

A^2Pt : Anti-Associative Prompt Tuning for Open Set Visual Recognition

Hairui Ren , Fan Tang , Xingjia Pan , Juan Cao , Weiming Dong , *Member, IEEE*, Zhiwen Lin , Ke Yan ,
and Changsheng Xu , *Fellow, IEEE*

Abstract—Multi-modality pre-trained models (PTMs) have considerably boosted the performance on a broad range of computer vision topics. Still, they have not been explored purposefully in open set recognition (OSR) scenarios when applying PTMs to downstream recognition tasks. Directly fine/prompt tuning PTMs on closed-set classification tasks will inevitably suffer from data bias and always learn more or less target class-irrelevant co-occurring contextual information, which leads to over-confident predictions on unknown samples. In this paper, we propose a simple yet effective approach, termed Anti-Associative Prompt Tuning (A^2Pt), toward learning compact and accurate class-related representation with few class-irrelevant associations from context using multi-modal priors. Specifically, a cross-modal guided activation module is adopted to refine the class-aware representation and suppress the associations from co-occurring contexts by involving text-modal information. We further design an anti-association calibration module to obtain compact class-aware and class-irrelevant representations, respectively, by introducing two additional object functions. Extensive experiments on publicly available benchmarks, including CIFAR series, TinyImageNet, and ImageNet-21K-P, show that the proposed A^2Pt achieves substantial and consistent performance gains compared with both SOTA OSR and PTM prompt tuning approaches.

Index Terms—Multi-modality Pre-trained models (PTMs), open set recognition (OSR), class-aware representation, anti-associative prompt tuning (A^2Pt).

I. INTRODUCTION

DESPITE the remarkable progress in advanced data-driven image understanding approaches, the conventional closed-set training/testing paradigm limits further applications of

intelligent models to real-world occasions. A typical closed-set classifier optimized by cross-entropy objective is by nature not designed for open set recognition (OSR) where the testing inputs may not belong to any predefined historical category. Large-scale pre-trained models (PTMs) have shown potential for zero-shot/open-world image understanding. However, fine/prompt tuning PTMs on downstream tasks or specific datasets further limits the performance of open set recognition.

Previous OSR methods have tried to fit the distributions of features for known classes based on pretrained networks and then reject novel classes [1], [2]. Serving as passive detectors for outliers, the performance of these post-processing approaches is limited since the decision is based on the features from the model trained on closed set samples. By contrast, researchers have focused on positive OSR approaches to learning more efficient features by additional constraints [3] or generating pseudo-unknown data [4], [5], [6]. Although encouraging results were obtained, the recent study by Vaze et al. [7] reveals that most of the passive/positive OSR methods are suboptimal and somewhat go around in circles. By carefully choosing hyper-parameters on different datasets for training cross-entropy models, a significant improvement could be observed by simply removing logit normalization when testing.

Such observation motivates us to rethink how traditional OSR methods work. No matter whether passive or positive, the critical aspect of open set recognition is to find and enlarge the differences between known classes and novel classes. However, the above-mentioned recognition pipeline trained on closed set images is usually confused by co-occurrence contexts. Such co-occurrence contexts, also known as prior or biases, could not undermine the close set performance since the model could learn to reduce incorrect co-occurrence. Whereas when applying to the open set setting, the remaining irrelevant contexts could lead to a strong confusion between open set and closed set categories. As shown in Fig. 1(a), as one of the SOTA methods of conventional OSR approaches, AGC [7] mistakenly classifies a ‘torch’ as a closed set class of a ‘neck brace’ with a confidence level of 81% due to the presence of human association observed on both the torch and the neck brace. Additionally, as a typical model of PTMs, CLIP [8] also misclassifies ‘torch’ as a ‘neck brace’ associated with the human body, and CoOp [9] confidently recognizes ‘torch’ as ‘binoculars’ associated with eyes. As shown the Fig. 1(b), both traditional methods and PTMs face the same association issue. We aim to develop a model that

Manuscript received 29 January 2023; revised 3 June 2023 and 24 October 2023; accepted 30 November 2023. Date of publication 5 December 2023; date of current version 31 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62102162, U20B2070, 61832016, and 61832002, in part by Beijing Natural Science Foundation under Grant L221013. The Guest Editor coordinating the review of this manuscript and approving it for publication was Prof. Nicu Sebe. (*Corresponding author: Fan Tang.*)

Hairui Ren is with the School of Artificial Intelligence, Jilin University, Jilin 130012, China (e-mail: renhr22@mails.jlu.edu.cn).

Fan Tang and Juan Cao are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100045, China (e-mail: tangfan@ict.ac.cn; caojuan@ict.ac.cn).

Xingjia Pan, Zhiwen Lin, and Ke Yan are with the Youtu Lab, Tencent, Shanghai 201103, China (e-mail: xjia.pan@gmail.com; xavier.lin@foxmail.com; kerryan@tencent.com).

Weiming Dong and Changsheng Xu are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100045, China (e-mail: weiming.dong@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

Code is available at <https://github.com/Hrren/A2Pt>.

Digital Object Identifier 10.1109/TMM.2023.3339387

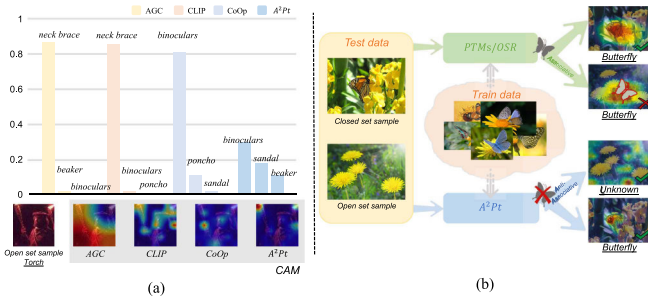


Fig. 1. Illustrations for open set associations. (a) Shows the top three predictions for the input open set samples by different methods. AGC misclassifies the ‘torch’ as a closed set class ‘neck brace’ with high confidence due to the CAM to the human neck in the torch, allowing the model to associate it with the neck in the neck brace. The same misclassification problem occurs for the PTMs models CLIP and CoOp. CLIP associates with a neck brace due to the cam of the body, and CoOp associates with ‘binoculars’ due to the cam of the eyes. However, our A^2Pt can refuse to focus on regions prone to association, and output a more even confidence for each closed set class, correctly recognizing the image into open sets. (b) Further indicates the problem of association that causes the phenomenon of misclassification, while our A^2Pt can suppress the influence of association and improve the model’s open set recognition.

effectively eliminates such irrelevant co-occurrence contexts, such as ‘flowers’ in the ‘butterfly’ class, and improve the open set recognition performance of the network.

The irrelevant co-occurrence, or closed-set association, is hard to get rid of without additional information. In this paper, we propose to involve multi-modal PTM priors for OSR tasks to enlarge the differences between known classes and novel classes. Learning directly from the raw text about images has shown its breakthrough potential, which leverages a much broader source of supervision [10], [11], [12], [13]. The pretrained multi-modal encoder aligns images and texts in a common feature space [9] and thus can be used to recognize the category of images by calculating the similarities between aligned multi-modal features. When adapting the PTMs on different downstream datasets, the choice of text prompts plays an important role. Zhou et al. [9] proved that learning contextual text prompts for different datasets could improve the closed-set classification accuracy. However, the adapting between closed-set and open set samples remains unsettled.

Inspired by recent advances in prompt learning, we introduce a novel **Anti-Associative Prompt Tuning** strategy, namely A^2Pt , to tuning pretrained networks. Our A^2Pt contains a cross-modal guided activation (CGA) module to eliminate closed set association and correspondingly an anti-association calibration (AAC) module is designed with cross-entropy and two additional loss functions to supervise the class-related target feature and association feature generated from CGA. We conduct extensive experiments on the CIFAR series [14], Tiny-ImageNet [15] and ImageNet-21K-P [16]. Results show that PTMs perform better than traditional OSR approaches, and the proposed A^2Pt obtain significant gains compared with SOTA prompt tuning methods. The contributions of this study include:

- We unveil that the associations caused by the irrelevant co-occurrence context lead to confusion for conventional OSR approaches.

- We propose a multi-modal A^2Pt approach consisting of a cross-modal guided activation module and an anti-association calibration module to eliminate the associations for precisely unknown class recognition.
- Our proposed approach shows consistent and substantial gains across CIFAR, TinyImageNet, and ImageNet-21K-P datasets for refusing the novel unknown categories.

