Deep Image Retrieval is not Robust to Label Noise

Stanislav Dereka
Tinkoff
st.dereka@gmail.com

Ivan Karpukhin
Tinkoff
i.a.karpukhin@tinkoff.ru

Sergey Kolesnikov
Tinkoff
scitator@gmail.com

Abstract

Large-scale datasets are essential for the success of deep learning in image retrieval. However, manual assessment errors and semi-supervised annotation techniques can lead to label noise even in popular datasets. As previous works primarily studied annotation quality in image classification tasks, it is still unclear how label noise affects deep learning approaches to image retrieval. In this work, we show that image retrieval methods are less robust to label noise than image classification ones. Furthermore, we, for the first time, investigate different types of label noise specific to image retrieval tasks and study their effect on model performance.

1. Introduction

Over the last decade, deep learning achieved impressive results in multiple computer vision domains, including image classification [16], image retrieval [2], and face verification [6]. One reason behind the success of deep learning is the ability of deep neural networks to extract useful information from large amounts of annotated data. It was shown that increasing the amount of data along with model size leads to an improvement in prediction quality [13, 31]. To do so, datasets with tens of thousands and even millions of items were collected for deep learning [8, 18]. While the majority of data can usually be obtained from the Internet, annotating this data is the most laborious part of the data preparation process [28]. When provided with millions of items, it is almost impossible to manually annotate all of them. Multiple semi-supervised approaches were created to reduce the amount of manual work required in data annotation [24]. However, both manual annotation and semi-supervised algorithms can introduce errors to final results. It was shown recently that the annotation error rate in large-scale datasets can exceed 40% [33].

Previous research on label noise in computer vision was primarily focused on image classification [1, 26]. In this paper, we study the performance of deep convolutional neural networks in image classification and retrieval tasks when working with label noise. The core contributions of the paper can be summarized as follows:

1. We study the effects of label noise in image retrieval on Stanford Online Products [25], In-shop [21] and face recognition datasets [8, 11] along with image classification on ImageNet [28]. Our results show that image retrieval models are less robust to label noise than image classification models under similar training conditions.

2. We study different retrieval-specific noise types with respect to the number of mislabeled samples and corrupted classes. Our findings show that the amount of corrupted classes affects performance more than the proportion of corrupted samples in each class.

2. Related Work

In deep computer vision, label noise was primarily studied in the context of image classification [1, 26]. Many techniques were proposed to reduce the effects of label noise on the final models' performance. For example, one of these methods involves automatically detecting mislabeled items and excluding them from training [17, 27, 29]. Another way to deal with label noise is to design specialized training algorithms [3, 5, 20]. While these methods can, to some degree, reduce the destructive impact of annotation errors, they are still affected by the labels' quality. While we find these approaches valuable for the community, we argue for deeper model robustness analysis in image retrieval.

Several previous works addressed the general question of deep image classification robustness to label noise [13, 15, 31, 37]. It was shown that modern convolutional neural network (CNN) architectures, such as ResNet [10], are robust to label noise in image classification. While the models' final accuracy reduces with the growing noise level, the model still trains even with five times more noisy samples than clean ones. Similar studies were recently performed for image retrieval and face recognition [4, 12, 19, 33], showing that metric learning algorithms are more sensitive to label noise than classification methods. While there is a fundamental difference between classification and metric learning approaches when it comes to the effects of label noise, the reason for this difference is still unclear. To address the issue of earlier works using different network architectures, training strategies, and noise types, we compared retrieval, classification, and face recognition methods under similar training conditions. Also, we study novel types of noising schemes designed for image retrieval.

3. Label Noise in Image Retrieval

Many image retrieval datasets are gathered using unreliable methods such as search engine crawling [4, 9, 33], which leads to label noise and degrades learning [33]. Manual cleaning of large-scale datasets with millions of samples can be expensive and resource demanding [33]. According to previous research, data annotation techniques used in image retrieval can produce specific types of label noise [19, 33]. In particular, noise can be concentrated in clusters of similar images leading to non-uniform distribution of noisy labels among classes. Moreover, well-studied datasets, e.g. Labeled Faces in the Wild (LFW) [1] [11], include ”trash” classes which completely consist of mislabeled samples. We hereby raise a question of how the number of corrupted (trash) classes affects final model performance? We call class corrupted if almost all items in this class have the wrong label. This question differs from previous research in image classification, where some amount of clean data needs to be provided for each class.

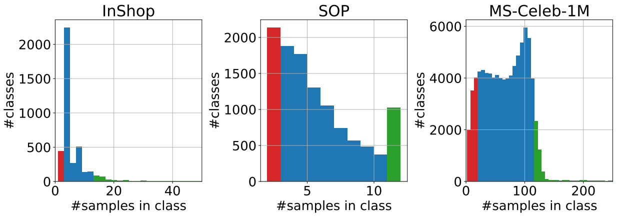

Figure 1. Class size distributions in popular image retrieval datasets. Examples of classes affected by small and large class label noise are highlighted in red and green respectively.

