Out-of-Distribution Detection for Reliable Face Recognition

Chang Yu, Xiangyu Zhu, *Member, IEEE*, Zhen Lei, *Senior Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—In real applications, face recognition systems are always faced with non-face inputs and low-quality faces due to the complicated conditions like mis-detections by face detectors. However, in deep learning based methods, these outliers are always ignored during training phase and the models tend to make unreasonable decisions on these images. For example, matching a texture-rich patch to an old-man face overconfidently. We formulate this challenge on the task of out-of-distribution detection (OOD), where a network must determine whether or not an input is outside of the set on which the network can safely perform. In this paper, we propose to detect out-of-distribution samples based on uncertainty prediction and the L2-norm of features, so as to effectively filter out non-face and low-quality faces. We demonstrate that the proposed method can reliably detect out-of-distribution samples and improve the performance of face recognition, without the need of labelled OOD data.**

***Index Terms*—Face recognition, low-quality, non-face, out-of-distribution detection.**

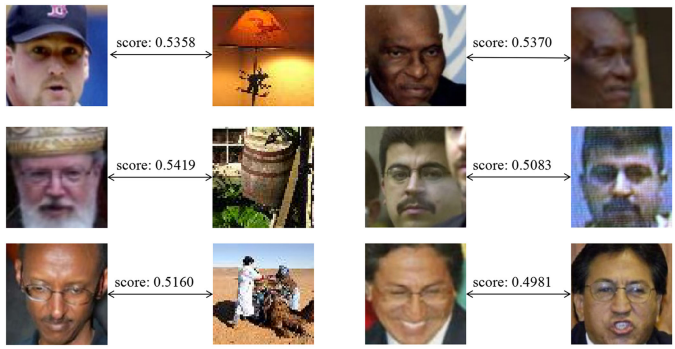


Fig. 1. The scores (the cosine-similarity of the two features) between faces and non-faces. It can be seen that the scores between faces and some background patches may be higher than that between intra-identity faces.

I. Introduction

In RECENT years, face recognition has witnessed great improvements due to the application of deep learning [1]–[9]. However, the data-driven strategy also brings a challenge: The images sent into the face recognition systems are not always appropriate for recognition. There may be low-resolution faces, motion-blurred faces, occluded faces and even background patches due to the unconstrained scenarios and the failure of the face detection. Unfortunately, since the face recognition engine has never seen such low-quality faces before, it probably makes unreasonable decisions on these outliers. As shown in Fig. 1, the similarity scores between faces and some background patches may be higher than that between intra-identity faces. One common solution is to train an extra model to judge the quality of images and filter out the bad ones. However, the image quality for face recognition models is somewhat subjective and not easy to label, which makes the quality prediction unreliable.

In this paper, we propose to detect low-quality inputs under the framework of out-of-distribution (OOD) detection [10]. It requires networks to be aware of uncertainty when out-of-distribution examples such as unrecognizable and irrelevant ones appear. Based on OOD detection, the network can not only extract the feature vectors, but can also output the certainty of features.

There are OOD detection works by integrating generative models. Lee *et al.* [11] use GAN with specially-designed loss to generate OOD samples for training. CCU [12] chooses Gaussian mixture model (GMM) for density estimation and uses TinyImageNet [13] as a proxy of all possible examples to estimate the density of OOD examples during training. Besides generative methods, Shalev *et al.* [14] combine the outputs of five similar networks to detect OOD samples. Yu *et al.* [15] utilize additional OOD data during the training process for unsupervised learning. ODIN [16] adds disturbances during the test process to distinguish OOD data.

Most of these methods concentrate on the close-set classification problem, where the training set and testing set share the same categories and the predicted probabilities of each category is a strong indicator for OOD detection. Differently, face recognition is usually an open-set problem. The nearest neighborhood classifier with cosine-similarity is usually used to perform classification without the probability of the predefined classes. Therefore, applying OOD in face recognition is not a trivial problem.