II. RELATED WORK

A. Open Set Recognition.

Open set recognition refers to refusing unknown classes while maintaining the ability to recognize known classes [1], [2], [17], [18], [19], [20], [21], [22]. Commonly popular methods include passive approaches and positive approaches. Passive approaches focused on maintaining the feature space and recognizing unknown classes by fitting coexisting features or decision boundaries to the output distribution of known class data. Open-Max [1] utilizes Weibull distribution to fit the distribution of unknown classes with anomalous data of known classes. C2AE [2] model reconstruction errors with the Extreme Value Theory (EVT) of statistical modeling to find the threshold for recognizing known/unknown class samples. CGDL [23] detects unknown samples and classifies known samples by forcing different latent features to approximate different Gaussian models. GFROS [3] urges the network to recognize how the image has been transformed to help the network learn accurate feature representations. CVAECapOSR [24] involved a capsule network to learn more compact features. Recently, AGC [7] promotes the network to achieve SOTA only by adjusting the normal cross-entropy training of closed set samples. Different from passive approaches, the positive approach [3], [6], [25] is more concerned with the additional restriction of the feature space or generating open set samples through GAN (Generative Adversarial Networks) [25] or autoencoders. G-OpenMax [4] attempts to train a classifier on the synthesis low-confidence images generated by noise as open set data. GCM-CF [5] generated counterfactual instances that are close to the real world by class-agnostic features between different classes. ARPL [6] generates reciprocal points to squeeze the feature space of closed sets to learn more compact and accurate features. The latest approach DIAS [26] generates fakes with diverse difficulty levels to simulate the real world. The above two mainstream models have achieved competitive results but ignored that the misclassified samples are often related to the co-occurrence context in the closed set, which caused the association. And in this paper, we propose a method to tackle this problem with cross-modal information and large-scale pretrained models.

B. Prompt Learning.

Recently prompt tuning [27], [28], [29], [30], [31], [32], [33] was proposed to make the most multi-modal knowledge adapt to the downstream task. Compared with the fine-tune paradigm, prompt can achieve competitive performance with fewer optimization parameters. With task-specific continuous vectors, CoOp [9] achieves SOTA on 11 datasets in

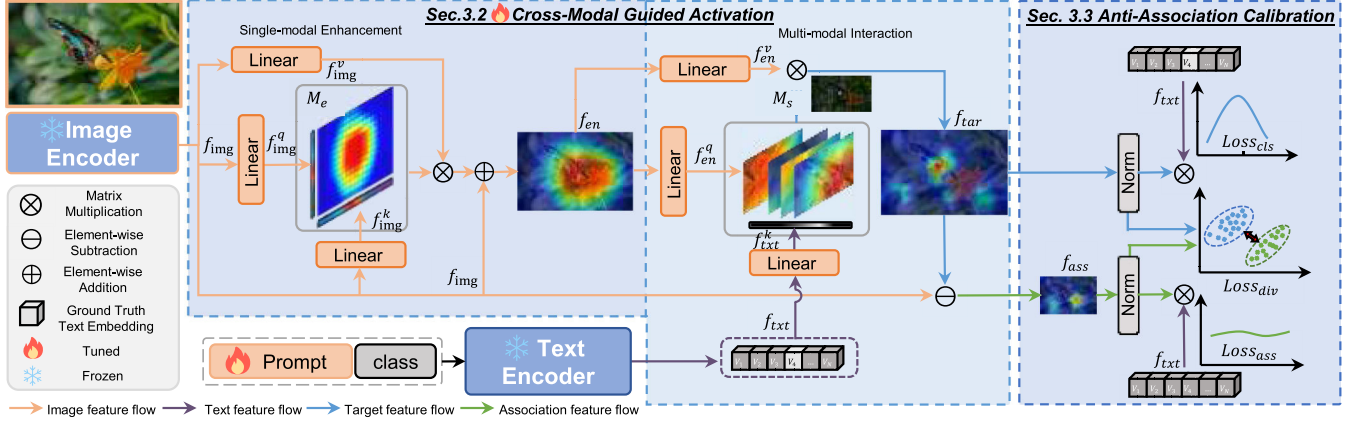


Fig. 2. Overview of our proposed anti-associative prompt tuning A^2Pt . A^2Pt consists of two parts: Cross-modal guided activation (CGA) and anti-association calibration (AAC). CGA is composed of modules with single-modal enhancement and multi-modal interaction. Single-modal enhancement is achieved by applying a self-attention mechanism to the image feature f_{img} . Multi-modal interaction is achieved by applying a cross-attention mechanism to the enhancement image feature f_{en} and text feature f_{txt} . AAC is a calibration of the association completed by the composition of three loss functions. Only parameters with ‘fire’ are tuned in the training phrase, with ‘snow’ being frozen.

zero-shot and few-shot settings. CoCoOp [34] alleviates the generalization challenge of CoOp by adopting Meta-Net, thus obtaining a trade-off performance between base classes and novel classes. Furthermore, DetPro [35] proposes a context grading scheme to optimize individual prompts for open vocabulary detection. CRIS [29] introduces prompt learning to propagate fine-structured textual features into pixel-level features for image segmentation. TPT [36] tunes the prompt on the fly using only the given test sample without requiring any task-specific training data or annotations. Wei et al. [30] proposes a unified hair editing framework by leveraging the powerful image text representation. R-Tuning [37] (Regularized prompt Tuning) mitigates the label bias of prompt learning in OSR, it extends the range of words forming the texts from only closed set classes to more to simulate open set scenarios. Instead of focusing on expanding the input label names to enhance open set recognition performance, we keep the label of the input unchanged and improve the network structure from the perspective of eliminating association. Apart from the above text prompts, there is also some work to try to help the network ‘understand’ downstream from the perspective of the visual prompt. VPT [38] achieves significant performance gains with two different visual prompt embeddings inserted ways as VPT-Shallow and VPT-Deep. Besides, Bar et al. [39] investigates visual prompting by giving input-output image examples to adapt a pre-trained visual model to novel downstream tasks. Different from these works, we explore prompt learning in a novel anti-associative perspective, which can promote the model to remove the irrelevant associations for open set recognition.

III. OUR APPROACH

To eliminate the irrelevant co-occurrence context for open set recognition, we propose anti-associative prompt tuning (A^2Pt) built on cross-modal pertaining models [8]. Fig. 2 shows the overall framework, which contains two parts: a cross-modal guided activation (CGA) module and an anti-association calibration (AAC) module. The CGA consists of the single-modal

enhancement module, which uses the self-attention mechanism, and the multi-modal interaction module, which uses the cross-attention mechanism. We start from introducing the overall prompt tuning pipeline in Section III-A. Then we move on to the details of CGA in Section III-B, which aims to expose the high response of the target area with single-modal enhancement and suppresses the expression of association precisely through multi-modal interaction. Finally, we show how AAC could cooperate with CGA to further calibrate the generated feature with carefully designed loss functions in Section III-C.

A. Problem Definition

When training classification models, a set of data of known classes $\mathcal{D}_k = \{1, \dots, K\}$ is fed to the model. When testing, additional unknown classes $\mathcal{D}_u = \{1, \dots, U\}$ are involved. The goal of OSR methods is to recognize testing samples in $\mathcal{D}_k$ classes while rejecting the rest in $\mathcal{D}_u$.

As suggested by CoOp [9], we take advantage of the multi-modal pretrained model and propose to optimize textual prompts by cross-entropy loss on the downstream dataset. We inherit the practice of CoOp and random initialize the prompt as a text vector for given class c as follows:

$$P_k = [v_1, v_2, \dots, v_l] + [v_k], \quad (1)$$

where l denotes the length of the shared prompts among different classes and v_k indicates the class token of class $k \in \mathcal{D}_k$, which is fixed during optimization. Then P_k will pass through text encoder E_T to get the text feature f_{txt} as:

$$f_{txt} = E_T(P_k), \quad (2)$$

similar to text feature, given an image input img , image feature f_{img} will be obtained by image encoder E_I as:

$$f_{img} = E_I(img). \quad (3)$$

Then f_{txt} and f_{img} will be sent to the later module for optimization.

