Revisiting Cross-Modal Knowledge Distillation: A Disentanglement Approach for RGBD Semantic Segmentation

Roger Ferrod[1,4](✉), Cássio F. Dantas[2,3], Luigi Di Caro[1], and Dino Ienco[2,3]

[1] University of Turin, Turin, Italy
(roger.ferrod,luigi.dicaro)@unito.it
[2] INRAE, UMR TETIS, Univ. Montpellier, Montpellier, France
[3] EVERGREEN, Univ. Montpellier, Inria, Montpellier, France
(cassio.fraga-dantas,dino.ienco)@inrae.fr
[4] LIPADE, Univ. Paris Cité, F-75006 Paris, France

Abstract. Multi-modal RGB and Depth (RGBD) data are predominant in many domains such as robotics, autonomous driving and remote sensing. The combination of these multi-modal data enhances environmental perception by providing 3D spatial context, which is absent in standard RGB images. Although RGBD multi-modal data can be available to train computer vision models, accessing all sensor modalities during the inference stage may be infeasible due to sensor failures or resource constraints, leading to a mismatch between data modalities available during training and inference. Traditional Cross-Modal Knowledge Distillation (CMKD) frameworks, developed to address this task, are typically based on a teacher/student paradigm, where a multi-modal teacher distills knowledge into a single-modality student model. However, these approaches face challenges in teacher architecture choices and distillation process selection, thus limiting their adoption in real-world scenarios. To overcome these issues, we introduce CroDiNo-KD (Cross-Modal Disentanglement: a New Outlook on Knowledge Distillation), a novel cross-modal knowledge distillation framework for RGBD semantic segmentation. Our approach simultaneously learns single-modality RGB and Depth models by exploiting disentanglement representation, contrastive learning and decoupled data augmentation with the aim to structure the internal manifolds of neural network models through interaction and collaboration. We evaluated CroDiNo-KD on three RGBD datasets across diverse domains, considering recent CMKD frameworks as competitors. Our findings illustrate the quality of CroDiNo-KD, and they suggest reconsidering the conventional teacher/student paradigm to distill information from multi-modal data to single-modality neural networks. Source code is available here.

Keywords: Knowledge Distillation · Cross-modal · Disentanglement Learning · RGBD · Semantic Segmentation

1 Introduction

Multi-modal information, such as RGB and Depth (RGBD) imagery, is becoming predominant in a plethora of diverse domains including robotics, autonomous driving, augmented reality, healthcare and remote sensing. The combination of these complementary sources of information significantly enhances environmental perception by enriching traditional 2D images with 3D spatial context provided by the Depth modality.

Despite the advantages of multi-modal learning, real-world deployment faces practical challenges. While multi-modal data may be available during training, operational constraints often limit modality availability at inference time due to sensor failures or budget restrictions. This can result in a mismatch between training and testing data, which can impede the practical deployment of an RGBD multi-modal model. To address this challenge, it is essential to design frameworks that are resilient to missing modalities at test time, transferring multi-modal knowledge available during training into single-modality models that operate solely on either RGB or Depth information at inference time. To this purpose, Cross-Modal Knowledge Distillation (CMKD) frameworks have been introduced [1]. Conversely to traditional knowledge distillation techniques, which typically transfers knowledge from a large model to a smaller one using the same input data [2], CMKD enables the transfer of information across modalities. Existing CMKD frameworks typically adopt a teacher/student paradigm, transferring knowledge from a multi-modal teacher to a single-modality student. However, these methods are sensitive to design choices such as teacher architecture, fusion mechanisms and knowledge distillation techniques. Moreover, they require substantial computational resources associated with the training of multiple neural network models: a multi-modal teacher and separate single-modality students, one for each target modality.

With the aim to advance cross-modal knowledge distillation for RGB and Depth imagery, we introduce CroDiNo-KD (Cross-Modal Disentanglement: a New Outlook on Knowledge Distillation), a novel framework that goes beyond conventional teacher/student paradigm, dominant in the CMKD field. Rather than relying on a multi-modal teacher model to guide single-modality RGB or Depth models, CroDiNo-KD relaxes the need for a teacher model through a collaborative training strategy where single-modality models interact with each other via carefully designed loss functions. Our approach removes design decisions related to the teacher architecture and fusion mechanism and teacher/student knowledge distillation techniques. Furthermore, CroDiNo-KD reduces training resources in terms of computational time and parameter size while achieving superior results to recent approaches based on the common teacher/student paradigm.

Specifically, CroDiNo-KD jointly trains two single-modality neural networks using disentangled representation and contrastive learning. This process structures each model's internal manifold into modality-invariant and modality-specific features, capturing both shared and unique information from RGB and Depth modalities. Finally, the training process enables a flexible data augmentation strategy, eliminating the constraints of conventional CMKD framework that require paired augmentation techniques between modalities.

In summary, our contributions are threefold:

(i) We introduce a novel framework for cross-modal knowledge distillation based solely on the joint training of two single-modality models, offering an alternative to the traditional multi-modal teacher/student paradigm;

(ii) We are the first to explore disentanglement representation learning jointly with contrastive learning for RGBD cross-modal knowledge distillation, demonstrating the benefits of structuring internal models manifold into modality-invariant and modality-specific information;

(iii) We provide insights and discussion on the advantages of our framework beyond classification results, analyzing resource efficiency in terms of both computational training time and model size (parameters count).

We validate the effectiveness of CroDiNo-KD on three RGBD benchmarks for semantic segmentation across different application domains, demonstrating superior performance compared to recent state-of-the-art methods especially designed for semantic segmentation under cross-modal knowledge distillation.

