Pace-Adaptive and Noise-Resistant Contrastive Learning for Multimodal Feature Fusion

Xiaobao Guo, Alex Kot, *Life Fellow, IEEE*, and Adams Wai-Kin Kong, *Member, IEEE*

***Abstract*—Multimodal feature fusion aims to draw complementary information from different modalities to achieve better performance. Contrastive learning is effective at discriminating coexisting semantic features (positive) from irrelative ones (negative) in multimodal signals. However, positive and negative pairs learn at separate rates, which undermines the overall performance of multimodal contrastive learning (MCL). Moreover, the learned representation model is not robust, as MCL utilizes supervision signals from potentially noisy modalities. To address these issues, a novel multimodal contrastive learning objective, Pace-adaptive and Noise-resistant Noise-Contrastive Estimation (PN-NCE), is proposed for multimodal fusion by directly using unimodal features. PN-NCE encourages the positive and negative pairs reaching to their optimal similarity scores adaptively and shows less susceptibility to noisy inputs during training. A theoretical analysis is performed on its robustness. Maximizing modality invariance information in the fused representation is expected to benefit the overall performance and therefore, an estimator that measures the difference between the fused representation and its unimodal representations is integrated into MCL to obtain a more modality-invariant fusion output. The proposed method is model-agnostic and can be adapted to various multimodal tasks. It also bears less performance degradation when reducing the number of training samples at the linear probing stage. With different networks and modality inputs from three multi-modal datasets, experimental results show that PN-NCE achieves consistent enhancements compared with previous state-of-the-art approaches.**

***Index Terms*—Multimodal fusion, multimodal contrastive learning, modality invariance.**

I. Introduction

HUMAN perception and reasoning about the world are usually through processing multimodal information. Although combining high-dimensional multimodal signals is natural and easy for humans, it is a challenge for current AI systems [1], [2], [3]. It is crucial for an intelligent system to effectively fuse useful information from multiple modalities for various downstream tasks. With the encouraging success of contrastive learning in image recognition [4], [5], [6], [7], [8], natural language processing [9], [10], [11], [12] and audio signal processing [13], [14], [15], recent works [15], [16], [17] have started to explore its application on multimodal data and showed its possibility for multimodal fusion.

Similar to traditional contrastive learning on unimodal representation learning, the discrimination between positive and negative pairs relies on the hypothesis that modalities from the same scenes contain identical or highly positive correlated semantic information while modalities from different scenes are negative correlated [20]. Multimodal contrastive learning (MCL) distinguishes the positive pairs from the negative ones, thus learning the shared information from different modalities in a self-supervised manner. In other words, paired modalities provide supervision to each other during training. Traditional contrastive learning methods [16], [17], [21], [22], [23] leverage on InfoNCE (Information Noise-Contrastive Estimation) loss [15] to learn the distribution of positive samples by comparing it against a noisy distribution. More concretely, the contrastive learning is conducted by pulling the elements in a positive pair closer while pushing the positive pairs further from the negative ones (see an illustration in Fig. 1).

However, the positive and negative pairs learning usually converge at separate rates. MCL models are prone to shortcut learning, which undermines the overall performance [24]. As modalities from the same scene are usually more correlated and easier to learn than the negative ones, MCL using traditional InfoNCE may lead to suboptimal solutions as it learns the positive and negative pairs by the same way. Table I shows the training accuracies of positive and negative pairs at the MCL stage and also the linear probing accuracies on the test set. InfoNCE has a large gap between the training accuracies of the positive and negative pairs. The large difference between their average gradients indicates the imbalanced learning paces. Table I also shows that the linear probing accuracy of InfoNCE on the test set is lower. Moreover, the supervisions between modalities in a pair may not always be reliable when a certain modality is noisy or

Manuscript received 12 September 2022; revised 2 January 2023; accepted 21 February 2023. Date of publication 3 March 2023; date of current version 15 December 2023. This work was supported by the Ministry of Education, Singapore, through the Academic Research Fund Tier 1, RG73/21. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Yadong Mu. *(Corresponding author: Adams Wai-Kin Kong.)*

Xiaobao Guo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, and also with the Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore 639798 (e-mail: XIAOBAO001@e.ntu.edu.sg).

Alex Kot is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eackot@ntu.edu.sg).

Adams Wai-Kin Kong is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: AdamsKong@ntu.edu.sg).

Digital Object Identifier 10.1109/TMM.2023.3252270

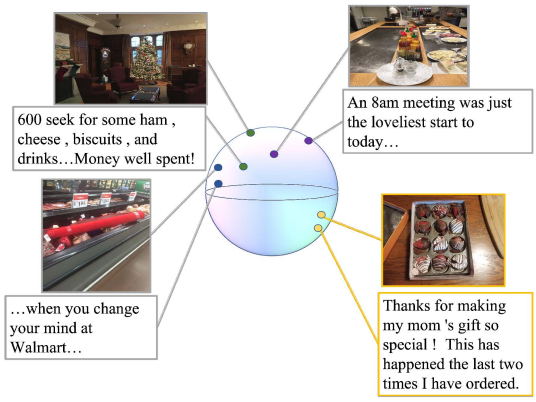

Fig. 1. Multimodal contrastive learning. Modalities from the same scene are pulled closer while modalities from different scenes are pushed away in the embedding space (e.g., the positive pair is highlighted in yellow, which is a meaningful yet sarcastic image-text pair, and the negative pairs are highlighted in gray). Examples are taken from Sarcasm dataset [19].

TABLE I
TRAINING ACCURACIES (ACC %) AND AVERAGE GRADIENTS (GRAD) OF THE LAST TRAINING EPOCH OF POSITIVE AND NEGATIVE PAIRS AT THE MCL STAGE. TEST ACCURACIES AT THE LINEAR PROBING STAGE. EXPERIMENTS ARE CONDUCTED ON VIOLIN DATASET [18]

Method	Pos-Acc (Grad)	Neg-Acc (Grad)	Test Acc
InfoNCE	66.04 (0.9374)	52.28 (3.1245e-6)	61.25
Ours	72.45 (0.2187)	69.37 (1.5620e-2)	67.25

corrupted. Thus, it is critical to balance the learning pace for positive and negative samples and improve robustness of MCL. At the same time, from the perspective of multimodal fusion [17], [21], [25], [26], [27], [28], [29], modality-invariant information [30] is expected to be maintained and maximized, but this problem is largely unexplored in MCL.

To address the aforementioned issues, in this work, a novel MCL objective named Pace-adaptive and Noise-resistant Noise-Contrastive Estimation (PN-NCE) is presented for multimodal feature fusion. Our goal is to produce effective yet robust multimodal fused representations via MCL by directly utilizing extracted multimodal features as inputs. For a particular multimodal fusion task, the learning process is divided into two stages. In the first stage, self-supervised MCL is conducted on trained features to produce fused representations. Particularly, MCL is regarded as a binary classification problem with noisy labels, where PN-NCE is introduced as the objective function. PN-NCE is able to mitigate the imbalanced learning pace of positive and negative pairs. It adjusts the learning pace of positive and negative pairs dynamically, according to pace factors estimated in each iteration. PN-NCE is also effective at improving the robustness of MCL against noisy inputs, where we propose a cooperative classifier ensemble for robust MCL with theoretical proof. Instead of training each binary classifier with individual positive or negative input, we propose a cooperative classifier ensemble that regards the MCL problem as training K binary classifiers with an orderless score list as input for each classifier, where each classifier satisfies the symmetric property. To obtain a more modality-invariant fusion output, an explicit distance estimation between the fused positive representation and its unimodal representations is integrated with PN-NCE. To achieve this, the fused representation of the positive sample is mapped back to its unimodal representations via transformation networks. Then the similarity between the output and each unimodal representation is evaluated by an *L1* distance. This operation serves as an auxiliary task in MCL, which is discarded at the second stage.

