e.g., 675,170 training images for 5089 categories [51]) with challenging long-tailed distributions. Compared to other FGVC benchmarks, the data distribution on the ePillID benchmark imposes a low-shot setting (one image for most of the classes) with a large number of classes (8k appearance classes). Among many algorithms [59, 55, 65] proposed for FGVC tasks, bilinear models [28, 32, 64] achieved remarkable performances by capturing higher-order interactions between feature channels. B-CNN [33] is one of the first approaches, which obtains full bilinear features by calculating outer product at each location of the feature map, followed by a pooling across all locations; however, the full bilinear features can be very high dimensional (e.g., over 250k when the input has 512 channels). Compact Bilinear Pooling (CBP) [15] addresses the dimensionality issue by approximating bilinear features with only a few thousand dimensions, which was shown to outperform B-CNN in few-shot scenarios. Another line of work is metric learning [40, 10, 45], where an embedding space that captures semantic similarities among classes and images is learned. Metric learning has been also successfully used in few- and low-shot settings [54, 46], making it suitable for our ePillID benchmark.

3. ePillID Benchmark

We construct a new pill identification benchmark, ePillID, by leveraging the NIH challenge dataset [49] and the NIH Pillbox dataset [36]. We use the challenge dataset, which offers consumer images, as a base dataset and extend it with the reference images from the Pillbox dataset. In total, the ePillID dataset includes 3728 consumer images for 1920 appearance classes (two sides for 960 pill types) and 8192 reference images (two sides for 4092 pill types). This requires a fine-grained low-shot setup, where models have access to one reference image for all the 8192 appearance classes; however, there only exists a few consumer images for 1920 appearance classes (Figure 3).

Experimental Setup: The consumer images are split into 80% training and 20% holdout sets in such a way that the pill types are mutually exclusive. The training set is further split on pill types for 4-fold cross-validation. The models have access to reference images for all of the 4092 pill types, but consumer images are unavailable for most pill types during training. To evaluate the performance in situations where both front- and back-sides are available as input, we construct two-sided queries by enumerating all possible consumer image pairs for each pill type.

Evaluation Metrics: For each query, a model calculates an ordered list of pill types with confidence scores. Note that, for the experiments with both sides of a pill, a query consists of a pair of front and back images. We consider Mean Average Precision (MAP) and Global Average Precision (GAP) for the model performance evaluation. For MAP, the average precision score is calculated separately for each query, and the mean value is calculated. MAP measures the ability to predict the correct pill types given queries. For GAP, all the query and pill-type pairs are treated independently, and the average precision score is calculated globally. GAP measures both the ranking performance and the consistency of the confidence scores i.e. the ability to use a common threshold across different queries [39]. We also calculate MAP@1 and GAP@1, where only the top pill type per query is considered.

4. Experiments

We first introduce our baseline approaches, then present quantitative and qualitative results.

4.1. Baseline Models

We use ResNet [18] and DenseNet [21] as base networks pretrained on ImageNet [41] for initial weights. In addition to the global average pooling layer as features, we evaluate two bilinear methods, B-CNN [33] and CBP [15], applied to the final pooling layer of the base network. For CBP, we use their Tensor Sketch projection with 8192 dimensions, which was suggested for reaching close-to maximum accuracy. We insert a 1x1 convolutional layer before the pooling layer to reduce the dimensionality to 256.

Plain Classification: As a first set of baselines, we train the models with the standard softmax cross-entropy loss. For regularization, we add a dropout layer [44] with probability 0.5 before the final classification layer. We use the appearance classes as target labels during training, and take the max of the softmax scores for calculating pill-type confidence scores. For the two-sides evaluation, the mean confidence score is used for a score between a two-sided input and a pill type.

