Deep Multimodal Sequence Fusion by Regularized Expressive Representation Distillation

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

***Abstract*—Multimodal sequence learning aims to utilize information from different modalities to enhance overall performance. Mainstream works often follow an intermediate-fusion pipeline, which explores both modality-specific and modality-supplementary information for fusion. However, the unaligned and heterogeneously distributed multimodal sequences pose significant challenges to the fusion task: 1) to extract both effective unimodal and crossmodal representations and 2) to overcome the overfitting issue in joint multimodal sequence optimization. In this work, we propose regularized expressive representation distillation (RERD) that aims to seek effective multimodal representations and to enhance the generalization of fusion. First, to improve unimodal representation learning, unimodal representations are assigned to multi-head distillation encoders, where the unimodal representations are iteratively updated through distillation attention layers. Second, to alleviate the overfitting issue in joint crossmodal optimization, a multimodal sinkhorn distance regularizer is proposed to reinforce the expressive representation extraction and to reduce the modality gap before fusion adaptively. These representations produce a comprehensive view of the multimodal sequences, which are utilized for downstream fusion tasks. Experimental results on several popular benchmarks demonstrate that the proposed method achieves state-of-the-art performance, compared with widely used baselines for deep multimodal sequence fusion, as shown in https://github.com/Redaimao/RERD.**

***Index Terms*—Multimodal sequence fusion, multimodal sentiment analysis, regularization.**

I. Introduction

With the advance of modality representation learning in language [1]–[4], audio [5]–[7], and vision [8]–[11], multimodal sequence learning that aims to improve overall performance by fusing multiple sensory data, has drawn much attention recently. Multimodal sequence learning bridges the gaps between different modalities and is expected to provide a more reliable solution with high generalization ability when involving more modalities.

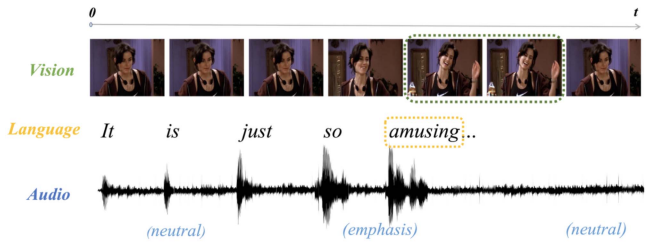


Fig. 1. A typical unaligned multimodal sequence. The video and audio features are not word-aligned. The highlighted expressive features are sparsely distributed.

Various fusion techniques [12]–[14] have been proposed. Among them, intermediate-fusion pipeline has shown significant advantages. On account of this, considerable progress [15]–[19] has been achieved in multimodal fusion for sentiment analysis tasks. As people's sentiment expression is usually through language, voice, and facial movements, recent works [20]–[23] have been proposed to enhance the effectiveness of fusion by probing both modality-specific and modality-supplementary information. This approach in multimodal sentiment analysis has achieved promising results to a large extent.

However, two main concerns, which incur a bottleneck of performance improvement, exist. One is the unaligned[1] nature of multimodal sequences, as pointed out in many previous works [20], [25]. As presented in Fig. 1, the misalignment of multimodal sequences is inherent in real-world scenarios in which the most expressive features are not synchronous. Additionally, in each unimodal sequence data, the prominent feature that most contributes to prediction is sparsely distributed. Such distinctive distributions often impair the modality-specific information extraction.

Meanwhile, although the presence of multiple modalities provides additional information, the unaligned sequences and their

Manuscript received 24 August 2021; revised 11 December 2021 and 29 December 2021; accepted 3 January 2022. Date of publication 13 January 2022; date of current version 7 June 2023. The work was supported by Nanyang Technological University, through NTU Internal Funding - Accelerating Creativity and Excellence (NTU–ACE2020-03). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. *(Corresponding author: Adams Wai-Kin Kong.)*

Xiaobao Guo is with the School of Computer Science and Engineering, Nanyang Technological University Singapore, Singapore 639798, Singapore, and also with the Rapid-Rich Object Search (ROSE) Laboratory, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore 639798, Singapore (e-mail: xiaobao001@e.ntu.edu.sg).

Adams Wai-Kin Kong and Alex Kot are with the Rapid-Rich Object Search (ROSE) Laboratory, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore 639798, Singapore (e-mail: AdamsKong@ntu.edu.sg; eackot@ntu.edu.sg).

Digital Object Identifier 10.1109/TMM.2022.3142448

[1]According to Baltrusaitis *et al.*'s definition [24], the term, unaligned multimodal sequences, refers to the multimodal features that are extracted without word-level alignment [15], [20], [24]. Multimodal signals are implicitly aligned with respect to time.

1520-9210 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

TABLE I
PRELIMINARY EXPERIMENTS WITH SELF-ATTENTION BASED INTERMEDIATE FUSION MODELS ON MOSI AND MOSEI. THE BEST BIMODAL FUSION ACCURACIES OUTPERFORM UNIMODAL, BUT THE TRIMODAL NETWORKS CONSISTENTLY PERFORM WORSE THAN THE BIMODAL ONES

Dataset	Best unimodal	Best bimodal	Trimodal
MOSI	78.05	79.12 (↑ 1.07)	78.51 (↓ 0.61)
MOSEI	77.98	78.95 (↑ 0.97)	78.18 (↓ 0.77)

sparsely distributed prominent features pose a significant challenge to the optimization [24]. Each modality generalizes at a different rate [26], and it is hard to fuse different modalities without overfitting and achieve the most desirable generalization performance when more sensory data is involved. As revealed in Table. I, the best bimodal accuracy may surpass the trimodal one, leading to an unsatisfactory generalization performance. As a result, the trained models may perform inadequately and lead to suboptimal solutions. Although prior works [20]–[23] have achieved admirable results, they either lack of full exploration the unaligned nature of multimodal sequences explicitly or weaken the capability to generalization.

To address the aforementioned problems, we propose deep multimodal sequence fusion by regularized expressive representation distillation (RERD) for efficient representation and joint optimization. We use "distillation" to indicate removing noise and retaining useful information in the unimodal encoders. Two hypotheses were proposed: (1) models must learn expressive unimodal representations before fusion and (2) models must promote joint optimization by encouraging a more regularized distribution with respect to multimodal sequence feature representations during training.

Guided by the hypotheses above, the proposed RERD is comprised of the two major components based on an intermediate-fusion pipeline: (1) a multi-head distillation encoder is proposed to enhance unimodal representations from unaligned multimodal sequences, where the distillation attention layers dynamically capture and extract the most expressive unimodal features and (2) a novel multimodal sinkhorn distance regularizer is introduced to aid the joint optimization in training. The distribution gap between latent unimodal embeddings is shrunk and regulated, which is proved effective for fusion.

To recap, the contributions of this paper are summarized below:

- A multi-head distillation encoder is proposed to extract and enhance the unimodal representations from unaligned sequences;
- A multimodal sinkhorn distance regularizer is introduced to improve the multimodal generalization capability for joint optimization, which strengthens the unimodal representation extraction and reduces the modality gap before fusion; and
- Experimental results from MOSI, MOSEI, and CH-SIMS for both classification and regression tasks demonstrate that the proposed method can achieve state-of-the-art performance when compared with widely used baselines and strong benchmark methods for multimodal sequence fusion.

