Privileged Modality Learning via Multimodal Hallucination

Shicai Wei, Chunbo Luo, *Member, IEEE,* Yang Luo, Jialang Xu

***Abstract*—Learning based on multimodal data has attracted increasing interest recently. While a variety of sensory modalities can be collected for training, not all of them are always available in practical scenarios, which raises the challenge to infer with incomplete modality. This paper presents a general framework termed multimodal hallucination (MMH) to bridge the gap between ideal training scenarios and real-world deployment scenarios with incomplete modality data by transferring the complete multimodal knowledge to the hallucination network with incomplete modality input. Compared with the modality hallucination methods that restore privileged modalities information for late fusion, the proposed framework not only helps to preserve the crucial cross-modal cues but relates the study in complete modalities and in incomplete modalities. Then, we introduce two strategies called region-aware distillation and discrepancy-aware distillation to transfer the response-based and joint-representation-based knowledge of pre-trained multimodal networks, respectively. Region-aware distillation establishes and weights knowledge transferring pipelines between the response of multimodal and hallucination networks at multiple regions, which guides the hallucination network to focus on discriminative regions and avoid wasted gradients. Discrepancy-aware distillation guides the hallucination network to mimic the local inter-sample distance of multimodal representations, which enables the hallucination network to acquire the inter-class discrimination refined by multimodal cues. Extensive experiments on multimodal action recognition and face anti-spoofing demonstrate the proposed multimodal hallucination framework can overcome the problem of incomplete modality input in various scenes and achieve state-of-the-art performance. Code is available at https://github.com/shicaiwei123/TMM-MMH**

***Index Terms*—Privileged modality, incomplete modality, multimodal hallucination, knowledge distillation.**

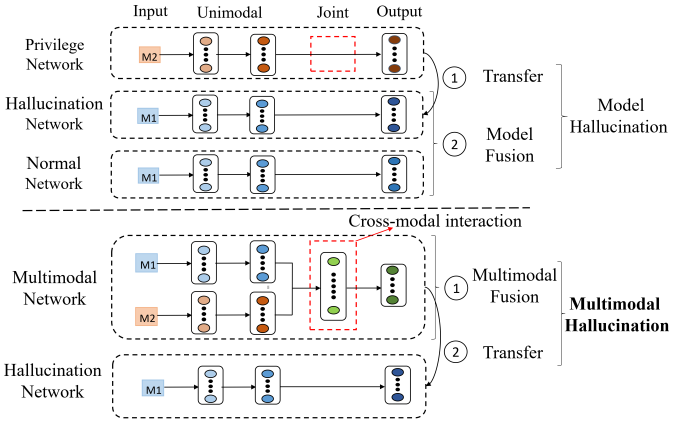


Fig. 1. Comparison of modality hallucination and the proposed multimodal hallucination frameworks. Here $M1$ and $M2$ denote two different inputs of the multimodal networks, respectively, and $M2$ is not available during the inference. Modality hallucination [11] hallucinates the privileged information of $M2$ from the available $M1$ for late fusion to overcome the problem of incomplete modality inference. Multimodal hallucination directly inherits the complete multimodal cues from the pre-trained multimodal model to bridge the gap between ideal training scenarios and real-world deployment scenarios. This not only helps to preserve the cross-modal interactions extracted in the joint representation but also relates the study in complete modalities and in incomplete modalities.

I. Introduction

In recent years, joining the success of deep learning, deep multimodal learning gathers growing attention from the research community and shows great power in practice, such as medical image analysis [1], [2], action recognition [3], [4] and face anti-spoofing [5], [6]. While deep multimodal learning has significant potentials power in improving the robustness and performance of models, it is difficult to meet the setting of multimodal data in practice due to the limitation of devices [7], [8] or user privacy [9], [10]. Therefore, how to bridge the gap between the ideal training scenarios with complete modalities and the real-world deployment scenarios with incomplete modalities is of great significance for multimodal tasks.

Generally, the additional information only available at the training stage is defined as privileged information [12] or side information [13]. Thus, the modality data only available at training time is called the privileged modality [14]. Different from other privileged information, such as the future frames in online action recognition [15] and the user behaviors in recommendation systems [16], the privileged modalities share the same semantics with the available inferring modalities, and their sample or representation can be generated by the available inferring modalities [17]. Therefore, privileged modalities learning methods such as modality imputation [18]–[21] and modality hallucination [11], [14], [22], [23] have been proposed to address the challenge to infer with incomplete modality. The imputation-based methods utilize the generative model to reconstruct the privileged modality sample and then combine it with the inference modality as the input to a common multimodal model. The hallucination-based methods train a hallucination model to reconstruct the privileged modality representation and fuse it with the model trained with the inference modality to make the final decision.

While these methods achieve fairly good performance, they still suffer from some limitations. The generative model of imputation-based methods is trained independently of the subsequent classification task, which may limit the discriminative capacity of the generated modality for the target task [18]. Although the hallucination-based methods can alleviate this issue by fusing the hallucination and normal models at the output layer, they ignore the intermediate cross-modal inter-

actions. Note that the cross-modal cues are the cruciality for multimodal learning to realize better performance than the single modality [17], [24]. More importantly, both of them are focusing on restoring the privileged modality information from the available modalities while ignoring the model trained for the scenarios with complete modalities. This disadvantages the community to build a unified framework to integrate the study with incomplete modalities as well as the study with complete modalities. Note that the study on the scenarios with complete modalities is more comprehensive and advanced than the scenarios with incomplete modalities [17], [23].

In this paper, we propose a novel privilege modality learning framework called multimodal hallucination to overcome the aforementioned challenges. As shown in Fig. 1, the proposed framework guides the hallucination network to learn the complete multimodal information from the pre-trained multimodal network directly. Compared with existing methods that preserve the privileged modality information for late processing, this not only helps to preserve the intermediate cross-modal interaction extracted in multimodal networks but relates the study in complete modalities and in incomplete modalities. However, compared with the conventional modality hallucination methods, the multimodal hallucination also introduces the additional architecture and information gap because the architecture of multimodal and hallucination networks is different, and the hallucination network needs to restore the information of $\mathcal{U}$ modalities with only $\mathcal{V}$ modalities ($\mathcal{V}$ $<\mathcal{U}$). These challenges will increase the difficulty of knowledge transfer [25], [26] and limit the optimization of the primary task at hand, such as recognition or segmentation with incomplete modality data.

To tackle the flaws of the multimodal hallucination framework, we propose two strategies called region-aware distillation and discrepancy-aware distillation, to learn the response-based and joint-representation-based knowledge of multimodal networks, respectively. Specifically, the region-aware distillation establishes and weights knowledge transferring pipelines between the multimodal and hallucination network at multiple regions. This enables the hallucination network to perceive the local patterns that are crucial for multimodal task [27], [27] and guides it to focus on the informative regions, which avoids wasted gradients from unimportant regions and alleviates the learning pressure [28], [29]. Besides, discrepancy-aware distillation encourages the hallucination network to acquire the sample representation that has the same inter-class discrimination as the multimodal network. The prior knowledge behind this is straightforward: it is the inter-class discrimination refined by the cross-modal interactions that make the multimodal model performs better than the unimodal one [30]–[32]. Besides, compared with the previous hallucination distillation methods that match the global value [11], [14], [22] or distribution [23] of the representations, the proposed strategy only requires to satisfy the local distance consistency between two arbitrary samples, which contributes to a simpler optimization objective. In summary, our main contributions are three-fold.

