FOOVF: A NOVEL MODEL OWNERSHIP VERIFICATION FRAMEWORK BASED ON FEATURE FINGERPRINTING AND OUTLIER DETECTION

Zhe Sun[⋆†] , *Yu Zhang*[⋆†‡] , *Zhongyu Huang* [⋆†], *Zongwen Yang*[⋆†] , *Wangqi Zhao*[⋆†] , *Jianzhong Zhang*[⋆†]

[⋆]Tianjin Key Laboratory of Network and Data Security Technology
[†]College of Cyber Science, College of Computer Science, Nankai University, Tianjin, China

ABSTRACT

Model intellectual property is gradually becoming a hot topic. However, although effective in protecting model ownership, existing works mainly focus on classification models. In addition, most proposed methods impair model utility. To address these shortcomings, this paper proposes a novel model ownership verification framework based on feature fingerprint, called FOOVF, which splits ownership verification and model tasks. Our key idea is model knowledge can be uniquely characterized by the extracted features. The extracted features for each input sample may be high-dimensional, complicated, and difficult to compare, but they are inherent information that cannot be discarded for piracy models. To measure the inheritance of our fingerprint, we are the first work to introduce outlier detection into model ownership verification. The experimental results show that our framework has significant effects and has more universality.

Index Terms— Model ownership verification, fingerprinting, outlier detection

1. INTRODUCTION

In recent years, DNN model architecture has become more complicated, and the training cost of models with great effects has gradually increased. As a result, the pair of offensive and defensive issues, model stealing and ownership protection, has received more attention from researchers.

A large number of model watermarking works [1, 2] have been widely proposed and used to verify model ownership. Uchida et al. [3] propose to embed the watermark in the parameters, and the operation of removing the watermark leads to the degradation of model performance. The robustness of watermarking [4] is enhanced by using neuron alignment. However, the proposed watermarking inevitably brings a degradation of model performance.

Fingerprinting [5, 6, 7] is used to verify model ownership using the inherent information invariance of the model. Cao et al. [8] propose that decision boundaries can be used as evidence of stealing, and they describe decision boundaries using adversarial examples. Fingerprinting works without affecting the model utility, but the decision boundaries are concentrated on classification models. Pan et al. [9] propose the first model task-agnostic ownership verification scheme to design fingerprinting algorithms using the piecewise linear property of the ReLU function, but the ReLU function reduces generality in another aspect. Therefore, no impact on model performance (zero contamination) and independence from the model task are two challenges we need to address.

Model fingerprinting, which differs from watermarking, is mining the model's inherent internal information as a basis for verifying model ownership. To address the first challenge, we propose to adopt the fingerprinting scheme as the basis of our framework. To address the second challenge, we reject the previous idea of focusing on describing classification boundaries (decision boundaries) and instead turn to feature extraction.

Therefore, we propose that the output of the model feature extraction layer can be used as the fingerprint, which is called the feature fingerprint. The data used to generate the feature fingerprint are referred to as feature fingerprint data. Previous work [10] consider measuring the differences in the output of the entire layer in the DNN models, but the curse of dimensionality from comparing a large number of high-dimensional features is difficult to overcome. Therefore, we need a method to quickly compare extracted features from models with significant results. This paper is the first attempt to introduce an outlier detection algorithm to compare model fingerprints. As shown in the Fig.1, we use the outlier detection algorithm [11] based on the empirical cumulative distribution to compute the deviation degree of feature fingerprint. Finally, we propose a novel model ownership framework based on the feature fingerprint called FOOVF. FOOVF can be used to verify whether the knowledge of the suspicious model comes from the victim model or not. In summary, our contributions are as follows:

- We devise the feature fingerprint to represent model knowledge abstractly. In contrast to describing classification boundaries previously, our feature fingerprint applies to most models rather than only classification models.
- We are the first work to compute the deviation de-

‡ Corresponding author: zhangyu1981@nankai.edu.cn.

