MetaCleaner: Learning to Hallucinate Clean Representations for Noisy-Labeled Visual Recognition

Weihe Zhang*[1] Yali Wang*[1] Yu Qiao†[1,2]
[1] Shenzhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2] The Chinese University of Hong Kong

Abstract

Deep Neural Networks (DNNs) have achieved remarkable successes in large-scale visual recognition. However, they often suffer from overfitting under noisy labels. To alleviate this problem, we propose a conceptually simple but effective MetaCleaner, which can learn to hallucinate a clean representation of an object category, according to a small noisy subset from the same category. Specially, MetaCleaner consists of two flexible submodules. The first submodule, namely Noisy Weighting, can estimate the confidence scores of all the images in the noisy subset, by analyzing their deep features jointly. The second submodule, namely Clean Hallucinating, can generate a clean representation from the noisy subset, by summarizing the noisy images with their confidence scores. Via MetaCleaner, DNNs can strengthen its robustness to noisy labels, as well as enhance its generalization capacity with richer data diversity. Moreover, MetaCleaner can be easily integrated into the standard training procedure of DNNs, which promotes its value for real-life applications. We conduct extensive experiments on two popular benchmarks in noisy-labeled recognition, i.e., Food-101N and Clothing1M. For both datasets, our MetaCleaner significantly outperforms baselines, and achieves the state-of-the-art performance.

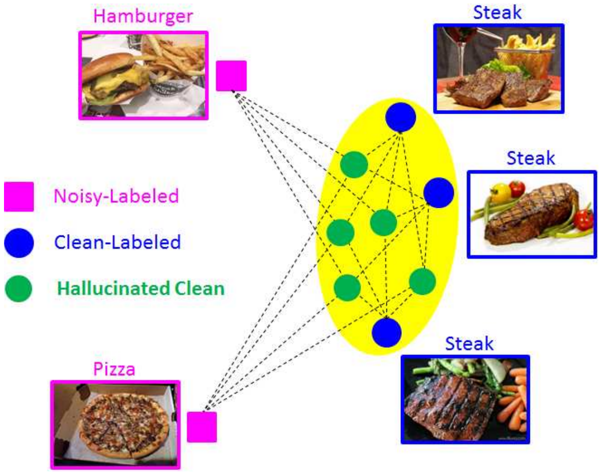


Figure 1. *MetaCleaner*. Suppose that there are 5 images altogether in the *steak* category, where 2 images are noisy-labeled, i.e., *hamburger* and *pizza*. In a training batch, we feed a randomly-sampled subset of this category (e.g., 4 images) into DNNs, and generate the semantic representations of these images. First, *MetaCleaner* compares these representations in the feature space, which can discover relations between images, and thus generate the confidence score of each image in the subset. Second, *MetaCleaner* summarizes the importance of different images in the subset to hallucinate a 'clean' representation of *steak*. This imaginary representation can improve the robustness of DNNs to label noise as well as generalize the capacity of DNNs with richer data diversity.

1. Introduction

Over the past years, visual recognition has been driven by Deep Neural Networks (DNNs) [10, 11, 32]. However, these models rely on large-scale data sets with manually-annotated labels. Collecting such data sets is expensive or time-consuming. Alternatively, a practical solution is to crawl images automatically from the internet. But these web images are noisy-labeled, e.g., a *pizza* image is often collected into the *steak* category. Many studies have shown that, such corrupted labels tend to deteriorate the classification performance of DNNs [5, 26].

One solution is label noise correction [23, 42]. However, these approaches mainly work on the label space, which requires a confusion matrix to build connections between clean and noise labels. In practice, it is often difficult and labor-intensive to obtain such statics information for large-scale data sets. To further improve effectiveness, an alternative solution has been proposed by weighting [12, 15, 25], i.e., assigning the confidence score of each image into the corresponding training loss. However, such mechanism

*Equally-contributed first authors ({wh.zhang1, yl.wang}@siat.ac.cn)
†Corresponding author (yu.qiao@siat.ac.cn)

may suffer from the following limitations. First, the confidence score is independently estimated [12, 25]. Hence, it may ignore the relations between images in this category, i.e., a key factor to depress the noisy-labeled confusion. Second, it is often challenging to distinguish between hard clean images and noisy ones. In this case, simple weighting may reduce data diversity by decreasing the importance of hard clean images [12, 15, 25]. Finally, these approaches either require the complex design of curriculum learning [12], or extra supervision of clean/verification set [15, 25]. This may limit their power in real-life applications.

To address these difficulties, we propose a conceptually simple but effective *MetaCleaner*, which can learn to hallucinate clean representations for noisy-labeled visual recognition. First, we propose a *Noisy Weighting* submodule. Instead of assigning a confidence score of each image independently into training loss, our *Noisy Weighting* can compare all the images in a small noisy subset of a specific category. This allows to discover the important relations between images, and thus provides a better estimation of confidence score for each image in the subset. Second, we introduce a *Clean Hallucinating* submodule, which can leverage the importance of different images in the subset to summarize a clean representation of the corresponding category. Via our *MetaCleaner*, DNNs can improve its robustness to label noise, as well as generalize its model capacity with richer data diversity. Moreover, *MetaCleaner* can be flexibly integrated into the standard training procedure of DNNs without any difficulties. This conciseness significantly promotes its value in practice. Finally, we evaluate *MetaCleaner* on two popular benchmarks in noisy-labeled recognition, i.e., Food-101N and Clothing1M. For both datasets, we achieve the state-of-the-art performance.

