UMSS : Towards Unsupervised Multi-modal Semantic Segmentation

Haitian Zhang[1], Thai Duy Nguyen[1], Xiangyuan Wang[2], Mohan Liu[1], and Lin Wang[1]

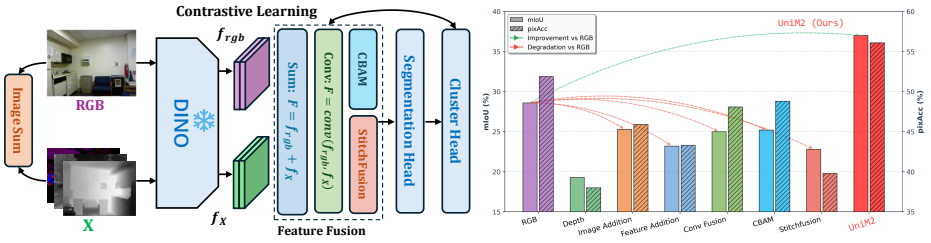
[1] Nanyang Technological University, Singapore
{haitian003,nguyendu003}@e.ntu.edu.sg
{mohan.liu,linwang}@ntu.edu.sg
[2] The University of Hong Kong, Hong Kong SAR, China
xiangyuan.wang@connect.hku.hk
https://github.com/Hatins/UniM2

Abstract. Multi-modal semantic segmentation (MSS) is essential for robust perception in complex environments, yet its potential remains largely untapped due to the prohibitive cost of human annotations. While unsupervised semantic segmentation (USS) has seen success on single RGB modality, **its naive extension to multi-modal data is hampered by fusion degradation**. This is because, in the absence of explicit supervision, existing frameworks struggle to reconcile the heterogeneous structural patterns captured by different sensors, failing to effectively exploit their complementary information. In this paper, we make the **first** attempt to address the novel problem of **Unsupervised Multi-modal Semantic Segmentation (UMSS)**, aiming to effectively exploit complementary sensor information in a fully label-free setting. To this end, we propose **UniM2** (**Uni**fied **M**ulti-**M**odal), a novel framework built upon DINOv3 that transforms conventional fusion methods into consistent performance gains. **Our key idea** is to learn a unified latent space driven by **Cross-modal Correspondence Synergy (CMCS)** to extract intrinsic shared semantic cues, bypassing the need for label-guided adaptive fusion. To mitigate inherent inter-modal conflicts, we introduce a **Cross-modal Harmonizer (CMH)** that designates RGB as a stable reference, effectively suppressing inconsistent relational supervision while guiding the model to exploit complementary structural features. Extensive experimental results on NYU-Depth-v2 and MFNet show that **UniM2** improves mIoU by **6.4%** and **9.8%**, respectively, demonstrating clear advantages over existing frameworks in UMSS task.

Keywords: Unsupervised Learning · Segmentation · Modal Fusion

1 Introduction

Multi-modal semantic segmentation (MSS) [23, 31, 79, 84] is crucial for robust perception in various safety-critical applications, including autonomous driving [5, 75, 82], robotic navigation [46, 52], and embodied intelligence [11, 13, 37].



(a) Directly Apply Conventional Fusion in UMSS **(b)** Performance Comparison

Fig. 1: Analysis of multi-modal integration in unsupervised semantic segmentation. (a) We explore various fusion schemes, including naive **Image Addition**, **Feature Addition**, and **Conv Fusion**, as well as SOTA fusion methods in MSS such as **CBAM** [67] and **StitchFusion** [31]. (b) Quantitative results on NYU-Depth-v2 [49] demonstrate that existing advanced fusion strategies in multi-modal segmentation inevitably lead to performance degradation compared to the single RGB baseline in the unsupervised setting, while our **UniM2** achieves significant mIoU gains.

By integrating complementary signals [19, 49, 69, 78] such as depth [49] or thermal/infrared [19, 35], MSS enhances perception in challenging environments where RGB-only perception often fail [55, 59]. Despite its importance, the progress of MSS is largely driven by massive human-annotated datasets [9, 34]. These pixel-level labels are not only prohibitively expensive to produce but also constrain the learning process to a limited set of predefined semantic categories [1, 10, 51], creating a barrier to utilizing the vast amounts of uncurated multi-modal data in label-free settings [10, 17, 44].

