Out-of-distribution Detection via Nearest Neighbors and Null Vectors on Principal Feature Space

Trang Nguyen *
Viettel High Tech
Hanoi, Vietnam
trangnt236@viettel.com.vn

Minh Nguyen Tong *
Viettel High Tech
Hanoi, Vietnam
minhnt425@viettel.com.vn

Hung Manh Lu
Viettel High Tech
Hanoi, Vietnam
hunglm26@viettel.com.vn

Bang Van Le
Viettel High Tech
Hanoi, Vietnam
banglv1@viettel.com.vn

***Abstract*—Out-of-distribution (OOD) detection is crucial for ensuring the reliability of deep learning models in open-world environments, especially in safety-critical applications such as autonomous driving. Existing OOD detection methods typically focus on either classifier-based scores, which suffers from over-confidence, or distance-based approaches that lack fine-graininess due to their class-agnostic nature. Recently, some fused scoring functions have been proposed but be limited to single type of feature-based information. To address these challenges, we propose Nearest Neighbors with Null Space Analysis (k-NNuSA), a novel distance-focused approach that combines both distance to nearest neighbors within the in-distribution (ID) data and the residual against the ID principal subspace. Our method scales the generalized distance using classification confidence, enhancing fine-grained detection capability. We extensively evaluate k-NNuSA on different CNN-based and transformer-based architectures, using ImageNet-1K as the ID reference set and a variety of OOD datasets, demonstrating its state-of-the-art performance of 96.23% AUROC and 19.34% FPR95. Our proposed scoring function consists of post-hoc operations which is instantly applicable without re-training.**

***Index Terms*—Out-of-distribution Detection, Deep Learning, Classification, Uncertainty**

I. Introduction

The open-world environment is full and unknowns, posing great challenges for deep networks that are required to handle diverse inputs reliably. Out-of-distribution (OOD) problem arises when a network encounters data beyond the semantics of instances with which it was trained. The detection of such anomalous examples is crucial for preventing the malfunctions and ensuring the error-free delivery of deep networks, which cannot be neglected in safety-critical applications such as autonomous driving [1]–[4] and medical analysis [5]–[7]. Recently, a plethora of researches have studied the issue of out-of-distribution detection [8]–[13].

The standard concept for OOD detection is to derive a scoring function from the mature networks such that the OOD samples relatively exhibit higher scores than in-distribution (ID) samples. Researchers have been designing many of such functions by exploiting the uniqueness of ID-ness or vice versa. One popular paradigm is to derive scores according to the output signal of classification networks, including: (i) "*probability*", such as maximum softmax probabilities [8], minimum KL-divergence between softmax and mean class-conditional distributions [12]; (ii) "*logit*", such as energy function [10], maximum logits [12]. These classifier-based approaches fully leverage the class-dependent information of ID samples for fine-grained detection capability. However, their ignorance of the feature space disregards class-agnostic information [14] which may not affect classification but depict the distinction between ID and OOD samples. Hence, the classifier-based scores are possibly vulnerable to model overconfidence, hindering OOD detection [15].

On the other hand, the distance-based approaches, such as the similarity to nearest neighbors [16], the norm of residual between the feature and the pre-image of its low-dimensional embedding [17], detect OOD samples based on their distance to ID data in the feature space. These approaches exclusively exploit features which provide class-agnostic information to address the overconfidence issues in OOD detection. Unfortunately, these scores, without class-dependent nature, struggle to distinguish subtle differences between samples and ID classes, limiting fine-grained detection capability [18]. Furthermore, we argue that utilizing solely either the residual or the distance to nearest neighbors may not capable of revealing the "sufficient" distance of a sample to ID data. Intuitively, the residual relies on assumptions about the data distribution since the feature space can only be induced using observed data, while nearest neighbors is non-parametric but be fragile to noise and bounded to the locality.

Motivated by such diverse factors that manifests in OOD detection, we propose Nearest Neighbors with Null Space Analysis (k-NNuSA), a generalized distance-focused approach that leverages not only the distance to nearest neighbors within ID distribution but also the residual of feature against ID principal subspace to filter OOD samples out of the database. We then scale the generalized distance based on the confidence of classification to improve its fine-graininess. Built upon this construction, k-NNuSA will tend to assign a sample, which more uncertain and distant to either hyperplane or neighborhood of ID data, as OOD.

