IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 March 2024, accepted 17 April 2024, date of publication 22 April 2024, date of current version 9 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391808

RESEARCH ARTICLE

C2I-CAT: Class-to-Image Cross Attention Transformer for Out-of-Distribution Detection

JAEHO CHUNG[1,*], SEOKHO CHO[2,*], HYUNJUN CHOI[1], DAEUNG JO[3], YOONHO JUNG[1], AND JIN YOUNG CHOI[1], (Member, IEEE)

[1]Automation and Systems Research Institute (ASRI), Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea
[2]Medintech, Seoul 03100, South Korea
[3]Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, Suwon 16678, South Korea

Corresponding author: Jin Young Choi (jychoi@snu.ac.kr)

This work was supported by the Institute for Information & communication Technology Planning & evaluation (IITP) grant funded by Korean Government [Ministry of Science, ICT (MSIT)], Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis, under Grant B0101-15-0266, and the Artificial Intelligence Graduate School Program (Seoul National University) under Grant 2021-0-01343.

*Jaeho Chung and Seokho Cho contributed equally to this work.

ABSTRACT In our work, we have empirically found that Vision Transformer (ViT) could not extract object-centric features when applied to out-of-distribution (OOD) detection. To make object-centric attention, we design an additional module that employs a cross-attention between class-wise token proxy and feature token sequence of an input image. For inference suitable to our cross-attention structure with multiple class-wise token proxies, we propose a score ensemble that can be applied to any scoring function. Compared to ViT, the proposed inference scheme achieves outperforming performance by synergizing with our cross-attention structure. Through experiments, we demonstrate that the proposed cross-attention structure with score ensemble inference improves largely near OOD detection performance, where FPR95 improvement in near OOD detection compared to the state-of-the-art method becomes 2.55% for CIFAR-10 and 2.67% for CIFAR-100, keeping competitive classification accuracy.

INDEX TERMS Near out-of-distribution (OOD) detection, vision transformer, class-wise cross attention.

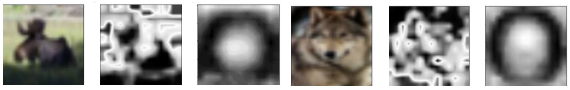
I. INTRODUCTION

Advancements in deep learning have demonstrated outstanding technological improvements, such as residual learning [1], large-scale image data learning [2], transformers [3], and resistant models to adversarial attack [4], in various fields, whereas the advanced AI techniques have created innovative products in data processing and analysis, such as SpectralGPT [5] and cross-city semantic segmentation system [6]. Recently, out-of-distribution (OOD) detection attracts attention in safety-critical fields such as military defense, system safety, autonomous driving, and surveillance.

The baseline approach for OOD detection focuses on proposing scoring functions or training schemes. Score-based methods measure the likelihood of how far a given sample originates from in-distribution (ID) ([7], [8], [9], [10], [11], [12]). Training-based methods are based on representation learning in embedding space ([13], [14]) or reduce overconfidence in a network [15] by normalizing model's logits during training. Thus, OOD performance of these methods based on both scoring functions and training heavily depends on the quality of the features or model's outputs (i.e., logits) obtained from the trained model.

To obtain highly representative features, a vision transformer (ViT) [3] has been employed as a backbone for OOD detection ([16], [17]). Although the self-attention in ViT can effectively capture the global context of an image, it does not pay attention to a target region as depicted in Figure 1, where the self-attention (the middle column of each subfigure in Figure 1) forms ambiguous attention map on the overall image, including both the foreground object

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.



(a) ID of CAFAR-10 (b) OOD of CAFAR-100

FIGURE 1. Illustrations of the attention map visualizations of the class tokens. In each sub-figure, the left columns are the original images of the corresponding datasets, the middle columns depict ViT's attention map, and the right columns show the proposed method's attention map.

and background. In contrast, the proposed proxy-feature cross-attention scheme forms an object-centric attention map (the right column of each subfigure in Figure 1), which is important because object's features are a key factor in distinguishing ID and OOD.

Motivated from our observations, we propose a novel cross-attention module to focus on object-centric information for OOD detection, referred to as a class-to-image cross attention transformer (C2I-CAT). The conventional cross-attention ViT (i.e., CrossViT [18]) uses a dual branch to learn multi-scale representations of an image. In contrast, our C2I-CAT utilizes class-wise token proxies and ViT's feature token sequences to pay attention to an object of an input image.

To this end, we introduce a class-wise token proxy as an additional input to focus on an object of a specific class during OOD detection, without the need for any additional branch. By averaging feature token sequences in each class, the class-wise token proxy can serve as a representative for each class. By utilizing the class-wise token proxy as the criterion, the model can capture object-centric features and learn the correlation between images and classes.

In addition, we newly introduce an inference method designed to fit our cross-attention structure using class-wise token proxies. Since the label of a test sample is unavailable during inference, we utilize all the class-wise token proxies to obtain the test sample's result. When considering our structure and an inference process, we leverage the ensemble of output scores for all class-wise token proxies. The proposed inference scheme can directly utilize the outputs of specific components, such as a classifier or a penultimate layer, without the need for modifying scoring functions. This inference method can be applied to various scoring functions for OOD detection. We summarize our contributions as follows.

- We design a new cross attention transformer (C2I-CAT) suitable for OOD detection. To the best of our knowledge, our work is the first attempt to apply a cross-attention mechanism for OOD detection. Our C2I-CAT leverages the proposed cross-attention module, using class-wise token proxies, to learn the correlation between the feature tokens of images and classes. As a result, C2I-CAT detects OOD samples in various OOD datasets by focusing on object-centric features.
- We introduce a novel inference method suitable for our cross-attention mechanism utilizing class-wise token proxies. The proposed inference scheme, synergizing with our cross-attention structure, shows outstanding OOD performance compared to the self-attention-based method. In addition, our inference method can be applied to various scoring functions without modification.
- We validate the effectiveness of the proposed method through extensive experiments on various OOD datasets, including both far and near OOD cases. The proposed method outperforms the state-of-the-art method in the near OOD case without significant degradation of classification accuracy.

II. RELATED WORKS

A. SCORE BASED OUT-OF-DISTRIBUTION DETECTION

MSP [7] is the baseline paper that proposes the framework of OOD detection for the first time and presents a maximum softmax probability (MSP) score. MaxLogit [9] improves OOD performance by using maximum logit value, compared to the MSP scoring function. Mahalanobis distance (MD) [10] proposes a distance-based scoring function that detects OOD samples using Mahalanobis distance. Energy [8] proposes a new scoring function that is called energy score which is based on the Energy-based Model (EBM) [19]. In addition to proposing an energy score, the authors improve OOD performance by using an outlier fine-tuning scheme. kNN [11] proposes an OOD detection method based on k-Nearest Neighbor (kNN) with supervised contrastive learning [20]. ViM [12] suggests a scoring function by utilizing feature and logit space information on a large scale.

B. VISION TRANSFORMER BASED OUT-OF-DISTRIBUTION DETECTION

Recently, [21] has demonstrated that the transformer-based model (i.e., Vision Transformer (ViT) [3]) is effective for capturing global information. Since ViT's self-attention module takes into account the correlation between all local patches in an image, it extracts features that effectively reflect the global context of the image. Therefore, OODformer [16] and Exploring [17] have achieved state-of-the-art performance by utilizing the features of ViT. These works have shown that using ViT-based features is more effective than CNNs for OOD detection. In particular, they show the outstanding ability of the ViT in near OOD detection and use the MD scoring function. RMD [22] proposes a distance-based method called relative mahalanobis distance (RMD) scoring function for near OOD detection.