2. Related Work

Noisy-Labeled Recognition. The recent studies have shown that, DNNs often suffer from overfitting on noisy labels [5, 26, 31]. To alleviate such problems, a number of approaches have been introduced by outlier removal [19, 29, 41], weakly/semi-supervised learning [34, 40, 45, 46], knowledge distillation and transfer learning [15, 17, 18, 22], robust loss function design [8, 21, 35, 44], label prediction and correction [14, 23, 24, 31, 33, 36, 42], sample weighting [7, 12, 15, 25], and so on. One well-known solution is label prediction and correction, where the predicted label is either used as extra training supervision of DNNs [14, 24, 33, 36], or passed through a label confusion matrix to reconstruct the noisy label [23, 31, 42]. However, the reliability of predicted label is often limited, and the ground truth confusion matrix is difficult to obtain for real-life applications. To improve the effectiveness, several approaches have been recently proposed by weighting training loss of each training samples. They implicitly inherit the spirit of meta learning, i.e., learning to evaluate the importance of different images by curriculum learning [7, 12], gradient direction [25], similarity matching [15], etc. However, [7, 12] often requires a complex training procedure or predefined curriculum, which may limit its application in practice. Furthermore, [12, 25] assigns the weight of each image independently into the corresponding training loss, which may ignore the important relations between images in a category. [15] can alleviate this difficulty to some degree, by matching similarity between a query and reference embedding. However, this approach requires extra verification sets as supervision. Different from existing approaches, our *MetaCleaner* can effectively exploit the relations between different images in a random subset of a category, and leverage the importance of images to hallucinate diversified clean representations for noise reduction.

Meta Learning. Meta learning monitors the automatic learning process itself, in the context of the learning problems it encounters and tries to adapt its behavior to perform better [16]. Hence, it is also named as learning to learn. Recently, it has been highlighted for optimization and initialization [2, 4, 20], reinforcement learning [38], few-shot image recognition [9, 27, 30, 37, 39], etc. In particular, [9, 39] and our *MetaCleaner* follows the similar insight of hallucination. However, [9, 39] aims at few-shot learning, while our *MetaCleaner* works on noisy-labeled recognition. This leads to different hallucination mechanisms and training procedures. Additionally, our *MetaCleaner* shares the spirit of prototype in [30]. But different from prototypical network, our *MetaCleaner* mainly develops a robust classifier to reduce confusion of noisy labels. Hence, it adaptively uses the weighted prototype as a 'clean' representation to generalize softmax classifier, instead of using the mean prototype to construct a metric classifier for low-shot learning.

3. MetaCleaner

This section describes the proposed *MetaCleaner* for noisy-labeled image classification. Our method is partly inspired by the remarkable ability of human vision system to extract vision concepts from noisy images. Cognitive studies have shown that, humans can perceive and learn novel concepts from input images under unsupervised and noisy conditions [28]. Specifically, humans can analyze the relationship of input samples and identify which ones are reliable and important to the target concept. Then humans can leverage this knowledge to summarize the input samples for learning the target concept.

To mimic this process, we introduce a conceptually simple and generic *MetaCleaner* for training deep CNNs under noisy labels. The key idea is to hallucinate a clean representation $\mathbf{v}_c$ of the k-th object category, by randomly sampling a small subset of N noisy-labeled images in the same category, i.e., $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N}$, where $\mathbf{v}_i$ can be the feature vec-