- We present a multimodal hallucination framework to leverage the privileged modality data by transferring the complete multimodal knowledge to the hallucination network with incomplete modality. This not only helps to relate the study with complete and incomplete modalities data but also preserves the crucial cross-modal interaction.
- We propose region-aware distillation to guide the hallucination network to learn response-based knowledge of the multimodal network by establishing the knowledge transferring pipelines between them at multiple regions. This helps to guide the hallucination network to focus on the informative regions and avoid wasted gradients.
- We propose discrepancy-aware distillation to guide the hallucination network to learn the multimodal joint-representation-based knowledge by learning the representation distance between multimodal samples. This helps it to acquire the inter-class discrimination refined by multimodal cues.
- Extensive experiments on two typical multimodal tasks demonstrate the proposed multimodal hallucination framework can overcome the problem of inferring with incomplete multimodal data in various scenes and achieve state-of-the-art performance.

II. Related Work

A. Knowledge Distillation

The knowledge distillation aims to transfer the representation capability of a large model (teacher) to a small one (student) to improve its performance [33]. Generally, transferred knowledge can be divided into three types: response-based knowledge, representation-based knowledge, and relation-based knowledge. Response-based knowledge refers to the neural response of the last output layer of the teacher model and it is usually transferred to the student network by matching the soft output distribution of teacher and student networks [34]. While the idea of the response-based knowledge is straightforward and easy to understand, it only considers the output of the last layer and thus fails to address the intermediate-level supervision from the teacher model [33]. Therefore some researchers propose the representation-based methods to model the knowledge of intermediate feature maps and transfer it to the student network by minimizing the discrepancy between the value [35], attention map [36], [37], and attention project [38] of their feature maps. Because the representation-based knowledge only considers specific layers in the teacher model, researchers further introduce relation-based knowledge to model the relationships between different layers or data samples. This knowledge is transferred from the teacher network to the student network by minimizing the discrepancy between the similarity map [39], distribution [40], inter-data relations [41] of feature pairs from different layers or samples.

B. Privileged Information Learning

The privileged information learning focuses on exploiting auxiliary information that is only available in the training stage to assist the optimization of the model and improve the inference performance [42], which has been used for multiple

tasks. For example, Lee *et al.* take the intermediate representation of reconstructed high-resolution ground truth image to assist the image super-resolution task [43]. Feyereisl *et al.* leverage auxiliary segmentation labels and attributes to improve the performance of object detection [44]. Xu *et al.* take the predicted pedestrian attributes and the semantics-preserving deep embeddings as the privileged information to assist the metric learning for person re-identification [45]. Generally, these methods utilize the privileged information by multi-task learning, i.e, introducing a constraint term for privileged information via an additional task [16]. However, in multi-task learning, each task does not necessarily satisfy the no-harm guarantee (i.e., privileged features can harm the learning of the original model). Besides, from the practical point of view, it may be a challenge to tune all the tasks when using dozens of privileged features at once.

C. Privileged Modality Learning

The privileged modality learning is proposed to exploit auxiliary modalities data that is only available in training. Unlike the privileged learning tasks that utilize the auxiliary information as extra input to assist the model optimization, privileged modalities learning focuses on restoring the privileged information from available inference modalities based on the consistency among multimodal data. For example, Jiang *et al.* utilize the CycleGAN [46] to generate the MRI image from the CT image to aid mediastinal lung tumor segmentation [19]. Pan *et al.* impute the missing PET images based on their corresponding MRI scans using a hybrid generative adversarial network and leverage them to aid the diagnosis of brain disease [18]. Liu *et al.* extend the CycleGAN for multimodal face reconstruction and generate the infrared face image from the corresponding RGB image to assist the face anti-spoofing task [8]. Because the training of GAN-based models is unstable and the generation of the privileged modality is independent of the subsequent classification task [11], [47], the imputation-based methods may limit the discriminative capacity of the generated modality. Thus some researchers propose to restore privileged information via knowledge distillation. Hoffman *et al.* present a modality hallucination architecture for improving the RGB object detection performance by distilling the depth information to the hallucination model and fusing it with the RGB model to make the final prediction [11]. Garcia *et al.* further extend the hallucination architecture for the video action recognition to model the motion flow explicitly to improve its performance [22]. Li *et al.* propose the dynamic-hierarchical attention distillation to hallucinate the SAR image feature from RGB images to aid the land cover classification [48]. In summary, these works aim to train a hallucination model to restore the privileged modalities information by matching the feature maps of hallucination and privileged networks. This may lead to overfitting and limit the performance of the hallucination model since the inputs of the hallucination and privileged networks are different.

D. Incomplete Multi-view Clustering

Incomplete multi-view clustering is proposed to address the challenge of missing some view data in clustering tasks. Generally, they can be categorized into two types: grouping-based methods and imputation-based methods. Grouping-based methods [49], [50] aim to group the data according to the existence of views and then divide them into multiple learning tasks. Imputation-based methods, such as SURE [51], COMPLETER [52], COMIC [53], and OS-LF-IMVC [54], focus on focus on recovering the missing-view data and then leverage them with the in the common multi-view clustering methods. While these methods also focus on addressing the problem of missing data, they are different from the privileged modality learning in the following aspects. First, they are proposed for the clustering task, which is unsupervised. However, to our best knowledge, the privilege modality learning methods are focusing on the supervised task, such as classification [11], [14], [48], detection [55]–[57], and segmentation [58]–[60]. Second, they take the features extracted from the pre-trained model or other handcrafted operators such as HOG, as the input. In contrast, the privilege modality learning methods take the original image as input and learn the features through training.

III. Method

In this section, we first introduce the proposed multimodal hallucination framework that transfers complete multimodal information to the hallucination model with incomplete modality input. This helps to preserve the cross-modal interaction and ensures that the hallucination model benefits from the development of multimodal learning because it learns from the pre-trained multimodal directly. Then, we propose the region-aware distillation and discrepancy-aware distillation strategies to alleviate the optimization drawback caused by input and architecture differences and transfer the response-based and joint-representation-based knowledge to the hallucination model, respectively.

A. The Framework of Multimodal Hallucination

For a task, let Θ denotes the collected training dataset whose sample consists of full modalities $\{M_1, \ldots, M_U\}$. $\mathcal{T}$ denotes the target application scenario that can only access the incomplete data in the dataset Ψ whose sample consists of modalities $\{M_1, \ldots, M_V\}$. And the rest $\{M_{(V+1)}, \ldots, M_U\}$ are the missing modalities.

The overall framework is shown in Fig. 1, the training pipeline of the proposed multimodal hallucination framework consists of three stages. 1) Training a multimodal model M_m with the data in Θ to perform the task with complete multimodal data. 2) Training a hallucination model M_i with the data in Ψ and the guidance of the pre-trained multimodal model M_m whose weights are frozen. 3) Inferring with the hallucination model M_i in target scenario $\mathcal{T}$. The details of our method are introduced in the following sections.

Network architecture: Stage one is exactly the procedure of a general multimodal task and the network can be implemented with any structure proposed for multimodal learning (e.g. CNN, Transformer, or heterogeneous one) [17]. Stage two leverages the knowledge extracted from complete multimodal data to alleviate the information loss caused by incomplete

modality data and improve performance in the target application scenario $\mathcal{T}$. Besides, because the effectiveness of knowledge transfer is highly dependent on structural similarity [**?**], [**?**], we propose to minimize structural differences by removing the module only used by the missing modality. For example, here we remove the feature extraction module of the modality M_2. Finally, to further cope with remaining structural and input differences, we propose propose the joint distribution distillation module to transfer the modality-shared knowledge from multimodal network to hallucination network, The detail will be introduced in Section III-C.