C. TRANSFORMER-BASED ADVANCED MODELS

Since the advent of ViT [21], attention mechanism-based methods have been employed in various fields to enhance performance in specific tasks, such as classification and segmentation. CrossViT [18] is a dual-branch transformer model applying cross-attention. In each branch, token sequences are extracted from image patches of different sizes, and the class tokens of the token sequences are exchanged. Then, cross-attention is performed between a class token of

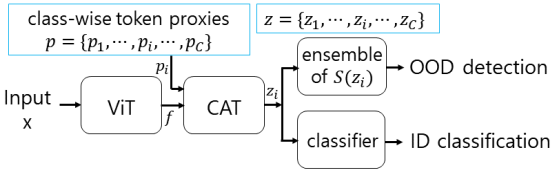


FIGURE 2. Entire structure for the proposed method. $S(\cdot)$ means a scoring function and p is class-wise token proxies, where C is the number of classes and p_i is i-th class token proxy. z_i represents the class token in the output tokens of CAT for p_i and f.

a branch (i.e., small or large branch) and the other branch's token sequence, enabling the transformer to learn multi-scale feature representation.

SpectralGPT [5] proposes a foundation model based on a self-attention-based encoder and a benchmark dataset for handling spectral data in the remote sensing field. The proposed method addressed the issues of limited spectral data extraction and utilization, enabling the recognition of objects or scenes in remote sensing applications. HighDAN [6] proposes a model using an adversarial domain adaptation module for cross-city semantic segmentation tasks, along with multi-modal remote sensing benchmark datasets. This method not only tackled the challenges in multiple urban environments with spatio-temporal and regional changes but also improved the generalization ability of the semantic segmentation across regions.

III. PROPOSED METHOD

In this section, we explain our method for out-of-distribution (OOD) detection, named **Class-to-Image Cross Attention Transformer (C2I-CAT)** depicted in Figure 2. First, we describe the process of extracting class-wise token proxies, which are fed into C2I-CAT (Section III-A). Then, we provide the details of C2I-CAT (Section III-B). Finally, we explain the training scheme for C2I-CAT and the inference method for OOD detection (Section III-C).

A. CLASS-WISE TOKEN PROXIES

The key idea of the proposed C2I-CAT is to detect OOD samples by focusing on object-centric features via the proposed cross-attention module. As shown in Figure 2, considering that the proposed method requires two inputs for cross-attention, we introduce *feature token sequence* and *class-wise token proxy*. The *feature token sequences* contain all token features for input images and are extracted from the penultimate layer (i.e., before a classifier) of vision transformer (ViT).

The *class-wise token proxies* for an in-distribution (ID) dataset are defined by $p = \{p_1, p_2, \cdots, p_C\}$, where C is the number of classes in ID and p_i is i-th class-wise token proxy. p_i is obtained by token-wise averaging all sample features in the i-th class, that is,

$$p_i = \{p_{i,1}, \cdots p_{i,t}, \cdots, p_{i,T}\}, \quad p_{i,t} = (1/N)\Sigma_{j=1}^{N} f_{i,t,j}, \quad (1)$$

where N is the number of training samples for i-th class and T is the number of tokens, whereas $f_{i,t,j}$ represents the t-th feature token of the j-th sample in the i-th class. The p_i can be regarded as the representative token proxy for i-th class in an ID dataset. Note that when obtaining the class-wise token proxies, we do not use a test set, but use a train set of the ID dataset.

As a result, the dimension of p_i is (1, T, D), and the dimension of p becomes (C, T, D), where D represents the feature dimension for each sample. The class-wise token proxies (p) are imputed to the proposed C2I-CAT along with the *feature token sequence* of each input sample in the train or test set.

B. CROSS ATTENTION TRANSFORMER

1) OVERALL ARCHITECTURE

As depicted in Figure 2, we use ViT as a feature extractor because the feature token sequences extracted by ViT contain abundant global information. As mentioned in Section III-A, our C2I-CAT has two inputs. One is a token proxy and the other is a feature token sequence extracted by ViT. When determining two inputs for applying cross-attention, it is basically considered to employ cross-attention between two sampled feature token sequences of ViT. However, we do not consider cross-attention between sampled feature token sequences by using a sampling method. The performance may vary depending on the sampling method used to select feature token sequences. Furthermore, there is a challenge in selecting the token proxy of a class between two sampled feature token sequences. To be more specific, it is difficult to determine a suitable feature token sequence to be the class-wise token proxy for cross-attention with the other token sequence.

For these reasons, we use feature token sequences and token proxies that average feature token sequences for each class as inputs. The difference from the existing ViT is that the proposed method does not require position embedding and patch embedding. As a result, we introduce a method of cross-attention between feature token sequences and class-wise token proxies.

2) CROSS-ATTENTION BLOCK

The structure of the cross-attention block (CAB) is depicted in Figure 3. Similar to ViT, CAB is composed of layer normalization, multi-head cross-attention module, residual connection, and feed-forward layer.

In addition to the difference mentioned in subsection III-B1, another difference is cross-attention module. While current ViT-based OOD detection methods use self-attention, we utilize a cross-attention module, and the details for the cross-attention module are explained in subsection III-B3. The other components except for the cross-attention module are applied in a similar way as existing ViT and transformer [23] components.

The proposed CAB consists of alternating layers of multi-head cross-attention module and feed-forward. The layer normalization is applied to inputs of CAB and feed-forward

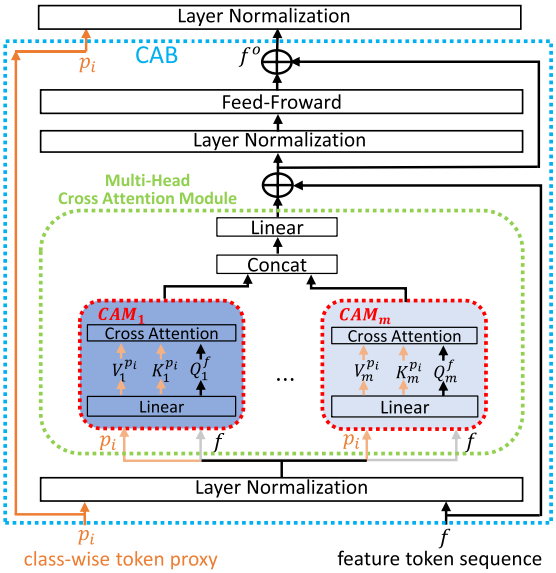


FIGURE 3. Cross-Attention Block (CAB) in the proposed cross attention transformer. The feature token sequence is generated by the trained ViT for an input image. CAM_m means m-th single-head cross-attention module. Multi-head cross-attention module and m-th single-head cross-attention module are indicated by a dotted green line and red line, respectively. CAT in Figure 2 is composed of layered 12 CABs. f^o represents the output feature token sequence of CAB and becomes an input to the following block.

layer. We employ residual connection around multi-head cross-attention module and feed-forward layer. In order to utilize residual connections like ViT and transformer, all the sub-layers in the CAB produce outputs with the same dimension. The residual connections are effective in learning the residual, which prevents the gradient vanishing problem. The feed-forward layer is performed to represent the attention result that is passed through the multi-head cross-attention module.