B. Cross-Modal Guided Activation

Single-modal enhancement: For the purpose of generating the most valid target feature, we first enhance the expression of the f_{img} to expose the high response of the target area by self-attention mechanism. Given the image feature $f_{img} \in \mathbb{R}^{B,C}$, we first extend the dimensionality of f_{img} to $\mathbb{R}^{B,N,C}$ to match the N classes in $\mathcal{D}_k$. Then we feed f_{img} to two linear layers, and generate f_{img}^q and f_{img}^k . The single-modal enhanced map $M_e \in \mathbb{R}^{N,N}$ is formulated as:

$$M_e = \text{softmax}(f_{img}^q \otimes (f_{img}^k)^T), \quad (4)$$

where $\otimes$ represents the matrix multiplication. Meanwhile, we feed f_{img} to another linear layer to generate f_{img}^v . Then the obtained single-modal enhance map M_e is applied to f_{img}^v through a matrix multiplication and element-wise addition to getting the enhancement visual feature:

$$f_{en} = M_e \otimes f_{img}^v + f_{img}. \quad (5)$$

Multi-modal interaction: To alleviate the irrelevant associative contexts in f_{en} , we introduce the multi-modal interaction to suppress the association and focus on the target area. The enhancement visual feature $f_{en} \in \mathbb{R}^{N,C}$ obtained in (5) is first fed to a linear layer and generates f_{en}^q . Meanwhile, we feed text feature $f_{txt} \in \mathbb{R}^{N,C}$ to another projection layer to generate f_{txt}^k . Since the target knowledge has a response in the text feature f_{txt}^k but the association does not exist in any of the known closed set classes, thus the multi-modal suppression map $M_s \in \mathbb{R}^{N,N}$ is formulated as:

$$M_s = \text{softmax}(f_{en}^q \otimes (f_{txt}^k)^T). \quad (6)$$

We adjust the suppression map M_s through a matrix multiplication and an element-wise addition to obtain the target feature:

$$f_{tar} = M_s \otimes f_{en}^v + f_{en}, \quad (7)$$

where f_{en}^v is the linear projection of f_{en} .

C. Anti-Association Calibration

The target feature obtained from (7) is not quite discriminative due to the absence of additional constraints to supervise, thus we propose three loss functions to calibrate the target feature.

Classification Loss: Text feature $f_{txt} \in \mathbb{R}^{N,C}$ generated by the text encoder derived from a prompt follows the form of '[Prompt][CLASS]', where the class token is replaced by the class name in $\mathcal{D}_k$ in Section III-A. Then the prediction confidence of the target feature f_{tar} is then computed as:

$$p(y = k|x) = \frac{\exp(\cos(f_{txt}^k, f_{tar})/\tau)}{\sum_{j=1}^N \exp((\cos(f_{txt}^j, f_{tar})/\tau))}, \quad (8)$$

where τ is a temperature parameter learned by CLIP and $\cos(\cdot, \cdot)$ denotes cosine similarity. We adhere to the prompt learning paradigm, freeze the image encoder and text encoder, and optimize only the prompt vector and CGA module with parameters θ_P, θ_C respectively to stretch the target feature f_{tar} with ground

truth class by cross-entropy loss:

$$\mathcal{L}_{cls}(\{\theta_P, \theta_C\}) = - \sum_{k=1}^N \Pi_{y_k} \log[p(x)], \quad (9)$$

where Π_{y_k} is an indicator function for ground truth class y_k and $p(x)$ is confidence from (8).

Association Loss: We introduce the association loss to provide a confrontation against the target feature. We first generate a pseudo association feature f_{ass} with f_{img} as:

$$f_{ass} = f_{img} \ominus f_{tar}, \quad (10)$$

where $\ominus$ represents element-wise subtraction, then an association loss is constrained on association feature f_{ass} which should be independent of each known closed set class to make it more genuine as:

$$\mathcal{L}_{ass}(\{\theta_P, \theta_C\}) = - \frac{1}{N} \sum_{k=1}^N y_k \log[p(x)], \quad (11)$$

where $\sum_{k=1}^N y_k$ represents the confidence to be calculated for association feature f_{ass} assigned to each close set classes, aims to push f_{ass} farther away from each classes in $\mathcal{D}_k$.

Division Loss: We introduce two loss functions in (9) and (11) to supervise the target feature and association feature. Since target and association essentially belong to different categories, so to further divide target feature f_{tar} from association feature f_{ass} , we propose the division loss as:

$$\mathcal{L}_{div}(\{\theta_P, \theta_C\}) = \cos(f_{tar}, f_{ass}). \quad (12)$$

The total loss function is formulated as follows:

$$\mathcal{L}_{total}(\{\theta_P, \theta_C\}) = \mathcal{L}_{cls} + \lambda_{ass} \mathcal{L}_{ass} + \lambda_{div} \mathcal{L}_{div}, \quad (13)$$

where λ_{ass} and λ_{div} represents the weight of $\mathcal{L}_{ass}$ and $\mathcal{L}_{div}$ respectively. In general, only the parameters in the prompts and CGA module are optimized in the training phase, while the text encoder and image encoder are both kept frozen. single-modal guided activation makes sufficient advantage of the homo-modal and multi-modal information to generate target feature, loss functions in anti-associations calibration further constrain and divide the target with the association, which jointly facilitate A^2Pt to get rid of the irrelevant co-occurrence contexts and generate the target features that are closest to the essence of the category.

IV. EXPERIMENTS

A. Experimental Settings

Evaluation Metrics: Following previous works [6], [7], [26], [40], we adopt two main metrics to evaluate validity: AUROC and OSCR. AUROC (Area Under the Receiver Operating Characteristic) is a threshold-independent classification metric that evaluates the binary classification known/unknown performance of the model. OSCR (Open Set Classification Rate) [20] measures the trade-off between the classification of known classes and the recognition of unknown classes.

Implementation Details: We follow text and image encoder ViT-B/32 of CLIP with a batch size of 100 and trained for 150

TABLE I
AUROC SCORES $\uparrow$ (%) FOR UNKNOWN RECOGNITION ON SEVERAL DATASETS

Method	Publish	Year	CIFAR10 $O^*=22.54\%$	CIFAR+10 $O^*=46.54\%$	CIFAR+50 $O^*=72.78\%$	TinyImageNet $O^*=68.37\%$
OpenMax † [1]	CVPR	2016	69.5 $^{+25.5}$	81.7 $^{+16.5}$	79.6 $^{+18.2}$	57.6 $^{+33.2}$
G-Openmax † [4]	BMVC	2017	67.5 $^{+27.5}$	82.7 $^{+15.5}$	81.9 $^{+15.9}$	58.0 $^{+32.8}$
OSRCI † [5]	ECCV	2018	69.9 $^{+25.1}$	83.8 $^{+14.4}$	82.7 $^{+15.1}$	58.6 $^{+32.2}$
C2AE † [2]	CVPR	2019	71.1 $^{+23.9}$	81.0 $^{+17.2}$	80.3 $^{+17.5}$	58.1 $^{+32.7}$
GFROSR † [3]	CVPR	2020	83.1 $^{+11.9}$	91.5 $^{+6.7}$	91.3 $^{+6.5}$	65.7 $^{+25.1}$
CGDL † [23]	CVPR	2020	68.1 $^{+26.9}$	79.4 $^{+18.8}$	79.4 $^{+18.4}$	65.3 $^{+25.5}$
RPL † [41]	ECCV	2020	78.4 $^{+16.6}$	88.5 $^{+9.7}$	88.1 $^{+9.7}$	71.1 $^{+19.7}$
PROSER [40]	CVPR	2021	89.1 $^{+5.9}$	96.0 $^{+2.2}$	95.3 $^{+2.5}$	69.3 $^{+21.5}$
CVAECapOSR [24]	ICCV	2021	83.5 $^{+11.5}$	88.8 $^{+9.4}$	88.9 $^{+8.9}$	71.5 $^{+19.3}$
ARPL [6]	TPAMI	2021	91.0 $^{+4.0}$	97.1 $^{+1.1}$	95.1 $^{+2.7}$	78.2 $^{+12.6}$
DIAS ‡ [26]	ECCV	2022	89.6 $^{+5.4}$	93.5 $^{+4.7}$	92.3 $^{+5.5}$	70.1 $^{+20.7}$
AGC [7]	ICLR	2022	91.0 $^{+4.0}$	95.4 $^{+2.8}$	93.9 $^{+3.9}$	82.6 $^{+8.2}$
CLIP [8]	ICML	2021	92.0 $^{+3.0}$	97.4 $^{+0.8}$	96.2 $^{+1.6}$	84.6 $^{+6.2}$
CoOp [9]	IJCV	2022	92.8 $^{+2.2}$	97.1 $^{+1.1}$	96.8 $^{+1.0}$	89.0 $^{+1.8}$
CoCoOp [34]	CVPR	2022	94.4 $^{+0.6}$	97.2 $^{+1.0}$	96.4 $^{+1.4}$	87.8 $^{+3.0}$
Ours	-	-	95.0	98.2	97.8	90.8

$\dagger$ indicates the results are provided by [24]. $\ddagger$ indicates we run the results following standard splits according to the official released code. The green numbers in the upper right show the improvement of our method compared to the traditional OSR methods. The blue numbers in the upper right show the improvement of our method compared to the prompt paradigm methods. The best results are in bold, the second best results under the traditional and prompt paradigm approaches are marked with underline.

epochs. Following the standard practice in the OSR literature [6], [7], [41], we report the experimental results averaged across five ‘known/unknown’ class splits for each dataset. We train the model on two Tesla V100. The input image is resized to 224×224 . We use SGD optimizer with the learning set to 0.002 and the momentum is 0.9.

