DisCoM-KD: Cross-Modal Knowledge Distillation via Disentanglement Representation and Adversarial Learning

Dino Ienco, Cassio Fraga Dantas

DisCoM-KD: Cross-Modal Knowledge Distillation via Disentanglement Representation and Adversarial Learning

Dino Ienco[1,2]
dino.ienco@inrae.fr
Cassio Fraga Dantas[1,2]
cassio.fraga-dantas@inrae.fr

[1] INRAE, UMR TETIS
500, rue Jean Francois Breton
34090, Montpellier, France

[2] INRIA, University of Montpellier
860, rue Saint Priest
34095, Montpellier, France

Abstract

Cross-modal knowledge distillation (CMKD) refers to the scenario in which a learning framework must handle training and test data that exhibit a modality mismatch, more precisely, training and test data do not cover the same set of data modalities. Traditional approaches for CMKD are based on a teacher/student paradigm where a teacher is trained on multi-modal data with the aim to successively distill knowledge from a multi-modal teacher to a single-modal student. Despite the widespread adoption of such paradigm, recent research has highlighted its inherent limitations in the context of cross-modal knowledge transfer.

Taking a step beyond the teacher/student paradigm, here we introduce a new framework for cross-modal knowledge distillation, named *DisCoM-KD* (Disentanglement-learning based Cross-Modal Knowledge Distillation), that explicitly models different types of per-modality information with the aim to transfer knowledge from multi-modal data to a single-modal classifier. To this end, *DisCoM-KD* effectively combines disentanglement representation learning with adversarial domain adaptation to simultaneously extract, for each modality, domain-invariant, domain-informative and domain-irrelevant features according to a specific downstream task. Unlike the traditional teacher/student paradigm, our framework simultaneously learns all single-modal classifiers, eliminating the need to learn each student model separately as well as the teacher classifier. We evaluated *DisCoM-KD* on three standard multi-modal benchmarks and compared its behaviour with recent SOTA knowledge distillation frameworks. The findings clearly demonstrate the effectiveness of *DisCoM-KD* over competitors considering mismatch scenarios involving both overlapping and non-overlapping modalities. These results offer insights to reconsider the traditional paradigm for distilling information from multi-modal data to single-modal neural networks. Our code is available at this link.

1 Introduction

The modern landscape is characterized by a large diversity of devices consistently sensing their environments. This influx of information poses new challenges in terms of data analysis

and understanding, particularly for machine learning and computer vision models, which must work seamlessly across a spectrum of platforms. From wearable gadgets to autonomous vehicles, an array of sensors continuously collect data about the surroundings [9].

In this context, the same object or entity can be described by multiple modalities, necessitating new learning paradigms to handle the collected heterogeneous information [15]. Many multi-modal learning models assume that data modalities between the training and deployment stages remain exactly the same [15]. However, due to the wide range of sensor data available, systematically accessing data across all sensor modalities may be infeasible.

Specifically, a set of modalities may be available during the training stage, while another set of modalities, either overlapping with the former or not, may be accessible during the deployment stage [18]. In such a scenario, strategies are required to operate under cross-modal scenarios, leveraging the multi-modal information available during the training stage to enhance classification capabilities on the modalities accessible at deployment stage.

Recently, Cross-Modal Knowledge Distillation (CMKD) has proven to be effective for multi-modal applications characterized by modalities mismatch [16]. The majority of existing solutions rely on a teacher/student paradigm [5, 11, 19, 24], where a teacher model trained on one or several data modalities is then used to supervise a single-modal student model trained on the modality available at deployment stage. However, this paradigm has several shortcomings, such as the arbitrary choice of modalities used to train the teacher model, the computational burden associated with the training of multiple models for specific downstream applications, and the need to set up a separate process each time a single-modal student model needs to be trained. Additionally, recent studies in [28] clearly highlights the inherent limitations of this paradigm in automatically distilling useful information such as modality-discriminative features for cross-modal knowledge transfer.

In this study, we aim to address the challenge of multi-modal learning in applications characterized by modalities mismatch. We explore an alternative approach based on disentanglement representation and adversarial learning, overcoming mainstream teacher/student paradigms and their associated limitations. Specifically, we introduce a novel framework for cross-modal knowledge distillation, called *DisCoM-KD* (Disentanglement-learning based Cross-Modal Knowledge Distillation). This framework explicitly models different types of per-modality information to transfer knowledge from multi-modal data to a single-modal classifier. *DisCoM-KD* effectively combines disentanglement representation learning with adversarial domain adaptation to simultaneously extract domain-invariant, domain-informative, and domain-irrelevant features for each modality, tailored to a specific downstream task. Conversely to traditional teacher/student paradigms, our framework learns all single-modal classifiers simultaneously, eliminating the need to train each student model separately. To evaluate the effectiveness of our framework, we consider three standard multi-modal benchmarks and recent state-of-the-art knowledge distillation frameworks, demonstrating *DisCoM-KD*'s ability to outperform previous strategies in scenarios of modalities mismatch, covering both overlapping and non-overlapping modalities between training and deployment stages.

