

Product Image Recognition with Guidance Learning and Noisy Supervision

Qing Li^{1,2}, Xiaojiang Peng², Liangliang Cao⁴, Wenbin Du², Hao Xing³, Yu Qiao²

¹Southwest Jiaotong University, Chengdu, China

²Shenzhen Institutes of Advanced Technology, CAS, China

³Vipshop Inc., Guangzhou, China

⁴University of Massachusetts at Amherst, U.S.A.

Abstract

This paper considers to recognize products from daily photos, which is an important problem in real-world applications but also challenging due to background clutters, category diversities, noisy labels, etc. We address this problem by two contributions. First, we introduce a novel large-scale product image dataset, termed as Product-90. Instead of collecting product images by labor-and time-intensive image capturing, we take advantage of the web and download images from the reviews of several e-commerce websites where the images are casually captured by consumers. Labels are assigned automatically by the categories of e-commerce websites. Totally the Product-90 consists of more than 140K images with 90 categories. Due to the fact that consumers may upload unrelated images, it is inevitable that our Product-90 introduces noisy labels. As the second contribution, we develop a simple yet efficient guidance learning (GL) method for training convolutional neural networks (CNNs) with noisy supervision. The GL method first trains an initial teacher network with the full noisy dataset, and then trains a target/student network with both large-scale noisy set and small manually-verified clean set in a multi-task manner. Specifically, in the stage of student network training, the large-scale noisy data is supervised by its guidance knowledge which is the combination of its given noisy label and the soften label from the teacher network. We conduct extensive experiments on our Products-90 and public datasets, namely Food101, Food-101N, and Clothing1M. Our guidance learning method achieves performance superior to state-of-the-art methods on these datasets.

1. Introduction

This paper studies a crucial problem in real-world application: recognize products from consumer photos without much supervision. More specifically, we want to recognize the fine-grained products taken by consumer's mobile cam-

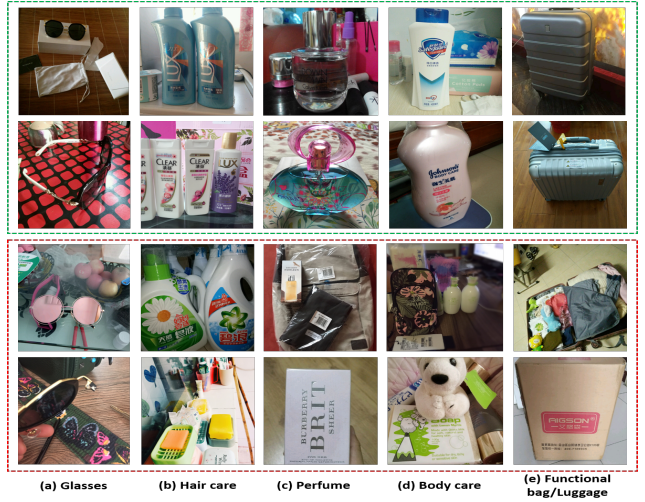


Figure 1. Example images from our Products-90. We illustrate 5 different categories in column. Visually correct images are shown in the first two rows. Visually confused or unrelated images are shown in the last two rows.

eras, with unconstrained viewing directions, cluttered background, and different lighting conditions. One can imagine an application that you are recommended where the products can be found and what the prices are by recognizing your casually-captured product photos.

To address this real-world product image recognition task, we build a novel large-scale dataset, termed as Product-90, which consists of 90 generic product categories. Instead of collecting daily images by labor-and time-intensive image capturing, we take advantage of the web and download images from the reviews of several e-commerce websites where the images are casually captured by consumers. Totally, we collected more than 140k product images from the customer reviews. The associated 90 categories are borrowed the categories of e-commerce websites. Figure 1 shows some examples of this dataset. We can see there are several challenges brought by Product-90 dataset: i) The visual contents in Product-90 contains a wide

range of subjects. ii) Some categories are very similar in appearance, e.g. Hair Care vs. Body Care. iii) Some photos are not related to the category, which suggests a significant level of noise exists in the dataset.

To evaluate product image recognition algorithms on our Product-90, we build a small manually-clean subset for traditional training and testing, and remain the rest of Product-90 as noisy data which can be used for extra training. To take full advantage of the small clean training subset and the massive noisy labeled data for daily product recognition, we propose a novel *guidance learning framework* for noisy data learning. It mainly includes two training stages. At the first stage, we train a baseline CNN model, or a teacher model, on the full Product-90 dataset (without the clean test set). At the second stage, we train a student or target network on the large-scale noisy set and the small clean training set with multi-task learning. Specifically, in the stage of student training, the large-scale noisy data is supervised by the guidance knowledge which consists of two supervision signals, namely the noisy ground truths and the soften labels from the teacher network. We fuse these one-hot ground truths with the soften multi-hot labels, and optimize the network by Kullback–Leibler Divergence (KLDiv) loss.