Relation to prior work: As shown in Fig. 1, compared with the *modality hallucination* [11], the multimodal hallucination framework is designed to transfer the complete multimodal cues directly, which not only allows it to preserve the intermediate cross-modal cues but also relates the study in complete modalities and in incomplete modalities. Compared with the *modality imputation* [18], the multimodal hallucination framework focuses on the task at hand and the trained hallucination model is exactly the inference model, which would not limit the discriminative capacity of the inference model.

Relation to general multimodal tasks: The multimodal hallucination framework can be regarded as a novel attempt to push multimodal learning into general scenarios because it bridges the gap between ideal training scenarios with complete multi-modal data and real deployment scenarios that can only access partial modality data.

B. Region-aware Distillation

The existing response-based distillation methods calculate the logit output from the global feature vector of inputs, with the implication that each region of the input image contributes equally to classification. However, the input images may contain regions that are irrelevant to the category information, e.g. background, and it may be sub-optimal to transfer the response-based knowledge by matching global response directly. Especially in the multimodal hallucination task where both the input and network structure of teacher and student networks differ, the gradients from the invalid regions would limit the learning of the hallucination model. To address this issue, we propose the region-aware distillation (Fig. 2) to establish the knowledge transferring pipelines at multiple regions and guide the hallucination network to focus on the informative regions.

We first introduce our notation. Let $x = (x_m, x_i)$ be a paired training sample with the same label y, where x_m and x_i are randomly chosen from the sample set Θ and Ψ, respectively. $\phi_m(x_m) \in R^{w_m \times h_m \times c_m}$ and $\phi_i(x_i) \in R^{w_i \times h_i \times c_i}$ denote the feature maps in the penultimate layer of multimodal network and hallucination network, respectively. Here, h and w are spatial dimensions, c is the number of feature channels. Then, $f_m(j,k)$ and $f_i(j,k)$ denote the feature vector $\phi_m(x_m)(j,k,:) \in R^{1\times1\times c_m}$ and $\phi_i(x_i)(j,k,:) \in R^{1\times1\times c_i}$, respectively. d is the downsampling factor between the input and the final feature map.

According to the analysis in [61], [62], $f_m(j,k)$ and $f_i(j,k)$ can be regarded as the representation of the region

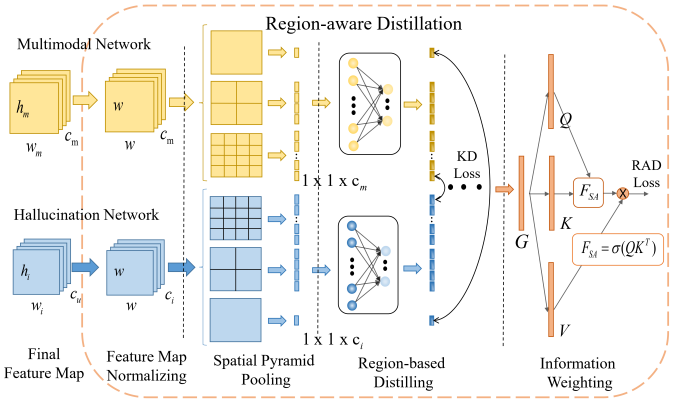


Fig. 2. Illustration of the region-aware distillation. The response-based knowledge of the multimodal network is modeled at different regions by spatial pyramid representation. The knowledge is transferred to the hallucination network by minimizing the weighting sum of the distillation loss between different selected region pairs.

$(tx, ty, tx + d, ty + d)$ in x_m and x_i, respectively, where $tx = d * j$, $ty = d * k$. Therefore, we can guide the hallucination network to learn the region pattern of multimodal networks by measuring KL divergence of the logits outputs of $f_i(j,k)$ and $f_m(j,k)$ as the training loss. Furthermore, we can adjust the downsampling factor and the position of feature vectors to acquire knowledge at multiple regions. Compared with existing methods to segment multiple regions at the input space [27], [63], the proposed strategy calculates the representation for multiple regions via the feature pyramid, reducing the computational burden significantly.

As shown in Fig. 2, the region-aware distillation (RAD) consists of four parts: feature map normalizing, spatial pyramid pooling, region-based distilling, and information weighting. Specifically, we firstly introduce a normalizing pooling to restraint the width and height of $\phi_m(x_m)$ and $\phi_i(x_i)$ to the same size $w \times w$ to align their downsampling factors. It ensures each pixel of them represents the same size region in the input space. Then, we run the spatial pyramid pooling [61] on the normalizing feature maps, which splits the feature map into cells at different scales and performs pooling operations to aggregate the representations in each cell. This helps to acquire the representations that cover different regions in the input space. In contrast to the traditional spatial pyramid pooling, we perform average pooling instead of max pooling to draw the cell representation, which helps preserve the complete information of covered regions. Let $(a(s,n), b(s,n))$ denotes the spatial position of the n_{th} cell at s_{th} scale, $\mathcal{Z}_(s,n)$ denotes the input region corresponding to this cell, $\pi_m(s,n) \in R^{1\times1\times c_m}$ denotes the representation for $\mathcal{Z}_(s,n)$ in x_m, which is the aggregate representation of this cell,

$$\begin{cases} \pi_m(s,n) = \sum_{j=a(s,n)}^{a(s,n)+s-1} \sum_{k=b(s,n)}^{b(s,n)+s-1} \frac{1}{s^2} f_m(j,k) & (1) \\ a(s,n) = s(\lfloor (n-1)/s \rfloor) & (2) \\ b(s,n) = s((n-1) \bmod s) & (3) \end{cases}$$

where $\lfloor . \rfloor$ denotes floor operations. And the paired represen-

tations for the same region $\mathcal{Z}_(s,n)$ in x_i is the $\pi_i(s,n) \in R^{1\times1\times c_i}$,

$$\pi_i(s,n) = \sum_{j=a(s,n)}^{a(s,n)+s-1} \sum_{k=b(s,n)}^{b(s,n)+s-1} \frac{1}{s^2} f_i(j,k) \tag{4}$$

where s and n are the same as those in $\pi_i(s,n)$. For each paired representations, the region-based distillation loss $\mathcal{L}_r(s,n)$ that transfers multimodal feature pattern at region $\mathcal{Z}_(s,n)$ to the hallucination network is defined as follow,

$$\mathcal{L}_r(s,n) = \mathcal{KL}(\sigma(\mathcal{F}_m(\pi_m(s,n)))\| \sigma(\mathcal{F}_i(\pi_i(s,n)))) \tag{5}$$

where $\mathcal{KL}(.,.)$ denotes the KL divergence and $\sigma(.)$ denotes the softmax function, $\mathcal{F}_m(.)$ and $\mathcal{F}_i(.)$ denote the fully-connected layer function of multimodal and hallucination networks, respectively. By traversing all the scales s and their corresponding cells $N_s = (\frac{w}{s})^2$, we can get $N = \sum_s N_s$ loss components in $\mathcal{G} = \{L_r(1,1), L_r(1,2)..., L_r(w,1)\} \in R^{N\times1}$ for N regions. In order to ensure the hallucination networks can always focus on the informative regions containing target cues, we further introduce an information weighting layer to calculate the importance for each component in $\mathcal{G}$ and assign it as the weight to aggregate the distillation loss in $\mathcal{G}$. Here, we implement this via the self-attention mechanism [64] that calculates the importance based the value rather than the position of the loss components, which can adapt to varying input with varying informative regions. Thus, the region-aware distillation loss is defined as follow,

$$\mathcal{L}_{RAD} = \sum_N \sigma(\mathcal{F}_1(\mathcal{G})\mathcal{F}_2(\mathcal{G})^T)\mathcal{F}_3(\mathcal{G}) \tag{6}$$

where $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ means different fully-connect layers.

