W-DOE: Wasserstein Distribution-Agnostic Outlier Exposure

Qizhou Wang, Bo Han, *Senior Member, IEEE*, Yang Liu, *Senior Member, IEEE*, Chen Gong, *Senior Member, IEEE*, Tongliang Liu, *Senior Member, IEEE*, and Jiming Liu, *Fellow, IEEE*

***Abstract*—In open-world environments, classification models should be adept at identifying out-of-distribution (OOD) data whose semantics differ from in-distribution (ID) data, leading to the emerging research in OOD detection. As a promising learning scheme, *outlier exposure* (OE) enables the models to learn from *auxiliary OOD data*, enhancing model representations in discerning between ID and OOD patterns. However, these auxiliary OOD data often do not fully represent real OOD scenarios, potentially biasing our models in practical OOD detection. Hence, we propose a novel OE-based learning method termed *Wasserstein Distribution-agnostic Outlier Exposure* (W-DOE), which is both theoretically sound and experimentally superior to previous works. The intuition is that by expanding the coverage of training-time OOD data, the models will encounter fewer unseen OOD cases upon deployment. In W-DOE, we achieve additional OOD data to enlarge the OOD coverage, based on a new data synthesis approach called *implicit data synthesis* (IDS). It is driven by our new insight that perturbing model parameters can lead to implicit data transformation, which is simple to implement yet effective to realize. Furthermore, we suggest a general learning framework to search for the synthesized OOD data that can benefit the models most, ensuring the OOD performance for the enlarged OOD coverage measured by the Wasserstein metric. Our approach comes with provable guarantees for open-world settings, demonstrating that broader OOD coverage ensures reduced estimation errors and thereby improved generalization for real OOD cases. We conduct extensive experiments across a series of representative OOD detection setups, further validating the superiority of W-DOE against state-of-the-art counterparts in the field.**

***Index Terms*—Out-of-distribution detection, reliable machine learning, open-set learning.**

Received 28 March 2024; revised 11 October 2024; accepted 9 January 2025. Date of publication 17 January 2025; date of current version 3 April 2025. The work of Qizhou Wang and Bo Han were supported in part by the NSFC General Program under Grant 62376235, in part by RGC Young Collaborative Research under Grant C2005-24Y, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011652 and Grant 2024A1515012399, in part by HKBU Faculty Niche Research Areas under Grant RC-FNRA-IG/22-23/SCI/04, and in part by HKBU CSD Departmental Incentive Scheme. The work of Chen Gong was supported in part by the NSF of China under Grant 62336003 and Grant 12371510, and in part by the NSF for Distinguished Young Scholar of Jiangsu Province under Grant BK20220080. Recommended for acceptance by R. Cucchiara. *(Corresponding author: Bo Han.)*

Qizhou Wang, Bo Han, Yang Liu, and Jiming Liu are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: csqzwang@comp.hkbu.edu.hk; bhanml@comp.hkbu.edu.hk; csygliu@comp.hkbu.edu.hk; jiming@comp.hkbu.edu.hk).

Chen Gong is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: chen.gong@sjtu.edu.cn).

Tongliang Liu is with the Sydney AI Centre and the School of Computer Science, Faculty of Engineering, University of Sydney, Camperdown, NSW 2050, Australia (e-mail: tongliang.liu@sydney.edu.au).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2025.3531000, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3531000

I. Introduction

DEEP learning systems operating in open-world environments frequently encounter out-of-distribution (OOD) data, which differ in label space from the in-distribution (ID) samples. Since these models cannot make valid predictions for semantic shifts raised by OOD cases, it is necessary to refrain from making label predictions. This issue has spurred recent attention in the area of OOD detection, where the model should identify anomalies raised by OOD data yet make accurate predictions for ID data [1]. Nowadays, OOD detection has attracted intensive attention in reliable machine learning, offering an anomaly-handling mechanism for numerous safety-critical applications, such as auto-driving, medical analysis, and financial security [2], [3], [4], [5].

OOD detection remains challenging as models can predict with arbitrary-high confidence for unknown data, thereby hindering the direct use of model confidence to identify OOD cases [6], [7]. Consequently, considerable efforts have been dedicated to pursuing improved OOD detection methods, primarily falling into two main categories, namely, *post-hoc* and *fine-tuning*. Post-hoc approaches utilize well-trained models with fixed parameters and focus on designing various scoring functions to assess the confidence of being ID [6], [8], [9], [10], [11], [12]. These methods presume that the original models are adequate for OOD detection, but they require the proper extraction of informative representations through scoring functions to be effective. On the other hand, fine-tuning approaches further allow for model training, which directly enhances the model representation and thereby boosts its capability to detect OOD data [2], [13], [14], [15], [16], [17], [18]. These methods benefit from explicit knowledge of unknowns, thus typically revealing more reliable performance in real-world scenarios over the post-hoc approaches [3].

Among fine-tuning methods, *outlier exposure* (OE) [13], [19], [20], [21] represents an effective strategy, which enhances OOD detection by incorporating *auxiliary OOD data* into training. It directly trains the model to differentiate between ID and auxiliary OOD data, refining model representations to boost OOD detection. It is one of the most potent learning strategies nowadays, as the model can benefit from explicit knowledge about OOD. However, despite their effectiveness, OE-based methods still have their limitations. One of the main challenges

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

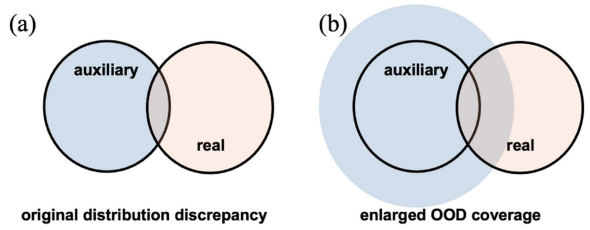


Fig. 1. *W-DOE Motivation*. By enlarging the OOD coverage as in (b), the distance between training and real OOD data is shrunk compared with the original OE as in (a). Therefore, W-DOE can mitigate OOD distribution discrepancy and thus improve OOD detection.

stems from data openness, where one cannot know what types of OOD data we will encounter in the open world. As a result, the auxiliary OOD data may differ arbitrarily from the real ones, indicating an *OOD distribution discrepancy* between training and test. This discrepancy generally has detrimental effects on OOD performance in real-world scenarios (cf., Section III), while seldom has been made for such a common yet important problem, motivating the main focus of this paper.

To overcome this issue, we introduce a novel OE-based learning method named *Wasserstein Distribution-agnostic Outlier Exposure* (W-DOE). Our method aims to broaden the coverage of training OOD data by synthesizing new OOD examples that are distinct from the (original) auxiliary ones. By training our model to perform well upon such an expanded distribution, we shrink the OOD distribution discrepancy and thereby mitigate its negative impacts (cf., Fig. 1). To realize our W-DOE, (a) how to synthesize OOD data and (b) how to guarantee overall performance on such a data distribution are the key questions to be answered.

For the first question, we present a new approach for synthesizing OOD data, termed *implicit data synthesis* (IDS). This method can effectively transform existing data into very different ones, based on our novel insight that model perturbation implicitly leads to data transformation. Accordingly, one can make the model learn from such *implicit (transformed) data* by model updating after its perturbation. IDS is simple to implement and flexible for synthetic data that deviate from the original ones. In Section VI-A, we justify two facets that can support our effectiveness: (a) the implicit data follow a different distribution from that of the original ones (cf., Theorem 4), and (b) the discrepancy between the original and transformed data distributions can be large (cf., Proposition 1). Overall, our analysis verifies that IDS can implicitly synthesize diverse data that are largely different from the original ones, thus having the potential to benefit OOD detection.

For the second question, we propose an advanced learning framework that guarantees performance on expanded OOD coverage with synthesized data. Therein, we define OOD coverage through a Wasserstein ball centered on the auxiliary OOD distribution, which can facilitate both our theoretical analysis (cf., Section VI-B) and the practical effectiveness (cf., Section VII). Then, we introduce a worst-case learning scheme upon the OOD coverage, training on data that exhibit poorest performance, thereby upper bounding the worst-case performance. It leads to a neat realization of W-DOE when combined with IDS in Algorithm 1, which largely improves upon OE in practice. In theory, we also demonstrate that W-DOE can mitigate the OOD distribution discrepancy, where a broader OOD coverage results in improved detection performance and a tighter generalization bound over OE.

We conduct a series of experiments in Section VII on commonly used benchmarks, verifying the superiority of our W-DOE over competitive baselines when facing OOD distribution discrepancy. For example, our W-DOE decreases the average FPR95 by 8.33%, 20.27%, and 24.55% compared with the original OE on the CIFAR-10, the CIFAR-100, and ImageNet benchmarks, respectively. We summarize the contributions of this paper as follows:

- *Algorithm:* In Section IV, we propose IDS to augment the training-time OOD data in a simple yet effective way. In Section V, we also suggest W-DOE as a systematic way to mitigate the impacts of OOD distribution discrepancy.
- *Theory:* In Section VI-A, we verify the validity of IDS in generating diverse data. In Section VI-B, we reveal the effectiveness of W-DOE when facing the OOD distribution discrepancy.
- *Experiment:* In Section VII, we conduct extensive experiments to verify the effectiveness of W-DOE under representative OOD detection setups with OOD distribution discrepancy.

Difference from the conference version: Comparing with the preliminary version published in ICLR 2023 [22], we have made the substantial extensions, summarizing in the following:

- *Further Formalization:* Besides providing only the heuristics as in our conference version, we further formalize the impact of OOD distribution discrepancy on OE in Section III. Our results justify that larger distribution discrepancy truly hinders the reliability of OE-based methods.
- *Previous Drawback:* We identify a key drawback for the method in our conference version, where the training dynamic is relatively unstable due to the unconstrained worst-case search. We analyze such a problem based on our theoretical analysis from Section VI-B, which are further summarized in Section VII-F with empirical justification.
- *Improved Methodology:* To address the above drawback in our conference version, we suggest a systematic framework in Section V that incorporates the Wasserstein constraint during the worst-case data search. It not only improves the stability of our learning dynamics and also enhances the overall performance in OOD detection.
- *In-depth Analysis:* From the statistic learning theories, we prove that our method can properly mitigate the impacts of distribution discrepancy and lead to better open-world performance for unseen OOD data in Section VI-B. It complements our conference version, where we only verify the convergence with respect to the auxiliary OOD distribution.

We offer new insights to comprehend the distribution discrepancy issue, introducing new algorithms and new theories. We also go beyond our evaluation setups in our conference version, evaluating more challenging datasets (e.g., Oxford-Pets) and

TABLE I
NOTATIONS AND ASSOCIATED DESCRIPTIONS

Notation	Description
Variable and Space	
$\mathcal{X}$ and $\mathcal{Y}$	feature and ID label spaces
$\mathcal{F}$ and $\mathcal{W}$	hypothesis and parameter spaces
$X_{\mathrm{I}}, X_{\mathrm{A}}$, and X_{O}	ID, auxiliary OOD, and real OOD variables
Y_{I} and Y_{O}	ID and OOD labels
Distribution and Measurement	
$D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $D_{X_{\mathrm{I}}}$	ID joint and ID margin distributions
$D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$ and $D_{X_{\mathrm{O}}}$	OOD joint and OOD margin distributions
$D_{X_{\mathrm{A}}}$ and $D_{X_{\mathrm{S}}}$	auxiliary and synthesized margin distribution
$\mathtt{KL}(\cdot\|\|\cdot)$ and $\mathtt{WD}(\cdot,\cdot)$	KL divergence and Wasserstein distance
Data and Model	
n, m, and c	numbers of ID, auxiliary OOD, and classes
i, j, and k	indices of data, labels, and layers
$\mathbf{x}$, y, and $\mathbf{z}$	instances, labels, and embeddings
$\mathbf{w}$ and $\mathbf{W}$	vector and matrix forms of parameters
$\mathbf{f}(\cdot;\mathbf{w})$ and $\boldsymbol{\phi}(\cdot;\mathbf{w})$	model outputs and embedding outputs
$\mathfrak{R}_n(\mathcal{F})$ and $C_{\mathcal{F}}$	Rademacher complexity and the estimation
Loss and Risk	
α	trade-off parameter
$\ell_{\mathrm{ID}}(\cdot,\cdot)$ and $\ell_{\mathrm{OOD}}(\cdot)$	ID and OOD loss functions
$R_{\mathrm{I}}(\cdot)$, $R_{\mathrm{A}}(\cdot)$, and $R_{\mathrm{O}}(\cdot)$	ID, auxiliary OOD, and real OOD risks
$R_{\mathrm{S}}(\cdot;D_{X_{\mathrm{S}}})$	expected synthesized OOD risk
$R_{\mathrm{OE}}(\cdot)$ and $R_{\mathrm{W}}(\cdot;\rho)$	expected OE and W-DOE risks
L_{ID} and L_{OOD}	Lipschitz constants for ID and OOD risks
A_{ID} and A_{OOD}	bounding constants for ID and OOD risks
$\epsilon(C,L,A)$	generalization error for PAC
W-DOE	
$\mathbf{p}$ and $\mathbf{g}$	perturbation parameters and its gradients
$\mathbf{A}$ and $\mathbf{P}$	two matrix forms of perturbations
λ and β	perturbation strength and sliding parameter
$\mathbf{t}(\cdot,\mathbf{w},\lambda\mathbf{p})$	implicit transform function
$\mathfrak{B}_{\mathrm{A}}^{\rho}$	auxiliary OOD coverage
ρ and γ	radius and regularization strength
$\mathtt{OR}(\cdot;D_{X_{\mathrm{S}}})$ and $\mathtt{WOR}(\cdot;\rho)$	OOD regret and worst-case OOD regret
$\kappa(\cdot;\mathbf{w})$	gradient norm-based OR estimation

more realistic setups (e.g., wild OOD and medical OOD) to reveal our effectiveness. Our work represents substantial enhancements over the conference version, thereby fulfilling the criteria of IEEE Transactions on Pattern Analysis and Machine Intelligence.