In summary, the contributions of our work are the following:

- A novel CMKD framework based on disentanglement representation and domain adversarial learning;
- An alternative CMKD strategy that circumvents traditional teacher/student paradigms to distill knowledge from multi-modal data to single-modal neural networks;
- An extensive comparison of *DisCoM-KD* across three computer vision multi-modal

benchmarks with state-of-the-art knowledge distillation approaches, highlighting the broad applicability of our framework in the field of multi-modal learning.

The rest of the manuscript is organized as follows: related works are reviewed in Section 2. Section 3 introduces the proposed framework based on feature disentanglement and adversarial learning. The experimental assessment and the related results are reported and discussed in Section 4, while Section 5 draws the conclusions.

2 Related Work

In this section, we firstly provide a brief recap of general knowledge distillation strategies, then we focus on knowledge distillation for multi-modal learning and, finally, we conclude by discussing elements on disentanglement representation-based learning.

Knowledge Distillation. Knowledge Distillation (KD) [7] was introduced to transfer "dark" knowledge from a teacher model to a lightweight student model by learning the student model from the soft labels generated by the teacher. The standard KD loss formulation used to train the student model is defined as follows:

$$L = \alpha L_{task} + (1-\alpha) L_{KD} \tag{1}$$

Here, L_{task} represents the downstream loss and L_{KD} represents the distillation loss enforcing the knowledge transfer from teacher to student, with α determining the balance between the two terms. Different approaches vary in how they implement the L_{KD} term, which can be logit-based [10], feature-based [4], or relation-based [8].

Recent studies have shown that logit-based approaches outperform other strategies [21]. Logit-based approaches implement the L_{KD} component by exploiting the Kullback-Leibler divergence between the teacher and student logits. For example, [31] proposes to explicitly decouple the target class from non-target classes in the knowledge distillation process. In [14], a curriculum learning process is introduced to estimate the temperature value to be used in the Kullback-Leibler divergence. [10] employs a multi-level approach to perform logit distillation at different granularity levels (instance, class, and batch), considering multiple temperature values to make the knowledge transfer more robust. Recently, [21] proposes a plug-in extension that can be combined with any of the previous frameworks in which teacher and student logits are standardized prior to the analysis, ensuring a more coherent and consistent comparison.

Cross/Multi-Modal Knowledge Distillation. Cross-modal KD extends traditional KD approaches to encompass multi-modal learning [28]. While cross-modal KD does not assume any overlap between the modalities accessed by the teacher and student models, in the multi-modal KD scenario [16], the information used by the student is a subset of the modalities used by the teacher model, thus making the former a more general scenario than the latter. However, in both scenarios, the student model typically has access only to a single data modality. The majority of the proposed frameworks for both cross-modal and multi-modal KD [5, 11, 19, 24] are tailored for task specific use cases with lack of a generic solution. This is primarily due to the fact that, despite the empirical success demonstrated by prior works, the mechanisms behind cross-modal KD still remains loosely understood [28]. An initial investigation towards understanding this mechanism has been proposed in [28], where the authors emphasize the significance of modality-discriminative features as key components for cross-modal KD. However, their study provides preliminary experiments that are heavily

reliant on data-specific characteristics, thus limiting the generic value of the obtained findings.

Disentanglement Representation Learning. Disentangled representation learning aims to identify and separate hidden information in the underlying data [25]. In contrast to standard learning processes, which focus solely on learning domain-invariant features, disentanglement-based methods explicitly decompose the learned representation into domain-specific and domain-invariant features, thus paving the way to the extraction of task-relevant and task-irrelevant information [22]. In the context of multi-modal learning, disentanglement-based strategies are used to extract both multi-modal and modality-specific factors [23, 30]. Recent approaches have focused on disentangling shared information among modalities to perform various downstream tasks [17, 27, 29]. Despite its contributions to numerous settings [3, 12, 25], disentangled representation learning still remains unexplored in the realm of knowledge distillation.

3 Method

The proposed architecture, depicted in Figure 1, consists of two independent branches, one for each modality, extracting several per-modality representations. These representations are then used by per-modality task classifiers to make the final prediction. Furthermore, auxiliary classifiers, acting at intermediate stages, are leveraged to ensure that the extracted per-modality representations cover different complementary facets of the underlying information while also carrying information related to the downstream task.

