Revisiting Deep Ensemble Uncertainty for Enhanced Medical Anomaly Detection

Yi Gu[†1], Yi Lin[†1], Kwang-Ting Cheng[2], and Hao Chen[✉1,3,4]

[1]Department of Computer Science and Engineering, HKUST, Hong Kong, China
[2]Department of Electronic and Computer Engineering, HKUST, Hong Kong, China
[3]Department of Chemical and Biological Engineering, HKUST, Hong Kong, China
[4]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China
yguax@connect.ust.hk, jhc@cse.ust.hk

Abstract. Medical anomaly detection (AD) is crucial in pathological identification and localization. Current methods typically rely on uncertainty estimation in deep ensembles to detect anomalies, assuming that ensemble learners should agree on normal samples while exhibiting disagreement on unseen anomalies in the output space. However, these methods may suffer from inadequate disagreement on anomalies or diminished agreement on normal samples. To tackle these issues, we propose D2UE, a Diversified Dual-space Uncertainty Estimation framework for medical anomaly detection. To effectively balance agreement and disagreement for anomaly detection, we propose Redundancy-Aware Repulsion (RAR), which uses a similarity kernel that remains invariant to both isotropic scaling and orthogonal transformations, explicitly promoting diversity in learners' feature space. Moreover, to accentuate anomalous regions, we develop Dual-Space Uncertainty (DSU), which utilizes the ensemble's uncertainty in input and output spaces. In input space, we first calculate gradients of reconstruction error with respect to input images. The gradients are then integrated with reconstruction outputs to estimate uncertainty for inputs, enabling effective anomaly discrimination even when output space disagreement is minimal. We conduct a comprehensive evaluation of five medical benchmarks with different backbones. Experimental results demonstrate the superiority of our method to state-of-the-art methods and the effectiveness of each component in our framework. Our code is available at https://github.com/Rubiscol/D2UE.

Keywords: Anomaly detection · Ensemble learning · Diversity

1 Introduction

Anomaly detection (AD) is an essential task in medical image analysis, encompassing early detection of medical diseases [1,27] and pathological localization [24]. The primary objective of visual medical AD is to identify images

---

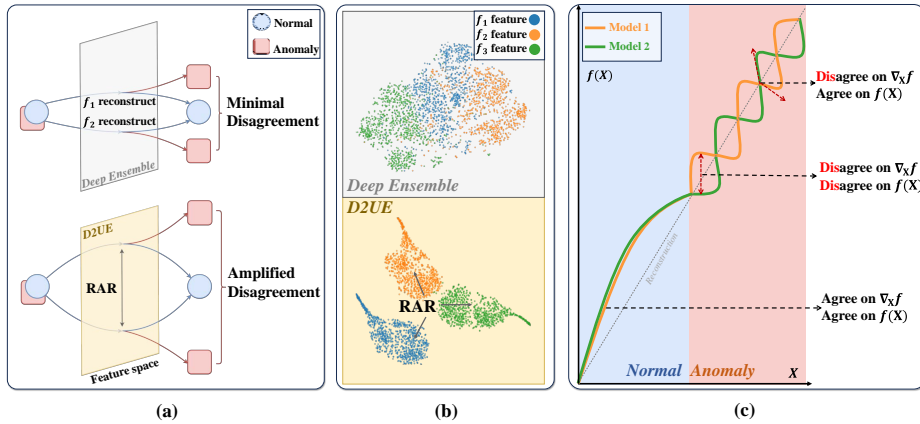[†] Equal contribution; ✉ corresponding author.

Fig. 1: **(a)**: An illustration of redundancy-aware repulsion (RAR). Disagreement on anomalies is amplified between different learners' feature spaces, while normal input converges to similar reconstructions guided by reconstruction training. **(b)**: A t-SNE [20] plot of feature spaces from three learners on the anomaly. Feature spaces are pushed away by RAR during training. **(c)**: An illustration of dual-space uncertainty (DSU) in 1D regression with two learners. Utilizing output space uncertainty fails to differentiate the anomaly at the upper point. In comparison, DSU utilizes the disagreement on $\nabla_X f$ to detect such anomalies.

containing diseases and pinpoint anomalous pixels within them. However, obtaining a sufficient number of anomalous samples that cover the vast spectrum of disease types can be challenging, as these samples often require specialized annotations [6]. Consequently, AD tasks are often formulated as one-class classification problem, wherein only normal data is utilized for model training [8].