II. PRELIMINARY

Let $\mathcal{X}$ be the feature space and $\mathcal{Y}=\{1,\ldots,c\}$ be the ID label space. We consider the ID distribution $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$, a joint distribution defined over $\mathcal{X}\times\mathcal{Y}$, where X_{I} and Y_{I} are random variables whose outputs are from $\mathcal{X}$ and $\mathcal{Y}$, respectively. We also have an OOD joint distribution $D_{X_{\mathrm{O}}Y_{\mathrm{O}}}$, where X_{O} is a random variable from $\mathcal{X}$ and Y_{O} is an unknown variable whose values do not belong to $\mathcal{Y}$, i.e., $Y_{\mathrm{O}}\notin\mathcal{Y}$ [23]. Furthermore, we consider the classification model $\mathbf{f}(\cdot;\mathbf{w}):\mathcal{X}\to\mathbb{R}^c$ with logit outputs, parameterized by $\mathbf{w}$ from the parameter space $\mathcal{W}$. The key notions are in Table I.

A. OOD Scoring

Building upon $\mathbf{f}(\cdot;\mathbf{w})$, our goal is to use a *scoring function* $s(\cdot;\mathbf{w}):\mathcal{X}\to\mathbb{R}$ to discern test-time inputs drawn from $D_{X_{\mathrm{I}}}$ and $D_{X_{\mathrm{O}}}$. Typically, if the *score value* $s(\mathbf{x};\mathbf{w})$ is larger than a threshold $\tau\in\mathbb{R}$, the associated input $\mathbf{x}\in\mathcal{X}$ is taken as ID, otherwise an OOD case. Here, we present two typical examples.

Maximum Softmax Prediction (MSP): As a well-known baseline, [6] uses the maximal dimension of the softmax predictions to indicate OOD and the scoring function is of the form:

$$s_{\mathrm{MSP}}(\mathbf{x};\mathbf{w})=\max_j\ \mathtt{softmax}_j\ \mathbf{f}(\mathbf{x};\mathbf{w}), \tag{1}$$

where $\mathtt{softmax}_j(\cdot)$ indicates the j-th softmax output. Ideally, the labels of OOD data do not align with any dimension of model predictions, and thus softmax outputs should be low across all dimensions. However, due to the well-known over-confidence issues [6], [10], MSP often makes mistakes in practice.

MaxLogit Prediction (MLP): [24] suggests a simple solution that can improve MSP, using maximal logits instead of softmax predictions to construct the scoring function, namely,

$$s_{\mathrm{MLP}}(\mathbf{x};\mathbf{w})=\max_j\mathbf{f}^j(\mathbf{x};\mathbf{w}), \tag{2}$$

where $\mathbf{f}^j(\mathbf{x};\mathbf{w})$ denotes the j-th element of model outputs. The MLP score works better than MSP, especially when facing large ID label space, and thus widely adopted in practice.

B. Outlier Exposure

Due to calibration failures [25], normal-trained models often exhibit a lot of mistakes, even when combined with those advanced scoring strategies. To this end, OE [19] suggests that models should be further trained, directly learning to differentiate between ID and OOD patterns by engaging so-called auxiliary OOD distribution $D_{X_{\mathrm{A}}}$. We formalize the problem as follows.

Definition 1 (OE Setup): Let $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$, $D_{X_{\mathrm{O}}}$, and $D_{X_{\mathrm{A}}}$ be the ID joint, the real OOD margin, and the auxiliary OOD margin distributions. Then, based on the ID data $\{(\mathbf{x}_{\mathrm{I}}^i,y_{\mathrm{I}}^i)\}_{i=1}^n$ and the auxiliary OOD data $\{\mathbf{x}_{\mathrm{A}}^i\}_{i=1}^m$ i.i.d. drawn from $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $D_{X_{\mathrm{A}}}$ respectively, the goal of OE is to learn from such data, so that the model $\mathbf{f}(\cdot;\mathbf{w})$ is good at OOD detection, i.e., for any input $\mathbf{x}$,

- if $\mathbf{x}$ is an observation from $D_{X_{\mathrm{I}}}$, the model $\mathbf{f}(\cdot;\mathbf{w})$ can classify $\mathbf{x}$ into its correct ID label;
- if $\mathbf{x}$ is an observation from $D_{X_{\mathrm{O}}}$, the model $\mathbf{f}(\cdot;\mathbf{w})$ can detect $\mathbf{x}$ as an OOD case.

Remark 1: OE considers the open-world setting, where the auxiliary distribution $D_{X_{\mathrm{A}}}$ may differ arbitrarily from the real distribution $D_{X_{\mathrm{O}}}$. It reflects our limited knowledge about real OOD data during training since we cannot anticipate and enumerate all OOD cases that we will encounter in the future. It is a standard setting widely adopted in practice [19], [21].

Learning Strategy: Generally speaking, OE takes OOD detection as a binary classification problem, learning to discern ID and OOD patterns. Overall, the empirical OE risk can be written as

$$\widehat{R}_{\mathrm{OE}}(\mathbf{w})=\widehat{R}_{\mathrm{I}}(\mathbf{w})+\alpha\widehat{R}_{\mathrm{A}}(\mathbf{w}), \tag{3}$$

with α the trade-off parameter. The first term $\widehat{R}_{\mathrm{I}}(\mathbf{w})$ handles the ID cases, making scores for ID data higher; the second term $\widehat{R}_{\mathrm{A}}(\mathbf{w})$ handles the OOD cases, making scores for auxiliary OOD data lower. We also define the expected counterpart of (3)

as

$$R_{\mathrm{OE}}(\mathbf{w}) = R_{\mathrm{I}}(\mathbf{w}) + \alpha R_{\mathrm{A}}(\mathbf{w}).$$

Loss Functions: The expansion forms of $R_{\mathrm{I}}(\mathbf{w})$ and $R_{\mathrm{A}}(\mathbf{w})$ are

$$R_{\mathrm{I}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim D_{X_{\mathrm{I}}Y_{\mathrm{I}}}} \ell_{\mathrm{ID}}\left(\mathbf{f}(\mathbf{x};\mathbf{w}), y\right)$$

and

$$R_{\mathrm{A}}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}} \ell_{\mathrm{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right),$$

respectively, where ℓ_{ID} and ℓ_{OOD} denote associated loss functions. Similarly, the expansions for the empirical risks are

$$\widehat{R}_{\mathrm{I}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{ID}}(\mathbf{f}\left(\mathbf{x}_{\mathrm{I}}^{i};\mathbf{w}\right), y_{\mathrm{I}}^{i})$$

and

$$\widehat{R}_{\mathrm{A}}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathrm{OOD}}\left(\mathbf{f}(\mathbf{x}_{\mathrm{A}}^{i};\mathbf{w})\right).$$

We can define the empirical OE risk $\widehat{R}_{\mathrm{OE}}$ accordingly. Following [19], we employ the cross entropy loss for ID cases, i.e.,

$$\ell_{\mathrm{ID}}(\mathbf{f}\left(\mathbf{x};\mathbf{w}\right), y) = -\log \mathtt{softmax}_y\ \mathbf{f}(\mathbf{x};\mathbf{w}),$$

and the KL-divergence between the uniform distribution and the softmax prediction for the OOD cases, namely,

$$\ell_{\mathrm{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) = \mathtt{KL}\left(\mathbf{1}/c \,||\, \mathtt{softmax}\ \mathbf{f}(\mathbf{x};\mathbf{w})\right),$$

where $\mathtt{KL}(\cdot\,||\,\cdot)$ denotes the KL-divergence. It directly makes high / low MSP scores for ID / OOD data and works empirically well for many other scoring strategies, including MLP scoring.

III. Distribution Discrepancy

As aforementioned, OE considers an open-world setting, where the auxiliary OOD margin $D_{X_{\mathrm{A}}}$ differs from the real OOD margin $D_{X_{\mathrm{O}}}$ [19]. This discrepancy leads to the discrepancy between training (i.e., $D_{X_{\mathrm{A}}}$) and test (i.e., $D_{X_{\mathrm{O}}}$) OOD distributions, named as the *OOD distribution discrepancy* in our following.

Distributional Metric: To understand the impacts of OOD distribution discrepancy, it is essential to have an appropriate metric that can quantify the distance or divergence between data. In this paper, we suggest the use of the Wasserstein distance [26], which has been widely adopted in the field of statistical learning.

Definition 2 (Wasserstein Measurement): The Wasserstein distance between two distributions D and D' is given by:

$$\mathtt{WD}(D, D') = \inf_{\pi\in\Pi(D,D')} \mathbb{E}_{(\mathbf{x},\mathbf{x}')\sim\pi}||\mathbf{x}-\mathbf{x}'||,$$

where $\Pi(D, D')$ is the coupling space and $||\cdot||$ is the L2 norm.

For two distributions denoted by D and D', a large $\mathtt{WD}(D, D')$ indicates their discrepancy is also large. Therefore, $\mathtt{WD}(D_{X_{\mathrm{O}}}, D_{X_{\mathrm{A}}}) > 0$ holds due to the OOD distribution discrepancy in the open-world setting. Moreover, to ease our discussion, we define the OOD risk w.r.t. the real OOD distribution $D_{X_{\mathrm{O}}}$, i.e.,

$$R_{\mathrm{O}}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{O}}}} \ell_{\mathrm{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right).$$

Then, by the following theorem, we quantify the negative impacts of OOD distribution discrepancy for OE-based methods.

Theorem 1: We adhere to Assumption 1, which posits bounded Rademacher complexity, bounded loss functions, and continuity—conditions commonly satisfied by deep models. We define $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \widehat{R}_{\mathrm{OE}}(\mathbf{w})$. With probability at least $1-\delta$, we have

$$R_{\mathrm{I}}(\widehat{\mathbf{w}}) + \alpha R_{\mathrm{O}}(\widehat{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} R_{\mathrm{OE}}(\mathbf{w})$$
$$\leq \alpha L\ \mathtt{WD}(D_{X_{\mathrm{O}}}, D_{X_{\mathrm{A}}}) + (1+\alpha)\frac{2\epsilon(C_{\mathcal{F}}, L, A)}{\sqrt{\min\{m,n\}}},$$

where $L = \max\{L_{\mathrm{ID}}, L_{\mathrm{OOD}}\}$, $A = \max\{A_{\mathrm{ID}}, A_{\mathrm{OOD}}\}$, and $C_{\mathcal{F}}$ relates to the Rademacher complexity of the model space $\mathcal{F}$. Moreover, the term $\epsilon(C_{\mathcal{F}}, L, A)$ serves as the key part for the upper bound of the generalization error, as defined in (14).

To keep clarity throughout our main discussion, we defer the proof to the later text within Section VI-B.

Remark 2: The expression $R_{\mathrm{I}}(\mathbf{w}) + \alpha R_{\mathrm{A}}(\mathbf{w})$ quantifies the expected performance on real data for models trained on auxiliary OOD data; the term $\epsilon(C_{\mathcal{F}}, L, A)/\sqrt{\min\{m,n\}}$ relates to the Rademacher complexity of the hypothesis space as well as the sample sizes of m and n. Due to this upper bound, the impacts of the OOD distribution discrepancy are reflected by the Wasserstein distance between the auxiliary and the real OOD data, i.e., $\mathtt{WD}(D_{X_{\mathrm{O}}}, D_{X_{\mathrm{A}}})$. Consequently, the detection performance on unseen data degrades linearly with respect to the Wasserstein distance between the auxiliary and the real OOD distributions.

Overall, although OE-based methods can largely enhance OOD detection, the OOD distribution discrepancy limits its reliability in the open world. It underscores the possibility to further improve OE, motivating the primary focus of this paper.

IV. OOD Data Synthesis

The OOD distribution discrepancy stems from our limited awareness of real OOD cases. A natural solution is to expose the model to a broad range of OOD data beyond the auxiliary ones, thereby expanding the OOD coverage for OE and bridging the gap between training and test. However, manually gathering such data is hard, often requiring impractical amounts of time and labor. Consequently, we recommend synthesizing such data, freeing from cumbersome data crawling and tedious human efforts.

Implicit Data Synthesis: Previous works have demonstrated that data synthesis is hard, requiring meticulous human design [20], resource-demanding generative models [13], and unstable feature optimization [27]. In response to these challenges, we suggest a novel strategy named *Implicit Data Synthesis* (IDS), which is simple to implement yet powerful to synthesize.