In the second stage, linear probing is conducted by training simple linear layers. By freezing the parameters of the MCL model trained at the first stage, the linear layers take only the fused representations as input to perform downstream tasks. With simple classifiers, the proposed method shows the superior capability of fusing multimodal features. The experiments are conducted on several multimodal fusion tasks that cover video, language, and audio modalities [18], [19], [31]. Using the same extracted multimodal features, the fused representations learned with PN-NCE achieve better linear probing results, compared with InfoNCE and its variants. The proposed method also outperforms the compared methods, even when the number of training samples is reduced at the linear probing stage. Furthermore, the proposed method for multimodal fusion is independent of network architecture, and hence can be applied to various multimodal fusion tasks.

In short, the contributions of this paper are summarized as:

- a novel PN-NCE objective is presented for efficient and robust MCL, which is model-agnostic and directly deals with unimodal features;
- the pace-adaptive learning for positive and negative pairs is proposed to address the imbalanced learning issue;
- the cooperative classifier ensemble is proposed for robustness with a theoretical proof;
- modality invariance estimator is explicitly integrated into MCL for more reliable fused outputs;
- experimental results on three multimodal fusion tasks demonstrate that the proposed method achieves consistent improvement compared with other state-of-the-art methods.

II. RELATED WORK

A. Multimodal Contrastive Learning

Contrastive learning [32], [33], [34] aims to learn useful representations through directly comparing positive and negative samples. It is a promising approach to representation learning. Previous works [4], [5], [7], [9], [10], [12], [13], [14], [23] enhanced contrastive learning for unimodal representation learning from different perspectives. Some works [10], [12] focused on generating more diverse negative samples to improve the discriminating capability of the trained models, while others adjusted the training strategies [4], [5] or modified the objective functions [7], [9], [13], [14], [23] to boost performance.

With the gradual deepening of multimodal research, recent works [15], [16], [17], [21] extended contrastive learning to multimodal learning tasks. CPC [15] combined predictive coding with a probabilistic contrastive loss [35] on a variety of data

modalities. CMC [16] proposed to maximize the mutual information among multiple views to capture the shared semantics. MMV-FAC [17] created a modality embedding graph where semantic comparisons were performed by dot products of different modalities. TupleInfoNCE [21] proposed tuple disturbing for multimodal data augmentation and extended InfoNCE loss for a better comparison between modality vectors. Unlike the previous MCL models that take raw data as input, we consider MCL directly on the extracted features for multimodal fusion tasks. Although previous works attempted to explore the shared information across modalities in MCL, the methods fail to account for the imbalanced convergence rates between positive and negative samples or resist noisy inputs. This may jeopardize the performance of the multimodal fusion for downstream tasks.

B. Multimodal Fusion

Multimodal fusion is one of the core research topics in multimodal representation learning [20]. It aims to draw the underlying useful information from multiple modalities for various tasks. Previous study [36] has summarized the basic fusion strategies, including early fusion and late fusion. Early fusion usually concatenates the input-level features before learning a concept, while late fusion focuses on learning the concept from individual modalities and then integrating the results. Recently, research works [21], [25], [27], [29], [37], [38] proposed more promising fusion mechanisms using sophisticated strategies. LMF [37] produced the multimodal representation by performing low-rank multimodal fusion with modality-specific factors. MFN [25] took both view-specific and cross-view interactions into consideration and proposed a fusion architecture based on LSTM [39]. MulT [29] addressed the issue of cross-modal fusion via the directional pairwise crossmodal attention technique based on transformer encoders [40]. Bottleneck fusion [38] combined the unimodal representations through multiple attention layers to collate and condense the shared information across modalities. CEN [27] proposed to measure the individual channel importance and dynamically exchange channels between sub-networks of different modalities to learn the crossmodal semantics.

Unlike previous works designing sophisticated architectures for fusion, our proposed method leverages a simple fusion module with concatenated unimodal representations. We focus on producing better-fused output by improving MCL, which is model-agnostic. The modality-invariant information is explored within MCL stage followed by simple linear probing layers as opposed to complicated architectures for different downstream tasks.

III. Approach

In this section, we begin with describing the multimodal contrastive learning problem and revisiting the classic InfoNCE loss for MCL. We then present the PN-NCE objective to deal with the imbalanced learning paces of positive and negative samples. We explain the robust property of PN-NCE against noisy modalities, where a cooperative classifier ensemble is proposed. We also introduce a modality invariance estimator (MIE) to produce a modality-invariant fused output. The proposed MCL for multimodal fusion is shown in Fig. 2, where PN-NCE objective and MIE are applied.

A. Preliminary

Given two modalities M and N, multimodal contrastive learning between a positive pair and its K negative pairs is regarded as a binary classification problem, where the label is assigned to *1* if the modalities are sampled from the joint distribution, i.e., $(m,n)\sim P_{MN}$, and is assigned to *-1* if the modalities are from the marginal distribution product, i.e., $(m,\tilde{n})\sim P_M P_N$. For one positive pair (m,n) and its corresponding negative pairs $\{(m,\tilde{n}_i)\}_{i=1}^K$, the positive score is denoted as s^+ and the negative scores are denoted as $\{s_i^-\}_{i=1}^K$. The task is to select the only positive pair from the set $\Omega=\{(m,n),(m,\tilde{n}_1),(m,\tilde{n}_2),\ldots,(m,\tilde{n}_K)\}$, which contains $(K+1)$ pairs. The classic InfoNCE loss [15] function is minimized as follows:

$$\mathcal{L}_{InfoNCE}(\mathbf{s})=-\mathop{\mathbb{E}}_{\Omega}\left(log\frac{e^{s^+}}{e^{s^+}+\sum_{i=1}^K e^{s_i^-}}\right), \tag{1}$$

where $\mathbf{s}$ is the set of similarity scores between the modalities, i.e., $\mathbf{s}=\{s^+,s_1^-,s_2^-,\ldots,s_k^-\}$. Specifically, $s^+=h_1(m)^\intercal h_2(n)/\tau$ and $s_i^-=h_1(m)^\intercal h_2(\tilde{n}_i)/\tau$. h_1 and h_2 are two different encoders to produce the embeddings for modality M and N while τ is the temperature parameter to adjust the dynamic range. Therefore, for MCL, the trained model learns to classify the positive pair from all the input pairs by maximizing s^+ while minimizing $\{s_i^-\}_{i=1}^K$.

However, the positive pair and the negative pairs reach their optimal scores at separate rates. On one hand, this is because K is often set as a large value in practice to obtain promising results; on the other hand, modalities from the same scene are more correlated and easier to learn than the negative pairs. These may lead to sub-optimal solutions. To mitigate this issue, we introduce PN-NCE objective function, which is elaborated in the following sections.

B. Pace-Adaptive Noise-Contrastive Estimation for MCL

To alter the learning paces for positive pairs and negative pairs in MCL, we first introduce P-NCE (Pace-adaptive NCE) objective that enables each pair to update its similarity score in an adaptive manner. We define the optimal score for positive pair as $O^+=1$ and that for K negative pairs as $\{O_i^-=0+\epsilon\}_{i=1}^K$, where ϵ is a small positive constant. The P-NCE objective is formulated by:

$$\mathcal{L}_{P-NCE}(\mathbf{s})=-\mathop{\mathbb{E}}_{\Omega}\left(log\frac{e^{\alpha s^+}}{e^{\alpha s^+}+\sum_{i=1}^K e^{\beta_i s_i^-}}\right),$$
$$\text{s.t. }\ \alpha+\sum_{i=1}^K\beta_i=1, \tag{2}$$

in which α and β_i are the non-negative pace estimation factors for reweighting the positive and negative scores. Concretely, α