Prevailing approaches mainly focus on reconstruction-based anomaly detection employing Autoencoders [14,21] or Generative Adversarial Networks [24,1]. These methods endeavor to maximize the likelihood of normal samples derived from training data. During inference, anomalies are detected based on per-pixel reconstruction error or model probability distribution. Nevertheless, these methods are limited by imprecise reconstructions or poorly calibrated likelihoods [5].

To circumvent a direct estimation of normal probability distributions, an alternative framework leveraging deep ensembles' uncertainty has emerged. This framework comprises multiple learners that perform self-supervised tasks [6,4] or acquires surrogate labels through a pretrained encoder [3,23]. Typically, learners undergo randomized training with distinct weight initializations [6] or Monte-Carlo dropout [13]. The underlying hypothesis posits that diverse learners should agree on normality while disagreeing on unseen anomalies in the output space.

However, balancing the trade-off between agreement and disagreement is challenging. Randomized training may not guarantee sufficient disagreement on anomalies, as learners in ensemble learning inherently tend to adopt the sim-

plest decision boundary [2]. This phenomenon, known as simplicity bias [25], inhibits learners' diversity and subsequently results in minimal disagreements on anomaly outputs [19]. To address simplicity bias, previous methods attempted to induce repulsion among learners in output space [22] or weight space [9]. Nevertheless, these approaches may culminate in either underfitting of individual models [11] or neural network redundancy [18], where models possess distinct weights yet output the same [12]. Consequently, learners' agreement on normal samples would be compromised.

In this paper, we propose a novel ensemble-based uncertainty estimation framework for medical anomaly detection called Diversified Dual-space Uncertainty Estimation (**D2UE**). To enhance learners' disagreement on anomalies, we introduce a Redundancy-Aware Repulsion (**RAR**), which encourages learners to reconstruct training samples from more diversified feature spaces. To promote this diversification without succumbing to neural network redundancy, RAR regulates ensemble training using a similarity kernel invariant to both isotropic scaling and orthogonal transformation. During inference, disagreement on anomalies is amplified between different learners' feature spaces (see Fig. 1(a)). Unlike output space repulsion, feature space repulsion does not result in underfitting for normal samples in output space. Consequently, normal features converge to similar reconstructions guided by reconstruction training. Moreover, to emphasize anomalous regions, we develop a Dual-Space Uncertainty (**DSU**) that combines uncertainties in both input and out spaces. In input space, we calculate gradients of the reconstruction error with respect to inputs, which are further combined with outputs to estimate the final uncertainty. DSU discriminates anomalies through input space disagreement even if learners exhibit minimal disagreement in output space (see Fig. 1(c)). Our primary contributions are as follows:

- We address medical anomaly detection from an uncertainty estimation perspective. We undertake a pioneering exploration of diversity in deep ensembles' uncertainty and propose D2UE, a novel Diversified Dual-space Uncertainty Estimation approach for medical anomaly detection.
- We propose Redundancy-Aware Repulsion (RAR) to strike an effective balance between agreement and disagreement, thereby enhancing anomaly detection accuracy.
- We design a Dual-Space Uncertainty (DSU) to emphasize anomalous regions, particularly when anomalies exhibit minimal disagreement in output space.
- We conduct comprehensive experiments on five medical benchmarks with different backbones. Experimental results demonstrate the superiority of our method to state-of-the-art methods and the effectiveness of each component.