The key motivation is that *perturbing model parameters has the same impact as transforming input data*, where specific model perturbations indicate particular input transformation. Formally speaking, given $\mathbf{p}$ the *perturbation parameters* and $\lambda > 0$ the *perturbation strength*, the following two events are equivalent:

- *Model Perturbation:* perturbing model parameters from $\mathbf{w}$ to $\mathbf{w}+\lambda\mathbf{p}$, given the input data $\mathbf{x}$;
- *Data Transformation:* transforming input data from $\mathbf{x}$ to $\mathbf{t}(\mathbf{x};\mathbf{w},\lambda\mathbf{p})$, given the model with parameters $\mathbf{w}$,

where $\mathbf{t}(\cdot;\mathbf{w},\lambda\mathbf{p})$ is an *implicit transform function*, depending on $\mathbf{w}$ and $\lambda\mathbf{p}$. Thus, when updating the model after its parameter perturbation, we can implicitly make the model learn from its transformed forms, named *implicit data* in the following. The transformed data can arbitrarily differ from the original ones, thus enlarging the coverage of OOD cases during OE training. Please refer to Section VI-A for a formal discussion.

Informal Justification: The underlying insight is quite simple. As an illustration, we consider only the k-th layer of the model, denoting $\mathbf{z}^{(k)}$ the inputs, $\mathbf{W}^{(k)}$ the matrix form of parameters, $\mathbf{P}^{(k)}$ the matrix form of perturbation, $\boldsymbol{\sigma}(\cdot)$ the activation function, and $\mathbf{f}^{(k)}_{\mathbf{W}^{(k)}}(\mathbf{z}^{(k)})=\boldsymbol{\sigma}(\mathbf{W}^{(k)}\mathbf{z}^{(k)})$ the outputs. To ease our discussion, we consider the equivalent form of $\mathbf{W}^{(k)}+\lambda\mathbf{P}^{(k)}$:

$$\mathbf{W}^{(k)}(\mathbf{I}+\lambda\mathbf{A}^{(k)}),$$

where $\mathbf{A}^{(k)}$ is the perturbation matrix with $\mathbf{P}^{(k)}=\mathbf{A}^{(k)}\mathbf{W}^{(k)}$. Then, we observe that perturbation in the parameter space corresponds to perturbation in the feature space, following

$$\begin{aligned}\mathbf{f}^{(k)}_{\mathbf{W}^{(k)}(\mathbf{I}+\lambda\mathbf{A}^{(k)})}(\mathbf{z}^{(k)}) &= \boldsymbol{\sigma}\left([\mathbf{W}^{(k)}(\mathbf{I}+\lambda\mathbf{A}^{(k)})]\mathbf{z}^{(k)}\right)\\ &= \boldsymbol{\sigma}\left(\mathbf{W}^{(k)}[(\mathbf{I}+\lambda\mathbf{A}^{(k)})\mathbf{z}^{(k)}]\right)\\ &= \mathbf{f}^{(k)}_{\mathbf{W}^{(k)}}\left((\mathbf{I}+\lambda\mathbf{A}^{(k)})\mathbf{z}^{(k)}\right). \qquad (4)\end{aligned}$$

The above equation connects model perturbation, namely, $\mathbf{W}^{(k)}(\mathbf{I}+\lambda\mathbf{A}^{(k)})$, to feature transformation, namely, $(\mathbf{I}+\lambda\mathbf{A}^{(k)})\mathbf{z}^{(k)}$. Overall, we can implicitly make the model learn from transformed data, where we just need to perturb model parameters yet keep the original inputs fixed. Such a novel data synthesis scheme is flexible across various model structures and requires less computational efforts over generation-based methods [13], [27], which depend on the tedious training of explicit data generators. It also provides greater extensibility than manual designs [20], [28], which rely on meticulously crafted assumptions about data distributions. We provide a rigorous discussion in Section VI-A. Specifically, in Theorem 4, we extend (4) to encompass perturbations for the whole model. In Proposition 1, we argue that IDS can benefit from layer-wise non-linearity, suggesting that deeper models can lead to more complex transformations.

V. Distribution-Agnostic OOD Learning

IDS allows the model to implicitly learn from additional OOD data; however, not all implicit data can benefit the current model. For example, when the model already performs well on certain synthesized data, further learning from them could be redundant. How to ensure that synthesized / implicit data are beneficial remains to be answered, which will be discussed as follows.

A. Worst-Case Data Search

We propose *Wasserstein Distribution-agnostic Outlier Exposure* (W-DOE), cf., Algorithm 1, an effective OE-based learning approach powered by IDS to mitigate the OOD distribution discrepancy. During training, we enlarge the coverage of OOD data beyond the auxiliary ones, defined by the *distribution coverage*.

Definition 3 (Distribution Coverage): A distribution coverage $\mathfrak{B}^{\rho}_{D}$ centered on the distribution D is given by

$$\mathfrak{B}^{\rho}_{D}=\{D':\mathtt{WD}(D,D')\leq\rho\},$$

where ρ is the maximal-allowed distance between distributions.

Remark 3: The distribution coverage encompasses a set of distributions rather than a single one, as making the model work well across distributions is more likely to guarantee its performance in unknown cases. Furthermore, we consider the distribution coverage centered on the auxiliary OOD distribution, named the OOD distribution coverage, denoted by $\mathfrak{B}^{\rho}_{\mathrm{A}}$ in the following.

W-DOE aims to guarantee the performance in the expanded OOD distribution coverage, such that the model can learn from all those data following various distributions within the coverage. Fig. 1 illustrates our heuristics in mitigating the OOD distribution discrepancy—learning from the enlarged OOD coverage shrinks the distance between the auxiliary and the real OOD distribution, thus mitigating their discrepancy. Such a motivation is inspired by Theorem 1, with further justifications in Section VI-B.

Performance Measurement: To ensure the overall OOD performance within the OOD distribution coverage, we should upper bound the worst-case detection performance therein. In W-DOE, the worst-case performance is measured by the *worst-case OOD regret* (WOR), which is formalized in the following.

Definition 4 (t-case OOD Regret): Given the model $\mathbf{f}_{\mathbf{w}}$ and the OOD distribution coverage $\mathfrak{B}^{\rho}_{\mathrm{A}}$, the worst-case OOD regret is

$$\mathtt{WOR}(\mathbf{w};\rho)=\sup_{D_{X_{\mathrm{S}}}\in\mathfrak{B}^{\rho}_{\mathrm{A}}}\left[R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})-\inf_{\mathbf{w}}R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})\right],$$

where $R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})=\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{S}}}}\ell_{\mathrm{OOD}}(\mathbf{f}(\mathbf{x};\mathbf{w}))$ is the OOD performance for the synthesized OOD distribution $D_{X_{\mathrm{S}}}$.

Remark 4: In the place of the OOD regret, one may measure the performance directly by the risk, i.e., $\sup_{D_{X_{\mathrm{S}}}\in\mathfrak{B}^{\rho}_{\mathrm{A}}}R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})$. However, we emphasize that our regret-based measurement is better due to the limited fitting power of the model. In particular, there exist some OOD data with high risk values that the model cannot adequately fit. In this case, OOD regret can better reflect the most beneficial data that the model can improve upon. Therefore, using WOR is better than that of the risk counterpart.

B. Learning Objective

We emphasize that minimizing the WOR ensures the upper bound on the overall model performance within the enlarged OOD coverage. Therefore, implicit OOD data that lead to WOR are of interest, and learning from them can benefit the model most. To mitigate the OOD distribution discrepancy, W-DOE

enhances the classical OE with the refined learning objective as

$$R_{\mathrm{W}}(\mathbf{w};\rho) = R_{\mathrm{I}}(\mathbf{w}) + \alpha \sup_{D_{X_{\mathrm{S}}}\in\mathfrak{B}_{\mathrm{A}}^{\rho}} \underbrace{[R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}}) - \inf_{\mathbf{w}} R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})]}_{\mathtt{OR}(\mathbf{w};D_{X_{\mathrm{S}}})}. \quad (5)$$

Directly solving (5) is non-trivial in practice, due to the regret estimation and the constraint optimization for $\mathtt{WOR}(\mathbf{w};\rho)$. Therefore, we propose a series of modifications to ease its computation.

Regret Estimation: The computation of the exact regret, $\mathtt{OR}(\mathbf{w};D_{X_{\mathrm{S}}})$, is hard due to the necessity to enumerate the optimal risk for an infinite number of distributions D_{S} within an infinite distribution space. Instead, following [29], [30], we suggest its effective estimation, involving the computations of the gradient norms w.r.t. the risk $R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})$, as outlined below:

$$\mathtt{OR}(\mathbf{w};D_{X_{\mathrm{S}}}) \approx ||\nabla_{t|t=1.0}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{S}}}}\ell_{\mathrm{OOD}}\left(t\cdot\mathbf{f}(\mathbf{x};\mathbf{w})\right)||^2. \quad (6)$$

Intuitively, a larger gradient norm indicates that the current model is far from its optimal, and thus the corresponding regret is large. It leads to an efficient estimation of the regret values.

Constraint Optimization: $\mathtt{WOR}(\mathbf{w};\rho)$ in (5) subjects to the distribution constraint meanwhile requires infinite-dimensional search. Generally, it is intractable to directly solve $\mathtt{WOR}(\mathbf{w};\rho)$, and thus we introduce the dual theorem as follows.

Theorem 2 (Dual Theorem [31], [32]): Consider the loss $\ell(\cdot)$ with the coverage constraint $\mathfrak{B}_D^{\rho} = \{D' : \mathtt{WD}(D,D') \le \rho\}$:

$$\sup_{D_{X_{\mathrm{S}}}\in\mathfrak{B}_D^{\rho}} \mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{S}}}}\ell\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right),$$

it equals the dual problem

$$\inf_{\gamma\ge 0}\left\{\gamma\rho + \mathbb{E}_{\mathbf{x}\sim D}\sup_{\mathbf{x}'\in\mathcal{X}}\left[\ell\left(\mathbf{f}(\mathbf{x}';\mathbf{w})\right) - \gamma||\mathbf{x},\mathbf{x}'||^2\right]\right\}, \quad (7)$$

for the associated $\rho > 0$.

Remark 5: Theorem 2 changes the infinite-dimensional searching to a finite-dimensional space, which can ease computation. Equation (7) can be simplified when assuming predefined γ [33], thus

$$R(\mathbf{w};\rho) = \mathbb{E}_{\mathbf{x}\sim D}\sup_{\mathbf{x}'\in\mathcal{X}}\left\{\ell\left(\mathbf{f}(\mathbf{x}';\mathbf{w})\right) - \gamma||\mathbf{x},\mathbf{x}'||^2\right\}. \quad (8)$$

W-DOE Objective: Theorem 2 cannot be applied to the regret estimation in (6) directly, while it is feasible for its upper bound, i.e., $\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{S}}}}||\nabla_{t|t=1.0}\ell_{\mathrm{OOD}}(t\cdot\mathbf{f}(\mathbf{x};\mathbf{w}))||^2$. Substituting $\ell(\mathbf{f}(\mathbf{x};\mathbf{w})) = ||\nabla_{t|t=1.0}\ell_{\mathrm{OOD}}(t\cdot\mathbf{f}(\mathbf{x};\mathbf{w}))||^2$ to (8), we propose the final W-DOE learning objective, which is given by

$$R_{\mathrm{I}}(\mathbf{w}) + \alpha\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}}\sup_{\mathbf{x}'\in\mathcal{X}}\left\{\kappa(\mathbf{x};\mathbf{w}) - \gamma||\mathbf{x},\mathbf{x}'||^2\right\}, \quad (9)$$

where we define $\kappa(\mathbf{x};\mathbf{w}) = ||\nabla_{t|t=1.0}\ell_{\mathrm{OOD}}(t\cdot\mathbf{f}(\mathbf{x};\mathbf{w}))||^2$.

C. Realization

In W-DOE, data searching is realized by IDS, and model updating is conducted in a stochastic manner. The overall realization of our method is summarized in Algorithm 1, where we emphasize several important points in the following.

Data Synthesis: IDS is used in (9) for the worst-case data search, leading to the following W-DOE objective with IDS:

$$R_{\mathrm{I}}(\mathbf{w}) + \alpha\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}}\sup_{\mathbf{p}\in\mathcal{W}}\left\{\kappa(\mathbf{x};\mathbf{w}+\lambda\mathbf{p}) - \gamma||\mathbf{x},\mathbf{t}(\mathbf{x};\mathbf{w},\lambda\mathbf{p})||^2\right\},$$

where $\mathcal{W}$ denotes the search space of model perturbation and $\mathbf{t}(\mathbf{x};\mathbf{w},\lambda\mathbf{p})$ is the implicit transformed data. However, deriving the exact $\mathbf{t}(\cdot;\mathbf{w},\lambda\mathbf{p})$ is hard, thus we approximate its behaviors in the embedding space of the model: Considering embedding features in the penultimate layer as $\boldsymbol{\phi}(\mathbf{x};\mathbf{w})$, one can measure the difference between original data and implicit data as $||\boldsymbol{\phi}(\mathbf{x};\mathbf{w}),\boldsymbol{\phi}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p})||^2$. Thereby, we have

$$R_{\mathrm{I}}(\mathbf{w}) + \alpha\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}}\sup_{\mathbf{p}\in\mathcal{W}}\left\{\boldsymbol{\kappa}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p}) - \gamma||\boldsymbol{\phi}(\mathbf{x};\mathbf{w}),\boldsymbol{\phi}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p})||^2\right\},$$

where OOD data are implicitly synthesized in the input space while explicitly measured in the embedding space.