We extensively validate the OOD detection capability of our proposed method across various models and datasets, using ImageNet-1K as the ID reference. The model architectures range from CNN-based networks, including legacy ResNet-50 [19], recent BiT [20] and RepVGG [21], to transformer-based networks, including ViT [22], DeiT [23] and Swin Transformer [24]. The experiments on four different OOD

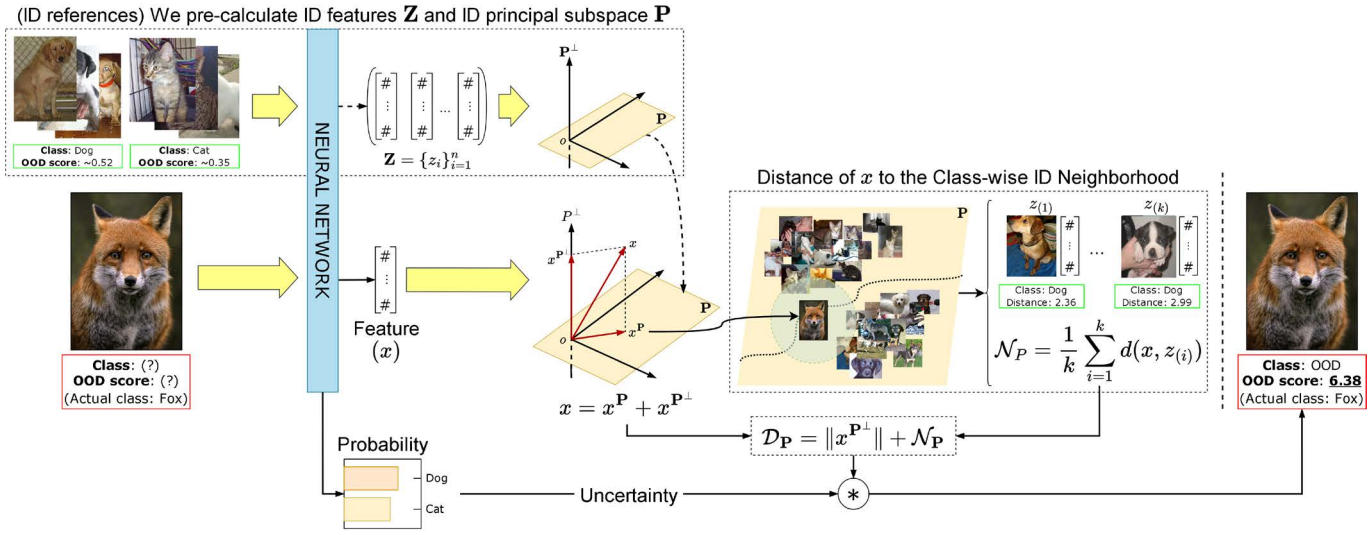

Fig. 1. ***The architecture of k-NNuSA***. The principal subspace $\mathbf{P}$ is induced by the features of ID reference data $\mathbf{Z} = \{z_i\}_{i=1}^n$ beforehand, with n as the sample size. During inference, the model produces a feature x for the input image. The distances to the hyperplane $\mathbf{P}$ and k nearest neighbors in projection are measured and scaled with the confidence of model prediction. The final output is a score indicating the increasing possibility that the input is OOD.

datasets, including OpenImage-O, ImageNet-O, Texture and iNaturalist, demonstrates that k-NNuSA achieves the state-of-the-art performance of average 92.09% AUROC and 39.53% FPR95 against previous approaches over a wide variety of scenarios of OOD detection.

Our contributions are twofold. (i) We propose a generic scoring function, abbreviated as Nearest Neighbors with Null Space Analysis (k-NNuSA), that incorporates both distances to nearest neighbors and ID manifold scaled with the network confidence to reliably detect OOD for a large range of datasets and models. k-NNuSA is instant-applied, requiring neither extra OOD reference nor expensive re-training. (ii) We conduct comprehensive experiments on the ImageNet-1K benchmark with multiple OOD datasets and model architectures, consisting of CNN-based and transformer-based vision models, to assess thoroughly the detection capability of k-NNuSA. Here, our proposed method manages to achieve the state-of-the-art results over many previous works.