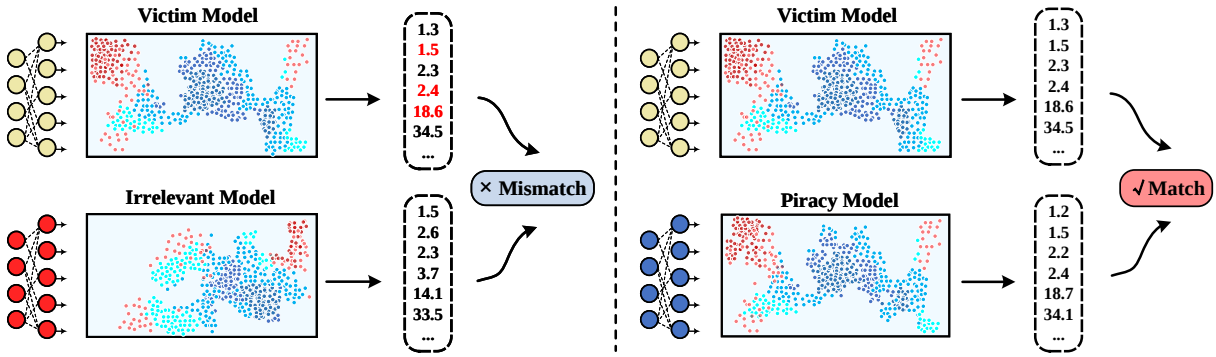


Fig. 1. Feature fingerprint matching and model ownership verification..

gree of anomalous data to measure feature fingerprinting. We propose a novel model ownership verification framework characterized by zero contamination, easy deployment, and high efficiency.

- Extensive experimental results show that our framework achieves highly competitive results in mainstream model stealing attacks. Compared to previous fingerprinting works, our framework has more universality.

2. THREAT MODEL

In the model ownership verification, the owner of victim model M_v provides the feature fingerprint data, and we need to compute the deviation degree of the suspicious model M_s feature fingerprint in a white-box environment. Based on the comparison results, we can confirm whether the knowledge of the suspicious model M_s comes from the victim model M_v, i.e., verifying the model ownership. In our scenario, the verification in a white-box setting is considered. Although most existing researches focus on black-box environments, they are always limited to operating only on training data and inferred results. However, we argue that in the future, there will be laws or third-party organizations which guarantee the proposed model owner needs to submit a piece of information as a unique identifier, and then register it for "patenting". Therefore, our proposed feature fingerprint can be used as a model identifier. In this scenario, a white-box verification approach is perfectly feasible.

3. THE PROPOSED FRAMEWORK

3.1. Fingerprint Generation

Definition 1 (The Feature Fingerprint). Given a dataset D_f of size N and a trained model M. Let $x_i \in D_f$ input to M. Suppose the l-th layer is the feature extraction layer of M, then its output $M^l(x_i)$ can be regarded as a high-dimensional data point, whose dimensionality is determined by the network architecture. The output $\{M^l(x_1), M^l(x_2), \cdots, M^l(x_N)\}$ at the feature extraction layer is called the feature fingerprint of model M under the feature fingerprint data D_f.

Two different model owners, say u and w, even though they use the same training data and model architecture, have very similar but considered independent models M_u and M_w, which we call irrelevant models. In our framework, their feature fingerprints $\{M_u^l(x_1), M_u^l(x_2), \cdots, M_u^l(x_N)\}$ and $\{M_w^l(x_1), M_w^l(x_2), \cdots, M_w^l(x_N)\}$ should be mismatched. Just the opposite, if the suspicious model M_s comes from the victim model M_v, the feature fingerprints of their outputs should match.

3.2. The Framework Working Mechanism

We expect the outlier detection algorithm for matching fingerprints can satisfy: (1) low computational cost, (2) unsupervised, (3) without tuning parameters, (4) support high dimensional data, and (5) high efficiency and accuracy. Therefore, we introduce empirical-cumulative-distribution-based outlier detection (ECOD) [11]. ECOD is a simple and effective algorithm that first estimates the underlying distribution of the input data in a nonparametric manner by computing the empirical cumulative distribution of the data in each dimension. And then it uses these empirical distributions to estimate the tail probabilities of each dimension for each data point. Finally, it computes the outlier scores for each data point by aggregating the estimated tail probabilities on all dimensions.