Our guidance learning framework considers the following issues. The first stage of our guidance learning ensures that we can obtain a powerful teacher model instead of using the noisy set or clean set only like in [25]. We fuse both ground truths and soften labels for the large-scale noisy data, since i) the teacher model provides useful information but is far from perfect, and ii) there exist both false and correct labels in noisy labels.

In summary, our contributions can be concluded as follows:

- We introduce a new task, i.e. daily product image recognition, and a novel large-scale dataset, termed as Product-90 which is collected from the reviews of e-commerce websites.
- To advance the performance of daily product image recognition, we propose a generic guidance learning method to take full advantage the small clean subset and the large-scale noisy data in Product-90.
- We conduct comprehensive evaluations with our guidance learning method on our Products-90, Food-101 [1], Food-101N [14], and Clothing1M [33], and achieve state-of-the-art results.

2. Related Work

Our work is related to product image recognition and noisy data learning. In this section, we first review some related product image datasets and noisy datasets, and then present existing noisy data learning methods.

2.1. Related Datasets

Product image datasets. As for product images in computer vision and multimedia community, researchers mainly focus on the products of retails and groceries such as Supermarket [27], Grocery Products [5], Gorzi-120 [20], Feribur Groceries [12], RPC [32]. **Supermarket** [27] is introduced for automatic fruit and vegetable classification from images. It has 15 product categories with 2,633 images captured under diverse conditions. **Grocery Products** [5] is another dataset aiming at grocery product recognition. It contains 80 grocery product categories with 8,350 training images and 680 test images. **Grozi-120** [20] is a dataset proposed for groceries recognition in natural environment. It contains 120 grocery product categories. For each product category, two types of images are collected, one from the web, the other from inside a grocery store. In total, 11,870 images are collected with 676 from the web and 11,194 from the store. **Freiburg Groceries** [12] is another grocery dataset comprising 5,021 images of 25 grocery classes. The images are divided into two sets: a training set that consists of 4,947 images taken by smartphone cameras, each containing one or more instances of one class; a test set with 74 images of 37 clutter scenes, each containing objects of multiple classes. **RPC** [32] is a recently-published retail product image dataset aiming at automatic checkout application. This dataset also provides images of two different types. One type is taken in a controlled environment and only contains a single product. Another type represents images of user-purchased products and these images usually include multiple products. In total, it contains 83,739 images of 200 fine-grain classes. Different from these retail or grocery product image datasets, our proposed Product-90 is a *label noisy* dataset collected by mobile cameras in daily life, and contains categories from retail products to clothing and shoes. Liu *et al.* [17] also propose a daily photo dataset but it limited on clothing images.

Noisy datasets. In recent research, both synthetic and real-world noisy datasets are widely used. For example, MNIST and CIFAR-10 are used as synthetic noisy datasets in [25, 29]. Synthetic label noise usually mimics random class noise and confusing class noise by corrupting the original clean datasets. To explore noisy data learning methods, three real-world noisy datasets are introduced more recently. Xiao *et al.* [33] present the Clothing1M fashion image dataset which consists of 14 classes with more than a million images crawled from online shopping websites. Li *et al.* [15] introduce the WebVision dataset which contains 2.4M noisy labeled images crawled from Flickr and Google using the ILSVRC taxonomy [3]. Lee *et al.* [14] collect the Food-101N dataset which contains 310k images from Google, Bing, Yelp, and TripAdvisor using the Food-101 taxonomy [1]. Our Products-90 is related to Clothing1M

but contains much more categories including clothing, bags, jewelry, shoes, home products, personal care products, stationery, etc. Meanwhile, the Products-90 is crawled from the customer reviews of online shopping websites which includes more complex background clutter and noise.

2.2. Noisy Data Learning Methods

We focus on the label noise problem and refer to [4] for a comprehensive overview. Methods on learning with label noise can be roughly grouped into three categories: noise-robust methods, semi-supervised noisy data learning methods, and noise-cleaning methods.

Noise-robust methods. The noise-robust or noise-tolerance learning methods are assumed to be not too sensitive to the presence of label noise, which directly learn models from the noisy labeled data [11, 13, 22, 25, 18]. Nettleton *et al.* [24] show that the Naive Bayes probabilistic learner is less sensitive to label noise. Manwani [19] present a noise-tolerance algorithm under the assumption that the corrupted probability of an example is a function of the feature vector of the example. Mnih *et al.* [23] propose two robust loss functions to deal with label noise. With synthetic noisy labeled data, Rolnick *et al.* [28] demonstrate that deep learning is robust to noise when training data is sufficiently large with large batch size and proper learning rate. Guo *et al.* [6] develop a curriculum training scheme to learn noisy data from easy to hard. Jiang *et al.* [10] design a MentorNet to adjust the loss weights of noisy samples in the training process.