In order to make the proposed intuition more rigorous and help quantify precisely the way in which vanilla knowledge distillation and region-aware distillation differ, we rewrite the vanilla knowledge distillation loss [34] as follow,

$$\begin{cases} L_{\mathcal{KD}} = \mathcal{KL}(\sigma(\mathcal{F}_m(f_m^g))\| \sigma(\mathcal{F}_i(f_i^g))) & (7) \\ f_m^g = \sum_{j=0}^{w-1}\sum_{k=0}^{w-1} \frac{1}{w^2} f_m(j,k) & (8) \\ f_i^g = \sum_{j=0}^{w-1}\sum_{k=0}^{w-1} \frac{1}{w^2} f_i(j,k) & (9) \end{cases}$$

Specifically, we can derive that $f_m^g = \pi_m(s,n)$ and $f_i^g = \pi_i(s,n)$ when $s = w$, $n = 1$. Thus, the vanilla knowledge distillation loss can be regarded as a term of the region-aware distillation loss, covering the entire image (0,0,$w*d$,$w*d$). It encourages the hallucination network to learn the global response of the multimodal network. Particularly, the region-aware distillation loss also contains other terms calculated in fine-grained levels that cover different local regions when $1 \leq s<w$. Weighted them by their importance can guide the hallucination network to focus on the informative local patterns from the multimodal network. Combined with the label supervision, the total training loss for the hallucination networks to leverage local patterns from multimodal networks to improve its performance is defined as follow,

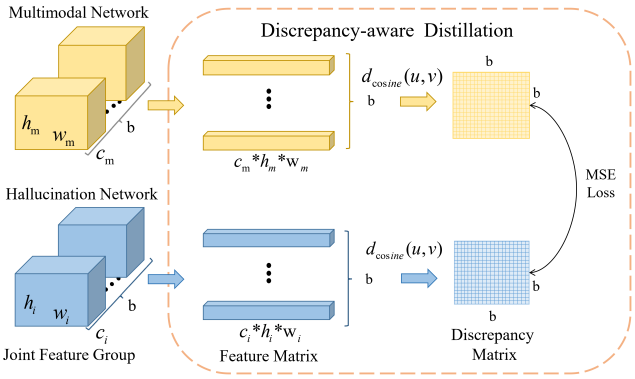


Fig. 3. Illustration of the discrepancy-aware distillation. The joint-representation-based knowledge of the multimodal network is modeled as the local inter-sample distance of multimodal representations. The knowledge is then transferred to the hallucination network by matching the discrepancy matrix between the multimodal network and the hallucination network.

$$\mathcal{L}_1 = \mathcal{L}_{CE} + \alpha\mathcal{L}_{RAD} \tag{10}$$

where $\mathcal{L}_{CE}(.,.)$ denotes the the label supervision loss for the task at hand, α is a balancing factor.

C. Discrepancy-Aware Distillation

The conventional modality hallucination methods transfer the privileged representation knowledge by matching the feature map directly. However, it is not feasible for the hallucination model to generate the same feature map as the multimodal model and this may lead to overfitting and harm the performance of the multimodal hallucination model since both the input and network structure of multimodal and hallucination networks are different. To overcome this issue and preserve the sample discrimination refined by multimodal cues, we present discrepancy-aware distillation (DAD) approach that guides the hallucination networks to learning the joint representation knowledge from the multimodal network by mimicking its inter-sample representation distance.

As shown in Fig. 3, given a paired input mini-batch $X = \{X_m, X_i, Y\}$, we define the joint representation of multimodal networks at l_{th}^m layer as $\phi_m^{l^m}(x_m) \in R^{b\times w_m\times h_m\times c_m}$ and the paired joint representation of hallucination networks as $\phi_i^{l^i}(x_i) \in R^{b\times w_i\times h_i\times ci}$. Here, l^m means the layers of multimodal network where multimodal feature are fused. Similar to modality hallucination [11], if the depths of multimodal and hallucination networks are the same, l^i represents the layer at the same depth as l^m. Otherwise, it represents the layer at the end of the same cell. For each mini-batch, the proposed distillation strategy will expand the feature groups of multimodal and hallucination input into feature matrix $H_m^{l^m} \in R^{b\times w_m\cdot h_m\cdot c_m}$ and $H_i^{l^i} \in R^{b\times w_i\cdot h_i\cdot c_i}$, respectively, in order to calculate the representation distance between different samples. Because the dimension of the feature vectors of the multimodal and the hallucinated network could be very high,

to eliminate the curse of dimensionality, we choose cosine distance as the discrepancy metric,

$$d_{cosine}(\mathbf{u},\mathbf{v}) = \frac{1}{2} - \frac{1}{2}\frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} \tag{11}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two arbitrary feature vectors with same dimension. For each mini-batch, let $D_m^{l^m} \in b \times b$ denotes the discrepancy matrix of the multimodal network at the l_{th}^m layer, whose component at j_{th} row and k_{th} column can be expressed as follow,

$$D_m^{l^m}(j,k) = d_{cosine}((H_m^{l^m})_j, (H_m^{l^m})_k) \tag{12}$$

And $D_i^{l^i} \in b \times b$ denotes the corresponding discrepancy matrix of the hallucination network at the l_{th}^i layer, whose component at j_{th} row and k_{th} column can be expressed as follow,

$$D_i^{l^i}(j,k) = d_{cosine}((H_i^{l^i})_j, (H_i^{l^i})_k) \tag{13}$$

Then, the discrepancy-aware distillation loss is defined as,

$$\mathcal{L}_{\text{DAD}} = \sum_{(l^m,l^i)\in\mathcal{O}} \left\| D_m^{l^m} - D_i^{l^i} \right\|_2 \tag{14}$$

where $\mathcal{O}$ contains the (l^m, l^i) layer pairs (e.g. layers at the same depth, as discussed above). This loss can guide the hallucination network to mimic the representation distance between different multimodal samples and preserve multi-modal complementary clues. Finally, the total training loss guiding the hallucination networks to leverage the inter-class discrimination refined by multimodal networks is defined as follow,

$$\mathcal{L}_2 = \mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{DAD}} \tag{15}$$

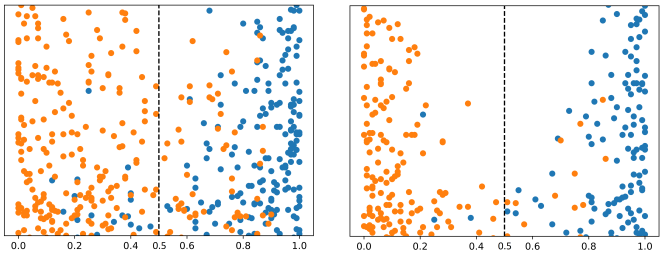
where β is a balancing factor.

