Cross-modal Data Augmentation for Tasks of Different Modalities

Dong Chen, Yueting Zhuang, *Senior Member, IEEE*, Zijin Shen, Carl Yang, Guoming Wang*, Siliang Tang, Yi Yang, *Senior Member, IEEE*

***Abstract*—Data augmentation has become one of the keys to alleviating the over-fitting of models on training data and improving the generalization capabilities on testing data. Most existing data augmentation methods only focus on one modality, which is incapable when facing multiple data modalities. Some prior works try to interpolate with random coefficients in the latent space to generate new samples, which can generically work for any data modality. However, these works ignore the extra information conveyed by multimodality data. In fact, the extra information in one modality can provide semantic directions to generate more meaningful samples in another modality. This paper proposes Cross-modal Data Augmentation (CMDA), a simple yet effective data augmentation method to alleviate the over-fitting issue and improve the generalization performance. We evaluate CMDA on unsupervised and supervised tasks of different modalities, on which CMDA consistently and significantly outperforms baselines. For instance, CMDA improves the unsupervised anomaly detection baseline in vision modality from the AUROC $76.46\%, 73.07\%$ and 64.36% to $83.25\%, 76.22\%$ and 70.57% on three different datasets, respectively. Besides, extensive experiments demonstrate that CMDA is applicable to various neural network architectures. Furthermore, prior methods that interpolate in the latent space need to work with downstream tasks to construct the latent space. In contrast, CMDA can work with or without downstream tasks, which makes the applicability of CMDA more extensive. Our source code is publicly available for non-commercial or research use at https://github.com/Anfeather/CMDA.**

***Index Terms*—Cross-modal, Data Augmentation.**

I. Introduction

WITH the rapid gains in computational resources, the trained neural network is prone to having much more parameters compared to the number of training samples, thus, over-fitting might happen and weaken its generalization ability. For example, a learned model may describe random error or noise instead of the underlying data distribution [1], and it may exhibit good performance on the training data but fail drastically on the testing data.

Most data augmentation methods only focus on a single modality, which is modality-specific. Popular methods for vision modality include geometric changes, photometric changes, information dropping, etc. As for language modality, the popular methods include random swap, random insertion, random deletion, random synonym replacement [2], etc. [3], [4] and [5] propose to apply interpolation in the latent space with the output of neural networks, which is a generic way to augment data for different modalities. Although performing data augmentation in the latent space is effective for different individual modalities, all the aforementioned methods do not consider the possibility of data augmentation with extra information in multimodality data. At first, we define *extra information* as *the information conveyed by one sample in one modality that matches multiple different samples from other modalities.* For example, as illustrated in Fig. 1, the sentence "A dog is running fast on the lawn as if it is flying" is paired with image 1 (*e.g.*, a running dog), but it can also match image 2 (*e.g.*, a jumping dog) well because of the extra information "flying". [6] finds that the specific direction in the latent space has special semantics. For example, we can add facial hair to a male face by translating the corresponding latent representation towards the direction of facial hair, which suggests that interpolating with a specific direction in the latent space can generate new semantic samples. In multimodality datasets, the specific direction in one modality can be provided by extra information in other modalities. Thus, we propose to perform data augmentation with extra information that can provide specific directions (expressed with matching scores) to generate more meaningful samples. Specifically, we regard the

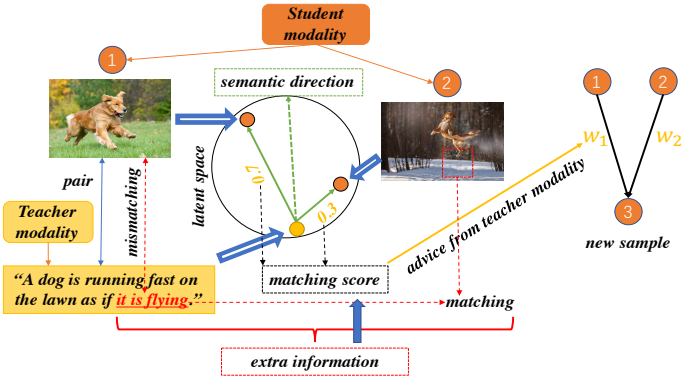


Fig. 1. Cross-modal Data Augmentation (CMDA) tries to augment data for the student modality with the advice of teacher modality. Specifically, we first select the student modality (*e.g.*, vision modality in this case) and teacher modality (*e.g.*, language modality in this case) according to the downstream tasks (*e.g.*, image classification in this case). Then, samples from different individual modalities will be projected into the common latent space, and samples (sentences) in the teacher modality will compute the matching score with samples (images) in the student modality. Finally, we can interpolate with advice (matching scores) in the raw space to get new semantic samples.

*Corresponding author
D. Chen, Y. Zhuang, Z. Shen, G Wang, S. Tang and Y. Yang are with the College of Computer Science and Technology, Zhejiang University, China (E-mail: chendongcs@zju.edu.cn; zijinshen@zju.edu.cn; yzhuang@zju.edu.cn; NB21013@zju.edu.cn; siliang@zju.edu.cn; yangyics@zju.edu.cn)
C. Yang is with the Emory University, USA. E-mail: yangji9181@gmail.com

matching scores as advice to generate samples, the modality that provides advice as the teacher modality, and the modality that accepts advice as the student modality. As illustrated in Figure 1, the sentence "A dog is running fast on the lawn as if it is flying" may have a matching score of 0.7 with image 1, while that of 0.3 with image 2. Thus, we may get a new image where a dog is flying above the lawn as 3 with the advice of language modality (*e.g.*, matching scores, 0.3 with image 2 and 0.7 with image 1). In the above example, there are two modalities. One is the student modality where we want to get more samples, and the other is the teacher modality that gives advice (*i.e.*, matching scores) to the student modality. Note that there may be multiple teacher modalities and only one student modality.

This paper proposes Cross-modal Data Augmentation (CMDA), a simple yet effective data augmentation method to alleviate the over-fitting issue and improve the generalization performance of the vanilla method. When performing data augmentation for the student modality, CMDA first projects samples from different individual modalities into the common latent space and computes the matching scores. Then, CMDA perform interpolating with matching scores in the raw space (i.e., student modality data space), where the generated samples will be more meaningful. Therefore, the proposed method is adaptable for various modalities, that is, modality-agnostic.

Besides, traditional interpolation methods randomly select samples from one batch to augment the dataset, which ignores the relationships among them [7]. We argue that random selection may degrade the semantic representation of new samples. As illustrated in Figure 2, image1 (papaver nudicaule with a spider) can be interpolated with image2 (snapdragon) and generate a new sample (a) that shows snapdragon with a spider. However, if image1 interpolates with image3 (sailboat), the new sample (b) can be meaningless, as no triangle will grow on the flower. Therefore, we propose to compute the matching score in CMDA by Double Cross-Attention (DCA), which computes matching scores with inter-modal relation weight and intra-modal relation weight. Inter-modal and intra-modal relations denote the relationship between different modalities and clusters of the same modality, respectively. Moreover, we use inter-modal relation weight times intra-modal relation weight as the overall matching scores to interpolate in the raw space, where the intra-modal relation will select close clusters, thus alleviating the meaningless interpolation.

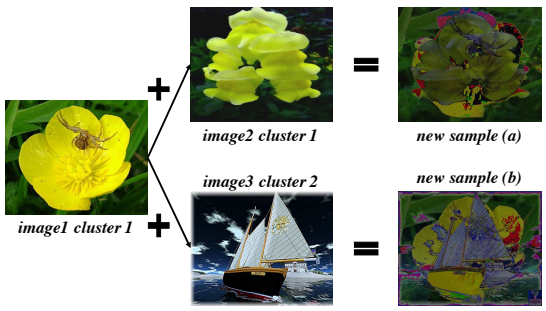


Fig. 2. Data augmentation with images from different clusters. In this paper, we use clusters to represent the relationship between samples, and samples in the same cluster are more similar than those in different clusters.