Suppose the feature fingerprint from the victim model M_v is treated as N data points $M_v^l(x_1), M_v^l(x_2), \cdots, M_v^l(x_N) \in \mathcal{R}^d$. Given the feature fingerprint data D_f, we assign an outlier score $\mathcal{O}(M_v^l(x_i)) \in \mathcal{R}$ to each data point $M_v^l(x_i)$, where a higher score indicates a higher deviation.

Compute Outlier Scores. Specifically, let $\mathcal{F} : \mathcal{R}^d \to [0, 1]$ be denoted as the joint cumulative distribution function (CDF) of all dimensions/features. Thus $M_v^l(x_1), M_v^l(x_2), \cdots, M_v^l(x_N)$ can be seen as sampled i.i.d. from a distribution with joint CDF $\mathcal{F}$. We write the j-th entry of $M_v^l(x_i)$ as $M_v^l(x_i^{(j)})$.

First, the left and right tail probability for each dimension of a single data point are computed as shown in Eq.1 and 2:

$$\hat{\mathcal{F}}^{(j)}_{left}(z) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{M^l_v(x^{(j)}_i) \leq z\}, z \in \mathcal{R} \quad (1)$$

$$\hat{\mathcal{F}}^{(j)}_{right}(z) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{M^l_v(x^{(j)}_i) \geq z\}, z \in \mathcal{R} \quad (2)$$

where $1\{\cdot\}$ is the indicator function that is 1 when its argument is true and is 0 otherwise. Note that univariate CDF's can be accurately estimated simply by using the empirical CDF.

Therefore, under the independence assumption, we can estimate the left and right-tail ECDFs for all d dimensions by estimating

$$\begin{aligned} \hat{\mathcal{F}}_{left}(M^l_v(x_i)) &= \prod_{j=1}^{d} \hat{\mathcal{F}}^{(j)}_{left}(M^l_v(x^{(j)}_i)) \\ \hat{\mathcal{F}}_{right}(M^l_v(x_i)) &= \prod_{j=1}^{d} \hat{\mathcal{F}}^{(j)}_{right}(M^l_v(x^{(j)}_i)) \end{aligned} \quad (3)$$

Then compute the sample skewness coefficient for the j-th dimension's distribution:

$$\gamma_j = \frac{\frac{1}{N}\sum_{i=1}^{N}(M^l_v(x^{(j)}_i) - \overline{M^l_v(x^{(j)})})^3}{[\frac{1}{N-1}\sum_{i=1}^{N}(M^l_v(x^{(j)}_i) - \overline{M^l_v(x^{(j)})})^2]^{3/2}} \quad (4)$$

where $M^l_v(x^{(j)}) = \frac{1}{N}\sum_{i=1}^{N} M^l_v(x^{(j)}_i)$ is the sample mean of the j-th dimension.

For each data point we aggregate its tail probability to obtain the outlier score:

$$\mathcal{O}_{left}(M^l_v(x_i)) = -\sum_{j=1}^{d} log(\hat{\mathcal{F}}^{(j)}_{left}(M^l_v(x^{(j)}_i))) \quad (5)$$

$$\mathcal{O}_{right}(M^l_v(x_i)) = -\sum_{j=1}^{d} log(\hat{\mathcal{F}}^{(j)}_{right}(M^l_v(x^{(j)}_i))) \quad (6)$$

In addition, it is possible to decide whether to use the left or right tail probability depending on the value of γ_j:

$$\begin{aligned} \mathcal{O}_{auto}(M^l_v(x_i)) = -\sum_{j=1}^{d} [&\mathbb{1}\{\gamma_j < 0\} log(\hat{\mathcal{F}}^{(j)}_{left}(M^l_v(x^{(j)}_i))) \\ &+ \mathbb{1}\{\gamma_j \geq 0\} log(\hat{\mathcal{F}}^{(j)}_{right}(M^l_v(x^{(j)}_i)))] \end{aligned} \quad (7)$$