As shown in Figure 3, the class-wise token proxy is fed into every CAB's input. The feature token sequence extracted by ViT is forwarded to the first CAB in our model. Then, the output for the cross-attention block becomes an input to the following cross-attention block.

3) CROSS ATTENTION MODULE

Multi-head cross-attention module and single-head cross-attention module are shown in Figure 3. In the case of m-th single-head cross-attention module, query Q_m^f is extracted from the feature token sequence. Key $K_m^{p_i}$ and value $V_m^{p_i}$ are extracted from the class-wise token proxy through a linear layer in m-th single-head cross-attention module (CAM_m).

Considering Figure 3, let f be a feature token sequence that is input to the cross-attention block and p_i be the class-wise token proxy corresponding to the label of f. Then, f is linearly mapped to the query Q_m^f and p_i is linearly mapped to the key $K_m^{p_i}$ and value $V_m^{p_i}$ in the CAM_m. Each of query, key, and value is a matrix since the inputs are a token sequence. Cross-attention is performed in a similar way to self-attention [23], except that the cross-attention is employed

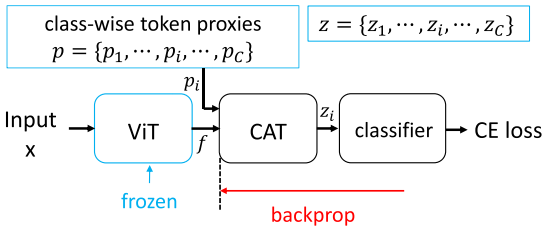


FIGURE 4. Training scheme for the proposed C2I-CAT. CE loss is a cross entropy loss. The class token proxy corresponding to the label of input is forwarded to CAT along with the feature token sequence of the input.

between two sequences. The expression for m-th single-head cross-attention module is given by

$$CAM_m(f, p_i) = Softmax(\frac{Q_m^f (K_m^{p_i})^T}{\sqrt{d}}) V_m^{p_i}, \quad (2)$$

where $\sqrt{d}$ is dimension of Q_m^f and $K_m^{p_i}$. As shown in Eq. (2), the similarity between the class-wise token proxy (p_i) and feature token sequence (f) is calculated by scaled dot product [23].

In the case of the multi-head cross-attention module, it can be expanded from m-th single-head cross-attention module. In a similar way as multi-head self-attention [23], the expression for the multi-head cross-attention module is described in Eq.(3).

$$MultiHeadCAM(f, p_i) = Concat(CAM_1(f, p_i), \cdots, CAM_M(f, p_i))W_{lin}, \quad (3)$$

where M is the number of cross-attention modules (i.e., the number of heads) and W_{lin} is the weight matrix of the linear layer.

After passing through layer normalization, the output token sequence of the multi-head cross-attention module is fed into the feed-forward layer. Therefore, the cross-attention block extracts the final output token sequence reflecting the class proxy information.

C. TRAINING AND INFERENCE

1) TRAINING SCHEME

As shown in Figure 4, we design a training method to learn target class relationships using the cross-attention module. We train C2I-CAT using a supervised manner that can utilize labels of training data. For simplicity, we consider that a training batch size is 1. Letting f be the feature tokens of a training sample with i-th label, then f and p_i are fed into our C2I-CAT.

After that, we train C2I-CAT using cross entropy (CE) loss between the classifier outputs and the target labels. At this time, the first token (i.e., class token) among the output tokens of CAT is used as input to the classifier. Note that we do not additionally train ViT, which is pre-trained on an ID dataset (e.g., CIFAR-10 or CIFAR-100 [24]) during training C2I-CAT.

Therefore, the model learns which part of input tokens for C2I-CAT should be focused more by referring to the given

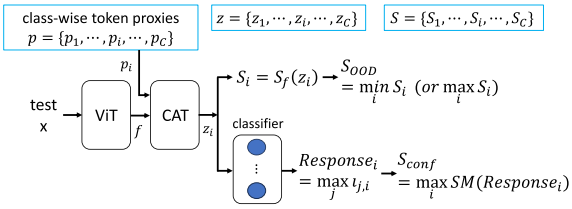


FIGURE 5. Score ensemble inference for the class-wise cross-attention structure. z_i is the class token in the output tokens of CAT inputted by p_i, where p_i denotes i-th class token proxy. $S_f(\cdot)$ represents a scoring function and S_i means i-th OOD score for z_i. S_{OOD} is the final OOD score for the test input. When obtaining S_{OOD}, the min (or max) operation depends on the scoring function. $\iota_{j,i}$ is the j-th logit of the classifier output vector for z_i. S_{conf} means a confidence score to evaluate classification performance. SM(·) is the softmax function for an input.

class-wise token proxy. Furthermore, this training process can be extended to arbitrary batch size.

2) INFERENCE SCHEME

We newly introduce an inference scheme referred to as a *score ensemble inference* method, which is suitable for a scenario of cross-attention with class-wise token proxies. During inference for OOD detection and ID classification, since the labels of test samples are unknown, we provide all the class-wise token proxies as shown in Figure 5. C2I-CAT performs cross-attention between a feature token sequence (denoted by f) of a test sample and every class-wise token proxy (denoted by p_i). In the same way as C2I-CAT training scheme, we evaluate OOD detection and ID classification performance using only the class token among the output tokens of CAT.

For evaluating ID classification, we use the highest softmax score among all outputs of the classifier for every class-wise token proxy. To be more specific, let ι_i be the classifier's logit vector for i-th class token proxy, $F(\cdot)$ be the classifier, and $CAT(\cdot)$ be the proposed module. Then, $\iota_i = F(z_i)$, where $z_i = CAT(f, p_i)$. After taking the maximum of ι_i, we use the maximum value as the resulting response (denoted by $Response_i$ in Figure 5) for the i-th class token proxy. Since the maximum logit value can be considered as a representative result for the corresponding class token proxy, we use the maximum value as the response result. When the classifier generates outputs for all class-wise token proxies, the outputs form a matrix with the dimension of (batch size, C), where C is the number of classes. Subsequently, the maximum values applying softmax to the outputs are used as the final confidence scores (i.e., S_{conf} in Figure 5) of ID samples, and the dimension of S_{conf} becomes (batch size, 1). Using the result of softmax on the outputs, C2I-CAT predicts the classes of the test samples.

In OOD inference, the process also performs similarly to ID classification. For testing OOD detection performance, we utilize Mahalanobis distance (MD) scoring function to determine whether the test sample is OOD or not. As shown in Figure 5, we put z_i as the input to a scoring function (i.e., $S_f(\cdot)$) and then obtain OOD scores for z_i. After obtaining OOD scores for all class-wise token proxies, we calculate the final OOD scores using min or max operation. Using min or max operation depends on a scoring function. For example, we use the max operation for MSP scoring function and the min operation for MD scoring function. Although we mainly utilize MD scoring function, this inference scheme can be applied to various scoring functions. Note that we do not use noise like [10] for MD scoring function.