Datasets: Before introducing the dataset, we first introduce openness to measure the open grade of unknown classes in the testing phase. we define O^* to represent openness [3], [5], [17], [43] as:

$$O^* = 1 - \sqrt{\frac{N_{train}}{N_{test}}}, \quad (14)$$

where N_{train} and N_{test} correspond to the number of classes during training and testing.

For open set recognition, the following datasets are involved:

- **CIFAR10** [14] is a dataset of 60 k images, each size is 32×32 and all images fall into ten different classes. Six classes are selected for training as known classes and the rest are employed for testing as unknown, $O^* = 22.54\%$.
- **CIFAR+10** and **CIFAR+50** [14] are created to allow our model under a larger openness. Four classes are selected from the CIFAR10 as known classes, and 10 and 50 classes from the CIFAR100 dataset are selected as unknown classes. The corresponding O^* are 46.54% and 72.78%.
- **TinyImageNet** [15] contains 200 classes sub-sampled from ImageNet [44]. 20 classes are used as known classes and the rest as unknown classes, $O^* = 68.37\%$.

We compare our A^2Pt with a range of state-of-the-art methods: OpenMax [1] replaces softmax with openmax operation to recognize unknown class; G-Openmax [4] generates fake

open set samples with GAN and adopts openmax for evaluating; C2AE [2] utilize the reconstruction errors difference between known class and unknown class to recognize open set sample; CGDL [23] introduce conditional gaussian distribution for recognition; PROSER [40] propose placeholders to pull the classification decision boundary between closed set categories; ARPL [6] proposes adversarial reciprocal points for open set learning; DIAS [26] grades the pseudo open set data from difficulty aware; AGC [7] achieves SOTA by carefully choosing hyper-parameters for training standard cross-entropy models. Due to the different class splits of DIAS [26] with recent SOTA methods RPL [41], ARPL [6] and AGC [7], we re-train DIAS following standard splits according to the officially released code.

B. Experimental Results

Comparison with AUROC: As shown in Table I, our method is significantly superior to previous traditional methods on the CIFAR series datasets and TinyImageNet dataset. Specifically, on the most challenging TinyImageNet benchmark, we get an 8.2% improvement and outperform other traditional methods by a large margin. We also compare with recent SOTA prompt paradigm methods CLIP [8], CoOp [9] and CoCoOp [34], our model still achieves the best results on each dataset. These results significantly demonstrate that our approach addressed the association problem that occurred not only in traditional open set methods but also in prompt methods. Fully exploits the cross-modal prior and effectively suppresses the expression of irrelevant co-occurrence information, expanding the difference between the known and unknown classes.

Comparison with OSCR: To conduct a more comprehensive analysis, we report comparative experiments on the

TABLE II
OSCR SCORES[†] (%) FOR EVALUATING THE TRADE-OFF PERFORMANCE BETWEEN CLOSED SET CLASSIFICATION AND OPEN SET RECOGNITION

Method	Publish	Year	CIFAR10 $O^*=22.54\%$	CIFAR+10 $O^*=46.54\%$	CIFAR+50 $O^*=72.78\%$	TinyImageNet $O^*=68.37\%$
GCPL [42]	CVPR	2018	84.3 ^{+7.1}	91.0 ^{+5.1}	88.3 ^{+6.9}	59.3 ^{+26.6}
RPL [41]	ECCV	2020	85.2 ^{+6.2}	91.8 ^{+4.3}	89.6 ^{+5.6}	53.2 ^{+32.7}
ARPL [6]	TPAMI	2021	87.9 ^{+3.5}	94.7 ^{+1.4}	92.9 ^{+2.3}	65.9 ^{+20.0}
AGC [†] [7]	ICLR	2022	88.9 ^{+2.5}	93.9 ^{+2.2}	94.2 ^{+1.0}	76.1 ^{+9.8}
CLIP [8]	ICML	2021	84.5 ^{+6.9}	91.4 ^{+4.7}	90.4 ^{+4.8}	73.3 ^{+12.6}
CoOp [9]	IJCV	2022	89.7 ^{+1.7}	94.3 ^{+1.8}	93.1 ^{+2.1}	82.3 ^{+3.6}
CoCoOp [34]	CVPR	2022	91.2 ^{+0.2}	95.1 ^{+1.0}	94.2 ^{+1.0}	82.1 ^{+3.8}
Ours	-	-	91.4	96.1	95.2	85.9

[†]indicates that we run the results following standard splits according to the official released code. The green numbers in the upper right show the improvement of our method compared to the traditional OSR methods. The blue numbers in the upper right show the improvement of our method compared to the prompt paradigm methods. The best results are in bold, the second best results under the traditional and prompt paradigm approaches are marked with underline.

OSCR metric over several recent SOTA methods. As shown in Table II, A^2Pt suppresses other traditional SOTA methods over all datasets. For the CIFAR dataset, we improved by 2.5%, 1.4%, and 1.0%, respectively. For the TinyImageNet dataset, we improved even more by 9.8% compared to the SOTA method. As well compared with recent SOTA prompt paradigm methods, Our models also obtained an outstanding performance. Demonstrating that for the trade-off capability between the classification of known classes and the recognition of unknown classes, our model still achieves substantial and consistent performance. Besides, this result also illustrates that removing irrelevant co-occurrence provides a boost for open set recognition as well as calibrates the features learned by the network for common closed set classification, promoting the network to learn more discriminative features.

Performance on large scale dataset Imagenet-21k-P: Apart from the above conventional datasets, we also experiment on a large-scale dataset ImageNet-21k-P [16]. 200 classes from TinyImageNet are employed for training, then we select non-overlapping ‘Easy’ and ‘Hard’ splits from Imagenet-21k-P for evaluation. Each split contains 1000 classes, and the split manner is based on semantic distance consistent with the splitting setting of AGC [7] for a fair comparison.

As shown in Fig. 3, our method outperforms the SOTA in both AUROC and OSCR metrics despite traditional OSR methods or prompt learning methods, demonstrating that the association problem caused by irrelevant co-occurrence in large-scale scenarios still exists, while our A^2Pt is not limited by the size of the dataset, and still maintains strongest performance even in this environment which is much closer to the real world. We do not report CoCoOp’s results because CoCoOp occupies much higher VRAM than our method, which cannot support the training of the CoCoOp on a large-scale dataset with two Tesla V100.

Performance under different openness: To explore the effectiveness of our method at different levels of openness. We select the same 20 classes from TinyImageNet as known classes for training and choose from 60 to 180 classes for testing, corresponding to a change in openness from 50.1% to 68.4%. Fig. 4 shows the recognition ability of our method with other methods (AGC [7], CLIP [8], CoOp [9] and CoCoOp [34]) under

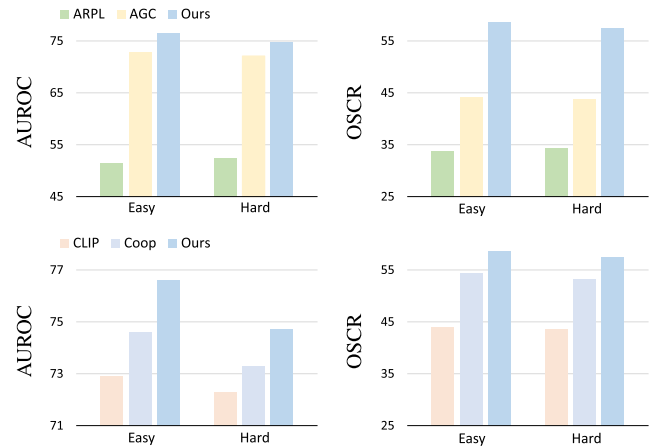


Fig. 3. OSR results on the large scale datasets ImageNet-21K-P. The model was trained on TinyImageNet with 200 classes and evaluated on ImageNet-21K-P with ‘Easy’/‘Hard’ splits. Each split contains 1000 classes. We compare AUROC and OSCR, respectively.

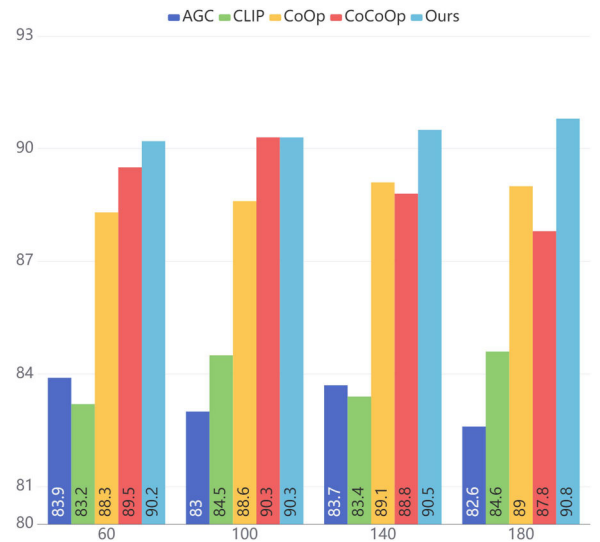


Fig. 4. AUROC comparison with TinyImageNet dataset under different Openness.