II. RELATED WORK

A. Multimodal Sequence Learning

Multimodal sequence learning utilizes both modality-specific dynamics and modality-complementary information for fusion. A wide range of neural network approaches has been proposed to learn multimodal sequence representation by fusing input unimodal features per timestep or at the utterance level. Among them, many methods captured the sequential information by taking advantage of long short-term memory (LSTM) [27] or its variants [28]. Zadeh *et al.* [29] used a hybrid LSTM module to learn multimodal information. Zadeh *et al.* [16] employed a recurrent model in which the crossmodal interaction is extracted via a dynamic memory module. Rajagopalan *et al.* [30] modeled both view-specific and cross-view interactions temporally using an LSTM variant. Other methods [17]–[19], [31] adopted a tensor fusion based mechanism to improve the efficiency of multimodal fusion. They exploited tensor representations to capture inter-modality interactions. Chauhan *et al.* [32] introduced an RNN-based model that explores context-aware attention. Subspace learning based methods such as MISA [23], explicitly projected multimodal representations onto modality-specific and modality-invariant subspaces. The learned features were then further utilized for fusion tasks. Beyond that, translation-based methods assumed that one modality could be translated to another by explicit modulation or implicit pairwise learning. Shenoy *et al.* [21] took context information into account and adopted pairwise fusion to learn the inter-dependencies between modalities. Tsai *et al.* [20] used an attention-based mechanism to perform fusion by pairwise translation on both aligned and unaligned sequences.

However, LSTM is difficult to train, when some intermediate representations may lead to conflicts [27]. Besides, LSTM may not be adequate to project all multimodal sequences [20] into a common space by applying pairwise modality translation directly. Although some methods utilized the unaligned data, none of them explicitly modeled the unaligned sequences, which is uniquely addressed in this paper.

B. Fusion Schemes

Previous studies [12]–[14] have indicated that the fusion schemes greatly affect the overall performance. Early fusion is a technique that usually concatenates the input-level features before learning a concept. However, early fusion [33]–[35] may suppress the modality-specific dynamics since the features are combined at an early stage. In contrast to early fusion, late fusion [14], [36], [37] focuses on the individual strength of modalities, where the unimodal feature is used to learn a concept separately, and the outputs are then integrated with a mechanism such as voting. Late fusion may not be effective for learning the modality-supplementary interactions.

To avoid the limitations above, recent methods [18], [20], [21], [23] for multimodal sequence fusion usually take both modality-specific dynamics and modality-supplementary interactions into consideration to achieve promising fusion performances. Following this line of research, in this paper, the intermediate-fusion

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

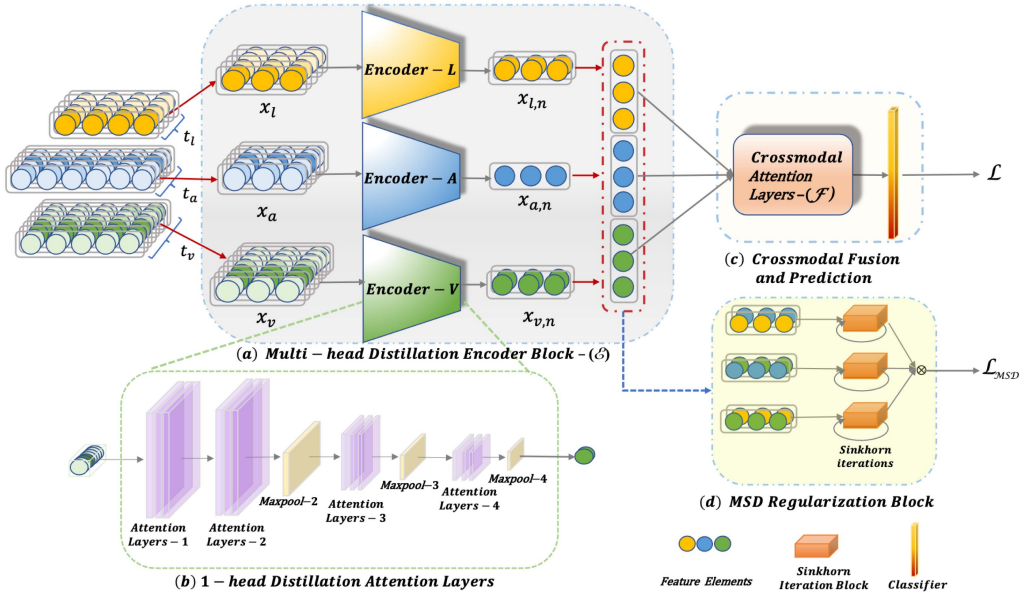


Fig. 2. The architecture of RERD. (a) is the Multi-head Distillation Encoder Block $\mathcal{E}$. x_l, x_a, and x_v are the unaligned unimodal features, where t_l, t_a, and t_v are the corresponding time dimensions. They are first projected to a fixed feature dimension and then encoded through n-layer multi-head distillation encoders, where these unimodal representations are split to each head and then concatenated to produce distilled results. Subsequently, the unimodal representations are projected and concatenated for fusion. The projection layers are presented by the red arrows. (b) are the 1-head Distillation Attention Layers, which are labeled as Attention Layers-1, 2, 3, 4 and Maxpool-2, 3, 4. Each head includes a stack of attention and maxpooling layers to distill the features along the time dimension. (c) is the Crossmodal Fusion and Prediction block. The fused representation is obtained by the full-attention layers $\mathcal{F}$. (d) is the MSD Regularizer. The modality gap is reduced through sinkhorn iterations. The network is jointly optimized by backpropagation of the errors from $\mathcal{L}$ and $\mathcal{L}_{MSD}$ via end-to-end training.

pipeline is also adopted. However, the proposed method aims to further enhance the modeling ability and to circumvent the overfitting issue.

C. Attention Mechanism

The attention mechanism is widely adopted in multimodal sequence learning. Prior works [20], [23] leverage the strong and effective network, transformer [38], to conduct sequence modeling. It shows superiority in training speed as well as performance compared with recurrence modeling. With an encoder-decoder structure, the attention mechanism weights the input sequence and determines the importance of each part at per step, adaptively. Many variants of Transformer have been applied for different tasks [39]–[42]. In multimodal sentiment prediction, both Tsai *et al.* [20] and Hazarika *et al.* [23] employ transformer encoders to fuse unimodal representations. With the benefit from the attention mechanism, multimodal sequence learning yields a more satisfying performance.

In this work, the attention mechanism is also utilized. However, unlike traditional self-attention, the distillation attention layers serve as an effective component to extract expressive unimodal representations from unaligned multimodal sequences.

III. Approach

A. Problem Definition

The i-th training sample is defined as X^i, $X^i = \{x_m^i \in R^{d_m \times t_m}; m \in \{L, A, V\}\}$, where L, A, V represent language, audio, and vision, respectively, and m denotes one of the input modalities. d_m and t_m denote the feature dimension and the length of time sequence, respectively. The goal of multimodal sequence learning can be defined as determining a deep multimodal fusion network $F(X^i|\mathbf{\Theta})$ by minimizing an empirical loss:

$$\min_{F} \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(\hat{y}^i,\ y^i), \tag{1}$$

where $\hat{y}^i$ is the output of $F(X^i|\mathbf{\Theta})$ and K is the number of training samples, so that the output $\hat{y}^i$ approaches the target y^i during the learning process.

B. Overall Network Description

The proposed network aims at first dealing with the unaligned multimodal sequences, and then improving the generalization capability of the fusion. As shown in Fig. 2, based on the intermediate-fusion pipeline, the proposed RERD is comprised of the multi-head distillation encoder block, the crossmodal fusion block, and the Multimodal Sinkhorn Distance (MSD) regularization block.

The first part (Fig. 2(a)) conducts unimodal representation learning, where the unaligned unimodal features are assigned to separate multi-head distillation encoders. Each encoder is composed of multi-head distillation attention layers. Fig. 2(b) shows the 1-head distillation attention layers, containing attention and maxpooling layers. The unimodal representations are distilled along the time dimension, indicating that the timestep lengths are reduced by each attention layer. The multi-head distillation

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

encoders are used to extract the expressive unimodal representations from the input features, which are sparsely distributed in input features.

The second part integrates the learned unimodal representations to produce crossmodal representations, followed by the classifier to output the final predictions. The crossmodal attention layers (Fig. 2(c)) are a stack of self-attention layers, which take the distilled unimodal representations as input.