Model	GAP	GAP@1	MAP	MAP@1
Plain Classification: Both-sides input				
DenseNet121	30.05 ± 2.04	38.53 ± 4.42	60.03 ± 1.05	41.21 ± 1.89
DenseNet201	33.30 ± 1.27	42.64 ± 3.40	64.38 ± 0.85	46.61 ± 1.11
DenseNet161	34.78 ± 2.80	44.47 ± 4.26	64.87 ± 2.27	46.85 ± 2.72
DenseNet161 B-CNN	36.35 ± 1.67	47.08 ± 2.16	67.07 ± 1.36	50.84 ± 2.47
DenseNet161 BCP	39.47 ± 1.90	48.72 ± 1.92	69.77 ± 1.01	52.72 ± 1.42
ResNet18	24.13 ± 1.99	33.13 ± 2.34	54.70 ± 2.13	35.70 ± 2.79
ResNet34	29.07 ± 2.45	38.53 ± 2.54	58.88 ± 2.26	40.34 ± 2.01
ResNet50	35.16 ± 3.46	45.43 ± 4.22	65.79 ± 1.85	48.32 ± 2.34
ResNet50 B-CNN	38.15 ± 2.73	48.13 ± 4.17	68.68 ± 2.23	51.92 ± 3.23
ResNet50 CBP	38.97 ± 2.74	48.63 ± 3.94	68.35 ± 3.08	51.60 ± 4.24
ResNet101	36.15 ± 3.21	46.96 ± 5.29	65.50 ± 2.95	48.20 ± 5.02
ResNet152	39.57 ± 1.23	49.97 ± 1.58	68.51 ± 1.09	51.64 ± 2.31
ResNet152 B-CNN	44.75 ± 2.11	55.14 ± 3.67	72.65 ± 1.69	56.75 ± 2.51
ResNet152 CBP	41.45 ± 2.15	52.43 ± 2.07	69.12 ± 2.10	52.64 ± 2.29
Multi-head Metric Learning: Both-sides input				
DenseNet121	74.04 ± 2.32	90.15 ± 2.63	93.42 ± 0.42	87.58 ± 0.87
DenseNet201	74.19 ± 2.08	87.69 ± 2.58	92.52 ± 0.91	86.42 ± 1.56
DenseNet161	76.60 ± 2.35	88.86 ± 2.88	93.41 ± 0.72	87.78 ± 1.55
DenseNet161 B-CNN	75.27 ± 4.65	88.24 ± 2.83	93.11 ± 0.81	87.22 ± 1.57
DenseNet161 BCP	77.15 ± 2.89	89.11 ± 2.36	94.03 ± 1.03	88.98 ± 1.78
ResNet18	66.58 ± 1.91	87.87 ± 2.86	91.16 ± 0.75	84.38 ± 1.36
ResNet34	73.61 ± 3.17	89.22 ± 3.35	92.74 ± 1.95	87.10 ± 2.96
ResNet50	78.27 ± 2.35	90.71 ± 2.55	94.02 ± 0.66	89.06 ± 1.05
ResNet50 B-CNN	79.95 ± 1.96	90.26 ± 1.56	94.59 ± 0.23	89.66 ± 0.62
ResNet50 CBP	78.44 ± 1.48	91.36 ± 2.83	95.27 ± 0.43	91.01 ± 0.97
ResNet101	79.83 ± 2.11	92.23 ± 1.12	94.99 ± 0.40	90.65 ± 0.60
ResNet152	78.63 ± 2.75	90.54 ± 1.96	95.71 ± 0.83	91.93 ± 1.56
ResNet152 B-CNN	80.45 ± 1.09	89.61 ± 1.83	95.01 ± 0.43	90.65 ± 0.81
ResNet152 CBP	81.20 ± 1.47	91.19 ± 0.28	95.76 ± 0.40	92.01 ± 0.67
Multi-head Metric Learning: Single-side input				
ResNet34	54.70 ± 2.58	78.69 ± 2.37	80.52 ± 1.76	70.61 ± 2.11
ResNet50	61.75 ± 1.40	81.63 ± 2.15	82.34 ± 0.79	72.56 ± 1.35
ResNet101	63.93 ± 1.71	84.51 ± 1.68	84.18 ± 1.23	75.72 ± 1.48
ResNet152	64.61 ± 2.45	82.67 ± 1.17	85.25 ± 0.91	76.42 ± 1.54

Table 1: **Recognition results on the ePillID benchmark.** Mean and standard deviations of the holdout metrics from the 4-fold cross-validation are reported in percentages.