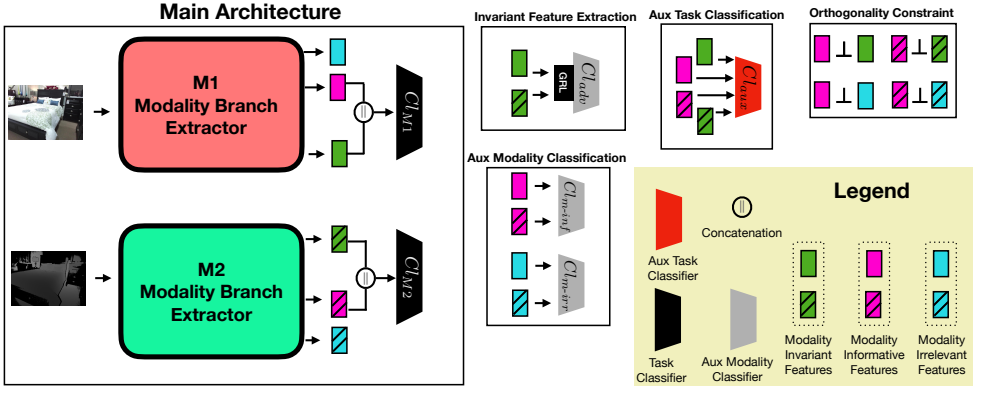

Figure 1: Schematic overview of *DisCoM-KD*: On the left, there are two per-modality branch extractors for modalities $M1$ and $M2$, along with two per-modality task classifiers to obtain the final prediction. On the right, several auxiliary classifiers, acting on intermediate representations, help disentangling per-modality information and make representations task informative. The training of the two parallel architectures is performed jointly, but at inference time, each model is deployed independently.

Given an image x_*, where x_* can be either x_{M1} or x_{M2} (with $M1$ and $M2$ being two different modalities), *DisCoM-KD* extracts three per-modality embeddings $z_*^{inv}, z_*^{inf}, z_*^{irr}$ referred as *modality-invariant*, *modality-informative* and *modality-irrelevant* representation, respectively. All the embeddings have the same dimensionality $z_*^{inv}, z_*^{inf}, z_*^{irr} \in \mathbb{R}^D$. Subsequently,

z_*^{inv} and z_*^{inf} are fed to the (per-modality) main task classifiers, while z_*^{irr} is discarded, as its objective is to collect/attract per-modality information that should not contribute to the downstream task. Specifically, for each of the input modalities, we have a task classifier Cl_* that outputs class probabilities $\hat{y}_* = Cl_*([z_*^{inv}||z_*^{inf}]) \in \mathbb{R}^C$ for the C existing classes and $||$ denotes the concatenation operation. This means that two main task classifiers Cl_{M1} and Cl_{M2} are trained during the process, with $\hat{y}_{M1} = Cl_{M1}([z_{M1}^{inv}||z_{M1}^{inf}])$ and $\hat{y}_{M2} = Cl_{M2}([z_{M2}^{inv}||z_{M2}^{inf}])$.

Beyond the main architecture, we introduce several modules to ensure that the extracted embeddings represent complementary information derived from the multi-modal input data. These additional modules are: i) A modality classifier Cl_{adv} coupled with gradient reversal layer [2] to facilitate the extraction of modality-invariant representations; ii) Two auxiliary modality classifiers, $Cl_{m\text{-}inf}$ and $Cl_{m\text{-}irr}$, ensuring that modality-informative (z_*^{inf}) and modality-irrelevant (z_*^{irr}) embeddings contain modality-specific information and iii) An auxiliary task classifier Cl_{aux} enforcing modality-invariant (z_*^{inv}) and modality-informative (z_*^{inf}) embeddings to be discriminative for the downstream task. During inference, only the per-modality extractors and the main task classifiers (Cl_{M1} and Cl_{M2}) are retained, resulting in two distinct models that have been jointly learnt and can be deployed independently of each other.

3.1 Training losses

To train our cross-modal knowledge distillation framework *DisCoM-KD*, we design a set of loss functions that explicitly model several properties beyond the main downstream classification task with the aim to enforce disentanglement across complementary per-modality representations. Specifically, the training procedure optimizes five different loss terms.