The unimodal representations are regularized by the MSD regularizer to improve the generalization capability. MSD (Fig. 2(d)), which is a normalized regularization term in the joint optimization process, aims at progressively reducing the modality gap.

C. Expressive Unimodal Feature Learning

As mentioned in the introduction, within each modality representation, the prominent feature that makes the largest contribution to the prediction is sparsely distributed along the timesteps (as shown in Fig. 1), which induces an obstacle for learning expressive unimodal representations. Although traditional transformer encoders provides a successful paradigm in sequence modeling, it may not be efficient at deriving the most expressive unimodal representation since these expressive features are sparsely distributed.

To capture the most expressive feature dynamically, inspired by [39], the *distillation attention layers* are introduced. Considering a unimodal representation $x_m \in R^{d_m \times t_m},\ m \in \{L, A, V\}$, the distilled representation $x_{m,n} \in R^{d_m \times (t_m/2^{(n-1)})}$ is obtained through n distillation attention layers. The distilled representation $x_{m,n}$ is denoted as:

$$x_{m,n} = \begin{cases} Attn(x_m|\mathbf{\Theta}_a), & n = 1 \\ MaxPool(Attn(x_{m,n-1}|\mathbf{\Theta}_a)), & 2 \le n \le \lfloor \log_2 t_m \rfloor + 1 \end{cases} \tag{2}$$

where $\mathbf{\Theta}_a$ are the learned attention parameters and $Attn(\cdot|\mathbf{\Theta}_a)$ is a self-attention operation guided by an active estimation of the unimodal representations per timestep. Given the i-th query $\mathbf{q_i} \in R^{d_m}$ and the j-th key $\mathbf{k_j} \in R^{d_m}$ in $\mathbf{x_m}$, the active estimation is defined as:

$$\mathcal{M}(\mathbf{q_i}, \mathbf{x_m}) = \max_j \left\{ \frac{\mathbf{q_i}\mathbf{k_j}^\top}{\sqrt{d_m}} \right\} - \frac{1}{t_m} \sum_{j=1}^{t_m} \frac{\mathbf{q_i}\mathbf{k_j}^\top}{\sqrt{d_m}}. \tag{3}$$

The unimodal attention of q_i on all other k_j is determined by the dot-product operation in which the most expressive representation attention should hold a larger value in the active estimation. Since the expressive representations are sparsely distributed, to extract the most expressive representation along all timesteps, the operation $Attn$ is conducted using the selected queries q_i corresponding to the largest k $\mathcal{M}(\mathbf{q_i}, \mathbf{x_m})$ following by the $MaxPool$ operation. More clearly, the top-k selected $\mathbf{q_i}$ will form the $\tilde{\mathbf{x}}_{\mathbf{m,k}}$ in the attention operation of the *multi-head distillation encoder*.

The *multi-head distillation encoder* $\mathcal{E}$ is comprised of a projection layer $\mathcal{P}_{ro}$, a stack of distillation attention layers $\mathcal{D}_n$, and a post-alignment layer $\mathcal{P}_a$. Let H be the number of heads defined in the multi-head distillation encoder, and the $Attn(\cdot|\mathbf{\Theta}_h)$ operation in each head be:

$$Attn(\mathbf{x_m}|\mathbf{\Theta}_h) = Softmax\left(\frac{\mathbf{w_{h,1}}\tilde{\mathbf{x}}_{\mathbf{m,k}}\mathbf{x_m}\mathbf{w_{h,2}}}{\sqrt{d_{m/H}}} \right) \mathbf{w_{h,3}}\mathbf{x_m}, \tag{4}$$

where $\tilde{\mathbf{x}}_{\mathbf{m,k}}$ is a matrix with all the top-k queries selected by the active estimation in Eq. 3, h is an integer in $[1, H]$, and $\mathbf{\Theta}_h = (\mathbf{w_{h,1}}, \mathbf{w_{h,2}}, \mathbf{w_{h,3}})$ are the learned parameter matrices in each head. Thus, for each multi-head distillation encoder, the output is the concatenation of the attention results from each head, which is further used for crossmodal fusion.

Given a unimodal representation $x_m \in R^{d_m \times t_m},\ m \in \{L, A, V\}$, the encoded unimodal representation can be denoted as:

$$x^{\mathcal{E}}_{m,n} = \mathcal{E}(x_m|n, \mathbf{\Theta}_e) = \mathcal{P}_a \circ \mathcal{D}_n \circ \mathcal{P}_{ro}(x_m), \tag{5}$$

where $x^{\mathcal{E}}_{m,n} \in R^{\tilde{d}_m \times \tilde{t}_m}$, $\tilde{d}_m$ is the projected feature dimension, $\tilde{t}_m = min\{t_m/2^{(n-1)}, m \in \{L, A, V\}\}$, $\circ$ denotes the composite operator, and $\mathbf{\Theta}_e = (\mathbf{\Theta}_{\mathcal{P}_a}, \mathbf{\Theta}_{\mathcal{D}_n}, \mathbf{\Theta}_{\mathcal{P}_{ro}})$ are the optimized parameters in the encoder. Specifically, $\mathcal{P}_{ro}(\cdot|\mathbf{\Theta}_{\boldsymbol{\mathcal{P}_{ro}}})$ is a 1d-convolutional layer that projects the feature dimension of each modality to a fixed size, which is set as the latent feature dimension in the distillation attention layers. $\mathcal{D}_n(\cdot|H, \mathbf{\Theta}_{\boldsymbol{\mathcal{D}_n}})$ are the H-head n-layer distillation attention layers. $\mathcal{P}_a(\cdot|\mathbf{\Theta}_{\boldsymbol{\mathcal{P}_a}})$ is also a 1d-convolutional layer that aligns the distilled timestep to the minimum one in all modalities after distillation attention layers. The reason for this operation is to enable the concatenation of multimodal sequences along the feature dimension before crossmodal learning.

The expressive feature distillation is efficient at extracting the unimodal dynamics, particularly for unaligned multimodal sequences, since it does not leverage the aligned information from the pair of modalities as input. The distillation attention layers contribute to a better unimodal representation in the unimodal encoder, while the multi-head distillation encoder can be optimized to permit expressive unimodal feature learning.

D. Crossmodal Representation Learning

The crossmodal representation learning adopts self-attention layers [38] instead of distillation attention. Since the most expressive unimodal representations have been obtained through the multi-head distillation encoders, the crossmodal fusion $\mathcal{F}$ aims at learning the modality-common information through all distilled unimodal features. The fused representation $x^{\mathcal{F}}_f$ is denoted as:

$$x^{\mathcal{F}}_f = \mathcal{F}(\oplus\, x^{\mathcal{E}}_{m,n}|\mathbf{\Theta}_f), \tag{6}$$

where $\mathbf{\Theta}_f$ are the optimized parameters in the crossmodal fusion module and $f \in \{L, A, V, (L, A), (A, V), (V, A), (L, A, V)\}$ denotes any possible combinations of the three multimodal inputs. $\oplus$ denotes the concatenation operation. The joint multimodal representation is produced through the crossmodal attention layers, which is vital to the performance improvement

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

for downstream tasks. Thus, to produce better joint representations while dealing with the overfitting issue, the Multimodal Sinkhorn Distance (MSD) Regularization is proposed in the following section.

E. Multimodal Sinkhorn Distance Regularization

The inherent heterogeneity of multimodal sequence distributions obstructs the joint optimization in multimodal fusion, leading to an overfitting problem and a suboptimal fusion result (see the results of the best bimodal and trimodal accuracies in Table I). To overcome this issue, it is natural to think about reducing the modality gap by regularizing the heterogeneous distributions. To achieve this, the **MSD regularizer** is introduced.