2 Related Work

Knowledge Distillation (KD) is the process of transferring information from a large model (teacher) to a smaller one (student). Originally envisioned in [2] for classification tasks with the aim to provide a compact, smaller and faster model, yet performing comparably to the wider teacher model, it has been further refined and formalized by [3], where KD has been commonly implemented via a Kullback-Leibler (KL) divergence between teacher and student predictions. The KD framework can be formulated as follows:

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

where $\mathcal{L}_{task}$ is the task-specific loss and $\mathcal{L}_{KD}$ the KL divergence between student and teacher predictions. By changing the way the KD loss is used, one could distill different kinds of knowledge: response-based [4], feature-based [5] or relation-based [6].

Beyond traditional approaches, KD has also been successfully applied to multimodal learning [1]. Taking inspiration from the the standard KD process, one can distill the knowledge from a multi-modal teacher to single-modality students [7], or from a single-modality teacher to a student working on a different modality [8]. Considering semantic segmentation, cross-modal KD has been proven to be effective over different applications [9]. For example, studies such as [10–16] explored RGBD segmentation with standard KD frameworks, while [17, 18] performed similar experiments on RGBT (RGB+Thermal) dataset. Following works extended the standard cross-modal knowledge distillation approach by adding a generative task [19], via prototype learning [20] or by decomposing the KD loss function into magnitude and angular terms [21].

Differently from standard learning processes, disentanglement representation learning aims to explicitly decompose the feature representation into semantic factors carrying explainable and meaningful information [22]. Leveraged also in multi-modal scenarios (e.g., [23, 24]) it can be used to learn modality-specific and modality-invariant features for the downstream task [25–27]. In particular, in [28] the authors successfully exploited disentanglement —together with adversarial learning— for cross-modal knowledge distillation in the context of scene classification. Inspired by this pioneering work, we further extended this research path onto dense classification, more precisely semantic segmentation.

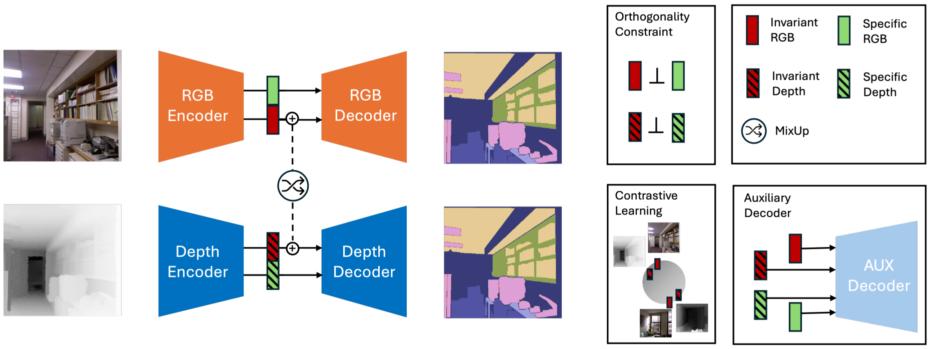

Fig. 1: Overview of the CroDiNo-KD architecture, composed by two encoder-decoder models, for both RGB and Depth modalities. In addition an auxiliary decoder and a set of loss functions are adopted to enforce the desired disentanglement properties between modalities, i.e., modality-invariant and modality-specific features for both RGB and Depth information.

3 Method

With the objective to overcome the limitations of current teacher/student paradigm, here we introduce CroDiNo-KD, a new cross-modal knowledge distillation framework that combines disentanglement representation learning, contrastive learning and decoupled data augmentation. Our approach simultaneously trains two single-modality models – one for RGB and another for Depth imagery – by exploiting modality interaction and collaboration during the training stage.

3.1 Proposed framework

The overall framework, depicted in Figure 1, consists of: i) two separate encoder-decoder models and ii) an auxiliary decoder, all trained with a set of carefully designed loss functions to structure the internal manifold representation of the single-modality models into modality-invariant and modality-specific features.

Given a batch of RGB images X_{RGB} and the corresponding Depth images X_D, with $X_{RGB} \in \mathbb{R}^{B \times H \times W \times 3}$ and $X_D \in \mathbb{R}^{B \times H \times W \times 1}$, we first encode them, via convolutional neural networks, into embedding representations Z_{RGB} and Z_D, respectively. Denoting generically $Z_m \in \mathbb{R}^{B \times h \times w \times F}$ with $m \in \{RGB, D\}$ we have:

$$Z_m = Enc_m(X_m) \tag{2}$$

where B is the batch size, $H \times W$ the spatial dimension of the RGB and Depth images, $h \times w$ the spatial dimension of the embedding representations and F the number of output channels.

For each modality, once the encoded representation Z_m is obtained, we divide it into two separate embeddings Z_m^{inv} and Z_m^{spc}, with $Z_m^{inv}, Z_m^{spc} \in \mathbb{R}^{B \times h \times w \times F/2}$. During training, we then encourage Z_m^{inv} (resp. Z_m^{spc}) to encode modality-invariant (resp. modality-specific) information.

To generate segmentation outputs, the decoder takes as input the concatenated representation $[Z_m^{inv} : Z_m^{spc}]$, where [:] denotes concatenation along the feature dimension. The auxiliary decoder, used only during training, follows a similar architecture but takes only half the channel dimension as input. While the decoder included in the main model relies on $[Z_m^{inv} : Z_m^{spc}]$ as input, the auxiliary one works separately on Z_{RGB}^{inv}, Z_{RGB}^{spc}, Z_D^{inv} and Z_D^{spc} to enforce every individual embedding representation to encode relevant information for the segmentation task.

With the aim to encourage invariant representation across modalities, we introduce a feature mixup strategy [29]. Precisely, we blended the RGB and Depth invariant embeddings, following the equations below, with $\lambda \in [0, 1]$:

$$\begin{aligned} \tilde{Z}_{RGB}^{inv} &= \lambda Z_D^{inv} + (1-\lambda) Z_{RGB}^{inv} \\ \tilde{Z}_D^{inv} &= \lambda Z_{RGB}^{inv} + (1-\lambda) Z_D^{inv} \end{aligned} \tag{3}$$

The augmented images $\tilde{Z}_m^{inv}$ are then processed by the main decoder and contribute to the final loss computation together with the original ones.