IV. EXPERIMENTS

In this section, we show our experimental results. In IV-A section, we explain datasets, evaluation metrics, model structure, and implementation details. In IV-B section, we report our results for structure variants, inference scheme, and visualizations. In IV-C section, we compare our method with ViT-based methods, including the state-of-the-art (SOTA) method, for near and far OOD detection tasks. Note that, for a fair comparison, we only compare with ViT-B/16-based methods, including the SOTA method using R50-ViT-B/16 model. In addition, when implementing the experiments, we fix all randomness factors for fair comparison and evaluation (i.e., fix all random seeds). We implement 5 independent training runs, setting random seeds from 0 to 4.

A. EXPERIMENTAL SETUP

1) IN-DISTRIBUTION DATASETS

We use CIFAR-10 and CIFAR-100 [24] as in-distribution (ID) dataset. When training a model, we use a standard split with 50,000 training images and 10,000 test images. All ID images are resized to 224 × 224.

2) OUT-OF-DISTRIBUTION DATASETS

For near out-of-distribution (OOD) detection task, we use CIFAR [24] dataset as OOD dataset. In other words, if CIFAR-10 is ID dataset, then we use CIFAR-100 as OOD dataset and vice versa. For far OOD datasets, we use SVHN [25], LSUN (resize/crop) [26], Texture [27], Places365 [28], iSUN [29], iNaturalist [30], SUN [31], STL10 [32], MNIST [33], K-MNIST [34], and fashion-MNIST [35]. ODIN [26] authors constructed LSUN (resize/crop) by resizing and cropping LSUN [36] dataset to 32 × 32. For MNIST family datasets (i.e., MNIST, K-MNIST, and fashion-MNIST), we follow MOOD [37] setting. In addition to these OOD datasets, we also validate our proposed method on synthetic data. To be more specific, we utilize Gaussian noise (σ=0.5), Rademacher noise, and Blob. We follow OE [38] settings for these synthetic data. All OOD images are resized to 224 × 224.

3) EVALUATION METRICS

We mainly use the false positive rate at true positive rate 95% (FPR95) and area under the receiver operating characteristic curve (AUROC). We also utilize a supportive evaluation metric that is area under the precision-recall curve (AUPR). AUROC and AUPR metrics show binary classification performance and higher these values indicate

TABLE 1. Variation results of attention structure on CIFAR-10 (ID). We denote self-attention as SA (i.e., ViT-B/16 [3]) and cross-attention as CA (i.e., our C2I-CAT). Training time [s] indicates one epoch training time for the corresponding structure in the measurement of seconds. We measure one epoch training time for each seed and then average the measurements for all seeds. Class tokens refer to the use of class tokens from class-wise token proxies (i.e., p_i) and feature token sequences (i.e., f), while all tokens refer to the use of entire tokens from p_i and f. We use MD scoring function and every structure consists of 12 layers. We report mean and std based on 5 independent training runs. The bolded result is the best result.

Structure	# Param.*	Training time [s]	ID Acc.↑	Near OOD result			Far OOD result		
				FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
SA† [3]	85.81M	$162.27_{\pm 0.11}$	$98.67_{\pm 0.05}$	$7.79_{\pm 0.25}$	$98.34_{\pm 0.05}$	$98.28_{\pm 0.07}$	$5.64_{\pm 0.16}$	$98.39_{\pm 0.08}$	$98.48_{\pm 0.07}$
CA†	85.08M	$201.33_{\pm 0.27}$	$98.47_{\pm 0.17}$	$4.34_{\pm 0.11}$	$98.95_{\pm 0.04}$	$98.87_{\pm 0.05}$	$\mathbf{3.44}_{\pm 0.15}$	$99.25_{\pm 0.03}$	$99.38_{\pm 0.02}$
CA‡		$91.09_{\pm 0.95}$	$98.43_{\pm 0.20}$	$4.65_{\pm 0.49}$	$98.90_{\pm 0.08}$	$98.82_{\pm 0.08}$	$3.75_{\pm 0.11}$	$99.16_{\pm 0.03}$	$99.30_{\pm 0.03}$
SA==CA†	106.31M	$255.97_{\pm 0.54}$	$98.60_{\pm 0.05}$	$4.23_{\pm 0.13}$	$98.97_{\pm 0.03}$	$98.89_{\pm 0.04}$	$3.52_{\pm 0.08}$	$99.24_{\pm 0.03}$	$99.38_{\pm 0.03}$
SA==CA‡		$114.49_{\pm 1.74}$	$98.56_{\pm 0.08}$	$4.55_{\pm 0.19}$	$98.93_{\pm 0.05}$	$98.84_{\pm 0.06}$	$3.48_{\pm 0.18}$	$99.25_{\pm 0.03}$	$99.39_{\pm 0.02}$
CA==SA†	106.31M	$259.84_{\pm 0.97}$	$98.67_{\pm 0.05}$	$4.30_{\pm 0.11}$	$98.99_{\pm 0.02}$	$\mathbf{98.91}_{\pm 0.02}$	$3.57_{\pm 0.06}$	$99.24_{\pm 0.03}$	$99.38_{\pm 0.04}$
CA==SA‡		$116.70_{\pm 0.74}$	$98.59_{\pm 0.07}$	$4.28_{\pm 0.12}$	$98.99_{\pm 0.03}$	$\mathbf{98.91}_{\pm 0.03}$	$3.55_{\pm 0.07}$	$99.25_{\pm 0.04}$	$99.39_{\pm 0.04}$
SA‖CA†	170.15M	$350.01_{\pm 1.15}$	$98.74_{\pm 0.09}$	$4.39_{\pm 0.19}$	$98.94_{\pm 0.03}$	$98.85_{\pm 0.04}$	$3.74_{\pm 0.07}$	$99.19_{\pm 0.03}$	$99.34_{\pm 0.04}$
SA‖CA‡		$137.06_{\pm 1.19}$	$\mathbf{98.76}_{\pm 0.06}$	$\mathbf{4.20}_{\pm 0.14}$	$\mathbf{99.00}_{\pm 0.03}$	$\mathbf{98.91}_{\pm 0.04}$	$3.52_{\pm 0.11}$	$\mathbf{99.28}_{\pm 0.04}$	$\mathbf{99.42}_{\pm 0.03}$

∗: "# Param." refers to the number of trainable model parameters for each attention structure, measured in millions (M).
†: All tokens are used for inputs to the structure. ‡: Class tokens only are used for inputs to the structure.
"SA==CA" denotes serial connection in order of SA and CA. "CA==SA" denotes serial connection in order of CA and SA.
"SA‖CA" denotes parallel connection of SA and CA, and their outputs are concatenated.

that a model performance is better. FPR95 is a strict threshold-based measurement when comparing performance with other models and the smaller this value is, the better model performance is. These metrics are widely used for evaluating OOD detection performance.

4) MODEL STRUCTURE

We use a ViT-B/16 as a baseline model for a feature extractor of our C2I-CAT and comparison methods (e.g., ViM or OODformer). The ViT-B/16 has 12 layers, a feature dimension of 768, an input image size of 224×224, an image patch size of 16×16, and a token sequence length of 197 by adding a class token to $224/16 \times 224/16 = 196$ tokens. The number of head is 12. Our C2I-CAT has a structure (e.g., the number of layers, feature dimension, the number of tokens) similar to the ViT-B/16, except for the attention mechanism and model's input. We denote our model as C2I-CAT-B/16-12. C2I-CAT refers to our cross attention transformer, B/16 denotes the feature extractor (ViT-B/16), and 12 represents the number of layers.