Multi-head Metric Learning: As another set of baselines, we employ a combination of four losses to learn an embedding space optimized for fine-grained low-shot recognition. The set up is a multi-task training procedure:

$$L_{final} = \lambda_{SCE} L_{SCE} + \lambda_{\eta} L_{\eta} + \lambda_{\rho} L_{\rho} + \lambda_{\Gamma} L_{\Gamma}, \quad (1)$$

where L_{SCE} indicates softmax cross-entropy, L_{η} cosine-softmax loss (ArcFace [13]), L_{ρ} triplet loss [42], and L_{Γ} contrastive loss [8]. The loss weights, λ_{SCE}, λ_{η}, λ_{ρ}, and λ_{Γ} are chosen empirically with a ResNet50 model (Section 4.1). In order to compute the loss for every mini-batch, the triplet and contrastive loss requires additional sampling and pairing procedures. We apply online hard-example mining to find informative negatives for the triplets and pairs respectively evaluated by these losses. The trained model is used for generating embeddings for the query consumer images and all the reference images. We calculate cosine similarities between the query and reference embeddings and use them as confidence scores.

Implementation Details: The images are cropped and resized to 224×224. Extensive data augmentation is employed to mimic consumer-image-like variations, including

Contrastive (λ_Γ)	Triplet (λ_ρ)	SCE (λ_{SCE})	CSCE (λ_η)	Validation GAP
1.0	1.0	1.0	0.1	77.18 ± 1.90
1.0	1.0	1.0	0.3	74.07 ± 2.29
1.0	1.0	1.0	0.0	75.48 ± 1.02
1.0	0.0	1.0	0.0	74.40 ± 2.93
0.0	1.0	1.0	0.0	73.89 ± 2.16
1.0	1.0	0.0	0.1	71.38 ± 4.52
0.0	0.0	1.0	0.0	61.78 ± 3.24

Table 2: **Validation GAP scores from the corss-validation for comparing the metric-learning loss weights.** The ResNet50 baseline model is used.

rotation and perspective transformation. The Adam optimizer [27] is used with an initial learning rate of 1×10^{-4}. The learning rate is halved whenever a plateau in the validation GAP score is detected. The model hyper-parameters are chosen to optimize the average validation GAP score using the 4-fold cross-validation. The loss weights for the metric learning are determined based on the ResNet50 experiment (Table 2). The training is done with the mini-batch of 48 images in a machine equipped with Intel Xeon E5-2690, 112GB RAM, one NVIDIA Tesla P40, CUDA 8.0 and PyTorch 0.4.1.

4.2. Quantitative Results

In Table 1, we report the baseline results on the ePillID benchmark. The plain classification baselines performed much worse than the metric learning baselines, suggesting the difficulty of the low-shot fine-grained setting. The performance differences among the models are consistent with the ImageNet pretraining performance [4]. The multi-head metric leraning models performed remarkably well, achieving over 95% MAP and 90% GAP@1. In most of the cases, bilinear pooling methods outperformed the global average pooling counterpart, showing the representation power of the bilinear features. As an ablation study, we report single-side experiment results i.e. only one image per query. The ResNet152 metric learning approach achieved over 85% MAP and 82% GAP@1; however, the results indicate that both sides are required for accurate identification.

4.3. Qualitative Results

Figure 4 depicts qualitative comparisons from examples of the ePillID holdout dataset with confidence scores. In (a), the plain classification approaches misclassified, whereas the metric learning approaches identified successfully, even with the challenging lighting and background variations in the query images. In (b), only the metric learning with CBP approach identified correctly, suggesting CBP was effective for capturing the small difference in the imprinted text. For (c) and (d), all the four models failed to identify the correct types. In (c), the consumer images are affected by the lighting variations with the shiny pill material. In (d), the pill types share extremely similar appearances, except the one

	query	correct	wrong

(a) NDC:00781-1452
plain-avg [0.137, 0.074, 0.068, **0.066**, 0.047, …]
plain-cbp [0.080, **0.076**, 0.068, 0.065, 0.045, …]
metric-avg [**0.658**, 0.586, 0.547, 0.540, 0.510, …]
metric-cbp [**0.725**, 0.666, 0.566, 0.562, 0.514, …]

(b) NDC:00093-0135
plain-avg [0.286, 0.263, 0.238, **0.085**, 0.027, …]
plain-cbp [0.421, **0.382**, 0.036, 0.033, 0.022, …]
metric-avg [0.739, **0.732**, 0.651, 0.649, 0.561, …]
metric-cbp [**0.866**, 0.780, 0.672, 0.662, 0.619, …]