When modeling label noise patterns, there are multiple ways to choose classes for corruption. As shown in Figure 1, image retrieval datasets are highly class-imbalanced. Large classes can contain thousands of times more samples than small ones. By selecting a fixed amount of noisy items, we can choose either many small classes or a small number of large classes. Based on this observation, we propose two new label noising schemes for image retrieval.

In the proposed **large class label noise** we select the most frequent classes for a given proportion of samples in a dataset and shuffle the labels among the samples of the selected classes. The number of classes to corrupt is selected to match the expected number of corrupted samples. **Small class label noise** is similar to large class label noise, but the rarest classes are corrupted instead of large ones. The dataset parts which were corrupted by each noise pattern are illustrated on histograms in Figure 1. The proposed patterns are opposites regarding the number of totally corrupted classes. Large class label noise minimizes the number of corrupted classes while the number of corrupted samples is fixed. In contrast, small class label noise maximizes the number of corrupted classes.

As a baseline, we study traditional **uniform label noise** [1], also known as label flipping [33]. In this noise pattern, a given proportion of samples in a dataset is randomly selected, and a random label from all the available labels is assigned to the selected samples. This noise pattern slightly corrupts each class in the dataset with a probability proportional to the number of samples in a class.

4. Experiment Setup

4.1. Datasets and Evaluation Metrics

The scope of this paper covers both image retrieval and image classification datasets. We include the image classification task in our work to compare its robustness to label noise with image retrieval. Closed-set classification datasets share the same set of labels between training and testing. The goal of an algorithm in this case is to predict label for each input image. Accuracy is usually used to evaluate classification performance. For *image classification* we use the ImageNet [28].

The goal of *image retrieval (IR)* is to find images from the gallery collection which are similar to the query image [23, 25]. Let’s call retrieval successful for some query if the most similar image in the gallery has the same label. We measure Recall@1 (R@1) during evaluation, which is equal to the fraction of successful retrievals.

Datasets for image retrieval usually have class-disjoint training and testing parts. This setup is also called open-set problem. In our experiments we use Stanford online products (SOP) [30] and In-shop Clothes (InShop) [21] datasets. Both of these datasets have severe class imbalance as shown in Figure 1.

In addition to the above, we extend the scope of our work by adding *face recognition (FR)* datasets. The majority of face recognition datasets has class imbalance similar to other retrieval datasets. Furthermore, face recognition tasks are open-set and can be subject to common image retrieval problems [35]. Finally, public FR datasets are large-scale and make it possible for the experiment setup to be close to real-world practical applications. For training, we use the MS-Celeb-1M [8] dataset, which contains over 10M images of around 100K individuals. For face recognition evalua-

[1] See known LFW errors on the official website: http://vis-www.cs.umass.edu/lfw/

tion, we chose the LFW [11] dataset. In addition to computing R@1, we evaluate *face verification* performance. LFW verification test set includes 6K image pairs labeled with 0 (same person) or 1 (different people). During testing, a similarity in model embedding space is computed for each pair. For face verification, we then compute TPR@FPR metric for 10^{-3} FPR by using a common FR approach [14, 35].

4.2. Training Details

In image retrieval and classification datasets (SOP, Inshop, ImageNet) we preprocess images using approaches from previous papers [10, 23]. Images in ImageNet, SOP, and InShop datasets are resized to 256 pixels by the shortest side. During training, each image is randomly cropped to 224x224 pixels. During testing, we use central crop. In MS-Celeb-1M v2 dataset [6], the size of aligned images is 112x112 pixels in both training and testing. We use random horizontal flip augmentation for all datasets.

In all experiments, we use ResNet50 [10] CNN architecture as a backbone network. ResNet50 is initialized with ImageNet pretrain weights for all the tasks, except for ImageNet and MS-Celeb-1M, where random initialization is used. The embedding size is set to 512 in all experiments.

Our models are trained with the ArcFace [6] loss function. According to recent works [6, 23], this is one of the best-performing loss functions in considered benchmarks. We use the default ArcFace scale and margin hyperparameters from the original paper for training on SOP, InShop, and MS-Celeb-1M. For ImageNet, we use Normalized Softmax loss [34].

We use the stochastic gradient descent (SGD) optimizer with momentum 0.9 and weight decay 0.0001. The initial learning rate is set to 0.1. We follow the pipeline from [6] and perform 16 epochs of training, decreasing the learning rate by a factor of 0.1 at the 9th and 14th epochs. In all experiments we set batch size to 256.