Some work has been done for the confidence of the networks. DeVries *et al.* [17] propose to train a confidence branch unsupervisely by interpolating between the original softmax prediction and the label, where the degree of interpolation is indicated by the confidence. Liu *et al.* [18] propose the decoupled convolution operators to weaken the inputs with small norms, which are regarded as the low-confident ones.

Manuscript received January 23, 2020; revised April 1, 2020; accepted April 12, 2020. Date of publication April 20, 2020; date of current version May 21, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61806196 and Grant 61876178. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sumohana S. Channappayya. *(Chang Yu and Xiangyu Zhu contributed equally to this work.) (Corresponding author: Zhen Lei.)*

The authors are with the CBSR & NLPR, Institute of Automation, Chinese Academy of Science and the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: yuchang2019@ia.ac.cn; xiangyu.zhu@nlpr.ia.ac.cn; zlei@nlpr.ia.ac.cn; szli@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/LSP.2020.2988140

In this paper, we propose to perform OOD detection in face recognition based on uncertainty and the L2-norm of features. The uncertainty of each sample is unsupervisely predicted by attenuating the influence of outliers during training. Moreover, we find the L2-norm of feature vectors are also useful for the OOD detection. With the uncertainty and the feature norm, the network can learn out-of-distribution samples without the complex generative models or extra OOD labels. This method can effectively filter out non-face images and improve the baseline of general OOD detection, so as to improve the robustness of face recognition system.

The main contributions of this work includes:

1) This paper proposes an unsupervised method to predict the uncertainty of each sample by attenuating the influence of low-quality samples during training, so as to filter out the non-face images.

2) We find that when incorporating uncertainty, the L2-norm of the feature vector becomes useful to detect OOD samples, which can be combined with the uncertainty to further improve the OOD detection.

3) The proposed OOD detection method can not only effectively filter out non-face images, but also improve the performance of face recognition.

II. Method

This part introduces the uncertainty learning in face recognition training and the OOD detection using the L2-norm of features in face recognition.

A. Uncertainty Prediction

Taking softmax loss as an example. Given an input feature vector x_i, and its corresponding label y_i, the softmax loss can be formulated as:

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_{y_j}}}, \tag{1}$$

where N denotes the number of training samples, f_{y_i} refers to the i-th output of the classification layer, and C denotes the number of classes.

In the process of network learning, the softmax loss tries to classify all the samples correctly, including the out-of-distribution (OOD) samples. This training strategy forces the model to give every sample a clear prediction even when it is impossible. As a result, when encountering a real OOD sample during testing, the model still outputs a high probability, which confuses the classification process.

To reduce the impact of OOD samples during training and identify them during testing, we introduce uncertainty [19] to the softmax loss and use the maximum of softmax as the measure of certainty following Hendrycks *et al.* [10]. The softmax loss with uncertainty can be formulated as:

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -e^{\lambda s}\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{c} e^{f_{y_j}}} \tag{2}$$

where $s = \max_{k\in C} \frac{e^{f_{y_k}}}{\sum_{j=1}^{c} e^{f_{y_j}}}$, and λ is a hyper-parameter that can be adjusted according to the degree of attenuation. The samples with low uncertainty s will adaptively decrease its loss L_s and attenuate its influence in the training.

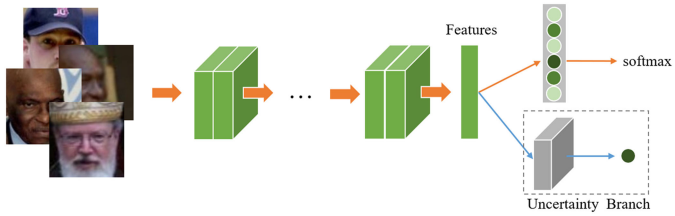


Fig. 2. The architecture for reliable face recognition with the uncertainty branch. It is used to regress the maximum value of softmax so as to predict the uncertainty s for OOD detection during testing.

Besides, the deep networks use the loss function for feature learning, so that the samples of the same identities are as close as possible in the mapping space. The low-quality or OOD inputs might affect the aggregation within the classes. By weakening the influence of low-confidence samples during training, the in-distribution samples within the classes will be closer, achieving a higher classification accuracy in classification and the uncertainty s can identify the OOD samples during testing.