The distilled unimodal representations from different modalities have different distributions. It is crucial to choose a suitable distance measure for heterogeneous distributions. Wasserstein distance [43] offers a desirable solution to measure the distance between different distributions. It holds a favorable property that the distance can still be measured, even without overlapping distributions. To apply the distance measure between two modality distributions in an end-to-end training process, its approximation via Sinkhorn iterations [44] is needed.

The *p-Wasserstein distance* between probability distributions μ and ν over a metric space $\mathcal{X}$ is:

$$\mathcal{W}_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{X}} d(x_1,x_2)^p d\pi(x_1,x_2)\right)^{(1/p)}, \tag{7}$$

where the infimum is taken over all the possible joint probability distributions $\Pi(\mu,\nu)$ that distribute the mass in μ to match that in ν, with the p-th power of the *ground metric* $d(x_1,x_2)$ on $\mathcal{X}$ giving the cost of moving a unit of mass from $x_1\in\mathcal{X}$ from μ to $x_2\in\mathcal{X}$ from ν. The *p-Wasserstein distance* is the optimal cost of matching two marginal distributions μ and ν [45].

To make the optimization problem solvable iteratively in an end-to-end process, *sinkhorn iterations* [44] with an entropic regularization term:

$$\mathcal{H}(\pi)=-\sum_{\pi\in\Pi(\mu,\nu)}\pi\log\pi. \tag{8}$$

are used. The *sinkhorn distance* can be formulated as:

$$\mathcal{L}_s(\mu,\nu)=\min_{\pi\in\Pi(\mu,\nu)}\mathcal{W}_1(\mu,\nu)-\epsilon\mathcal{H}(\pi) \tag{9}$$

where ϵ is the trade-off parameter of the two terms and the subscript **1** means the *1-Wasserstein distance*.

Formally, let p_a and p_b be two marginal distributions from distinct modalities in language, audio, and visual. By embedding the unimodal representations as probability distributions in the Wasserstein space with the regularization, the *multimodal sinkhorn distance* for the n-modal fusion is defined as:

$$\mathcal{L}_{MSD}=\frac{1}{n}\sum_{n}\mathcal{L}_s(p_a,p_b), \tag{10}$$

where $a,b\in\{L,A,V\}, a\neq b$ and n indicates the total number of combinations of a and b.

The goal of $\mathcal{L}_{MSD}$ is to progressively minimize the multimodal embedding distributions, which serve as a regularization term in the joint multimodal optimization:

$$\min_F \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}(\hat{y}^i=F(X^i,\mathbf{\Theta}),\ y^i)+\lambda\mathcal{L}_{MSD}, \tag{11}$$

where λ is a trade-off parameter, and $\mathcal{L}$ is either classification or regression loss depending on the type of task. It is vital to choose a proper distance measure for heterogeneous distributions. As there is no prior knowledge of the distributions, the Wasserstein distance is preferred. The *MSD regularizer* provides flexibility during modeling because it does not force the modal to choose a particular parametric distribution. Reducing the modality gaps helps produce better representations. The information is projected to the embedding space, where the modality complementarities are extracted more efficiently. The MSD iterations can be executed efficiently on GPUs and the total loss is fully differentiable, making it a desirable measurement for the modality gap. Therefore, the proposed MSD regularizer is efficient at mitigating the overfitting issue in joint multimodal optimization, which is evinced by experiments in the following section.

IV. Experiment

To test the performance of the proposed RERD network on classification and regression tasks for unaligned multimodal sequence fusion, extensive experiments were conducted to compare with multiple benchmark methods. To validate the effectiveness of the MSD regularization, the generalization comparison is reported by comparing with several recent state-of-the-art methods.

A. Datasets

1) MOSI: The Multimodal Corpus of Sentiment Intensity (MOSI) [46] dataset consists of 2,199 short monologue video clips, which are collected from YouTube. Each video clip is a movie review in which a speaker expresses his/her opinions. Each opinion video is annotated with sentiment in the range [-3,3] in which −3/+3 represents strongly negative/positive sentiments.

2) MOSEI: The CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) [15] dataset[2] is an improvement dataset on MOSI and is the largest dataset for multimodal sentiment analysis and emotion recognition. The dataset is made up of 23,453 movie review video clips from 1000 distinct speakers. The sentence utterances are chosen randomly from 250 different topics.

3) CH-SIMS: The Chinese Single- and Multi- Modal Sentiment Analysis (CH-SIMS) dataset [47] contains 2,281 refined video segments with fine-grained annotations of modalities. The data is collected from movies, TV serials, and variety shows. The annotation for each sample ranges from -1 (strongly negative) to 1 (strongly positive).

[2]Please refer to [Online]. Available: https://github.com/A2Zadeh/CMU-MultimodalSDK for more information.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

B. Feature Extraction

1) Language: Two methods are utilized for language feature extraction. The first method, GloVe [1], generates language modality features with a 300-dimensional embedding for each token in an utterance. As recent works [23], [48] use a pre-trained BERT [3] as the feature extraction method for textual information, in this work, we also utilize a BERT pre-trained model, which consists of 12 Transformer layers. The resultant utterance feature is an average representation of the tokens, which is a 768-dimensional vector. For CH-SIMS dataset[3], the language features are extracted using Chinese-BERT [3]. For a fair comparison, experiments are conducted on both GloVe-based and BERT-based language features in this work.

2) Audio: The acoustic features for MOSI and MOSEI are extracted by COVAREP [5], an acoustic analysis tool. The features contain 12 Mel-Frequency Cepstral Coefficients (MFCCs), pitch, Voiced/UnVoiced segmenting features (VUV), glottal source parameters, and other features related to the tone of speech, such as peak slope parameters, resulting a dimension of 74. LibROSA [49] speech toolkit with default setting is adopted to extract acoustic feature from CH-SIMS, which results in a 33-dimensional feature vector.

3) Vision: Visual features of MOSI and MOSEI are extracted by Facet [50] for each sampled frame in the video clips. The facial expression features include 35 facial action units to represent facial muscle movements based on the Facial Action Coding System (FACS). For CH-SIMS visual features, MTCNN [51] and MultiComp OpenFace2.0 toolkit [24] are used to extract 709-dimensional features, including facial landmarks, facial action units, head pose, etc.

C. Tasks and Evaluation Criteria

To validate the effectiveness of our method, extensive experiments are conducted on both classification and regression tasks on MOSI, MOSEI, and CH-SIMS datasets. Following the benchmark evaluation methods for the classification task, weighted F1 score (F1-Score), binary classification accuracy (Acc-2), and seven-class accuracy (Acc-7) ranging from -3 to 3 are reported. Since CH-SIMS only has 2-class and 5-class labels, the Acc-7 performance index is not suitable for it. For the regression task, the mean absolute error (MAE) and Pearson's correlation (Corr) are adopted as the metrics. Previous works use two distinct measurements for binary accuracy scores, namely negative/non-negative classification [17] (*i.e.*, sentiment labels are assigned as <0 and ≥ 0) and negative/positive classification [20] (*i.e.*, sentiment labels are assigned as <0 and >0). For a fair comparison in this work, results on both these metrics are reported using the segmentation marker /, where the left-side result is for neg./non-neg. and the right-side result is for neg./pos. classification.

D. Compared Benchmarks

Most of the state-of-the-art methods follow the intermediate-fusion pipeline and many of them extract both the modality-specific and modality-complementary information. They can be categorized as follows:

- RNN-based models: **BC-LSTM** [52];
- Tensor fusion based, low rank fusion and the variants models: **TFN** [17], **LMF** [18], **HFFN** [31], and **LMFN** [19];
- Utterance and temporal modeling methods: **MV-LSTM** [30], **MARN** [29], **MFN** [16];
- Nonverbal subword sequences learning: **RAVEN** [53];
- Context based and inter-utterance attention method: **CIA** [32];
- Graph-based method: **GRAFH-mfn** [15]
- Adversarial auto-encoder based model: **ARGF** [22];
- Canonical correlation analysis model: **ICCN** [48];
- Invariant and specific subspace learning: **MISA** [23];
- Translation-based fusion and pairwise learning models: **MulT** [20], **Mu-net** [21]

Among the methods above, MulT, Mu-net, and MISA have been published recently and were found to outperform the others either on MOSI or MOSEI datasets. Specifically, MulT [20] uses explicit source-target modality translation by extending the transformer architecture with pairwise crossmodal attention. Mu-Net [21] adopts a pairwise attention mechanism for fusion. MISA [23] is proposed to learn modality-invariant and modality-specific representations by minimizing a combination of losses, including distributional similarity, orthogonal loss, reconstruction loss, and task prediction loss. In our experiments, to compare the generalization performance and stability for unaligned multimodal sequences, MulT, Mu-net, and MISA are utilized as baseline models, as they are considered as the recent state-of-the-art (SOTA) methods.