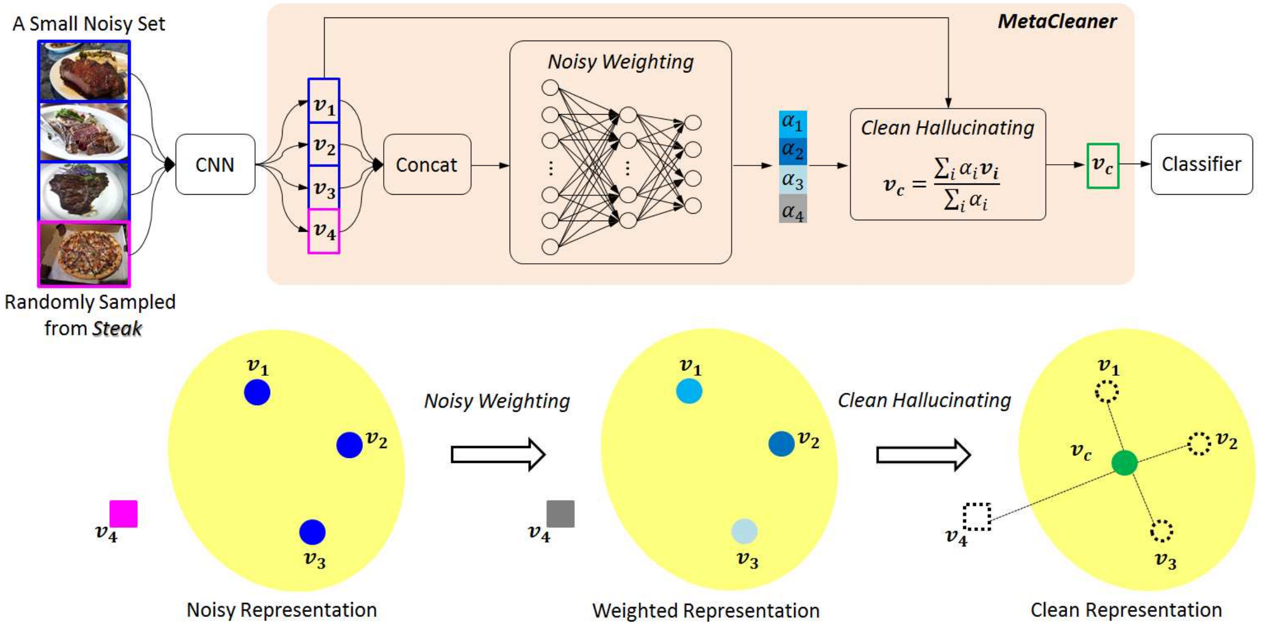


Figure 2. The whole framework of *MetaCleaner*. It can mimic the learning procedure of humans to hallucinate a clean representation from a noisy-labeled set of an object category (e.g., *steak*). To achieve this goal, *MetaCleaner* consists of two submodules, i.e., *Noisy Weighting* and *Clean Hallucinating*. First, *Noisy Weighting* is used to estimate the importance of each image, by comparing all the representations in this set. Second, *Clean Hallucinating* summarizes the weighted representations as a clean representation for classification. It is worth mentioning that, no prior knowledge (i.e., which images are noisy-labeled) is required for our *MetaCleaner*. It can adaptively depress the noisy-labeled images and highlight the clean images in an end-to-end learning framework.

tor from CNNs. Ideally, we hope that $MetaCleaner$ can identify the correctly labeled samples in $\mathbf{V}$ and summarize their representations into $\mathbf{v}_c$. Subsequently, $\mathbf{v}_c$ severs as a reliable and rich (generated from multiple samples) representation for training CNNs. Moreover, the subset $\mathbf{V}$ has a small size N for the k-th object category. Hence, it is possible to construct a huge number of subsets (i.e., C_M^N), where M is the total number of training samples for the k-th object category. This allows to construct the diversified $\mathbf{v}_c$ for learning.

Mathematically, we formulate hallucination by

$$\begin{aligned} \mathbf{v}_c &= MetaCleaner(\mathbf{V}) \\ &= \mathbf{E}[\mathbf{v} \cdot p_{\text{clean}}(\mathbf{v}|\mathbf{V})] \\ &\approx \frac{1}{\sum_{i=1}^{N} p_{\text{clean}}(\mathbf{v}_i|\mathbf{V})} \sum_{i=1}^{N} \mathbf{v}_i \cdot p_{\text{clean}}(\mathbf{v}_i|\mathbf{V}), \end{aligned} \tag{1}$$

where $p_{\text{clean}}(\mathbf{v}|\mathbf{V})$ is the conditional density function of the clean representation given the subset $\mathbf{V}$. The main challenge of using the above formula comes from how to estimate $p_{\text{clean}}(\mathbf{v}|\mathbf{V})$, since the subset $\mathbf{V}$ only includes a small number of samples with noisy labels. Statistically, it is difficult to estimate a precise p_{clean} with $\mathbf{V}$. However, $p_{\text{clean}}(\mathbf{v}_i|\mathbf{V})$ can be seen as a confident score of sample $\mathbf{v}_i$ being with correct label given subset $\mathbf{V}$. So instead of estimating the density p_{clean}, we circumvent this difficulty by directly calculating the values of $p_{\text{clean}}(\mathbf{v}_i|\mathbf{V})$.

Based on the analysis above, we design $MetaCleaner$ with two modules, i.e., *Noisy Weighting* for confidence score estimation, *Clean Hallucinating* for clean representation generation. The whole framework of *MetaCleaner* is shown in Fig. 2.

3.1. Noisy Weighting

As stated above, *Noisy Weighting* aims at estimating a confidence score α_i about whether v_i is a correctly labeled sample. For example, a *pizza* image is mistakenly collected into the *steak* category (Fig. 2). To effectively avoid confusion when learning *steak*, *Noisy Weighting* should assign a low confident score for the *pizza* image, while applying the high scores for other *steak* images.

To achieve this goal, we propose to apply a multi-layer network for confidence score estimation,

$$[\alpha_1, ..., \alpha_N] = f_{\text{Noisy Weighting}}([\mathbf{v}_1, ..., \mathbf{v}_N]), \tag{2}$$

where the input of the network is the concatenation of feature vectors in $\mathbf{V}$, and the outputs are the predicted confidence scores. Note that, $f_{\text{Noisy Weighting}}$ is a nonlinear mapping in general, which does not depend on the specific training category. In our experiment, we investigate different choices of this mapping, and empirically find that a simple two-layer MLP works well.

Furthermore, although sample weighting methods [12, 25] have been studied for noisy labels, our *MetaCleaner* differs these methods in two aspects. First, the previous methods independently estimate the confidence score of each image, w.r.t., the evaluation factors such as training loss [12], gradients [25]. They often ignore the relations between images, i.e., an important clue to discover the noisy-labeled confusion. Alternatively, our method applies a subset of one category as input, and leverages MLP for relation comparison. As a result, it can depress the noisy labels with adaptive weighting. Second, instead of reweighting the training loss with the importance of samples [25], we use the confidence scores to construct novel clean samples. Hence, our *MetaCleaner* tends to generalize DNNs with richer data diversity.

3.2. Clean Hallucinating

After obtaining the confidence scores of the noisy-labeled images in the subset, we can hallucinate (generate) a clean representation by summarizing these noisy images with their weights. We term this process as a *Clean Hallucinating* submodule in *MetaCleaner*. According to Eq. (1) and Eq. (2), we can obtain a clean representation $\mathbf{v}_c$,

$$\mathbf{v}_c = \frac{\sum_i \alpha_i \mathbf{v}_i}{\sum_i \alpha_i}, \tag{3}$$

which is treated as a training sample for classifier.

3.3. Training & Testing of MetaCleaner

Our *MetaCleaner* is a general and flexible module, which can be easily integrated into any deep classification architecture with mini-batch SGD training. In the training phase, we use *MetaCleaner* as a new layer before classifier. For each batch, we randomly select K categories. Then, for each selected category, we randomly select N examples as subset. For each subset, we can use *MetaCleaner* to hallucinate a clean representation for training. In this way, we can generate diversified samples with reliable labels for different batches. Additionally, when we perform concatenation, the order of image representations is random in a subset. As the number of batches increases, all possible orders tend to be enumerated. This allows our MetaCleaner to be generalized well in the training, avoiding the influence of particular orders.