Losses: To enhance the performance of single-modality models through mutual interaction and collaboration, we design a set of loss functions that shape the models' internal manifold. The first term is a task-specific segmentation loss, modeled through Cross Entropy. More formally, we have:

$$\mathcal{L}_{seg}^m = CE\left(Dec_m([Z_m^{inv} : Z_m^{spc}]), Y\right) \tag{4}$$

where CE denotes the pixel-wise cross-entropy loss, $Y \in \{1, \ldots, C\}^{B \times H \times W}$ is the ground-truth segmentation map over C classes and Dec_m the decoder for the modality $m \in \{RGB, D\}$.

Then, to explicitly constrain embeddings to encode complementary information (i.e., modality-invariant and modality-specific) we enforced orthogonality between the modality-invariant and modality-specific embeddings of the same modality $m \in \{RGB, D\}$ as follows:

$$\mathcal{L}_{\perp}^m = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{h} \sum_{j=1}^{w} sim(Z_m^{inv}[b, i, j, :], Z_m^{spc}[b, i, j, :]) \tag{5}$$

where $Z_m[b, i, j, :] \in \mathbb{R}^{F/2}$ is the feature vector at spatial location ij in the feature map corresponding to the b-th sample in the batch and $sim(u, v) = \frac{u \cdot v}{||u||_2 ||v||_2}$ denotes the cosine similarity between vectors u and v.

Furthermore, we introduce a contrastive term to bring Z_{RGB}^{inv} and Z_D^{inv} closer together, to force the representation to be invariant with respect to the modality. To this end, we relied on the InfoNCE loss [30] with a negative Euclidean distance, contrasting a positive example (i.e., an RGB-Depth pair of the same instance) with in-batch negatives (i.e., all the remaining invariant embeddings inside the batch, both RGB and Depth). Let $p_m^{(b)} \in \mathbb{R}^{F/2}$ be the L2 normalized feature vector of the b-th instance obtained via spatial average pooling from Z_m^{inv}, that is:

$$p_m^{(b)} = \frac{\rho_m^{(b)}}{\|\rho_m^{(b)}\|_2}, \quad \text{with } \rho_m^{(b)} = \frac{1}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} Z_m^{inv}[b, i, j, :], \tag{6}$$

the contrastive loss is then formulated as:

$$\begin{aligned} \mathcal{L}_{con}^{RGB} &= -\frac{1}{B} \sum_{i=1}^{B} log \frac{exp(-\|p_{RGB}^{(i)} - p_D^{(i)}\|_2/\tau)}{\sum_{m\in\{RGB,D\}} \sum_{j\neq i} exp(-\|p_{RGB}^{(i)} - p_m^{(j)}\|_2/\tau)} \\ \mathcal{L}_{con}^{D} &= -\frac{1}{B} \sum_{i=1}^{B} log \frac{exp(-\|p_D^{(i)} - p_{RGB}^{(i)}\|_2/\tau)}{\sum_{m\in\{RGB,D\}} \sum_{j\neq i} exp(-\|p_D^{(i)} - p_m^{(j)}\|_2/\tau)} \end{aligned} \tag{7}$$

where τ is the temperature parameter. In the first equation, the RGB modality serves as the anchor, while in the second equation, the Depth modality takes this role. In both cases, the numerator represents the positive pair, which corresponds to the embeddings of the same instance across different modalities. The denominator contains the negative samples, comprising all other invariant embeddings from both modalities within the batch, excluding the anchor itself.

To ensure that the embeddings independently encode relevant information for the segmentation task, we added an auxiliary segmentation loss that processes each embedding individually:

$$\mathcal{L}_{aux}^{m} = CE\left(Dec_{Aux}(Z_m^{inv}), Y\right) + CE\left(Dec_{Aux}(Z_m^{spc}), Y\right) \tag{8}$$

with Dec_{Aux} the auxiliary decoder.

Finally, the loss optimized by CroDiNo-KD is an unweighted combination of all the loss terms previously introduced:

$$\mathcal{L}_{tot} = \sum_{m\in\{RGB,D\}} \mathcal{L}_{con}^{m} + \mathcal{L}_{seg}^{m} + \mathcal{L}_{\perp}^{m} + \mathcal{L}_{aux}^{m} \tag{9}$$

Training procedure: The training process, outlined in Algorithm 1, runs over a predefined number of epochs (N_{ep}). For each batch in an epoch, it starts by augmenting the RGB and Depth images, as commonly done in prior works [10,31].However, unlike conventional CMKD frameworks, which enforce paired transformations for both RGB and Depth images, our approach relaxes this constraint by allowing independent per-modality augmentations, a strategy we term as *decoupled augmentation* (lines 3–4). Since RGB and Depth losses are computed separately, this strategy enables greater augmentation flexibility compared to approaches based on the standard teacher/student paradigm, where augmentation consistency across modalities is required.

Next, we extract both domain-invariant and domain-specific embeddings for RGB and Depth images using their respective encoders (lines 5–6). To enhance domain-invariant representations, we leverage feature mixup (line 9), which blends RGB and Depth features, enriching the decoder's training samples (lines 10–12). Additionally, an auxiliary decoder is used to enforce task discrimination for both modality-invariant and modality-specific feature representations, independently (lines 13–15).

To accommodate disentanglement representation learning properties, we use an orthogonality constraint between domain-invariant and domain-specific embeddings (line 16) and we rely on a contrastive loss (line 17) to encourage the RGB and Depth representations of the same instance to be closer to each other while ensuring separation from other instances within the same batch, regardless of the modality. Finally, the total loss is computed as an unweighted sum of all the previously computed losses across RGB and Depth modalities (line 19), back propagating the signal and updating the framework components accordingly.