E. Quantitative Results Compared With Benchmarks

Extensive experiments are conducted on the two multimodal sentiment analysis datasets for both classification and regression tasks for unaligned sequence fusion. To rule out randomness, the average results over 10-round runs with different initializations of the proposed method are listed. Table II shows the results on MOSI, and our method leads by a large margin in the classification results. For the regression task, our method also achieves competitive results. Table III exhibits the results on MOSEI in which our method reaches the highest performance on classification while having comparable or better performance on regression. Table IV shows the results on CH-SIMS. The compared methods use BERT-based language features provided by the benchmark dataset. RERD achieves the best performance compared with the baselines[4]. Overall, RERD achieves the best performance and outperforms the state-of-the-art methods in most of the metrics on these benchmark datasets.

It is worth noting that RERD consistently improves the performance of both GloVe-based and BERT-based language features.

[3]The dataset features are provided by [Online]. Available: https://github.com/thuiar/MMSA

[4]The compared results are from [Online]. Available: https://github.com/thuiar/MMSA

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

TABLE II
SENTIMENT PERFORMANCE COMPARISON ON MOSI. F1 AND ACC ARE PERCENTAGES (%). (B) MEANS THE LANGUAGE FEATURES ARE BASED ON BERT, [a] FROM [20], [b] FROM [22], AND [c] FROM [48]

Method	F1 (↑)	Acc-2 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)
BC-LSTM	73.9 / -	73.9 / -	1.079	0.581	28.7
TFN	73.4 / -	73.9 / -	0.970	0.633	32.1
LMF	75.7 / -	76.4 / -	0.912	0.670	32.8
MARN	77.0 / -	77.1 / -	0.968	0.625	34.7
MFN[a]	77.3 / -	77.4 / -	0.965	0.632	34.1
RAVEN[a]	76.6 / -	78.0 / -	0.915	0.691	33.2
CIA	79.5 / -	79.8 / -	0.914	0.689	38.9
HFFN[b]	- / 80.3	- / 80.2	-	-	-
LMFN[b]	- / 80.9	- / 80.9	-	-	-
MulT	- / 81.0	- / 81.1	0.870	0.690	39.1
ARGF	- / 81.4	- / 81.3	-	-	-
Mu-net	- / 80.1	- / 81.2	-	-	-
RERD (Ours)	**81.64 / 82.52**	**81.55 / 82.47**	**0.868**	**0.690**	**39.6**
Δ_{SOTA}	2.14 ↑ / 1.12 ↑	1.75 ↑ / 1.17 ↑	0.002 ↓	-	0.5 ↑
TFN(B)[c]	- / 80.7	- / 80.8	0.901	0.698	34.9
LMF(B)[c]	- / 82.4	- / 82.5	0.917	0.695	33.2
MFM(B)[c]	- / 81.6	- / 81.7	0.877	0.706	35.4
ICCN(B)	- / 83.02	- / 83.07	0.862	0.714	39.0
MISA(B)	81.82 / 83.58	81.84 / 83.54	0.776	**0.778**	41.3
RERD(B) (Ours)	**83.30 / 83.66**	**83.33 / 84.52**	**0.709**	0.723	**41.9**
Δ_{SOTA}	1.48 ↑ / 0.08 ↑	1.49 ↑ / 0.98 ↑	0.067 ↓	0.055 ↓	0.6 ↑

TABLE III
SENTIMENT PERFORMANCE COMPARISON ON MOSEI. F1 AND ACC ARE PERCENTAGES (%). (B) MEANS THE LANGUAGE FEATURES ARE BASED ON BERT, [a] FROM [20], AND [c] FROM [48]

Method	F1 (↑)	Acc-2 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)
LMF	55.80 / -	59.41/ -	-	-	-
LMFN	- / 59.48	- / 61.31	-	-	-
ARGF	- / 59.03	- / 60.77	-	-	-
HFFN	- / 59.07	- / 60.07	-	-	-
MFN[a]	76.0 / -	76.0 / -	-	-	-
MV-LSTM[a]	76.4 / -	76.4 / -	-	-	-
GRAFH-mfn[a]	77.0 / -	76.90 / -	0.71	0.54	45.0
RAVEN	79.50 / -	79.10 / -	0.614	0.662	50.0
CIA	78.23 / -	80.37 / -	0.68	0.59	50.1
MulT	- / 81.60	- / 81.60	**0.59**	**0.69**	50.7
Mu-net	- / 80.01	- / 82.10	**0.59**	0.50	-
RERD (Ours)	**80.7 / 81.96**	**80.54 / 82.25**	0.632	0.676	**50.9**
Δ_{SOTA}	1.20 ↑ / 0.36 ↑	0.17 ↑ / 0.15 ↑	0.042 ↑	0.014 ↓	0.2 ↑
TFN(B)[c]	- / 82.1	- / 82.5	0.593	0.70	50.2
LMF(B)[c]	- / 82.1	- / 82	0.623	0.677	48.0
MFM(B)[c]	- / 84.3	- / 84.4	0.568	0.717	51.3
ICCN(B)	- / 84.2	- / 84.2	0.565	0.713	51.6
MISA(B)	81.12 / 83.66	80.67 / 83.87	0.558	0.752	52.1
RERD(B) (Ours)	**83.40 / 84.78**	**83.19 / 84.75**	**0.550**	**0.761**	**52.5**
Δ_{SOTA}	2.28 ↑ / 0.91 ↑	2.52 ↑ / 0.88 ↑	0.008 ↓	0.009 ↑	0.3 ↑

TABLE IV
PERFORMANCE COMPARISON ON CH-SIMS. (B) MEANS THE LANGUAGE FEATURES ARE BASED ON BERT

Method	F1 (↑)	Acc-2 (↑)	MAE (↓)	Corr (↑)
TFN(B)	75.66 / 52.79	75.32 / 53.56	0.432	0.591
LMF(B)	77.59 / 53.83	77.99 / 57.06	0.441	0.576
GRAPH-mfn(B)	78.92 / 54.66	79.21 / **57.99**	0.445	0.579
MulT(B)	78.07 / 54.26	78.07 / 56.34	0.453	0.564
MISA(B)	77.70 / 53.99	78.07 / 57.27	-	-
RERD (B) (Ours)	**80.34 / 55.12**	**80.48 / 57.99**	**0.428**	**0.594**
Δ_{SOTA}	1.42 ↑ / 0.46 ↑	1.27 ↑ / -	0.004 ↓	0.003 ↑

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

TABLE V
GENERALIZATION PERFORMANCE COMPARISON OF BIMODAL TO TRIMODAL IMPROVEMENT ON MOSI AND MOSEI FOR NEG./POS. CLASSIFICATION TASK. RESULTS IN () ARE FROM BERT-BASED TEXT FEATURES. THE UNDERLINED PER-TYPE BIMODAL ACCURACIES ARE OVERFITTED COMPARED TO THE RESPECTIVE TRIMODAL RESULTS