Semi-supervised noisy data learning methods. Semi-supervised methods aim to improve performance using a small manually-verified clean set. These methods usually obtain higher performance than the other methods since extra human supervision is added. Lee *et al.* [14] train an auxiliary CleanNet using manually-verified data to detect label noise and adjust the final sample loss weights. Similarly, Veit *et al.* [31] also use the clean set to train a label cleaning network but with a different architecture. These methods assume there exists such a label mapping from noisy labels to clean labels. Xiao *et al.* [33] mix the clean set and noisy set, and train an extra label noise type CNN and a classification CNN to estimate the posterior distribution of the true label. *Our guidance learning belongs to the semi-supervised noisy data learning which leverages a teacher-student training strategy to take full use of the whole data space (noisy set and clean set) and the student network trades off the noisy ground truths and soften labels by guidance knowledge.*

Noise-cleaning methods. Noise-cleaning methods aim to identify and remove or relabel noisy samples with filter approaches [21]. Brodley *et al.* [2] propose to filter noisy samples using ensemble classifiers with majority and consensus voting. Sukhbaatar *et al.* [29] introduce an ex-

tra noise layer into a standard CNN which adapts the network outputs to match the noisy label distribution. Daiki *et al.* [30] propose a joint optimization framework to train deep CNNs with label noise, which updates the network parameters and labels alternatively. Based on the consistency of the noisy groundtruth and the current prediction of the model, Reed *et al.* [26] present a ‘Soft’ and a ‘Hard’ bootstrapping approach to relabel noisy data. Similarly, Li *et al.* [16] relabel noisy data using the noisy groundtruth and the current prediction adjusted by a knowledge graph constructed from DBpedia-Wikipedia. Our guidance learning framework is also related to [26] and [16] but differs in that *i) we consider a small clean set instead of noisy data only which inherits the advantage of semi-supervised methods and ii) we train the teacher model with the full dataset and the student model with guidance knowledge in a multi-task learning manner.*

3. The Product-90 Dataset

Considering the applications of generic product image recognition, we collect the Products-90 dataset. It is collected by crawling images in consumer reviews from several e-shopping websites. Labels are assigned by categories provided by the sellers like in [33]. Several image samples along with the annotation are shown in Figure 2. The 90 classes are mainly selected according to the categories of these websites and filtered by their definition ambiguity. We keep the original fine-grain classes since it is useful in practice. These product classes can be mainly grouped into 13 meta categories, namely tools, baby and kids, home, beauty and personal care, shoes, clothing, accessories, sports/outdoors, grocery, health, luggage, electronics, and background. There exist fine-grain classes in most of the meta categories. For example, there are 21 categories in the beauty and personal care including eye makeup, lip makeup, cheek makeup, facial care, eye care, lip care, etc.

Our current version of Products-90 consists of 142,466 images, with hundreds or thousands of samples for each class. Figure 3 shows the statistics of Products-90 where each color corresponds to a certain meta categories. All categories are relatively balanced except for the ‘facial care’ and ‘women tops’. To separate product classes from non-product class, the background category is added by crawling scene images (keywords such as landscape, building) in several searching engines. The number of images for the background is comparable to the one of the largest class (i.e. facial care).

Protocols. Due to the fine-grain classes in Products-90, we find that it is hard to relabel or refine manually as in [33]. Instead, we manually verify about 17K of all the images in the dataset which are correctly-labeled. We further split it into training (D_c) and test sets, which contains 8,795 and

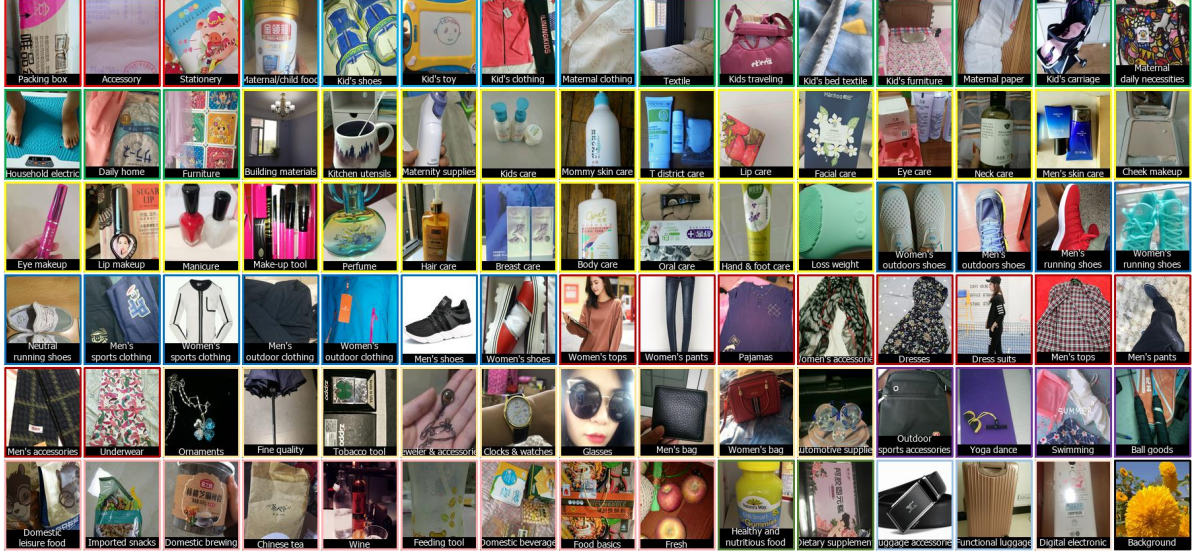


Figure 2. Illustration of the Products-90 dataset. Each image represents one class which is selected from clean data. Different image boundary colors correspond to different meta categories. (Zoom in for better view.)