We can obtain the final outlier scores of $M^l_v(x_i)$:

$$\begin{aligned} \mathcal{O}(M^l_v(x_i)) = \max\{&\mathcal{O}_{left}(M^l_v(x_i)), \mathcal{O}_{right}(M^l_v(x_i)), \\ &\mathcal{O}_{auto}(M^l_v(x_i))\} \end{aligned} \quad (8)$$

Finally, we obtain the outlier scores of all feature fingerprints $\{\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_N\}$.

Anomalous Data Generation. We match the fingerprint by computing the deviation degree. Therefore, for better fingerprint matching, some anomalous data needs to be generated as feature fingerprint data input to the model. Intuitively, we also need normal data as a baseline to enable the results to be compared within the same scale. The normal data can be chosen from the origin data of victim model, and practically any data can be used as a baseline for comparison. We argue that anomalous data can be constructed in two ways. The first is to embed exogenous features in normal data, an idea derived from [12]. The second is data completely independent of the model task and the original dataset. As shown in the Fig.2, two exogenous features, image steganography [13] and style transformation [14], are chosen in this paper to compute the deviation degree as anomalous data together with the irrelevant data [15]. Thus, after computing the outlier scores for the feature fingerprints, we get $\{\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_m, \mathcal{O}_{m+1}, \cdots, \mathcal{O}_N\}$, where m is the number of anomalous data. Finally, we test the effect of different anomalous data on FOOVF in our experiments.

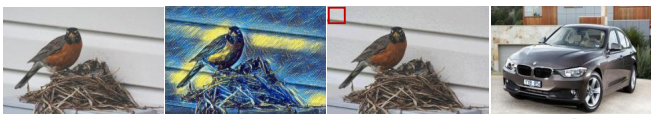


Fig. 2. (a) Training data, (b) Style transformation, (c) Image steganography, and (d) Irrelevant data.

3.3. Model Ownership Verification

We design an interesting feature fingerprint data. As discussed in Section 2, we assume the existence of a third-party organization that keeps the unique identifying information of the model as an identifier. Therefore, we assume that the model owner submits the feature fingerprint data D_f and the outlier score vector $\mathcal{O}(M^l_v)$. When the owner suspects that a model M_s is stolen from M_v, a request is made to the organization to compute $\mathcal{O}(M^l_s)$. To compare both, we need to do some transformations. As shown in Eq.9, we take the normal data in D_f as the baseline and compute the comparison-outlier score vector. We can generally set a threshold value to determine the suspicious model.

$$\mathcal{CO}(M^l_v(x_i)) = |\mathcal{O}(M^l_v(x_i)) - \frac{\sum_{j=m+1}^{N}\mathcal{O}(M^l_v(x_j))}{N-m}| \quad (9)$$