The first term is directly related to the downstream classification task. Both modality invariant (z_*^{inv}) and modality informative (z_*^{inf}) embeddings are fed to the main task classifiers. Then, we use standard Cross-Entropy loss (CE) between the output of each per-modality task classifier and the associated ground-truth y:

$$\mathcal{L}_{cl} = \sum_{m \in \{M1,M2\}} CE(Cl_m([z_m^{inv}||z_m^{inf}]), y) \tag{2}$$

The second term has the goal to enforce the learning of modality-invariant representations. We adopt a CE loss over the output of a classifier that discriminates across representations of samples from different modalities. Here, we use an adversarial training strategy, implemented via Gradient Reversal Layer (GRL) [2] over the modality invariant (z_*^{inv}) representations:

$$\mathcal{L}_{adv} = \sum_{m \in \{M1,M2\}} CE(Cl_{adv}(GRL(z_m^{inv})), m) \tag{3}$$

The third term guides the learning of modality-aware representations via modality classifiers. We use two classifiers, one for modality informative ($Cl_{m\text{-}inf}$) and one for modality irrelevant ($Cl_{m\text{-}irr}$) embeddings in order to predict from which branch the embedding originates:

$$\mathcal{L}_{mod} = \sum_{m \in \{M1,M2\}} CE(Cl_{m\text{-}inf}(z_m^{inf}), m) + \sum_{m \in \{M1,M2\}} CE(Cl_{m\text{-}irr}(z_m^{irr}), m) \tag{4}$$

The fourth term aims to enhance the task discriminative information carried by the per-modality embeddings. Here, we employ an auxiliary task classifier over the set of modality

invariant (z_*^{inv}) and modality informative (z_*^{inf}) embeddings:

$$\mathcal{L}_{aux} = \sum_{m\in\{M1,M2\}} \sum_{i\in\{inv,inf\}} CE(Cl_{aux}(z_m^i), y) \tag{5}$$

The last term explicitly constrains embeddings from the same modality to contain complementary information. We implement a double disentanglement process, enforcing orthogonality [12] between modality-invariant (z_*^{inv}) and informative (z_*^{inf}) representations, as well as between modality-informative (z_*^{inf}) and irrelevant (z_*^{irr}) embeddings. This guides the network to explicitly separate different per-modality contributions:

$$\mathcal{L}_{\perp} = \sum_{m\in\{M1,M2\}} \frac{\langle z_m^{inv}, z_m^{inf}\rangle}{\|z_m^{inv}\|_2 \|z_m^{inf}\|_2} + \sum_{m\in\{M1,M2\}} \frac{\langle z_m^{inf}, z_m^{irr}\rangle}{\|z_m^{inf}\|_2 \|z_m^{irr}\|_2} \tag{6}$$

The final loss function is defined as the sum of all the previous terms:

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{adv} + \mathcal{L}_{mod} + \mathcal{L}_{aux} + \mathcal{L}_{\perp} \tag{7}$$

Implementation details. Each modality branch extractor, as reported in Figure 2, is composed of two encoders: a modality specific encoder and a modality invariant encoder. As encoder backbone we use a ResNet-18 model [8].

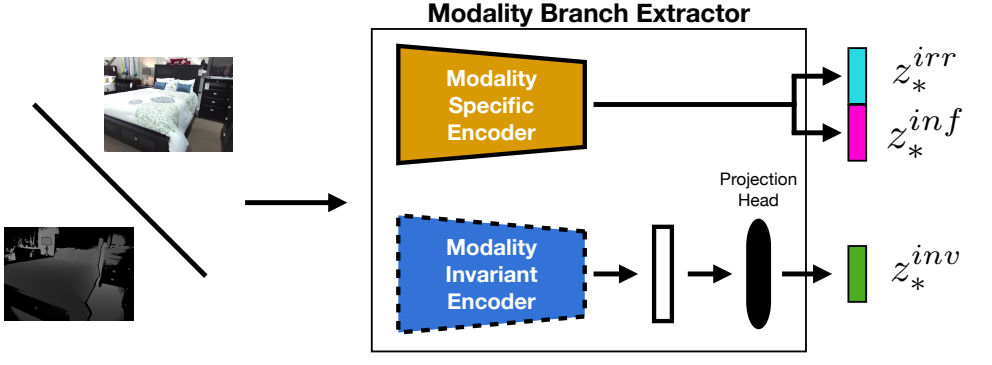

Figure 2: Details of the Modality Branch Extractor. It consists of two encoders, one extracting modality-specific (z_*^{irr}, z_*^{inf}) and one deriving modality-invariant (z_*^{inv}) representations. A projection head is used on the output of the modality-invariant encoder to obtain embeddings of the same size as the other representations.