Perturbation Estimation: Gradient-based optimization is adopted to find the proper $\mathbf{p}$. Given the data $\mathbf{x}$, each updating step of $\mathbf{p}$ is

$$\mathbf{p} \leftarrow \mathbf{p} + \nabla_{\mathbf{p}}\left[\boldsymbol{\kappa}(\mathbf{x};\mathbf{w}+\alpha\mathbf{p}) - \gamma||\boldsymbol{\phi}(\mathbf{x};\mathbf{w}),\boldsymbol{\phi}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p})||^2\right],$$

Determining the appropriate $\mathbf{p}$ for each $\mathbf{x}$ results in prohibitively high computing costs, requiring at least m times more computation over OE. Instead, we propose using a unified $\mathbf{p}$ that is searched to induce the overall worst OOD performance across the original data. Accordingly, we have the updating rule of

$$\mathbf{p} \leftarrow \mathbf{p} + \left(\nabla_{\mathbf{p}}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}}\boldsymbol{\kappa}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p}) - \nabla_{\mathbf{p}}\mathbb{E}_{\mathbf{x}\sim D_{X_{\mathrm{A}}}}\gamma||\boldsymbol{\phi}(\mathbf{x};\mathbf{w}),\boldsymbol{\phi}(\mathbf{x};\mathbf{w}+\lambda\mathbf{p})||^2\right). \quad (10)$$

Stochastic Learning: We consider a stochastic realization for model updating, where random batches $B_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $B_{X_{\mathrm{A}}}$ are sampled i.i.d. from $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $D_{X_{\mathrm{A}}}$ respectively, in each training step. Perturbation estimation and model updating are then conducted on a batch-wise basis subsequently. However, determining the optimal $\mathbf{p}$ may necessitate multiple rounds of forward / backward propagation as per (10), which is computationally expensive. Instead, we employ the *sliding moving average* to approximate the optimal $\mathbf{p}$. It requires only a single round of forward / backward propagation per step for (10) while allowing for the accumulation of more gradient information over time from previous estimations. The key modifications to implement the sliding moving average are articulated as follows:

- $\mathbf{p}$ should be initialized by its current moving average;
- the gradient $\mathbf{g}$ w.r.t. $\mathbf{p}$ is calculated with one step of forward / backward propagation;
- the moving average of $\mathbf{p}$ can be updated via

$$\mathbf{p} \leftarrow (1-\beta)\mathbf{p} + \beta\mathbf{g},$$

where β is the sliding hyper-parameter.

After finding the appropriate $\mathbf{p}$, the second term within the WOR formulation, namely, $\inf_{\mathbf{w}} R_{\mathrm{S}}(\mathbf{w};D_{X_{\mathrm{S}}})$, becomes a constant that can be omitted. Consequently, the learning objective for updating model parameters is established as follows:

$$\widehat{R}_{\mathrm{I}}(\mathbf{w}) + \alpha\widehat{R}_{\mathrm{A}}(\mathbf{w}+\lambda\mathbf{p}),$$

Algorithm 1: Wasserstein Distribution-agnostic Outlier Exposure.

Input: ID and OOD samples from $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $D_{X_{\mathrm{A}}}$, resp;
$\mathbf{p}$ initialized by $\mathbf{0}$;
for $\mathtt{ns} = 1$ **to** $\mathtt{num\_step}$ **do**
  Sampling $B_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $B_{X_{\mathrm{A}}}$ from $D_{X_{\mathrm{I}}Y_{\mathrm{I}}}$ and $D_{X_{\mathrm{A}}}$;
  if $\mathtt{ns} > \mathtt{num\_warm}$ **then**
    $\mathbf{g}_{\mathrm{OR}} = \nabla_{\mathbf{p}} \sum_{\mathbf{x}\in B_{X_{\mathrm{A}}}} \kappa(\mathbf{x}; \mathbf{w} + \lambda\mathbf{p})$;
    $\mathbf{g}_{\mathrm{WA}} = \nabla_{\mathbf{p}} \sum_{\mathbf{x}\in B_{X_{\mathrm{A}}}} ||\boldsymbol{\phi}(\mathbf{x}; \mathbf{w}) - \boldsymbol{\phi}(\mathbf{x}; \mathbf{w} + \lambda\mathbf{p})||^2$;
    $\mathbf{g} = \mathbf{g}_{\mathrm{OR}} - \gamma\mathbf{g}_{\mathrm{WA}}$;
    $\mathbf{p} \leftarrow (1-\beta)\mathbf{p} + \beta\mathbf{g}$;
  end if
  $\mathbf{w} \leftarrow \mathbf{w} - \mathtt{lr}\nabla_{\mathbf{w}}[\widehat{R}_{\mathrm{I}}(\mathbf{w}) + \alpha\widehat{R}_{\mathrm{A}}(\mathbf{w} + \lambda\mathbf{p})]$;
end for
Output: model parameters $\mathbf{w}$.

which can be optimized by mini-batch gradient descent.

Scoring Strategy: We adopt the MaxLogit scoring [24], i.e.,

$$s_{\mathrm{MLP}}(\mathbf{x}; \mathbf{w}) = \max_j \mathbf{f}^j(\mathbf{x}; \mathbf{w}), \tag{11}$$

to conduct OOD detection, which is more effective than MSP for many complex detection setups. We further test W-DOE with more advanced scoring strategies in the appendix, available online.

VI. Theoretical Analysis

We provide theoretical justifications for our W-DOE, certifying the validity of IDS in generating diverse data and the effectiveness of W-DOE under the OOD distribution discrepancy.

A. Implicit Data Synthesis

We reveal that model perturbation can implicitly lead to data transformation in the input space, where the implicit data follow a new data distribution compared with the original ones. To begin with, we consider the recursive definition of a K-layer network

$$\mathbf{z}^{(k+1)} = \boldsymbol{\sigma}(\mathbf{W}^{(k)}\mathbf{z}^{(k)}) \ \text{ for } k = 1, \ldots, K,$$

with $\mathbf{W}^{(k)}$ the k-th layer weights, $\boldsymbol{\sigma}(\mathbf{z})$ the activation function, and $\mathbf{z}^{(k)}$ the k-th layer outputs. We have $\mathbf{z}^{(1)} = \mathbf{x}$ the inputs and $\mathbf{z}^{(K)} = \mathbf{f}(\mathbf{x}; \mathbf{w})$ the outputs. Moreover, we consider the multiply forms of perturbation as in (4), which is more convenient than the additive form, i.e., $\mathbf{w} + \lambda\mathbf{p}$, in theoretical analysis.

Definition 5 (Multiply Perturbation): For model $\mathbf{f}(\cdot; \mathbf{w})$, its k-th layer is multiply perturbed if $\mathbf{W}^{(k)}$ is changed into

$$\mathbf{W}^{(k)}(\mathbf{I} + \lambda\mathbf{A}^{(k)}),$$

where $\alpha > 0$ is the perturbation strength and $\mathbf{A}^{(k)}$ is the perturbation matrix. Moreover, the whole model is multiply perturbed if all its layers undergo multiple perturbations.

Layer-level Perturbation: We first consider the layer-level perturbation as in Section IV, connecting multiply perturbation to data transformation in any layer of the model, summarized as follows.

Theorem 3: Assuming the multiply perturbation $\mathbf{W}^{(k)}(\mathbf{I} + \lambda\mathbf{A}^{(k)})$ in the k-th layer. Then, measuring in the k-th layer feature space, the multiply perturbation transforms the original distribution D to a new distribution D'. If the eigenvalues of $\mathbf{A}^{(k)}$ are all greater than 0, we further have $\mathtt{KL}(D, D') > 0$.

Proof: Based on (4), we know that the multiply perturbation $\mathbf{W}^{(k)}(\mathbf{I} + \lambda\mathbf{A}^{(k)})$ transforms the features $\mathbf{z}^{(k)}$ by an affine transformation $(\mathbf{I} + \lambda\mathbf{A}^{(k)})\mathbf{z}^{(k)}$. Define the k-th layer feature space as $\mathcal{Z}^{(k)}$, we assume that original data are i.i.d. drawn from the distribution with the associated density function of $f_{\mathcal{Z}^{(k)}}(\mathbf{z}^{(k)})$. Then, the transformed data $\mathbf{z}'^{(k)}$ are i.i.d. drawn from a new distribution with the probability density function of

$$f_{Z^{(k)}}(\mathbf{z}'^{(k)}) = f_{Z^{(k)}}(\mathbf{z}^{(k)})\left|\mathbf{I} + \lambda\mathbf{A}^{(k)}\right|^{-1}.$$

Moreover, when measuring by the KL divergence, we have

$$\mathtt{KL}(f_{Z^{(k)}}(\mathbf{z}^{(k)})||f_{Z^{(k)}}(\mathbf{z}'^{(k)})) = \log|\mathbf{I} + \lambda\mathbf{A}^{(k)}|.$$

By the Jordan matrix decomposition, we have $\mathbf{A}^{(k)} = \mathbf{T}^{(k),-1}\mathbf{J}^{(k)}\mathbf{T}^{(k)}$, where $\mathbf{J}^{(k)}$ is upper triangular. Assuming Γ different eigenvalues $\upsilon^{(r)}$ for the matrix $\mathbf{A}^{(k)}$, then we have

$$\begin{aligned}|\mathbf{I} + \lambda\mathbf{A}^{(k)}| &= |\mathbf{T}^{(k),-1}(\mathbf{I} + \lambda\mathbf{J}^{(k)})\mathbf{T}^{(k)}| \\ &= |\mathbf{I} + \lambda\mathbf{J}^{(k)}| = \prod_{r=1}^{\Gamma}(1 + \lambda\upsilon^{(r)})^{n_r}.\end{aligned}$$

where n_r is the number of eigenvectors with the eigenvalue $\upsilon^{(r)}$. Thus, if the eigenvalues of $\mathbf{A}^{(k)}$ are all greater than 0, we have $|\mathbf{I} + \lambda\mathbf{A}^{(k)}| > 1$ and thus $\mathtt{KL}(f_{Z^{(k)}}(\mathbf{z}^{(k)})||f_{Z^{(k)}}(\mathbf{z}'^{(k)})) > 0$. Therefore, the proof is completed. □

Model-level Perturbation: The above theorem can be generalized to multiply perturbation for the whole model, which leads to data transformation in the input space, illustrated as follows.

Theorem 4: Assuming the multiply perturbations $\{\mathbf{W}^{(k)}(\mathbf{I} + \lambda\mathbf{A}^{(k)})\}_{k=1}^{K}$ for the ReLU model. Then, measuring in the input space $\mathcal{X}$, the multiply perturbation transforms the original distribution D to a new distribution D'. If the eigenvalues of $\mathbf{A}^{(k)}$ are greater than 0 and $|\mathbf{W}^{(k)}| \neq 0$ for all k, then $\mathtt{KL}(D, D') > 0$.

Proof: We consider an induction proof, justifying that: the multiply perturbation with $\mathbf{A}^{(k)}$ can be transformed into an equivalent multiply perturbation $\bar{\mathbf{A}}^{(k-1)}$ in the $(k-1)$-th layer for all k. Moreover, if $|\mathbf{A}^{(k)}| > 0$ indicates $|\bar{\mathbf{A}}^{(k-1)}| > 0$, the determinant of the equivalent perturbation is greater than 0. Then, based on Theorem 3, we will complete the proof by induction. To find the equivalent perturbation matrix $\bar{\mathbf{A}}^{(k-1)}$ in the $(k-1)$-th layer for $\mathbf{A}^{(k)}$ in the l-th layer, we should solve:

$$\begin{aligned}&\mathbf{W}^{(k)}(\mathbf{I} + \lambda\mathbf{A}^{(k)})\boldsymbol{\sigma}(\mathbf{W}^{(k-1)}\mathbf{z}^{(k-1)}) \\ &= \mathbf{W}^{(k)}\boldsymbol{\sigma}\left(\mathbf{W}^{(k-1)}(\mathbf{I} + \lambda\bar{\mathbf{A}}^{(k-1)})\mathbf{z}^{(k-1)}\right).\end{aligned}$$

It can be further rewritten as

$$\begin{aligned}&\mathbf{A}^{(k)}\boldsymbol{\sigma}(\mathbf{W}^{(k-1)}\mathbf{z}^{(k-1)}) \\ &= \boldsymbol{\sigma}'(\mathbf{W}^{(k-1)}\mathbf{z}^{(k-1)})\mathbf{W}^{(k-1)}\bar{\mathbf{A}}^{(k-1)}\mathbf{z}^{(k-1)},\end{aligned}$$

where we apply the Taylor Theorem for the right-hand side, with the usual adjustments that the equations only hold almost

everywhere in parameter space. Then, since the ReLU activation is applied, we can approximate the above formulation as

$$\mathbf{A}^{(k)}\mathbf{W}^{(k-1)} = \mathbf{W}^{(k-1)}\bar{\mathbf{A}}^{(k-1)},$$

holding almost sure when outputs are not sparse. Furthermore,

$$\begin{aligned}|\mathbf{I}+\lambda\mathbf{A}^{(k)}||\mathbf{W}^{(k-1)}| &= |\mathbf{W}^{(k-1)}+\lambda\mathbf{A}^{(k)}\mathbf{W}^{(k-1)}| \\ &= |\mathbf{W}^{(k-1)}+\lambda\mathbf{W}^{(k-1)}\bar{\mathbf{A}}^{(k-1)}| \\ &= |\mathbf{W}^{(k-1)}||\mathbf{I}+\lambda\bar{\mathbf{A}}^{(k-1)}|.\end{aligned}$$