(c) NDC:00555-1883
plain-avg [0.888, 0.042, 0.031, 0.007, **0.005**, …]
plain-cbp [0.986, 0.005, 0.002, 0.001, **0.001**, …]
metric-avg [0.842, **0.831**, 0.768, 0.715, 0.686, …]
metric-cbp [0.870, **0.812**, 0.764, 0.725, 0.719, …]

(d) NDC:51672-4030
plain-avg [0.878, **0.114**, 0.006, 0.0, 0.0, …]
plain-cbp [0.993, **0.005**, 0.0, 0.0, 0.0, …]
metric-avg [0.867, **0.865**, 0.687, 0.652, 0.649, …]
metric-cbp [0.887, **0.811**, 0.675, 0.642, 0.538, …]

Figure 4: **Qualitative results from the ePillID holdout dataset.** For each query, top confidence scores are shown from the four ResNet152 baselines. The reference images of **correct** and wrong pill types are shown with the input consumer images in the left.

character difference in the imprinted text.

5. Conclusion

We introduced ePillID, a low-shot fine-grained benchmark on pill image recognition. To our knowledge, this is the first publicly available benchmark that can be used to develop pill identification systems in a real-world low-shot setting. We empirically evaluated various baseline models with the benchmark. The multi-head metric learning approach performed remarkably well; however, our error analysis suggests that these models still cannot distinguish confusing pill types reliably. In the future, we plan to integrate optical character recognition (OCR) models. OCR integration has been explored for storefronts and product FGVC tasks [23, 2] and recent advances in scene text recognition are promising [24, 16, 34]; however, existing OCR models are unlikely to perform reliably on pills as they stand. Challenging differences include low-contrast imprinted text, irregular-shaped layouts, lexicon, and pill materials such as capsules and gels. Finally, we plan to extend the benchmark further with more pill types and images as we collect more data. By releasing this benchmark, we hope to support further research in this under-explored yet important task in healthcare.

References

[1] Jeffrey K Aronson. Medication errors: what they are, how they happen, and how to avoid them. *QJM: An International Journal of Medicine*, 102(8):513–521, 2009. 1

[2] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 4

[3] Sara Beery, Dan Morris, and Pietro Perona. The iwildcam 2019 challenge dataset. *arXiv preprint arXiv:1907.07617*, 2019. 2

[4] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018. 4

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 2

[6] Jesus J Caban, Adrian Rosebrock, and Terry S Yoo. Automatic identification of prescription drugs using shape distribution models. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1005–1008, 2012. 2

[7] Zhiyuan Chen and Sei-ichiro Kamata. A new accurate pill recognition system using imprint information. In *Proceedings of the International Conference on Machine Vision*, volume 9067, page 906711, 2013. 2

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546. IEEE, 2005. 3

[9] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4109–4118, 2018. 2

[10] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1153–1162, 2016. 2

[11] António Cunha, Telmo Adão, and Paula Trigueiros. Helpmepills: A mobile pill recognition tool for elderly persons. *Procedia Technology*, 16:1523–1532, 2014. 2

[12] Natalia Larios Delgado, Naoto Usuyama, Amanda K Hall, Rebecca J Hazen, Max Ma, Siva Sahu, and Jessica Lundin. Fast and accurate medication identification. *npj Digital Medicine*, 2(1):10, 2019. 2

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[14] Fieldguide. Butterflies and Moths. https://sites.google.com/view/fgvc6/competitions/butterflies-moths-2019, 2019. CVPR Workshop: FGVC. 2

[15] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016. 2, 3

[16] Raul Gomez, Baoguang Shi, Lluis Gomez, Lukas Numann, Andreas Veit, Jiri Matas, Serge Belongie, and Dimosthenis Karatzas. Icdar 2017 robust reading challenge on coco-text. In *International Conference on Document Analysis and Recognition*, volume 1, pages 1435–1443. IEEE, 2017. 4

[17] Xintong Han, Sheng Guo, Weilin Huang, Lingshu Kong, and Matt Scott. iMaterialist Challenge. https://github.com/malongtech/imaterialist-product-2019, 2019. CVPR Workshop: FGVC. 2

[18] 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*, 2016. 1, 3

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[20] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179–187, 1962. 2