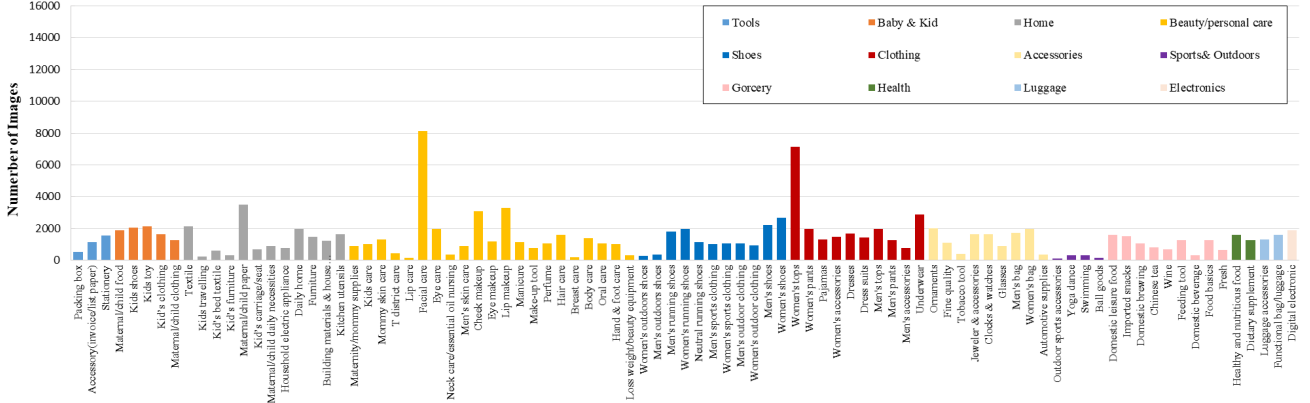


Figure 3. Statistics of the collected Products-90. Each color indicates a meta category. (Zoom in for better view.)

8,787 images, respectively. This small training subset is used as clean set. The remaining of Products-90 is used as noisy training set (D_n). We report the overall accuracy on the clean test set.

4. Guidance Learning

We propose the guidance learning framework to deal with the problem of product image recognition with noisy supervision. The guidance learning framework is illustrated in Figure 4. Our framework consists of two stages: 1) teacher network training 2) student network training. In the first stage, we use all the training data to train a basic CNN model, which is called the teacher model. In the second stage, we use the noisy training dataset and the clean training dataset to train a student model in a multi-task learning manner.

4.1. Teacher Network

The teacher network is trained on the full dataset which contains mislabeled and correctly-labeled samples. More specifically, given dataset $D = \{(x_i, y_i) \mid i = 1 \dots N\}$, where x_i denotes the i -th observed image and the corresponding label $y_i \in \{1, \dots, C\}$ and C is the number of the categories, we use all the training data D to learn the teacher model.

At this stage, the teacher network training is considered identical with the classical classification problem, assuming all the samples are correctly labeled. The loss function of teacher network is the cross entropy between the softmax output p and the ground-truth distribution over labels q .