Since $|\mathbf{I}+\lambda\mathbf{A}^{(k)}| > 1$ and $|\mathbf{W}^{(k-1)}| \neq 0$, we can easily know that $|\mathbf{I}+\lambda\bar{\mathbf{A}}^{(k-1)}| > 1$ holds. We have shown that multiply perturbation in the k-th layer can be transformed to the $(k-1)$-th layer. Then, the equivalent perturbation $\bar{\mathbf{A}}^{(k-1)}$ and the original perturbation in the $(k-1)$-th layer, i.e., $\mathbf{A}^{(k-1)}$ can formulate a joint perturbation $\bar{\bar{\mathbf{A}}}^{(k-1)}$, namely, $\mathbf{I}+\lambda\bar{\bar{\mathbf{A}}}^{(k-1)}$, with

$$\bar{\bar{\mathbf{A}}}^{(k-1)} = \bar{\mathbf{A}}^{(k-1)} + \mathbf{A}^{(k-1)} + \lambda\mathbf{A}^{(k-1)}\bar{\mathbf{A}}^{(k-1)}. \tag{12}$$

Then, the joint perturbation $\bar{\bar{\mathbf{A}}}^{(k-1)}$ satisfies:

$$\begin{aligned}|\mathbf{I}+\lambda\bar{\bar{\mathbf{A}}}^{(k-1)}| &= |(\mathbf{I}+\lambda\mathbf{A}^{(k-1)})(\mathbf{I}+\lambda\bar{\mathbf{A}}^{(k-1)})| \\ &= |\mathbf{I}+\lambda\mathbf{A}^{(k-1)}||\mathbf{I}+\lambda\bar{\mathbf{A}}^{(k-1)}| \\ &> |\mathbf{I}+\lambda\mathbf{A}^{(k-1)}| > 1.\end{aligned} \tag{13}$$

Thus we complete our proof. □

Remark 6: Using KL divergence to justify that transformed data follow a new distribution may lead to a small theoretical gap for W-DOE. The reason is that W-DOE searches OOD distributions w.r.t. the Wasserstein distance. However, the Wasserstein distance cannot be simply lower-bounded by the KL divergence, thus $\mathtt{WD}(D, D') > 0$ does not hold in general. On the other side, KL divergence-based constraints are not as effective as that of the Wasserstein distance in many applications [34]. Therefore, as a trade-off, we use the KL divergence to facilitate our derivation and the Wasserstein measurement to pursue practical effectiveness. Moreover, these two metrics are intricately linked by Talagrand's inequality [35], asserting that the square of the Wasserstein distance is upper bounded by the corresponding KL divergence. Hence, we have the necessary conditions for assessing changes in distribution as measured by the Wasserstein distance.

Theorem 4 further leads to the proposition below, indicating that our IDS can benefit from layer-wise architectures of models.

Proposition 1: Consider a K-layer ReLU network with multiply perturbation $\{\mathbf{A}_K^{(k)}\}_{k=1}^{K}$ and transformed distribution D'_K. Then, there exists a $K+1$-layer ReLU network with multiply perturbation $\{\mathbf{A}_{K+1}^{(k)}\}_{k=1}^{K+1}$ and transformed distribution D'_{K+1}, such that $\mathtt{KL}(D, D'_{K+1}) \geq \mathtt{KL}(D, D'_K)$, measured in the input space.

Proof: For the $(K+1)$-layer ReLU network, we assume its model parameters and the model perturbation are the same as that of the corresponding layers for the K-layer ReLU network (except for the $K+1$-th layer). Then, by inspecting (13), the perturbation from the $K+1$-th layer can make the perturbation matrices for the $K+1$-layer network no smaller than that of the K-layer network regarding each layer (including the input space) of the joint multiplicative perturbation. Thus, we complete our proof. □

We summarize the above derivations as follows: In Theorem 4, we conclude that model perturbation can implicitly transform data that follow new distributions. In Proposition 1, we find that the transform function is complex enough with layer-wise non-linearity, where deeper models induce stronger transformations.

B. Distribution Robustness

We begin with standard assumptions that ease our discussions.

Assumption 1: Denote the model space by $\mathcal{F}$, we assume its Rademacher Complexity $\mathfrak{R}_n(\mathcal{F})$ is large enough yet upper bounded, i.e., there is a $C_{\mathcal{F}}$ such that $\mathfrak{R}_n(\mathcal{F}) \leq C_{\mathcal{F}}/\sqrt{n}$. Moreover, we assume the ID loss is bounded by A_{ID} and is L_{ID} Lipschitz continuous; the OOD loss is bounded by A_{OOD} and is L_{OOD} Lipschitz continuous.

Remark 7: The bounded Rademacher complexity is a standard assumption that holds for many deep models [36], [37]. We can justify that the cross entropy loss and the KL divergence can be bounded and Lipschitz with respect to $\mathbf{w}$ in practice if they satisfy

- the activation functions are Lipschitz (holding for ReLU);
- the input space $\mathcal{X}$ is bounded (holding for images);
- the parameter space is bounded (holding for regularization).

Specifically, when parameters are constrained by the F-norm, the softmax outputs remain continuous and are prevented from reaching infinity. Further assuming the bounded feature space, the model then becomes a continuous function therein. It implies that the function $\mathbf{f}(\cdot;\mathbf{w})$ is both upper and lower bounded.

We further introduce an useful lemma in the following.

Lemma 1: Given Assumption 1, then for any $\delta \geq 0$,

$$\begin{aligned}&\sup_{D_X \in \mathfrak{B}_{\text{A}}^{\rho+\delta}} \mathbb{E}_{\mathbf{x}\sim D_X} \ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) \\ &\qquad \leq \sup_{D_X \in \mathfrak{B}_{\text{A}}^{\rho}} \mathbb{E}_{\mathbf{x}\sim D_X} \ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) + L_{\text{OOD}}\delta.\end{aligned}$$

Proof: For any $\epsilon > 0$, we set $D_X^{\delta,\epsilon}$ satisfies that

$$\sup_{D_X \in \mathfrak{B}_{\text{A}}^{\rho+\delta}} \mathbb{E}_{\mathbf{x}\sim D_{X'}} \ell_{\text{OOD}}(\mathbf{x};\mathbf{f}_{\mathbf{w}}) \leq \mathbb{E}_{\mathbf{x}\sim D_X^{\delta,\epsilon}} \ell_{\text{OOD}}(\mathbf{x};\mathbf{f}_{\mathbf{w}}) + \epsilon,$$

and

$$\mathtt{WD}(D_X^{\delta,\epsilon}, D_{X_{\text{A}}}) \leq \rho + \delta.$$

If $\mathtt{WD}(D_X^{\delta,\epsilon}, D_{X_{\text{A}}}) \leq \rho$, then

$$\begin{aligned}&\sup_{D_X \in \mathfrak{B}_{\text{A}}^{\rho+\delta}} \mathbb{E}_{\mathbf{x}\sim D_X} \ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) \\ &\qquad \leq \sup_{D_X \in \mathfrak{B}_{\text{A}}^{\rho}} \mathbb{E}_{\mathbf{x}\sim D_X} \ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) + \epsilon.\end{aligned}$$

If $\mathtt{WD}(D_X^{\delta,\epsilon}, D_{X_{\text{A}}}) > \rho$, we introduce the distribution $D'_X = (1-u)D_X^{\delta,\epsilon} + uD_{X_{\text{A}}}$ for $u \in [0,1]$. Then, we have

$$\mathtt{WD}(D'_X, D_{X_{\text{A}}}) \leq (1-u)\mathtt{WD}(D_X^{\delta,\epsilon}, D_{X_{\text{A}}}) \leq (1-u)(\rho+\delta).$$

If $u = \delta/(\rho+\delta)$, then $\texttt{WD}(D'_X, D_{X_\text{A}}) \le \rho$ holds. By Kantorovich–Rubinstein duality [38], we have

$$\begin{aligned}
&\texttt{WD}(D'_X, D_X^{\delta,\epsilon}) \\
&= \sup_{\|f\|_{\text{Lip}}\le 1} \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}D'_X(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}D_X^{\delta,\epsilon}(\mathbf{x}) \\
&= u \sup_{\|f\|_{\text{Lip}}\le 1} \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}D_{X_\text{A}}(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}D_X^{\delta,\epsilon}(\mathbf{x}) \\
&= u\texttt{WD}(D_{X_\text{A}}, D_X^{\delta,\epsilon}) = \delta.
\end{aligned}$$

We also obtain that

$$\begin{aligned}
&\sup_{D_X\in\mathfrak{B}_\text{A}^{\rho+\delta}} \mathbb{E}_{\mathbf{x}\sim D_X}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) \\
&- \sup_{D_X\in\mathfrak{B}_\text{A}^{\rho}} \mathbb{E}_{\mathbf{x}\sim D_X}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) \\
&\le \mathbb{E}_{\mathbf{x}\sim D_X^{\delta,\epsilon}}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) - \mathbb{E}_{\mathbf{x}\sim D'_X}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) + \epsilon \\
&\le L_{\text{OOD}}\delta + \epsilon,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\sup_{D_X\in\mathfrak{B}_\text{A}^{\rho+\delta}} \mathbb{E}_{\mathbf{x}\sim D_X}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) \\
&\qquad \le \sup_{D_X\in\mathfrak{B}_\text{A}^{\rho}} \mathbb{E}_{\mathbf{x}\sim D_X}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) + L_{\text{OOD}}\delta.
\end{aligned}$$

Combining the cases with $\texttt{WD}(D_{X'}^{\delta,\epsilon}, D_{X_\text{A}}) \le \rho$ and $\texttt{WD}(D_{X'}^{\delta,\epsilon}, D_{X_\text{A}}) > \rho$, we complete the proof. □

OOD Distribution Discrepancy for OE: As a direct application of Lemma 1, we can prove Theorem 1. Assuming

$$\epsilon(C, L, A) = 2CL + A\sqrt{0.5\log 1/\delta}, \tag{14}$$

we have the following derivation.

Proof: By the Rademacher Bound, with the probability at least $1-\delta$, we can prove that

$$|R_\text{I}(\mathbf{w}) - \widehat{R}_\text{I}(\mathbf{w})| \le \epsilon(C_\mathcal{F}, L_\text{ID}, A_\text{ID})/\sqrt{n},$$

and

$$|R_\text{A}(\mathbf{w}) - \widehat{R}_\text{A}(\mathbf{w})| \le \epsilon(C_\mathcal{F}, L_\text{OOD}, A_\text{OOD})/\sqrt{m},$$

for all $\mathbf{w}$. Denote $\mathbf{w}^* = \inf_\mathbf{w} R_\text{I}(\mathbf{w})$ and $\widehat{\mathbf{w}} = \inf_\mathbf{w} \widehat{R}_\text{I}(\mathbf{w})$, then, for any $\epsilon > 0$, there exists $\mathbf{w}^\epsilon$ such that $R_\text{I}(\mathbf{w}^\epsilon) \le R_\text{I}(\mathbf{w}^*) + \epsilon$. Thus, using the fact that $\widehat{R}_\text{I}(\widehat{\mathbf{w}}) \le \widehat{R}_\text{I}(\mathbf{w}^\epsilon)$, we have

$$\begin{aligned}
&R_\text{I}(\widehat{\mathbf{w}}) - R_\text{I}(\mathbf{w}^*) \\
&= R_\text{I}(\widehat{\mathbf{w}}) - R_\text{I}(\mathbf{w}^\epsilon) + R_\text{I}(\mathbf{w}^\epsilon) - R_\text{I}(\mathbf{w}^*) \\
&\le R_\text{I}(\widehat{\mathbf{w}}) - R_\text{I}(\mathbf{w}^\epsilon) + \epsilon \\
&= R_\text{I}(\widehat{\mathbf{w}}) - \widehat{R}_\text{I}(\widehat{\mathbf{w}}) + \widehat{R}_\text{I}(\widehat{\mathbf{w}}) - R_\text{I}(\mathbf{w}^\epsilon) + \epsilon \\
&\le R_\text{I}(\widehat{\mathbf{w}}) - \widehat{R}_\text{I}(\widehat{\mathbf{w}}) + \widehat{R}_\text{I}(\mathbf{w}^\epsilon) - R_\text{I}(\mathbf{w}^\epsilon) + \epsilon \\
&\le 2\sup_\mathbf{w}|R_\text{I}(\mathbf{w}) - \widehat{R}_\text{I}(\mathbf{w})| + \epsilon,
\end{aligned} \tag{15}$$

holding for any $\mathbf{w}$ and $\epsilon > 0$. Accordingly, we have,

$$R_\text{I}(\widehat{\mathbf{w}}) \le R_\text{I}(\mathbf{w}^*) + 2\epsilon(C_\mathcal{F}, L_\text{ID}, A_\text{ID})/\sqrt{n}. \tag{16}$$

Similarly, we can derive

$$R_\text{A}(\widehat{\mathbf{w}}) \le R_\text{A}(\mathbf{w}^*) + 2\epsilon(C_\mathcal{F}, L_\text{OOD}, A_\text{OOD})/\sqrt{m}.$$

and thus,

$$R_\text{OE}(\widehat{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} R_\text{OE}(\mathbf{w}) \le (1+\alpha)\frac{2\epsilon(C_\mathcal{F}, L, A)}{\sqrt{\min\{m,n\}}}, \tag{17}$$

where $L = \max\{L_\text{ID}, L_\text{OOD}\}$ and $A = \max\{A_\text{ID}, A_\text{OOD}\}$.