2   Method

The proposed Diversified Dual-space Uncertainty Estimation (D2UE) framework consists of N learners $f(\cdot)$, each possessing identical Autoencoder architectures. Anomaly detection is achieved through uncertainty estimation, where learners
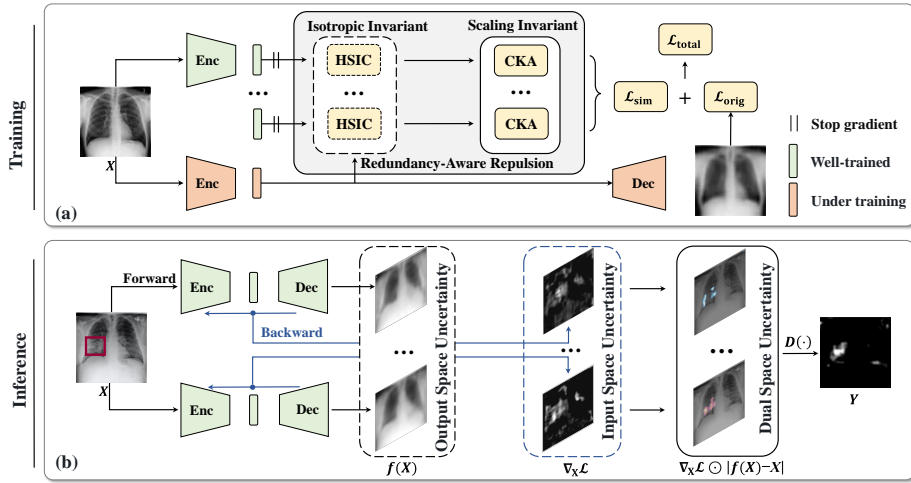
Fig. 2: Overview of D2UE. In the training stage, the redundancy-aware repulsion (RAR) module amplifies the diversity of different models with both isotropic and scaling invariance. In the inference stage, the dual-space uncertainty is calculated, utilizing both $f(X)$ in the output space and $\nabla_X \mathcal{L}$ in the input space.

should agree on normal samples while disagreeing on anomalies. To this end, we propose redundancy-aware repulsion (RAR) to enhance learners' disagreement on anomalies while maintaining agreement on normal samples. Moreover, to further emphasize anomalous regions, dual-space uncertainty (DSU) is designed to combine uncertainties in output and input spaces during inference. In the following, we will detail each component.

Redundancy-Aware Repulsion for Feature Space. Existing methods mainly train ensemble learners under the reconstruction loss $\mathcal{L}_{\mathrm{orig}}$, such as mean square error loss, without any repulsion encouraging constraint. To encourage learners' disagreement on anomalies, our approach incorporates a similarity constraint loss $\mathcal{L}_{\mathrm{sim}}$ that induces repulsion in feature space. As depicted in Fig. 2, learners undergo sequential training and the learner under training is denoted as f_n. The $\mathcal{L}_{\mathrm{sim}}$ of f_n, except for 0 in the first one, is optimized to minimize the Centered Kernel Alignment (CKA similarity) [16] between its feature vector Q^n and well-trained learner's feature P^j in total $n-1$ trained learners:

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{n-1} \sum_{j=1}^{n-1} \mathrm{CKA}(P^j, Q^n) \tag{1}$$

$$\mathrm{CKA}(P, Q) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K)}\sqrt{\mathrm{HSIC}(L, L)}} \tag{2}$$
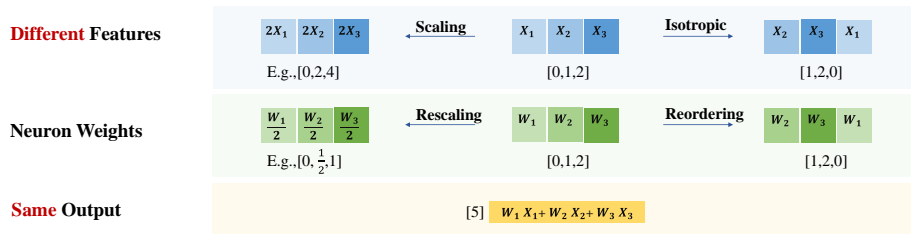
Fig. 3: An illustration of neural network redundancy. Different features may output the same by weight re-scaling or spatial reordering.

In Eq. (2), $K = PP^T$ and $L = QQ^T$, HSIC is Hilbert-Schmidt Independence Criterion [15] measuring the independence between variables:

$$\text{HSIC}(K, L) = \frac{1}{(r-1)^2} tr(KHLH) \tag{3}$$

where r is the rank of K and L, H is the centering matrix, and tr is the matrix trace. In the following, we will elaborate the motivation behind our design.