Overall, CMDA performs data augmentation for student modality with the advice of teacher modalities to improve the performance of downstream tasks in different modalities. Without bells and whistles, CMDA significantly improves the performance of unsupervised anomaly detection and supervised classification [8], [9] in different modalities over COCO, Wikipedia, and Oxford-102. Moreover, we run experiments on various neural network architectures, and CMDA consistently and significantly improves the results in all these cases. Furthermore, unlike prior latent space augmentation works, CMDA (offline) can learn the common latent space without downstream tasks, thus avoiding the computational overhead of data augmentation when applied to different downstream tasks. Meanwhile, CMDA (offline) also clears the suspicion that the improvement comes from better mapping function, as the downstream task model is only trained with augmented data in the student modality.

The main contributions of this paper can be summarized as follows:

- We propose a novel framework, CMDA, which can perform data augmentation for various student modalities and tasks with the advice of teacher modality.
- We propose a cross-modal attention method named DCA, which utilizes inter-modal and intra-modal relations of multimodality datasets to alleviate the meaningless interpolation.
- We extensively evaluate CMDA on unsupervised anomaly detection, supervised classification and image caption tasks in vision and language modalities over various datasets and neural network architectures.

II. Related Work

A. Cross-modal Learning

There are many modalities in datasets in practice, such as vision modality and language modality. To learn a common latent space, where the similarity between the samples from different modalities can be measured, a variety of cross-modal approaches have been proposed. And these cross-modal approaches can be roughly divided into two categories: 1) Cross-modal hashing approaches [10]–[12]. Cross-modal hashing approaches map the heterogeneous data into a common Hamming space, in which the representations are encoded to binary codes. However, the similarity will be slightly inaccurate due to the loss of information [13]. 2) Real-valued approaches [14]–[16] that includes three subclasses. 1. Unsupervised approaches [17] that only use co-occurrence information to learn the latent space for all modalities. 2. Pairwise approaches [18], [19] that utilize more similar pairs to learn a meaning metric to learn the latent space. 3. Supervised approaches [16] that exploit label information to distinguish the samples from different categories. The supervised methods target to pull each instance close to the instances from the same category while pushing away from other instances from different categories. This paper focuses on both unsupervised and supervised cross-modal methods with contrastive learning and supervised contrastive learning to construct a common latent space.

B. Data Augmentation

Data augmentation is a popular strategy that can improve the generalization of neural networks. Specifically, data augmentation usually augments datasets by transforming raw data while preserving semantics globally, following humans' cognitive intuition. The most popular data augmentation methods include geometric changes (horizontal flip and vertical flip), photometric changes (color jitter and gaussian blur), and information dropping (random erasing [20] and cutout [21]). In addition, Mixup augments a dataset by interpolating two images [7], which also has been empirically shown to improve test performance substantially. However, most of the aforementioned methods are only suitable for vision tasks. Instead of improving the model performance in a specific modality, this paper proposes a generic way to perform data augmentation that can work for any data modality.

C. Data Augmentation in the Latent Space

Many tasks need to process different modalities, such as vision and language [22]–[25], 3D [25]–[27], video [28]), etc. It is effective to project all samples into the common latent space [29]–[33], where all modalities become the same pattern. Hence, [3] applies data augmentation in the latent space for different modalities. Moreover, Manifold Mixup generates new samples by interpolating the outputs from different hidden layers [34]. LSI also interpolates in the latent space for image classification [4]. Besides, [5] also explores data augmentation in the latent space for text data. Although performing data augmentation in the latent space is applicable for various modalities, these methods ignore the extra information conveyed by multimodality that provides semantic directions to generate more meaningful samples in other modalities. Therefore, we propose Cross-modal Data Augmentation (CMDA) and try to augment data in the student modality with the advice of teacher modalities.

III. Method

This section presents the proposed Cross-modal Data Augmentation (CMDA) method that aims to perform data augmentation in the student modality with advice of teacher modalities.

Assume that a given data $x := \left(x^{(1)}, \cdots, x^{(K)}\right)$ consists of K modalities, and $x^{(k)} \in \mathcal{D}^{(k)}$, where $\mathcal{D}^{(k)}$ denotes the k-th modality of input space $\mathcal{D} = \mathcal{D}^{(1)} \times \mathcal{D}^{(2)} \times \cdots \times \mathcal{D}^{(K)}$. Moreover, let student modality and teacher modalities be $\hat{\mathcal{D}}$ and $\bar{\mathcal{D}}$ ($\bar{\mathcal{D}} = \mathcal{D} - \hat{\mathcal{D}}$), respectively. We use $\mathcal{Z}$ to denote a latent space, and use $g^\star : \mathcal{D} \mapsto \mathcal{Z}$ to denote the true mapping from input space $\mathcal{D}$ to latent space $\mathcal{Z}$. Besides, for a downstream task, each data x will correspond to a target $\mathbf{y} \in \mathcal{Y}$, and $f^\star : \mathcal{Z} \mapsto \mathcal{Y}$ denotes the true mapping from latent space $\mathcal{Z}$ to downstream task space $\mathcal{Y}$.

Intuitively, multimodality usually conveys extra information. For example, as shown in Figure 1, the same caption could correspond to multiple images with different matching scores, which is helpful for generating new samples. Hence, we propose to perform data augmentation for the student modality with weights computed in the latent space. However, it also works to interpolate images in the latent space [3] or input space [7] with a single modality, *i.e.*, student modality. Therefore, we follow [35] to theoretically show the advantage of multimodality latent space compared to single modality latent space at first.

Let $\mathcal{G}$ and $\mathcal{F}$ denote function class that contains the mapping from $\mathcal{D}$ to $\mathcal{Z}$ and $\mathcal{Z}$ to $\mathcal{Y}$, respectively.

$$\mathcal{G} \triangleq \{g : \mathcal{D} \mapsto \mathcal{Z}\}, \mathcal{F} \triangleq \{f : \mathcal{Z} \mapsto \mathcal{Y}\} \tag{1}$$

and $\hat{g}$ is the learned latent representation in the student modality, $g^\star, \hat{g} \in \mathcal{G}$. Given a data set S, the objective is, following the Empirical Risk Minimization [**?**], to learn g and f to minimize

$$\min \quad \mathcal{L}(f \circ g) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell(f \circ g(x_i)) \tag{2}$$

where ℓ is the loss function, it could be contrastive loss and cross-entropy loss for unsupervised and supervised tasks, respectively.

We further define the corresponding population risk [35]–[37] as

$$r(f \circ g) = \mathbb{E}_S[\mathcal{L}(f \circ g)] \tag{3}$$

And we introduce latent space quality as

$$\eta(g) = \inf_f [r(f \circ g) - r(f^\star \circ g^\star)] \tag{4}$$

where $\eta(g)$ measures the loss incurred by the distance between g and $g^\star$.

We use Rademacher complexity to quantify the population risk performance based on different modalities. Specifically, for a class of vector-valued function, $F : \mathbb{R}^d \mapsto \mathbb{R}^{\hat{d}}$, the Rademacher complexity is

$$\mathfrak{R}_m(F) = \mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(Z_i)\right]\right] \tag{5}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ with $\sigma_i \overset{iid}{\sim}$ unif $\{-1, 1\}$.

Then, we present an assumption [37] and Theorem 3.1 as follows

Assumption 3.1: The loss function ℓ is L-smooth with respect to the first coordinate and is bounded by a constant C.

Theorem 3.1: Let S be a dataset with m examples. Assuming we have produced the empirical risk minimizers $\left(\bar{f}, \bar{g}\right)$ and $\left(\hat{f}, \hat{g}\right)$, training with the K teacher modalities and 1 student modality separately. Then, for all $1 > \delta > 0$, with probability at least $1 - \frac{\delta}{2}$:

$$\begin{aligned} & r\left(\bar{f} \circ \bar{g}\right) - r\left(\hat{f} \circ \hat{g}\right) \\ & \leq \gamma_{\mathcal{S}}(K, 1) + 8L\mathfrak{R}_m(\mathcal{F} \circ \mathcal{G}) + \frac{4C}{\sqrt{m}} + 2C\sqrt{\frac{2\ln(2/\delta)}{m}} \\ & \text{where} \quad \gamma_{\mathcal{S}}(K, 1) \triangleq \eta(\bar{g}) - \eta(\hat{g}) \end{aligned} \tag{6}$$

$\gamma_{\mathcal{S}}(K, 1)$ in Eq.(6) compares the quality between latent space learning from K teacher modalities and 1 student modality of S, which bounds the difference of population risk and validates the advantage of multimodality latent space.