After training with the hallucinated clean features from *MetaCleaner*, we expect that softmax classifier has been gradually generalized well for learning how to recognize an object in the image. Hence, we propose to remove the *MetaCleaner* layer in the testing phase, and feed the feature vector of a testing sample directly into softmax classifier for visual recognition. But if the test input is a noisy-labeled set (not individual samples), one can still apply the trained *MetaCleaner* to improve the performance.

3.4. Prototypical Interpretation of MetaCleaner

MetaCleaner aims at learning to hallucinate a representative embedding from a noisy subset of one category. It implicitly inherits the spirit of *Prototypical Network* [30], where the cluster center of a category is used as a prototype for low-shot learning. Next, we interpret the connections and clarify the difference between these two meta learners.

In the meta training period, both meta learners exploit deep CNNs to generate the semantic representations. However, due to different learning goals, these two meta learners introduce different training strategies. *Prototypical Network* aims at low-shot learning. Hence, it mainly establishes a metric classifier to reduce overfitting, i.e., this classifier directly assigns the class probabilities of a query, based on its distance from the mean representation (prototype) of each selected category. Different from *Prototypical Network*, *MetaCleaner* aims at noisy-labeled recognition. Hence, it estimates the confidence scores of input samples, and hallucinates a clean representation by using these scores. During the end-to-end training, the weights of different images can be adaptively adjusted to generalize softmax classifier for noisy-labeled images.

In the meta testing period, *Prototypical Network* aims at classifying test images of new categories, given a low-shot support set of these categories. To achieve this goal, it leverages the support set to generate the prototypes of new categories, and uses the metric based classifier to recognize test images. Alternatively, *MetaCleaner* aims at improving the robustness of softmax classifier to noisy labels. Hence, in its testing phase, one can feed the feature vector of an image directly into softmax classifier for visual recognition.

4. Experiment

Data Sets. In this paper, we mainly evaluate our *MetaCleaner* on two popular benchmarks for noisy-labeled visual recognition, i.e., **Food-101N** [15] and **Clothing1M** [42]. (1) Food-101N consists of 310k/25k train/test images, with 101 food categories of the original Food-101 data set [1]. Around 80% of train set is correctly labeled, and 55k/5k training/test images contain extra noise verification labels. (2) Clothing1M consists of 1M/14k/10k train/val/test images with 14 fashion classes. Around 61.54% of train set is correctly labeled. Furthermore, there is an extra clean training set with around 50k images, where around 25k images contain both clean and noisy labels. Due to the fact that the categories in Food-101N and Clothing1M are fine-grained with large confusions, they are two challenging benchmarks for noisy-labeled recognition.

Implementation Details. Unless stated otherwise, we perform *MetaCleaner* as follows. **First**, we use ResNet50 as CNN backbone. For each image, we generate a semantic representation (after global pooling), i.e., 2048-dim fea-

Method	Food-101N	Clothing1M
Baseline	81.44	68.94
MetaCleaner	**82.52**	**72.50**

Table 1. *Baseline* vs. *MetaCleaner*. Baseline is the standard CNN without *MetaCleaner*.

Operations	Food-101N	Clothing1M
Constant	81.24	70.41
Attention$_{sig}$	81.44	71.08
Attention$_{exp}$	81.78	70.68
FC-FC-Sig	82.09	71.15
FC-FC-ReLU-Sig	**82.52**	**72.50**
FC-ReLU-FC-ReLU-Sig	82.18	71.67

Table 2. *Noisy Weighting*. (I) Constant. We assign the importance score $\alpha_i = 1$ for each image in the small subset. In this case, the hallucinated representation in Eq. (3) is reduced as the mean of noisy-labeled representations. (II) Attention. We use two widely-used attention mechanisms as the weighting operation, i.e., $\alpha_i = Sigmoid[\mathbf{a} \cdot tanh(\mathbf{W}\mathbf{v}_i + \mathbf{b})]$ and $\alpha_i = Exp[\mathbf{a} \cdot tanh(\mathbf{W}\mathbf{v}_i + \mathbf{b})]$. (III) FC. We use the FC layers as the weighting operation. Since α_i is the confidence score, we use sigmoid (Sig) in the last layer as a soft gate. Furthermore, we explore the role of ReLU in different layers. More explanations can be found in Section 4.1.

Operations	Food-101N	Clothing1M
(I) Loss Reweight	78.77	68.89
(II) Hallucination	**82.52**	**72.50**

Table 3. *Clean Hallucinating*. (I) Loss Reweight. Since α_i is the confidence score of each image, we multiply α_i with the training loss of the corresponding image for reweighting. (II) Hallucination. It is the hallucinated representation in Eq. (3), which is the weighted sum of original representations. As expected, (II) outperforms (I). It illustrates that, the proposed hallucination is more robust to noisy labels.

ture vector. The neural network structure in *Noisy Weighting* is two-layer, i.e., FC-FC-ReLU-Sigmoid. The input / output dimension of the 1st FC layer is $N \times 2048/384$ (Food-101N), $N \times 2048/512$ (Clothing1M). The input / output dimension of the 2nd FC layer is $384/N$ (Food-101N), $512/N$ (Clothing1M). For Food-101N/Clothing1M, the size of small subset N in *MetaCleaner* is 4/4 per category, and the batch size is 480/256. **Second**, for both data sets, we just use the noisy-labeled train set for training our model, and report the classification accuracy on the test set. **Third**, we implement our network by PyTorch, where we use the standard SGD, the momentum is 0.9, weight decay is 0.001/0.005 for Food-101N/Clothing1M. Furthermore, the initial learning rate is 0.01. It is divided by 10 at each 20/5 epoches, and the training procedure is finished with 80/20 epoches for Food-101N/Clothing1M.