IV. Experiments

We conduct experiments on the two-stream architecture task (multimodal action recognition) that fuses multimodal information at the output space and the joint architecture task (multimodal face anti-spoofing) that fuses multimodal information at the feature space. In the following, we first compare the proposed multimodal hallucination framework with the previous state-of-the-arts on the two tasks. Then, we ablate the important design elements of multimodal hallucination. Particularly, all experiments set the general RGB modality as the available modality in the inference like previous work [8], [11], [14], [22], [23], [65]. Still, due to its general design, the proposed framework can also be used to assist other modalities as well.

A. Face Anti-spoofing Performance and Comparison.

Settings: For multimodal face anti-spoofing, we report the result on the CASIA-SURF [66] and CASIA-CeFA [70] datasets with the intra-testing protocol as well as cross-ethnicity and cross-attack protocol suggested by authors, respectively. For a fair comparison, we follow the same benchmark model and data augmentation strategy with CASIA-SURF [66]. The models are trained with the SGD optimizer,



(a) Vanilla (b) DAD

Fig. 4. The prediction distribution of the Vanilla and DAD models with only RGB modality on the test set of CASIA-SURF dataset [66]. the X-axis represents the normalized logit output and x=0.5 is the classification boundary. Orange and blue dots denote two different classes. Compared with the vanilla model trained with incomplete modalities only, the hallucination model trained by the DAD strategy has better semantic separation.

where learning rate is 0.001, momentum is 0.9, batch size is 64, and maximum epoch is 50. The metric used is Average Classification Error Rate (ACER) [66].

Comparison methods: We compare our method with two groups of methods. The first group includes the baseline methods that are trained and tested with the same modality data (marked as 'complete'). Here, we take SURF, the benchmark method of the CASIA-SURF dataset, as the baseline. Another is the state-of-the-art privileged modality learning methods (marked as 'incomplete'). Here, we consider the MARS [22], MERS [22], ADDA [23], MT-Net [8], WCoRD [67], TAKD [68], and CMKD [69]. Specifically, MARS transfers the privilege modality knowledge from the privileged model to the hallucination model by matching their feature maps directly. MERS further fine-tunes MARS on the available modality data. Taking the framework of MERS as the baseline, ADDA proposes to transfer the privileged modality knowledge by matching the feature distribution via adversarial discriminative distillation. WCoRD proposes to match the feature distribution via contrastive learning with the 1-Lipschitz constraint. This can leverage global and local information simultaneously. TAKD introduces an assistant network to match the logit outputs between teacher and student networks. CMKD proposes to match the feature map and logit outputs simultaneously. Besides, MT-Net proposes to reconstruct the privileged modality sample from the available ones by the CycleGAN with the complexion regularizer and subspace-based discriminator.

Finally, for a fair comparison, we follow the implementation of the baseline method for all the privileged learning methods, e.g. we unify their modality encoders as the ResNet18 used in SURF.

Result and analysis: As shown in Table I, the proposed framework implemented with RAD and DAD strategies outperforms all other privileged modality learning methods in the multimodal face anti-spoofing task. This is because the proposed multimodal hallucination framework is designed to learn from the pre-trained multimodal network directly, which can acquire the crucial cross-modal interactions that make the multi-modal model better than the single-modal model. Besides, in both datasets, the DAD strategy designed to learn the complete joint-representation-based knowledge from the

TABLE I
THE PERFORMANCE ON THE MULTIMODAL FACE ANTI-SPOOFING TASK. THE METRIC IS ACER(%) AND THE ↓ MEANS THAT THE LOWER THE VALUE, THE BETTER THE PERFORMANCE.

	Method	Training Modalities	Testing Modalities	CASIA-SURF(↓)	CASIA-CeFA (↓)
Complete	SURF [66]	RGB	RGB	10.32	36.12
	SURF [66]	RGB, D, IR	RGB, D, IR	3.53	28.62
Incomplete	MARS [22]	RGB, D, IR	RGB	9.83	33.50
	MERS [22]	RGB, D, IR	RGB	8.95	33.89
	ADDA [23]	RGB, D, IR	RGB	8.64	33.20
	MT-Net [8]	RGB, D, IR	RGB	8.45	30.60
	WCoRD [67]	RGB, D, IR	RGB	8.08	29.64
	TAKD [68]	RGB, D, IR	RGB	7.78	29.33
	CMKD [69]	RGB, D, IR	RGB	7.45	28.90
	RAD (Ours)	RGB, D, IR	RGB	**7.25**	**28.70**
	DAD (Ours)	RGB, D, IR	RGB	**6.75**	**27.60**
	RAD+DAD (Ours)	RGB, D, IR	RGB	**5.92**	**27.14**

pre-trained multimodal model outperforms the RAD strategy designed to learn the complete response-based knowledge. This is because the joint representation contains more information than the output logits [33]. However, this does not mean the complete response-based knowledge is redundant because the ensemble of the models trained by DAD and RAD strategies, respectively, outperforms the single model trained with DAD strategy by 0.83% in CASIA-SURF and 0.46% in CASIA-CeFA. Particularly, the proposed framework even outperforms the baseline model with complete multimodal information in CASIA-CeFA by 1.48%, which will be discussed in Section IV-E in detail.

B. Action Recognition Performance and Comparison

Settings: For multimodal action recognition, we conduct experiments on four classic datasets with two different modality combinations. Specifically, NW-UCLA [71] and NTUD60 [72] consist of RGB and Depth modalities. HMDB51 [73] and UCF101 [74] consist of RGB and Flow modalities. For the NW-UCLA, we follow the cross-view protocol in the original paper, using two views for training and the rest for testing. For NTUD60, we follow the cross-subject protocol provided in the original paper, using twenty individuals for training and the remaining for testing. Besides, the clip length for NW-UCLA is set as 8 frames and for NTUD60 is 16 frames. For both HMDB51 and UCF101, we conduct experiments using their first split with 16 frames. Besides, we use the TV-L1 method [22] to extract optical flow, with the default parameter setting from OpenCV.

For a fair comparison, we take a basic two-stream architecture that consists of classic I3D [75] as the backbone of multimodal networks. It is trained with the SGD optimizer, where the learning rate is 0.001, momentum is 0.9, batch size is 16, and maximum epoch is 250. Particularly, we use random flipping, and cropping for data augmentation and initialize all models with weights pre-trained on the Kinetics400 dataset [76]. The metric is the proportion of the accurate classification.

Comparison methods: We take the classic I3D [75] model as the baseline. We compare our method with MARS [22], MERS [22], ADDA [23], WCoRD [67], TAKD [68], and CMKD [69] except for the MT-Net, since it is specially designed for multimodal face reconstruction. For a fair comparison, we follow the implementation of the baseline method for all the privileged learning methods, e.g., we set their modality encoders as the I3D-ResNet50.

Result and analysis: Table II shows the results on the NW-UCLA and NTUD60 datasets that consist of RGB and Depth modalities. Specifically, the top part of the table presents the performance of baseline methods. Here we can see that the multimodal method outperforms the unimodal one by 0.9%∼5.7%. The bottom part of the table refers to the methods that leverage the privileged modalities to assist the network that can only access the incomplete modalities. Here, the multimodal hallucination framework implemented with the RAD strategy outperforms all other privileged modality learning methods, including the recent state-of-the-art, TAKD, and CMKD. This demonstrates that learning from the pre-trained multimodal model is still the most effective way to leverage the privileged modality even if it does not explicitly model the cross-modal interaction. This is because the proposed framework focuses on improving the inference performance with incomplete modality data directly. This avoids the sub-optimal results caused by the two-stage procedures that aim to preserve privilege modality information for fine-tuning. Particularly, the proposed framework even outperforms the baseline model with complete multimodal information in NW-UCLA by 0.47%, which will be discussed in Section IV-E in detail. Also, this verifies its scalability to different multimodal tasks (e.g. action recognition and face anti-spoofing). Note that we do not show the result of the multimodal hallucination framework implemented with DAD here, because the two-stream architecture that fuses multimodal information at the output layer does not have the joint representation.