Different from prior methods that interpolate in the latent space [3], [34], CMDA can work with or without downstream tasks (*i.e.*, CMDA and CMDA (offline)). CMDA and CMDA (offline) learn the latent space $\mathcal{Z}$ in the same way that is based on contrastive learning. Moreover, CMDA can work with or without labels, and there may be a big gap between unsupervised latent space and supervised latent space for the same sample. Hence, we use different contrastive losses to project all samples into the common latent space in different situations. For unsupervised dataset

$$\ell_i = -\log \frac{\sum_{p=1}^{K}\sum_{q=1}^{K} 1_{[p\neq q]} \exp\left(\text{sim}\left(\boldsymbol{z}_i^{(p)}, \boldsymbol{z}_i^{(q)}\right)/\tau\right)}{\sum_{p=1}^{K}\sum_{q=1}^{K}\sum_{l=1}^{m} 1_{[p\neq q]}1_{[l\neq i]} \exp\left(\text{sim}\left(\boldsymbol{z}_i^{(p)}, \boldsymbol{z}_l^{(q)}\right)/\tau\right)} \tag{7}$$

where $sim(\cdot,\cdot)$ is cosine similarity, $1_{[p\neq q]} \in \{0,1\}$ is an indicator evaluating to 1 iff $p \neq q$ and τ denotes a temperature parameter. Different from other contrastive methods, Eq.(7) try to pull each modality of the same sample close while pushing away from other samples.

For supervised dataset

$$\ell_i = \frac{-1}{m_{y_i}-1}\sum_{j=1}^{m} 1_{[y_i=y_j]} \log \frac{\sum_{p=1}^{K}\sum_{q=1}^{K} 1_{[p\neq q]} e(i,j)^{p,q}}{\sum_{p=1}^{K}\sum_{q=1}^{K}\sum_{l=1}^{m} 1_{[p\neq q]}1_{[l\neq i]} e(i,l)^{p,q}}$$
$$\text{where} \quad e(i,j)^{p,q} = \exp\left(\text{sim}\left(\boldsymbol{z}_i^{(p)}, \boldsymbol{z}_j^{(q)}\right)/\tau\right) \tag{8}$$

In Eq.(8), m_{y_i} is the number of samples belonging to class y_i. Eq.(8) tries to pull each modality of the same class close while pushing away from other classes.

To lean common latent space $\mathcal{Z}$ with a downstream task, the loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_f + \lambda \mathcal{L}_g, \quad \text{where} \quad \mathcal{L}_g = \sum_{i=1}^{m} \ell_i \tag{9}$$

where $\mathcal{L}_f$ is the loss of downstream task. By minimizing the joint loss function, the latent space and downstream task can be iteratively optimized in a batch-by-batch manner with a stochastic gradient descent optimization algorithm. We fix $\lambda = 0.1$ in all following experiments. As for learning latent space without downstream tasks (*i.e.*, CMDA (offline)), there is no task loss in the loss function and no other differences. Intuitively, learning the common latent space with a downstream task (CMDA) always performs better than that without a downstream task, as the augmented dataset is more suitable for the downstream task. However, the cost of training downstream task with CMDA (offline) is lower as the dataset is augmented at first.

With the learned latent space, teacher modalities can guide student modality to generate new samples as

$$\widehat{x_{new}} = \sum_{j=1}^{m'} \frac{z_i^t(\hat{z}_j)^{\mathrm{T}}}{\sum_{j=1}^{m'} z_i^t(\hat{z}_j)^{\mathrm{T}}} \hat{x}_j \tag{10}$$

where $\hat{x}$ is the raw sample of student modality; $\hat{z}$ and z_i^t is the sample of student modality and one teacher modality in the latent space, respectively; m' is the number of samples used for interpolation.

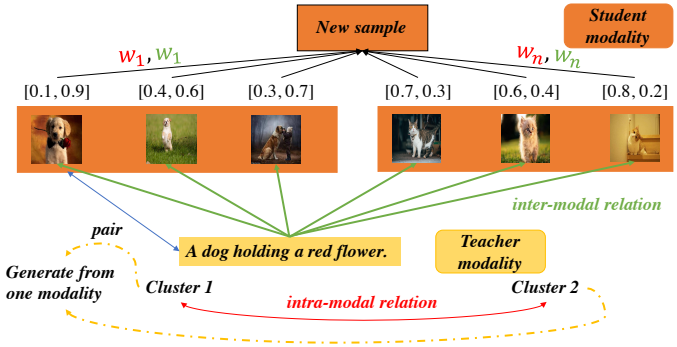


Fig. 3. Double Cross-Attention (DCA) that includes intra-modal and inter-modal relations.

Moreover, there is not only inter-modal information but also intra-modal information in multimodal datasets, which helps alleviate the meaningless interpolation. Therefore, we propose Double Cross-Attention (DCA). As shown in Figure 3, we first perform cluster in the student modality. Then, the relation between different clusters will be calculated and used as weights to generate new samples. For example, we use one sample z_i^t in the latent space from teacher modality to augment the student modality dataset as

$$\widehat{x_{new}} = \sum_{j=1}^{m'} \frac{z_i^t(\hat{z}_j)^{\mathrm{T}}}{\sum_{j=1}^{m'} z_i^t(\hat{z}_j)^{\mathrm{T}}} w_{ij}\hat{x}_j,$$
$$\text{where} \quad w_{ij} = \frac{\alpha_{j\in c_1}\phi\left(x_i \mid \theta_{j\in c_1}\right) + \alpha_{i\in c_2}\phi\left(x_i \mid \theta_{i\in c_2}\right)}{\sum_{h=1}^{M}\alpha_h\phi\left(x_i \mid \theta_h\right)} \tag{11}$$

In Eq.(11), M is the number of clusters, $j \in c_1$ denotes sample j belongs to cluster c_1, and $\alpha_{j\in c_1}$ denotes the probability that observations belong to the c_1th cluster, θ is the parameters of different clusters (i.e., mean and variance), $\phi(x|\theta)$ denotes the Gaussian distribution density function, w_{ij} is the intra-modal relation between sample i and j that calculated by the probability sample i comes from cluster $i \in c_2$ and $j \in c_1$. When implementing the algorithm, we use Expectation Maximization (EM) [38] to compute the aforementioned hyperparameters iteratively. The proposed algorithm is summarized in Algorithm 1.

IV. Experiments

In our experiments, we aim to (1) validate the effectiveness of CMDA for different tasks and modalities, (2) validate CMDA can combine with other data augmentation methods, (3) validate the effectiveness of advice given by the teacher modality, (4) validate the robustness of CMDA on various neural network architectures, (5) validate the effectiveness of DCA in CMDA by ablation experiments, (6) validate that CMDA (offline) is as valid as CMDA. In this paper, our experiments only focus on vision modality and language modality, which are the most common. Moreover, if we run tasks of vision modality, vision modality will be the student

Algorithm 1 CMDA

Input: dataset X with $K+1$ modalities and m samples. Assume $\hat{x} \in X^{(s)}$ is the student modality data and $x \in X^{\{K\}}$ is the teacher modality data.
Output: augmented data $\widehat{x_{new}}$.
For CMDA (offline), learn the latent space $\mathcal{Z}$ by Eq.(7) or Eq.(8).
For CMDA, learn the latent space $\mathcal{Z}$ by Eq.(9).
Cluster one modality of X into M clusters, c, and get distribution parameters θ, Gaussian distribution density function $\phi()$. α denotes the probability that observations belong to c.
Compute all the intra-modal weights in student modality, such as $w_{ij}^{intra} = \frac{\alpha_{j\in c_1}\phi(\hat{x_i}|\theta_{j\in c_1})+\alpha_{i\in c_2}\phi(\hat{x_i}|\theta_{i\in c_2})}{\sum_{h=1}^{M}\alpha_h\phi(\hat{x_i}|\theta_h)}$, $\hat{x_i}, \hat{x_j} \in X^{(s)}$
for $X^{(n)}$ in $X^{\{K\}}$ **do**
  for $x_i^{(n)}$ in $X^{(n)}$ **do**
    $Z^{(s)} = g^{(s)}(X^{(s)})$
    $z_i^{(n)} = g^{(n)}(x_i^{(n)})$
    Compute all the inter-modal weight as: $w^{inter} = z_i^{(n)}(Z^{(s)})^T / sum(z_i^{(n)}(Z^{(s)})^T)$
    $\widehat{x_{new}} = \sum_{j=1}^{m} w_{ij}^{inter} w_{ij}^{intra} \hat{x_j}$, WHERE $\hat{x_j} \in X^{(s)}$
  end for
end for

modality, and language modality will be the teacher modality and vice versa.