II. Methodology

We present Nearest Neighbors with Null Space Analysis (k-NNuSA), a generalized distance-focused scoring function for OOD detection that measures whether an image is OOD using the unified distances to nearest neighbors and ID principal subspace. Fig. 1 depicts the general framework of k-NNuSA. For a feature x, we (i) compute the residual $x^{\mathbf{P}^\perp}$ against the principal subspace $\mathbf{P}$; (ii) compute the average distance to k nearest neighbors; (iii) estimate how confident the model predicts the input image; (iv) aggregate and scale the distance and residual by the confidence of model prediction. The output is k-NNuSA score that is expected to be higher if the input image is potentially OOD. Without specified, we consider $\mathbf{Z} \in \mathbb{R}^{n\times d}$ as the ID reference set; $\mathbf{W}$ and $\mathbf{b}$ as the weight and bias of classification head, respectively.

A. Residual from the Principal Subspace

We consider a classification problem of C classes, whose logit $\mathbf{l} \in \mathbb{R}^C$ is projected from the feature $z \in \mathbb{R}^d$, i.e. $\mathbf{l} = \mathbf{W}^\top z + \mathbf{b}$. We safely omit the bias term $\mathbf{b}$ by shifting the coordinate system of feature space to a new origin $o \triangleq -(\mathbf{W}^\top)^+\mathbf{b}$, where $(\cdot)^+$ is the Moore-Penrose inverse. We abuse $\mathbf{Z}$ to define the shifted ID reference set, where rows are features in the new coordinate system with origin o. Accordingly, we formulate the eigendecomposition on the matrix $\mathbf{Z}^\top\mathbf{Z}$, as follows:

$$\mathbf{Z}^\top\mathbf{Z} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} \tag{1}$$

where $\mathbf{Q}$ is the square of $n \times n$ matrix whose i-th column is the eigenvector $\mathbf{q}_i$ of $\mathbf{Z}^\top\mathbf{Z}$ and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues $\mathbf{\Lambda}_{ii} = \lambda_i$, sorted descendingly. Therefore, we obtain the D-dimensional principal subspace $\mathbf{P}$ spanned from the first D eigenvectors in $\mathbf{Q}$. The principal subspace $\mathbf{P}$ presents only principal components of the original ID feature space and possibly reduces its noises and unnecessary details.

Given an image feature $x \in \mathbb{R}^d$, we decompose $x = x^{\mathbf{P}} + x^{\mathbf{P}^\perp}$, where the residual $x^{\mathbf{P}^\perp}$ is the projection of x onto the null space $\mathbf{P}^\perp$. We can compute the residual using the rest of eigenvectors from $(D+1)$-th to the last column in $\mathbf{Q}$, which is extracted as a new matrix $\mathbf{R} \in \mathbb{R}^{d\times(d-D)}$. Here, the residual $x^{\mathbf{P}^\perp}$ is of the closed form:

$$x^{\mathbf{P}^\perp} = \mathbf{R}\mathbf{R}^\top x \tag{2}$$

We use the norm of the residual $\|x^{\mathbf{P}^\perp}\|$ to measure how distant the image feature x is to the ID manifold, which reveals the OOD-ness of the corresponding input.

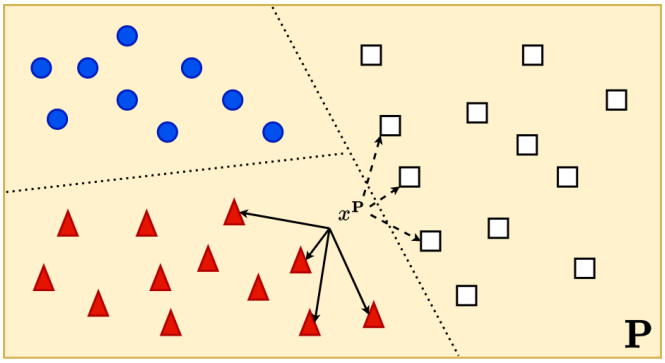

Fig. 2. ***The class-wise nearest neighbors of principal feature*** $x^{\mathbf{P}}$. The model predicts x^{P} as red triangle with low confidence. The dotted lines represent the decision boundaries between ID classes. The dashed arrows are the distances of $x^{\mathbf{P}}$ to nearest neighbors regardless of different classes. The solid arrows are the distances of $x^{\mathbf{P}}$ to the nearest neighbors of the same class (i.e., red triangle). One can see that the norm of the solid lines are larger than that of dashed ones.