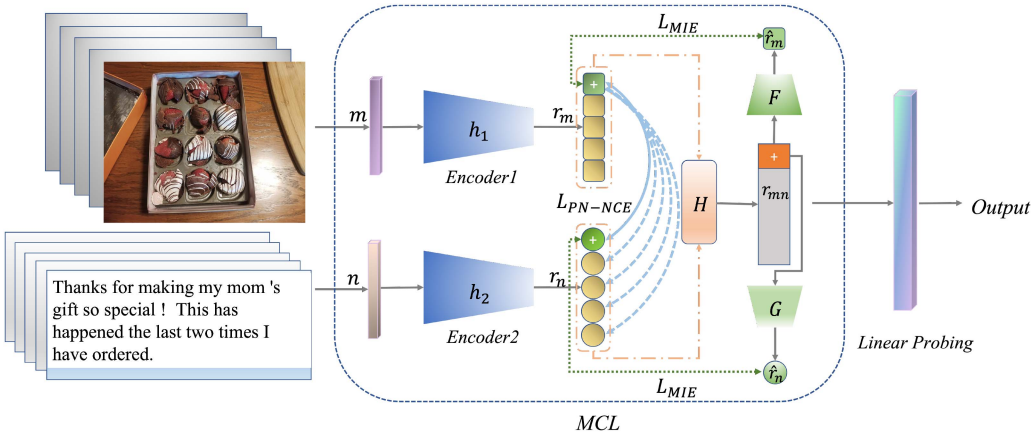

Fig. 2. MCL for multimodal fusion using *PN-NCE*. The part with the blue dashed square is the proposed MCL method. The extracted features m and n are from different input modalities (image and text). h_1 and h_2 are two encoders for different modalities. r_m are image embeddings and r_n are text embeddings. MCL is conducted between the positive and the negative pairs in the embedding space. The blue arrows show the process, i.e., the modalities from the positive pair are pulled closer (the solid arrow) while the modalities between the positive pair and the negative pairs are pushed away (the dashed arrows). H is the fusion module. r_{mn} is the fused embedding. F and G are transformation networks. Modality Invariance Estimator (*MIE*) is introduced between the positive embeddings for each modality, i.e., $(r_m^+, \hat{r}_m^+)$ and $(r_n^+, \hat{r}_n^+)$, which are denoted by '+'. At the linear probing stage, the MCL model is fixed and simple linear layers are trained for prediction.

and β_i are calculated by:

$$\alpha = \left[\frac{O^+}{s^+}\right]_+ \text{ and } \beta_i = \left[\frac{s_i^-}{O_i^-}\right]_+, \tag{3}$$

where $[\cdot]_+$ indicates the non-negative truncation operation. During training, when the similarity score is far from its optimum, the gradient regarding that pair will be multiplied by a larger pace estimation factor α and β_i in back-propagation, thus accelerating the learning pace. Conversely, when the similarity score is near its optimum, the learning pace will be eased. Such pace estimation factors are normalized between positive and negative pairs, which provides a relative adjustment. In practice, for each batch of samples, we have one positive pair and K negative pairs. If batchsize is set to B, α and β_i are the average values over the batchsize, i.e., $\alpha = (\sum_B \alpha_b)/B$ and $\beta_i = (\sum_B \beta_{ib})/B$. For each iteration, the pace estimation factors are recalculated to adjust the learning paces for both positive and negative pairs.

We note that reweighting similarity scores is a conventional practice in many loss functions for classification problems. Traditional methods [15], [16], [17], [21], [41], [42] often take equal scaling factors rather than adaptive values to reweight the similarity scores. This is because, in the objective function of a classification problem, the softmax value is regarded as the probability of an input belonging to each category. However, in MCL, equal scaling is not preferred as it restricts the learning pace adjustment for each positive and negative pair. The proposed method exploits adaptive reweighting factors for greater flexibility, therefore providing a more desirable optimization strategy.

C. Towards Robust MCL With Noisy Supervision

When a certain modality is noisy or corrupted, the supervision between modalities in a pair may not always be reliable. However, traditional InfoNCE loss [35] is not able to deal with noisy labels of training data. To overcome this issue, Chuang et al. proposed a robust InfoNCE [23] that achieves robustness against noisy views, which is defined as[1]

$$\mathcal{L}_R(\mathbf{s}) = -\mathbb{E}\left(e^{s^+} - \mu\sum_{i=1}^{K} e^{s_i^-}\right), \tag{4}$$

where μ is a hyperparameter balancing the score from the positive pair and the scores from negative pairs. They exploit the robustness classification theorem [43] derived by Ghosh et al. to prove that $\mathcal{L}_R$ is robust against label noise. We can extend (4) to be pace-adaptive i.e.,

$$\mathcal{L}_{PN-NCE}(\mathbf{s}) = -\mathbb{E}\left(e^{\alpha s^+} - \mu\sum_{i=1}^{K} e^{\beta_i s_i^-}\right),$$
$$s.t.\ \alpha + \sum_{i=1}^{K}\beta_i = 1, \tag{5}$$

however, it no longer enjoys the theoretical guarantee given by Chuang et al., because it cannot be decomposed to K classifiers. Each of them takes only one input pair to make the decision without influencing by other input pairs. More clearly, Chuang et al.'s theory requires that the loss function has the form

$$\mathcal{L}(\mathbf{s}) = l_+(s^+, 1) + \lambda\sum_{i=1}^{K} l_-(s^-, -1), \tag{6}$$

where l is a binary classification loss that satisfies the symmetric condition, i.e., $l(s,1) + l(s,-1) = const., \forall s \in \mathbb{R}$. Each of the individual loss functions only takes a score from a pair of inputs, l_+ for positive pairs and l_- for negative pairs. Due to the normalization $\alpha + \sum_{i=1}^{K}\beta_i = 1$, $\mathcal{L}_{PN-NCE}$ does not fit the form. As

[1]Note that Eq. 4 is the robust InfoNCE with parameter q equals to 1, achieving robustness against noisy views in the same manner as binary classification with noisy labels. (2 in Section 4 of Chuang et al. [23])

a result, we cannot have the robustness guarantee from Chuang et al.'s theory. However, it does not mean that $\mathcal{L}_{PN-NCE}$ is not robust against the noisy labels. To study the robustness of $\mathcal{L}_{PN-NCE}$, we first define a cooperative classifier ensemble.

Definition 1: Let $\mathbf{F}$ be a cooperative classifier ensemble i.e., $\mathbf{F} = [f_1, \ldots, f_u]$, where each f_i is a binary classifier taking a score list $\mathbf{s} = [s_1, s_2, \ldots, s_u]$ as input to classify s_i.

For $\mathcal{L}_{PN-NCE}$, let $l(f^+(\mathbf{s}, s^+), y) = ye^{(\alpha s^+)}$ and $l(f_i^-(\mathbf{s}, s_i^-), y) = ye^{(\beta_i s_i^-)}$ be the losses of the individual classifiers in a cooperative classifier ensemble $\mathbf{F} = [f^+, f_1^- \ldots, f_K^-]$, where each f classifies a score from a pair of inputs based on scores from other pairs. Note that α and β_i are functions of the score list, $\mathbf{s} = [s^+, s_1^-, \ldots, s_K^-]$ defined before. Using $l(f^+(\mathbf{s}, s^+), y)$ and $l(f_i^-(\mathbf{s}, s_i^-), y)$, $\mathcal{L}_{PN-NCE}$ can be rewritten as

$$\mathcal{L}(\mathbf{F}(s)) = -\mathbb{E}\left(l(f^+(\mathbf{s}, s^+), 1) + \mu \sum_{i=1}^{K} l(f_i^-(\mathbf{s}, s_i^-), -1)\right). \tag{7}$$

Note that $l(f^+(\mathbf{s}, s^+), y)$ and $l(f_i^-(\mathbf{s}, s_i^-), y)$ fulfil the symmetry property i.e., $l(f^+(\mathbf{s}, s^+), 1) + l(f^+(\mathbf{s}, s^+), -1) = 0$ and $l(f_i^-(\mathbf{s}, s_i^-), 1) + l(f_i^-(\mathbf{s}, s_i^-), -1) = 0$. Since f_i^- is identical for all i, in the order of $s_1^-, \ldots, s_K^-$ would not change the classification result. In the following discussion, f^- is used to denote f_i^- to simplify the notation. Theorem 1 indicates that $\mathcal{L}_{PN-NCE}$ is robust to noisy labels.