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 3

[22] Institute of Medicine. *Preventing Medication Errors*. The National Academies Press, Washington, DC, 2007. 1

[23] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5):1063–1076, 2016. 4

[24] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *International Conference on Document Analysis and Recognition*, pages 1156–1160. IEEE, 2015. 4

[25] Parneet Kaur, , Karan Sikka, Weijun Wang, serge Belongie, and Ajay Divakaran. Foodx-251: A dataset for fine-grained food classification. *arXiv preprint arXiv:1907.06167*, 2019. 2

[26] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fgvc: Stanford dogs. In *CVPR Workshop: FGVC*, volume 1, 2011. 1

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[28] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 365–374, 2017. 2

[29] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, pages 554–561, 2013. 1

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[31] Young-Beom Lee, Unsang Park, Anil K Jain, and Seong-Whan Lee. Pill-id: Matching and retrieval of drug pill images. *Pattern Recognition Letters*, 33(7):904–910, 2012. 2

[32] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–955, 2018. 2

[33] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, pages 1449–1457, 2015. 2, 3

[34] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5685, 2018. 4

[35] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[36] National Library of Medicine (NLM). Pillbox. https://pillbox.nlm.nih.gov, 2016. 1, 2

[37] Mateus A Vieira Neto, Joao WM de Souza, Pedro P Reboucas Filho, and W de O Antonio. Cofordes: An invariant feature extractor for the drug pill identification. In *IEEE International Symposium on Computer-Based Medical Systems*, pages 30–35. IEEE, 2018. 2

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1

[39] Florent Perronnin, Yan Liu, and Jean-Michel Renders. A family of contextual measures of similarity between distributions with application to image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2358–2365. IEEE, 2009. 3

[40] Qi Qian, Rong Jin, Shenghuo Zhu, and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3716–3724, 2015. 2

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 3

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. 3

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[45] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *ECCV*, pages 805–821, 2018. 2

[46] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2

[47] Siroratt Suntronsuk and Sukanya Ratanotayanon. Pill image binarization for detecting text imprints. In *International Joint Conference on Computer Science and Software Engineering*, pages 1–6. IEEE, 2016. 2

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[49] U.S. National Library of Medicine (NLM). Pill image recognition challenge. https://pir.nlm.nih.gov/challenge/, 2016. 2

[50] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 1

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018. 2

[52] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1

[53] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 2

[54] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[55] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4148–4157, 2018. 2

[56] Jan D Wegner, Steven Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6014–6023, 2016. 1

[57] Yuen Fei Wong, Hoi Ting Ng, Kit Yee Leung, Ka Yan Chan, Sau Yi Chan, and Chen Change Loy. Development of fine-grained pill identification algorithm using deep convolutional network. *Journal of biomedical informatics*, 74:130–136, 2017. 2

[58] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3973–3981, 2015. 1

[59] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV*, pages 420–435, 2018. 2

[60] Ziv Yaniv, Jessica Faruque, Sally Howe, Kathel Dunn, David Sharlip, Andrew Bond, Pablo Perillan, Olivier Bodenreider, Michael J Ackerman, and Terry S Yoo. The national library of medicine pill image recognition challenge: An initial report. In *IEEE Applied Imagery Pattern Recognition Workshop)*, pages 1–9. IEEE, 2016. 2

[61] Jiye Yu, Zhiyuan Chen, and Sei-ichiro Kamata. Pill recognition using imprint information by two-step sampling distance sets. In *Proceedings of the International Conference on Pattern Recognition*, pages 3156–3161. IEEE, 2014. 2

[62] Jiye Yu, Zhiyuan Chen, Sei-ichiro Kamata, and Jie Yang. Accurate system for automatic pill recognition using imprint information. *IET Image Processing*, 9(12):1039–1047, 2015. 2

[63] Xiao Zeng, Kai Cao, and Mi Zhang. MobileDeepPill: A Small-Footprint Mobile Deep Learning System for Recognizing Unconstrained Pill Images. In *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services*, pages 56–67, Niagara Falls, NY, USA, 2017. 2

[64] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning deep bilinear transformation for fine-grained image representation. In *Advances in Neural Information Processing Systems*, pages 4279–4288, 2019. 2

[65] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019. 2

[66] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014. 1