$$\mathcal{L}_{teacher} = - \sum_{i=1}^C q_i \log p_i \quad (1)$$

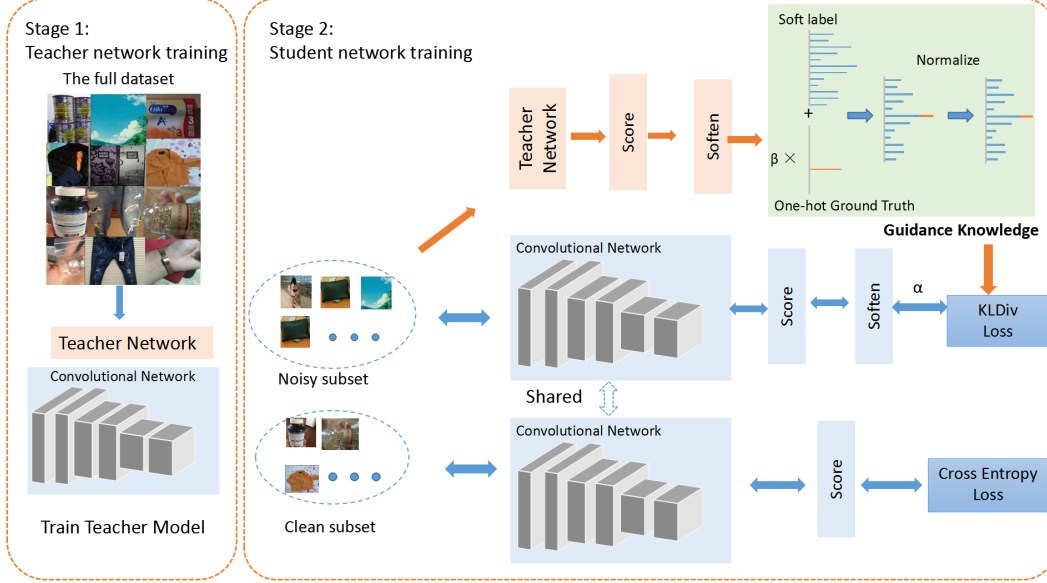


Figure 4. The proposed guidance learning framework. At the first stage, we utilize all training data to train a teacher model. At the second stage, we separate the training data into a noisy subset and a clean subset to train the student network with a multi-task learning mechanism.

where q_i is the ground-truth distribution of the i th true class label.

4.2. Student Network

Once we finished the first stage, we obtain a teacher model which contains implicit information of the dataset. At the second stage, we train a target network in a multi-task manner with the large-scale noisy set and a small clean training set. We refer this set as the clean subset and the remaining noisy dataset as the noisy subset. We denote the clean training subset $D_c = \{(\hat{x}_i, y_i) \mid i = 1 \dots K\}$ and noisy subset $D_n = \{(x_i^*, y_i^*) \mid i = 1 \dots M\}$. In our case, the clean training data is a small portion of the training data. We have $N = |D_c| + |D_n|$ with $|D_c| \ll |D_n|$. Since the clean subset is built based on the selected samples with the right annotation, the noisy subset and the clean subset could have different distribution. However, the clean training subset share the same distribution with the clean test set. To take full advantage of the whole training dataset, we propose the guidance learning method which leverages different supervision information for the noisy subset and clean subset, and combine them with a multi-task learning (MTL) framework.

For the clean subset, it is natural to choose the traditional cross entropy loss for training since we are confident about the sample labels. For the noisy subset, instead of using the original label which may be totally wrong, we aim to alleviate the effects of the noisy label and to maintain the right label functioning well at the same time. To achieve this goal, we resort to the noisy labels as well as the knowledge included in the softened predictions of the teacher network.

Specifically, we first input the noisy subset into the teacher network to obtain logit scores $z_i \in R^C$ for the i -th sample. Then, we feed the predictions into the softmax layer with a score-soften operation to obtain the target soft probabilities p_i . In the soften operation, inspired by [8] we introduce a temperature to transfer the knowledge of teacher. The target soft probability p_i is defined as follows,

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \quad (2)$$

where z_i is the pre-softmax activations of teacher network and T is the temperature which softens the signals. As mentioned in [8], the soften operation can provide more information or knowledge about the model's prediction.

To achieve the final soft target of a sample, we fuse the knowledge from the teacher network p_i and its noisy label y_i which is a one-hot vector as follows,

$$g_i = \frac{1}{(1 + \beta)}(p_i + \beta y_i), \quad (3)$$

where β is a trade-off weight of the two parts. We call g_i as guidance knowledge in the paper since it contains both transferred and noisy clues.

Once we obtain the soft targets, the noisy subset is supervised by KL-divergence Loss as implemented in most of deep learning toolbox. Formally, it is defined as,

$$\mathcal{L}_g(\theta) = \sum_i^N g_i \log\left(\frac{g_i}{q_i}\right), \quad (4)$$

where q denotes the softened prediction of networks, g is the soft target label.

Table 1. Comparison on Products-90.

Model #	Method	Training Data	Initialization	Test Accuracy
1	ResNet-101	noisy data D_n	ImageNet	60.97%
2	ResNet-101	clean data D_c	ImageNet	62.15%
3	ResNet-101	D_n and D_c	ImageNet	66.78 %
4	Guidance Learning	D_n and D_c	model#3	68.86 %

Finally, the student network is trained by integrating the KL-divergence loss for the noisy subset and cross entropy loss for the clean subset. The total loss is as follows:

$$\mathcal{L}_{total}(\theta) = \alpha T^2 \mathcal{L}_g + \mathcal{L}_c \quad (5)$$

where T^2 is used to compensate the impact of soften operation in Eq.(2), α is the hyper-parameter which balances the importance between these two tasks, and \mathcal{L}_c is the cross-entropy loss on clean dataset whose formulation is the same with Eq.(1).

5. Experiments

In this section, we first present the implementation details, and then conduct extensive evaluations with our guidance learning method on the Products-90, and finally apply our method on Food-101 and Food-101N.