TABLE III
COMPARISON WITH DIFFERENT PROMPT METHODS WITH ACCURACY SCORES $\uparrow$ (%)

Method	Publish	Year	CIFAR10	CIFAR+10	CIFAR+50	TinyImageNet
CLIP [8]	ICML	2021	90.6 ^{+5.9}	93.5 ^{+4.0}	93.5 ^{+4.2}	82.7 ^{+9.7}
CoOp [9]	IJCV	2022	95.5 ^{+1.0}	97.2 ^{+0.3}	96.9 ^{+0.8}	90.9 ^{+1.5}
CoCoOp [34]	CVPR	2022	95.8 ^{+0.7}	97.3 ^{+0.2}	97.4 ^{+0.3}	91.3 ^{+1.1}
Ours	-	-	96.5	97.5	97.7	92.4

The blue numbers in the upper right show the improvement of our method compared to the current method. The best results are in bold while the second best results are marked with underline.

TABLE IV
COMPARISON AUROC SCORES (%) WITH DIFFERENT ADAPTER METHOD

Method	$Split_0$	$Split_1$	$Split_2$	$Split_3$	$Split_4$	Average
CLIP [8]	83.4	84.2	83.6	84.8	87.0	84.6
CoOp [9]	89.2	86.4	89.8	90.6	89.0	89.0
CLIP+adapter [45]	86.8	87.7	86.6	88.5	89.7	87.8
CoOp+adapter [45]	87.9	88.7	85.2	88.9	91.0	88.3
Ours	91.0	90.1	90.3	90.7	92.0	90.8

CLIP+adapter resents the text prompt is fixed to ‘a photo of’ and only the adapter is optimized. CoOp+adapter represents the text prompt and the adapter is optimized together. The best results are in bold.

different levels of openness. This result demonstrates that A^2Pt maintains its excellent performance at any level of openness, the target information is not disturbed by more open and more categories, and our approach is stronger in solving the open set problem in real-world occasions.

Performance under closed set classification against PTMs: We conduct this experiment to investigate whether A^2Pt is also effective on closed set classification compared with other methods under the same prompt paradigm. As shown in Table III, compared with the recent SOTA methods CLIP [8], CoOp [9] and CoCoOp [34], our model still achieves the best results. For Cifar series datasets, we obtain an average increase of 0.4%, and for the Tinyimagenet, we achieve a 1.1% improvement. Demonstrating that the association problem has implications not only for open set scenarios but also for closed-set scenarios, and our A^2Pt can effectively reduce the impact of association.

Comparison with PTMs + ADAPTER finetuning paradigm: To explore the performance differences between our approach and the adapter paradigm. We added an adapter to the image encoder following the setting of [45], which consists of two layers of linear and activation functions. We conducted five different splits on the TinyImagenet dataset under the same configuration following [7] and report the results of the open set metric AUROC in Table IV. CLIP+adapter represents the text prompt is fixed to ‘a photo of’ and only the adapter is optimized, CoOp+adapter represents the text prompt and adapter are optimized together. As the results shown in Table IV, for CLIP, adding an adapter increases the trainable parameters in the model, allowing the model to improve its performance on open set recognition. However, for CoOp, the addition of an adapter does not improve the performance of open set recognition, and there is even a slight decrease. We believe that the reason is that the role of the adapter is the same as that of optimizable text prompt, which is to improve the generalization ability of the network to the downstream task, but due to the lack of fine

TABLE V
COMPARISON AUROC SCORES (%) WITH DIFFERENT FINETUNE METHOD

Method	AGC	AGC+F	AGC+L	Ours
CIFAR10	91.0	-	92.2	95.0
CIAFR+10	95.4	71.6	88.2	98.2
CIFAR+50	93.9	70.5	89.8	97.4
Tinyimagenet	82.6	-	85.1	90.8

- Represents the model collapses when finish training phase. AGC+F represents the model’s backbone replaced with CLIP image encoder and finetuned with full model. AGC+L represents the model’s backbone replaced with CLIP image encoder and finetuned with only last linear. The best results are in bold.

module design and loss optimization settings, simply increasing the optimizable parameters lead to the overfitting problem which causes the degradation of the performance. Compared to the model with the addition of adapter, our A^2Pt still maintains the best open set recognition performance, indicating that the module we designed for the adaptation of large language-vision models to downstream tasks has been effective, suppressing the influence of association problems during the training phase of the network.

Comparison with PTMs + OSR finetuning paradigm: To verify that the improvement of our method is not directly derived from the CLIP encoders, we invite at present the most effective method AGC [7] for another comparison experiment. The VGG32 backbone of AGC is replaced with the image encoder of CLIP, then we inherit the optimization strategy of AGC for fair comparison and train the model with fully finetuned and only the last linear layers finetune respectively.

The results shown in Table V indicate that for the only linear layers finetune approach, CLIP does help the model improve the open set recognition performance by 1.2% and 2.5% on CIFAR10 and TinyImageNet respectively. However, on CIFAR+10 and CIFAR+50 the performance decreased. We consider for these two datasets, four classes for training are inadequate to leverage the CLIP model, compared with six and twenty classes involved in training in CIFAR10 and TinyImageNet.

Although we have trained for adequate 600 epochs following AGC, the same flaw also appears in the full finetune approach. For the fully finetuned paradigm, models trained with CIFAR10 and TinyImageNet both turn to collapse. Models on CIFAR+10 and CIFAR+50 only succeed on just 3 out of the 5 splits. We show the average score under these three splits. We consider small-scale data inadequate for prying large model parameters, causing the network to conflict between learning new knowledge and existing priors, which causes the model to collapse.

TABLE VI
COMPARISON TO OTHER METHODS WITH PARAMETERS, SPEED AND VRAM

Method	Parameters (M)	Trainable Parameters (M)	Training Time (s/epoch)	Test Time (s/image)	VRAM (M)
CoOp	125.9896	0.0081	17.30	0.01	6718
CoCoOp	126.0219	0.0415	190.66	0.02	64622
Ours	130.2202	4.2388	21.00	0.01	6840

In general, comparing the above two finetune methods, our A^2Pt substantial to achieve the best results on various datasets, verify that the improvement of our A^2Pt is not derived from the pretrained CLIP encoders, but comes from the contribution of cross-modal guided activation and the matched anti-association calibration.

Comparison of parameters, speed, and VRAM: To explore the efficiency of our model, we compared A^2Pt with other prompt methods following the standard training process. We employ TinyImageNet as training and testing data. As shown in Table VI, VRAM denotes the memory occupation of the model during the training process. Although the total and trainable parameters are slightly larger than that of CoOp and CoCoOp, our training time and test time are basically the same as CoOp, much smaller than CoCoOp. These results indicate that the large parameters do not affect our model's training and inference speed of our model. Besides, the low VRAM, similar to CoOp, also indicates that our model does not require a high configuration and can still work with lower memory. While the improvement of our A^2Pt compared to CoCoOp [34] is not as significant as over CLIP [8] and CoOp [9], with only an average 1.5% point improvement in the AUROC metric on the four conventional datasets, and a corresponding 1.5% point improvement in the OSCR metric. This is a relatively small improvement compared to the 2.3% point improvement in CoOp and the 7.2% point advantage over CLIP. However, our model consumes less memory, only 6840 M which is nearly 90% reduction compared to CoCoOp's 64622 M, and our model consumes less time for training, 21 s/epoch which is nearly 10 times faster compared to CoCoOp's 190 s/epoch. From this perspective achieving similar performance but at a lower cost is a major advantage of our model.

C. Ablation Studies

Ablation of loss functions: The proposed A^2Pt framework consists of two modules: a cross-modal guided activation module and the anti-association calibration loss. We first compare the AUROC results among different datasets with and without each loss in anti-association calibration. As shown in Table IX, using only classification loss ($\mathcal{L}_{cls}$) will result in a significant decrease in the open set recognition ability of the model, due to the degradation of the model to a simple closed-set classification network. When the association loss $\mathcal{L}_{ass}$ or division loss $\mathcal{L}_{div}$ is adopted in the training phase, the performance is significantly enhanced, cause the association is separate from the full co-occurrence feature.

Then, we combine three losses jointly to encourage the model explicitly suppress association influence and evaluate the performance with the most class-related target features, the AUROC

risers to a new level at about 17%–27% compared to the baseline, which demonstrates the loss functions in anti-association calibration is all indispensable.