Note that the uncertainty s depends on the classification layer $[f_{y_1}, f_{y_2}, ..., f_{y_C}]$. Different from general classification tasks, face recognition always contains hundreds of thousands of or even millions of classes. Thus, uncertainty prediction operation is computational cost. To improve efficiency, we add an uncertainty branch to predict the uncertainty s, as illustrated in Fig. 2. The final loss is formulated as:

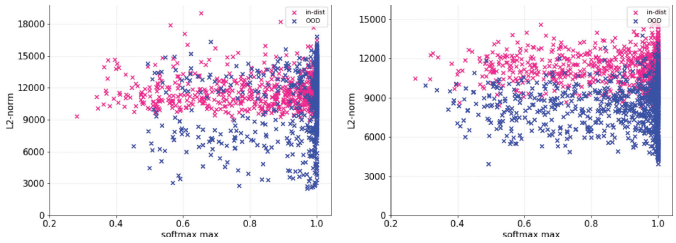
$$\begin{aligned} L_s &= \frac{1}{N}\sum_{i=1}^{N} -e^{\lambda s}\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{c} e^{f_{y_j}}} \\ L_u &= \frac{1}{N}\sum_{i=1}^{N} \mid f_c - s \mid^2, \\ L &= L_s + rL_u \end{aligned} \tag{3}$$

where f_c refers to the output of the uncertainty branch, L_s denotes the softmax loss with uncertainty, L_u denotes the loss of uncertainty branch, and r is a hyper-parameter to balance L_s and L_c during training.

With this loss, we not only improve the efficiency, but also reduce the model parameters by discarding the classification layer $[f_{y_1}, f_{y_2}, ..., f_{y_C}]$ during testing.

B. OOD in Face Recognition

For out-of-distribution detection in face recognition, face samples are regarded as in-distribution, and non-face samples are regarded as OOD. Ranjan *et al.* [20] observe that a high-quality face tends to have a higher L2-norm than a blur one. As there is no baseline for OOD face detection, we intuitively choose the L2-norm of features $\|f(x)\|_2$ as the OOD detector. However, there still exist out-of-distribution faces with high L2-norm. As illustrated in Fig. 2, by introducing uncertainty to the training process, it is found that the correlation between the L2-norm and OOD becomes more significant and it is more reliable to use L2-norm as a clue for OOD detection. Moreover, by combining uncertainty and L2-norm, the performance can be further improved. Thus, we can obtain the OOD detector for



(a) w/o uncertanty learning (b) with uncertainty learning

Fig. 3. The correlation between the uncertainty s and the L2-norm of features. The blue cross represents the OOD samples, and the magenta cross represents the in-distribution samples. Compared with (a) the distribution of L2-norm between OOD samples and in-distribution samples in (b) is more separable. After introducing uncertainty to the training process, the correlation between L2-norm and OOD become more significant.

face recognition as

$$g(x) = \begin{cases} 1, & \text{if } s\|f(x)\|_2 \geq \delta \\ 0, & \text{if } s\|f(x)\|_2 < \delta \end{cases} \tag{4}$$

where s denotes the uncertainty predicted by the uncertainty branch and $\|f(x)\|_2$ refers to the L2-norm of features. The hyper-parameter δ is chosen on the basis of the recognition performance in our experiment.

III. Experiment

In this part, we examine the effectiveness of the proposed OOD detection method in face recognition problem.

A. Datasets

MS-Celeb-1M MS-Celeb-1M [21], is a wild dataset for Large-Scale Face Recognition. It remains about 5 million images after cleaning [3].

LFW Labeled Faces in the Wild (LFW) [22] is a widely used database for unconstrained face verification. It contains 13,233 images and each identity has one or several samples.

CPLFW Cross-Pose LFW (CPLFW) [23] is a renovated dataset of LFW. CPLFW emphasizes pose difference between images from the same identity which makes verification more challenge. It has 12,000 samples, and each identity has one or two samples.