Dataset	Method	Per-type bi-Acc L+A/L+V/A+V	bimodal avg Acc	trimodal Acc	Δ tri-bi (↑)
MOSI	MulT	81.25 (82.60) / 79.60 (81.26) / 75.65	78.83 (79.84)	81.1 (82.21)	2.27 (2.37)
	Mu-net	80.68 (81.76) / 79.25 (80.63) / 76.12	78.68 (79.50)	81.2 (82.05)	2.52 (2.55)
	MISA	81.55 (83.86) / 80.75 (82.37) / 76.95	79.75 (81.06)	82.05 (83.54)	2.3 (2.48)
	RERD (Ours)	81.75 (83.25) / 80.26 (82.88) / 77.48	**79.83** (**81.20**)	**82.47** (**84.52**))	**2.64** ↑ (**3.32** ↑)
MOSEI	MulT	80.94 (81.75) / 78.57 (80.91) / 74.31	77.94 (78.99)	79.07 (81.26)	1.13 (2.27)
	Mu-net	80.18 (81.90) / 80.06 (81.27) / 75.16	78.47 (79.44)	78.77 (81.65)	0.30 (2.21)
	MISA	79.55 (82.51) / 81.87 (83.97) / 77.52	79.08 (**81.33**)	78.68 (83.87)	1.19 (2.54)
	RERD (Ours)	80.68 (82.58) / 79.75 (81.59) / 77.99	**79.47** (80.72)	**82.25** (**84.75**)	**2.78** ↑ (**4.03** ↑)

This further validates that the proposed fusion method can be effective regardless of different feature extraction methods.

The results demonstrate that expressive unimodal representation distillation is vital for modality fusion. They also show the competitive performance because of the ample exploration of unimodal dynamics from multi-head distillation encoders and the effective crossmodal representation by the MSD regularization. RERD is effective at adaptively shrinking the modality gap and extracting expressive representations from unaligned multimodal sequences.

F. Generalization Comparison

1) Bimodal to Trimodal Fusion: Multimodal fusion aims to use information from different modalities to enhance system performance, and therefore, generalization performance under the different number of modalities is vital. To validate the strength of RERD in terms of generalization, Table V exhibits the bimodal to trimodal performance gains of neg./pos. classification accuracies. The experiments are conducted on both MOSI and MOSEI. The compared methods MulT[5], Mu-net[6], and MISA[7] are strong benchmark methods on MOSI and MOSEI.

Specifically, the models are trained with L+A, A+V, and L+V inputs for 10 rounds with different initializations. The per-type bimodal accuracy is the average over the 10 rounds results, and the bimodal average accuracy is the average value over all per-type bimodal accuracies. For the trimodal case, with the same experimental settings, our results are averaged by running 10 rounds with different initializations. All the baseline methods are conducted under the same environment for fair comparisons. Δ stands for the average gain from bimodal to trimodal fusion. Compared with the state-of-the-art methods, in addition to the better average results on both bimodal and trimodal fusions, RERD shows better performance on bi-to-tri gain, which demonstrates the strong multimodal generalization capability.

Furthermore, the compared baseline methods suffer from the overfitting issue while RERD is not affected, as shown in V. By comparing the per-type bimodal accuracies (*i.e*, the underlined values) with the respective trimodal performances, it is worth noting that the best per-type bimodal accuracies of MulT, Mu-net, and MISA outperform their trimodal ones thus revealing the deficiency of generalization by adding a certain modality for fusion. In other words, by adding more modality, the baseline methods face degradation, especially when using GloVe-based language features. Although it is often the case that the BERT-based language features perform better than GloVe-based features, the baseline methods still have fewer bi-to-tri gains. In contrast, RERD is not only free from overfitting but also improves the performance on both GloVe-based and BERT-based features. Through MSD regularization, RERD can easily overcome this issue to achieve a better generalization capability.

2) Learning Curve Analysis: The proposed RERD can handle the overfitting issue when more modalities are involved in the fusion process. To show this point, RERD is compared with a baseline model that has the same fusion pipeline. However, it neither uses MSD regularization nor the proposed distillation attention layers. Fig. 3 shows the training and validation curves from bimodal (*i.e.*, L+A) and trimodal (*i.e.*, L+A+V) to demonstrate the overfitting issues. Fig. 3(a) shows learning curves of the baseline model, which has only self-attention layers in encoders and the fusion module. Fig. 3(b) shows the learning curves of RERD. Fig. 3 demonstrates that the baseline model suffers from overfitting seriously in trimodal setting but RERD can overcome it and effectively use information from the additional modality *i.e.*, V to achieve better performance. In other words, RERD is effective for improving the generalization capability when dealing with more modalities and causing an improvement in the overall performance compared to the baseline method.

3) Stability and Convergence Time: The assessment of training is equally important for interpreting model quality. In this sense, multiple indicators, including average accuracy, standard deviation, and average convergence time, are compared in the trimodal fusion setting. The average neg./pos. classification accuracies from GloVe-based language features are taken as an example for clarity in this section. For fair comparisons, the experiments are conducted on one GPU card (*i.e*, Tesla M40 with 24 GB memory) with their best settings. As shown in Table VI, RERD enjoys a higher average accuracy, less fluctuation, and higher stability. This may give credit to the reliable expressive feature distillation mechanism and the

[5][Online]. Available: https://github.com/yaohungt/Multimodal-Transformer
[6][Online]. Available: https://github.com/amanshenoy/multilogue-net
[7][Online]. Available: https://github.com/declare-lab/MISA

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

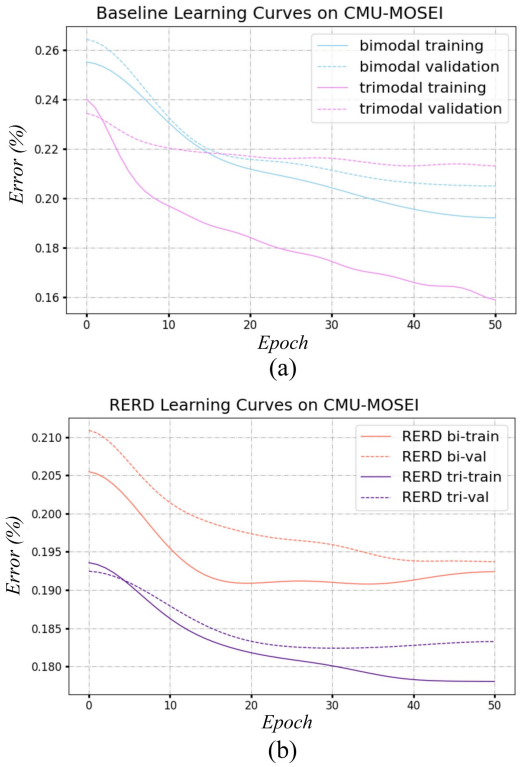


Fig. 3. Learning curves for training and validation sets on MOSEI for neg./pos. classification task. (a) is the curves of a baseline with only self-attention layers for all the encoders and the fusion module. (b) is the learning curves of RERD, with the distillation encoders and the MSD regularization.

TABLE VI
MULTIPLE INDICATOR ASSESSMENT: AVERAGE ACCURACY (AVG ACC), STABILITY (STD), AND AVERAGE CONVERGENCE TIME (ACT)

Method	avg Acc	Std	ACT (mins)
MulT	79.07	2.26	728.33
Mu-net	78.77	1.34	394.17
MISA	78.68	0.94	378.33
RERD (trimodal)	**82.25**	**0.52**	**322.67**

efficient MSD regularization. Beyond that, RERD has a higher training speed, compared with the other methods, which also shows the superiority of the proposed method.