Moreover, based on Lemma 1, we know that

$$\begin{aligned}
&R_\text{A}(\widehat{\mathbf{w}}) - R_\text{O}(\widehat{\mathbf{w}}) \\
&\le \sup_{D_{X_\text{S}}\in\mathfrak{B}_\text{A}^{\delta}} \mathbb{E}_{\mathbf{x}\sim D_{X_\text{S}}}\ell_{\text{OOD}}\left(\mathbf{f}(\mathbf{x};\mathbf{w})\right) - R_\text{O}(\widehat{\mathbf{w}}) \le L\delta.
\end{aligned} \tag{18}$$

where $\delta = \texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$ Substituting (18) into (17), we complete the proof.

OOD Distribution Discrepancy for W-DOE: Now, we study the effectiveness of our W-DOE in mitigating the OOD distribution discrepancy, summarized by the following theorem.

Theorem 5: Given Assumption 1 and $\widehat{\mathbf{w}} = \arg\min_\mathbf{w} \widehat{R}_\text{W}(\mathbf{w};\rho)$, then with the probability at least $1-\delta$, we have

$$\begin{aligned}
&R_\text{I}(\widehat{\mathbf{w}}) + \alpha R_\text{O}(\widehat{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} R_\text{W}(\mathbf{w};\rho) \\
&\le \alpha L\ \max\left\{\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) - \rho, 0\right\} \\
&+ (2+4\alpha)\frac{\epsilon(C_\mathcal{F}, L, A)}{\sqrt{\min\{m,n\}}} + \inf_\mathbf{w} R_\text{O}(\mathbf{w}),
\end{aligned} \tag{19}$$

where $L = \max\{L_\text{ID}, L_\text{OOD}\}$, $A = \max\{A_\text{ID}, A_\text{OOD}\}$, and $\epsilon(C_\mathcal{F}, L, A)$ is defined as in (14).

Proof: For any $\mathbf{w}$, with the probability at least $1-\delta$,

$$\begin{aligned}
&\texttt{WOR}(\widehat{\mathbf{w}};\rho) \\
&= \sup_{D_{X_\text{S}}\in\mathfrak{B}_\text{A}^{\rho}} R_\text{S}(\widehat{\mathbf{w}}; D_{X_\text{S}}) - R_\text{S}(\mathbf{w}^*; D_{X_\text{S}}) \\
&\le \sup_{D_{X_\text{S}}\in\mathfrak{B}_\text{A}^{\rho}} \widehat{R}_\text{S}(\widehat{\mathbf{w}}; D_{X_\text{S}}) - \widehat{R}_\text{S}(\mathbf{w}^*; D_{X_\text{S}}) + \frac{2\epsilon(C_\mathcal{F}, L, A)}{\sqrt{m}} \\
&\le \sup_{D_{X_\text{S}}\in\mathfrak{B}_\text{A}^{\rho}} \widehat{R}_\text{S}(\mathbf{w}; D_{X_\text{S}}) - \widehat{R}_\text{S}(\mathbf{w}^*; D_{X_\text{S}}) + \frac{2\epsilon(C_\mathcal{F}, L, A)}{\sqrt{m}} \\
&\le \sup_{D_{X_\text{S}}\in\mathfrak{B}_\text{A}^{\rho}} R_\text{S}(\mathbf{w}; D_{X_\text{S}}) - R_\text{S}(\mathbf{w}^*; D_{X_\text{S}}) + \frac{4\epsilon(C_\mathcal{F}, L, A)}{\sqrt{m}} \\
&\le \texttt{WOR}(\mathbf{w};\rho) + \frac{4\epsilon(C_\mathcal{F}, L, A)}{\sqrt{m}}
\end{aligned}$$

where $\mathbf{w}^* = \inf_\mathbf{w} R_\text{S}(\mathbf{w}; D_{X_\text{S}})$. Further combining with (16), one can easily derive

$$R_\text{W}(\widehat{\mathbf{w}};\rho) \le \min_{\mathbf{w}\in\mathcal{W}} R_\text{W}(\mathbf{w};\rho) + (2+4\alpha)\frac{\epsilon(C_\mathcal{F}, L, A)}{\sqrt{\min\{m,n\}}}. \tag{20}$$

Then, if $\rho \geq \texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$, we have

$$R_\text{O}(\widehat{\mathbf{w}}) - R_\text{O}(\mathbf{w}^*) - \texttt{WOR}(\widehat{\mathbf{w}}; \rho) \leq 0.$$

Otherwise, by Lemma 1, we have

$$\begin{aligned} &R_\text{O}(\widehat{\mathbf{w}}) - \inf_{\mathbf{w}} R_\text{O}(\mathbf{w}) - \texttt{WOR}(\widehat{\mathbf{w}}; \rho) \\ &\leq L\left(\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) - \rho\right). \end{aligned}$$

Thus, we have

$$\begin{aligned} &R_\text{O}(\widehat{\mathbf{w}}) - \texttt{WOR}(\widehat{\mathbf{w}}; \rho) \\ &\leq L \max\left\{\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) - \rho, 0\right\} + \inf_{\mathbf{w}} R_\text{O}(\mathbf{w}). \end{aligned}$$

Combining with (20), we complete the proof. □

Remark 8: In (19), the first term in the right-hand-side plays the key role, in that larger ρ leads to tighter bounds in the open world. Thus, it verifies that our W-DOE can lead to improved performance when facing OOD distribution discrepancy. Moreover, $\inf_{\mathbf{w}} R_\text{O}(\mathbf{w})$ is typically a negligible constant, especially for deep models with relatively large capacities. Then, we can easily observe that our bound is much tighter than that of OE in (1) demonstrating the superiority of our W-DOE over previous works.

VII. Experiments

Besides theoretical analysis, we provide empirical evaluations under a series of representative OOD detection setups.

- *Simulated OOD Detection:* We begin with illustrative experiments, visualizing the impacts of the OOD distribution discrepancy and the worst-case learning, respectively.
- *Common OOD Detection:* We present our main experiments on the CIFAR benchmarks [39], revealing the superiority of W-DOE over its many advanced counterparts.
- *Challenging OOD Detection:* We then test W-DOE for more challenging setups with wild [40] and hard [12] OOD data. We also test on ImageNet-1K [41], aligning with real-world cases with large label spaces and complex feature patterns.
- *Medical OOD Detection:* To substantiate the real-world applicability of our method, we test on the CheXpert X-ray Lung Pathology dataset [42] for medical OOD detection,

In our supplementary materials, we further conduct ablation studies and present hyper-parameter analysis, aiming to study the respective powers of our algorithmic designs.

A. Evaluation Setups

We first present the default setups used in empirical evaluations.

Baseline Methods: We compare our W-DOE with advanced fine-tuning methods, including OE [19], ATOM [17], POEM [21], DAL [32], and WOOD [40]. In the supplementary materials, we further consider post-hoc and representation-based methods [14], [43], [44], [45]. We adopt their default setups while unifying backbone models and auxiliary OOD datasets (if used) for fairness.

Pre-training Setups: On CIFAR benchmarks, we employ Wide ResNet-40-2 [46] trained for 200 epochs based on empirical risk minimization, with batch size 64, momentum 0.9, and learning rate 0.1. The learning rate is divided by 10 after 100 and 150 epochs. On the ImageNet-1 K, we employ ResNet-50 [47] with the pre-trained parameters released by the PyTorch official.

OOD Datasets: On CIFAR benchmarks, we adopt Tiny-ImageNet-200 [48] as the auxiliary OOD dataset; Textures [49], SVHN [50], Places365 [51], LSUN-Crop [52], LSUN-Resize [52], and iSUN [53] as the common OOD datasets; Oxford-Pets [54], ImageNet-Fix [41], and ImageNet-Resize [41] as the hard OOD datasets. On the ImageNet-1 K benchmark, we adopt ImageNet-21K-P [55] as the auxiliary OOD dataset; iNaturalist [56], SUN [53], Places365 [51], and Texture [49] as the test OOD datasets. Data that coincide with ID semantics are removed.

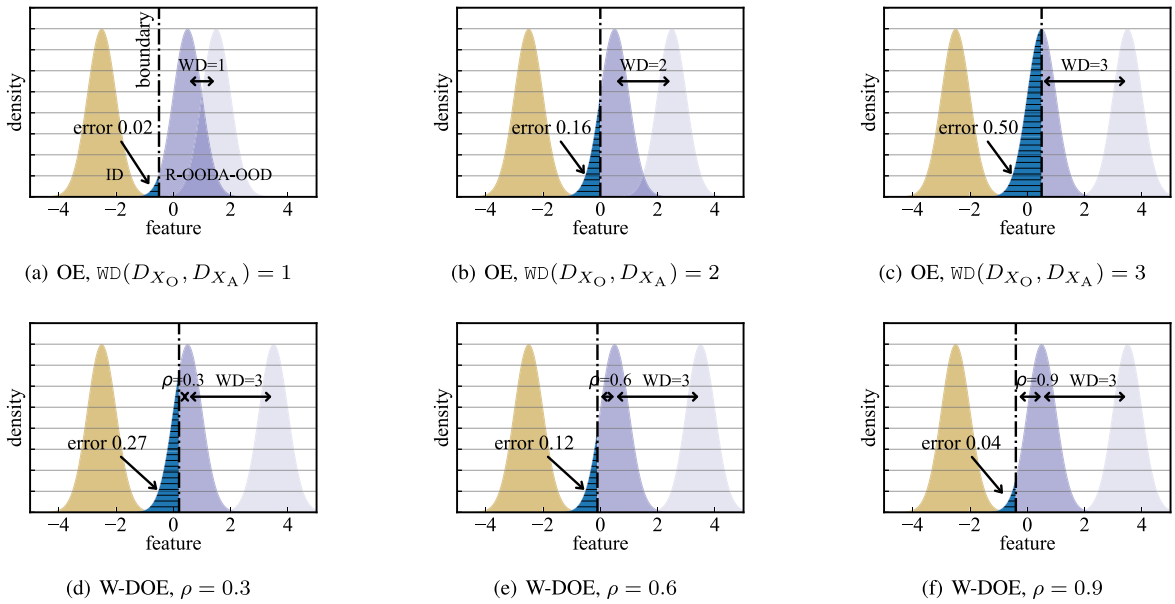
Evaluation Metrics: The detection performance is measured by two threshold-independent metrics: The false positive rate of OOD data when the true positive rate is 95% (FPR95); the area under the receiver operating characteristic curve (AUROC). Higher AUROC and lower FPR95 are preferred in OOD detection.

W-DOE Default Setups: Hyper-parameters in W-DOE are tuned by grid search concerning validation data, which are separated from ID and auxiliary OOD data. Such a tuning setup is a common practice [17], [19]. For CIFAR benchmarks, W-DOE is run for 15 epochs with an initial learning rate of 0.005 and cosine decay [57]. The batch size is 128 for ID and 256 for OOD. $\texttt{num\_warm} = 10$, $\beta = 0.6$, $\gamma = 4$. For the ImageNet-1K, W-DOE is run for 6 epochs with the learning rate 0.0001. The batch sizes are 64 for both ID and OOD. $\texttt{num\_warm} = 6$, $\beta = 0.1$, and $\gamma = 5$. For both CIFAR and ImageNet-1K benchmarks, λ is uniformly sampled from $\{0.1, 0.01, 0.001, 0.0001\}$ in each training step, allowing us to cover wider OOD cases meanwhile easing its tuning procedures.

B. Simulated OOD Detection

We commence by visualizing the impact of OOD discrepancies on model performance through simulated experiments. Specifically, we consider Gaussian features with the mean of -2.5 for the ID distribution and another set of Gaussian features with the mean of 0.5 as the real OOD distribution. Additionally, we assume various Gaussian distributions for the auxiliary OOD distributions, with means of 1.5, 2.5, and 3.5 as examples, to assess their different impacts. All their standard deviations are fixed at 0.5. Data are sampled within the 99.7% confidence intervals of their respective Gaussian distributions to ensure separability. The 0-1 loss is employed as the objective to identify the optimal boundary that can separate between ID and OOD cases.