Theorem 1: Given a cooperative classifier ensemble, score list $\mathbf{s}$, and the loss l satisfying the symmetric property, and the optimal $\mathbf{F}^*$ obtained from clean data i.e., $\mathcal{L}(\mathbf{F}^*(s)) < \mathcal{L}(\mathbf{F}(s)), \forall \mathbf{F}$, if s^+ and $s_J^-, \ldots, s_K^-$'s labels have η probability being wrong and

$$\eta < \min\left\{0.5, \sup_{\mathbf{F}}\left(\frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right)\right)\right\}, \tag{8}$$

where $T_1(\mathbf{F})$ and $T_2(\mathbf{F})$ are defined as

$$T_1(\mathbf{F}) = \int_{\mathbf{s}} \mu \sum_{i=1}^{J-1} l(f^-(\mathbf{s}, s_i^-), y_i^-) dp(\mathbf{s}), \tag{9}$$

$$T_2(\mathbf{F}) = \int_{\mathbf{s}} l(f^+(\mathbf{s}, s^+), y^+) dp(\mathbf{s}) + \mu \sum_{i=J}^{K} l(f^-(\mathbf{s}, s_i^-), y_i^-) dp(\mathbf{s}), \tag{10}$$

we have

$$\mathcal{L}^\eta(\mathbf{F}^*(s)) < \mathcal{L}^\eta(\mathbf{F}(s)), \forall \mathbf{F}, \tag{11}$$

where $\mathcal{L}^\eta(\cdot)$ is the loss defined in (7) with s^+ and $s_J^-, \ldots, s_K^-$'s labels having η probability being wrong.

In other words, if the noisy probability is smaller than

$$\min\left\{0.5, \sup_{\mathbf{F}}\left(\frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right)\right)\right\}, \tag{12}$$

we can still obtain the same $\mathbf{F}^*$ even when some labels are corrupted. Note that $y_i^- = -1$ and $y^+ = 1$.

We attached the proof of Theorem 1 in the Appendix. The three possible cases analyzed in the proof demonstrate the robustness of $\mathbf{F}$ given the noisy probability condition in (8). Discussion on the term $\sup_{\mathbf{F}}\left(\frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right)\right)$ is also given in the Appendix. Therefore, the noise-resistant property is achieved for MCL. We use $\mathcal{L}_{PN-NCE}$ to denote the proposed objective that combines pace-adaptive factors in the cooperative classifier ensemble.

D. Modality Invariance Estimator

As our goal is to produce better-fused representations for downstream tasks, we explicitly measure the modality invariance between the fused representation of the positive pairs with its unimodal representations in the embedding space. This relies on transformation networks that take the fused embedding as input. The outputs from the networks are then measured with the unimodal embeddings by *L1* distance.

Formally, let $r_m \in \mathcal{R}^{d_m}, r_n \in \mathcal{R}^{d_n}$ represent the unimodal embeddings of modality M and N, where $r_m = h_1(m)$, $r_n = h_2(n)$, and d_m and d_n are their dimensions. r_m^+ and r_n^+ denote the positive unimodal embeddings. Let $r_{mn} = H(r_m, r_n; \theta_H)$ be the fused embedding, $r_{mn} \in \mathcal{R}^{d_h}$, where H is the fusion module, d_h is the fused dimension, and θ_H is the trainable parameters. r_{mn}^+ denotes the positive fused embedding. The outputs of the transformation networks with respect to M and N are:

$$\begin{cases} \hat{r_m^+} = F(r_{mn}^+; \theta_F) \\ \hat{r_n^+} = G(r_{mn}^+; \theta_G), \end{cases} \tag{13}$$

where F and G are the two transformation networks with trainable parameters θ_F and θ_G. The modality invariance estimator (MIE) between r^+ and $\hat{r^+}$ is defined as:

$$\mathcal{L}_{MIE}(m, n) = \mathop{\mathbb{E}}_{(m,n)\sim P_{MN}} log\left((D_1(r_m^+, \hat{r_m^+}) + 1)(D_1(r_n^+, \hat{r_n^+}) + 1)\right) \tag{14}$$

where $D_1(r_m^+, \hat{r_m^+}) = (\sum_{d_m} |r_m^+ - \hat{r_m^+}|)/d_m$ and $D_1(r_n^+, \hat{r_n^+}) = (\sum_{d_n} |r_n^+ - \hat{r_n^+}|)/d_n$. For each additional modality, an additional $(D_1(\cdot) + 1)$ term must be appended in the $log(\cdot)$ term.

We expect that the fused embedding can maintain the modality-invariant information to a large extent. In other words, it can be reverted back to the corresponding unimodal embeddings as much as possible. During training, minimizing $\mathcal{L}_{MIE}(m, n)$ is integrated to MCL as an auxiliary task, which is subsequently abandoned in the linear probing stage.

We treat each modality equally important. To build a complete contrast, the proposed MCL objective is applied on set $\Omega = \{(m, n), (m, \tilde{n}_1), (m, \tilde{n}_2), \ldots, (m, \tilde{n}_K)\}$ and set $\Omega' = \{(n, m), (n, \tilde{m}_1), (n, \tilde{m}_2), \ldots, (n, \tilde{m}_K)\}$. The loss for each set is combined during MCL stage.

E. Extension to Three Modalities

The proposed method can be generalized to more modalities. Here, we take three-modality scenario as an example. We denote the modalities as M, N, and Q. The MCL among M, N, Q is to select the positive tuple (m, n, q) from the set $\Omega_{m,n,q} = \{(m, n, q), \{(m, \tilde{n}^*, \tilde{q}^*)\}_{i=1}^{K}\}$ that contains $(K + 1)$ tuples with noisy supervision. This is achieved by splitting $\Omega_{m,n,q}$ into subsets (i.e., $\Omega_{m,n}$, $\Omega_{n,q}$, and $\Omega_{q,m}$) and performing MCL on the subsets concurrently.

Hence, the proposed method for multimodal feature fusion is formulated as optimizing the following objective function:

$$\underset{\Omega_{m,n,q}}{\mathcal{L}} = \underset{\Omega_{m,n}}{\mathcal{L}_{PN-NCE}} + \underset{\Omega_{n,q}}{\mathcal{L}_{PN-NCE}} + \underset{\Omega_{q,m}}{\mathcal{L}_{PN-NCE}} + \mathcal{L}_{MIE}(m, n, q), \tag{15}$$

where $\mathcal{L}_{MIE}(m, n, q)$ is the modality invariance estimator for three modaltities (see equation 14).

IV. Experiment

The proposed method is evaluated on three multimodal datasets with visual, audio and language modalities for joint multimodal understanding and inference. The multimodal fusion tasks are: *Video-and-Language Inference* on Violin [18], *Multimodal Sentiment Classification* on CMU-MOSEI [31], and *Multimodal Sarcasm Detection* on Sarcasm dataset [19].

A. Dataset

Violin [18] is a large scale dataset for Video-and-Language Inference.[2] It contains 95,322 video-hypothesis pairs from 15,887 video clips. Each video clip is paired with a subtitle and three pairs of positive/negative natural language hypotheses. The task is to infer whether the hypothesis is entailed or contradicted by the given video clip.

CMU-MOSEI [31] is a dataset for multimodal sentiment analysis and emotion recognition.[3] The dataset is made up of 23,454 movie review video clips. The video clips cover a variety of topics. Each video clip is annotated with the sentiment on a [−3, 3] Likert scale of [−3: highly negative, −2 negative, −1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 highly positive]. The metric Acc-7 represents the 7-class classification accuracy. The metric Acc-2 evaluates the positive and negative sentiment accuracy by separating the labels into [−3, −2, −1] and [0, 1, 2, 3] groups.

Sarcasm Detection [19] is a public dataset for multimodal sarcasm detection.[4] It is collected from English tweets containing pictures, attributes descriptions, and texts. The training, validation, and test sets have 19,816, 2,410, and 2,409 sentences, respectively. The task is to determine whether the input is sarcastic by combining the information from images and texts.

[2][Online]. Available: https://github.com/jimmy646/violin

[3][Online]. Available: https://github.com/A2Zadeh/CMU-MultimodalSDK for more information.