5. Experiment Results

5.1. Robustness Comparison of Classification and Retrieval Tasks

The goal of our experiments is to measure the robustness of image retrieval tasks for comparison with image classification. We quantify robustness as a drop in performance when a model is trained on a noisy dataset compared to training on a clean one. We compare models trained with multiple noise levels ranging from 0 (clean dataset) to 10%. The results are presented in Table 1.

Dataset	Metric	Label noise level			
		clean	0.01	0.05	0.1
ImageNet	Accuracy	67.32	67.09	66.42	65.80
InShop	R@1	84.61	82.09	74.70	72.70
SOP	R@1	63.49	63.68	60.72	58.78
LFW	R@1	67.22	67.36	≈0.0	≈0.0
LFW	TPR@10^{-3}	99.27	99.23	≈0.0	≈0.0

Table 1. Uniform label noise effects on model performance for image classification (ImageNet) and verification tasks compared to image retrieval tasks. All metric values are shown in %. We label cases where the model was unable to generalize on test set with ≈.

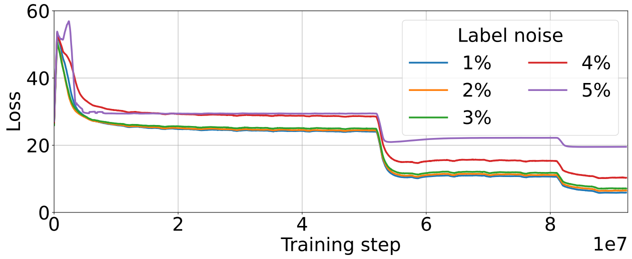

Figure 2. Training curves for MS-celeb-1M dataset with different label uniform noise levels.

On all retrieval tasks, we observe a rapid decrease of quality with the increase of uniform label noise level. Even 5% of label noise causes a relative performance drop of up to 10%. In contrast, classification accuracy under equal conditions shows a much smaller drop. These results show that image classification is more robust to uniform label noise than image retrieval.

One of the most interesting findings from our experiments is that, when a model is trained on the MS-Celeb-1M dataset with at least 5% label noise, its score is close to that of an untrained model. In order to clarify this phenomena, we show training curves for several noise levels in Figure 2. It can be seen that at 5% label noise, loss function optimization still occurs, which means that the model training still converges. These findings allow us to conclude that image retrieval setups are extremely sensitive to label noise and their performance can be affected drastically by only a small fraction of label noise.

5.2. Effects of Noise Patterns

According to the previous experiment, image retrieval is not robust to label noise. Therefore, we can raise a question of how the number of corrupted classes affects retrieval performance. To answer this question, we corrupted a training set using new types of label noise for image retrieval, previously described in section 3. Given a fixed set of noisy items, we can either corrupt a small number of large classes or many small classes. These noising approaches are called large-class noise and small-class noise respectively. The total amount of corrupted items and corrupted classes in our experiments is given in Table 2.

It can be seen from Figure 3 that the more training set

Noise type	Dataset	Noise level		
		0.01	0.05	0.1
		Corrupted classes, %		
Uniform	InShop	6.0	26.0	44.3
Large Classes	InShop	0.05	0.5	1.6
Small Classes	InShop	2.5	11.1	19.2
Uniform	SOP	5.0	22.6	40.0
Large Classes	SOP	0.4	2.2	4.4
Small Classes	SOP	2.6	13.2	23.8
Uniform	MS-Celeb-1M	45.5	88.0	95.5
Large Classes	MS-Celeb-1M	0.2	1.4	4.2
Small Classes	MS-Celeb-1M	7.3	19.5	29.1

Table 2. Percentage of corrupted classes for each dataset and noise pattern for three noise levels. A class is considered corrupted if at least one label in this class has been flipped during the noising procedure.

classes are affected by label noise, the higher is the drop in performance, even if the total amount of corrupted elements doesn't change. For example, the noise level 5% on the InShop dataset leads to a 10% quality drop for uniform noise with 26% corrupted classes. On the other hand, large and small class noises with the same 5% noise level produce only 2% and 4.5% quality drop, while the numbers of corrupted classes for these noises are 0.05% and 2.5% respectively (Table 2).

As a conclusion of these experiments, small class label noise leads to more significant degradation than large class label noise. The effect of uniform label noise is the strongest, as it affects more classes than the other noise types. It's worth noting that thoroughly corrupted classes affect performance less than a large amount of slightly damaged classes.