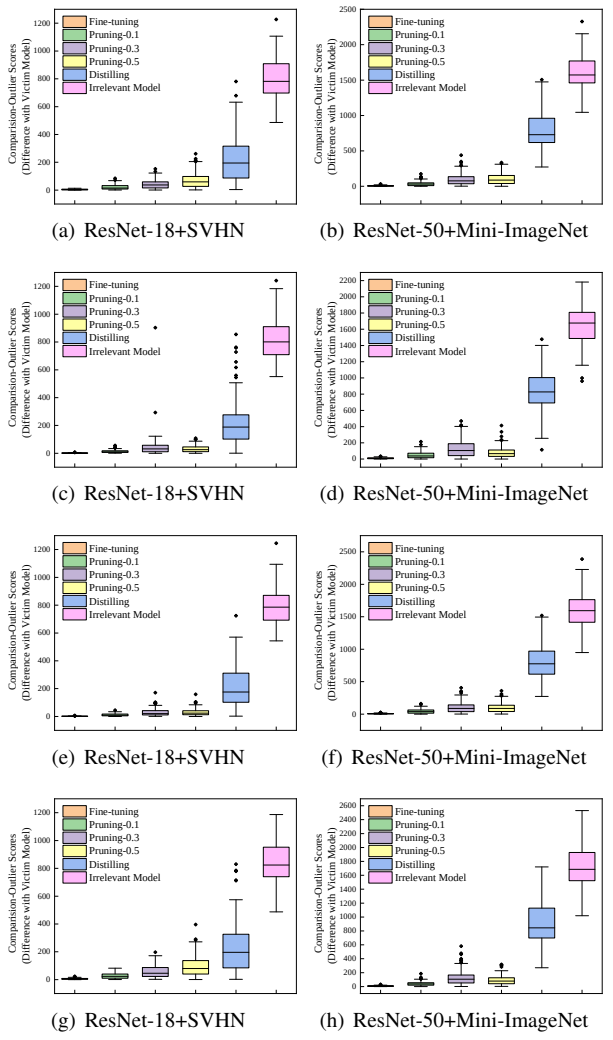
4. EXPERIMENTS

4.1. Baselines

In this paper, we evaluate the performance of FOOVF on SVHN [16] and Mini-ImageNet [17]. We employ ResNet-18 and ResNet-50 as victim model architectures. To evaluate the performance of our proposed framework against model stealing, we implement popular model stealing techniques such as model fine-tuning, pruning (various rates), and distillation.

We implement all experiments using Pytorch on NVIDIA GeForce RTX 3060, and implement an irrelevant model based on the GoogLeNet [18] architecture (from the same dataset) for comparison. We provide specific model parameter information and available code[1].

[1]https://github.com/nkunetlab/nkunetlab-FOOVF



(a) ResNet-18+SVHN (b) ResNet-50+Mini-ImageNet

(c) ResNet-18+SVHN (d) ResNet-50+Mini-ImageNet

(e) ResNet-18+SVHN (f) ResNet-50+Mini-ImageNet

(g) ResNet-18+SVHN (h) ResNet-50+Mini-ImageNet

Fig. 3. Difference of comparison-outlier scores between the piracy (irrelevant) and victim model feature fingerprints.

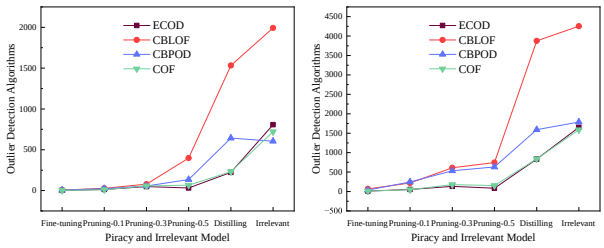
4.2. Evaluating Effectiveness

In this subsection, we discuss the effectiveness of FOOVF. We design four types of anomalous data, including (1) Style transformation(Fig.3-a,b), (2) Image steganography(Fig.3-c,d), (3) Irrelevant data(Fig.3-e,f), and (4) An average mixture of the above three(Fig.3-g,h). In our experiments, we randomly select a set of size 30 from the original dataset of victim model as normal data, with anomalous data to form a feature fingerprint data of size 100. We evaluate the effectiveness of different combinations. As shown in the Fig.3, our framework is significantly effective for different techniques, and the deviation results of the piracy and the victim model perform almost the same. Even though the results of our framework for the model distillation technique are lower than other stealing techniques, FOOVF can still strongly prove that the model is a piracy model compared to the irrelevant model. In conclusion, FOOVF can be used to claim model ownership.

Comparison to Existing Works. By adjusting the pruning rate, our framework can support more than 0.67 pruning rate (on SVHN and Mini-ImageNet) by comparing with the previous fingerprinting work IPGuard [8], which exceeds their performance (0.5).