B. Distance to the Class-wise Neighborhood

Based on the formulation in Eq. 1 and 2, the residual $x^{\mathbf{P}^\perp}$ depends on the observed ID reference set $\mathbf{Z}$ with an assumption that the principal subspace $\mathbf{P}$ is capable of representing the vast majority of ID population. In other words, the residual is a parametric scoring method and may not generalize well when the ID data turns dynamic (e.g., translated, rotated, ...) over time. Therefore, we address this potential pitfall by assessing additional distances of an image feature x to its nearest ID neighbors.

To make the measurement robust, we derive the distance of x to its neighborhood upon the principal space $\mathbf{P}$ which consists of more essential and less noisy information. Particularly, the OOD-ness of the image feature x is determined by the average dissimilarity of $x^{\mathbf{P}}$ to the k projected nearest neighbors $\{z_1^{\mathbf{P}}, ..., z_k^{\mathbf{P}}\}$, that have the same class c_x predicted for x, in the ID reference set $\mathbf{Z}$:

$$\mathcal{N}_{\mathbf{P}}(x) = \frac{1}{k}\sum_{i=1}^{k} d(x^{\mathbf{P}}, z_{(i)}^{\mathbf{P}}),\ \forall z_{(i)} \in \mathbf{Z}_x \tag{3}$$

where d is the dissimilarity function and the subset $\mathbf{Z}_x \in \mathbf{Z}$ consists of ID references of class c_x. We simply use L2-distance for d in experiments. The reordered index (i) is given in the ascending order of distance from the reference $z_{(i)}$ to the principal feature $x^{\mathbf{P}}$. We recognize the presence of near-OOD instances [18] which may occur in small intermediate regions between ID classes. Fig. 2 depicts a calling example: $x^{\mathbf{P}}$ are close to the decision boundary, meaning that x is likely to be OOD. Here we can observe that the norm of dashed arrows and the shortest solid arrow of common nearest neighbors are smaller than that of solid lines of the class-wise nearest neighbors. Hence, the latter can exhibit a higher score for x and be more capable of fine-grained detection. Inspired from this observation, we prefer class-wise nearest neighbors as in the formula in Eq. 3.

C. Nearest Neighbors with Null Space Analysis

Given the residual norm $\|x^{\mathbf{P}^\perp}\|$ and the average distance to class-wise neighborhood $\mathcal{N}_{\mathbf{P}}(x)$, we define the generalized distance-focused score $\mathcal{D}_{\mathbf{P}}$ to ID distribution as follows:

$$\mathcal{D}_{\mathbf{P}}(x) = \mathcal{H}(\mathbf{p}(x)) \cdot (\mathcal{N}_{\mathbf{P}}(x) + \|x^{\mathbf{P}^\perp}\|) \tag{4}$$

where $\mathcal{H}(\mathbf{p}) = \sum_{c=1}^{C} p_c^{\gamma}(1-p_c)^{\gamma}$ [26], given the probability vector $\mathbf{p} = (p_c)_{c=1}^{C}$ with $\mathbf{p} = \text{softmax}(\mathbf{l})$. $\mathcal{H}(\mathbf{p})$ is the entropy measuring the confidence of predicted probability $\mathbf{p}(x)$ for the image feature x. Thanks to the entropic term $\mathcal{H}(\mathbf{p})$, the two distances in $\mathcal{D}_{\mathbf{P}}$ of higher confidence x are scaled down and vice versa, providing the fine-graininess for the scoring function in OOD detection. Eq. 4 reveals that if a feature x deviates from either the ID manifold or its neighborhood, it is more likely to be OOD. The computational overhead is minimized as k-NNuSA solely utilize the features and outputs of the classification head which the model has pre-computed during its inference.