TinyImageNet The Tiny ImageNet (TinyImageNet) [13] is a subset of ImageNet [24] . It contains 10,000 test images from 200 different classes.

LSUN The Large-scale Scene Understanding dataset (LSUN) [25] consists of 10,000 test images from 10 different scenes.

SVHN The Street View House Numbers (SVHN) contains 26,032 test images from 10 digit classes, which obtained from house numbers in Google Street View images.

In face recognition, we choose the MS-Celeb-1M as the training set. LFW and CALFW as regarded as in-distribution testing datasets, and TinyImageNet, LSUN and SVHN as OOD datasets. The samples in the OOD datasets are resized to 120×120 to keep the same size as the face images.

B. CNN Architecture

All the CNN models in the experiments follow the same ResNet64 architecture similar to Zhu *et al.* [2]. The additional uncertainty branch for OOD detection in face recognition contains two fc layers. All the networks are trained on 2 GeForce 1080 GPUs with batch size 128 and momentum 0.9. The learning rate begins with 0.01 and is divided by 10 when the loss does not decrease until the learning rate is 0.00001. We first train the network with the loss L_s until the network converges for initialization and then use the loss L to train it jointly. The hyper-parameter r and λ are both set to 1 during training. All methods are implemented on PyTorch [26].

C. Evaluation Metrics

This paper follows the same metrics used by Hendrycks *et al.* [10] to measure the effectiveness of the OOD detector.

FPR@95% TPR This shows the probability that an OOD sample is misclassified as in-distribution when the true positive rate is 95%. Networks with better performance should have lower FPR at 95% TPR.

AUROC AUROC is the Area under the Receiver Operating Characteristic curve. A perfect OOD detector should correspond to an AUROC score of 100%.

AUPR AUPR is the Area under the Precision-Recall curve. A perfect network has an AUPR of 100%. During testing, in-distribution images are regarded as positives.

D. OOD Detection Results

Table I shows the results of OOD detection. The Baseline is trained by only softmax and the Baseline with uncertainty is trained by Eqn.(2). Both of them directly use L2-norm to filter out the OOD samples. Our method combines the uncertainty prediction and the L2-norm and employs the strategy in Eqn.(4) to filter out the OOD samples. We can see that after introducing the uncertainty learning during training, the FPR@TPR=95% can be reduced by 60% to 40%, since the correlation between L2-norm and OOD become more significant. Moreover, our method further improves the FPR @TPR=95% by 5.1% to 1%, and other evaluation metrics are also increased by about 1%, which makes the OOD detection in Face Recognition work well. It shows that the performance of OOD detection has been greatly improved by introducing the uncertainty prediction and the L2-norm of features.

In real face recognition systems, we want to recognize a test set with OOD samples and get a higher recognition accuracy. However, current methods separate the evaluations of OOD detection and recognition. In this paper, we mix CPLFW and other three OOD datasets separately and directly use TPR@FPR $= 1\%$ and TPR@FPR $= 0.1\%$ to evaluate the accuracy of face recognition on the mixed datasets. For each sample filtered out by the OOD detector, we set its feature to zero. In Table II, the FR refers to a common face model trained by softmax and the Baseline directly uses L2-norm to filter out OOD samples. Taking the results on CPLFW+TinyImageNet as an example, the common face model degrades seriously when dealing with OOD data and only achieves 5.20%@FPR $= 0.1\%$. After introducing L2-norm to filter out OOD samples, we get a slightly better accuracy 8.78%@FPR $= 0.1\%$. Finally, our method gets the best performance and achieves 50.45%@FPR $= 0.1\%$, which demonstrates the effectiveness of our method. Besides, we can