[4][Online]. Available: https://github.com/wrk226/pytorch-multimodal_sarcasm_detection

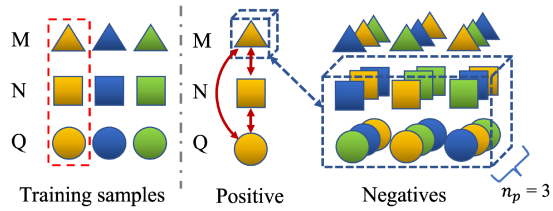

Fig. 3. Negative feature augmentation for MCL. On the left-hand side, training samples are from three modalities M, N, and Q. In the middle, the triangle, square and circle with the same color is a positive sample from the three modalities. On the right-hand side, negative samples are generated using training samples by random permutation within one modality and then grouped across modalities vertically. The number of permutations is set as a hyper-parameter n_p (e.g., $n_p = 3$). The contrast between positive and negative samples is illustrated by the blue dashed arrow. The contrast within positive samples is illustrated by the red solid arrows.

B. Implementation Details

1) Feature Generation: The unimodal features are generated by the protocol from each multimodal dataset.

For *Video-and-Language Inference* task on Violin dataset, the global visual features for each frame are extracted using ResNet-101 [44] pre-trained on ImageNet [45], and are down-sampled to 3 frames per second, resulting in a 2,048-dimensional feature for each frame. Text features are encoded by finetuning a pre-trained BERT-based model [46] on the statements and subtitles on the Violin, which yields a feature dimension of 768.

For *Multimodal Sentiment Classification* task in CMU-MOSEI, the inputs include visual, vocal, and verbal features. The visual features contain 35 facial action units, 68 facial landmarks, and general face features [47]. The audio features are extracted using COVAREP [48], and the text features are extracted by Glove [49].

For *Multimodal Sarcasm Detection* task, the image and the attribute features are extracted using ResNet-50 [44] pre-trained on ImageNet [45]. The text features are extracted using Glove [49].

2) Negative Feature Augmentation for MCL: Different from traditional data augmentation for contrastive learning, our data augmentation for MCL is illustrated in Fig. 3. The training samples cover three modalities, M, N, and Q, which are represented by triangles, squares, and circles, respectively. A set of a triangle, a square, and a circle arranged vertically constitutes a positive sample or a negative sample (e.g., a positive pair is illustrated in a red-dashed box). The same color set constitutes a positive sample. The negative samples are generated by first randomly permutating within each modality and then grouped across modalities. The modalities from a positive sample are pulled closer (illustrated by the red solid arrows). The contrast between positive and negative samples happens between one modality (e.g., M /N /Q) from the positive sample and other modalities (e.g., (N, Q)/(M, Q)/(M, N)) from negatives (illustrated by the blue dashed arrow and the blue dashed cubes). In our experiments, the number of permutations was set as a hyper-parameter n_p (e.g., $n_p = 3$). This intuitively introduced more diversified negatives, which were beneficial to MCL performance.

TABLE II
ENCODER SETUP FOR MCL

Modality	Violin	CMU-MOSEI	Sarcasm
image/video/attribute	RNN	Transformer	ResNet
audio	-	Transformer	-
subtitle/text	RNN	Transformer	bi-LSTM

3) Encoder Setup: The encoders for different modalities are summarized in Table II. The encoder setup for the Violin dataset was the same as [18]. For sarcasm detection, encoders were chosen from [19]. On CMU-MOSEI, we used the encoders from [29]. The encoders trained from scratch via supervised learning were set as the supervised baseline. For a fair comparison, all the MCL models were also trained using the same encoder setup in Table II.

4) Fusion Module and Transformation Networks: Our proposed method does not rely on a delicate fusion network. We adopted the widely used transformer fusion module in our implementation. To be specific, the Transformer encoders in [29] were adopted as the fusion module and the transformation networks. However, the numbers of heads and layers were set differently for each dataset. The fusion module had 2 heads and 4 layers for the Violin dataset, and 2 heads and 3 layers for both CMU-MOSEI and Sarcasm datasets. The transformation networks had 4 heads and 3 layers for all the tasks.

5) Linear Probing: The second stage of the proposed method for multimodal fusion was linear probing. The inputs were original multimodal features without augmentations. We kept the trainable parameters from the MCL stage unchanged and trained simple linear layers in a supervised manner for downstream tasks.

6) Compared Approaches: To demonstrate the effectiveness of the proposed PN-NCE objective, we compared it with the supervised trained-from-scratch baseline as well as several self-supervised MCL methods including CPC (InfoNCE) [15], CMC [16], MMV-FAC [17], TupleInfoNCE [21], and RINCE [23]. These methods proposed different learning objectives based on InfoNCE for self-supervised learning, which are competitive SOTA methods.

C. Results

The proposed method was examined in three multimodal fusion tasks. As presented in Table III, we compared the performances with the supervised method and several state-of-the-art self-supervised methods. For fair comparison with baseline methods, we followed the same encoder setup in Table II. Our method consistently outperformed previous self-supervised MCL baselines. Furthermore, it was comparable to the supervised baseline, even surpassing it in some metrics. This showcased the versatility of the proposed method for different tasks. The reported results in Table III were based on $n_p = 40$, i.e., we generated our negative samples by 40 times random permutation on all the modalities. Specifically, for *Video-and-Language Inference* task on Violin, the trained MCL model was able to explore and relate underlying multimodal information. Our method reaches 67.25% of overall accuracy, which was close to the performance of the supervised baseline of 67.60%. On the *multimodal sentiment classification* task, the diversity of input modalities, including image, text, and audio signals, made fusion a complex challenge. However, the proposed method also outperformed all the baselines and surpassed the supervised baseline in terms of two-class accuracy. On the *sarcasm detection* task, the proposed method was effective at exploring and fusing multimodal features to discriminate the semantic differences.

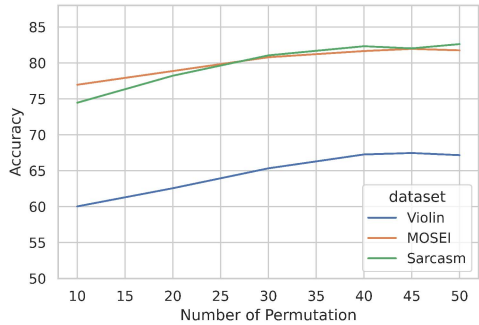

Fig. 4. Accuracy (%) with different number of permutations (n_p). Performances in different multimodal fusion tasks increase constantly when n_p becomes larger. After $n_p = 30$, performance improvement slows down gradually. We limit n_p up to 40 to address the scarcity of negative samples without spending excessive training time.

1) Ablation Study: We show the ablation study results in Table IV. PN-NCE achieved the best results among its variants. Specifically, PN-NCE reached higher performance compared to P-NCE, which indicated that the noisy resistant property of PN-NCE was beneficial for improving the performance. By comparing PN-NCE to PN-NCE w/o MIE (and P-NCE to P-NCE w/o MIE), we show that MIE was able to effectively improve the performance for both P-NCE and PN-NCE objectives. More intuitively, please refer to the improvements of PN-NCE over PN-NCE w/o MIE in parentheses.

2) Effect of Size and Diversity of Training Data: It is worth investigating the effect of size and diversity of training data on model performance. Fig. 4 shows the accuracy curves regarding the number of permutations n_p. We randomly permuted the unimodal features n_p times and combined them by permuted indexes, which diversified the training samples and increased the number of negative samples. With the increasing number of permutations, performance in the three tasks steadily improved. After $n_p = 30$, a more gradual increase in accuracy was observed. We stopped permutation at $n_p = 40$, as the performances were approaching or exceeding the supervised baselines. When $n_p = 45$ or higher, the performances on different datasets did not show significant and meaningful improvement. We suggest that a proper value of n_p can effectively address the scarcity of negative samples for MCL without excessively increasing training time.