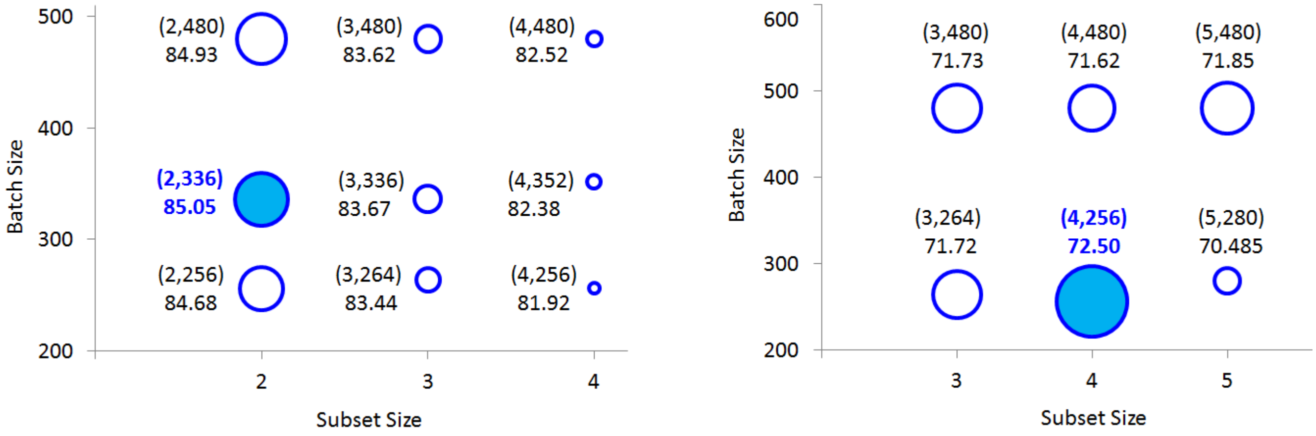
4.1. Ablation Studies

To investigate the properties of our *MetaCleaner*, we mainly evaluate its key model submodules. For fairness, when we explore different strategies of one submodule, others are with the basic strategy in the implementation details.

Baseline* vs. *MetaCleaner. First of all, we compare our *MetaCleaner* with *Baseline*, i.e., the standard CNN without *MetaCleaner*. As shown in Table 1, *MetaCleaner* significantly outperforms *Baseline*, showing the essentiality of *MetaCleaner*.

Noisy Weighting. We investigate different weighting operations for *Noisy Weighting*. (I) Constant. We does not use any weight operations, and assign the importance score $\alpha_i = 1$ for each image in the small group. In this case, the hallucinated representation in Eq. (3) is reduced as the mean of noisy-labeled representations. (II) Attention. We use two widely-used attention mechanisms as the weighting operation. Attention$_{sig}$ and Attention$_{exp}$ are respectively formulated as i.e., $\alpha_i = Sigmoid[\mathbf{a} \cdot tanh(\mathbf{W}\mathbf{v}_i + \mathbf{b})]$ and $\alpha_i = Exp[\mathbf{a} \cdot tanh(\mathbf{W}\mathbf{v}_i + \mathbf{b})]$ with the parameter set $\{\mathbf{a} \in \mathbb{R}^{1\times384}, \mathbf{W} \in \mathbb{R}^{384\times2048}, \mathbf{b} \in \mathbb{R}^{384\times1}\}$ for Food-101N, $\{\mathbf{a} \in \mathbb{R}^{1\times512}, \mathbf{W} \in \mathbb{R}^{512\times2048}, \mathbf{b} \in \mathbb{R}^{512\times1}\}$ for Clothing-1M. (III) FC. We use the FC layers as the weighting operation. Since α_i is the importance score, we use sigmoid (Sig) in the last layer as a soft gate. Furthermore, we explore the role of ReLU in different layers. The results are shown in Table 2. **First**, the attention setting outperforms the constant setting. It illustrates that, it is necessary to estimate the importance of different images before hallucinating the clean representation. **Second**, the FC setting outperforms the attention setting in general. The main reason is that, the input of attention is an individual representation of each image. Hence, it may lack the capacity of discovering the relations between different images. Alternatively, the input of FC is the concatenation of all the representations in a subset. As a result, the FC layers can produce the importance of different images by similarity comparison. **Finally**, we investigate ReLU in the FC layers. In Table 2, FC-FC-ReLU-Sig outperforms FC-FC-Sig. The main reason is that, Food-101N and Clothing 1M often exhibit the fine-grained characteristics, i.e., many noisy-labeled images look similar to the clean-labeled images in a category. In other words, the importance α_i of a noisy-labeled image may not be low in the subset. The design of FC-FC-ReLU-Sig allows α_i to be ranged from 0.5 to 1, which reasonably takes the importance of noisy-labeled images into account. On the contrary, α_i of FC-FC-Sig is ranged from 0 to 1. The neural network tends to underestimate the importance of noisy-labeled images. Furthermore, FC-FC-ReLU-Sig outperforms FC-ReLU-FC-ReLU-Sig. It may be because the first layer is used for dimensionality reduction, which can be effectively achieved by linear transformation.