5) IMPLEMENTATION DETAILS

We use an ImageNet [39] pre-trained ViT-B/16 [3] as a feature extractor for our C2I-CAT and fine-tune the ViT-B/16 on ID datasets. When fine-tuning the ViT on ID datasets, we train the model for 50 epochs using cross entropy loss. We set the initial learning rate of 0.001 with Cyclic learning rate scheduler [40], batch size of 32, weight decay of 0.0001, dropout rate of 0.1, and SGD optimizer with momentum 0.9. For training C2I-CAT, we train 10 and 15 epochs for CIFAR-10 and CIFAR-100, respectively. We set the initial learning rate of 0.001 with a Cosine learning scheduler [41], batch size of 32, weight decay of 0.0001, and dropout rate of 0. We also use SGD optimizer with momentum 0.9 for CIFAR-10 and momentum 0.95 for CIFAR-100. We do not use any data augmentation.

We re-implemented the comparison methods except for Exploring[1] [17]. When re-implementing other methods, we search hyper-parameters from the papers and publicly available Git-Hub codes. For OODformer [16], we fine-tune an ImageNet [39] pre-trained ViT-B/16 on ID datasets for 50 epochs, using cross entropy loss. The initial learning rate is 0.01 with Cyclic learning rate scheduler [40]. We set batch size of 32, weight decay of 0, dropout rate of 0.1, and SGD with momentum 0.9. For ViM [12] and RMD [22], they evaluate on a large-scale OOD (i.e., ImageNet-1k is ID) or do not specify training hyper-parameters. Therefore, we evaluate their OOD performance on ViT-B/16 model trained with OODformer settings.

In addition, we follow all hyper-parameters by the proposed methods such as k for the kNN scoring function and temperature (*Temp*) for the Energy scoring function. To be more specific, we set k=50 for CIFAR-10 (ID) and k=200 for CIFAR-100 (ID) in all experiments. For the Energy scoring function, we set *Temp*=1 for all experiments. Note that we do not use any auxiliary data to train the models for all experiments. Furthermore, we utilize softmax function to evaluate the model's ID classification accuracy for all experiments.

B. ANALYSIS OF THE PROPOSED METHOD

1) VARIATION RESULTS OF ATTENTION STRUCTURE

Table 1 shows the results of structural variants in attention modules and input types for CIFAR-10 (ID). SA means the self-attention structure (i.e., ViT-B/16 [3]) and CA represents the cross-attention structure (i.e., our C2I-CAT). == indicates a serial connection from the left one and ‖ indicates a parallel connection. Like CA structure, the inputs of the serial and parallel structures are feature token sequences of ViT and class-wise token proxies.

[1]The Git-Hub code has an issue when attempting to reproduce the method. Furthermore, the authors only specify hyper-parameters for outlier training, while not addressing hyper-parameters for training on ID datasets.

TABLE 2. Variation results of inference scheme on CIFAR-10 (ID). We compare three variations of proposed inference scheme. We use MD scoring function and all tokens for all inference methods. We report mean and std based on 5 independent training runs. The bolded result is the best one.

Inference Method	Ensemble	Near OOD result			Far OOD result		
		FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
Layer-average Concatenation	Feature	$91.70_{\pm 1.30}$	$51.10_{\pm 1.56}$	$52.68_{\pm 1.41}$	$95.34_{\pm 1.29}$	$47.47_{\pm 6.42}$	$59.33_{\pm 3.92}$
Layer-score Ensemble	Feature/Score	$92.29_{\pm 1.63}$	$55.09_{\pm 1.95}$	$54.54_{\pm 1.67}$	$91.11_{\pm 4.04}$	$61.33_{\pm 3.70}$	$66.05_{\pm 3.36}$
Feature Concatenation	Feature	$93.16_{\pm 4.06}$	$49.85_{\pm 8.35}$	$49.75_{\pm 6.80}$	$93.78_{\pm 7.56}$	$45.40_{\pm 15.05}$	$56.82_{\pm 8.39}$
Proposed Ensemble	Score	$\mathbf{4.34}_{\pm 0.11}$	$\mathbf{98.95}_{\pm 0.04}$	$\mathbf{98.87}_{\pm 0.05}$	$\mathbf{3.44}_{\pm 0.15}$	$\mathbf{99.25}_{\pm 0.03}$	$\mathbf{99.38}_{\pm 0.02}$

"Proposed Ensemble": minimum confidence score among class-wise scores on the features in the penultimate layer.
"Feature Concatenation": confidence score of the concatenated feature for all class-wise features in the penultimate layer.
"Layer-score Ensemble": minimum confidence score among scores for all layers, obtained by "Feature Concatenation" for each layer.
"Layer-average Concatenation": confidence score of the concatenated feature for all class-wise averaged features through all layers.

Regarding input types, we conduct structure variation experiments on two input types. One uses all tokens (i.e., $p_i = \{p_{i,t}\}_{t=1}^{t=T}$ and $f = \{f_t\}_{t=1}^{t=T}$) and the other utilizes only class tokens (i.e., $p_{i,1}$ and f_1 in p_i and f) for inputs to all structure variants, excluding SA.

When compared to SA (i.e., ViT-B/16 [3]), all of our structural variants significantly improve OOD performance, particularly in terms of FPR95. When considering the training time, all of our structures with only ''class tokens'' require less training time compared to SA structure. Among all variants, our SA‖CA structure with only ''class tokens'' shows outstanding improvements in both near OOD performance and classification accuracy, demonstrating enhancements of 0.09% and 3.59% for ID accuracy and FPR95, respectively. In addition, our CA structure with only ''class tokens'' still shows remarkable OOD improvements in both near and far OOD tasks while requiring the least training time among all our structural variants and SA. When compared to SA, our CA structure with only ''class tokens'' enhances FPR95 performance by 3.14% and by 1.89% for near and far OOD cases, respectively.

Regarding the number of model parameters, our CA has slightly fewer parameters than SA, while the other structures combining SA and CA have more parameters. Although more model parameters lead to increased training time, OOD performance is significantly improved compared to SA. In addition, despite having a similar number of model parameters as SA, our CA still demonstrates remarkable OOD performance compared to SA.

Therefore, considering training time, model parameters, and improvements in OOD performance, our CA with only ''class tokens'' is effective in OOD detection compared to SA. On the other hand, SA‖CA with only ''class tokens'' achieves the best OOD performance but requires longer training time and more parameters.

2) VARIATION RESULTS OF INFERENCE SCHEME

Table 2 illustrates the comparison between the proposed class-wise proxy score ensemble and feature-level ensemble variants that are based on feature concatenation over all classes instead of class-wise score ensemble. As shown in Table 2, scoring the concatenated feature over all class is not useful for OOD detection, compared to the proposed class-wise score ensemble.

The reason is that ''Proposed Ensemble'' method calculates the score of an input based on each class-wise proxy and ensembles these scores to determine the final score, which can clearly distinguish ID and OOD samples by the precisely fitted in-distribution boundary from multiple class-wise proxy kernels. That is, OOD input images can be easily discriminated from the kernel of the nearest class proxy determined by the proposed ensemble. However, in the case of concatenated feature, the features of other classes within the concatenated feature affect scoring, giving a rough in-distribution boundary based on one kernel in high-dimensional space and so degrading OOD detection performance. Therefore, ensemble of scores for all class proxies is more suitable for our structure than feature-level ensemble methods.