First, it is crucial to establish an optimization objective that explicitly encourages repulsion in learners' feature spaces. To this end, we endow a more straightforward target to the training of a learner: to reconstruct normal samples using $\mathcal{L}_{\text{orig}}$ through more different paths supervised by all well-trained learners via $\mathcal{L}_{\text{sim}}$. By passing anomalous samples through more diversified feature spaces, learners are encouraged to exhibit more significant disagreements while maintaining consistency on normal samples during inference.

Second, it is essential to identify a suitable similarity kernel $S(\cdot)$ that eliminates neural network redundancy. To achieve this, we establish two properties for $S(\cdot)$: 1. **isotropic scaling invariance**: $S(P, Q) = S(\alpha P, \beta Q)$ for $\alpha, \beta \in \mathbb{R}^+$. 2. **orthogonal transformation invariance**: $S(P, Q) = S(PU, QV)$ for $U, V \in$ orthogonal transformation. Scaling invariance implies that neurons cannot deceive well-trained neurons into perceiving different features merely by re-scaling their weights during training. Similarly, orthogonal invariance prevents neurons from deceiving well-trained neurons into perceiving different features merely by spatial reordering during training. As illustrated in Fig. 3, different features may still output the same by weight re-scaling or reordering. Therefore, both isotropic and orthogonal invariance are necessary to effectively promote feature space diversity without succumbing to neural network redundancy.

In Eq. (3), $K = PP^T$ and $L = QQ^T$ ensure orthogonal invariance since $(PU)(PU)^T = PP^T$. Normalized term $\sqrt{\text{HSIC}(K, K)}\sqrt{\text{HSIC}(L, L)}$ ensures scaling invariance in $\mathcal{L}_{\text{sim}}$. Finally, the total loss $\mathcal{L}_{\text{total}}$ is formulated as follows:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{orig}}}_{\text{agree on normal}} + \underbrace{\lambda \mathcal{L}_{\text{sim}}}_{\text{disagree on anomaly}} \tag{4}$$

where λ controls the strength of the repulsion to the overall loss function[†].

Dual-Space Uncertainty. To identify anomalous samples, an image X is input to all learners to estimate uncertainty, generating a pixel-level anomaly score map Y. Previous methods calculated Y based on models' outputs $f(X)$:

$$Y = D(f_i(X)), i = 1....N \tag{5}$$

wherein $D(\cdot)$ signifies the deviation function. However, relying solely on the uncertainty of $f(X)$ may fail to discriminate anomalies in some cases. In the context of the same reconstruction task, learners can sometimes output similar reconstructions even for anomalies, as depicted in Fig. 1(c). Despite minimal disagreement in output space, a model can hold unique first-order derivatives $\nabla_X f$ in input space [28]. Typically, $\nabla_X f$ is used to construct saliency maps and visually interpret models' divergent attention on input pixels [26]. Inspired by this, we devise DSU that correlates output space uncertainty with input space uncertainty to better reveal learners' disagreement towards anomalies. Specifically, in input space, we calculate $\nabla_X \mathcal{L}$, the gradient of reconstruction error with respect to the input, as opposed to a large Jacobian matrix $\nabla_X f$ in the reconstruction model. The $\nabla_X \mathcal{L}$ is further elementally multiplied with normalized output $|f(X) - X|$ to calculate the final uncertainty:

$$Y = D(\nabla_X \mathcal{L}_i \odot |f_i(X) - X|), i = 1...N \tag{6}$$

where $\odot$ symbolizes the element-wise multiplication. Consequently, even if distinct learners inadvertently agree in output space, the anomalous region can still be accentuated by input space disagreement.

3   Experiment

Datasets and Evaluation metrics. We conduct experiments on five medical datasets, encompassing various modalities such as chest X-ray, magnetic resonance imaging, and retinal fundus. Datasets include 1. RSNA: RSNA Pneumonia Detection Challenge dataset[†]. 2. VinDr-CXR: VinBigData Chest X-ray Abnormalities Detection dataset[†]. 3. CXAD: Chest X-ray Anomaly Detection dataset [6]. 4. Brain MRI: Brain Tumor MRI dataset[†]. 5. LAG: Large-scale Attention-based Glaucoma dataset [17]. To ensure consistency with other studies, we adhere to the split criteria outlined in [4,6,7] and details can be seen in supplementary materials. We adopt the area under the ROC curve (AUC) and average precision (AP) for image-level classification.