TABLE I
OOD DETECTION FOR FACE RECOGNITION ON RESNET64

	OOD dataset	FPR@TPR=95% ↓	AUROC ↑	AUPR ↑
		Baseline / Baseline with uncertainty / Ours		
ResNet64 LFW	TinyImageNet	75.99 / 26.15 / 12.80	44.86 / 91.97 / 97.14	52.20 / 91.86 / 97.48
	LSUN	50.61 / 16.33 / 9.54	70.34 / 96.05 / 98.02	69.32 / 96.39 / 98.33
	SVHN	54.54 / 27.44 / 15.85	55.73 / 92.46 / 95.87	32.96 / 83.27 / 90.77
ResNet64 CPLFW	TinyImageNet	70.36 / 15.90 / 11.08	74.79 / 96.53 / 97.85	81.73 / 96.89 / 98.40
	LSUN	42.33 / 8.11 / 8.22	88.04 / 98.42 / 98.39	90.26 / 98.65 / 98.81
	SVHN	51.03 / 15.42 / 14.15	75.68 / 96.78 / 97.24	62.75 / 93.46 / 95.29

TABLE II
FACE VERIFICATION ON THE MIXED DATASET

Mixed Dataset	Method	TPR @FPR=1%	TPR @FPR=0.1%
CPLFW+TinyImageNet	FR	41.43	5.20
	Baseline	55.05	8.78
	Ours	**72.03**	**50.45**
CPLFW+LSUN	FR	81.30	54.77
	Baseline	**82.12**	54.95
	Ours	81.63	**64.53**
CPLFW+SVHN	FR	77.98	32.28
	Baseline	78.50	32.15
	Ours	**83.48**	**65.98**

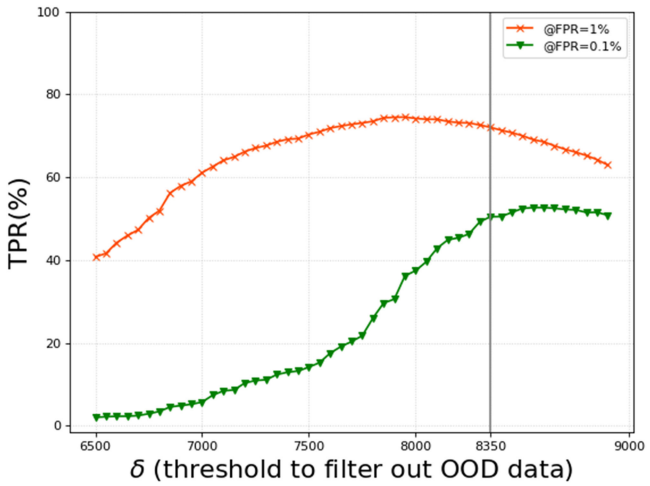


Fig. 4. The face recognition performance of our method on the mixed dataset (CPLFW+TinyImageNet) with different threshold δ. With the increase of δ, both of the TPR@FPR $= 1\%$ and the TPR@FPR $= 0.1\%$ on the mixed dataset gradually rise to the peak and then decline.

see that the improvement of TPR@FPR $= 0.1\%$ is more significant than TPR@FPR $= 1\%$ since the OOD samples mainly affect the face pairs with relatively lower cosine-similarity.

The effect of the hyper-parameter δ in Eqn.(4) is illustrated in Fig. 3. With the increase of δ, the face recognition performance on the mixed dataset gradually rises to the peak and then declines. The decline of the curves is because the in-distribution face samples will also be filtered out if δ is too large. Besides, TPR@FPR $= 1\%$ and TPR@FPR $= 0.1\%$ reach their peaks at different thresholds δ. For this mixed dataset, we set the δ to 8350.

Another important hyper-parameter is λ, which is used to weaken the uncertainty samples. In the experiments above, λ is set to 1. We conduct an experiment to explore the impact of λ. By setting $\lambda = 0.5, 1, 2, 5$, we train our model on MS-Celeb-1M