Table 3 and Table 4 show influence of scoring function type in our inference scheme, comparing with ViT-B/16 model. When evaluating OOD performance for our method and ViT-B/16, we utilize widely used scoring functions based on the model's output (i.e., logit) or feature space. In other words, we use scoring functions based on logit space (MSP, Energy, MaxLogit), feature space (kNN, MD), and a combination of feature and logit space (ViM).

As demonstrated in Table 3, the proposed inference method (i.e., score ensemble inference) can be applied to various scoring functions. Compared to ViT-B/16 in each scoring function result, our C2I-CAT significantly improves near OOD performance across all scoring functions. Considering the average result over all scoring functions, we improve the performance by 2.51% (FPR95 average), 0.83% (AUROC average), and 4.52% (AUPR average) for CIFAR-10 (ID), compared to ViT-B/16. We also improve OOD performance for CIFAR-100 (ID).

Table 4 shows the far OOD results. Similar to the near OOD results in Table 3, the proposed inference method is still effective for far OOD datasets, considering the average results over all scoring functions. Compared to ViT-B/16, our C2I-CAT improves far OOD performance by 1.01% (FPR95 average), 0.73% (AUROC average), and 4% (AUPR average) for CIFAR-10 (ID). In addition, we also achieve improvements in OOD performance for CIFAR-100 (ID).

TABLE 3. Near OOD detection results of scoring functions on CIFAR (ID). ↓ (or ↑) indicates that the smaller (or bigger) the value, the better the performance. We report mean and std based on 5 independent training runs. The bolded result is the best result.

Score	Model	Near OOD result on CIFAR-10 (ID)			Near OOD result on CIFAR-100 (ID)		
		FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP	ViT-B/16 [3]	11.32±0.27	96.72±0.20	76.29±2.17	46.54±0.98	90.80±0.08	89.70±0.22
	C2I-CAT-B/16-12 (Ours)	**7.37**±0.29	**98.15**±0.13	**97.83**±0.37	**19.93**±3.81	**95.49**±0.42	**95.50**±0.40
Energy	ViT-B/16 [3]	8.87±0.41	96.97±0.21	95.42±0.47	26.50±0.44	93.82±0.16	92.82±0.41
	C2I-CAT-B/16-12 (Ours)	**7.78**±0.55	**97.93**±0.17	**97.41**±0.37	**17.76**±4.02	**95.85**±0.49	**95.73**±0.46
MaxLogit	ViT-B/16 [3]	8.88±0.40	96.96±0.21	95.41±0.47	26.97±0.45	93.75±0.15	92.78±0.41
	C2I-CAT-B/16-12 (Ours)	**7.78**±0.55	**97.93**±0.17	**97.41**±0.37	**17.76**±4.01	**95.85**±0.49	**95.73**±0.46
MD	ViT-B/16 [3]	7.79±0.25	98.34±0.05	98.28±0.07	28.33±0.86	94.18±0.24	94.22±0.27
	C2I-CAT-B/16-12 (Ours)	**4.34**±0.11	**98.95**±0.04	**98.87**±0.05	**16.06**±0.56	**96.29**±0.14	**96.11**±0.16
kNN	ViT-B/16 [3]	7.59±0.27	98.49±0.05	98.47±0.06	35.44±1.07	91.50±0.38	91.12±0.50
	C2I-CAT-B/16-12 (Ours)	**4.99**±0.32	**98.91**±0.07	**98.85**±0.10	**23.62**±1.28	**94.44**±0.21	**93.99**±0.23
ViM	ViT-B/16 [3]	7.47±0.20	98.32±0.06	98.15±0.08	29.81±0.81	94.23±0.17	94.37±0.21
	C2I-CAT-B/16-12 (Ours)	**4.59**±0.13	**98.87**±0.05	**98.78**±0.08	**16.78**±0.41	**96.21**±0.11	**96.13**±0.13
Average	ViT-B/16 [3]	8.65	97.63	93.67	32.27	93.05	92.50
	C2I-CAT-B/16-12 (Ours)	**6.14**	**98.46**	**98.19**	**18.65**	**95.69**	**95.53**

TABLE 4. Far OOD detection results of scoring functions on CIFAR (ID). ↓ (or ↑) indicates that the smaller (or bigger) the value, the better the performance. We report mean and std based on 5 independent training runs. The results are averaged over all OOD datasets (15 datasets). The bolded result is the best result.

Score	Model	Far OOD result on CIFAR-10 (ID)			Far OOD result on CIFAR-100 (ID)		
		FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑
MSP	ViT-B/16 [3]	9.11±0.34	96.58±0.17	77.08±2.29	44.86±4.51	90.46±0.68	89.04±0.81
	C2I-CAT-B/16-12 (Ours)	**7.41**±0.42	**97.89**±0.14	**97.92**±0.24	**31.17**±2.40	**92.62**±0.56	**92.72**±0.82
Energy	ViT-B/16 [3]	**6.94**±0.20	97.34±0.15	97.28±0.19	**24.42**±1.49	93.12±0.79	91.41±0.89
	C2I-CAT-B/16-12 (Ours)	7.40±0.64	**97.76**±0.25	**97.71**±0.37	28.65±2.43	**93.29**±0.59	**93.43**±0.78
MaxLogit	ViT-B/16 [3]	**6.95**±0.21	97.33±0.15	97.28±0.19	**24.84**±1.49	93.04±0.78	91.36±0.89
	C2I-CAT-B/16-12 (Ours)	7.40±0.64	**97.76**±0.25	**97.71**±0.37	28.64±2.43	**93.29**±0.59	**93.43**±0.78
MD	ViT-B/16 [3]	5.64±0.16	98.39±0.08	98.48±0.07	20.38±1.37	95.63±0.40	95.76±0.36
	C2I-CAT-B/16-12 (Ours)	**3.44**±0.15	**99.25**±0.03	**99.38**±0.02	**17.53**±0.69	**96.21**±0.24	**96.38**±0.21
kNN	ViT-B/16 [3]	6.34±0.23	98.02±0.07	98.11±0.09	27.25±1.03	93.22±0.47	93.11±0.42
	C2I-CAT-B/16-12 (Ours)	**4.97**±0.20	**98.76**±0.04	**98.93**±0.03	**25.25**±0.96	**93.82**±0.31	**94.04**±0.24
ViM	ViT-B/16 [3]	4.86±0.10	98.73±0.06	98.90±0.07	18.88±1.58	96.14±0.44	96.29±0.42
	C2I-CAT-B/16-12 (Ours)	**3.14**±0.13	**99.32**±0.04	**99.44**±0.04	**16.09**±0.79	**96.67**±0.28	**96.86**±0.26
Average	ViT-B/16 [3]	6.64	97.73	94.52	26.77	93.60	92.83
	C2I-CAT-B/16-12 (Ours)	**5.63**	**98.46**	**98.52**	**24.56**	**94.32**	**94.48**

TABLE 5. Near OOD detection results on CIFAR (ID). ↓ (or ↑) indicates that the smaller (or bigger) the value, the better the performance. Except for exploring [17], we reimplement all other methods. We use the results of Exploring, which are reported in the paper [17]. We report mean and std based on 5 independent training runs. The bolded result is the best.