The modality-specific encoder (shown in gold) extracts modality-specific information, namely: modality-irrelevant z_*^{irr} and modality-informative z_*^{inf} representations encoded separately into each half of the generated embedding vectors (depicted in light blue and purple, respectively). The modality-invariant encoder produces the modality-invariant representation z_*^{inv}. To ensure that all representations have the same size, a projection head, implemented via a fully connected layer, is used to project the output of the modality-invariant encoder to $z_*^{inv} \in \mathbb{R}^D$. For the two per-modality downstream task classifiers (Cl_{M1} and Cl_{M2}) as well as all other auxiliary classifiers (Cl_{adv}, $Cl_{m\text{-}inf}$, $Cl_{m\text{-}irr}$, Cl_{aux}) we use a single linear layer with as many neurons as the number of classes to predict.

4 Experimental Evaluation

To evaluate our framework, *DisCoM-KD*, we designed an experimental evaluation considering three different multi-modal benchmarks involving SOTA teacher/student strategies from the Knowledge Distillation field encompassing both cross-modal and multi-modal KD scenarios. Additionally, an ablation study of *DisCoM-KD* is proposed to analyze the interplay between its different components.

DATASETS. As datasets we consider: i) **SUNRGBD**, the version proposed in [1] for multi-modal RGB-D scene classification. We consider RGB and Depth images from the Kinect v2 domain, for a total of 2 105 pairs of RGB/Depth images, with 3 and 1 channels respectively, covering 10 classes; ii) **EuroSat-MS-SAR** proposed in [26] for multi-modal Multi-Spectral (MS) and Synthetic Aperture Radar (SAR) remote sensing scene classification. The dataset contains 54 000 pairs of MS and SAR images, with 13 and 2 channels respectively, for a land cover classification task spanning 10 classes; iii) **TRISTAR** proposed in [20] for multi-modal (RGB, Thermal and Depth) action recognition. According to results reported in [20], here we only consider the two most informative modalities (Thermal and Depth). The dataset contains 14 201 pairs of Thermal and Depth images, with 1 channel each, representing an action recognition task spanning 6 classes.

COMPETING METHODS. We adopt three recent state-of-the-art strategies: Decoupled Knowledge distillation (*DKD*) [31], Curriculum Temperature Knowledge Distillation (*CTKD*) [14] and Multi-Level Knowledge Distillation (*MLKD*) [10]. Furthermore, we integrate two baseline methods proposed in [28], referred to as *KDv*1 and *KDv*2. Both baselines implement the traditional knowledge distillation loss reported in Equation 1, with *KDv*1 setting the α hyper-parameter to 0, while *KDv*2 sets it to 0.5. While *KDv*1 only uses the soft label to train the student model, *KDv*2 equally weighs the information from the original hard labels and the teacher soft labels. We combine each of these five strategies with the plug-in logit standardization preprocessing (LSKD) proposed in [21]. Additionally, as references, we report the performance of the teacher model (referred to as TEACHER) and a student model that has not received any distillation supervision (referred to as STUDENT) for each evaluation scenario.

EVALUATION SCENARIOS. We adopt two evaluation scenarios: cross-modal KD and multi-modal KD. For the cross-modal KD scenario the teacher is trained on the richest, in terms of downstream task performances, modality and, successively a single-modal student is distilled leveraging the remaining modality. Here the teacher is implemented via a ResNet-18 [6] architecture. For the multi-modal KD scenario the teacher model is trained on the full set of per-dataset modalities and, successively, a single-modal student is distilled. For this scenario, the teacher model is a two-branch architecture with a per modality encoder implemented via a ResNet-18. The fusion is performed at the penultimate layer of the ResNet-18 architecture via feature element-wise addition. Finally, a linear layer exploits the fused representation for the classification decision. All the student models are implemented with a ResNet-18 architecture.

EXPERIMENTAL SETTINGS. For all the approaches the same training setup is used: 300 training epochs, a batch size of 128 and Adam [13] as parameters optimizer with a learning rate of 10^{-4}. For all the approaches we use online data augmentation via geometrical transformations (e.g. flipping and rotation). For the competing methods, we adopt the original hyper-parameter settings. The assessment of the models performance, on the test set, is done considering the weighted F1-Score, subsequently referred simply as F1-Score. Each dataset is divided into training, validation and test set with a proportion of 70%, 10% and 20% of

the original data, respectively. We repeat each experiment five times and report average results. Experiments are carried out on a workstation equipped with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, with 377Gb of RAM and four RTX3090 GPU. All the methods require only one GPU for training.