Besides, Table III shows the results on HMDB51 and UCF101 dataset that consist of RGB and Flow modalities. Also, the multimodal hallucination framework implemented with the RAD strategy outperforms all other privileged modality learning methods. Especially, it improves the performance of the RGB baseline by 3.69% and 2.11% in HMDB51 and UCF101 datasets, respectively. This demonstrates the robustness of the multimodal hallucination framework to different modality combinations (e.g. RGB+Depth and RGB+Flow).

TABLE II
THE PERFORMANCE ON THE MULTIMODAL ACTION RECOGNITION TASK WITH NW-UCLA AND NTUD60 DATASETS. THE METRIC IS ACCURACY(%) AND THE ↑ MEANS THAT THE HIGHER THE VALUE, THE BETTER THE PERFORMANCE.

	Method	Training Modalities	Testing Modalities	NW-UCLA(↑)	NTUD60 (↑)
Complete	I3D [75]	RGB	RGB	93.12	81.93
	I3D [75]	RGB, D	RGB,D	**94.03**	**87.41**
Incomplete	MARS [22]	RGB, D	RGB	93.33	82.90
	MERS [22]	RGB, D	RGB	93.45	83.89
	ADDA [23]	RGB, D	RGB	93.54	82.50
	WCoRD [67]	RGB, D	RGB	93.62	83.91
	TAKD [68]	RGB, D	RGB	93.89	84.11
	CMKD [69]	RGB, D	RGB	94.05	84.45
	RAD (Ours)	RGB, D	RGB	**94.50**	**85.64**

TABLE III
THE PERFORMANCE ON THE MULTIMODAL ACTION RECOGNITION TASK WITH HMDB51 AND UCF101 DATASETS. THE METRIC IS ACCURACY(%) AND THE ↑ MEANS THAT THE HIGHER THE VALUE, THE BETTER THE PERFORMANCE.

	Method	Training Modalities	Testing Modalities	HMDB51(↑)	UCF101(↑)
Complete	I3D [75]	RGB	RGB	68.2	90.02
	I3D [75]	RGB, Flow	RGB,Flow	**75.0**	**93.9**
Incomplete	MARS [22]	RGB, Flow	RGB	68.30	90.23
	MERS [22]	RGB, Flow	RGB	68.61	90.41
	ADDA [23]	RGB, Flow	RGB	69.30	91.22
	WCoRD [67]	RGB, Flow	RGB	69.88	91.50
	TAKD [68]	RGB, Flow	RGB	70.02	91.67
	CMKD [69]	RGB, Flow	RGB	70.23	91.73
	RAD (Ours)	RGB, Flow	RGB	**71.89**	**92.13**

TABLE IV
THE PERFORMANCE OF DIFFERENT SYSTEM FRAMEWORK ON THE FACE ANTI-SPOOFING TASK.

Framework	CASIA-SURF	CASIA-CeFA
MH	8.78	30.65
DMCL	7.65	29.70
MMH	**5.92**	**27.14**

TABLE V
THE PERFORMANCE OF DIFFERENT PRE-TRAINED MULTIMODAL MODELS AND MULTIMODAL HALLUCINATION MODELS LEARNING FROM THEM. 'CM(.)' REFERS TO THE DIFFERENT MULTIMODAL MODELS TRAINED WITH COMPLETE MODALITY DATA. 'RAD(.)' AND 'DAD(.)' REFER TO THE HALLUCINATION MODELS THAT LEARN FROM DIFFERENT PRE-TRAINED MULTIMODAL MODELS VIA THE RAD AND DAD STRATEGIES, RESPECTIVELY.

Model	CASIA-SURF	CASIA-CeFA
CM(SURF)	3.53	28.62
CM(PSMM)	2.21 (+1.31)	25.51 (+3.11)
RAD(SURF)	7.25	28.70
RAD(PSMM)	6.85 (+0.40)	26.24 (+2.46)
DAD(SURF)	6.75	27.60
DAD(PSMM)	5.56 (+0.79)	25.12 (+2.48)

C. Ablation Study

In this section, we ablate important design elements in the proposed method. To study both the RAD and DAD strategies, we report the result on the multimodal face anti-spoofing task with the joint representation architecture. The metric used in the ablation is the ACER(↓)

Impact of system framework: To study the effectiveness of the proposed MMH framework, we compare it with the conventional modality hallucination(MH) framework [11] and state-of-the-art modality hallucination framework, DMCL [77] on the face anti-spoofing task. Note that both of them distill knowledge from the privileged model to the hallucination model by matching the feature map. For a fair comparison, their distillation methods are replaced with the proposed RAD and DAD strategies to provide a common basis. The result is shown in Table IV, the DMCL framework introduces the multiple-choice learning to fuse modality adaptively and outperforms the conventional MH framework by 1.13% in CASIA-SURF and 0.95% in CASIA-CeFA, respectively. Furthermore, the proposed MMH framework outperforms the DMCL framework by 1.73% in CASIA-SURF and 2.56% in CASIA-CeFA, respectively. This is because the MMH framework guides the hallucination network to learn from the pre-trained multimodal model directly. Compared with the modality hallucination framework that restores the information of missing modality for late fusion, this helps to preserve the crucial cross-modal interaction for the multimodal tasks.

Impact of pre-trained multimodal model: To study whether our framework can benefit from the development of multimodal learning with complete modality data, we introduce the PSMM model [78] that introduces an advanced multimodal fusion module on the basis of the SURF model. Then we guide the hallucination network to learn from the pre-trained SURF and PSMM model, respectively. As shown in Table V, in both datasets, the PSMM model outperforms the SURF model. Also, in both datasets, the hallucination

TABLE VI
THE ABLATION EXPERIMENTS FOR RAD STRATEGY. WE REPORT THE RESULT OF THE COMPLETE RAD STRATEGY (FULL), THE RAD WITHOUT INFORMATION WEIGHTING (IW), AND THE RAD WITHOUT INFORMATION WEIGHTING AND MULTIPLE KNOWLEDGE TRANSFERRING PIPELINES (MKTP).

Setting	CASIA-SURF	CASIA-CeFA
Full	**7.25**	**28.70**
- IW	7.96	29.52
- IW-MKTP	9.5	31.80

TABLE VII
THE PERFORMANCE OF DIFFERENT PRIVILEGE DISTILLATION STRATEGIES THAT GUIDE THE HALLUCINATION NETWORK TO LEARN THE MULTIMODAL JOINT-REPRESENTATION-BASED KNOWLEDGE.

Methods	CASIA-SURF	CASIA-CeFA
$\mathcal{L}_{CE}$	10.11	36.20
$\mathcal{L}_{CE}$ + $\mathcal{L}_{Hall}$	10.82	37.12
$\mathcal{L}_{CE}$ + $\mathcal{L}_{GAN}$	8.14	30.20
$\mathcal{L}_{CE}$ + $\mathcal{L}_{DAD}(ours)$	**6.75**	**27.60**

networks learning from the pre-trained PSMM model via RAD and DAD strategies outperform that learning from the pre-trained SURF, respectively.