6. Discussion and Future Work

According to our experiments, image retrieval is less robust to label noise than image classification. We used similar models, training objectives, and optimization methods for both tasks. The difference in robustness can lie in contrast between closed-set and open-set tasks. Image retrieval models need to generalize well to unseen classes, while classification tasks share the same set of labels between training and testing. Our experiments show that while models converge with all considered noise levels, the test set accuracy completely drops even for 5% label noise. These results bring us to the conclusion that image retrieval models fail to generalize to unseen classes in this case, highlighting the need for further investigation on label noise robustness in image retrieval and other open-set tasks.

The goal of this work is to highlight difference between classification and retrieval robustness to label noise as well as to study retrieval-specific types of annotation errors. We choose ArcFace as one of the best-performing models for our experiments. Future work can extend presented results to a variety of training objectives [22, 36] and model architectures [7, 32].

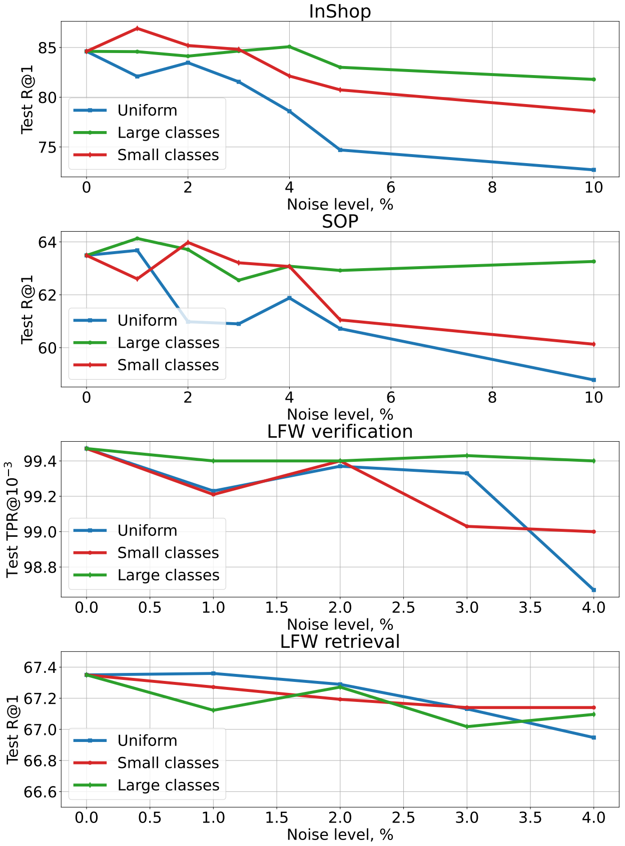

Figure 3. Label noise patterns' effect on test performance.

7. Conclusion

In this work, we studied image retrieval robustness to label noise. According to our experiments, image retrieval is less robust to label noise than image classification approaches. Therefore, more attention must be paid to training data quality for image retrieval tasks. We further showed that the performance is rather affected by the number of corrupted classes than by the proportion of corrupted samples in each class.

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021. 1, 2

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In

European conference on computer vision, pages 584–599. Springer, 2014. 1

[3] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, 33:4308–4319, 2020. 1

[4] Charles Corbiere, Hedi Ben-Younes, Alexandre Ramé, and Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2268–2274, 2017. 1, 2

[5] Bharath Bhushan Damodaran, Rémi Flamary, Vivien Seguy, and Nicolas Courty. An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images. *Computer Vision and Image Understanding*, 191:102863, 2020. 1

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 1, 2

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[11] Gary B Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, 14(003), 2014. 1, 2, 3

[12] Sarah Ibrahimi, Arnaud Sors, Rafael Sampaio de Rezende, and Stéphane Clinchant. Learning with label noise for image retrieval by selecting interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2181–2190, 2022. 1

[13] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016. 1

[14] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016. 3

[15] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016. 1

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[17] Mengmeng Kuang, Weiyan Wang, Zhenhong Chen, Lie Kang, and Qiang Yan. Efficient two-stage label noise reduction for retrieval-based tasks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 526–534, 2022. 1

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[19] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Noise-resistant deep metric learning with ranking-based instance selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6811–6820, 2021. 1, 2

[20] Hengwei Liu, Jinyu Ma, and Xiaodong Gu. Towards image retrieval with noisy labels via non-deterministic features. In *International Conference on Artificial Neural Networks*, pages 446–456. Springer, 2021. 1

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1, 2

[22] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 4

[23] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. pages 681–699, 2020. 2, 3

[24] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7044–7053, 2017. 1

[25] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 1, 2

[26] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 1

[27] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under

the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020. 1

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2

[29] Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I Zeki Yalniz. Noiserank: Unsupervised label noise reduction with dependence models. In *European Conference on Computer Vision*, pages 737–753. Springer, 2020. 1

[30] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[31] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[32] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 4

[33] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018. 1, 2

[34] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 3

[35] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 2, 3

[36] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009. 4

[37] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025, 2019. 1