D. Pace Adjustment in MCL

To deal with the imbalanced learning paces of positive and negative samples in MCL, we introduced hyper-parameters α

TABLE III
PERFORMANCE COMPARISONS WITH SUPERVISED AND SELF-SUPERVISED BASELINES ON VIOLIN, CMU-MOSEI AND SARCASM

Methods	Violin	CMU-MOSEI			Sarcasm	
	Acc (%)	Acc-2 (%)	F1 (%)	Acc-7 (%)	Acc (%)	F1 (%)
Supervised Baseline	*67.60*	*81.50*	*82.80*	*51.80*	*80.18*	*83.44*
CPC [15]	61.25	79.96	80.21	49.96	78.87	76.93
CMC [16]	59.08	79.25	77.60	50.82	75.82	73.89
MMV-FAC [17]	59.96	78.45	79.33	51.35	78.95	78.26
RINCE (q=1) [23]	65.21	79.94	80.66	50.87	79.88	81.02
TupleInfoNCE [21]	65.85	80.01	81.44	51.50	80.57	81.74
PN-NCE	**67.25**	**81.64**	**82.05**	**51.72**	**81.56**	**82.31**

TABLE IV
ABLATION STUDY ON VIOLIN, CMU-MOSEI, AND SARCASM

Methods	Violin	CMU-MOSEI			Sarcasm	
	Acc (%)	Acc-2 (%)	F1 (%)	Acc-7 (%)	Acc (%)	F1 (%)
P-NCE w/o MIE	64.81	79.56	80.29	49.80	79.15	80.77
P-NCE	66.04	80.66	81.20	50.69	80.08	81.62
PN-NCE w/o MIE	65.76	80.75	81.02	50.48	80.69	81.58
PN-NCE	**67.25** (↑ 1.49)	**81.64** (↑ 1.01)	**82.05** (↑ 1.03)	**51.72** (↑ 1.24)	**81.56** (↑ 0.87)	**82.31** (↑ 0.73)

↑ in the parentheses indicates the performance improvement over PN-NCE without MIE.

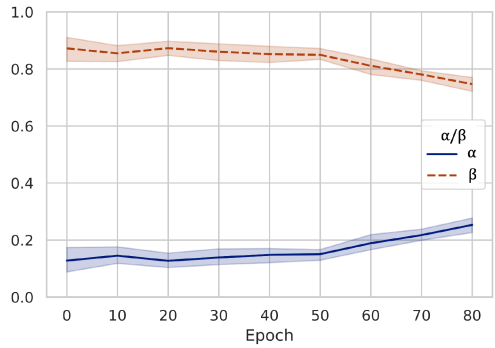

Fig. 5. Value changes of pace estimation factor (α for positive) and (β for negatives) on Violin dataset.

and β (the sum for all negative samples) in the PN-NCE objective to adjust the learning paces dynamically. In our experiment, we set the optimal scores $O^+ = 1$ and $O^- = 0.01$. To illustrate how α and β change during training, we present the value curves in Fig. 5, and thereby intuitively show the effectiveness of PN-NCE objective at providing a more balanced learning pace between positive and negative samples. We collected the average α and β values in each epoch (from epoch 1 to 80) with multiple different initializations on the Violin dataset. The curves were the mean estimates by aggregating multiple α and β values. The shading of the curves represented the 95% confidence interval of the estimates. As presented in Fig. 5, with the encoder setup and $n_p = 40$ on the Violin dataset, positive samples showed a lower value of pace estimation factor α, which adjusted the positive samples to slow down their learning paces. On the contrary, the higher pace estimation factor β adjusted the negative samples to accelerate their learning paces. We dynamically changed the learning paces of positive and negative samples, which mitigated the imbalanced learning issue illustrated in Table I.

During training, α and β were normalized to 1. We observed that β remained at a high value to accelerate the learning pace for negative samples at the first half of the training epochs while decreasing gradually to ease the learning as the negative samples were approaching their optimums. For positive samples, α kept lower than β. However, when negative samples learned relatively faster, α increased to adjust the learning pace for positive samples. PN-NCE achieved satisfying performances across all the multimodal fusion tasks because it can learn the rich underlying multimodal dynamics by balancing the learning paces of positive and negative samples at the MCL stage.

E. Robustness Comparison With Feature Perturbation

We evaluated the robustness of our method against different levels of perturbation in the training data. We defined perturbation level (*P-Level*) as the percentage of input training features added with a Gaussian noise, which was sampled from a normal distribution with zero mean and unit variance. We randomly selected one modality in a pair to partially introduce noisy supervision in positive pairs to perform perturbation. This ensured that the semantic information shared in positive pairs was not totally lost. This was also to create a harder scenario for MCL. We compared our method with CPC (InfoNCE) [15] as it was the most popular approach for MCL. We also compared it with the competitive method for multimodal fusion, TupleInfoNCE [21]. We took the same generated features and encoder setups as described in section 4.3 but added perturbations to the features. We also used the same linear probing layers for all the downstream tasks.

At all *P-Levels*, our method outperformed the compared methods on all the metrics. As shown in Table V, with the increasing *P-Level*, from 0% to 50%, the performances of both methods dropped. With a mild perturbation, we noticed that the accuracies/F1 decreased steadily. Overall, PN-NCE degraded less with

TABLE V
ROBUSTNESS COMPARISON WITH DIFFERENT LEVELS OF FEATURE PERTURBATION

Methods	0 %	10 %	20 %	30 %	40 %	50 %
CPC (InfoNCE) [15]	61.25 (6.0)	60.52 (4.50)	57.48 (5.81)	55.99 (**5.33**)	52.23 (3.78)	44.38 (**5.27**)
TupleInfoNCE [21]	65.85 (1.4)	61.06 (3.96)	58.25 (5.04)	55.60 (**5.72**)	50.33 (**5.68**)	44.02 (**5.63**)
PN-NCE	67.25	65.02	63.29	61.32	56.01	49.65

(a) Violin Dataset. The results are Acc (%)

Methods	0 %	10 %	20 %	30 %	40 %	50 %
CPC (InfoNCE) [15]	79.96 (1.68) 80.21 (1.84) 49.96 (1.76)	75.19 (2.85) 77.46 (0.49) 46.84 (0.77)	72.06 (1.93) 71.96 (2.65) 43.06 (0.82)	68.24 (1.41) 67.98 (0.67) 40.15 (1.91)	55.25 (**5.63**) 56.01 (**3.84**) 36.92 (**3.29**)	50.88 (2.08) 48.95 (2.11) 32.05 (**3.37**)
TupleInfoNCE [21]	80.01 (1.63) 81.44 (0.61) 51.50 (0.22)	75.36 (2.68) 75.22 (2.73) 45.25 (2.36)	69.90 (4.09) 70.66 (3.95) 41.54 (2.34)	65.28 (**4.37**) 64.07 (**4.58**) 39.76 (2.30)	53.44 (**7.44**) 54.21 (**5.64**) 33.88 (**6.33**)	47.06 (**5.90**) 45.32 (**5.74**) 28.48 (**6.76**)
PN-NCE	81.64 82.05 51.72	78.04 77.95 47.61	73.99 74.61 43.88	69.65 68.65 42.06	60.88 59.85 40.21	52.96 51.06 35.42

(b) CMU-MOSEI Dataset. The results are Acc-2 (%), F1 (%), and Acc-7 (%) in order from top to bottom.

Methods	0 %	10 %	20 %	30 %	40 %	50 %
CPC (InfoNCE) [15]	78.87 (2.78) 76.93 (5.38)	74.65 (1.57) 73.42 (2.03)	70.21 (2.17) 69.28 (2.33)	65.38 (1.98) 64.31 (1.53)	60.72 (2.23) 58.82 (2.20)	52.48 (**3.18**) 51.06 (3.01)
TupleInfoNCE [21]	80.57 (0.99) 81.74 (0.57)	75.06 (1.16) 73.68 (1.77)	68.45 (3.93) 69.06 (2.55)	63.64 (**3.72**) 62.75 (3.09)	58.66 (**3.93**) 56.36 (**4.66**)	49.95 (**5.71**) 50.39 (**3.68**)
PN-NCE	81.56 82.31	76.22 75.45	72.38 71.61	67.36 65.84	62.59 61.02	55.66 54.07

(c) Sarcasm Dataset. The results are Acc (%) and F1 (%).