Clean Hallucinating. We explore different hallucination



(a) Food-101N

(b) Clothing1M

Figure 3. Batch Size & Subset Size. Note that, batch size = subset size × the number of sampled categories in the batch, i.e., batch size has to be divided exactly by subset size. In this case, when we change subset size, we have to change batch size slightly. First, when using the comparable batch size, the performance with different subset sizes tends to have notable fluctuations for most cases. Hence, subset size may be more important for noisy-labeled recognition. Second, subset size can be strongly relevant to the level of noise. The noise level of Clothing1M ($\sim$ 40%) is higher than the one of Food-101N ($\sim$ 20%). As a result, Clothing1M requires a larger subset size than Food-101N. In Fig. 4, we further investigate the relation between subset size and noise level.

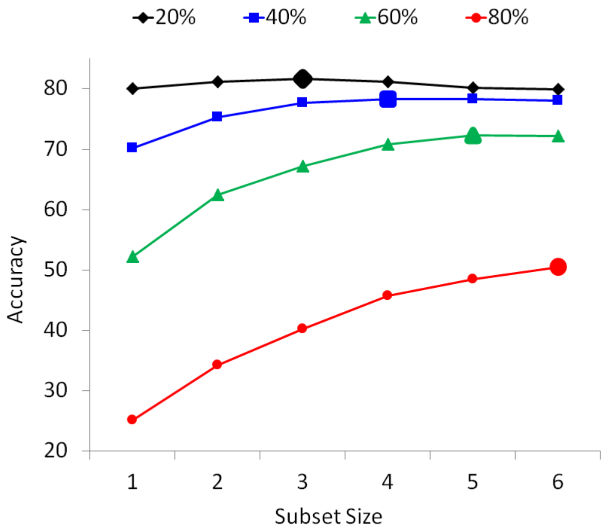


Figure 4. Noise Level. Specifically, we add random noise on the labels of original Food-101 data set. As expected, when the noise level is higher, larger subset size allows to hallucinate a more robust representation, and significantly boosts accuracy. More explanations can be found in Section 4.1.

operations. (I) Loss Reweight. Since α_i is the confidence score of each image, we multiply α_i with the training loss of the corresponding image for reweighting. (II) Hallucination. It is the hallucinated representation in Eq. (3), which is the weighted sum of original representations. As expected, the proposed hallucination outperforms Loss Reweight in Table 3. It illustrates that, the weighted sum of noisy representation is more robust to noisy labels.

Batch Size & Subset Size. We investigate two important hyper-parameters in our *MetaCleaner*, i.e., subset size and batch size. Note that, when exploring different sizes of subset, we should fix batch size. However, batch size = subset size × the number of sampled categories in the batch, i.e., batch size has to be divided exactly by subset size. In this case, when we change subset size, we have to change batch size slightly. The results are shown in Fig. 3. **First**, when using the same value of subset size, the performance with different batch sizes tends to be comparable. On the contrary, when using the comparable value of batch size, the performance with different subset sizes varies largely. It illustrates that, subset size is more important for noisy-labeled recognition. **Second**, subset size tends to be different among data sets, i.e., it is 2/4 for the best performance of Food-101N/Clothing1M. This observation illustrates that, subset size may be strongly relevant to the level of noise. As mentioned in the data description, the noise level of Clothing1M ($\sim$ 40%) is higher than the one of Food-101N ($\sim$ 20%). Hence, Clothing1M requires a larger subset size than Food-101N. In the next, we further investigate the relation between subset size and noise level.

Noise Level. To investigate the influence of noise level, we add random noise on the original Food-101 [1] data set. For example, 20% noise level denotes that, we uniformly sample 20% training set, and randomly flip the correct label into another category. The results are shown in Fig. 4, where batch size is 480 for all the cases. **First**, when the noise level is low, the performance tends to be compara-

Method	Data	Acc
Softmax [15]	Food-101	81.67
Softmax [15]	Food-101N	81.44
Weakly Supervised [46]	Food-101N	83.43
CleanNet (w_{hard}) [15]	Food-101N+VF(55k)	83.47
CleanNet (w_{soft}) [15]	Food-101N+VF(55k)	83.95
Our *MetaCleaner*	Food-101N	**85.05**

Table 4. Comparison with The-State-of-The-Art (Food-101N). VF(55k) is the noise-verification set used in CleanNet [15].

Method	Data	Acc
Softmax [23]	Noise1M	68.94
LossCorrect [23]	Noise1M	69.84
Weakly Supervised [46]	Noise1M	71.36
JointOptim [33]	Noise1M	72.23
Our *MetaCleaner*	Noise1M	**72.50**
CleanNet (w_{hard}) [15]	Noise1M+Clean(25k)	74.15
CleanNet (w_{soft}) [15]	Noise1M+Clean(25k)	74.69
Our *MetaCleaner*	Noise1M+Clean(25k)	**76.00**
CleanNet (w_{soft}) [15]	Noise1M+Clean(50k)	79.90
LossCorrect [23]	Noise1M+Clean(50k)	80.38
Our *MetaCleaner*	Noise1M+Clean(50k)	**80.78**

Table 5. Comparison with The-State-of-The-Art (Clothing1M). Clean(25k) is used in CleanNet [15] to obtain the verification set. To keep same data setting, we train our *MetaCleaner* on 1M noisy training set, and then fine-tune it on 25k clean images. Furthermore, we achieve the state-of-the-art performance on the setting of Noise1M+Clean(50k), even though other approaches use extra verification labels [15], extra label confusion information [23].

ble among different subset sizes. But still, the case with a subset of images is better than the case with individual images (subset size=1), i.e., baseline without *MetaCleaner*. For instance, the accuracy of (subset size=3) is the best (acc: 81.67) in the 20% noise setting. It outperforms the baseline (subset size=1, acc: 80.11%). This observation demonstrates, hallucination with small image sets is important for noisy-labeled recognition. **Second**, when the noise level is higher, larger subset size allows to hallucinate a more robust representation, and significantly boosts accuracy.