Following the post-analysis of ViM [13], we additionally attach the energy-based score [10] to $\mathcal{D}_{\mathbf{P}}(x)$, that formulates the overall k-NNuSA score, of the form:

$$k\text{-NNuSA}(x) = \mathcal{D}_{\mathbf{P}}(x) - \ln \sum_{c=1}^{C} e^{l_c} \tag{5}$$

On the one hand, we complete the energy score by adding diverse class-agnostic information, including the residual and the neighborhood. On the other hand, we observe that k-NNuSA achieves superior performance with the addition of this energy score that also emphasizes the fine-grained detection capability in scoring function.

III. Experiments and Results

This section illustrates our empirical investigations on the detection capability of k-NNuSA in comparison with the representative works, including ViM [13], residual [13], GradNorm [25], ReAct [11], Energy [10], Mahalanobis [9], KL-Matching [12], MaxLogit [12] and MSP [8]. We evaluate all the algorithms on large-scale OOD detection, ImageNet-1K benchmark, using Texture [27], iNaturalist [28], OpenImage-O [13] and ImageNet-O [29] as the OOD datasets. The feature extractors used consist of transformer-based and CNN-based pretrained vision models, that produces feature spaces of diverse quality and characteristics.

A. Evaluation Metrics

We compare the detection capability of scoring functions in terms of two popular metrics: AUROC (%) and FPR95 (%). AUROC is threshold-free and measures the area under the receiver operating characteristic curve. A larger area signifies better detection performance. FPR95 stands for the false positive rate at a 95% true positive rate, which means a smaller FPR95 indicates better performance.

TABLE I
THE PERFORMANCE OF k-NNUSA AND PREVIOUS WORKS FOR OOD DETECTION

	Dataset	Texture		iNaturalist		OpenImage-O		ImageNet-O		Average	
Model	Detection	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
BiT	MSP [8]	79.80	76.67	87.92	64.08	83.05	76.10	57.12	96.85	76.97	78.42
	MaxLogit [12]	81.65	73.72	86.76	70.60	82.30	79.86	63.01	96.85	78.43	80.26
	KL-Matching [12]	86.91	51.02	92.95	**33.28**	87.93	54.82	65.68	86.65	83.37	56.44
	Mahalanobis [9]	97.33	14.05	85.70	64.95	82.62	66.24	80.37	70.05	86.50	53.82
	Energy [10]	81.09	73.89	84.48	74.97	80.55	82.04	63.59	96.40	77.43	81.82
	ReAct [11]	90.64	50.25	91.45	48.60	85.37	67.66	67.07	91.70	83.63	64.55
	GradNorm [25]	83.06	55.68	85.95	58.74	70.64	79.44	53.82	91.80	73.37	71.41
	Residual [13]	97.66	11.18	76.76	80.43	80.18	68.09	81.57	65.55	84.04	56.31
	ViM [13]	**98.92**	4.69	89.30	55.68	89.94	49.23	83.87	**61.50**	90.51	42.77
	k-NNuSA	98.31	**4.18**	**93.03**	47.39	**92.27**	**42.37**	**84.75**	64.20	**92.09**	**39.53**
ViT	MSP [8]	87.10	48.55	96.11	19.05	92.16	34.99	81.86	64.90	89.31	41.87
	MaxLogit [12]	93.01	30.58	98.56	6.57	96.72	16.59	89.85	44.10	94.53	24.46
	KL-Matching [12]	88.76	44.09	96.87	14.78	93.45	29.59	84.12	55.65	90.80	36.03
	Mahalanobis [9]	94.24	25.17	**99.54**	**2.12**	97.34	14.18	**92.81**	36.95	95.98	19.60
	Energy [10]	93.38	28.24	98.65	6.15	96.99	14.77	90.46	41.25	94.87	22.60
	ReAct [11]	93.34	28.49	99.00	4.30	97.14	14.68	90.70	42.60	95.04	22.52
	GradNorm [25]	89.68	34.48	97.32	8.58	93.76	21.03	80.28	50.90	90.26	28.75
	Residual [13]	92.21	33.82	98.56	6.63	91.87	36.37	88.23	47.85	92.72	31.17
	ViM [13]	95.34	**20.31**	99.41	2.61	97.30	14.33	92.55	**36.75**	96.15	**18.50**
	k-NNuSA	**95.76**	23.40	98.95	2.76	**97.47**	**14.17**	92.69	37.05	**96.23**	19.34

(1) The experiments take ImageNet-1K as the ID dataset and Texture, iNaturalList, OpenImage-O, ImageNet-O as the OOD datasets.
(2) AUROC and FPR95 are measured in percentage (%).
(3) The best performance is in **bold** and 2nd ones are underlined.