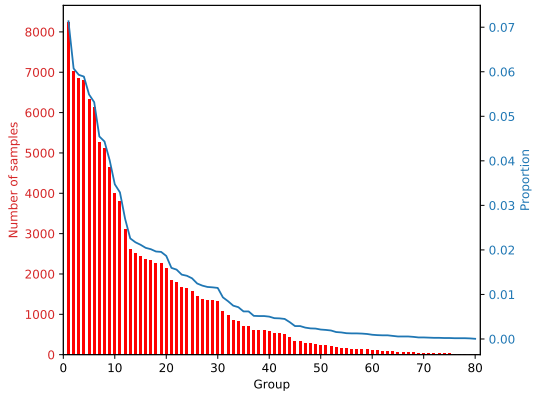


Fig. 4. Dataset of Class-COCO. There are 80 categories, each with a different number of samples.

A. Datasets and Settings

Class-COCO. A popular multimodal dataset MS COCO [39] contains images and the corresponding sentences annotated by Amazon Mechanical Turk. To run classification and anomaly detection tasks on MS COCO, we propose Class-COCO. The training set is grouped according to the objects in the images. Besides, we drop images that contain two or more objects. As shown in Figure 4, there are 126,055 samples divided into 80 groups . For image anomaly detection tasks, we select groups with a sample size greater than 6,000 as normal samples (there are 6 groups, totaling 45,205), while 15,150 samples of 50 groups are regarded as abnormal samples. As for text anomaly detection, there are 30 groups, totaling 110,905 normal samples, while 15,150 samples of 50 groups are regarded as abnormal samples. Dataset of classification tasks is divided in the same way.

Wikipedia [40] contains 2,866 image-text pairs that belong to 10 classes. For anomaly detection, we divide the dataset into normal samples (3 classes, totaling 921) and abnormal samples (7 classes, totaling 1,945). As for classification, we divide the dataset into 2 subsets: 2,273 and 593 pairs for training and testing, respectively. Note that we follow [41] to use the precomputed Wikipedia as input.

Oxford-102 [42] dataset of flower images contains 8,189 pairs of flowers from 102 different categories. For anomaly detection, we divide the dataset into normal samples (30 classes, totaling 4,018) and abnormal samples (72 classes, totaling 4,171). As for classification, we divide the ten classes with the largest amount of data into 2 subsets: 1,861 and 442 pairs for training and testing, respectively.

B. Implementation Details

In our experiments, all methods utilize the same features extractor and downstream task module (classifier for classification and detector for anomaly detection). More specifically, for Class-COCO, we use ResNet-50 and the pretrained bert [43] to extract features of vision modality and language modality, respectively. Followed by the extractor, two and three fully-connected layers are stacked to project all samples into the common latent space. Each layer of fully-connected layers follows a ReLU layer except the last one. The numbers of hidden units are 2,048, 128 and 4,096, 4,096, 128 for vision modality and language modality, respectively. We follow [44] to employ stochastic gradient descent as the optimizer with a learning rate of 0.01 for 200 epochs. Moreover, weight decay and batch size are set to 1e-4 and 128, respectively. As for Wikipedia, we follow [41] to adopt VGG-19 as the backbone for vision, and Doc2Vec [45] model as the backbone for language. The following fully-connected layers and other hyperparameters are the same as that of Class-COCO. As for Oxford-102, we follow [42] to extract the language features with a deep convolutional-recurrent text encoder, and the other settings are the same as that of Class-COCO. For all classification models, we additionally add a fully-connected layer that follows a softmax layer. Moreover, all Mixup experiments use the same hyperparameter of beta distribution, *e.g.*, $\alpha = 0.25$ that is the mean of IMAGENET experiments in [7]. It should be noted that if there is no special explanation, all experiments use the basic data augmentation methods like geometric transformation, photometric transformation, and information dropping.

C. Cross-modal Augmentation for Unsupervised Anomaly Detection

To validate the effectiveness of CMDA for unsupervised tasks, we run unsupervised anomaly detection based on the

Code and data are included in the supplemental material. Code and data will be released upon the paper acceptance.

TABLE I
RESULTS OF UNSUPERVISED ANOMALY DETECTION IN VISION MODALITY AND LANGUAGE MODALITY. THE ± SHOWS 95% CONFIDENCE INTERVAL OVER TASKS.

Dataset	Method	VISION		LANGUAGE	
		AUROC	AUPR	AUROC	AUPR
Class-COCO	Base	$76.46 \pm 1.16\%$	$88.59 \pm 0.69\%$	$73.34 \pm 0.74\%$	$78.18 \pm 0.38\%$
	Base + Mixup	$78.45 \pm 2.05\%$	$89.74 \pm 0.74\%$	$77.40 \pm 5.70\%$	$80.83 \pm 5.83\%$
	Base + CMDA (ours)	$\mathbf{83.25 \pm 1.25\%}$	$\mathbf{91.96 \pm 1.06\%}$	$\mathbf{79.64 \pm 1.64\%}$	$\mathbf{81.51 \pm 5.31\%}$
Wikipedia	Base	$73.07 \pm 1.27\%$	$95.48 \pm 0.28\%$	$89.69 \pm 1.99\%$	$98.36 \pm 0.36\%$
	Base + Mixup	$74.49 \pm 3.09\%$	$95.94 \pm 0.54\%$	$89.24 \pm 8.04\%$	$98.34 \pm 1.34\%$
	Base + CMDA (ours)	$\mathbf{76.22 \pm 2.62\%}$	$\mathbf{96.10 \pm 0.50\%}$	$\mathbf{91.48 \pm 1.18\%}$	$\mathbf{98.73 \pm 0.23\%}$
Oxford-102	Base	$64.36 \pm 0.76\%$	$89.58 \pm 0.98\%$	$63.87 \pm 0.87\%$	$90.29 \pm 0.29\%$
	Base + Mixup	$68.20 \pm 3.90\%$	$90.82 \pm 1.22\%$	$63.57 \pm 0.57\%$	$90.18 \pm 0.08\%$
	Base + CMDA (ours)	$\mathbf{70.57 \pm 1.07\%}$	$\mathbf{91.44 \pm 0.24\%}$	$\mathbf{65.60 \pm 0.60\%}$	$\mathbf{90.21 \pm 0.11\%}$

recent method, SSD [44] that learns latent space by self-supervised learning and performs detection by Mahalanobis distance. Note that text data is projected into latent space before augmentation.

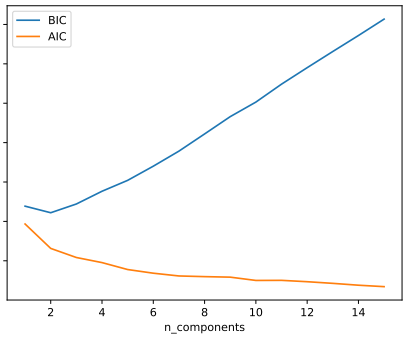


Fig. 5. BIC and AIC of Class-COCO. The blue curve denotes BIC, and the red curve denotes AIC.

We use the Akaike information criterion (AIC) and Bayesian information criterion (BIC) to decide the number of clusters in CMDA. As shown in Figure 5, for Class-COCO AIC achieves relatively stable values after 2, while BIC gets the lowest value when $n_{component} = 2$. Hence, we select 2 as the number of clusters. This hyperparameter for other datasets is computed in the same way. We get clusters in the embedding space of language modality by Gaussian mixture models (GMM) [46], where the number of EM iterations is 100, the convergence threshold is 1e-3, non-negative regularization added to the diagonal of covariance is 1e-6.