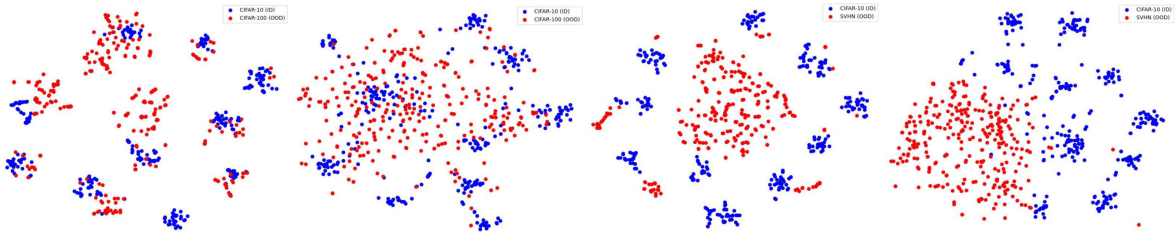
ID dataset	Method	Model / Score	ID Acc.↑	FPR95↓	AUROC↑	AUPR↑
CIFAR-10	RMD [22]	ViT-B/16 / RMD	98.67 ±0.05	9.35 ±0.27	97.75 ±0.06	97.68 ±0.07
	ViM [12]	ViT-B/16 / ViM	98.67 ±0.05	7.47 ±0.20	98.32 ±0.06	98.15 ±0.08
	OODformer [16]	ViT-B/16 / MD	98.67 ±0.05	7.79 ±0.25	98.34 ±0.05	98.28 ±0.07
	Exploring [17]	R50+ViT-B/16 / MD	**98.70**	6.89	98.52	98.70
	C2I-CAT (Ours)	C2I-CAT-B/16-12 / MD	98.47 ±0.17	**4.34** ±0.11	**98.95** ±0.04	**98.87** ±0.05
CIFAR-100	RMD [22]	ViT-B/16 / RMD	**92.01** ±0.13	34.28 ±0.48	92.75 ±0.14	93.31 ±0.16
	ViM [12]	ViT-B/16 / ViM	**92.01** ±0.13	29.81 ±0.81	94.23 ±0.17	94.37 ±0.21
	OODformer [16]	ViT-B/16 / MD	**92.01** ±0.13	28.33 ±0.86	94.18 ±0.24	94.22 ±0.27
	Exploring [17]	R50+ViT-B/16 / MD	91.71	18.73	96.23	**96.32**
	C2I-CAT (Ours)	C2I-CAT-B/16-12 / MD	91.40 ±0.08	**16.06** ±0.56	**96.29** ±0.14	96.11 ±0.16

From these results, our inference method synergizing with our cross-attention structure outperforms the OOD performance of the self-attention-based method. When considering an inference process and our structure, the ensemble of output scores obtained from specific components, such as a classifier or a penultimate layer, is more suitable for our structure and OOD detection than feature-based ensemble methods.

Furthermore, our inference method can be applied to various scoring functions without modification.

3) QUALITATIVE RESULTS

Figure 6 shows the penultimate layer's features of each model (i.e., C2I-CAT and ViT-B/16) in 2D embedding space. While ViT does not detect near OOD samples well, our C2I-CAT robustly identifies near OOD samples. For ViT, the near OOD samples are more closely located around the boundaries of ID samples than our C2I-CAT, considering that near OOD detection is a challenging task. For the far OOD case, the result of ViT shows that far OOD samples are widely spread.



(a) Near OOD result of C2I-CAT (Ours) (b) Near OOD result of ViT-B/16 (c) Far OOD result of C2I-CAT (Ours) (d) Far OOD result of ViT-B/16

FIGURE 6. t-SNE visualization of CIFAR-10 (ID). We visualize t-SNE results for near and far OOD cases. We use CIFAR-100 and SVHN as the near and far OOD datasets, respectively. Blue dots indicate ID samples and red dots represent OOD samples. We randomly select 300 samples from each dataset. For all models, we extract class token features from the penultimate layer (i.e., before a classifier) of a model.

TABLE 6. Far OOD detection results on CIFAR (ID). ↓ (or ↑) indicates that the smaller (or bigger) the value, the better the performance. We report mean and std based on 5 independent training runs. The results are averaged over all OOD datasets (15 datasets). The model for all comparison methods is ViT-B/16. The bolded result is the best.

ID dataset	Method	FPR95↓	AUROC↑	AUPR↑
CIFAR-10	RMD [22]	7.72 $\pm$0.14	97.16 $\pm$0.08	97.51 $\pm$0.11
	ViM [12]	4.86 $\pm$0.10	98.73 $\pm$0.06	98.90 $\pm$0.07
	OODformer [16]	5.64 $\pm$0.16	98.39 $\pm$0.08	98.48 $\pm$0.07
	C2I-CAT (Ours)	**3.44** $\pm$0.15	**99.25** $\pm$0.03	**99.38** $\pm$0.02
CIFAR-100	RMD [22]	36.46 $\pm$2.42	91.93 $\pm$0.48	93.16 $\pm$0.42
	ViM [12]	18.88 $\pm$1.58	96.14 $\pm$0.44	96.29 $\pm$0.42
	OODformer [16]	20.38 $\pm$1.37	95.63 $\pm$0.40	95.76 $\pm$0.36
	C2I-CAT (Ours)	**17.53** $\pm$0.69	**96.21** $\pm$0.24	**96.38** $\pm$0.21

However, C2I-CAT still discriminates ID and OOD samples more clearly than ViT in the far OOD case.

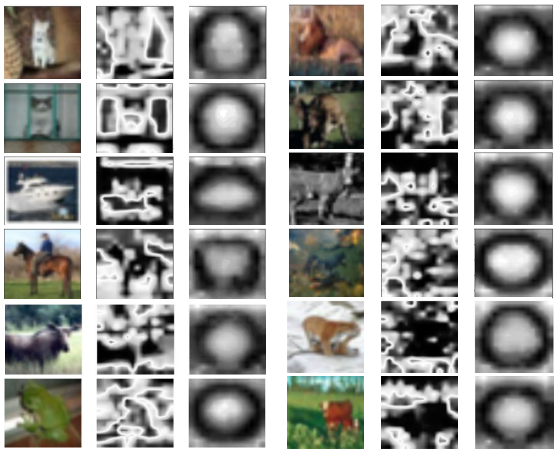
In addition to Figure 1, we provide additional attention map visualizations for CIFAR-10 (ID) and CIFAR-100 (OOD) in Figure 7. Therefore, the visualization results indicate that focusing on the object-centric features is a key factor in detecting OOD samples. Furthermore, our C2I-CAT, utilizing class-wise token proxies, captures object-centric information more effectively than ViT-B/16, demonstrating that our structure extracts the correlation between the feature tokens of images and classes.

C. COMPARISON WITH STATE-OF-THE-ART METHOD

Table 5 and Table 6 show the results of a comparison with other methods, including the SOTA methods (i.e., Exploring [17] and ViM [12]). Note that our primary goal is near OOD detection.

As shown in Table 5, our C2I-CAT demonstrates superior performance in terms of FPR95 and AUROC, keeping competitive ID classification accuracy. When compared to OODformer [16] that uses a ViT-B/16 model and the same scoring function, our C2I-CAT improves OOD performance by 3.45% (FPR95) and 0.61% (AUROC) for CIFAR-10 (ID). For CIFAR-100 (ID), our C2I-CAT also improves OOD performance by 12.27% (FPR95) and 2.11% (AUROC). In addition, compared to the SOTA method (i.e., Exploring), our C2I-CAT notably improves FPR95 by 2.55% and 2.67% for CIFAR-10 (ID) and CIFAR-100 (ID), respectively.