5.1. Implementation Details

We implement our method with Pytorch. For data augmentation, we resize images to scale 256×256 , and randomly crop regions of 224×224 with random flipping. We crop the middle 224×224 regions for testing. We use ResNet-101 [7] architecture on Products-90, and ResNet-50 on Food-101N and Clothing1M. All networks are pre-trained on the ImageNet dataset. In the teacher network training step, we initialize the learning rate (lr) to 10^{-3} , and divide it by 10 after 10, 15 and 20 epochs. We stop training after 25 epochs. In the student network training step, we set the lr to 10^{-4} , and divide it by 10 after 5, 8 and stop training after 11 epochs. We use the SGD method for optimization with a momentum of 0.9 and a weight decay of 10^{-3} . The batch size is set to 64 for all steps. For the hyper-parameters α , β , and T in our guidance learning framework, the default values are 0.1, 0.3, and 5, respectively.

5.2. Exploration of Guidance Learning on Products-90

In this section, we first compare our guidance learning method to several well-known baseline methods, and then evaluate the hyper-parameters.

Table 1 presents the test accuracy comparison between our methods and several baselines, i.e. model #1, #2, and #3. These baselines ignore the noisy label problem and view all labels as ground truth. Traditional cross-entropy loss are used for all these baselines. As shown in Table 1,

training on the noisy set D_n gets the worst result 60.97%, which even inferior to the one trained on the small clean set (i.e., 62.15%). It suggests that i) the noisy subset may show different distribution compared to the clean one and ii) noisy labels degrade CNNs significantly even there is a large scale of data. As found in [33], training on both clean and noisy sets is a better choice which achieves 66.78% on our collected dataset. We use this model as the teacher model of our guidance learning framework. With the same full dataset, our guidance learning framework further improves the teacher model by 2.08%. As another useful trick in noisy data learning [25, 6], fine-tuning the model trained with noisy data on clean set further boosts the final performance. This trick improves our guidance learning from 68.86% to 71.4%, and boosts the model#3 from 66.78% to 68.6% .

The effect of noise. To investigate the effect of noise for our guidance learning framework, we change the ratio of clean images. Specifically, we reduce the number of clean images to 10%, 30%, 50%, and 80% of the original clean training set (i.e. D_c). Figure 5 presents the results of the teacher models and student models on the test set. We also compare our method to the most related work in [16] which also uses knowledge distillation where the teacher model is trained on clean data and the student model on the full data with modified soft labels. We do not use knowledge graph to refine soft labels as [16] since we do not have. Our teacher model is slightly impacted by the decreasing of clean images (i.e. we remove partial clean training images). Our guidance learning framework consistently improves the teacher model by more than 2%. [16] is inferior to our method consistently and degrades significantly when the ratio of clean images is reduced. For example, reducing the number of clean images to 10% (about 10 images for each class) degrades about 20%, and both the teacher model and student model achieve less than 40%. However, it only leads to 2.22% degradation (66.64% vs. 68.86%) for our student model which demonstrates the robustness and efficiency of guidance learning framework even with tiny-scale of clean images. This can be explained by that i) training with tiny-scale clean data, leads to overfitting easily and ii) a bad teacher model impacts the final performance of its student models.

Evaluation of α , T and β . β is a trade-off weight between noisy labels and the predictions of teacher model in our guidance learning. Taking the default values of α and T , we evaluate different β from 0 to 1. The results are shown in Figure 6. We observe that increasing β boosts performance but saturates above 0.3. α balances the importance between the losses of noisy set and clean set, T is the temperature used for softening. We evaluate α and T by fixing β to 0.3. The results are illustrated in Figure 7. From Figure 7, several observations can be found. First, ‘ $T=5$ ’ consistently

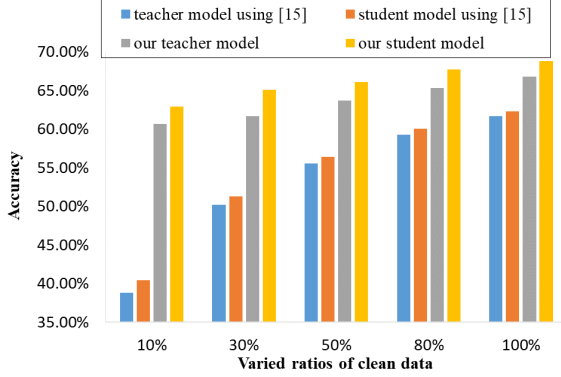


Figure 5. Evaluation of clean image ratios w.r.t. the original clean set.

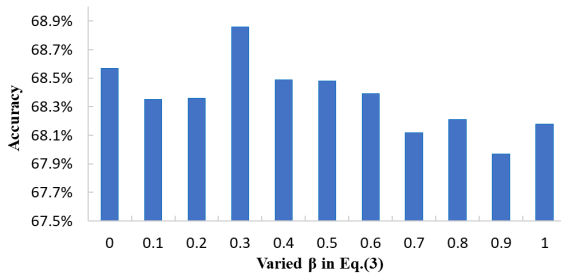


Figure 6. Evaluation of β in Eq.(3).