Results in Table I show that CMDA consistently outperforms baselines by a remarkable margin, especially on the largest dataset Class-COCO, CMDA improves AUROC in vision and language modalities by more than 6% compared to the Base. For the vision modality on Class-COCO, Mixup enhances the original AUROC 76.46% and AUPR 88.59% to 78.45% and 89.74%, respectively. CMDA further improves the AUROC and AUPR to 83.25% and 91.96%. Its distinguished performance from CMDA compared to that of Mixup verifies the efficacy of advice given by teacher modality under this challenging scenario. Moreover, for language modality on Wikipedia and Oxford-102, Mixup degrades the base method, which shows that interpolation with sampled weights in the latent space is easily corrupted. In contrast, with the advice of the teacher modality, the generated samples can significantly improve original results in each modality.

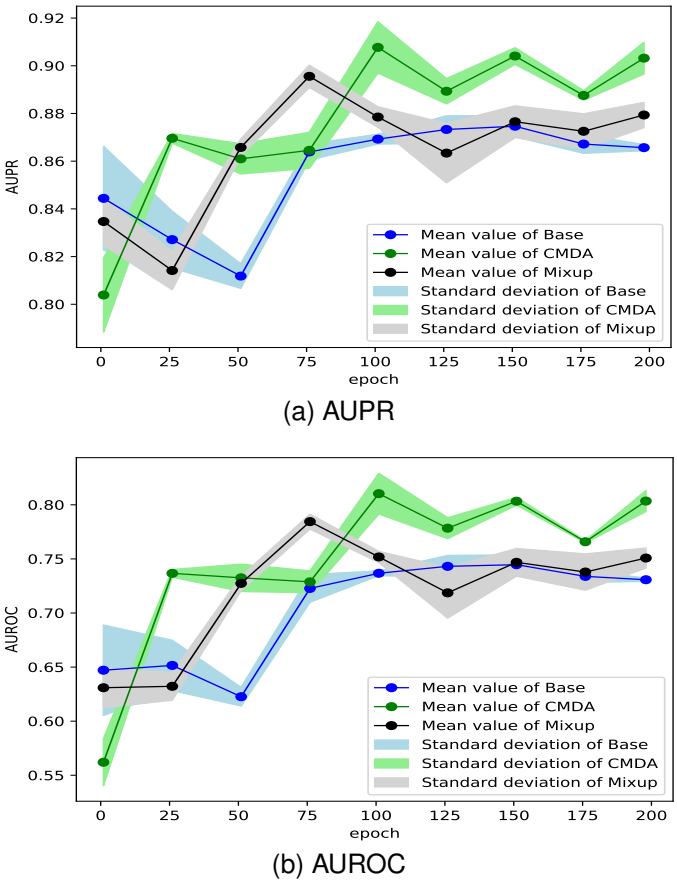


Fig. 6. Convergence of CMDA, Mixup and Base on anomaly detection task.

From the perspective of combining CMDA with other basic data augmentation methods, in Table I, Base uses geometric transformation, photometric transformation, information dropping, and adding noise, and Base + CMDA achieves better results. Thus, CMDA can be combined with other augmentation methods to achieve better results for unsupervised tasks.

We show the convergence by AUPR and AUROC in Figure 6. These promising results show that CMDA can stably exceed baselines after convergence, while Mixup and Base even degrade after about 125 and 75 epochs due to overfitting.

TABLE II
RESULTS OF SUPERVISED CLASSIFICATION IN VISION AND LANGUAGE MODALITY. THE ± SHOWS 95% CONFIDENCE INTERVAL OVER TASKS.

Dataset	Method	VISION Accuracy	LANGUAGE Accuracy
Class-COCO	Base	$15.74 \pm 0.64\%$	$83.93 \pm 0.33\%$
	+Mixup	$18.03 \pm 0.93\%$	$81.49 \pm 0.39\%$
	+CMDA	$\mathbf{18.65 \pm 0.95\%}$	$\mathbf{84.23 \pm 0.23\%}$
Wikipedia	Base	$53.39 \pm 2.49\%$	$70.82 \pm 1.92\%$
	+Mixup	$53.96 \pm 3.39\%$	$68.17 \pm 1.17\%$
	+CMDA	$\mathbf{54.79 \pm 1.09\%}$	$\mathbf{72.11 \pm 2.11\%}$
Oxford-102	Base	$27.43 \pm 2.13\%$	$51.03 \pm 0.83\%$
	+Mixup	$24.60 \pm 2.10\%$	$49.82 \pm 1.02\%$
	+CMDA	$\mathbf{28.43 \pm 1.23\%}$	$\mathbf{51.98 \pm 1.48\%}$

D. Cross-modal Augmentation for Supervised Classification

To validate the effectiveness of CMDA for supervised tasks, we run data augmentation methods with classification tasks in this section.

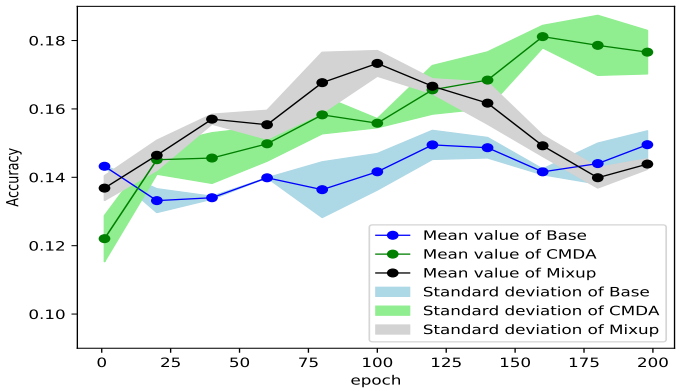
As shown in Table II, CMDA still significantly improves the results of Base in all modalities. However, Mixup severely degrades the results of language modality (text data is projected into latent space before interpolation) over all datasets, while it shows good performance in raw image space over Class-COCO. These results implicitly show that the generated samples from Mixup are corrupted with their classes, thus even degrading the classification boundary in the latent space. In contrast, CMDA can work well in both raw space and latent space with the advice of teacher modality for the supervised task. Besides, Mixup cannot work for the vision modality of Oxford-102. It may be that the hyperparameters α of the beta distribution (which decides the sampled weights for interpolation) are not applicable to this dataset. Compared to Mixup, CMDA performs well on all datasets, which shows that the robustness of CMDA is much better than that of Mixup.

From the perspective of combining CMDA with other basic augmentation methods, in Table II, Base uses geometric transformation, photometric transformation, information dropping, and adding noise, and Base + CMDA achieves better results. Thus, CMDA can be combined with other data augmentation methods to achieve better results for supervised tasks.

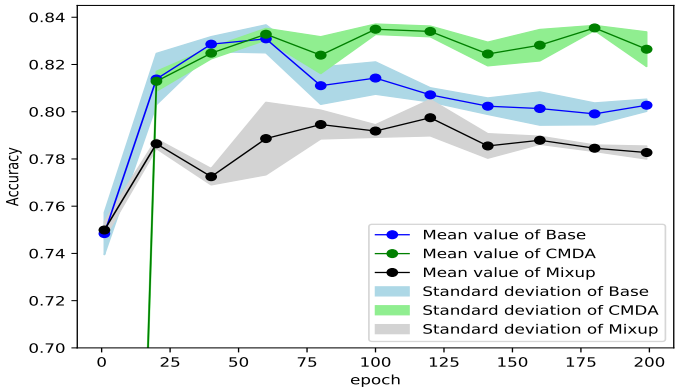
We show the convergence of CMDA, Mixup and Base on classification task for vision modality and language modality in Figure 7. In picture (a) of Figure 7, CMDA and Mixup significantly improve the Base after about 10 epochs. However, Mixup's performance gets progressively worse after about 100 epochs, while the results of CMDA still rise steadily. This may be caused by the fact that the random interpolation of Mixup cannot effectively introduce new information, resulting in the overfitting of the model. Moreover, picture (b) implicitly shows that Mixup (randomly sample weights to interpolate) cannot work well for language modality in the latent space as the green curve is much lower than the blue curve. In contrast, CMDA works well for both modalities. These results validate the effectiveness of the advice of the teacher modality.