4.2. Comparison with The-State-of-The-Art

For fairness, our comparisons are based on the same CNN backbone, i.e., all the approaches are built upon ResNet50 (for Food-101N / Clothing1M / ImageNet) and WideResNet-28-10 (for CIFAR-10).

Food-101N. As shown in Table 4, *MetaCleaner* significantly outperforms the softmax baseline. More importantly, it outperforms the state-of-the-art CleanNet [15], without using the extra 55k noise-verification images. It shows the robustness of *MetaCleaner* to noisy labels.

Approach	Data	Acc
Baseline [25]	Noisy	67.97
Reed-Hard [24]	Noisy	69.66
S-Model [6]	Noisy	70.64
MentorNet [12]	Noisy	76.60
Reweight [25]	Noisy+Clean(1k)	86.92
Our *MetaCleaner*	Noisy	**90.09**

Table 6. Comparison with The-State-of-The-Art (Cifar-10). We perform our *MetaCleaner* for Cifar-10, by adding 40% noise ratio with uniform flip [25].

Clothing1M. We mainly demonstrate comparisons in Table 5, according to the usage of different training sets. **First**, when only using 1M noisy training set, our *MetaCleaner* slightly outperforms the state-of-the-art JointOptim [33]. But we claim that, the training procedure of our *MetaCleaner* is the standard optimization of CNN, while JointOptim requires an alternating optimization procedure with careful regularization. Hence, our *MetaCleaner* is a more practical solution for noisy-labeled recognition. **Second**, we compare *MetaCleaner* with CleanNet [15], which requires 25k clean images to obtain the verification set. To keep the same data setting, we train our *MetaCleaner* on 1M noisy training set, and then fine-tune it on 25k clean images. In Table 5, our *MetaCleaner* outperforms CleanNet, showing its effectiveness. **Finally**, we compare *MetaCleaner* with different approaches, where all the clean training set is available. Same as before, we train our *MetaCleaner* on 1M noisy training set, and then fine-tune it on 50k clean images. As one can see that, *MetaCleaner* achieves the state-of-the-art performance in this setting, even though other approaches use extra verification labels [15], extra label confusion information [23].

Cifar-10. We perform our *MetaCleaner* on Cifar-10 [13], by adding 40% noise ratio with uniform flip [25]. We use the same backbone (WideResNet-28-10 with dropout 0.3) as [25]. Additionally, after training with *MetaCleaner*, we unload it and further fine-tune CNN for feature generalization. As shown in Table 6, our *MetaCleaner* outperforms all the state-of-the-art approaches. Moreover, it is better than [25], which uses an extra clean data set (1k). This illustrates that *MetaCleaner* is more robust to noisy labels.

ImageNet. We perform our *MetaCleaner* on ImageNet [3], by adding 40% noise ratio with uniform flip. After training with *MetaCleaner*, we unload it and further fine-tune CNN for feature generalization. The top-1 accuracy is 66.47 / 69.12 for ResNet50 without / with *MetaCleaner*. It further shows the power of *MetaCleaner* for large-scale noisy-labeled recognition.