Algorithm 1: CroDiNo-KD training procedure

```
   input: RGB-Depth labeled dataset $\mathcal{D} = \{(X_{RGB}, X_D, Y)^{(i)}\}_{i=1}^{N}$
1  for epoch $\in \{1, \ldots, N_{ep}\}$ do
2      forall batches $(X_{RGB}, X_D, Y) \in \mathcal{D}$ do
           // Decoupled Augmentations
3          $X_{RGB}$ = Aug($X_{RGB}$);
4          $X_D$ = Aug($X_D$);
           // Encoder
5          $Z_{RGB}^{inv}, Z_{RGB}^{spc} \leftarrow Enc_{RGB}(X_{RGB})$;
6          $Z_D^{inv}, Z_D^{spc} \leftarrow Enc_D(X_D)$;
7          for $m \in \{RGB, D\}$ do
               // Define complementary modality
8              $\overline{m} = D$ if $m = RGB$ else $\overline{m} = RGB$
               // Feature mixup
9              $\tilde{Z}_m^{inv} \leftarrow \lambda Z_{\overline{m}}^{inv} + (1 - \lambda) Z_m^{inv}$;
               // Main Semantic Segmentation task
10             $S_m \leftarrow Dec_m([Z_m^{inv} : Z_m^{spc}])$;
11             $\tilde{S}_m \leftarrow Dec_m([\tilde{Z}_m^{inv} : \tilde{Z}_m^{spc}])$;
12             Compute $\mathcal{L}_{seg}^m$ using $(S_m, \tilde{S}_m, Y)$ with eq. (4);
               // Auxiliary Semantic Segmentation task
13             $A_m^{inv} \leftarrow Dec_{Aux}(Z_m^{inv})$;
14             $A_m^{spc} \leftarrow Dec_{Aux}(Z_m^{spc})$;
15             Compute $\mathcal{L}_{aux}^m$ using $(A_m^{inv}, A_m^{spc}, Y)$ with eq. (8);
               // Disentanglement contrainsts
16             Compute $\mathcal{L}_{\perp}^m$ using $(Z_m^{inv}, Z_m^{spc})$ with eq. (5);
17             Compute $\mathcal{L}_{con}^m$ using $(Z_m^{inv}, Z_{\overline{m}}^{inv})$ with eqs. (6)-(7);
18         end
19         $\mathcal{L}_{tot} = \sum_{m \in \{RGB, D\}} \mathcal{L}_{seg}^m + \mathcal{L}_{\perp}^m + \mathcal{L}_{con}^m + \mathcal{L}_{aux}^m$
20         Update weights of $(Enc_{RGB}, Enc_D, Dec_{RGB}, Dec_D, Dec_{Aux})$ by
             back-propagating $\mathcal{L}_{tot}$
21     end
22 end
23 return $(Enc_{RGB}, Enc_D, Dec_{RGB}, Dec_D)$
```

4 Experiment

To assess the behavior of our framework, CroDiNo-KD, we conducted a comprehensive experimental evaluation using three RGBD benchmarks, comparing our approach against recent competitors in Cross-Modal Knowledge Distillation for semantic segmentation. Furthermore, we performed an ablation study to examine the contributions of individual CroDiNo-KD components and a sensitivity analysis on the hyperparameter λ, which controls the feature mixup strategy across modalities. Finally, we analyze and discuss the computational requirements of competing methods in terms of both training time and model size (parameter counts), emphasizing the advantages provided by CroDiNo-KD over competitors based on the conventional teacher-student paradigm.

Benchmarks: We selected three RGBD semantic segmentation datasets spanning diverse domains to ensure a broad evaluation: indoor scene segmentation, aerial imagery and synthetic drone flight data. Specifically, we considered the following benchmarks:

- **NYU Depth v2** [32]: dataset consisting of 1,449 pairs of indoor RGB and Depth images, labeled with 40 semantic classes. Each image has a resolution of 480×640 and is captured using Microsoft's Kinect. Following prior works [10, 31, 33], we split the dataset into 795 training pairs and 654 test pairs;
- **Potsdam** [34]: a remote sensing dataset comprising 38 scenes of true orthophotos with a ground resolution of 5 cm, annotated with 6 semantic classes. The dataset includes four-channel visual images (R-G-B-IR) and corresponding Digital Surface Models (DSM). For our experiments, we used IR-G-B images and the provided normalized DSMs. Each high-resolution 6,000 $\times$ 6,000 scene was divided into 500×500 crops with stride 1 and further resized to 256×256 due to computational constraints. This resulted in a total of 5,472 images. We followed the same training/test split as described in [35];
- **Mid-Air** [36]: a synthetically generated dataset designed for low-altitude drone flight segmentation, containing 79 minutes of flight data across different weather and seasonal conditions. It includes RGB images and stereo disparity depth maps annotated with 13 semantic classes. Given the large dataset size (over 400k frames) and computational limitations, we selected only a subset of images generated using Unreal Engine's PLE plugin during the spring season. We further subsampled the dataset by selecting one frame every 8 and downscaling the resolution from 1,024 $\times$ 1,024 to 256×256. This resulted in 6,859 images, which were split into training and test sets following the original benchmark.

Competing methods: We compare our approach against several baselines and state-of-the-art CMKD methods for semantic segmentation. Specifically, we evaluate:

- Single-modality, either RGB or Depth, models which do not receive any distillation supervision (referred to as single-modality);
- A full multimodal architecture, corresponding to the teacher model (referred to as *multimodal*);
- Two standard knowledge distillation (KD) baselines [1] (referred to as $KDv1$ and $KDv2$);

- Four state-of-the-art CMKD frameworks, especially tailored for semantic segmentation in multi-modal scenario.