The levels indicate the percentage of training samples with gaussian noise. The results in parentheses show the performance differences of PN-NCE over the compared methods.

increasing *P-Level*. We also found that the performances of InfoNCE suffered a significant decrease from *P-Level* 30% to 40%. For example, on CMU-MOSEI dataset, the Acc-2 drops 12.99%, and the F1 drops 11.97%. Under large perturbation, the compared methods suffered much degradation. PN-NCE was more resilient at large *P-Level*. The results in parentheses show the performance differences of PN-NCE over the compared methods. We focused on comparing the performance when positive pairs introduce a large number of noisy modalities, *i.e. P-Level* 30% to 50% (see the bolded results in Table V). Under such circumstances, the supervision signals in positive pairs were undesirably affected, which caused hard positive samples. However, it can be observed that PN-NCE was more robust against noise in training data and consistently outperforms the compared baseline.

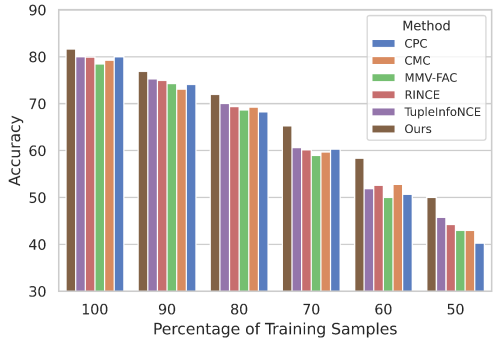

Fig. 6. Accuracy (%) with proportions of training samples (%) on CMU-MOSEI at the linear probing stage. As the training data is greatly reduced, e.g., from 70% to 60% (and 60% to 50%), the performance of the baseline models degrades to a greater extent. Compared with other MCL baselines, our method is the least affected.

F. Linear Probing With Proportionally Reduced Training Samples

We observed that the number of training data in MCL affected the performance significantly. However, few studies were found on performance degradation when using less training data in the linear probing stage. We evaluated the effect of reduced training samples on the model performance at the linear probing stage.

As shown in Fig. 6, with the same encoder and dataset setup for all the compared methods, the two-class accuracies on CMU-MOSEI dataset are reported. The number of training samples varied from 100% to 50% of the original sample size. While the trend of decreasing accuracy was expected with the reduction of training samples, our proposed method was the least affected. This effect was the most pronounced when there was a large reduction in training samples. Under the circumstance of losing a large proportion of training samples (70% to 50%), our method bore the least degradation compared with the other MCL baseline methods. As we only applied simple linear layers,

the performance improvement was attributed to the MCL stage, where we proposed PN-NCE objective function.

The results revealed that 1) linear probing with fewer training samples undermines the fusion performance to a large extent regardless of MCL models and 2) an effective MCL model can mitigate the performance loss on multimodal fusion tasks. The results provided strong evidence that validates the effectiveness of the proposed PN-NCE for multimodal fusion.

V. Conclusion

In this work, we proposed Pace-adaptive and Noise-resistant contrastive learning for multimodal feature fusion. To alleviate the issue of the imbalanced learning pace of positive and negative pairs, the PN-NCE objective function was proposed for efficient and robust self-supervised multimodal contrastive learning. Furthermore, modality-invariant information was maintained and maximized by explicit distance measure between the fused representation and the unimodal representations. The proposed method for multimodal fusion directly took extracted unimodal features as input and was model-agnostic. The experimental results on three multimodal fusion tasks showed the efficacy of our proposed method. In the future, we consider extending our method to other multimodal representation learning tasks, such as multimodal generation.

Appendix A

A. Proof of Theorem 1

The loss function of the cooperative classifier ensemble with noisy labels is defined as

$$\begin{aligned}\mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = &\int_{\mathbf{s}} l(f^{+}(\mathbf{s}, s^{+}), \hat{y}^{+}) dp(\mathbf{s}) \\ &+ \mu \int_{\mathbf{s}} \left(\sum_{i=1}^{J-1} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) \right. \\ &\left. + \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), \hat{y}_i^{-}) \right) dp(\mathbf{s}),\end{aligned}$$

where y_i^- represents clean label and $\hat{y}^+$ and $\hat{y}_i^-$ represent labels with η probability being wrong. Rewriting it,

$$\begin{aligned}\mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = &\int_{\mathbf{s}} \mu \sum_{i=1}^{J-1} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) dp(\mathbf{s}) \\ &+ \int_{\mathbf{s}} (1-\eta) l(f^{+}(\mathbf{s}, s^{+}), y^{+}) \\ &+ \eta l(f^{+}(\mathbf{s}, s^{+}), -y^{+}) dp(\mathbf{s}) \\ &+ \mu \int_{\mathbf{s}} (1-\eta) \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) \\ &+ \eta \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), -y_i^{-}) dp(\mathbf{s}).\end{aligned}$$

To simplify the notation, let

$$T_1(\mathbf{F}) = \int_{\mathbf{s}} \mu \sum_{i=1}^{J-1} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) dp(\mathbf{s}).$$

Using the symmetric condition

$$\begin{cases} l(f^{+}(\mathbf{s}, s_i), 1) + l(f^{+}(\mathbf{s}, s_i), -1) = C \\ l(f^{-}(\mathbf{s}, s_i), 1) + l(f^{-}(\mathbf{s}, s_i), -1) = C, \end{cases}$$

$$\begin{aligned}\mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = &T_1(\mathbf{F}) + \int_{\mathbf{s}} (1-2\eta) l(f^{+}(\mathbf{s}, s^{+}), y^{+}) \\ &+ \eta C dp(\mathbf{s}) + \mu \int_{\mathbf{s}} (1-2\eta) \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) \\ &+ \eta (K-J+1) C dp(\mathbf{s}),\end{aligned}$$

$$\begin{aligned}\mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = &T_1(\mathbf{F}) + (1-2\eta) \int_{\mathbf{s}} l(f^{+}(\mathbf{s}, s^{+}), y^{+}) dp(\mathbf{s}) \\ &+ (1-2\eta)\mu \int_{\mathbf{s}} \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) dp(\mathbf{s}) + Const.\end{aligned}$$

Considering

$$\begin{aligned}&\mathcal{L}^{\eta}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = T_1(\mathbf{F}^{*}) - T_1(\mathbf{F}) \\ &+ (1-2\eta) \int_{\mathbf{s}} l(f^{+*}(\mathbf{s}, s^{+}), y^{+}) - l(f^{+}(\mathbf{s}, s^{+}), y^{+}) dp(\mathbf{s}) \\ &+ (1-2\eta)\mu \int_{\mathbf{s}} \sum_{i=J}^{K} l(f^{-*}(\mathbf{s}, s_i^{-}), y_i^{-}) - l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) dp(\mathbf{s}),\end{aligned}$$

where $\mathbf{F}^*$ is optimal under clean data.

Let

$$\begin{aligned}T_2(\mathbf{F}) = &\int_{\mathbf{s}} l(f^{+}(\mathbf{s}, s^{+}), y^{+}) dp(\mathbf{s}) \\ &+ \mu \int_{\mathbf{s}} \sum_{i=J}^{K} l(f^{-}(\mathbf{s}, s_i^{-}), y_i^{-}) dp(\mathbf{s}),\end{aligned}$$

then we have

$$\begin{aligned}\mathcal{L}^{\eta}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) = &T_1(\mathbf{F}^{*}) - T_1(\mathbf{F}) \\ &+ (1-2\eta)(T_2(\mathbf{F}^{*}) - T_2(\mathbf{F})).\end{aligned}$$

Case 1: $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) < 0$ and $T_2(\mathbf{F}^*) - T_2(\mathbf{F}) < 0$

If $\eta < 0.5$, $\mathcal{L}^{\eta}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) < 0$. It implies the optimal $\mathbf{F}^*$ can be obtained from noisy training.

Case 2: $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) < 0$ and $T_2(\mathbf{F}^*) - T_2(\mathbf{F}) > 0$

Note $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) + T_2(\mathbf{F}^*) - T_2(\mathbf{F}) = \mathcal{L}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}(\mathbf{F}(\mathbf{s}))$ and $\mathbf{F}^*$ is optimal under clean data. In other words, $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) + T_2(\mathbf{F}^*) - T_2(\mathbf{F}) < 0$. As a result, any η between 0 and 1, $\mathcal{L}^{\eta}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}^{\eta}(\mathbf{F}(\mathbf{s})) < 0$. It also implies robustness.