G. Ablation Study

1) Quantitative Results: RERD is proposed to reduce the modality gap and explore the expressive representations for fusion. To validate this, an ablation study has been conducted for neg./pos. classification on MOSEI for clear comparisons. The quantitative results are summarized in Table VII. It can be observed that 1) RERD gains a large improvement through the multi-head distillation encoder; 2) RERD achieves higher performance with the distillation attention layers and 3) MSD regularization can boost the performance for fusion. Hence, the proposed multi-head distillation encoder is efficacious in exploring expressive unimodal representation from unaligned sequences, which evinces the hypothesis 1 (*i.e.*, Models must learn expressive unimodal representations before fusion). The MSD regularization is advantageous at reducing the modality gap and improving the overall performance, which supports the hypothesis 2 (*i.e.*, Models must promote joint optimization by encouraging a more regularized distribution with respect to multimodal sequence feature representations during training).

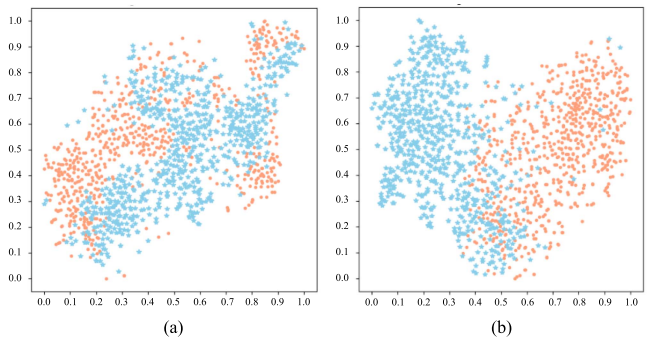


Fig. 4. Multimodal feature distributions in embedding space before the classifier. (a) is the distribution w/o expressive representation distillation layers and MSD regularization and (b) is the distribution of RERD. Blue and orange points stand for positive and negative classes, respectively.

2) Detailed Results of Modality Combinations: The detailed results on MOSEI are presented in Table VIII. Overall, the average performance from unimodal to trimodal is increasing, which shows RERD can improve the generalization performance when adding more modalities. For bimodal cases, RERD is capable of extracting modality-common information and achieve a satisfying average performance. The results for unimodal predictions achieve higher accuracies than some of trimodal predictions of the compared benchmark methods in Table III, showing that our method is suitable for exploring unimodal representations from unaligned sequences.

3) Visualization and Analysis:

a) t-SNE embeddings before classification: To further show the advantage of the proposed RERD on the multimodal sequence fusion problem intuitively, the visualization of multimodal feature distribution in the embedding space right before 2-class sentiment prediction is provided in Fig. 4. The t-SNE method is utilized to project the multimodal features to two dimensions. (a) is the visualization of the distribution without applying the distillation and regularization, while (b) uses distillation attention layers and MSD regularization. The proposed method substantially contributes to a more regularized and separable multimodal distribution. It reveals that expressive representation and MSD regularization can be beneficial for solving the multimodal fusion problem.

b) MSD visualization: The MSD regularizer causes a progressive reduction in the multimodal embedding distances by sinkhorn iterations during training. Fig. 5 shows the MSD in the training process. Specifically, the distance is measured between a pair of modalities. Each axis represents the time dimension of the language (L), audio (A), or visual (V) feature. The square lattice depicts the average distance of the latent representation per time step between two modalities in the training data. It is worth noting that the sequences are not aligned and the

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
ABLATION STUDY ON MOSEI FOR NEG./POS. CLASSIFICATION, REGRESSION, AND ACC-7 CLASSIFICATION. THE RESULTS IN PARENTHESES ARE BASED ON BERT-BASED LANGUAGE FEATURES. THE COMPARED MODELS ARE: THE RERD MODEL, RERD W/O MSD REGULARIZATION, RERD W/O DISTILLATION LAYERS, AND RERD W/O MULTI-HEAD DISTILLATION ENCODER

Model	avg F1 (↑)	avg Acc (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)
RERD	**81.96 (84.78)**	**82.25 (84.75)**	**0.632 (0.550)**	**0.676 (0.761)**	**50.9 (52.5)**
RERD w/o MSD regularization	80.64 (81.77)	80.52 (82.11)	0.666 (0.619)	0.643 (0.688)	48.7 (50.4)
RERD w/o distillation layers	79.78 (80.56)	79.39 (81.17)	0.670 (0.655)	0.599 (0.623)	45.1 (47.4)
RERD w/o multi-head distillation encoder	58.38 (66.32)	59.03 (64.50)	0.813 (0.788)	0.182 (0.286)	32.3 (35.0)

TABLE VIII
DETAILED RESULTS OF MODALITY COMBINATIONS ON MOSEI FOR NEG./POS. CLASSIFICATION WITH GLOVE-BASED LANGUAGE FEATURES

Modality	F1	Acc	MAE	Corr	avg F1	avg Acc	avg MAE	avg Corr
L+A+V	**81.96**	**82.25**	**0.632**	**0.676**	-	-	-	-
L+A	80.16	80.68	0.643	0.663	79.21	79.47	0.674	0.612
L+V	79.18	79.75	0.654	0.642				
A+V	78.30	77.99	0.726	0.527				
L	79.81	79.09	0.657	0.634	76.97	76.91	0.770	0.504
A	76.71	77.72	0.819	0.527				
V	74.38	73.93	0.835	0.352				

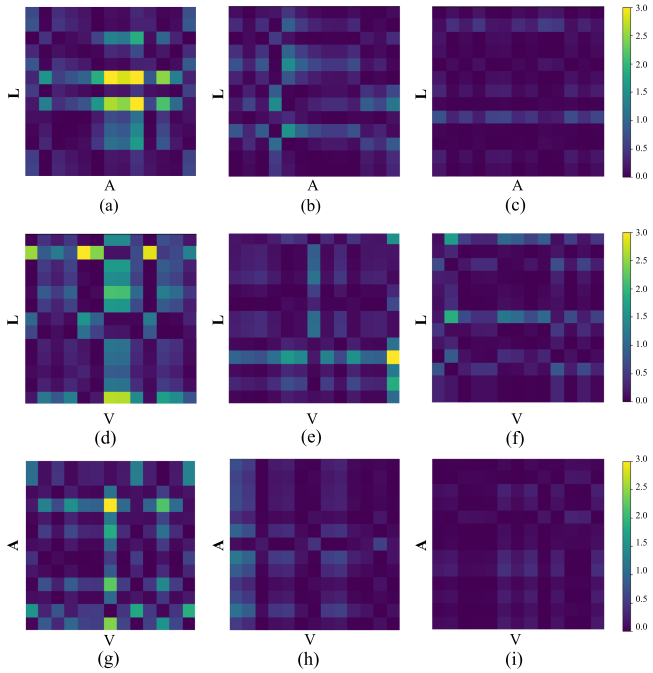


Fig. 5. Multimodal Sinkhorn Distance Martix Visualization. From (a) to (c), (d) to (f), and (g) to (i) show the modality embedding distances between LA, LV, and VA, respectively. The distances are getting smaller in the learning process (*i.e.*, epoch-5, 15, 25).

color with lighter intensity shows larger distance. For example, in sub-figure (a), the horizontal axis is the language modality and the vertical axis is the audio modality for each time step. The colored squares are the average distance between language and audio representations. The average distances are normalized between 0 to 3 for visualization. From sub-figure (a) to (c), the sinkhorn distance between language and audio representations declines as the color of the square lattices gets dimmer. Sub-figure (d) to (f) and (g) to (i) show the embedding distance of modality pair L+V and V+A, respectively. During training, MSD regularization reduces the modality gap and learns a better optimization result.

V. CONCLUSION

In this work, we proposed RERD for efficient, deep, multimodal sequence learning. Based on the intermediate-fusion pipeline, RERD is highly effective for extracting expressive unimodal representations from the unaligned sequences through multi-head distillation encoders. Furthermore, to avoid the overfitting issue in the joint multimodal optimization, the MSD regularizer is introduced in the end-to-end training process. It was validated that MSD is competent for supporting expressive representation learning and reducing the modality gap before fusion, which is preferred in the multimodal fusion task. Our experiments indicate that RERD can achieve state-of-the-art results on multiple multimodal datasets. In the future, to further explore multimodal learning, we will involve more multimodal tasks such as multimodal activity recognition. We will also investigate the robustness of this method against noisy modalities, which is also important for understanding of generalizability for fusion.