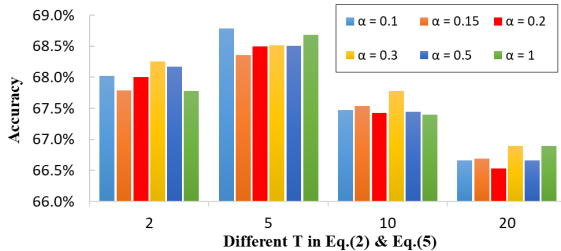


Figure 7. Evaluation of α and T in Eq.(5).

outperforms the others regardless of α . Second, increasing β boosts performance in the beginning but degrades after 5, which indicates that a highly-soften operation corrupts supervision knowledge. Third, α and T impact performance jointly which change the loss of noisy set in Eq. (5).

5.3. Experiments on Food-101 and Food-101N

Food-101 and Food-101N. The Food-101 dataset [1] is a benchmark for visual food evaluation. It contains 101 food categories, with 101,000 real-world food images totally. For each class, 750 images are used for training, the other 250 images for testing. It is a clean dataset reviewed manually. The training set of Food-101 is used as clean trainset D_c . To conduct experiments with label noise, we utilize the Food-101N noisy dataset as D_n . The Food-101N dataset [14] is collected from Google, Bing, Yelp, and TripAdvisor with the concepts of Food-101, and is filtered

Table 2. Comparison between our method and recent state-of-the-arts on Food-101. *60K extra images have been used which have both correct labels and noisy labels introduced in [14]. It is worth noting that 60K is comparable to the number of training images in Food-101. \dagger BNInception is used for backbone network which is pre-trained on the full ImageNet dataset with 21k classes. *The results in brackets are our best reimplementation of CleanNet using Pytorch.*

#	Method	Training Data	Init.	Accuracy
1	ResNet-50[14]	Food-101N	ImageNet	81.44
2	ResNet-50 [14]	Food-101	ImageNet	81.67
3	ResNet-50	Food-101 and Food-101N	ImageNet	85.80
4	ResNet-50	Food-101N*	ImageNet	79.83
5	CleanNet(hard) [14]	Food-101N*	ImageNet	83.47(82.39)
6	CleanNet(soft) [14]	Food-101N*	ImageNet	83.95(82.99)
7	Curriculum [6] \dagger	Food-101 and synthetic data	ImageNet	87.3
8	Guidance Learning	Food-101N*	model#4	84.20
9	Guidance Learning	Food-101 and Food-101N	model#3	87.36

out from foodspotting.com where the Food-101 is collected. We use all the 310k images of Food-101N as the noisy dataset in our experiments, and report the overall accuracy on the Food-101 test set.

We further validate our method on Food-101 and Food-101N. The performance comparison is presented in Table 2. The first two baselines are provided in [14]. Our teacher model trained on both the training set of Food-101 and Food-101N obtains 85.8% which outperforms the baselines and CleanNet [14] in a large margin. It demonstrates that more data is better even there has label noise. Our guidance learning method improves the baseline result from 85.8% to 87.36%, which outperforms the current state-of-the-art methods. It is worth noting that CleanNet, which is a state-of-the-art semi-supervised noisy data learning method, leverages additional 60K manually-verified images within the Food-101N. The number of 60K is comparable to 75K of the original training images on Food-101. For a fair comparison, we conduct an extra experiment with the same 60K manually-verified set, and obtain 84.2% which is better than both the hard (82.39%) and soft (82.99%) version of CleanNet.

5.4. Experiments on Clothing1M

Clothing1M. The Clothing1M dataset [33] is a public large-scale fashion dataset to evaluate recognition accuracy from noisy data with human supervision. It contains 1 million images with noisy class labels from 14 fashion classes and thousands of human-annotated images. All the images are crawled from several online shopping websites. The human-annotated set is used as the clean set which is fur-

Table 3. Performance comparison between our method and recent state-of-the-art methods on Clothing1M. *32K images have both correct labels and noisy labels which are used to train CleanNet. ‡BNInception is used for backbone network which is pre-trained on the full ImageNet dataset with 21k classes.

#	Method	Training Data	Initialization	Accuracy
1	ResNet-50	1M noisy	ImageNet	68.94
2	ResNet-50	50K clean	ImageNet	75.19
3	ResNet-50	1M noisy + 50K clean	ImageNet	71.61
4	CleanNet(hard) [14]	1M noisy*	ImageNet	74.14
5	CleanNet(soft) [14]	1M noisy*	ImageNet	74.69
6	CleanNet(soft) [14]+ D_c Finetuning	50k clean	model#5	79.90
7	Loss correction [25]	1M noisy	ImageNet	69.84
8	Loss correction [25] + D_c Finetuning	50k clean	model#7	80.38
9	Curriculum [6]‡	1M noisy	ImageNet	75.80
10	Curriculum [6]‡ + D_c Finetuning	50k clean	model#9	81.50
11	Guidance Learning	1M noisy + 50K clean	ImageNet	75.76
12	Guidance Learning + D_c Finetuning	50k clean	model#11	80.31
13	Guidance Learning ‡	1M noisy + 50K clean	ImageNet	78.77
14	Guidance Learning + D_c Finetuning ‡	50k clean	model#13	81.13

ther split into training D_c , validation and test sets with the size of 50k, 14k, 10k, respectively. A confusion matrix between the human annotations and the original noisy labels shows that the overall accuracy is 61.54% in [33]. We report the overall accuracy on the test set of Clothing1M.