E. Compare with More Data Augmentation Methods

We have shown that CMDA can be combined with other data augmentation methods in sections IV-C and IV-D. In this



(a) classification for vision modality



(b) classification for language modality

Fig. 7. Convergence of CMDA, Mixup and Base on classification task.

section, we compare CMDA with the aforementioned methods such as geometric transformation (GT), photometric transformation (PT), and information dropping (ID). The results are reported in Table III. Note that all the aforementioned methods only work for the raw images, and Wikipedia is precomputed [41]; thus, we only run experiments on Class-COCO and Oxford-102 in this section.

Table III shows that it is important to combine multiple methods for data augmentation, as a single method cannot improve the performance of the vanilla method well (e.g., GT and ID on Class-COCO; GT and PT on Oxford-102), while the combination of all methods achieves the best results. Meanwhile, using CMDA alone can consistently improve the performance on different datasets. In contrast, other augmentation methods do not always work. For example, using ID alone can hardly improve the performance on Class-COCO, while using ID alone can significantly improve the performance on Oxford-102. These results show the robustness of CMDA to different datasets compared to other methods.

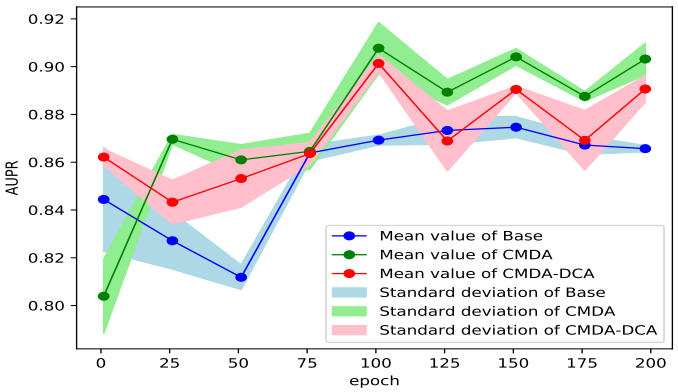
F. Ablation

We show the effectiveness of Double Cross-Attention (DCA) by Figure 8. It can be seen that DCA helps the training process more stable as the green curve is more stable than the red curve. These results validate the intuition that "randomly sample may degrade the semantic representation of new samples" in section I. In other words, intra-modal relation

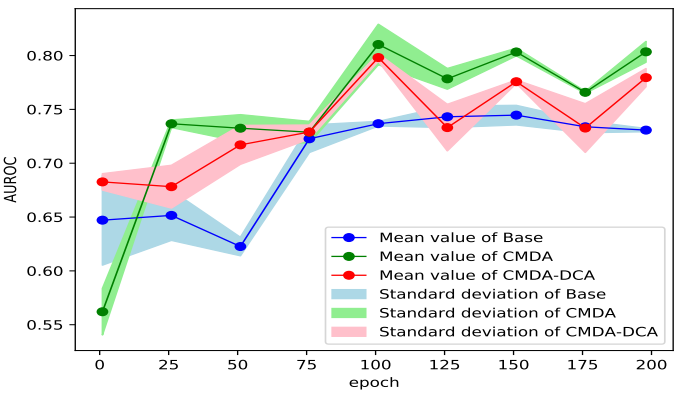
TABLE III
COMPARE CMDA WITH OTHER BASIC DATA AUGMENTATION METHODS, WHERE GT DENOTES GEOMETRIC TRANSFORMATION, PT DENOTES PHOTOMETRIC TRANSFORMATION, AND ID DENOTES INFORMATION DROPPING. THE ± SHOWS 95% CONFIDENCE INTERVAL OVER TASKS. NOTE THAT ALL THE AFOREMENTIONED METHODS ONLY WORK FOR THE RAW IMAGES, AND WIKIPEDIA IS PRECOMPUTED AS [41]; THUS, WE ONLY RUN EXPERIMENTS ON CLASS-COCO AND OXFORD-102. BESIDES, VANILLA DENOTES A METHOD WITHOUT DATA AUGMENTATION.

Dataset	Method	AUROC	AUPR
Class-COCO	Vanilla	$64.74 \pm 3.74\%$	$85.06 \pm 2.56\%$
	+ GT	$63.18 \pm 2.28\%$	$84.16 \pm 1.46\%$
	+PT	$66.17 \pm 5.17\%$	$84.76 \pm 1.46\%$
	+ID	$63.09 \pm 10.09\%$	$83.79 \pm 7.49\%$
	+CMDA	$73.39 \pm 5.39\%$	$87.52 \pm 3.52\%$
	+all	$\mathbf{83.25 \pm 1.25\%}$	$\mathbf{91.96 \pm 1.06\%}$
Oxford-102	Vanilla	$58.35 \pm 4.45\%$	$86.80 \pm 1.80\%$
	+ GT	$57.21 \pm 3.31\%$	$86.60 \pm 1.80\%$
	+PT	$57.01 \pm 0.71\%$	$86.44 \pm 0.34\%$
	+ID	$66.94 \pm 1.74\%$	$90.75 \pm 0.55\%$
	+CMDA	$63.39 \pm 2.29\%$	$88.99 \pm 0.99\%$
	+all	$\mathbf{70.57 \pm 1.07\%}$	$\mathbf{91.44 \pm 0.24\%}$

weight in DCA always selects samples from other clusters that are not too different from the original sample, which makes results more stable.



(a) AUPR



(b) AUROC

Fig. 8. Convergence of CMDA, CMDA-DCA and Base on anomaly detection task.

G. Augmentation with Different Architectures

We follow [47], [48] to run data augmentation experiments on more neural network architectures. Table IV shows the results of Oxford-102 on Deep residual network (ResNet, includes ResNet-18, ResNet-50, ResNet-101) [49], Visual Geometry Group (VGG, includes VGG-11, VGG-16, VGG-19) [50], and Dense Convolutional Network (DenseNet, includes DenseNet-121, DenseNet-169, DenseNet-201) [51]. Results indicate that models trained with CMDA have significant improvement, demonstrating that our method is applicable to various architectures. Specifically, on ResNet-18, ResNet-50 and ResNet-101, CMDA enhances the AUROC 66.74%, 64.36% and 63.92% to 71.96%, 70.57% and 66.38%, respectively. For three DenseNet architectures, CMDA enhances the original AUROC about 66% to about 70%. As for VGG, CMDA improve the 64.83% AUROC to 70.94% on VGG-11, while improve 59.95% and 58.36% to 65.21% and 62.97% on VGG-16 and VGG-19, respectively. The same trend for AUPR. These results verify that CMDA can significantly improve baseline performance on various network architectures.

H. Augmentation with More Complicated Task

This section compares CMDA with other methods on more complicated task. In section IV-C, we run unsupervised anomaly detection on Class-COCO, where normal samples belong to 6 groups. To demonstrate the potential of CMDA for more complicated problems, we set the normal samples to 30 groups while keeping other settings, and the results are presented in Table V.

As shown in Table V, when $Groups = 30$, the margin of AUROC and AUPR between CMDA and Base is 9.98% and 8.09%, respectively, while when $Groups = 6$, the margin is 6.79% and 3.37%. It can be seen that with the task complexity increasing, CMDA achieves better performance on Class-COCO. Although Mixup also gets better results, as the margin becomes 3.25% from 1.99% of AUROC and 2.27% from 1.15% of AUPR, the improvements of Mixup are much lower than that of CMDA.

I. Offline Data Augmentation

Data augmentation methods that interpolate in the latent space [3], [34] always need to learn latent space with the downstream task, which may significantly increase the cost of training downstream task models. Therefore, it is crucial to augment the dataset offline. Different from prior methods, CMDA augments the dataset by the extra information across different modalities, which enables offline data augmentation.