4.1 Results

Table 1 and Table 2 present the average F1-Score results of the competing methods on both cross-modal and multi-modal KD scenarios, respectively. We use green arrows to indicate when a model outperforms the STUDENT baselines, and red arrows otherwise.
In the cross-modal KD scenario, Table 1, we consider the following cases: RGB $\rightarrow$ DEPTH for SUNRGBD, MS $\rightarrow$ SAR for EuroSat-MS-SAR, and THERMAL $\rightarrow$ DEPTH for TRISTAR. Here, the left modality indicates the one used by the teacher model while the one on the right is leveraged by the student. We observe that *DisCoM-KD* outperforms all competitors on both SUNRGBD (47.69 vs. 42.87 achieved by the best competitor) and EuroSat-MS-SAR (80.03 vs. 78.89 achieved by the best competitor). Moreover, it achieves comparable performance with the best competitor on TRISTAR (92.86 vs. 93.06 achieved by the best competitor). Notably, our framework is the only one that consistently improves (as indicated by green arrows) over the STUDENT baseline across all considered cross-modal scenarios.

In the multi-modal KD scenario, Table 2, *DisCoM-KD* outperforms all state-of-the-art KD approaches, consistently improving classification performance compared to the STUDENT baseline. It is worth noting that our framework is the only one that achieves improvement on the TRISTAR dataset when the THERMAL modality is considered for the deployment stage, achieving a classification score of 97.06. On EuroSat-MS-SAR, all competitors are capable of distilling a student single-modal neural network that outperforms the TEACHER model. Also in this case, *DisCoM-KD* achieves the best classification performance with a score of 98.12. Interestingly, we observe that depending on the dataset, teachers trained on multiple modalities are not always the best choice for distilling a single-modal student. For example, in the TRISTAR case, when the deployment stage covers the DEPTH modality, all KD frameworks exhibited their best performances when the TEACHER has been only trained on the THERMAL modality (cross-modal KD scenario) rather than on the whole set of modalities (multi-modal KD scenario). This underscores that no strategy (neither cross-modal nor multi-modal) guarantees a systematic improvement, highlighting the arbitrary impact this inherent choice can have on the underlying distillation process.

Ablations. The first ablation study (Table 3) explores the importance of the different components on which our framework is built. We observe that Auxiliary Task Classifiers ($\mathcal{L}_{aux}$) and the Disentanglement Loss ($\mathcal{L}_{\perp}$) seem to play the most significant roles in the underlying process. Depending on the considered dataset, each loss term has different relative impacts, and on average, the highest performance is achieved when all components are involved, underscoring the rationale behind the proposed framework. The second ablation study (Table 4) investigates the interplay between the different representations extracted by the disentanglement process. Here, we note that only considering one of the two groups of information —modality-invariant (z_*^{inv}) or modality-informative (z_*^{inf})— systematically decreases the classification performances. Modality-informative features provide slightly better discrimination capability, with varying margins depending on the dataset. In summary, this analysis suggests the suitability of exploiting both modality-invariant and modality-informative representations for the downstream classification task.

		SUNRGBD	EuroSat-MS-SAR	TRISTAR
		RGB → DEPTH	MS → SAR	THER. → DEPTH
TEACHER	-	44.45	95.49	96.99
STUDENT	-	43.82	71.54	90.66
KDv1	ORIG	40.10 (↓)	78.38 (↑)	92.71 (↑)
	w/ LSKD	37.96 (↓)	78.29 (↑)	91.91 (↑)
KDv2	ORIG	41.61 (↓)	78.19 (↑)	92.92 (↑)
	w/ LSKD	42.08 (↓)	78.27 (↑)	91.68 (↑)
DKD	ORIG	42.44 (↓)	78.30 (↑)	92.53 (↑)
	w/ LSKD	41.88 (↓)	78.83 (↑)	92.02 (↑)
CTKD	ORIG	40.09 (↓)	78.89 (↑)	92.46 (↑)
	w/ LSKD	40.76 (↓)	78.89 (↑)	92.36 (↑)
MLKD	ORIG	44.43 (↑)	47.63 (↓)	**93.06** (↑)
	w/ LSKD	42.87 (↓)	78.13 (↑)	91.83 (↑)
DisCoM-KD	-	**47.69** (↑)	**80.03** (↑)	92.86 (↑)

Table 1: Avg. F1-Score performances on cross-modal KD evaluation scenario. We consider the scenarios RGB → DEPTH, MS → SAR and THERMAL → DEPTH for the SUNRGBD, EuroSat-MS-SAR and TRISTAR, respectively. ↑ (resp. ↓) indicates improved (resp. degraded) performances compared to the STUDENT baseline.