Implementation Details. We optimize our model using the Adam optimizer with an initial learning rate of $5e-4$ for AE and MemAE, and $1e-3$ for AEU. The default batch size is set to 64 with the image size of 64×64. Each learner is randomly initialized and trained for 250 epochs. λ is set to 1 in our experiment.

[†] Ablation studies of λ and the layer of $\mathcal{L}_{sim}$ are included in supplementary materials.
[†] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[†] https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection
[†] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Table 1: Comparison with SOTA methods. "Reconstruction" refers to "Reconstruction methods", "Uncertainty" refers to "Uncertainty estimates methods". The best result is highlighted in bold and the second best result is underline.

	Methods	RSNA AUC	RSNA AP	VinDr-CXR AUC	VinDr-CXR AP	CXAD AUC	CXAD AP	Brain MRI AUC	Brain MRI AP	LAG AUC	LAG AP	Average AUC	Average AP
Reconstruction	AE	66.9	66.1	55.9	60.3	55.6	59.6	79.7	71.9	79.3	76.1	67.5	66.8
	MemAE [14]	68.0	67.1	55.8	59.8	56.0	60.0	77.4	70.0	78.5	74.9	66.7	66.6
	AEU [21]	86.7	84.7	73.8	72.8	66.4	66.9	94.0	89.0	81.3	78.9	80.4	78.5
	IGD [27]	81.2	78.0	59.2	58.7	55.2	57.6	94.3	90.6	80.7	75.3	74.1	72.0
	f-AnoGAN [24]	79.8	75.6	76.3	74.8	61.9	67.3	82.5	74.3	84.2	77.5	76.9	73.9
	Ganomaly [1]	71.4	69.1	59.6	60.3	62.5	63.0	75.1	69.7	77.7	75.7	69.3	67.6
Uncertainty	DDAD [6]	87.3	86.4	74.3	71.5	69.2	71.7	84.5	83.3	75.3	75.1	78.1	77.6
	Multi-ST [23]	86.0	83.5	68.1	68.2	60.8	63.6	95.6	92.7	79.1	74.4	77.9	76.5
	RDAD [10]	85.7	82.9	69.4	66.3	55.3	56.9	96.0	92.5	82.7	78.5	77.8	75.4
	Destseg [29]	73.3	73.9	64.4	66.8	55.8	56.0	**96.7**	**95.6**	73.6	72.1	72.8	72.9
	Ours (AE)	84.1	82.4	76.6	74.4	65.2	66.4	89.2	83.0	82.5	79.0	79.5	77.0
	Ours (AEU)	**88.6**	**86.8**	**78.7**	**76.1**	**72.9**	**71.8**	96.2	92.0	**86.3**	**84.0**	**84.5**	**82.6**

Table 2: Ablation study for D2UE, where "Ens" and "Unc" stand for "ensemble reconstruction score" and "ensemble uncertainty estimation from output space", respectively. Results are in AUC and the best result is bold in each column.

Ens	Unc	RAR	DSU	RSNA AE	RSNA MemAE	RSNA AEU	VinDr-CXR AE	VinDr-CXR MemAE	VinDr-CXR AEU	CXAD AE	CXAD MemAE	CXAD AEU	Brain MRI AE	Brain MRI MemAE	Brain MRI AEU	LAG AE	LAG MemAE	LAG AEU	Average AE	Average MemAE	Average AEU
✓				66.9	68.0	86.7	55.9	55.8	73.8	55.6	56.0	66.4	79.7	77.4	94.0	79.3	78.5	81.3	67.5	67.1	80.4
✓	✓			69.4	68.0	87.3	60.1	59.5	74.3	59.8	59.4	69.2	59.8	52.6	84.5	72.1	70.4	75.3	64.2	62.0	78.1
✓	✓	✓		76.8	77.9	87.8	71.8	71.6	75.6	62.9	62.4	72.4	61.1	63.9	91.0	79.6	78.7	79.9	70.4	70.9	81.3
✓	✓		✓	80.8	81.6	88.5	68.5	66.2	77.5	61.1	62.6	72.6	88.9	86.2	95.1	82.4	78.5	83.7	75.5	75.0	83.5
✓	✓	✓	✓	**84.1**	**83.5**	**88.6**	**76.6**	**75.7**	**78.7**	**65.2**	**63.1**	**72.9**	**89.2**	**87.5**	**96.2**	**82.5**	**79.0**	**86.3**	**79.5**	**77.8**	**84.5**