For KD baselines, we adopt the approaches proposed in [1]. These follow a standard KD framework (Equation 1), where α controls the balance between task-specific loss and knowledge distillation. In particular, $KDv1$ sets $\alpha = 0$, meaning the student model learns exclusively from the teacher's soft labels, whereas $KDv2$ uses $\alpha = 0.5$, combining both the original ground-truth labels and the teacher's soft labels equally.

Regarding state-of-the-art CMKD competing frameworks, we consider the following approaches from the recent literature:

- **KD-Net** [9]: originally designed for medical imaging, KD-Net transfers knowledge from a multimodal teacher network to a single-modal student to handle missing modalities. It employs a generalized KD framework [37], utilizing both the teacher's soft labels and bottleneck logits alongside a task-specific loss (binary cross-entropy and Dice loss).
- Masked Generative Distillation (**Masked Dist.**) [19]: introduces a generative distillation task where the student learns to reconstruct a corrupted feature map using the teacher's features as a reference. The final loss consists of a task-specific segmentation loss and a generative distillation term. For the experimental evaluation, we use the encoder's output as feature map to reconstruct.
- **ProtoKD** [20]: combines prototype learning with traditional knowledge distillation and segmentation loss. This method captures semantic correlations across the entire dataset by modeling intra- and inter-class feature variations, transferring similarity maps from the teacher to the student. For the experimental evaluation, we consider the features of the decoder just before the logits computation.
- Layer-wise Angular Distillation (**LAD**) [21] and Channel-wise Angular Distillation (**CAD**) [21]: these methods extend conventional KD approaches by incorporating angular constraints on features. Similar to KD-Net, they perform distillation on both bottleneck features and logits. However, LAD applies layer-wise angular constraints, while CAD operates on channel-wise angular representations.

Experimental Settings: We adopted a consistent training setup across all methods and experiments, with 140 training epochs and batch size of 8. Optimization was performed using the AdamW optimizer with a staring learning rate of 10^{-8} and a learning rate schedule with 10 linear warmup epochs, reaching a target learning rate of 10^{-4}, followed by polynomial decay with power 0.9. Regarding CroDiNo-KD, the temperature τ and the feature mixup λ hyperparameters were set to 0.07 and 0.35, respectively. For data augmentation, we use random flipping, scaling, and cropping for both RGB and Depth images and color jittering only for RGB images, following common practices from previous research [10, 31]. Model performance on the test set has been evaluated using the mean Intersection over Union (mIoU) metric. All experiments were conducted on a single NVIDIA A40 GPU with 48 GB of memory.

Implementation details: To ensure a fair comparison, all the competing approaches share the same architecture, which follows a convolutional encoder-decoder design. For each modality, the encoder is based on a ResNet-50 network with dilated convolutions[5] and initialized with ImageNet-pretrained weights. In the Depth single-modality model, the first-layer weights are initialized by averaging the three-channel pretrained weights into a single-channel representation.

Segmentation outputs are generated using the DeepLabV3+ model [38], which integrates Atrous Spatial Pyramid Pooling (ASPP) and a skip connection linking the second convolutional layer of the ResNet backbone to the decoder.

The teacher model follows the ACNet [10] architecture, a commonly adopted multi-modal semantic segmentation framework for RGBD data. It consists of two ResNet-50 branches dedicated to RGB and Depth modalities, alongside a third ResNet-50 branch for fusing per-modality features. The fusion process is further refined through an Attention Complementary Module (ACM), which applies attention pooling, a 1×1 convolution followed by a sigmoid activation function, as introduced in ACNet. The DeepLabV3+ decoder then processes the fused representation. Figure 2 provides an overview of the teacher model architecture.

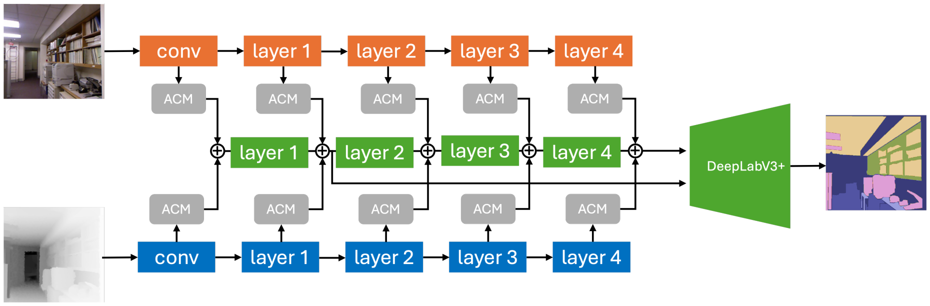

Fig. 2: Teacher model architecture used for the competing methods. It consists of ResNet50 branches for i) RGB ii) Depth and iii) fused representation encoding with an Attention Complementary Module (ACM) as proposed in ACNet [10]. A DeepLabV3+ decoder is added to generate semantic segmentation predictions.

4.1 Results

We present in Table 1 a comparison between the performances achieved by CroDiNo-KD and the competing methods described in the previous section, in terms of mIoU score. Specifically, we include the multi-modal teacher, single-modality models, standard KD baselines (Equation 1) and state-of-the-art competitors. We highlight models that outperform the single-modality baseline with a green arrow and those that underperform the same baseline with a red arrow. To ensure a comprehensive evaluation, we

[5] The final pooling operation is removed and replaced with a $conv_{1\times1}$ projection, followed by batch normalization and a ReLU activation, reducing the feature dimensionality from 2048 to 1024.

assess each benchmark in two cross-modal distillation scenarios, transferring knowledge from multi-modal RGBD data to either the RGB or Depth modality.