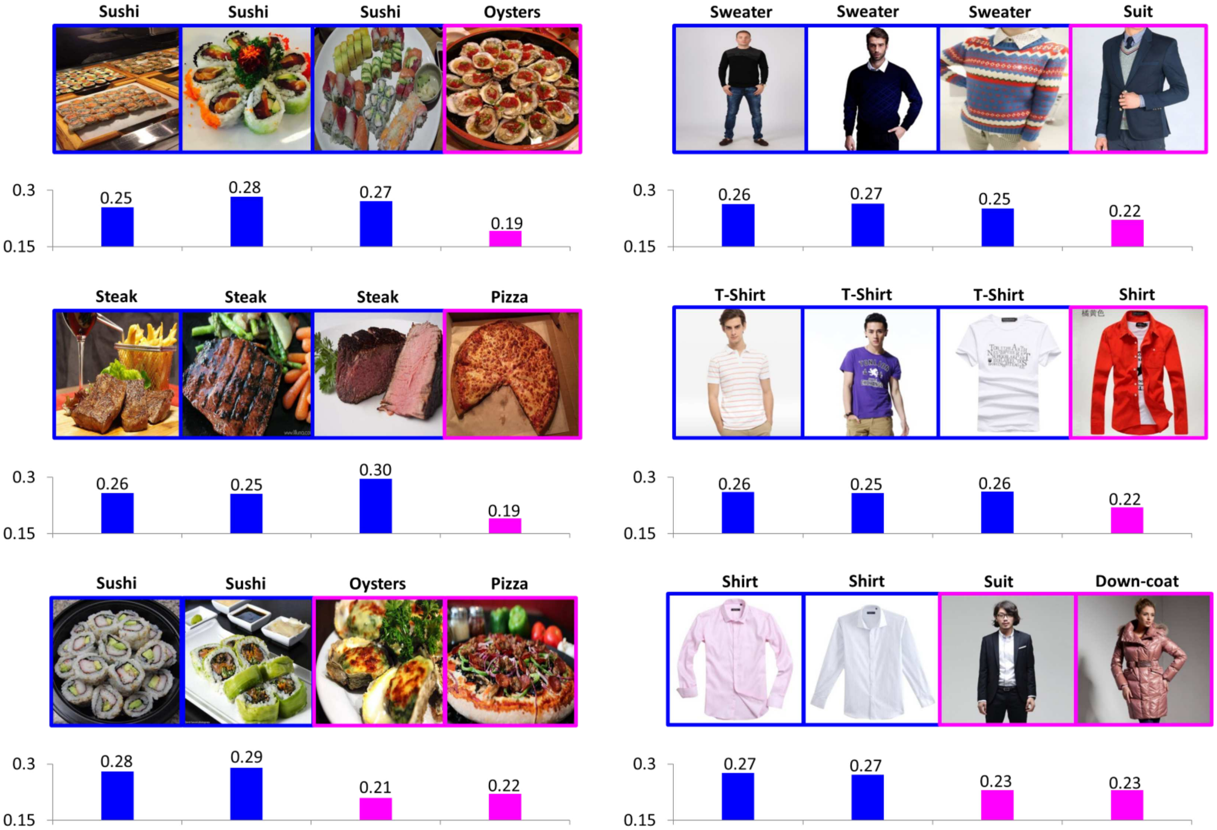


Figure 5. Visualization. We show the importance of different images in the subset, where we demonstrate the normalized score $\alpha_i/(\sum_i \alpha_i)$ for hallucination in Eq. (3). On one hand, our *MetaCleaner* can effectively depress the negative influence of noisy-labeled images, by reducing the importance scores of these images. On the other hand, our *MetaCleaner* can aware the fine-grained characteristics of Food-101N and Clothing1M, i.e., the noisy-labeled images look similar to the clean ones. Hence, it can reasonably assign the lower scores on these images, but do not delete their contribution completely.

4.3. Visualization

We visualize the importance of different images in the subset, where we demonstrate the normalized score $\alpha_i/(\sum_i \alpha_i)$ for hallucination in Eq. (3). The results are shown in Fig. 5. On one hand, our *MetaCleaner* can effectively depress the negative influence of noisy-labeled images, by reducing the importance scores of these images. On the other hand, our *MetaCleaner* can aware the fine-grained characteristics of Food-101N and Clothing1M, i.e., the noisy-labeled images look similar to the clean ones. Hence, it can reasonably assign the lower scores on these images, but do not delete their contribution completely.

5. Conclusion

In this paper, we propose a flexible *MetaCleaner*, which can learn to hallucinate clean representations for noisy-labeled visual recognition. It mainly consists of two submodules. First, *Noisy Weighting* compares semantic representations in a randomly-sampled image subset of a category. Via exploiting relations between images, it can estimate the importance of each image in the subset. Then, *Clean Hallucinating* summarizes a clean representation by taking the weight of different image representations into account. As a result, our *MetaCleaner* can improve the robustness of DNNs to noisy labels. More importantly, it can generalize the capacity of DNNs by richer data diversity and variations. We mainly evaluate *MetaCleaner* on Food-101N and Clothing1M, where it achieves the state-of-the-art performance on both benchmarks. In the future, it would be interesting to further investigate the theoretical aspects of *MetaCleaner*, such as Vicinal Risk Minimization in [43].

Acknowledgements. This work is partially supported by National Natural Science Foundation of China (61876176), Shenzhen Research Program (JCYJ 20150925163005055, CXB201104220032A), the Joint Lab of CAS-HK, State Joint Engineering Laboratory for Robotics and Intelligent Manufacturing funded by National Development and Reform Commission (NO. 2015581).

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[2] Yutian Chen, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matt Botvinick, and Nando de Freitas. Learning to learn without gradient descent by gradient descent. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 748–756, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[5] Benoit Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

[6] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2017.

[7] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *ECCV*, 2018.

[8] Bo Han, Ivor W. Tsang, and Ling Chen. On the convergence of a family of robust losses for stochastic gradient descent. In *ECML/PKDD*, 2016.

[9] Bharath Hariharan and Ross B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3037–3046, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *arXiv:1512.03385*, 2015.

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[12] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

[13] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[14] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.

[15] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *CVPR*, 2018.

[16] Christiane Lemke, Marcin Budka, and Bogdan Gabrys. Metalearning: a survey of trens and technologies. *Artificial Intelligence Review*, 2015.

[17] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] Or Litany and Daniel Freedman. Soseleto: A unified approach to transfer learning and training with noisy labels. *CoRR*, abs/1805.09622, 2018.

[19] Wei Liu, Gang Hua, and John R. Smith. Unsupervised one-class learning for automatic outlier removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[20] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2113–2122, Lille, France, 07–09 Jul 2015. PMLR.

[21] Volodymyr Mnih and Geoffrey E. Hinton. Learning to label aerial images from noisy data. In *ICML*, 2012.

[22] Li Niu, Qingtao Tang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Learning from noisy web data with category-level supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

[24] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep convolutional networks on noisy labels with bootstrapping. In *ICLR Workshop*, 2015.

[25] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018.

[26] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. In *arXiv:1705.10694*, 2018.

[27] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[28] Richard A. Schmidt and Robert A. Bjork. New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 1992.

[29] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, 1999.

[30] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPs*, 2017.

[31] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. In *ICLR Workshop*, 2015.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[33] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *CVPR*, 2018.

[34] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *CoRR*, abs/1706.00038, 2017.

[35] Brendan van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *NIPS*, 2015.

[36] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[37] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPs*, 2016.

[38] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matthew Botvinick. Learning to reinforcement learn. *CoRR*, abs/1611.05763, 2017.

[39] Yu-Xiong Wang, Ross B. Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *CoRR*, abs/1801.05401, 2018.

[40] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep learning via semi-supervised embedding. In *ICML*, 2008.

[41] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[42] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.

[43] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[44] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *CoRR*, abs/1805.07836, 2018.

[45] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.

[46] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: A weakly-supervised deep learning framework for learning from web data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.