4.3. Comparison with other outlier detection algorithms

In this subsection, several unsupervised outlier detection (OD) algorithms are implemented and compared with ECOD in terms of effectiveness, efficiency and accuracy. As shown in the Fig.4, this experiment compares the effectiveness of COF [19], CBLOF [20] and COPOD [21] methods in terms of model feature fingerprint matching. We can see that ECOD outperforms CBLOF and COPOD algorithms on all models, and the framework using ECOD algorithm can significantly distinguish irrelevant models from piracy models, and the difference between victim and piracy models is reduced. Moreover, ECOD is superior to other methods in terms of computational efficiency. As Table 1 indicated, the computation time of COF with the same model is much higher than the other algorithms, but the effectiveness of CBLOF and COPOD is lower than that of COF. In conclusion, the framework using ECOD has superior performance and effect on model ownership verification.



(a) ResNet-18+SVHN (b) ResNet-50+Mini-ImageNet

Fig. 4. Comparison of the effectiveness of different OD algorithms.

Victim Model	Computing Time(ms)			
	CBLOF	CBPOD	COF	ECOD
ResNet-18+SVHN	1.765	0.897	67.526	**0.005**
ResNet-50+Mini-ImageNet	25.891	17.657	452.670	**0.176**

Table 1. Computation time of different OD algorithms.

5. CONCLUSION

In this paper, we propose a new framework for model ownership verification based on the fact that feature fingerprints can uniquely characterize model knowledge. We conduct extensive experiments to demonstrate the proposed framework's effectiveness, universality and efficiency. Finally, experimental results show that our framework has the potential to be widely employed due to its high accuracy and efficiency.

6. REFERENCES

[1] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1615–1631.

[2] Guobiao Li, Sheng Li, Zhenxing Qian, and Xinpeng Zhang, "Encryption resistant deep neural network watermarking," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3064–3068.

[3] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, 2017, pp. 269–277.

[4] Fang-Qi Li, Shi-Lin Wang, and Yun Zhu, "Fostering the robustness of white-box deep neural network watermarks by neuron alignment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3049–3053.

[5] Zirui Peng, Shaofeng Li, Guoxing Chen, Cheng Zhang, Haojin Zhu, and Minhui Xue, "Fingerprinting deep neural networks globally via universal adversarial perturbations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13430–13439.

[6] Si Wang and Chip-Hong Chang, "Fingerprinting deep neural networks-a deepfool approach," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[7] Siyue Wang, Xiao Wang, Pin-Yu Chen, Pu Zhao, and Xue Lin, "Characteristic examples: High-robustness, low-transferability fingerprinting of neural networks.," in *IJCAI*, 2021, pp. 575–582.

[8] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong, "Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary," in *Proceedings of the 2021 ACM AsiaCCS*, 2021, pp. 14–25.

[9] Xudong Pan, Mi Zhang, Yifan Lu, and Min Yang, "Tafa: A task-agnostic fingerprinting algorithm for neural networks," in *European Symposium on Research in Computer Security*. Springer, 2021, pp. 542–562.

[10] Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song, "Copy, right? a testing framework for copyright protection of deep learning models," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 824–841.

[11] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen, "Ecod: Unsupervised outlier detection using empirical cumulative distribution functions," *IEEE TKDE*, 2022.

[12] Yiming Li, Linghui Zhu, Xiaojun Jia, Yong Jiang, Shu-Tao Xia, and Xiaochun Cao, "Defending against model stealing via verifying embedded external features," AAAI, 2022.

[13] Abdelfatah A Tamimi, Ayman M Abdalla, and Omaima Al-Allaf, "Hiding an image inside another image using variable-rate steganography," *IJACSA*, vol. 4, no. 10, 2013.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[17] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[18] Christian Szegedy, Wei Liu, and Yangqing et al. Jia, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[19] Jian Tang, Zhixiang Chen, Ada Wai-Chee Fu, and David W Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2002, pp. 535–548.

[20] Zengyou He, Xiaofei Xu, and Shengchun Deng, "Discovering cluster-based local outliers," *Pattern recognition letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[21] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu, "Copod: copula-based outlier detection," in *2020 ICDM*. IEEE, 2020, pp. 1118–1123.