APPENDIX A
IMPLEMENTATION DETAILS

A. Training and Testing

1) Training: The proposed RERD was implemented by the deep learning toolkit PyTorch and only one Tesla M40 with a 24 GB memory GPU card was used in the experiments. On the MOSI [46] dataset, the batch size was set to 16 and the basic learning rate was 1e-3. The Adam optimizer was adopted during the optimization with the default setting for hyper-parameters.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
HYPERPARAMETERS OF TRAINING RERD

Parameter Name	MOSI	MOSEI / CH-SIMS
Latent Dimension	64	64
Distillation Heads	6	4
Distillation Layers	3	3
Crossmodal Heads	5	5
Crossmodal Layers	3	3
Base Learning Rate	1e-3	1e-3
Max Learning Rate	-	5e-3
Optimizer	Adam	Adam
Batchsize	16	32
lambda of MSD	0.3	0.2
Out Dropout	0.5	0.5

The model was trained for 50 epochs. After 20 epochs, the learning rate was divided by 10. The early stop was used when test accuracy was remained unchanged for 7 epochs.

On the MOSEI [15] dataset, the batch size was set to 32, the basic learning rate was 1e-3, and the max learning rate was 5e-3. The cyclic learning rate was adopted as a scheduler in training. The dropout rate was set to 0.5 for all operations.

2) Testing: The testing process was conducted after each training epoch. Final results were collected by selecting the best-performing model in 10 trainings with different initializations. However, to rule out randomness and to show the overall performance, the average results were reported.

3) Hyper-parameters: Table IX shows the settings of training on different datasets. Specifically, the Latent Dimension indicates the latent feature dimension in the multimodal distillation encoders. Distillation Heads and Layers are the number of heads and layers in each encoder. The numbers are set the same for all unimodal distillation encoders. Lambda of MSD is the trade-off term of the MSD regularizer.

APPENDIX B
NETWORK STRUCTURE

The example is 2-class sentiment analysis on MOSEI with a batch size of 32. GloVe-based language input is taken as an example. For audio and vision inputs, the basic structure of the multimodal distillation encoder is the same. Specifically, the language feature has a size of (32, 300, 50), while audio and vision features are (32, 74, 500) and (32, 35, 500), respectively. The first dimension represents the batch size, the middle dimension represents time length, and the last dimension for each representation is the feature dimension. The input features are not aligned in the time dimension.

The first step in the multi-head distillation encoder is to transform each feature to a latent dimension using the conv1d operation, while the latent dimension is set to 64 in our experiments. Then, the unimodal representation is processed by multi-head attention layers. Here, the number of heads is set to 6 and the number of layers is 3. The output of the multi-head distillation encoder has a tensor size of (batch size, distilled time dimension, 30).

The MSD regularizer takes the distilled unimodal representations as input. They are first combined into three pairs. The distance between each pair is calculated by sinkhorn iterations. Subsequently, these distances are averaged as the regularization term in the end-to-end training process.

Crossmodal Attention Layers take distilled unimodal representations as input. The concatenation is along the feature dimension. The concatenated representations are then sent to multi-head attention layers. The output of crossmodal attention layers is (1, batch size, 90).

The learned multimodal representation is then sent to the classifier for prediction. The Linear-ReLU-Dropout-Linear-Linear layers are used as a classifier to map the feature to a tensor with the size of (1, batch size, 1) for 2-class sentiment prediction.

REFERENCES

[1] 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.

[2] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 437–10446.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chap. Assoc. Comput. Linguistics: Human Language Technol.*, 2019, pp. 4171–4186.

[4] T. B. Brown *et al.*, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[5] 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.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[7] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6675–6679.

[8] 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.

[9] H. Tian, Y. Tao, S. Pouyanfar, S.-C. Chen, and M.-L. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019.

[10] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, no. 1, pp. 1–25, 2020.

[12] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2005, vol. 4, pp. 3437–3443.

[13] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.

[14] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang, "Robust late fusion with rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3021–3028.

[15] 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 (vol. 1: Long Papers)*, 2018, pp. 2236–2246.

[16] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 5634–5641.

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

[18] P. Liu, X. Qiu, and X.-J. Huang, "Adversarial multi-task learning for text classification," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (vol. 1: Long Papers)*, 2017, pp. 1–10.

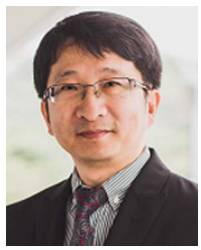
[19] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2019.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.

[20] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, Art. no. 6558.
[21] A. Shenoy, A. Sardana, and N. Graphics, "Multilogue-Net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation," *Proc. 2nd Grand-Challenge Workshop Multimodal Lang.*, pp. 19–28, 2020.
[22] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 01, pp. 164–172.
[23] 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.
[24] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Challenges and applications in multimodal machine learning," *Handbook Multimodal-Multisensor Interfaces: Signal Process. Architectures, Detection Emotion Cogn.-Vol. 2*, 2018, pp. 17–48.
[25] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.
[26] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12695–12705.
[27] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 473–479, 1997.
[28] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Neural Inf. Process. Syst. Workshop Deep Learn.*, 2014, pp. 1–9.
[29] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 5643–5649.
[30] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis*. Springer, 2016, pp. 338–353.
[31] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.
[32] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5651–5661.
[33] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.
[34] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 949–954.
[35] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.
[36] M. Glodek *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, Springer, 2011, pp. 359–368.
[37] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 569–576.
[38] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.
[39] H. Zhou *et al.*, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, 2021, pp. 11106–11115.
[40] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representat.*, 2021, pp. 1–12.
[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 213–229.
[42] Y. Wang *et al.*, "End-to-end video instance segmentation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recognit.*, 2021, pp. 8741–8750.
[43] Y. Rubner, L. J. Guibas, and C. Tomasi, "The earth mover's distance, multi-dimensional scaling, and color-based image retrieval," in *Proc. ARPA Image Understanding Workshop*, vol. 661, 1997, pp. 668–675.
[44] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 2292–2300, 2013.
[45] C. Villani, *Topics in Optimal Transportation*. American Mathematical Soc., 2021, vol. 58.
[46] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, pp. 82–88, 2016.
[47] W. Yu *et al.*, "CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
[48] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 05, pp. 8992–8999.
[49] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python Sci. Conf.*, Citeseer, 2015, vol. 8, pp. 18–25.
[50] N. Hamelin, O. El Moujahid, and P. Thaichon, "Emotion and advertising effectiveness: A novel facial expression analysis approach," *J. Retail. Consum. Services*, vol. 36, pp. 103–111, 2017.
[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[52] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. meeting Assoc. Comput. linguistics (vol. 1: Long papers)*, 2017, pp. 873–883.
[53] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 7216–7223.

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) Laboratory, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore. Her research interests include computer vision and multimodal learning.

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 current research interests include biometrics, forensics, image processing, and pattern recognition.



Alex Kot (Fellow, IEEE) has been with Nanyang Technological University, Singapore, since 1991. He was the Head of the Division of Information Engineering and the Vice Dean Research with the School of Electrical and Electronic Engineering. He was an Associate Dean for the College of Engineering for eight years. He is currently a Professor and the Director of Rapid-Rich Object SEarch (ROSE) Lab and Nanyang Technological University, Singapore, and Peking University, Beijing, China, Joint Research Institute. He has authored or coauthored mainly in the areas of signal processing, biometrics, image forensics and security, and computer vision and machine learning. He 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 coauthor 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.

Authorized licensed use limited to: Nanyang Technological University Library. Downloaded on February 13,2025 at 12:44:54 UTC from IEEE Xplore. Restrictions apply.