		SUNRGBD		EuroSat-MS-SAR		TRISTAR	
		RGB	DEPTH	MS	SAR	THER.	DEPTH
TEACHER	-	55.95		95.36		97.72	
STUDENT	-	44.45	43.82	95.49	71.54	96.99	90.66
KDv1	ORIG	49.88 (↑)	47.46 (↑)	97.92 (↑)	78.69 (↑)	96.82 (↓)	92.47 (↑)
	w/ LSKD	47.44 (↑)	42.90 (↓)	97.37 (↑)	78.45 (↑)	96.60 (↓)	91.45 (↑)
KDv2	ORIG	50.38 (↑)	46.08 (↑)	97.90 (↑)	78.86 (↑)	96.82 (↓)	92.54 (↑)
	w/ LSKD	47.38 (↑)	43.52 (↓)	97.88 (↑)	77.71 (↑)	96.22 (↓)	91.64 (↑)
DKD	ORIG	48.95 (↑)	46.38 (↑)	97.39 (↑)	78.45 (↑)	96.60 (↓)	91.54 (↑)
	w/ LSKD	49.01 (↑)	43.40 (↓)	97.84 (↑)	78.37 (↑)	96.54 (↓)	91.45 (↑)
CTKD	ORIG	48.27 (↑)	44.78 (↑)	97.40 (↑)	79.45 (↑)	91.84 (↓)	91.84 (↑)
	w/ LSKD	48.54 (↑)	43.57 (↓)	97.73 (↑)	79.03 (↑)	96.57 (↓)	91.19 (↑)
MLKD	ORIG	51.48 (↑)	42.57 (↓)	57.90 (↓)	36.17 (↓)	52.39 (↓)	92.27 (↑)
	w/ LSKD	48.92 (↑)	43.82 (↓)	97.78 (↑)	77.64 (↑)	91.64 (↓)	91.44 (↑)
DisCoM-KD	-	**53.63** (↑)	**47.69** (↑)	**98.12** (↑)	**80.03** (↑)	**97.06** (↑)	**92.86** (↑)

Table 2: Avg. F1-Score performances on multi-modal KD evaluation scenario. Here, the TEACHER model has access to all modalities for each dataset. ↑ (resp. ↓) indicates improved (resp. degraded) performances compared to the STUDENT baseline.

	SUNRGBD		EuroSat		TRISTAR		Avg.
	RGB	DEPTH	MS	SAR	THER.	DEPTH	-
w/o $\mathcal{L}_{adv}$	51.38	46.83	98.10	80.63	96.86	92.93	77.78
w/o $\mathcal{L}_{mod}$	53.73	46.86	98.11	80.44	96.93	92.79	78.14
w/o $\mathcal{L}_{aux}$	53.91	41.38	97.82	79.10	96.68	91.56	76.74
w/o $\mathcal{L}_{\perp}$	49.44	42.97	98.03	79.79	97.06	92.62	76.65
DisCoM-KD	53.63	47.69	98.12	80.03	97.06	92.86	**78.23**

Table 3: *DisCoM-KD* components ablation study. Analysis of the contributions of all the components on which our framework is built on in terms of Avg. F1-Score.

	SUNRGBD		EuroSat		TRISTAR		Avg.
	RGB	DEPTH	MS	SAR	THER.	DEPTH	-
Only z_*^{inv}	46.13	39.92	97.73	76.21	96.55	90.79	74.56
Only z_*^{inf}	50.82	43.29	97.60	78.09	96.20	90.68	76.11
DisCoM-KD	53.63	47.69	98.12	80.03	97.06	92.86	**78.23**

Table 4: *DisCoM-KD* modality representations ablation study. Analysis of the contribution of the modality-invariant and -informative representations in terms of Avg. F1-Score.

5 Conclusion

In this study we have introduced a new framework for cross-modal knowledge distillation, namely *DisCoM-KD*. Our aim is to transfer knowledge from multi-modal data to a single-modal classifier. To this end, our framework effectively combines disentanglement representation learning with adversarial domain adaptation. Experimental evaluation, considering both cross-modal and multi-modal knowledge distillation evaluation scenarios, demonstrates the quality of *DisCoM-KD* compared to recent state-of-the-art KD techniques based on the standard teacher/student paradigm. In addition to performance improvements, our framework offers several inherent advantages over the standard paradigm: i) it learns all single-modal classifiers simultaneously, eliminating the need to train each student model separately; ii) it avoids the use of a teacher model, thereby eliminating the need to select which set of data modalities must be used to train the teacher model. Furthermore, our research work introduces an alternative strategy that opens new opportunities beyond the traditional teacher/student paradigm commonly employed for cross-modal and multi-modal knowledge distillation.