Comparisons with State-of-the-Art Methods. Table 1 presents a comparison of our approach with an extensive assortment of state-of-the-art (SOTA) methods in AUC% and AP%. Compared SOTA methods include reconstruction-based methods such as MemAE [14], AEU [21], IGD [27], f-AnoGAN [24], Ganomaly [1], as well as ensemble uncertainty estimation methods such as DDAD [6], Multi-ST [23], RDAD [10], Destseg [29]. Employing the AEU backbone, our method exhibits exceptional performance in comparison to other methods across multiple medical image datasets, such as RSNA, VinDr-CXR, CXAD, and LAG, in addition to achieving the second-highest AUC in Brain MRI. Specifically, it surpasses SOTA results in AUC and AP by 1.3% and 0.4% (RSNA), 2.1% and 1.3% (VinDr-CXR), 3.7% and 0.1% (CXAD), and 2.1% and 5.0% (LAG), demonstrating our method's effectiveness and superiority.

Ablation Study. We conduct an ablation analysis utilizing three distinct Autoencoder architecture backbones: AE, MemAE, and AEU, to scrutinize the effectiveness of each constituent of D2UE. Table 2 showcases a comparison of

Table 3: From left to right: No constraint, Euclidean and Manhattan distance (We use $e^{-\text{distance}(P,Q)}$ to stabilize training), Cosine similarity, Pearson correlation coefficient, CKA similarity. ✓ indicates the presence of the given mathematical property, while ✗ indicates the opposite. The results are based on vanilla Autoencoder in the RSNA.

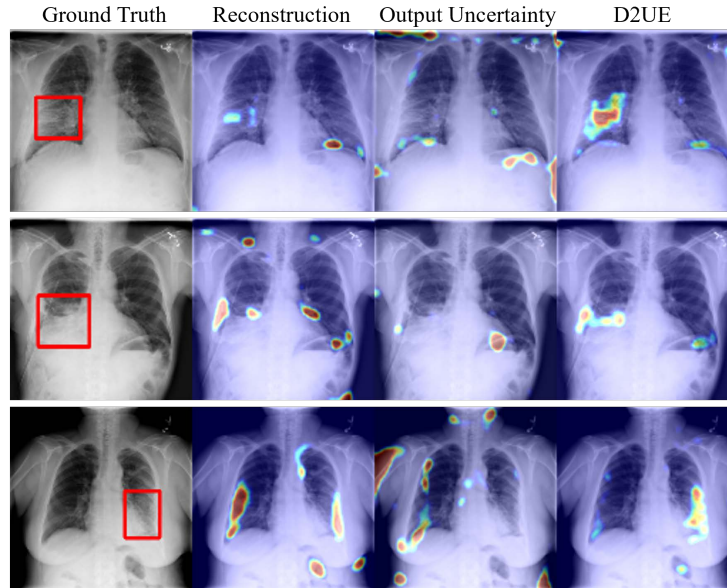|	Similarity metrics	None	Euclidean	Manhattan	Cosine	Pearson	CKA
Invariant to:	Isotropic scaling	—	✗	✗	✓	✓	✓
	Orthogonal transform	—	✗	✗	✗	✗	✓
	AUC	69.4	70.2	71.3	72.1	72.9	76.8

Fig. 4: Visualization results. Red bounding boxes indicate abnormal regions.

the AUC% for D2UE variations across five datasets. Variations include: 1) Ens, ensemble reconstruction score; 2) +Unc, ensemble uncertainty estimation from output space; 3) +RAR, incorporating RAR into training; 4) +DSU, utilizing DSU during inference. Results confirm that proposed components contribute positively to the enhancement of accuracy. For instance, on the RSNA, our RAR and DSU respectively improve the performance by 7.4% AUC and 11.4% within the AE model, demonstrating the effectiveness of our method.