It is simple to derive that when two Gaussian distributions have the same variances, their Wasserstein distance is the distance between their mean values. We can also calculate the optimal boundary between the ID and the auxiliary OOD data as the average of their means. The 0-1 loss for real OOD data can be calculated using the cumulative distribution function of



(a) OE, $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) = 1$ (b) OE, $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) = 2$ (c) OE, $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}}) = 3$

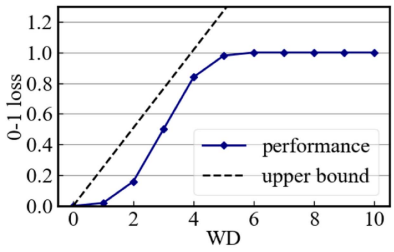
(d) W-DOE, $\rho = 0.3$ (e) W-DOE, $\rho = 0.6$ (f) W-DOE, $\rho = 0.9$

Fig. 2. The simulated experiments with one-dimensional Gaussian features. ID, R-OOD, and A-OOD represent the ID, real OOD, and auxiliary OOD distributions; the boundary is the decision boundary in discerning ID and OOD patterns; the WD denotes the Wasserstein distance between real and auxiliary OOD distributions; the error represents the 0-1 loss with respect to the unseen real OOD distribution.

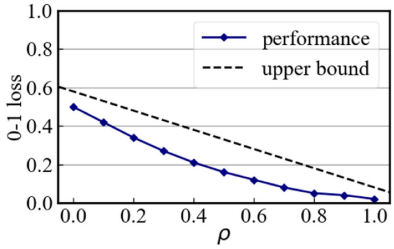
the Gaussian distribution representing the real OOD data. The results are summarized in Fig. 2(a)–(c). As observed, as the OOD distribution discrepancy increases, the identified boundaries for OOD detection make more mistakes and have larger errors.

We further illustrate decision boundaries searched with worst-case learning in Fig. 2(d)–(f), where we fix the mean of auxiliary OOD distribution to be 3.5 with the Wasserstein distance of 3. The intensity of worst-case learning is controlled by the radius of the Wasserstein ball, ρ, as prescribed by (5). Since the task is separable, $\inf_{\mathbf{w}} R_\text{S}(\mathbf{w}; D_{X_\text{S}})$ tends to be close to 0 when $\texttt{WD}(D_{X_\text{S}}, D_{X_\text{I}})$ is reasonably large. Then, we can explore varying ρ with its values being 0.3, 0.6, and 0.9, and deduce that the robust boundaries will adjust towards the ID distribution with a magnitude of ρ. As we can see, through worst-case data search, we ensure the detection performance of the model on unseen OOD cases, with larger ρ typically indicating stronger guarantees.

To echo our theoretical analysis for the learning guarantees of OE and W-DOE, additional simulated experiments are conducted across varying values of $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$ and ρ. The results, which highlight how these variables influence the overall detection performance as measured by the 0-1 loss, are detailed in Fig. 3. We also plot the crafted linear bounds for both cases, supporting our theoretical observations in Theorems 1 and 5 for the positive linear upper bound of OE with respect to $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$ and the negative linear upper bound of W-DOE for ρ.



(a) OE across varying Wasserstein distances



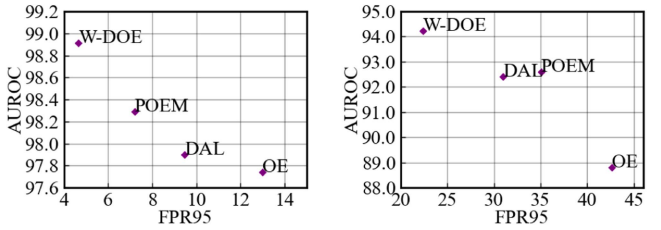
(b) W-DOE cross varying radius ρ

Fig. 3. The simulated experiments for OE and W-DOE with respect to different setups of $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$ and ρ. For (b), we fix the value of $\texttt{WD}(D_{X_\text{O}}, D_{X_\text{A}})$ to be 3 yet testing varying radius ρ.

C. Common OOD Detection

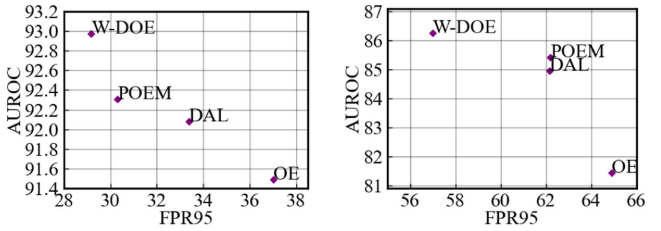
We provide the average results across various test-time OOD datasets on the CIFAR benchmarks, alongside some representative baseline methods, in Fig. 4. We summarize the detection performance across test OOD datasets, which have distribution discrepancies over auxiliary data. Please refer to the appendix, available online, for detailed detection results with more baseline methods.

The OOD distribution discrepancy naturally exists, making OE-based methods remain vulnerable. Fortunately, our W-DOE, which enlarges the OOD coverage during training, can largely improve the OOD performance. Compared with conventional



(a) CIFAR-10 (b) CIFAR-100

Fig. 4. Comparison between W-DOE and advanced methods on the CIFAR benchmarks. Please refer to the appendix, available online, for more results.



(a) CIFAR-10 (b) CIFAR-100

Fig. 5. Comparison between W-DOE and advanced methods on hard OOD detection. Please refer to the appendix, available online, for more results.

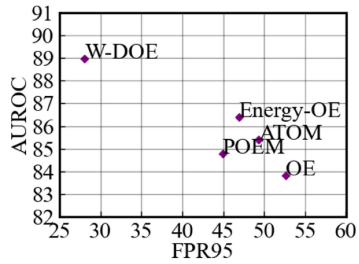
OE, our method reveals 8.33-20.27 average FPR95 improvements. For other works that similarly include OOD sampling, e.g., POEM, W-DOE also achieves better performance. The primary reason for the limited effectiveness of these methods is that they mainly focus on cases where model capacity is insufficient. It deviates from our considered situations for the discrepancy between training- and test-time OOD distributions. Finally, compared with existing methods that employ similar schemes of min-max learning, e.g., DAL, our W-DOE still exhibits superior performance, demonstrating our superiority in tackling the OOD distribution discrepancy.

D. Challenging OOD Detection

We consider two more challenging detection tasks on CIFAR benchmarks, evaluating W-DOE beyond the common setups.

Hard OOD detection: We consider a set of challenging test OOD datasets that possess different semantics yet similar styles over ID data. It reflects the reliability of models under very hard scenarios, thus drawing more practical interest. We present the results in Fig. 5. As we can see, W-DOE can overall beat previous methods, especially powerful in the CIFAR-100 case. The reason for the superiority of W-DOE stems from our worst-case learning scheme, which can cover some hard OOD data during training. Please refer to the appendix for detailed results, available online.

Wild OOD detection: Another interesting setup is wild OOD detection, where auxiliary OOD data consist of a portion of ID data. It is a practical setup since one cannot guarantee that the collected OOD data do not contain any ID semantic. As demonstrated in previous works [40], such wild OOD data may mislead our models in discerning ID and OOD patterns, thus detrimental to effective OOD detection. Here, we consider



(a) ImageNet

Fig. 6. Comparison between W-DOE and advanced methods on the ImageNet benchmarks.

different portions of ID data that are mixed with auxiliary OOD data, varying from 0.1 to 0.5, and summarize the experimental results in Table II. Here, we compare W-DOE with the original OE and the seminal work WOOD [40] that specifically tackles wild OOD detection, and scoring strategies are unified as MaxLogit for fairness. As revealed, the inherent noise in wild OOD data affects detection. However, W-DOE can mitigate their negative impacts when ID portions are relatively small (0.1 and 0.2), even better than WOOD that are specifically designed for wild OOD detection. Moreover, our W-DOE can still beat OE when ID portions further grow, since model updating with parameter noise, as in IDS, is a general strategy that can improve model robustness [58].

Real-world OOD detection: Nevertheless, we consider the experimental evaluations on the ImageNet-1 K benchmark, which is not fully covered in previous OE-based methods while critical for real-world evaluations. It has a large semantic space of 1,000 ID classes and the image patterns therein are complex and diverse, making it a challenging yet attractive OOD detection task. We summarize the experimental results in Fig. 6. As we can see, our W-DOE also reveals the best performance over the baselines, aligning with the CIFAR cases. It demonstrates that our W-DOE works still well in realistic detection setups. Please also refer to the appendix, available online, for the detailed results and more baselines.

E. Medical OOD Detection

Following [59], we utilize subsets from the X-ray Lung Pathology dataset to facilitate our evaluations of medical OOD detection. Our experiments consider two settings in the following:

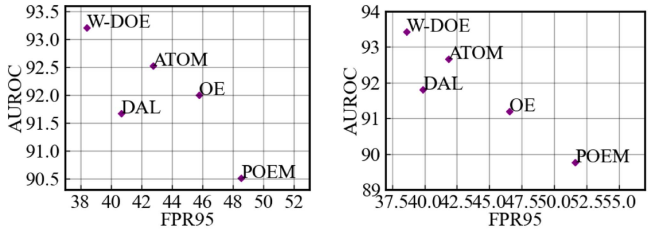
- *Setting 1:* Cardiomegaly and Pneumothorax are the ID classes and Fracture is related to the OOD class.
- *Setting 2:* Lung Opacity and Pleural Effusion are the ID classes, and Fracture and Pneumonia are the OOD classes.

For both of the above settings, we integrate data from other classes in the full dataset to construct the auxiliary OOD data. We choose ResNet-50 as our backbone model and use the hyper-parameter configurations as for the CIFAR benchmarks. We summarize our results in Fig. 7. Notably, even though the number of ID classes is relatively small, we observe the overall performance for most of the adopted methods is relatively low. It indicates that the setup is quite challenging. However, we still

TABLE II
COMPARISON BETWEEN W-DOE AND RELATED METHODS ON WILD OOD DETECTION

Method	0.1		0.2		0.3		0.4		0.5	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10										
OE	40.51	90.60	43.22	89.92	47.58	87.46	50.71	86.81	56.51	84.50
WOOD	36.57	91.95	36.54	92.17	**37.65**	**92.48**	**43.10**	**91.51**	**47.85**	**90.01**
W-DOE	**30.82**	**93.22**	**34.82**	**92.29**	42.16	90.10	51.81	86.48	60.87	84.32
CIFAR-100										
OE	9.13	97.20	9.52	97.19	9.48	97.32	11.35	97.16	14.73	96.74
WOOD	6.74	98.44	6.87	98.35	**7.43**	**98.36**	**7.91**	**98.26**	**8.31**	**97.97**
W-DOE	**6.03**	**98.62**	**6.56**	**98.55**	9.17	98.17	10.23	98.03	14.00	97.03

↓ (or ↑) indicates smaller (or larger) values are preferred, and a bold font indicates the best result in a column.



(a) Setting 1 (b) Setting 2

Fig. 7. Comparison between W-DOE and several advanced methods, focusing on the medical applications within the X-ray Lung Pathology dataset.

observe the superiority of W-DOE, revealing its reliability and real-world utility.

F. Performance Stability

Compared with our conference version of the algorithm design, named DOE, we further introduce the Wasserstein constraint to restrict the IDS searching region. It facilitates our theoretical analysis in Section VI, and we further claim that it can make the training procedure more stable and effective. Generally, W-DOE will degenerate to DOE when we assume ρ is very large in Theorem 5. At first glance, it seems that the first term in (19) will approach 0 and the upper bound will be tighter than that with small ρ. However, we notice that larger ρ also indicates larger $\min_{\mathbf{w}\in\mathcal{W}} R_{\mathrm{W}}(\mathbf{w};\rho)$, mainly due to the limited fitting power of $\mathbf{f}_{\mathbf{w}}$. Thus, the learning bound of W-DOE will be tighter than DOE, thus indicating better performance of our W-DOE.

As empirical justifications, we conduct five individual trials for OE, DOE, and W-DOE with random seeds of $\{1, 2, 3, 4, 5\}$, summarizing the results in Table III. Note that, for a fair comparison, we use the MaxLogit scoring after OE training, aligning with the cases of DOE and W-DOE. As we can see, DOE and W-DOE improve OOD performance over OE, while they are both less stable. However, with our newly introduced Wasserstein constraint, the training dynamics of W-DOE can be largely improved, with improved stability and further improved performance. We further consider various backbone models, including Wide ResNet-40-2 (WRN-40-2), ResNet-50 (RN-50), and DenseNet-101 [60] (DN-101), comparing the associated performance for OE, DOE, and W-DOE. The results in Table IV also reveal the stable improvement of our W-DOE method over OE and DOE, demonstrating the general applicability of our enhanced learning framework.