Simultaneously, we delve into the impact of different loss function designs on two pivotal components, the Cross-Modal Guided Activation (CGA) and Anti-Association Calibration (AAC) modules, throughout the training of the Single-modal Enhancement and Multi-modal Interaction modules. Starting with the Anti-Association Calibration loss, denoted as $\mathcal{L}_{ass}$, originally intended to optimize ‘association’ features, we introduce a modification to refine the network focusing on distancing the association features not just from every class but specifically from the ground truth label of the current sample. The proposed non-ground-truth loss variant, termed $\mathcal{L}_{ngt}$, is defined as follows:

$$\mathcal{L}_{ngt} = -\frac{1}{N} \sum_{k=1}^N \frac{1}{\Pi_{y_k} \log[p(f_{ass})]}, \quad (15)$$

where $\frac{1}{\Pi_{y_k} \log[p(x)]}$ represents the confidence to be calculated for association feature f_{ass} not assigned to the ground truth label class. As for the division loss $\mathcal{L}_{div}$, which is initially designed to create separation between target and association features using cosine similarity, we replaced the cosine similarity with the Euclidean distance loss function $\mathcal{L}_{euc}$, defined as:

$$\mathcal{L}_{euc} = \frac{1}{euc(f_{tar}, f_{ass})}, \quad (16)$$

where f_{tar} and f_{ass} represent the target feature and association feature, respectively. By leveraging the Euclidean distance, this adjusted loss aims to create a distinct distance between target and association features. Our comparative experiments, as presented in Table IX, shed light on the impact of modifying the loss functions in our studies. When we replaced $\mathcal{L}_{ass}$ with $\mathcal{L}_{ngt}$, the results were striking: a dramatic drop in performance occurred, with a decrease of 14.5% for TinyImageNet and 15.3% for CIFAR10. These findings suggest that the simple restriction of association features to stay away from the corresponding label class is insufficient. It becomes apparent that the attributes of the association feature itself must exhibit dissimilarity across different classes, not just with the ground truth class. Similarly, when $\mathcal{L}_{div}$ was substituted with $\mathcal{L}_{euc}$, we observed a notable decline in performance, with a decrease of 4.1% for TinyImageNet and 6.5% for CIFAR10. This result implies that using the Euclidean space to enforce a separation between target features and association features may be inadequate. The vectors in this potential space appear to be high-dimensional, and the advantages of constraining them from a cosine similarity perspective seem to outweigh those of using Euclidean distance.

TABLE VII
COMPARISON AUROC SCORES (%) WITH DIFFERENT λ_{ass} VALUE

λ_{ass}	13	17	21	25	29
AUROC	98.1	98.1	98.2	98.0	97.9

The best results are in bold.

TABLE VIII
COMPARISON AUROC SCORES (%) WITH DIFFERENT λ_{div} VALUE

λ_{div}	0.2	0.5	1	2	3
AUROC	97.9	98.0	98.2	98.0	97.7

The best results are in bold.

TABLE IX
AUROC SCORES (%) ON SEVERAL DATASETS ARE REPORTED TO
VERIFY THE EFFECTS OF DIFFERENT LOSSES

$\mathcal{L}_{cls}$	$\mathcal{L}_{ass}$	$\mathcal{L}_{div}$	Variants	CIFAR10	CIFARAdd10	CIFARAdd50	TinyImageNet
✓				68.9	81.2	69.2	78.2
✓		✓		92.5	94.0	93.0	87.3
✓	✓			93.2	89.2	88.6	86.1
✓		✓	$\mathcal{L}_{ass} \rightarrow \mathcal{L}_{ngt}$	79.7	91.9	92.1	76.3
✓	✓		$\mathcal{L}_{div} \rightarrow \mathcal{L}_{cuc}$	88.5	93.8	94.9	86.7
✓	✓	✓		95.0	98.2	97.8	90.8

TABLE X
COMPARISON AUROC SCORES (%) WITH AND WITHOUT SINGLE-MODAL
ENHANCEMENT MODULE AND MULTI-MODAL INTERACTION MODULE
RESPECTIVELY

Single	Multi	$Split_0$	$Split_1$	$Split_2$	$Split_3$	$Split_4$	Average
✓		86.3	86.9	85.8	84.9	89.5	86.6
	✓	90.1	89.0	88.0	89.0	89.7	89.1
✓	✓	91.0	90.1	90.3	90.7	92.0	90.8

Single represents the model's performance only with the single-modal enhancement module. Multi represents the model's performance only with the multi-modal interaction module. The best results are in bold.

Furthermore, we vary λ_{ass} from 13 to 29 and λ_{div} from 0.2 to 3 to test the corresponding AUROC metric. As shown in Table VII, for λ_{ass} the AUROC metric vary from 98.1 to 97.9, and in Table VIII, λ_{div} vary from 97.7 to 98.2. The difference between the maximum and minimum is only 0.2 and 0.5, indicating that $\mathcal{L}_{ass}$ and $\mathcal{L}_{div}$ are both not sensitive to the weighting parameter and can maintain a stable output. Therefore, we choose a weight value 21 for $\mathcal{L}_{ass}$ and 1 for $\mathcal{L}_{div}$ that gives a relatively higher result.

Ablation of cross-modal guided activation: Given that the cross-modal guided activation collaborates with the association loss and division loss in anti-association calibration, thus when we remove the cross-modal guided activation, the model degrades to the CoOp [9], in which only continuous prompt is optimized. And as the results shown in Table I, our A^2Pt suppress CoOp on each common dataset with closed-set accuracy and open set AUROC, OSCR metrics. This result proves that the association suppresses module suppressing the irrelevant co-occurrence information, which can further improve model performance. Further, we conducted ablation to verify the validity of the single-modal enhancement module and multi-modal interaction module respectively. We employ five standard sets of different close set splits on TinyImagenet following [7] for training, and test the performance of the AUROC of the model in the presence or absence of the single/multi-modal module on the corresponding open sets. As shown in Table X, the performance

TABLE XI
CLOSED-SET ACCURACY (%), AUROC SCORES (%) AND OSCR (%) ON
TINYIMAGENET DATASET ARE REPORTED TO VERIFY THE EFFECTS OF
DIFFERENT DATA AUGMENTATIONS

Random-FP	Random-GS	Mixup	Random-Er	Accuracy	AUROC	OSCR
✓				92.5	90.7	85.9
	✓			92.2	90.7	85.8
✓	✓			92.4	90.8	85.9
✓	✓	✓		91.8	90.3	84.9
✓	✓		✓	91.3	90.1	84.2
✓	✓	✓	✓	91.1	89.3	82.7

"Random-FP" stands for RandomHorizontalFlip, "Random-GS" stands for RandomGrayscale, Random-ER" stands for RandomErasing.

of the missing single-modal enhancement module is always inferior to that of the complete model regardless of the split method, indicating that the presence of single-modal enhancement plays a role in helping to express more accurate features for the subsequent multi-modal interaction processing. Similarly, the performance of the model decreases more when the network lacks multi-modal interaction, indicating that the introduction of text information is crucial to suppress association.

Our A^2Pt model utilizes two default data augmentation methods following [7], RandomHorizontalFlip and RandomGrayscale. To assess the impact of these default methods, as well as the influence of other augmentation techniques on the model's performance, we conducted a comparative experiment detailed in Table XI. In this experiment, we systematically removed the two default data augmentation methods on the TinyImageNet dataset and then sequentially introduced two additional data augmentation methods, Mixup [46] and RandomErasing [47] to the default configuration. We evaluated the model's performance across various dimensions, including closed-set Accuracy, open-set AUROC, and combined OSCR metrics. For RandomErasing, we adhered to the default configuration outlined in [47], setting the probability (p) to 1, and specifying the scale as (0.02, 0.33), the ratio as (0.3, 3.3), and the value as the default 'random'. Meanwhile, for Mixup, we followed the default configuration from [46], with $\alpha = \beta = 0.5$ for the Beta Distribution. The results, as presented in Table XI, reveal insightful observations. When it comes to the default augmentations' RandomHorizontalFlip, RandomGrayscale, even RandomErasing and Mixup alone, the model's performance demonstrates minor fluctuations. The changes in open-set AUROC and OSCR metrics are less than 0.7%, while closed-set accuracy varies by less than 1.1%. This suggests that our model exhibits stability and is not overly dependent on the specific data augmentation methods employed, maintaining consistent performance across various data augmentation approaches. The incorporation of both Mixup and RandomErasing has a somewhat amplifying effect on the model, resulting in a 1.3% variation in accuracy, a 1.5% reduction in open-set AUROC, and a 3.2% decrease in OSCR metrics. This observation suggests that the combined usage of Mixup and RandomErasing may lead to a mild degradation of the model's performance. The reasoning behind this phenomenon likely lies in the fact that the image transformations introduced by Mixup and RandomErasing may cause some additional damage to the target feature and association feature compared to the original data augmentation.