B. Experimental Settings

BiT (Big Transfer) [20] is an enhanced variant of the ResNet-v2 architecture, incorporating group normalization and weight standardization to improve performance. For CNN-based approach, we utilize the BiT-S model series, pretrained on the ImageNet-1K dataset, which is the BiT-S-R101×1 model on its officially released checkpoint. On the other hand, ViT (Vision Transformer) [22] represents a transformer-based approach to image classification, treating images as sequences of non-overlapping patches. For our analysis, we use the pretrained ViT-B/16 model, which has been fine-tuned on ImageNet-1K. We also report the performance of our method on other vision architectures, namely ResNet-50 [19], RepVGG [21], Swin [24] and DeiT [23]. All OOD algorithms in comparison do not require re-training the models.

For the setting k-NNuSA, we use $k = 10$ nearest neighbors to average the distance of image features to its neighborhood. The entropy $\mathcal{H}(\cdot)$ works with $\gamma = 0.1$ which is commonly set for softmax temperature [30], [31]. We configure other methods included in our experiments with the best settings mentioned in their papers.

C. Results and Discussions

In TABLE I, we summarize the results of OOD detection algorithms in ImageNet-1K benchmark using two vision models ViT-B/16 and BiT-S-R101×1. One can see that k-NNuSA achieves performance comparable to that of the state-of-the-art approaches in OOD detection.

a) BiT-S-R101×1: On average, our method k-NNuSA reaches 92.02% AUROC and 39.53% FPR95, which surpasses the second-performing method ViM by more than 2% and 3%, respectively. Over 5 of 8 comparisons across the four datasets, we achieve the largest AUROC and smallest FPR95, compared to other methods. In Eq. 4, k-NNuSA includes the combination of the residual and distance to nearest neighbors. TABLE I shows that we significantly outperform both Residual on all the datasets and Mahalanobis on three fourth of the datasets. Notably, iNaturalist has the least informative residual, since the average norm of residual is much less than that of residual on other datasets [13]. As expected, Residual struggles to detect iNaturalist images. Yet our proposed method which comes with the additional information on class-wise neighborhood alleviates the pitfall successfully and lands comparable performance to Mahalanobis. Interestingly, Mahalanobis utilizes a single source of distance, that is the minimum distance between the feature and the class centroids. Such a distance quite resembles our distance to class-wise neighborhood and explains the observation that the two methods can perform seamlessly on iNaturalist. Mahalanobis is, however, far behind k-NNuSA on the other datasets where we continue our domination thanks to the residual $\|x^{\mathbf{P}^\perp}\|$ and confidence scaling $\mathcal{H}(\mathbf{p})$. This result empirically verifies the effectiveness of our non-trivially distance-focused score $\mathcal{D}_{\mathbf{P}}$ in Eq. 4.

b) ViT-B/16: Since ViT-B/16 is pretrained on ImageNet-21K dataset, it offers much more rich contextual embeddings, compared to BiT. Hence, the performance of OOD detection algorithms relatively converge. Observing the results of ViM, Mahalanobis and k-NNuSA, we can see that these methods share similar detection capability across the datasets. Notably on average, both AUROC and FPR95 of all methods improves drastically thanks to that rich contextual embeddings. Despite performing poorly with BiT, Mahalanobis, a distance-based scoring function, surpasses the fused scoring function ViM on OpenImage-O, ImageNet-O and iNaturalist. Nevertheless,

TABLE II
ADDITIONAL RESULTS OF OOD DETECTION METHODS WITH REPVGG, RESNET-50D, SWIN TRANSFORMER AND DEIT.