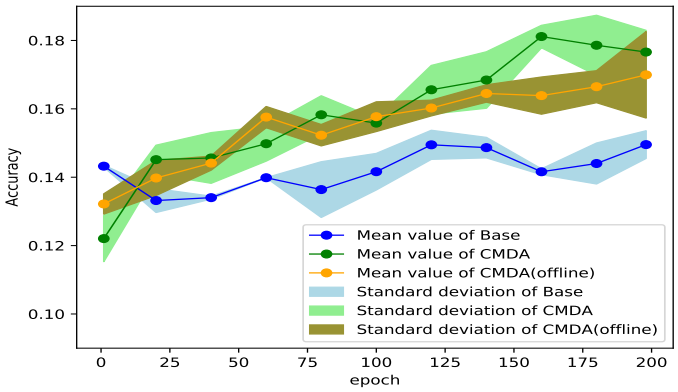
In this section, we augment the dataset at first. Then, we run downstream tasks based on the new dataset. As shown in Table VI, CMDA (offline) gets similar results as CMDA. More specifically, CMDA consistently gets better results than CMDA (offline), as the latent space learned by CMDA also works for downstream tasks. Thus, the new data will be more applicable. Besides, CMDA (offline) still significantly improves Base, except for language modality on Class-COCO, the results of CMDA (offline) are slightly lower than that of Base. This may result from the request for the high quality of latent space to perform text interpolation, as text classification results on Class-COCO are much higher than that of other datasets.

TABLE IV
MORE RESULTS OF ANOMALY DETECTION OVER OXFORD-102 ON DIFFERENT ARCHITECTURES

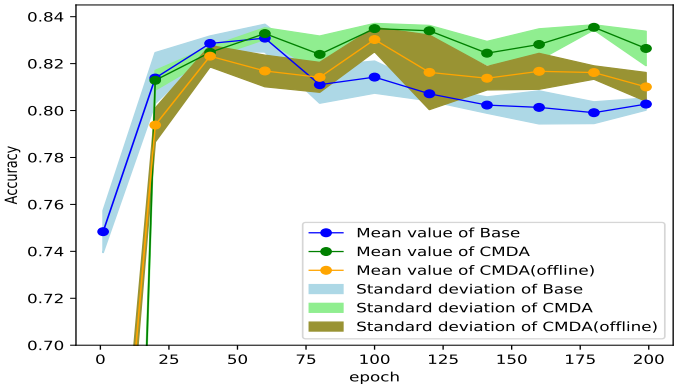
Model	Model Size	FLOPs	Base		CMDA	
			AUROC	AUPR	AUROC	AUPR
ResNet-18	$11.17M$	$27.28G$	$66.74 \pm 0.94\%$	$90.58 \pm 0.58\%$	$\mathbf{71.96 \pm 1.16\%}$	$\mathbf{92.12 \pm 0.72\%}$
ResNet-50	$23.50M$	$63.93G$	$64.36 \pm 0.76\%$	$89.58 \pm 0.98\%$	$\mathbf{70.57 \pm 1.07\%}$	$\mathbf{91.44 \pm 0.24\%}$
ResNet-101	$42.49M$	$123.49G$	$63.92 \pm 1.92\%$	$89.28 \pm 0.98\%$	$\mathbf{66.38 \pm 1.68\%}$	$\mathbf{90.27 \pm 1.07\%}$
DenseNet-121	$13.51M$	$2.89G$	$66.18 \pm 0.58\%$	$90.03 \pm 0.23\%$	$\mathbf{71.47 \pm 1.87\%}$	$\mathbf{91.59 \pm 0.29\%}$
DenseNet-169	$20.35M$	$3.42G$	$66.97 \pm 0.77\%$	$90.28 \pm 0.48\%$	$\mathbf{70.21 \pm 1.51\%}$	$\mathbf{91.24 \pm 0.84\%}$
DenseNet-201	$26.49M$	$4.37G$	$66.24 \pm 1.24\%$	$90.19 \pm 0.39\%$	$\mathbf{70.78 \pm 0.48\%}$	$\mathbf{91.27 \pm 0.27\%}$
VGG-11	$124.84M$	$7.62G$	$64.83 \pm 1.33\%$	$89.64 \pm 0.34\%$	$\mathbf{70.94 \pm 3.14\%}$	$\mathbf{91.92 \pm 0.92\%}$
VGG-16	$130.34M$	$15.50G$	$59.95 \pm 2.35\%$	$88.03 \pm 1.63\%$	$\mathbf{65.21 \pm 6.01\%}$	$\mathbf{90.24 \pm 1.84\%}$
VGG-19	$135.65M$	$19.67G$	$58.36 \pm 3.16\%$	$87.42 \pm 1.42\%$	$\mathbf{62.97 \pm 4.67\%}$	$\mathbf{90.68 \pm 5.68\%}$

TABLE V
RESULTS OF MORE COMPLICATED UNSUPERVISIED TASK IN VISION MODALITY. THE $\pm$ SHOWS 95% CONFIDENCE INTERVAL OVER TASKS.

Groups	Method	AUROC	AUPR
6	Base	$76.46 \pm 1.16\%$	$88.59 \pm 0.69\%$
	+Mixup	$78.45 \pm 2.05\%$	$89.74 \pm 0.74\%$
	+CMDA	$\mathbf{83.25 \pm 1.25\%}$	$\mathbf{91.96 \pm 1.06\%}$
30	Base	$65.50 \pm 2.90\%$	$67.68 \pm 2.18\%$
	+Mixup	$68.75 \pm 1.65\%$	$69.95 \pm 1.65\%$
	+CMDA	$\mathbf{75.48 \pm 1.38\%}$	$\mathbf{75.77 \pm 1.17\%}$



(a) vision modality



(b) language modality

Fig. 9. Convergence of CMDA, CMDA (offline) and Base on classification task.

We show the convergence of CMDA (offline), CMDA, and Base in vision modality on classification task in Figure 9. Note that CMDA (offline) performs 100 epochs ahead of time to learn the common latent space. When training the downstream task model, CMDA always gets better results than CMDA (offline), which validates the importance of latent space to generate new samples. Moreover, for vision modality, after about 110 epochs, CMDA and CMDA (offline) both significantly outperform the Base method. As for language modality, CMDA (offline) outperforms Base after 75 epochs, as Base drastically overfits on the trained dataset. Figure 9 shows that even if the result of CMDA (offline) is not as good as CMDA, it can significantly improve the performance of Base without increasing the cost of training downstream task. In addition, the results of CMDA (offline) also verify that the improvement profits from the augmented data.

TABLE VI
RESULTS INCLUDE CMDA (OFFLINE) OF SUPERVISED CLASSIFICATION IN VISION MODALITY AND LANGUAGE MODALITY. THE $\pm$ SHOWS 95% CONFIDENCE INTERVAL OVER TASKS.

Dataset	Method	VISION Accuracy	LANGUAGE Accuracy
Class-COCO	CMDA	$\mathbf{18.65 \pm 0.95\%}$	$\mathbf{84.23 \pm 0.23\%}$
	CMDA (offline)	$18.30 \pm 1.70\%$	$83.80 \pm 0.10\%$
Wikipedia	CMDA	$\mathbf{54.79 \pm 1.09\%}$	$\mathbf{72.11 \pm 2.11\%}$
	CMDA (offline)	$54.75 \pm 1.15\%$	$71.80 \pm 1.40\%$
Oxford-102	CMDA	$\mathbf{28.43 \pm 1.23\%}$	$\mathbf{51.98 \pm 1.48\%}$
	CMDA (offline)	$27.97 \pm 0.87\%$	$51.67 \pm 1.77\%$

J. Data Augmentation for Multi-modal Task

In this section, we validate the effectiveness of CMDA on the image caption task. We follow [52] to implement the image caption part, except the output size of an encoder is 2048 for convenience. As for the data augmentation part, we use the pre-trained CLIP [53] to align samples from vision modality and language modality.

Results are presented in Figure 10. Compared to established baselines, our approach yields significant performance improvements. More specifically, the lowest value of CMDA is higher than the highest value of Base. In addition, the results range of CMDA is significantly smaller than that of Base, which indicates that CMDA is more stable. All these results show that the proposed augmentation method is generic and can easily be incorporated into various tasks.