Study of RAD strategy: Here we ablate the multiple knowledge transferring pipelines and information weighting layer in the RAD to evaluate their impacts. Table VI shows the results. The RAD module without information weighting degrades the performance of the complete one by 0.71% in CASIA-SURF and 0.82% in CASIA-CeFA, showing the effectiveness to guide the hallucination network focus on discriminative regions. Then, the RAD module further degrades the performance by 1.54% in CASIA-SURF and 2.28% in CASIA-CeFA when removing the multiple multiple knowledge transferring pipelines at different scales and positions, which demonstrates the effectiveness to enable the hallucination network to perceive local patterns. Here, the information weight module introduces extra 882 parameters due to the fully connected layers for calculating the weighting factor. However, compared to the parameter number of the vanilla model,11172042, the 0.0078% (882/11172042) increase is generally economic.

Effectiveness of DAD strategy: To study the effect of the DAD strategy, we report and compare the performance of different privilege distillation strategies that guide the hallucination network to learn the joint-representation-based knowledge from the pre-trained multimodal model. Here $\mathcal{L}_{CE}$ represents the baseline label supervision without the assist of privilege distillation strategies. $\mathcal{L}_{Hall}$ [11], $\mathcal{L}_{GAN}$ [23], and $\mathcal{L}_{DAD}$ means to guide the hallucination network to learn from the pre-trained multimodal model by matching their feature map, feature distribution, and the discrepancy matrix, respectively.

As shown in Table VII, the model trained by matching the feature map degrades the performance of baseline methods in both datasets. This is because of the huge gap between the multimodal and hallucination networks caused by input and architecture differences, so do their feature maps. Thus matching the feature map directly would introduce extra noise

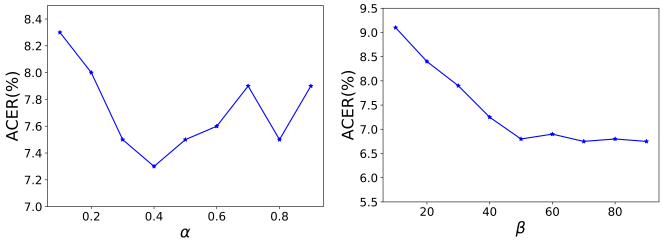


Fig. 5. The performance of the proposed distillation strategies with different parameters α and β on the CASIA-SURF dataset.

to the model training and limits the label supervision task at hand [23]. In contrast, the proposed DAD strategy brings an increase of 3.4% and 8.60% over the baseline method in CASIA-SURF and CASIA-CeFA, respectively, demonstrating it is an effective way to cope with the input and architecture gap and leverage the multimodal joint-representation-based knowledge. Also, the DAD strategy outperforms the existing feature-based privileged modality learning methods $\mathcal{L}_{GAN}$, which shows the superiority to transfer multimodal joint-representation-based by matching the discrepancy matrix. Finally, the prediction distribution in Fig. 4 demonstrates the DAD strategy can guide the hallucination network to inherit the inter-class discrimination refined by complete multimodal cues and acquire a more separable inter-class margin.

D. Hyper-parameter Analysis

As discussed above, the proposed distillation strategies have two hyper-parameters α and β. We tested their impact on the accuracy of the hallucination network on the validation set of the CASIA-SURF dataset. It can be seen from Fig. 5 that the hallucination networks trained with RAD and DAD achieve the lowest ACER(7.25% and 6.75%) on the validation set when α = 0.4 and β = 100, respectively. This is exactly the suggested point to maintain the multiple losses (distillation and cross-entropy loss) in the same order of magnitude [79]. Therefore, the hyper-parameters on other datasets are also determined via this principle. And (α, β) is set as (0.4,100),(0.25,100),(0.25,100) for CASIA-CeFA, NW-UCLA, NTUD60 datasets respectively.

E. Discussion

Why model with incomplete modality data outperforms that with complete modality data: Particularly, the multiple modality data is collected to leverage their complementary information to enhance the discrimination of the target [1], [2], [4], [8], [17]. However, previous studies have shown that multimodal models are not always better than single-modal models [80]–[82] because the black-box multimodal algorithms do not model the cross-modal interaction of multimodal input. Besides, the work in [7], [42] demonstrates that the distillation is an effective way to fuse extra information to the target black-box network. Thus, the proposed multimodal hallucination framework exactly provides the possibility of extracting cross-modal interaction from the pre-trained multimodal model and integrating it into the hallucination network. And the proposed

RAD and DAD strategies that focus on the informative regions and inter-class distance, respectively, are two ways to extract the cross-modal cues from the pre-trained multimodal model. This is because it is cross-modal cues from the multimodal data that refine the target regions and inter-class discrimination in multimodal input and network, respectively.

Limitation and future work: The proposed multimodal hallucination framework guides the hallucination network with incomplete modality input to learn from the model trained with complete modality data directly, which is independent of target tasks and has the potential to handle dense prediction task, such as segmentation [83]. However, the proposed RAD and DAD strategies are designed for the classification task that one image has only one semantics, which may not suitable for the dense prediction task [84]. Therefore, future work will extend the multimodal modality hallucination to dense prediction tasks by introducing new privilege distillation strategies.

V. Conclusion

This paper introduces a general framework called multimodal hallucination to bridge the gap between ideal training scenarios and real deployment scenarios with partial modality data by transferring the complete multimodal knowledge to the hallucination network with incomplete modality input. Then we propose two strategies called region-aware distillation and discrepancy-aware distillation to overcome the gap between multimodal and hallucination networks caused by input and architecture differences and transfer the multimodal response-based and joint-representation-based knowledge to the hallucination network, respectively. Region-aware distillation establishes and weights knowledge transferring pipelines at multiple scales and regions, encouraging the hallucination network to focus on informative regions, reducing the learning complexity. In addition, discrepancy-aware distillation guides the hallucination network to acquire the inter-class discrimination refined by multimodal cues by mimicking the local distance between the representation of two arbitrary samples. Finally, extensive experiments are conducted to verify the effectiveness of our methods.

References

[1] H. Müller and D. Unay, "Retrieval from and understanding of large-scale multi-modal medical datasets: a review," *IEEE transactions on multimedia*, vol. 19, no. 9, pp. 2093–2104, 2017.

[2] E. A. Bernal, X. Yang, Q. Li, J. Kumar, S. Madhvanath, P. Ramesh, and R. Bala, "Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 107–118, 2017.

[3] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, and S.-F. Chang, "Modeling multimodal clues in a hybrid deep learning framework for video classification," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3137–3147, 2018.

[4] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1159–1168.

[5] Z. Li, H. Li, X. Luo, Y. Hu, K.-Y. Lam, and A. C. Kot, "Asymmetric modality translation for face presentation attack detection," *IEEE Transactions on Multimedia*, 2021.

[6] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 919–928.

[7] J. C, D. A. Stroud, C. Ross, J. Sun, R. Deng, and Sukthankar, "D3d: Distilled 3d networks for video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2755–2764.

[8] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

[9] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7985–7993.

[10] K. Fan, W. Jiang, H. Li, and Y. Yang, "Lightweight rfid protocol for medical privacy protection in iot," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1656–1665, 2018.

[11] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.