Method	CNN				Transformer				Average	
	RepVGG		ResNet-50d		Swin		DeiT			
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP [8]	78.02	70.68	78.01	67.96	87.57	43.38	79.43	66.31	80.75	62.08
MaxLogit [12]	77.48	73.63	75.5	69.19	88.44	35.29	76.77	**64.43**	79.55	60.63
KL-Matching [12]	81.28	61.74	82.67	64.46	88.85	46.94	83.44	64.81	84.06	59.49
Mahalanobis [9]	85.93	59.72	88.12	56.43	92.1	40.81	85.07	72.8	87.80	57.44
Energy [10]	76.29	79.04	71.28	77.96	87.84	34.95	72.82	69.93	77.06	65.47
ReAct [11]	49.11	98.96	82.97	58.44	90.2	31.28	77.39	66.77	74.92	63.86
GradNorm [25]	52.97	95.03	44.04	96.03	41.65	84.53	32.1	97.44	42.69	93.26
Residual [13]	83.99	59.42	86.72	59.33	92.82	37.75	84.17	73.96	86.92	57.61
ViM [13]	87.66	**50.84**	89.03	53.36	94.08	31.23	85.27	69.75	89.01	51.29
k-NNuSA	**87.82**	51.35	**89.08**	**52.96**	**94.15**	**31.15**	**86.21**	68.76	**89.31**	**51.05**

(1) The experiments are in ImageNet-1K benchmark. Only the average AUROC and FPR95 are reported.
(2) AUROC and FPR95 are measured in percentage (%).
(3) The best performance is in **bold** and 2nd ones are underlined.

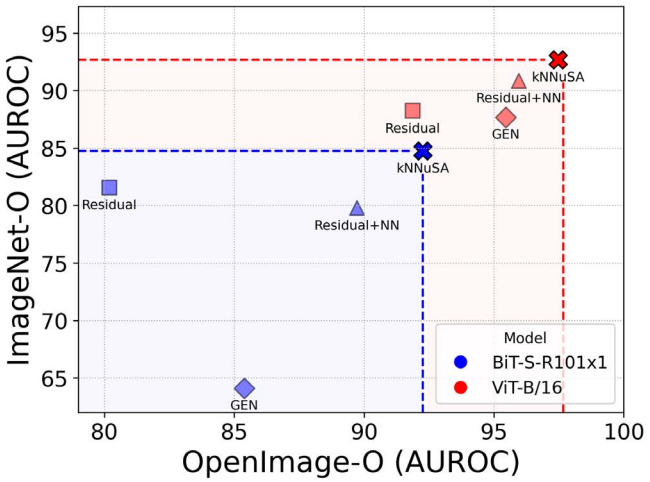

Fig. 3. ***Ablation study of k-NNuSA***. The unified scoring function in Eq. 5 (k-NNuSA) is compared with single components of it, including Residual ($\|x^{\mathbf{P}^\perp}\|$), GEN ($\mathcal{H}_{\mathbf{P}}$), Residual+NN ($\mathcal{D}_{\mathbf{P}}$). Reported in AUROC↑ (%). The blue marks represent the results of OOD detection with BiT and the red ones do with ViT. k-NNuSA stands out from its single components.

Mahalanobis fails shortly on Texture and OpenImage-O due to its lack of fine-graininess. Texture has a portion of images that are indistinguishable from ID samples [32] while OpenImage-O features have much more diverse distribution and semantics [13]. In such cases, classifier-based information is of importance to OOD detection; thus k-NNuSA scales its distances based on the confidence $\mathcal{H}(\mathbf{p})$ and varies following the energy of the prediction logsumexp. Empirically, our method remains competitive regardless of different data characteristics and feature quality.

c) Other Architectures: TABLE II demonstrates more results of ImageNet-1K benchmark on different recent model architectures. To be specific, we select two CNN-based representatives RepVGG [7] and ResNet-50d [11] and two transformer-based representatives Swin Transformer [26] and DeiT [33]. We report their average AUROC and FPR95 over the four OOD datasets. Our proposed method, k-NNuSA achieves the state-of-the-art performance regardless of the architectures used for feature space. On average, k-NNuSA achieves the largest AUROC 89.31% and the smallest FPR95 51.05% over all the OOD detection methods in comparison.