In addition to near OOD detection, we also compare our method with other methods for far OOD detection, as shown in Table 6. Similar to near OOD results, the proposed C2I-CAT improves far OOD performance when compared to ViT-B/16-based other methods. Compared to OODformer [16], our method achieves FPR95 improvements of 2.2% and 2.85% for CIFAR-10 (ID) and CIFAR-100 (ID), respectively. When compared to ViM [12] that is the best result for far OOD detection, we improve FPR95 by 1.42% and 1.35% for CIFAR-10 (ID) and CIFAR-100 (ID), respectively.



(a) ID attention map (b) OOD attention map

FIGURE 7. Attention map visualizations of the class tokens. Figure 7a and Figure 7b are the visualization results for CIFAR-10 (ID) and CIFAR-100 (OOD), respectively. In each sub-figure, the left columns are the original images, the middle columns depict ViT's attention map, and the right columns represent the proposed method's attention map.

Therefore, based on these experimental results, our C2I-CAT learns more informative features by extracting object-centric features via the cross-attention module, compared to the self-attention module (i.e., ViT-B/16). In addition, our C2I-CAT effectively detects outlier samples in various OOD datasets, which demonstrates outstanding robustness for OOD detection.

V. CONCLUSION

In this paper, we have proposed a new cross attention transformer, namely C2I-CAT, for OOD detection. Unlike existing ViT-based OOD methods, we have introduced a newly designed cross-attention module that employs a cross-attention between class-wise token proxy and feature token sequence of an input image. The proposed structure extracts object-centric features, which are a key factor in

discriminating ID and OOD samples. For inference suitable to our cross-attention structure with multiple class-wise token proxies, we have suggested a score ensemble that can be applied to any scoring function. Through experiments, we have demonstrated that our inference method can be applied to various scoring functions and outperforms ViT's OOD performance by synergizing with our cross-attention structure.

A. LIMITATION AND FUTURE RESEARCH

OOD detection is a classification task, where a localized image including only the target object (trained or untrained) should be given. To apply OOD detection to actual environments, it is essential to localize untrained objects from a natural scene. However, the localization of an untrained (unknown) object is a challenging task because most object detection algorithms (YOLO, etc.) mainly detect trained objects. As the future research for military and social purposes, an undefined foreground object (UFO) detection is required where unknown object localization and OOD detection tasks are tackled at the same time. For instance, in a military coastal security system, AI should localize an unknown object in a wide range of coastal scenes and determine whether it is a suspicious object (OOD) or not. In a social environment such as autonomous driving, UFO detection in 3D scenes is essential for highly safe driving even when undefined objects appear.

B. SOCIETAL IMPACT

Regarding societal impact, it is crucial to identify an invasion of an unidentified object or person that is not anticipated. However, actual applications of OOD detection to real-world environments are still limited because the performance of unidentified object localization along with OOD detection is not satisfactory. Thus, the low technical level might lead to harmful situations, such as enemy infiltration within the military or major accidents involving self-driving cars, etc. However, if the technical level of UFO detection increases to a satisfying level via future research, it can be applied to real-world scenarios to enhance human safety.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.
[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp.'1–7.
[3] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
[5] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," 2023, *arXiv:2311.07113*.
[6] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.
[7] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2016, *arXiv:1610.02136*.
[8] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21464–21475.
[9] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," 2019, *arXiv:1911.11132*.
[10] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[11] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," 2022, *arXiv:2204.06507*.
[12] H. Wang, Z. Li, L. Feng, and W. Zhang, "ViM: Out-of-distribution with virtual-logit matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4911–4920.
[13] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," 2021, *arXiv:2103.12051*.
[14] Y. Ming, Y. Sun, O. Dia, and Y. Li, "How to exploit hyperspherical embeddings for out-of-distribution detection?" 2022, *arXiv:2203.04450*.
[15] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 23631–23644.
[16] R. Koner, P. Sinhamahapatra, K. Roscher, S. Gunnemann, and V. Tresp, "OODformer: Out-of-distribution detection transformer," 2021, *arXiv:2107.08976*.
[17] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.
[18] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.
[19] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting Structured Data*, vol. 1, pp. 1–54, Aug. 2006.
[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.
[21] C. Zhang, M. Zhang, S. Zhang, D. Jin, Q. Zhou, Z. Cai, H. Zhao, X. Liu, and Z. Liu, "Delving deep into the generalization of vision transformers under distribution shifts," 2021, *arXiv:2106.07617*.
[22] J. Ren, S. Fort, J. Liu, A. Guha Roy, S. Padhy, and B. Lakshminarayanan, "A simple fix to Mahalanobis distance for improving near-OOD detection," 2021, *arXiv:2106.09022*.
[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," M.S. thesis, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2009.
[25] 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. NIPS*, 2011, p. 7.
[26] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2017, *arXiv:1706.02690*.
[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.
[28] B. Zhou, A. 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.
[29] 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*.

[30] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8769–8778.
[31] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. CVPR*, Jun. 2010, pp. 3485–3492.
[32] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[33] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[34] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," 2018, *arXiv:1812.01718*.
[35] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.
[36] 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*.
[37] Z. Lin, S. D. Roy, and Y. Li, "MOOD: Multi-level out-of-distribution detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15308–15318.
[38] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.
[40] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.
[41] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

JAEHO CHUNG received the B.S. degree in electrical and information engineering from Korea University, Sejong-si, South Korea, in 2021, and the M.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2023. His current research interests include computer vision, machine learning, and anomaly detection.

SEOKHO CHO received the B.S. degree in mechanical engineering from Inha University, Incheon, Republic of Korea, in 2020, and the M.S. degree in interdisciplinary program of artificial intelligence from Seoul National University, Seoul, Republic of Korea, in 2022. His research interests include machine learning, anomaly detection, and object detection.

HYUNJUN CHOI received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2017, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His current research interests include 3D object detection, autonomous driving, and anomaly detection.

DAEUNG JO received the B.S. and Ph.D. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2016 and 2022, respectively. He is currently a Staff Engineer with Samsung Electronics, Kyeong-Gi, South Korea. His research interests include machine learning, computer vision, the architecture of deep learning, and the autonomous driving.

YOONHO JUNG is currently pursuing the B.S. degree with the Department of Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea. His research interests include out-of-distribution and anomaly detection, reinforcement learning, and robotics and computer vision.

JIN YOUNG CHOI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1982, 1984, and 1993, respectively. From 1984 to 1989, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, where he was involved in the Project of Switching Systems. From 1992 to 1994, he was with the Basic Research Department, ETRI, where he was a Senior Member of Technical Staff involved in the neural information processing system. Since 1994, he has been with Seoul National University, where he is currently a Professor with the School of Electrical Engineering. From 1998 to 1999, he was a Visiting Professor with the University of California at Riverside, Riverside, CA, USA. He is also with the Automation and Systems Research Institute, Engineering Research Center for Advanced Control and Instrumentation, and the Automatic Control Research Center, Seoul National University. His current research interests include adaptive and learning systems, visual surveillance, motion pattern analysis, object detection, object tracking, and pattern recognition.