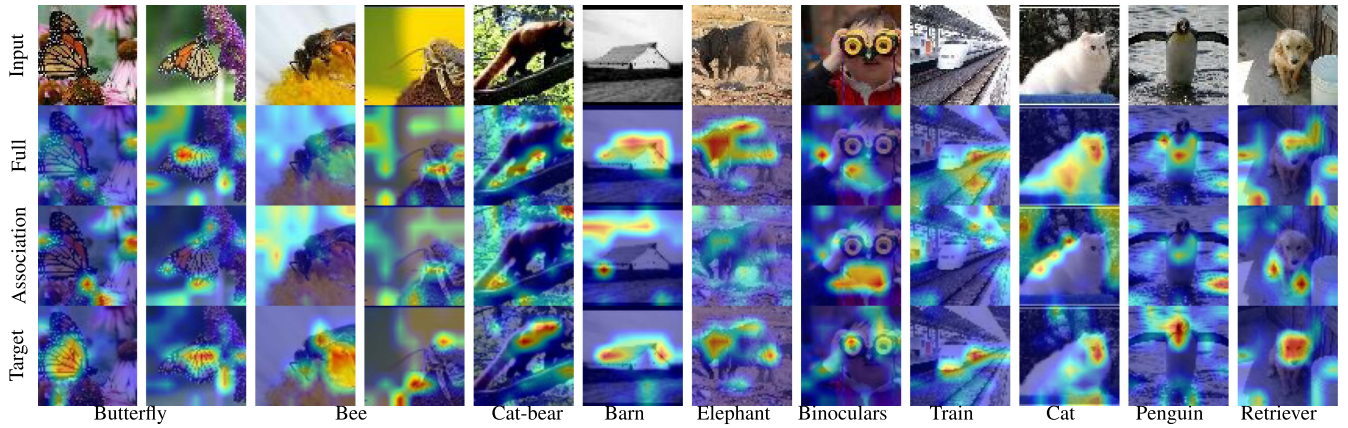


Fig. 5. Visualization of the full, association, and target feature of CAM, respectively. Full represents the feature output from the clip's image encoder, and target and association feature is the output from CGA in A^2Pt . In contrast with full and association features, the target feature activates the most class-related area.

This, in turn, could weaken the model's ability to learn features accurately, resulting in a slight decline in overall model performance.

However, the key takeaway from this analysis is that, regardless of which augmentation method is used, the final performance of our A^2Pt model is not substantially affected. This observation underscores that our model's enhancement is not contingent on the specific choice of data augmentation. Instead, it is primarily a result of the well-designed loss functions and the collaborative interaction of the Cross-Modal Guided Activation (CGA) and Anti-Association Calibration (AAC) modules. These components play a pivotal role in elevating the model's capabilities, making it robust and adaptable to variations in the data augmentation strategy.

D. Qualitative Analysis

Visualization of the activations: To verify that the target features output by our model is activated at the most class-related areas. We adopt CAM [48] to analyze concerns of the network. Fig. 5 shows some examples of the activation area with CAM [48]. The 4th row is the result of our method by using generated target feature, the 2nd and 3rd rows are full features (contains target and association information) and association feature respectively. Compared to CAM of full and association feature, the results of our method activates most class-related areas. Notably, the association activation area is not completely class-agnostic, for instance, the follower in class *butterfly* and the child in class *binoculars*. This explains why activating these associative information has less impact on closed set classification, since they are not completely meaningless, but harmful on open set recognition, causing semantic areas not related to the current class to lead to confusion between different known and unknown classes. We also visualized the difference between the CAM activations of the samples with and without the single-modal enhancement module. As shown in Fig. 6, although the CAM without the single-modal enhancement module can also remove the association in the image, such as the 'grass' in the 'European fire salamander', 'seawater' in the 'Dugong' and the 'table' in

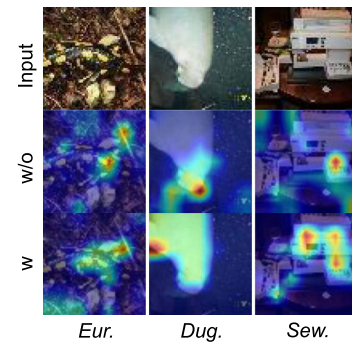


Fig. 6. Visualization of with and without single-modal enhancement module CAM respectively. 'Eur.' represents the 'European fire salamander,' 'Dug.' represents the 'Dugong,' 'Sew.' represents the 'Sewing machine.' W/o represents the CAM without the single-modal enhancement module. W represents the CAM with the single-modal enhancement module. The Cam with the single-modal enhancement module activates more accurate and complete features in the image.

the 'Sewing machine', their activation of the target region is not complete. The European fire salamander only activates the limb part missing the head information, the dugong only activates the head missing the body, and the sewing machine only activates a small part of the region, which leads to the model not learning the most complete and accurate target features, which in turn leads to the degradation of the open set recognition performance of the model in Table X. In contrast to the complete A^2Pt with the single-modal enhancement module, the model activates more accurate and complete target regions, and the processing with multi-modal interaction makes the best performance of open set recognition.

Visualization of closed-set and open set distribution: Our model differentiates the known and unknown classes by the value of max logits. A high value of logits indicates the testing samples belong to a closed set. To demonstrate the difference in the distribution of open set and closed-set samples, We show some examples of the max logits in Fig. 7 with joy plots, and comparisons between different known classes with unknown classes. 50 random samples are selected for each class. The blue part with lower logits indicates the distribution of unknown

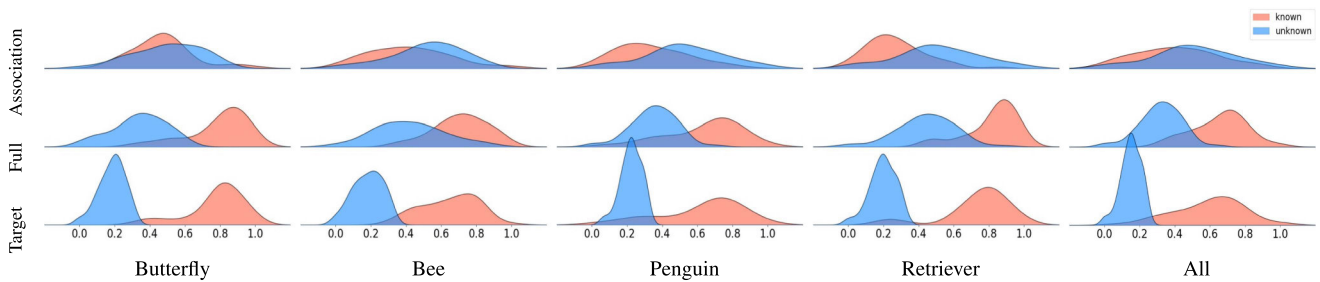


Fig. 7. Comparison of known and unknown distribution with the maximum logits. The horizontal coordinate indicates the range of normalized logits variation and the vertical coordinate indicates the number of samples. Blue color for unknown classes and the red for known.

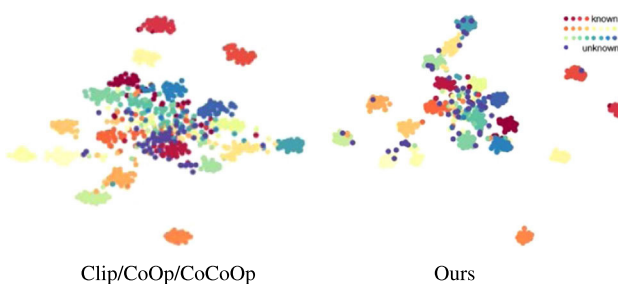


Fig. 8. Comparison with different prompt methods with t-SNE. Clip, CoOp and CoCoOp generates the same feature through the same image encoder. Purple refers to the unknown class and the others refer to the known class.

samples, and the red area with higher logits indicates known samples. For a critical comparison, we report the full, association, and target logits respectively. As shown in Fig. 7, both known and unknown samples in the association row overlap a considerable portion, indicating that the associative information does not effectively distinguish the known class from the unknown class. Influenced by the associative information, the logits of the full feature in 2nd row also have a considerable overlap area when making the distinction between known and unknown classes. However, for our generated target feature logits, the overlap between the unknown class regions in blue and the known classes in red is significantly mitigated, indicating that our isolated target information is better discriminatory for the unknown samples, both on each class and all classes in the last column.