d) Ablation Study: Fig. 3 illustrates the comparison between the full-fledge k-NNuSA in Eq. 5 and the scoring functions that are parts of it. For the datasets, we involve OpenImage-O, which has diverse contexts and the least informative residual, and ImageNet-O, which has the largest norm of residual on average [13]. As expected, Residual and GEN perform poorly on one of the two datasets due to their singleness of information source; i.e. features and probabilities, respectively. We then add the distance to class-wise neighborhood $\mathcal{N}_{\mathbf{P}}$ which then works seamlessly on the diversity of OpenImage-O. However, Residual+NN appears less robust and falls behind with using solely residual (approx. 1%) on lower-quality features of BiT. We thus improves the fine-graininess by adding GEN, followed by the energy-based score, to Residual+NN and construct the outstanding k-NNuSA which remains robustness and achieves the best performance over the former regardless of feature spaces.

IV. CONCLUSION

In this work, we introduce a generalized distance-focused scoring function for OOD detection: Nearest Neighbors with Null Space Analysis (k-NNuSA). This method utilizes both distance to nearest neighbors and residual from the principal feature subspace of ID reference. By scaling the combination of these two distances based on classification confidence, k-NNuSA provides more fine-grained detection capability. We hold a series of extensive experiments across a diverse set of models and datasets, which demonstrates that k-NNuSA consistently outperforms existing methods, achieving the state-of-the-art performance with an average 96.23% AUROC and 19.34% FPR95 using ViT architectures. The results verify the efficiency of our non-trivially combined scoring function regardless of datasets and models.

ACKNOWLEDGMENT

This work was supported by the Camera Center of Viettel High Tech (Vietnam). We thank for the technical advices and resources provided during the research. The authors also acknowledge that Mrs. Trang Nguyen and Mr. Minh Nguyen Tong, named before the asterisk (∗), have contributed equally to this work as co-first authors.

REFERENCES

[1] X. Ren, T. Yang, L. E. Li, A. Alahi, and Q. Chen, "Safety-aware motion prediction with unseen vehicles for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 731–15 740.

[2] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road anomaly detection by partial image reconstruction with segmentation coupling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 651–15 660.

[3] S. N. Rai, F. Cermelli, D. Fontanel, C. Masone, and B. Caputo, "Unmasking anomalies in road-scene segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4037–4046.

[4] T. Vojíř and J. Matas, "Image-consistent detection of road anomalies as unpredictable patches," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5491–5500.

[5] P. Seeböck, S. M. Waldstein, S. Klimscha, H. Bogunovic, T. Schlegl, B. S. Gerendas, R. Donner, U. Schmidt-Erfurth, and G. Langs, "Unsupervised identification of disease marker candidates in retinal oct imaging data," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1037–1047, 2018.

[6] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.

[7] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of covid-19 lesions in lung ct," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2808–2819, 2021.

[8] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Hkg4TI9xl

[9] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[10] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.

[11] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[12] S. Basart, M. Mantas, M. Mohammadreza, S. Jacob, and S. Dawn, "Scaling out-of-distribution detection for real-world settings," in *International Conference on Machine Learning*, 2022.

[13] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4921–4930.

[14] M. Cook, A. Zare, and P. Gader, "Outlier detection through null space analysis of neural networks," *arXiv preprint arXiv:2007.01263*, 2020.

[15] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7068–7081, 2021.

[16] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 827–20 840.

[17] I. Ndiour, N. Ahuja, and O. Tickoo, "Out-of-distribution detection with subspace techniques and probabilistic modeling of features," *arXiv preprint arXiv:2012.04250*, 2020.

[18] J. Park, Y. G. Jung, and A. B. J. Teoh, "Nearest neighbor guidance for out-of-distribution detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1686–1695.

[19] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 558–567.

[20] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.

[21] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[25] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," *Advances in Neural Information Processing Systems*, vol. 34, pp. 677–689, 2021.

[26] X. Liu, Y. Lochman, and C. Zach, "Gen: Pushing the limits of softmax-based out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 946–23 955.

[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[28] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[29] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 262–15 271.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[31] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[32] R. Huang and Y. Li, "Mos: Towards scaling out-of-distribution detection for large semantic space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8710–8719.