The first two baselines are provided in [25]. Our baseline trained on mixed noisy and clean data is 71.61% which is slightly lower than those in [25]. Both [14] and [25] use ResNet-50 as backbone network, and respectively obtain 74.69% and 69.84% without the final finetuning process on clean data. With the same backbone, our guidance learning method improves the teacher model from 71.61% to 75.76% which outperforms [14] and [25]. CurriculumNet [6] uses the BNInception [9] network which is pre-trained on the full ImageNet with 21k classes as its backbone model, and is implemented with Caffe¹. For a fair comparison, we replace the ResNet-50 as the same BNInception model, and obtain 78.77% without finetuning on clean set which is 2.93% better than CurriculumNet (78.77% vs. 75.8%).

For the results with a further finetuning on clean set, the accuracies are 79.9%, 80.38%, and 81.5% in [14], [25], and [6], respectively. We obtain 81.13% with the finetuning trick which is comparable to the state of the arts. [25] estimates a confusion matrix which indicates the probability of each class being corrupted into another. This method is based on the assumption that the noisy label can be corrected to another. [6] designs a curriculum to train noisy data from easy to hard which utilizes a clustering process and a training process repeatedly. Compared to these two state-of-the-art methods, our guidance learning method is more efficient and simpler.

¹<http://caffe.berkeleyvision.org/>

6. Conclusion

In this paper, we present a large-scale daily product image dataset, termed as Product-90, for recognizing product in daily life. Compared to existing product datasets, our product dataset is more diverse in product categories and owns more images. Since our Product-90 introduces noisy labels, we also propose a simple yet efficient guidance learning framework to address the problem of training CNNs from noisy data. It first trains an initial teacher network from the full dataset including both clean and noisy data, and then separates the noisy part and clean part, and finally trains a target network with multi-task learning. In the target network training step, the noisy data is supervised by the guidance knowledge which is the combination of its noisy label and soft label from the teacher network. Experiments on several public datasets and our dataset show that our guidance learning method improves the base model significantly and achieves state-of-the-art performance. All the code will be publicly available including the reimplementation of CleanNet.

References

- [1] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014.
- [2] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

- [4] B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2014.
- [5] M. George and C. Floerkemeier. Recognizing products: A per-exemplar multi-label image classification approach. In *ECCV*, pages 440–455. Springer, 2014.
- [6] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang. Curriculumnet: Weakly supervised learning from large-scale web images. *arXiv preprint arXiv:1808.01097*, 2018.
- [7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [8] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [9] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [10] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. Mentornet: Regularizing very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.
- [11] A. Joulin, L. van der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, pages 67–84. Springer, 2016.
- [12] P. Jund, N. Abdo, A. Eitel, and W. Burgard. The freiburg groceries dataset. *arXiv preprint arXiv:1611.05799*, 2016.
- [13] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, pages 301–320. Springer, 2016.
- [14] K.-H. Lee, X. He, L. Zhang, and L. Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. *arXiv preprint arXiv:1711.07131*, 2017.
- [15] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.
- [16] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li. Learning from noisy labels with distillation. In *ICCV*, pages 1928–1936, 2017.
- [17] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337. IEEE, 2012.
- [18] Y. Lu, C. Yuan, Z. Lai, X. Li, W. K. Wong, and D. Zhang. Nuclear norm-based 2dlpp for image classification. *IEEE Transactions on Multimedia*, 19(11):2391–2403, 2017.
- [19] N. Manwani and P. Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.
- [20] M. Merler, C. Galleguillos, and S. Belongie. Recognizing groceries in situ using in vitro training data. In *CVPR*, pages 1–8. IEEE, 2007.
- [21] A. L. Miranda, L. P. F. Garcia, A. C. Carvalho, and A. C. Lorena. Use of classification algorithms in noise detection and elimination. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 417–424. Springer, 2009.
- [22] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, pages 2930–2939, 2016.
- [23] V. Mnih and G. E. Hinton. Learning to label aerial images from noisy data. In *ICML*, pages 567–574, 2012.
- [24] D. F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4):275–306, 2010.
- [25] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 2233–2241, 2017.
- [26] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.
- [27] A. Rocha, D. C. Hauagge, J. Wainer, and S. Goldenstein. Automatic fruit and vegetable classification from images. *Computers and Electronics in Agriculture*, 70(1):96–104, 2010.
- [28] D. Rolnick, A. Veit, S. Belongie, and N. Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.
- [29] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.
- [30] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint optimization framework for learning with noisy labels. *arXiv preprint arXiv:1803.11364*, 2018.
- [31] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. J. Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pages 6575–6583, 2017.
- [32] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu. Rpc: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*, 2019.
- [33] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015.