We observe that CroDiNo-KD consistently outperforms all competitors across all benchmarks in both RGB and Depth cross-modal distillation scenarios. For the NYUDepth and Mid-Air datasets, the single-modality models are outperformed by the multimodal teacher. However, in the Postdam benchmark, both CroDiNo-KD and Masked Dist. produce RGB-based models that surpass the multimodal approach, achieving mIoU scores of 76.13 and 76.09, respectively, compared to the 74.98 mIoU achieved by the multimodal teacher. Notably, our framework stands out as the only one that consistently demonstrates improvements (green arrows) over the single-modality baselines across all cross-modal scenarios, delivering results that surpass the state-of-the-art methods in Cross-Modal Knowledge Distillation for the considered RGBD benchmarks.

Model	**NYUDepth**		**Potsdam**		**Mid-Air**	
	RGB	Depth	RGB	Depth	RGB	Depth
multimodal	46.92		74.98		51.21	
single-modality	42.64	36.01	75.73	42.47	47.84	47.07
KDv1 [39]	43.43 (↑)	36.44 (↑)	66.32 (↓)	39.20 (↓)	47.36 (↓)	45.80 (↓)
KDv2 [39]	43.86 (↑)	36.91 (↑)	66.24 (↓)	39.38 (↓)	47.62 (↓)	45.88 (↓)
KD-Net [9]	42.78 (↑)	36.36 (↑)	73.82 (↓)	41.85 (↓)	48.32 (↑)	46.22 (↓)
Masked Dist. [19]	40.97 (↓)	34.93 (↓)	76.09 (↑)	42.43 (↓)	47.60 (↓)	47.40 (↑)
ProtoKD [20]	43.82 (↑)	37.28 (↑)	66.64 (↓)	39.27 (↓)	47.11 (↓)	45.45 (↓)
LAD [21]	43.62 (↑)	36.86 (↑)	66.80 (↓)	39.31 (↓)	48.01 (↑)	46.98 (↓)
CAD [21]	43.48 (↑)	37.16 (↑)	66.43 (↓)	38.89 (↓)	48.21 (↑)	47.09 (↑)
CroDiNo-KD	**44.85** (↑)	**37.60** (↑)	**76.13** (↑)	**42.78** (↑)	**48.37** (↑)	**47.91** (↑)

Table 1: Mean Intersection over Union (mIoU) performances over the three considered benchmarks, comparing our model with the multi-modal teacher and single-modality models, as well as state-of-the-art competitors for CMKD semantic segmentation; green and red arrows indicate, respectively, improvement or reduction of scores with respect to the single-modality model.

Ablation Table 2 presents the results of our ablation study, examining the contribution of individual components and loss terms in CroDiNo-KD. Our analysis reveals that the most significant performance drops occur when removing the auxiliary loss ($\mathcal{L}aux$) and the contrastive loss ($\mathcal{L}con$), indicating their crucial role in the framework. The impact of other components and loss terms remains comparable, with variations depending on the dataset. Overall, the highest performance is consistently achieved when all components are included, highlighting the rationale behind CroDiNo-KD.

Sensitivity Analysis We explored the impact of varying the mixup hyperparameter λ (Equation 3) from 0.05 to 0.5, adjusting the degree of feature mixup between domain-invariant RGB and Depth features. As shown in Table 3, performance remains relatively

	NYUDepth		**Potsdam**		**Mid-Air**		**Avg**
	RGB	Depth	RGB	Depth	RGB	Depth	
w/o $\mathcal{L}_{\perp}$	43.98	37.24	75.88	42.48	48.15	47.58	49.22
w/o $\mathcal{L}_{con}$	43.10	**37.96**	75.53	42.31	48.46	47.26	49.11
w/o $\mathcal{L}_{aux}$	44.84	37.62	75.55	**42.99**	47.08	46.44	49.09
w/o *mixup*	44.82	37.48	75.52	42.36	48.19	47.47	49.31
w/o *dec. aug.*	43.62	37.49	75.92	42.37	48.31	47.30	49.17
Original	**44.85**	37.60	**76.13**	42.78	**48.37**	**47.91**	**49.61**

Table 2: Analysis of the contributions of all the components of CroDiNo-KD in terms of mIoU.

stable across this range, with no significant variation as highlighted by the standard deviation.

Segmentation examples Some qualitative segmentation examples on the Potsdam dataset are presented in Figure 3. Here, we compare our method with best performing competitors (KD-Net and Masked Dist.) and the single-modality baseline. The analysis focuses on the RGB modality, as it provides greater visual detail. The results clearly show that all the CMKD frameworks provide a more precise and reliable segmentation mask compared to the one produced by the single-modality baseline. Among the different approaches, we can observe that the quality of segmentation examples is consistent with the quantitative results we have reported above.

λ	**NYUDepth**		**Potsdam**		**Mid-Air**	
	RGB	Depth	RGB	Depth	RGB	Depth
0.05	44.55	37.66	75.90	42.66	48.16	46.82
0.1	44.67	37.72	76.09	42.71	47.95	46.96
0.2	44.64	37.82	76.06	42.59	48.17	47.04
0.35	44.85	37.60	76.13	42.78	48.37	47.91
0.5	44.53	37.50	75.99	42.39	48.13	47.41
std	0.13	0.12	0.09	0.15	0.15	0.44

Table 3: Sensitivity analysis on the feature mixup hyperparameter λ.

To further inspect the behavior of our model, in Figure 4 we depict the output of each per-modality branch, separately, on a few samples coming from the MidAir dataset. It could be noted that the input features may provide complementary information for the segmentation task, for example, in the second row the road is perfectly detected via the RGB sensor, while in the third row the Depth map provides useful information given the lack of visibility, due to fog, on the RGB image.