Choice for similarity metric. We examine various similarity metrics for RAR, and show results in Table. 3. It is discerned that the CKA similarity attains the highest AUC with 76.8% among all other similarity functions. This is followed

by the Pearson correlation coefficient (72.6%), Cosine similarity (72.1%), Manhattan distance(71.3%), Euclidean distance (70.2%), and no constraint (69.4%). The empirical results substantiate that both scaling invariance and orthogonal invariance contribute positively to the accuracy.

Visualization results. We visualize heat maps of ensemble reconstruction, ensemble uncertainty estimation from output space, and D2UE on RSNA in Fig. 4. Our method can significantly emphasize abnormal regions.

4   Conclusion

In this paper, we presented D2UE, a Diversified Dual-space Uncertainty Estimation framework for medical anomaly detection. To effectively balance the diversity among ensemble learners and reconstruction accuracy, we introduced redundancy-aware repulsion, which compels learners to disagree on anomalies without compromising agreement on normal inputs. Further, we propose dual-space uncertainty highlighting anomalous regions during inference to enhance the model's discrimination ability. The framework has been extensively tested on various medical benchmarks, and experimental results demonstrate the superiority of our method to state-of-the-art methods and the effectiveness of each component. In future work, we intend to explore the quantitative relationship between ensemble diversity and final performance, and reduce the time and computational cost of training model ensembles.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 622–637. Springer (2019)
2. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
3. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4183–4192 (2020)

4. Bozorgtabar, B., Mahapatra, D., Thiran, J.P.: Amae: Adaptation of pre-trained masked autoencoder for dual-distribution anomaly detection in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 195–205. Springer (2023)

5. Cai, Y., Chen, H., Cheng, K.T.: Rethinking autoencoders for medical anomaly detection from a theoretical perspective. arXiv preprint arXiv:2403.09303 (2024)

6. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy for anomaly detection in chest x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 584–593. Springer (2022)

7. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Medical Image Analysis **86**, 102794 (2023)

8. Cai, Y., Zhang, W., Chen, H., Cheng, K.T.: Medianomaly: A comparative study of anomaly detection in medical images. arXiv preprint arXiv:2404.04518 (2024)

9. D'Angelo, F., Fortuin, V.: Repulsive deep ensembles are bayesian. Advances in Neural Information Processing Systems **34**, 3451–3465 (2021)

10. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9737–9746 (2022)

11. Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In: International Conference on Machine Learning. pp. 1184–1193. PMLR (2018)

12. Entezari, R., Sedghi, H., Saukh, O., Neyshabur, B.: The role of permutation invariance in linear mode connectivity of neural networks. arXiv preprint arXiv:2110.06296 (2021)

13. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

14. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1705–1714 (2019)

15. Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. Advances in neural information processing systems **20** (2007)

16. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMLR (2019)

17. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10571–10580 (2019)

18. Lin, Y., Liu, Y., Chen, H., Yang, X., Ma, K., Zheng, Y., Cheng, K.T.: Lenas: Learning-based neural architecture search and ensemble for 3-d radiotherapy dose prediction. IEEE Transactions on Cybernetics (2024)

19. Lin, Y., Qu, Z., Chen, H., Gao, Z., Li, Y., Xia, L., Ma, K., Zheng, Y., Cheng, K.T.: Nuclei segmentation with point annotations from pathology images via self-supervised learning and co-training. Medical Image Analysis **89**, 102933 (2023)

20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

21. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–538 (2020)
22. Pagliardini, M., Jaggi, M., Fleuret, F., Karimireddy, S.P.: Agree to disagree: Diversity through disagreement for better transferability. arXiv preprint arXiv:2202.04414 (2022)
23. Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R.: Multiresolution knowledge distillation for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14902–14912 (2021)
24. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)
25. Shah, H., Tamuly, K., Raghunathan, A., Jain, P., Netrapalli, P.: The pitfalls of simplicity bias in neural networks. Advances in Neural Information Processing Systems **33**, 9573–9585 (2020)
26. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
27. Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S.H., Verjans, J.W., Singh, R., Carneiro, G.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 128–140 (2021)
28. Trinh, T., Heinonen, M., Acerbi, L., Kaski, S.: Input-gradient space particle inference for neural network ensembles. arXiv preprint arXiv:2306.02775 (2023)
29. Zhang, X., Li, S., Li, X., Huang, P., Shan, J., Chen, T.: Destseg: Segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3914–3923 (2023)