Several possible future avenues can be drawn. Our current process has only been assessed on cross-modal distillation tasks involving no more than two modalities. Extending *DisCoM-KD* to manage more than two modalities at once remains an open question. While most of the terms of the proposed loss function can be directly adapted to multiple modalities, how to modify the adversarial term to cope with more than two modalities is not straightforward. Another possible follow-up could investigate how to take inspiration from *DisCoM-KD* to design multi-modal distillation frameworks dealing with semantic segmentation and object detection tasks. For these tasks, the common methodologies are based on encoder/decoder neural network architectures that provide dense predictions as result. All these elements prevent the direct application of our methodology requiring to rethink how disentanglement and adversarial learning may be defined and implemented.

6 Acknowledgment

This work was supported by the French National Research Agency under the grant ANR-23-IAS1-0002 (ANR GEO ReSeT).

References

[1] Andrea Ferreri, Silvia Bucci, and Tatiana Tommasi. Multi-modal RGB-D scene recognition across domains. In *IEEE/CVF International Conference on Computer Vision*

Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, pages 2199–2208. IEEE, 2021.

[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[3] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W. Picard. DISSECT: disentangled simultaneous explanations via concept traversals. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[4] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11868–11877. IEEE, 2023.

[5] Frank M. Hafner, Amran Bhuyian, Julian F. P. Kooij, and Eric Granger. Cross-modal distillation for rgb-depth person re-identification. *Comput. Vis. Image Underst.*, 216: 103352, 2022.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[7] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

[8] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from A stronger teacher. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[9] Summaira Jabeen, Xi Li, Amin Muhammad Shoib, Bourahla Omar, Songyuan Li, and Abdul Jabbar. A review on methods and applications in multimodal deep learning. *ACM Trans. Multim. Comput. Commun. Appl.*, 19(2s):76:1–76:41, 2023.

[10] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24276–24285. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02325. URL https://doi.org/10.1109/CVPR52729.2023.02325.

[11] Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, and Cairong Zhao. Cross-modal distillation for speaker recognition. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence,*

AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 12977–12985. AAAI Press, 2023.

[12] Sang-Yeong Jo and Sung Whan Yoon. POEM: polarization of embeddings for domain-invariant representations. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 8150–8158. AAAI Press, 2023.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 1504–1512. AAAI Press, 2023.

[15] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Arav Agarwal, Yun Cheng, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multizoo and multibench: A standardized toolkit for multimodal deep learning. *J. Mach. Learn. Res.*, 24:234:1–234:7, 2023.

[16] Baolong Liu, Tianyi Zheng, Peng Zheng, Daizong Liu, Xiaoye Qu, Junyu Gao, Jianfeng Dong, and Xun Wang. Lite-mkd: A multi-modal knowledge distillation framework for lightweight few-shot action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 7283–7294. ACM, 2023.

[17] Joanna Materzynska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16389–16398. IEEE, 2022.

[18] Brandon McKinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T. Toshev. Robustness in multimodal learning under train-test modality mismatch. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 24291–24303. PMLR, 2023.

[19] Pritam Sarkar and Ali Etemad. XKD: cross-modal knowledge distillation with domain alignment for video representation learning. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial*

Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, pages 14875–14885. AAAI Press, 2024.

[20] Christian Stippel, Thomas Heitzinger, and Martin Kampel. A trimodal dataset: Rgb, thermal, and depth for human segmentation and temporal action detection. In Ullrich Köthe and Carsten Rother, editors, *Pattern Recognition - 45th DAGM German Conference, DAGM GCPR 2023, Heidelberg, Germany, September 19-22, 2023, Proceedings*, volume 14264 of *Lecture Notes in Computer Science*, pages 18–33. Springer, 2023.

[21] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. *CoRR*, abs/2403.01427, 2024.

[22] Josh Tenenbaum. Building machines that learn and think like people. In Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar, editors, *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, page 5. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018.

[23] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[24] Sijie Wang, Rui She, Qiyu Kang, Xingchao Jian, Kai Zhao, Yang Song, and Wee Peng Tay. Distilvpr: Cross-modal knowledge distillation for visual place recognition. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 10377–10385. AAAI Press, 2024.

[25] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *CoRR*, abs/2211.11695, 2022. doi: 10.48550/ARXIV.2211.11695. URL https://doi.org/10.48550/arXiv.2211.11695.

[26] Yi Wang, Hugo Hernández Hernández, Conrad M. Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *CoRR*, abs/2310.18653, 2023.

[27] Zipeng Xu, Tianwei Lin, Hao Tang, Fu Li, Dongliang He, Nicu Sebe, Radu Timofte, Luc Van Gool, and Errui Ding. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18208–18217. IEEE, 2022.

[28] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[29] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via CLIP. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 3637–3645. ACM, 2022.

[30] Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE Trans. Neural Networks Learn. Syst.*, 34(10):7956–7966, 2023.

[31] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11943–11952. IEEE, 2022.