It is noteworthy that neither CMDA nor other data augmentation algorithms can be directly applied to the image caption task due to the demanding requirements of fine-grained matching between modalities. However, CMDA works when it combines with an attenuation coefficient, making the

model more focused on raw data with increasing epochs. This phenomenon can be explained by curriculum learning [54]. When the epoch is low, the model trains with mixed samples, and there are lower requirements for fine-grained matching, making models learn more generalizable features. With the increasing epochs, the proportion of raw data is getting higher, and it is easier for the model to learn specific features based on the generalizable features.

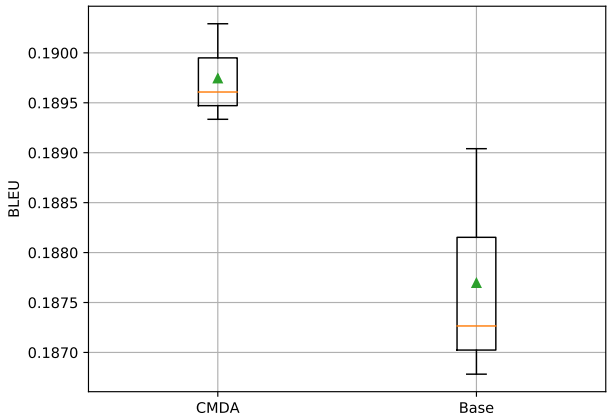


Fig. 10. Box-plot of CMDA and Base on the image caption task. The triangle denotes the median, the boundary of the lower whisker denotes the minimum value, and the boundary of the upper whisker denotes the maximum value.

K. Computational Complexity Analysis

In this paper, we mainly focus on vision modality and language modality, thus, we perform computational complexity analysis based on these modalities. As mentioned in section IV-J, we can use the pretrained CLIP [53] to align samples from vision modality and language modality, and perform data augmentation. Thus, the computationally intensive computation is clustering and computing similarity of samples.

As for clustering, we perform GMM in the embedding space in one modality, and the complexity is of $O(mM\hat{d}^3)$, where m is the number of data points in a batch, M is the number of Gaussian components and $\hat{d}$ is the embedding dimension (for CLIP, $\hat{d} = 512$).

The similarity of samples is computed by multiplying two matrices where each matrix is $m \times \hat{d}$, and the complexity is of $O(\hat{d}m^2)$.

Therefore, the overall complexity is $O(mM\hat{d}^3 + \hat{d}m^2)$. As m and $\hat{d}$ will not be too large, the complexity of CMDA is acceptable.

L. Hyperparameters Analysis

One of the most important hyperparameters is the number of clusters M; thus, we show the results with different M in Figure 11. It can be seen that both AUROC and AUPR get the best performance when $M = 2$, which is consistent with the conclusions of AIC and BIC in section IV-C. Besides, this experiment shows the effectiveness of DCA, as CMDA with $M > 1$ gets better results than that with $M = 1$. That is, DCA selects samples from close clusters and thus alleviates the meaningless interpolation.

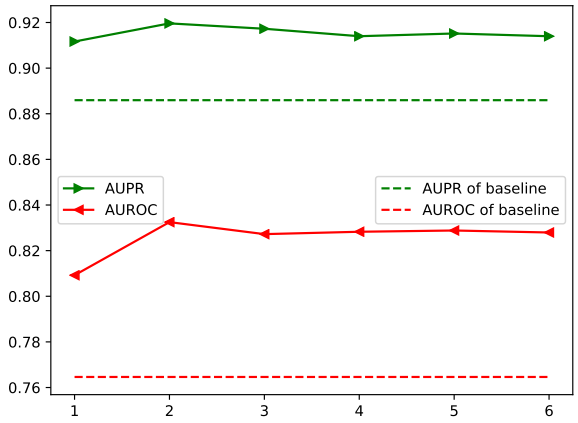


Fig. 11. Averaged results with different M on anomaly detection task.

V. Conclusion

This paper proposes a generic framework, CMDA, to perform data augmentation for different individual modalities and tasks to alleviate the over-fitting issue and improve the generalization performance. More specifically, CMDA uses extra information in one modality that can provide semantic directions to generate more meaningful samples in another modality. Extensive experiments on unsupervised anomaly detection task, supervised classification task and image caption in vision modality and language modality validate the effectiveness of CMDA. In addition, we run CMDA on various neural network architectures, which show the robustness of CMDA. Besides, CMDA (offline) that does not need to train with the downstream tasks module avoids the concerns about the computational cost of training downstream tasks.

Acknowledgments

This work has been supported in part by the Zhejiang NSF (LR21F020004), NSFC (No. 62272411), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Chinese Knowledge Center of Engineering Science and Technology (CKCEST).

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, no. 1, pp. 1–34, 2021.

[3] T.-H. Cheung and D.-Y. Yeung, "Modals: Modality-agnostic automated data augmentation in the latent space," in *International Conference on Learning Representations*, 2021.

[4] X. Liu, Y. Zou, L. Kong, Z. Diao, J. Yan, J. Wang, S. Li, P. Jia, and J. You, "Data augmentation via latent space interpolation for image classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 728–733.

[5] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell, "A closer look at feature space data augmentation for few-shot intent classification," *arXiv preprint arXiv:1910.04176*, 2019.

[6] P. Upchurch, J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Weinberger, "Deep feature interpolation for image content changes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7064–7073.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[8] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[9] F. Ye, C. Huang, J. Cao, M. Li, Y. Zhang, and C. Lu, "Attribute restoration framework for anomaly detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 116–127, 2022.

[10] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1445–1454.

[11] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3232–3240.

[12] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1059–1071, 2016.

[13] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 9, pp. 2372–2385, 2017.

[14] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 154–162.

[15] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks." in *IJCAI*, 2016, pp. 3846–3853.

[16] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.

[17] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.

[18] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, "Multiview metric learning with global consistency and local smoothness," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–22, 2012.

[19] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 27, no. 1, 2013, pp. 1198–1204.

[20] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[21] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[22] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[23] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.

[24] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1979–1988.

[25] S. Ye, D. Chen, S. Han, and J. Liao, "3d question answering," *arXiv preprint arXiv:2112.08359*, 2021.

[26] ——, "Learning with noisy labels for robust point cloud segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6443–6452.

[27] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European Conference on Computer Vision*. Springer, 2020, pp. 202–221.

[28] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[29] C. Yan, G. Pang, X. Bai, C. Liu, X. Ning, L. Gu, and J. Zhou, "Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss," *IEEE Transactions on Multimedia*, vol. 24, pp. 1665–1677, 2021.

[30] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2017.

[31] D. Wang, C. Zhang, Q. Wang, Y. Tian, L. He, and L. Zhao, "Hierarchical semantic structure preserving hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, 2022.

[32] X. Guo, W.-K. A. Kong, and A. C. Kot, "Deep multimodal sequence fusion by regularized expressive representation distillation," *IEEE Transactions on Multimedia*, 2022.

[33] S. Li, B. Zhang, L. Fei, S. Zhao, and Y. Zhou, "Learning sparse and discriminative multimodal feature codes for finger recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[34] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.

[35] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[36] M. R. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," *Advances in neural information processing systems*, vol. 22, 2009.

[37] N. Tripuraneni, M. Jordan, and C. Jin, "On the theory of transfer learning: The importance of task diversity," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7852–7862, 2020.

[38] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[40] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 251–260.

[41] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5403–5413.

[42] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[43] H. Xiao, "bert-as-service," https://github.com/hanxiao/bert-as-service, 2018.

[44] V. Sehwag, M. Chiang, and P. Mittal, "Ssd: A unified framework for self-supervised outlier detection," *arXiv preprint arXiv:2103.12051*, 2021.

[45] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," *arXiv preprint arXiv:1607.05368*, 2016.

[46] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[47] C. Gong, T. Ren, M. Ye, and Q. Liu, "Maxup: Lightweight adversarial training with data augmentation improves neural network training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2474–2483.

[48] J. Han, P. Fang, W. Li, J. Hong, M. A. Armin, I. Reid, L. Petersson, and H. Li, "You only cut once: Boosting data augmentation with a single cut," *arXiv preprint arXiv:2201.12078*, 2022.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[52] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[54] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.