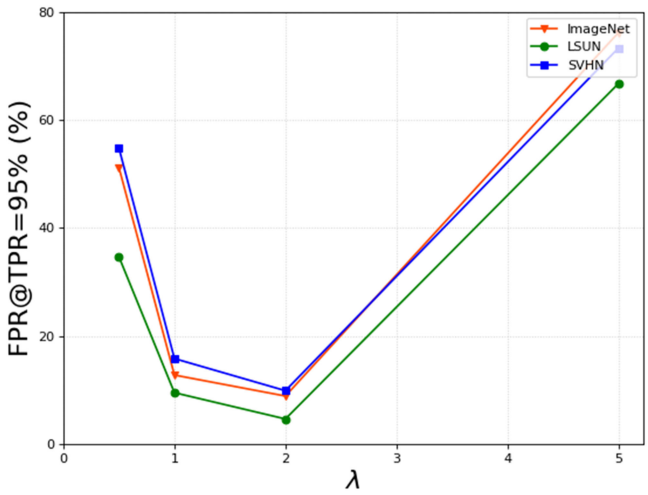


Fig. 5. The FPR@TPR $= 95\%$ of OOD detection on LFW and the corresponding OOD dataset with different λ. With the increase of λ, the loss attenuation for uncertainty samples is enhanced.

with the CNN architecture of ResNet64, and analyze the change of FPR@TPR $= 95\%$ of OOD detection. As shown in Fig. 3, with the increase of λ, the performance of the OOD detector is gradually improved and we get a best result when $\lambda = 2$. But if λ is too large, the network ignores the effective information and gets degraded performance.

IV. CONCLUSION

In this paper, we propose to detect out-of-distribution faces based on the uncertainty prediction and the L2-norm of features, so as to effectively filter out non-face and low-quality faces. We demonstrate that the proposed method can reliably detect out-of-distribution samples and improve the performance of face recognition. We hope that our method will bring some inspiration to the problem of OOD detection.

REFERENCES

[1] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2017, pp. 212–220.
[2] X. Zhu *et al.*, "Large-scale bisample learning on ID versus spot face recognition," *Int. J. Comput. Vision*, vol. 127, nos. 6/7, pp. 684–700, 2019.
[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4690–4699.
[4] H. Liu, X. Zhu, Z. Lei, and S. Z. Li, "AdaptiveFace: Adaptive margin and sampling for face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 11939–11948.

[5] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.
[6] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," 2016, *arXiv:1612.02295*.
[7] W. Liu *et al.*, "Learning towards minimum hyperspherical energy," 2018, *arXiv:1805.09298*.
[8] J. Guo, X. Zhu, Z. Lei, and S. Z. Li, "Face synthesis for eyeglass-robust face recognition," in *Proc. Chin. Conf. Biometric Recognit.*, 2018, pp. 275–284.
[9] J. Guo, X. Zhu, C. Zhao, D. Cao, Z. Lei, and S. Z. Li, "Learning meta face recognition in unseen domains," 2020, *arXiv:2003.07733*, 2020.
[10] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2016, *arXiv:1610.02136*.
[11] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," 2017, *arXiv:1711.09325*.
[12] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," 2019, *arXiv:1909.12180*.
[13] TinyImagenet. Accessed: Jan. 5, 2020. [Online]. Available: https://tiny-imagenet.herokuapp.com
[14] G. Shalev, Y. Adi, and J. Keshet, "Out-of-distribution detection using multiple semantic label representations," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Red Hook, NY, USA: Curran, 2018, pp. 7375–7385.
[15] Q. Yu and K. Aizawa, "Unsupervised out-of-distribution detection by maximum classifier discrepancy," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 9517–9525.
[16] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2017, *arXiv:1706.02690*.
[17] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," 2018, *arXiv:1802.04865*.
[18] W. Liu *et al.*, "Decoupled networks," 2018, *arXiv:1804.08071*.
[19] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Red Hook, NY, USA: Curran, 2017, pp. 5574–5584.
[20] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," 2017, *arXiv:1703.09507*.
[21] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 87–102.
[22] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[23] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying crosspose face recognition in unconstrained environments," Beijing Univ. Posts Telecommun., Beijing, China, Tech. Rep. 18-01, Feb. 2018.
[24] J. Deng, W. Dong, R. Socher, L. J. Li, and F. F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.
[25] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.
[26] Pytorch. Accessed: Jan. 2, 2020. [Online]. Available: https://pytorch.org/.