[12] Z. Wu, X. Xia, R. Wang, J. Li, J. Yu, Y. Mao, and T. Liu, "Lr-svm+: Learning using privileged information with noisy labels," *IEEE Transactions on Multimedia*, 2021.

[13] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.

[14] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 103–118.

[15] P. Zhao, L. Xie, Y. Zhang, Y. Wang, and Q. Tian, "Privileged knowledge distillation for online action detection," 2020.

[16] C. Xu, Q. Li, J. Ge, J. Gao, X. Yang, C. Pei, F. Sun, J. Wu, H. Sun, and W. Ou, "Privileged features distillation at taobao recommendations," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2590–2598.

[17] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[18] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2965–2975, 2020.

[19] J. Jue, H. Jason, T. Neelam, R. Andreas, B. L. Sean, D. O. Joseph, and V. Harini, "Integrating cross-modality hallucinated mri with ct to aid mediastinal lung tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 221–229.

[20] X. Gu, J. Yu, Y. Wong, and M. S. Kankanhalli, "Toward multi-modal conditioned fashion image translation," *IEEE Transactions on Multimedia*, vol. 23, pp. 2361–2371, 2020.

[21] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, 2021.

[22] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "Mars: Motion-augmented rgb stream for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7882–7891.

[23] N. C. Garcia, P. Morerio, and V. Murino, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2581–2593, 2019.

[24] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multimodal learning better than single (provably)," *arXiv preprint arXiv:2106.04538*, 2021.

[25] J. T. Zhou, S. J. Pan, and I. W. Tsang, "A deep learning framework for hybrid heterogeneous transfer learning," *Artificial Intelligence*, vol. 275, pp. 310–328, 2019.

[26] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *Journal of Big Data*, vol. 4, no. 1, pp. 1–42, 2017.

[27] Y. Pan, M. Liu, C. Lian, Y. Xia, and D. Shen, "Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2965–2975, 2020.

[28] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[30] X. Li, C. Wang, J. Tan, X. Zeng, D. Ou, D. Ou, and B. Zheng, "Adversarial multimodal representation learning for click-through rate

prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 827–836.

[31] N. Sengupta, C. B. McNabb, N. Kasabov, and B. R. Russell, "Integrating space, time, and orientation in spiking neural networks: A case study on multimodal brain data modeling," *IEEE Transactions on neural Networks and Learning systems*, vol. 29, no. 11, pp. 5249–5263, 2018.

[32] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1470–1474, 2020.

[33] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[36] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations*, 2017.

[37] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.

[38] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "Alp-kd: Attention-based layer projection for knowledge distillation," *arXiv preprint arXiv:2012.14022*, 2020.

[39] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[40] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 268–284.

[41] S. Lee and B. C. Song, "Graph-based knowledge distillation by multi-head attention network," *arXiv preprint arXiv:1907.02226*, 2019.

[42] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[43] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *ECCV*. Springer, 2020, pp. 465–482.

[44] J. Feyereisl, S. Kwak, J. Son, and B. Han, "Object localization based on structural svm using privileged information," *Advances in Neural Information Processing Systems*, vol. 27, pp. 208–216, 2014.

[45] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 791–805, 2017.

[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[47] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1158–1166.

[48] X. Li, L. Lei, Y. Sun, and G. Kuang, "Dynamic-hierarchical attention distillation with synergetic instance selection for land cover classification using missing heterogeneity images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[49] H. Zhao, H. Liu, and Y. Fu, "Incomplete multi-modal visual data grouping." in *IJCAI*, 2016, pp. 2392–2398.

[50] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, "Partial multi-view clustering via consistent gan," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1290–1295.

[51] M. Yang, Y. Li, P. Hu, J. Bai, J. C. Lv, and X. Peng, "Robust multi-view clustering with incomplete information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[52] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, "Completer: Incomplete multi-view clustering via contrastive prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 174–11 183.

[53] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "Comic: Multi-view clustering without parameter selection," in *International conference on machine learning*. PMLR, 2019, pp. 5092–5101.

[54] Y. Zhang, X. Liu, S. Wang, J. Liu, S. Dai, and E. Zhu, "One-stage incomplete multi-view clustering via late fusion," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2717–2725.

[55] Z. Gao, J. Chung, M. Abdelrazek, S. Leung, W. K. Hau, Z. Xian, H. Zhang, and S. Li, "Privileged modality distillation for vessel border detection in intracoronary imaging," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1524–1534, 2019.

[56] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 166–183.

[57] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[58] J. Jue, H. Jason, T. Neelam, R. Andreas, B. L. Sean, D. O. Joseph, and V. Harini, "Integrating cross-modality hallucinated mri with ct to aid mediastinal lung tumor segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 221–229.

[59] Z. Gu, L. Niu, H. Zhao, and L. Zhang, "Hard pixel mining for depth privileged semantic segmentation," *IEEE Transactions on Multimedia*, vol. 23, pp. 3738–3751, 2020.

[60] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Dada: Depth-aware domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7364–7373.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[62] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[63] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review," *Computer methods and programs in biomedicine*, vol. 187, p. 105242, 2020.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[65] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 166–183.

[66] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 919–928.

[67] L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin, "Wasserstein contrastive representation distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 296–16 305.

[68] H. Liu, Y. Qu, and L. Zhang, "Multispectral scene classification via cross-modal knowledge distillation," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[69] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 87–104.

[70] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.

[71] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.

[72] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[73] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[74] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[75] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[76] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
[77] N. C. Garcia, S. A. Bargal, V. Ablavsky, P. Morerio, V. Murino, and S. Sclaroff, "Distillation multiple choice learning for multimodal action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2755–2764.
[78] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1179–1187.
[79] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 794–803.
[80] A. Jabri, A. Joulin, and L. Van Der Maaten, "Revisiting visual question answering baselines," in *European conference on computer vision*. Springer, 2016, pp. 727–739.
[81] J. Hessel and L. Lee, "Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think!" *arXiv preprint arXiv:2010.06572*, 2020.
[82] I. Gat, I. Schwartz, and A. Schwing, "Perceptual score: What data modalities does your model perceive?" *arXiv preprint arXiv:2110.14375*, 2021.
[83] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.
[84] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7842–7851.

Shicai Wei received the B.S. degree in Communication Engineering from the University of Electronic Science and Technology of China,(UESTC), Chengdu, China in 2019. He is a Ph.D. candidate in Information and Communication Engineering at the same institution. His research interests focus on multimodal learning for the open world, including incomplete, unsupervised, multimodal learning, co-learning, and knowledge transfer.

Chunbo Luo (Member, IEEE) received the Ph.D. degree in high performance cooperative wireless networks from the University of Reading, Reading, U.K., in 2011. His research has been supported by RCUK, Royal Society, EU H2020, NSFC, and industries. His research interests focus on developing model-based and machine learning algorithms to solve networking and engineering problems, with a particular focus on networked unmanned vehicles.,Dr. Luo is a Fellow of the Higher Education Academy.

Yang Luo received the B.S., M.Sc., and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China(UESTC), Chengdu, China, in 2005, 2008, and 2016, respectively. He is currently an Assistant Professor within the same university. His research interests include wireless signal processing, the Internet of Things, and machine learning.

Jialang Xu received his Master's degree in Information and Communication Engineering from the University of Electronic Science and Technology of China (UESTC) in 2022. He is currently pursuing his Ph.D. degree in Medical Physics and Biomedical Engineering from University College London. His research interests include surgical robotics vision and deep learning.