Visualization of the t-SNE: We visualize the output with t-SNE [49] to analyze how the spatial distribution of our generated target features differs from the image features generated by other methods. Given the image features an output of CLIP [8], CoOp [9], and CoCoOp [34], all from the same image encoder (these methods focus on solving text prompt), their image features distribution keeps the same with the same image. As shown in Fig. 8, the color of the point in purple refers to the known classes, and the other refers to different known classes. The data distribution in our model is more compact and the inter-class distance is brighter than that of other prompt methods. Meanwhile, the unknown class points are more distributed between the gaps of known categories, and comparing other methods will be overlapped with the center of known categories, further

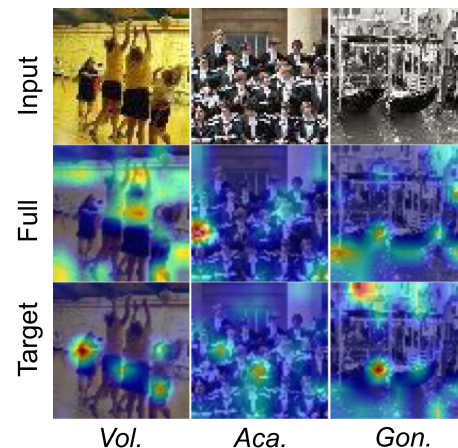


Fig. 9. Visualization of failure case of the full and target feature of CAM respectively. ‘Vol.’ represents ‘Volleyball’, ‘Aca.’ represents the ‘Academic gown’, ‘Gon.’ represents the ‘Gondola’. Full represents the feature output from the clip’s image encoder, and the target feature is the output from CGA in A^2Pt . Compared to the samples shown in Fig. 5, the activation of the target region for these failed samples containing complex irrelevant co-occurrence information is not a pure target region.

demonstrating that the target features generated by our model are more distinguishable than those mixed with associative information.

E. Discussion

Although our method achieves excellent performance, like current SOTA OSR methods, it is still challenging for A^2Pt to handle situations where multiple instances exist in one image. As shown in Fig. 9, when there is relatively more irrelevant co-occurrence information in the image, such as ‘people’ in the ‘volleyball’ class, very many ‘students’ inside the ‘academic gown’ and many ‘buildings’ in ‘gondola’, resulting in target instances that are not obvious and very much smaller in size relative to the contextual information. Our A^2Pt can also focus on the target region, but it is more disturbed by the irrelevant association context, and the anti-association performance degrades. Compared with only one or a small number of associations, this situation requires a higher recognition performance of the model. Thus how to pay more attention to the target class and at the same time suppress the multiple associations in one test sample will be our future study.

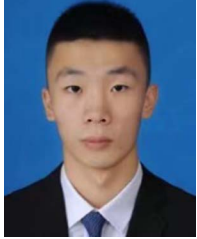
V. CONCLUSION

In this study, we unveiled that the unknown open set misclassified as a known close class is caused by the irrelevant co-occurrence association information. We accordingly proposed an anti-associative prompt tuning (A^2Pt) approach to precisely suppress the associative influence. A^2Pt leverages the cross-modal priors to generate the most class-related target feature and employ anti-association calibration with three loss functions to supervise the difference between target and association. Extensive experiments on CIFAR, TinyImageNet and ImageNet-21K-P benchmarks validated the effectiveness of the proposed approach, in striking contrast with the state-of-the-art. The A^2Pt approach provides fresh insight into the OSR problem.

REFERENCES

- [1] A. Bendale and T. E. Boulton, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1563–1572.
- [2] P. Oza and V. M. Patel, "C2AE: Class conditioned auto-encoder for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2307–2316.
- [3] P. Perera et al., "Generative-discriminative feature representations for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11814–11823.
- [4] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," in *Proc. Brit. Mach. Vis. Conf.*, 2017.
- [5] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 613–628.
- [6] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8065–8081, Nov. 2022.
- [7] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?," in *Proc. Int. Conf. Learn. Representations*, 2022.
- [8] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [9] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, pp. 2337–2348, 2022.
- [10] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proc. 1st Int. Workshop Multimedia Intell. Storage Retrieval Manage.*, 1999, pp. 1–9.
- [11] A. Quattoni, M. Collins, and T. Darrell, "Learning visual representations using images with captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
- [12] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012.
- [13] A. Joulin, L. v. d. Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 67–84.
- [14] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
- [15] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS231N*, vol. 7, no. 7, 2015, Art. no. 3.
- [16] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.
- [17] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.
- [18] D. O. Cardoso, F. França, and J. Gama, "A bounded neural network for open set recognition," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2015, pp. 1–7.
- [19] M. Hassen and P. K. Chan, "Learning a neural-network-based representation for open set recognition," in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 154–162.
- [20] A. R. Dhamija, M. Günther, and T. Boulton, "Reducing network agnostophobia," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2018.
- [21] I. Jo, J. Kim, H. Kang, Y.-D. Kim, and S. Choi, "Open set recognition by regularising classifier with fake data generated by generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2686–2690.
- [22] Y. Yu, W.-Y. Qu, N. Li, and Z. Guo, "Open-category classification by adversarial sample generation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:21277286>
- [23] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13480–13489.
- [24] Y. Guo, G. Camporese, W. Yang, A. Sperduti, and L. Ballan, "Conditional variational capsule network for open set recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 103–111.
- [25] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [26] Y. Moon, J. Park, H. S. Seong, C.-H. Cho, and J.-P. Heo, "Difficulty-aware simulator for open set recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 365–381.
- [27] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, *arXiv:2101.00190*.
- [28] X. Liu et al., "P-tuning V2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," 2021, *arXiv:2110.07602*.
- [29] Z. Wang et al., "CRIS: CLIP-driven referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11686–11695.
- [30] T. Wei et al., "HairCLIP: Design your hair by text and reference image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18072–18081.
- [31] H. Fang, P. Xiong, L. Xu, and Y. Chen, "CLIP2Video: Mastering video-text retrieval via image CLIP," 2021, *arXiv:2106.11097*.
- [32] H. Luo et al., "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.
- [33] A. Miech et al., "End-to-end learning of visual representations from uncensored instructional videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9879–9889.
- [34] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16816–16825.
- [35] Y. Du et al., "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14084–14093.
- [36] S. Manli et al., "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 14274–14289.
- [37] N. Liao, X. Zhang, M. Cao, Q. Tian, and J. Yan, "R-Tuning: Regularized prompt tuning in open-set scenarios," 2023, *arXiv:2303.05122*.
- [38] M. Jia et al., "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 709–727.
- [39] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. A. Efros, "Visual prompting via image inpainting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 25005–25017.
- [40] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4401–4410.
- [41] G. Chen et al., "Learning open set network with discriminative reciprocal points," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 507–522.
- [42] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Robust classification with convolutional prototype learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3474–3482.
- [43] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boulton, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.
- [44] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [45] P. Gao et al., "Clip-adapter: Better vision-language models with feature adapters," 2021, *arXiv:2110.04544*.
- [46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=r1Ddpl-Rb>
- [47] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

- [48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
- [49] L. V. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.



Hairui Ren received the B.Sc. degree in mechanical engineering in 2017 from Jilin University, Changchun, China, where he is currently working toward the Ph.D. degree in computer science with the School of Artificial Intelligence. His research interests include image recognition and object detection.



Fan Tang received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is currently an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.



Xingjia Pan received the B.S. degree in automation and finance from Nankai University, Tianjin, China, in 2015, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently a Senior Researcher with Youtu Lab, Tencent, Shanghai, China. His research interests include computer vision and machine learning.



Juan Cao received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. She is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences. She has more than 90 publications in international journals and conferences including IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, KDD, WWW, and CVPR. Her research interests include multimedia content analysis and fake multimedia detection.



thesis and image recognition. He is a Member of the ACM.

Weiming Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the University of Lorraine, Lorraine, France, in 2007. He is currently a Professor with the Sino-European Lab, Computer Science, Automation and Applied Mathematics (LIAMA) and National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include image syn-



Zhiwen Lin received the master degree in computer engineering from Zhejiang University, Hangzhou, China, in 2019. He is currently a Researcher of Tencent Youtu Lab, Shanghai, China. His research interests include computer vision and deep learning.



Ke Yan received the M.Sc. degree in computer science from Peking University, Beijing, China, in 2017. He is currently a Staff Researcher of Tencent Youtu Lab, Shanghai, China. His research interests include computer vision and deep learning.



Changsheng Xu (Fellow, IEEE) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and the Executive Director of the China-Singapore Institute of Digital Media. He was the recipient of the Best Associate Editor Award of ACM Transactions on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He was the Program Chair of ACM Multimedia 2009. He was an Associate Editor, the Guest Editor, General Chair, Program Chair, an Area/Track Chair, a Special Session Organizer, Session Chair, and a TPC member of more than 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is an Associate Editor for *ACM Transactions on Multimedia Computing, Communications and Applications* and the Editor-in-Chief of *ACM/Springer Multimedia Systems Journal*. He is an IAPR Fellow and ACM Distinguished Scientist.