Training time and model size To further emphasize the advantages of CroDiNo-KD, we compare models performance in terms of total training time and model size (param-

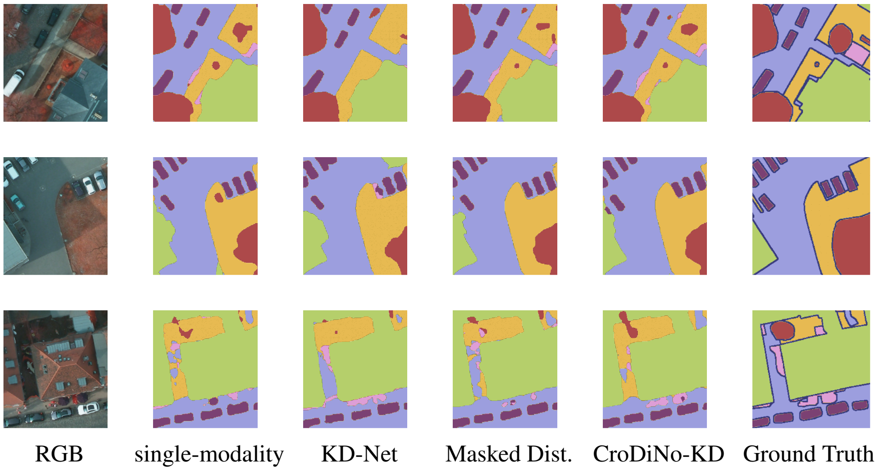

Fig. 3: Example of qualitative results from Potsdam dataset.

eters count). Table 4 presents the complete training time for all competing methods on the MidAir dataset for training both RGB and Depth single-modality models. We report the training time[6] for the distillation process (referred as **Main**), the one for the teacher training (referred as **Teacher**) and the total one (referred as **Tot.**) CroDiNo-KD exhibits the shortest training time for the distillation process, completing both RGB and Depth single-modality models training in less than twenty-one hours. Furthermore, unlike our approach, all CMKD methods require pre-training a teacher model, adding an extra 14 hours overhead to the total training time. Such analysis clearly demonstrates the advantage, in terms of training time, of CroDiNo-KD over standard teacher/student CMKD frameworks.

Table 5 compares the number of parameters required by competing frameworks during training. We categorize the parameters into: those required for the main architecture (**Main**), those used for auxiliary tasks which are discarded during inference such as the generative decoder in the *Masked Dist.* model (**Aux**), the parameters of the teacher model (**Teacher**) and the total per framework parameters (**Tot.**). CroDiNo-KD has fewer parameters in its main architecture compared to competing models, due to the practical choice to reduce the encoder's extracted features to accommodate the disentanglement representation process. The auxiliary parameters in CroDiNo-KD, associated with the auxiliary decoder, remain negligible compared to the overall model size. Furthermore, by eliminating the need for a computationally demanding multi-modal teacher, our approach requires less than half the parameters of the second smallest CMKD framework, thus highlighting the parameter-efficient design of CroDiNo-KD.

[6] Training times are reported in GPU hours, meaning the equivalent training duration without parallelization.

Model	**GPU hours**		
	Main	**Teacher**	**Tot.**
Single-Modality	14h 52m	-	14h 52m
KDv1 / KDv2	22h	14h	36h
KD-Net	22h 46m	14h	36h 46m
Masked Dist.	47h 22m	14h	61h 22m
ProtoKD	25h 02m	14h	39h 02m
LAD	22h 26m	14h	36h 26m
CAD	38h 24m	14h	52h 24m
CroDiNo-KD	20h 30m	-	20h 30m

Table 4: Training time in GPU hours.

Model	**Num. of Params.**			
	Main	**Aux**	**Teacher**	**Tot.**
Single-Modality	80M	-	-	80M
KDv1 / KDv2	80M	-	98M	178M
KD-Net	80M	-	98M	178M
Masked Dist.	80M	150M	98M	328M
ProtoKD	80M	-	98M	178M
LAD/CAD	80M	-	98M	178M
CroDiNo-KD	68M	5M	-	73M

Table 5: Models' size in terms of parameters counts at training time.

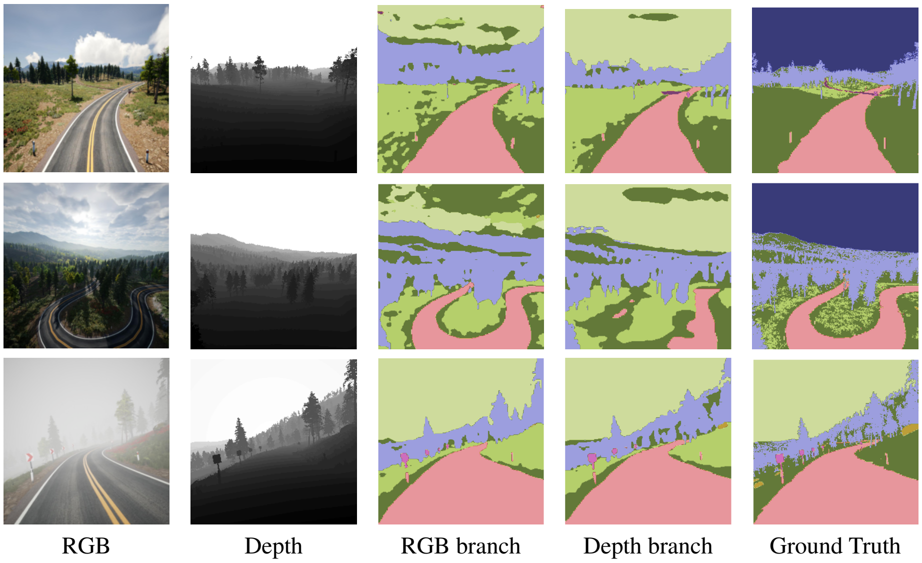

Fig. 4: Example of qualitative results from CroDiNo-KD predictions over the MidAir dataset.