Case 3: $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) > 0$ and $T_2(\mathbf{F}^*) - T_2(\mathbf{F}) < 0$

Since

$$\begin{aligned}&\mathcal{L}(\mathbf{F}^{*}(\mathbf{s})) - \mathcal{L}(\mathbf{F}(\mathbf{s})) < 0, \\ &T_1(\mathbf{F}^{*}) - T_1(\mathbf{F}) + T_2(\mathbf{F}^{*}) - T_2(\mathbf{F}) < 0,\end{aligned}$$

$$T_1(\mathbf{F}^*) - T_1(\mathbf{F}) < -(T_2(\mathbf{F}^*) - T_2(\mathbf{F})).$$

Combining $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) > 0$, we have

$$0 < T_1(\mathbf{F}^*) - T_1(\mathbf{F}) < -(T_2(\mathbf{F}^*) - T_2(\mathbf{F})). \ (*)$$

Considering

$$T_1(\mathbf{F}^*) - T_1(\mathbf{F}) + (1 - 2\eta_0)(T_2(\mathbf{F}^*) - T_2(\mathbf{F})) = 0,$$

$$\eta_0 = \frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right).$$

Using $(*)$, we know that

$$0 > \frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} > -1.$$

Thus, $0 < \eta_0 < 0.5$.

If

$$\eta < \sup_{\mathbf{F}}\left(\frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right)\right),$$

$\mathcal{L}^\eta(\mathbf{F}^*(\mathbf{s})) - \mathcal{L}^\eta(\mathbf{F}(\mathbf{s})) < 0$. It implies the optimal $\mathbf{F}^*$ can be obtained from noisy training.

Note that Case 4: $T_1(F^*) - T_1(F) > 0$ and $T_2(F^*) - T_2(F) > 0$ is impossible because $\mathcal{F}^*$ is optimal under clean data. Combining Cases 1–3, we complete the proof.

B. Discussion

We can observe from the proof of Theorem 1 that Case 2 requires the most relaxed condition for the noise probability η (i.e., $0 \leq \eta \leq 1$) compared with Cases 1 and 3.

According to Case 3 in the proof, if there exists an $\mathbf{F}$ fulfilling $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) > 0$ and $T_2(\mathbf{F}^*) - T_2(\mathbf{F}) < 0$,

$$\eta = \sup_{\mathbf{F}}\left(\frac{1}{2}\left(\frac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})} + 1\right)\right).$$

As $\dfrac{T_1(\mathbf{F}^*) - T_1(\mathbf{F})}{T_2(\mathbf{F}^*) - T_2(\mathbf{F})}$ is between -1 and 0, η is between 0 and 0.5 in this case. Note that $\mathbf{F}^*$ is optimal, meaning that $T_1(\mathbf{F}^*) + T_2(\mathbf{F}^*) < T_1(\mathbf{F}) + T_2(\mathbf{F}), \forall \mathbf{F}$.

Since all the l and f^- are identical and all s_i^- follows i.i.d, as long as μ is large enough, $l(f^-(\mathbf{s}, s_i^-), y_i^-)$ would dominate the training and $T_1(\mathbf{F}^*) - T_1(\mathbf{F}) < 0$. In other words, as long as μ is large enough, the optimal $\mathbf{F}^*$ would be obtained for noise probability $\eta < 0.5$. If all the labels of $s^+, s_1^-, \ldots, s_K^-$ have η probability of being corrupted, the theorem above is still valid because there is no $T_1(\mathbf{F}^*)$ and $T_1(\mathbf{F})$, and we will obtain the optimal $\mathbf{F}^*$ as long as $\eta < 0.5$.

To sum up, we defined a cooperative classifier ensemble and showed the robust property of $\mathcal{L}_{PN-NCE}$ in three possible cases.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017.

[3] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[6] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 18661–18673.

[7] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.

[8] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15750–15758.

[9] J. Giorgi, O. Nitski, B. Wang, and G. Bader, "DECLUTR: Deep contrastive learning for unsupervised textual representations," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 879–895.

[10] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6894–6910.

[11] H. Bao et al., "Unilmv2: Pseudo-masked language models for unified language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 642–652.

[12] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020,vol. 33, pp. 8765–8775.

[13] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3875–3879.

[14] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. InterSpeech*, 2019, pp. 3465–3469.

[15] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[16] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.

[17] J.-B. Alayrac et al., "Self-supervised multimodal versatile networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020,vol. 33, pp. 25–37.

[18] J. Liu et al., "Violin: A large-scale dataset for video-and-language inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10900–10910.

[19] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.

[20] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[21] Y. Liu et al., "Contrastive multimodal fusion with tupleinfonce," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 754–763.

[22] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1483–1492.

[23] C.-Y. Chuang et al., "Robust contrastive learning against noisy views," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16670–16681.

[24] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–17.

[25] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, pp. 5634–5641.

[26] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[27] Y. Wang et al., "Deep multimodal fusion by channel exchanging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020,vol. 33, pp. 4835–4845.

[28] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[29] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, pp. 6558–6569.

[30] R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. W. Chen, "Improving contrastive learning by visualizing feature transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10306–10315.
[31] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[32] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.
[33] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 1735–1742.
[34] N. Saunshi, O. Plevrakis, S. Arora, M. Khodak, and H. Khandeparkar, "A theoretical analysis of contrastive unsupervised representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5628–5637.
[35] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist., JMLR Workshop Conf. Proc.*, 2010, pp. 297–304.
[36] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion*, 2020, pp. 1–6.
[37] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[38] A. Nagrani et al., "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 14200–14213.
[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[40] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017,vol. 30, pp. 1–11.
[41] X. Zhang, F. X. Yu, S. Karaman, W. Zhang, and S.-F. Chang, "Heated-up softmax embedding," 2018, *arXiv:1809.04157*.
[42] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.
[43] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, pp. 1919–1925.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[45] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[46] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[47] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[48] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep–A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
[49] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical methods Natural Lang. Process.*, 2014, pp. 1532–1543.

Xiaobao Guo is currently working toward the Ph.D. degree with the School of Computer Science and Engineering, and with Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore. Her research interests include computer vision and multimodal learning.

Alex Kot (Life Fellow, IEEE) has been with the Nanyang Technological University, Singapore, since 1991. He was the Head of the Division of Information Engineering and Vice Dean Research with the School of Electrical and Electronic Engineering. Subsequently, he was the Associate Dean for College of Engineering for eight years. He is currently a Professor and the Director of Rapid-Rich Object Search (ROSE) Lab and NTU-PKU Joint Research Institute. He has authored or coauthored extensively in the areas of signal processing, biometrics, image forensics and security, and computer vision and machine learning. Dr. Kot was an Associate Editor for more than ten journals, mostly for IEEE transactions. He was the IEEE SP Society in various capacities such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice-President for the IEEE Signal Processing Society. He was the recipient of the Best Teacher of the Year Award and is a co-author for several Best Paper Awards including ICPR, IEEE WIFS and IWDW, CVPR Precognition Workshop and VCIP. He was elected as the IEEE Distinguished Lecturer for the Signal Processing Society and the Circuits and Systems Society. He is a Fellow of Academy of Engineering, Singapore.

Adams Wai-Kin Kong (Member, IEEE) received the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada. He is currently an Associate Professor with Nanyang Technological University, Singapore. His papers have been published in TPAMI, TIP, TIFS, TMM, TPWRS, NeurIPS, ICRL, CVPR, ICCV, ECCV, EMLP, IJCA, and pattern recognition. His research interests include pattern recognition, deep learning, and their applications in power systems, healthcare, and biometrics. One of his papers was selected as a spotlight paper by TPAMI and another one was selected as Honorable Mention by Pattern Recognition. With his students, he was the recipient of the Best Student Paper Awards at The IEEE Fifth International Conference on Biometrics: Theory, Applications, and Systems, 2012, and IEEE International Conference on Bioinformatics and Bioengineering, 2013. In 2016, he was also the recipient of the Best Reviewer Award from BTAS. He also was an Associate Editor for TIFS.