TABLE III
OE, DOE, AND W-DOE WITH 5 INDIVIDUAL TRAILS, MAXLOGIT SCORING IS USED FOR ALL METHODS

#	OE-MaxLogit		DOE		W-DOE	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
CIFAR-10						
1	5.81	98.47	5.63	98.71	4.56	98.83
2	6.14	98.49	5.87	98.54	4.50	98.83
3	6.05	98.50	5.36	98.79	4.86	98.95
4	6.11	98.47	5.88	98.58	4.67	98.80
5	5.97	98.53	5.28	98.78	5.17	98.71
mean	6.02	98.49	5.60	98.68	4.77	98.82
±std	±0.13	±0.03	±0.28	±0.12	±0.24	±0.09
CIFAR-100						
1	35.34	93.12	26.53	93.50	24.86	94.24
2	34.95	93.18	25.97	93.71	23.31	94.51
3	35.41	93.12	27.55	93.75	22.55	94.13
4	34.86	93.16	25.11	94.20	22.90	94.60
5	35.28	93.11	27.23	93.56	24.11	94.39
mean	35.17	93.14	26.48	93.74	23.55	94.37
±std	±0.25	±0.03	±0.98	±0.28	±0.94	±0.19

TABLE IV
OE, DOE, AND W-DOE WITH DIFFERENT BACKBONE ARCHITECTURES, MAXLOGIT SCORING IS USED FOR ALL METHODS

backbone	OE-MaxLogit		DOE		W-DOE	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
CIFAR-10						
WRN-40-2	5.81	98.47	5.63	98.71	**4.56**	**98.83**
RN-50	6.28	98.33	5.60	98.74	**4.78**	**98.85**
DN-101	5.60	98.44	6.37	98.57	**4.20**	**99.00**
CIFAR-100						
WRN-40-2	35.34	93.12	26.53	93.50	**24.86**	**94.24**
RN-50	38.68	92.51	29.60	93.17	**25.60**	**93.70**
DN-101	30.58	93.51	25.72	93.88	**25.13**	**94.20**

VIII. CONCLUSION

OE-based methods are powerful for OOD detection, while the issue of OOD distribution discrepancy hinders its open-world reliability. We study its negative impacts by modeling the OOD distribution discrepancy via the Wasserstein distance and propose a general learning framework, named W-DOE, that can mitigate its effects. Overall, the power of W-DOE in OOD detection is mainly attributed to two factors. First, we propose

IDS for OOD synthesis, based on the connection between model perturbation and input transformation. Synthetic data follow a diverse distribution compared to original ones, rendering models to learn from unseen data. Second, we suggest a min-max optimization scheme in searching for the worst-case regret, which can demonstrate better results than the risk-based counterpart. Our suggested learning scheme leads to certifiable OOD performance in the open world. We provide both theoretical justifications and empirical supports, demonstrating the reliability of W-DOE towards effective OOD detection. However, as in the supplementary materials, the optimal performance of our method depends on the proper selection of hyper-parameters, and we will explore AutoML and meta-learning methods in the future to ease the computation. Moreover, the proposed techniques in W-DOE, e.g., WOR and IDS, may contribute beyond OOD detection, and we will explore their usage scenarios in OOD generalization, adversarial training, and robust optimization.

REFERENCES

[1] S. Bulusu, B. Kailkhura, B. Li, P. Varshney, and D. Song, "Anomalous instance detection in deep learning: A survey," in *Proc. 42nd IEEE Symp. Secur. Privacy*, 2020.

[2] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: Learning what you don't know by virtual outlier synthesis," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=TW7d65uYu5M

[3] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," 2021, *arXiv:2110.11334*.

[4] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022.

[5] B. Barz, E. Rodner, Y. G. Garcia, and J. Denzler, "Detecting regions of maximal divergence for spatio-temporal anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1088–1101, May 2019.

[6] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl

[7] S. Wang et al., "$E^3$3Outlier: A self-supervised framework for unsupervised deep outlier detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2952–2969, Mar. 2023.

[8] S. Liang, Y. Li, and R. Srikant,, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1VGkIxRZ

[9] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7167–7177.

[10] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1802.

[11] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-distribution detection with rectified activations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, Art. no. 12.

[12] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 20827–20840.

[13] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ryiAv2xAZ

[14] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty detection via contrastive learning on distributionally shifted instances," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 11839–11852.

[15] S. Mohseni, M. Pitale, J. B. S. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5216–5223.

[16] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=v5gjXpmR8J

[17] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "ATOM: Robustifying out-of-distribution detection using outlier mining," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2021, pp. 430–445.

[18] J. Bitterwolf, A. Meinke, M. Augustin, and M. Hein, "Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 2041–2074.

[19] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HyxCxhRcY7

[20] J. Zhang, N. Inkawhich, R. Linderman, Y. Chen, and H. Li, "Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments," 2021, *arXiv:2106.03917*.

[21] Y. Ming, Y. Fan, and Y. Li, "POEM: Out-of-distribution detection with posterior sampling," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 15650–15665.

[22] Q. Wang et al., "Out-of-distribution detection with implicit outlier transformation," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=hdghx6wbGuD

[23] Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, and F. Liu, "Is out-of-distribution detection learnable?," 2022, *arXiv:2210.14707*.

[24] D. Hendrycks et al., "Scaling out-of-distribution detection for real-world settings," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 8759–8773.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[26] Y. Wang, W. Sun, J. Jin, Z. J. Kong, and X. Yue, "WOOD: Wasserstein-based out-of-distribution detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 944–956, Feb. 2024.

[27] Q. Wang et al., "Probabilistic margins for instance reweighting in adversarial training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, Art. no. 1781.

[28] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[29] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv: 1907.02893*.

[30] A. Agarwal and T. Zhang, "Minimax regret optimization for robust machine learning under distribution shift," in *Proc. 35th Conf. Learn. Theory*, 2022, pp. 2704–2729.

[31] J. H. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Math. Oper. Res.*, vol. 44, no. 2, pp. 565–600, 2019.

[32] Q. Wang, Z. Fang, Y. Zhang, F. Liu, Y. Li, and B. Han, "Learning to augment distributions for out-of-distribution detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, Art. no. 3203.

[33] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *J. Mach. Learn. Res.*, vol. 19, no. 13, pp. 1–48, 2018.

[34] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," 2019, *arXiv: 1908.05659*.

[35] N. Alon and J. H. Spencer, *The Probabilistic Method*. Hoboken, NJ, USA: Wiley, 2016.

[36] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in *Proc. Conf. Learn. Theory*, 2018, pp. 297–299.

[37] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.

[38] C. Villani, *Topics in Optimal Transportation*. Providence, RI, USA: Amer. Math. Soc., 2021.

[39] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.

[40] J. Katz-Samuels, J. B. Nakhleh, R. Nowak, and Y. Li, "Training OOD detectors in their natural habitats," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 10848–10865.

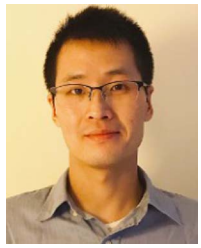
[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[42] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597.

[43] A. Djurisic, N. Bozanic, A. Ashok, and R. Liu, "Extremely simple activation shaping for out-of-distribution detection," 2022, *arXiv:2209.09858*.

[44] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 677–689.
[45] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb
[46] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 87.1–87.12. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[48] Y. Le and X. Yang, "Tiny ImageBet visual recognition challenge," *CS 231N*, vol. 7, no. 7, 2015, Art. no. 3.
[49] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.
[50] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NeurIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, Art. no. 4.
[51] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[52] 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*.
[53] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "TurkerGaze: Crowdsourcing saliency with webcam based eye tracking," 2015, *arXiv:1504.06755*.
[54] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3498–3505.
[55] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21k pretraining for the masses," 2021, *arXiv:2104.10972*.
[56] G. V. Horn et al., "The iNaturalist species classification and detection dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769–8778.
[57] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Skq89Scxx
[58] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6Tm1mposlrM
[59] C. Berger, M. Paschali, B. Glocker, and K. Kamnitsas, "Confidence-based out-of-distribution detection: A comparative study and analysis," in *Proc. Int. Workshop Uncertainty Safe Utilization Mach. Learn. Med. Imag.*, 2021, pp. 122–132.
[60] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

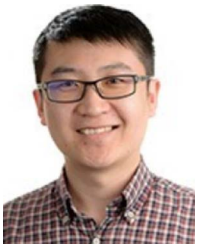
Qizhou Wang is currently working toward the PhD degree with the Department of Computer Science, Hong Kong Baptist University. His research primarily focuses on trustworthy machine learning, such as label noise learning, adversarial training, and OOD detection. Additionally, he is also exploring foundation model fine-tuning, including directions like unlearning, editing, and alignments. He has published more than 10 papers in prominent journals and conferences, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, NeurIPS, ICLR, etc.



Bo Han (Senior Member, IEEE) is currently an assistant professor in machine learning and a director of Trustworthy Machine Learning and Reasoning Group, Hong Kong Baptist University, and a BAIHO visiting scientist with RIKEN Center for Advanced Intelligence Project (RIKEN AIP). He has served as senior area chair of NeurIPS, and area chairs of NeurIPS, ICML and ICLR. He has also served as associate editors of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, MLJ and *Journal of Artificial Intelligence Research*, and Editorial Board members of *Journal of Machine Learning Research* and *MLJ*. He received Outstanding Paper Award at NeurIPS, Most Influential Paper at NeurIPS, Outstanding Student Paper Award at NeurIPS Workshop, Notable Area Chair at NeurIPS, Outstanding Area Chair at ICLR, and Outstanding associate editor with *IEEE Transactions on Neural Networks and Learning Systems*. He received the RGC Early CAREER Scheme, IEEE AI's 10 to Watch Award, IJCAI Early Career Spotlight, RIKEN BAIHO Award, Dean's Award for Outstanding Achievement, Microsoft Research StarTrack Scholars Program, and Faculty Research Awards from ByteDance, Baidu, Alibaba and Tencent.

Yang Liu (Senior Member, IEEE) received the BS and MS degrees in automation from the National University of Defense Technology, in 2004 and 2007, respectively, and the PhD degree in computing from the Hong Kong Polytechnic University, in 2011. He is currently an assistant professor with the Department of Computer Science, Hong Kong Baptist University and the associate director of the Health Informatics Center. During 2011–2012, he was a postdoctoral research associate with the Department of Statistics, Yale University. His research interests include artificial intelligence, machine learning, and applied mathematics, as well as their applications in high-dimensional/heterogeneous data analytics, complex systems modeling, computational epidemiology, and infectious disease modeling. He has published more than 90 peer-reviewed papers in reputable venues, including top-tier journals such as *Lancet Discovery Science*, *Artificial Intelligence Journal*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Automatic Control*, *IEEE Transactions on Autonomous Mental Development*, *ACM Transactions on Intelligent Systems and Technology*, *Pattern Recognition*, and *NeuroImage*, as well as top-tier conferences such as AAAI, IJCAI, SIGIR, ACMMM, WWW, and CIKM.

Chen Gong (Senior Member, IEEE) received the dual doctoral degrees from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS), respectively. Currently, he is a full professor of SJTU. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has published more than 130 technical papers at prominent journals and conferences such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Journal of Machine Learning Research*, *International Journal of Computer Vision*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Image Processing*, ICML, NeurIPS, ICLR, CVPR, ICCV, ECCV, AAAI, IJCAI, ICDM, etc. He serves as the associate editor of *IEEE Transactions on Circuits and Systems for Video Technology*, *Neural Networks*, and *NePL*, and also the area chair or senior PC member of several top-tier conferences such as AAAI, IJCAI, ICML, ICLR, ECML-PKDD, AISTATS, ICDM, ACM MM, etc. He won the ICDM Best Student Paper Runner-Up Award, the second prize of Natural Science Award of the Chinese Institute of Electronics, "Excellent Doctorial Dissertation Award" of Chinese Association for Artificial Intelligence, "Wu Wen-Jun AI Excellent Youth Scholar Award", and the Scientific Fund for Distinguished Young Scholars of Jiangsu Province. He was also selected as the "Global Top Chinese Young Scholars in AI" released by Baidu, and "World's Top 2% Scientists" released by Stanford University.



Tongliang Liu (Senior Member, IEEE) is the director of Sydney AI Centre, University of Sydney. He is an affiliated professor with MBZUAI and a visiting scientist with RIKEN AIP. He is broadly interested in the fields of trustworthy machine learning and its interdisciplinary applications, with a particular emphasis on learning with noisy labels, adversarial learning, causal representation learning, transfer learning, unsupervised learning, and statistical deep learning theory. He has authored and co-authored more than 200 research articles including ICML, NeurIPS, ICLR, CVPR, ICCV, ECCV, AAAI, IJCAI, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *Journal of Machine Learning Research*. He is/was a senior meta reviewer for many conferences, such as NeurIPS, ICLR, AAAI, and IJCAI. He is a co-editor-in-chief of the *Neural Networks*, an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *Transactions on Machine Learning Research*, and *ACM Computing Surveys*, and is on the Editorial Boards of *Journal of Machine Learning Research* and *MLJ*. He is a recipient of CORE Award for Outstanding Research Contribution in 2024, the IEEE AI's 10 to Watch Award in 2022, the Future Fellowship Award from Australian Research Council (ARC) in 2022, the Top-40 Early Achievers by The Australian in 2020, and the Discovery Early Career Researcher Award (DECRA) from ARC in 2018.

Jiming Liu (Fellow, IEEE) received the PhD degree in electrical engineering from McGill University. He is a chair professor in computer science and the associate vice-president (research development) with Hong Kong Baptist University. His research journey, for more than three decades, has covered multiple fields, including: Artificial Intelligence & machine learning, data-driven complex systems modeling, and health informatics & computational epidemiology. He was elected as a fellow of the IEEE in 2011, for his research contributions to Multi-Agent Autonomy-Oriented Computing (AOC) and Web Intelligence (WI). He was a recipient of the Chinese Association for Artificial Intelligence (CAAI) Wu Wenjun Artificial Intelligence Science and Technology Award in 2017. He has been invited to deliver Keynote or Plenary Talks at numerous international conferences in the fields of computer science and artificial intelligence, including the International Joint Conference on Artificial Intelligence (IJCAI).