5 Conclusion

In this paper, we propose CroDiNo-KD, a novel framework for RGBD Cross-Modal Knowledge Distillation (CMKD). Unlike conventional teacher/student approaches, our framework facilitates knowledge transfer between single-modality models without requiring a multi-modal teacher. This is achieved by leveraging disentanglement representation learning, contrastive learning and decoupled data augmentation. Through carefully designed loss functions, our method structures the internal manifolds of the single-modality models to account for both modality-invariant and modality-specific features. This approach harnesses the synergy between RGB and Depth modalities to enhance semantic segmentation performance in scenarios where mismatches exist between the data modalities accessible during training and inference. Our evaluation demonstrates the quality of CroDiNo-KD over baselines and state-of-the-art CMKD frameworks, considering both classification performance and computational efficiency during training. Furthermore, our findings invite reconsidering the traditional teacher/student paradigm for distilling information from multi-modal data to single-modality neural networks in the context of semantic segmentation.

References

1. Z. Xue, Z. Gao, S. Ren, and H. Zhao, "The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation," in *ICLR*, 2022.
2. C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, 2006.
3. G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, vol. abs/1503.02531, 2015.
4. Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," *CVPR*, pp. 24276–24285, 2023.
5. Z. Guo, H. Yan, H. Li, and X.L. Lin, "Class attention transfer based knowledge distillation," *CVPR*, pp. 11868–11877, 2023.
6. T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *ArXiv*, vol. abs/2205.10536, 2022.
7. B. Liu, T. Zheng, P. Zheng, D. Liu, X. Qu, J. Gao, J. Dong, and X. Wang, "Lite-mkd: A multi-modal knowledge distillation framework for lightweight few-shot action recognition," *ACM Multimedia*, 2023.
8. F. M. Hafner, A. H. Bhuyian, J. F. P. Kooij, and E. Granger, "Cross-modal distillation for rgb-depth person re-identification," *Comput. Vis. Image Underst.*, vol. 216, pp. 103352, 2018.
9. M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori, "Knowledge distillation from multi-modal to mono-modal segmentation networks," *ArXiv*, vol. abs/2106.09564, 2020.
10. X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," *ICIP*, pp. 1440–1444, 2019.
11. X. Xu, L. Kong, H. Shuai, and Q. Liu, "Frnet: Frustum-range networks for scalable lidar segmentation," *ArXiv*, vol. abs/2312.04484, 2023.
12. C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2016.
13. C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv*, 2013.
14. J. Yang, L. Bai, Y. Sun, C. Tian, M. Mao, and G. Wang, "Pixel difference convolutional network for rgb-d semantic segmentation," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 34, pp. 1481–1492, 2023.
15. S. Lee, S.J. Park, and K. S. Hong, "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation," *ICCV*, pp. 4990–4999, 2017.
16. J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," *ArXiv*, vol. abs/1806.01054, 2018.
17. Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6348–6360, 2024.
18. Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Trans. on Autom. Sci. and Eng.*, vol. 18, no. 3, pp. 1000–1011, 2021.
19. Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan, "Masked generative distillation," in *ECCV*, 2022.

20. S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, and R. Li, “Prototype knowledge distillation for medical segmentation with missing modality,” in *ICASSP*, 2023.
21. T. Liu, C. Chen, X. Yang, and W. Tan, “Rethinking knowledge distillation with raw features for semantic segmentation,” *WACV*, pp. 1144–1153, 2024.
22. Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al., “Disentangled representation learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
23. Y.H. H. Tsai, P. Pu Liang, A. Zadeh, L.P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” *ArXiv*, vol. abs/1806.06176, 2018.
24. Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, “Learning disentangled representation for multimodal cross-domain sentiment analysis,” *IEEE Trans. on Neural Net. and Learning Sys.*, vol. 34, no. 10, pp. 7956–7966, 2023.
25. Z. Xu, T. Lin, H. Tang, F. Li, D. He, N. Sebe, R. Timofte, L. Van Gool, and E. Ding, “Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pretrained vision-language model,” *CVPR*, pp. 18208–18217, 2021.
26. Y. Yu, F. Zhan, R. Wu, J. Zhang, S. Lu, M. Cui, X. Xie, X.-S. Hua, and C. Miao, “Towards counterfactual image manipulation via clip,” in *ACM Multimedia*, 2022.
27. J. Materzyńska, A. Torralba, and D. Bau, “Disentangling visual and written concepts in clip,” in *CVPR*, 2022.
28. D. Ienco and C.F. Dantas, “Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning,” in *BMVC*, 2024.
29. H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *ArXiv*, vol. abs/1710.09412, 2017.
30. A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *ArXiv*, vol. abs/1807.03748, 2018.
31. Z. Wan, Y. Wang, S. Yong, P. Zhang, S. Stepputtis, K. P. Sycara, and Y. Xie, “Sigma: Siamese mamba network for multi-modal semantic segmentation,” *ArXiv*, vol. abs/2404.04256, 2024.
32. P. Kohli N. Silberman, D. Hoiem and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012.
33. R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, “Omnivore: A single model for many visual modalities,” *CVPR*, pp. 16081–16091, 2022.
34. ISPRS, “Urban modelling and semantic labelling benchmark,” 2024, `https://www.isprs.org/education/benchmarks/UrbanSemLab/default.aspx`.
35. N. Kieu, K. Nguyen, S. Sridharan, and C. Fookes, “General-purpose multimodal transformer meets remote sensing semantic segmentation,” *ArXiv*, vol. abs/2307.03388, 2023.
36. M. Fonder and M. Van Droogenbroeck, “Mid-air: A multi-modal dataset for extremely low altitude drone flights,” in *CVPRW*, June 2019.
37. D. Lopez-Paz, L. Bottou, B. Scholkopf, and V. Naumovich Vapnik, “Unifying distillation and privileged information,” *ArXiv*, vol. abs/1511.03643, 2015.
38. L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018.
39. Z. Xue, Z. Gao, S. Ren, and H. Zhao, “The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation,” in *ICLR*, 2023.