To bridge the gap between label-free learning and multi-modal perception, we define the task of **Unsupervised Multi-modal Semantic Segmentation (UMSS)**. A straightforward way to address this task is to extend SOTA multi-modal fusion strategies [31, 79] to the USS framework [20, 27, 29, 42, 45] as illustrated in Fig. 1 (a). Although USS has achieved remarkable success on the single RGB modality propelled by the development of self-supervised Vision Transformers like the DINO family [4, 40, 50], simply extending these methods to multi-modal settings often results in performance degradation instead of the expected gains as shown in Fig. 1 (b). This phenomenon occurs because *existing frameworks struggle to reconcile the heterogeneous structural patterns and representation biases captured by different sensors without explicit guidance.* While ground-truth annotations in supervised learning implicitly arbitrate these inter-modal inconsistencies, such **Conflicting Signals** cannot be properly resolved in an unsupervised setting. The lack of label-driven arbitration disturbs optimization and leads to a disorganized latent space with degraded clustering quality, which prevents the effective use of complementary information.

To address these challenges, we propose **UniM2**, a unified framework designed to learn a shared latent space driven by **Cross-modal Correspondence Synergy (CMCS)** as intrinsic supervision (Sec. 3.2). Instead of enforcing rigid feature alignment, CMCS promotes structural consistency across modalities by encouraging agreement in cross-modal correspondences, enabling the discovery of shared semantic manifolds while preserving complementary cues. To further

mitigate inter-modal conflicts arising from heterogeneous sensing mechanisms, we designate RGB as the primary semantic reference and introduce a **Cross-modal Harmonizer (CMH)** (Sec. 3.3). The CMH adaptively regulates alignment strength, suppressing unreliable relational supervision while retaining informative auxiliary signals, thereby preventing negative transfer during fusion.

We evaluate **UniM2** on three representative multi-modal benchmarks using the latest **DINOv3** [50] architecture. Specifically, we investigate the "R+X" setting on bi-modal datasets, *i.e.*, **NYU-Depth-v2** [49] and **MFNet** [19], and further validate the scalability of our framework on the quad-modal **MCubeS** [32] dataset. Experimental results show that conventional supervised fusion strategies fail to fully exploit the **Inherent Complementarity** of heterogeneous modalities in the absence of labels, resulting in severe performance degradation. In contrast, **UniM2** achieves absolute mIoU gains of **6.4%** on NYU-Depth-v2 and **9.8%** on MFNet over the RGB-only baseline, consistently transforming cross-modal interference into synergistic improvements. Furthermore, the modular design of CMH facilitates its extension to N auxiliary modalities, yielding incremental gains on MCubeS. Our contributions are summarized as follows:

- **Task Definition.** We introduce the task of Unsupervised Multi-modal Semantic Segmentation, investigating how to leverage heterogeneous modalities for semantic segmentation without any human annotations.
- **UniM2 Framework.** We propose UniM2, learning a unified latent space via Cross-modal Correspondence Synergy while utilizing a Cross-modal Harmonizer to regulate alignment strength and mitigate modal conflicts.
- **Performance and Scalability.** Built upon DINOv3, UniM2 converts the performance degradation typical of conventional fusion into substantial mIoU gains, while naturally scaling to multiple auxiliary modalities.

2 Related Work

Multi-modal Semantic Segmentation. Multi-modal semantic segmentation [31, 79, 84] leverages complementary data from heterogeneous sensors to overcome the inherent limitations of single-modal RGB perception, particularly in visually degraded environments [47, 77, 85]. Existing methods typically design modality-aware fusion mechanisms [79] to integrate heterogeneous features, including hierarchical feature aggregation, cross-modal attention [67], and adaptive gating strategies [84]. These approaches rely on dense pixel-level annotations to learn effective modality alignment, resolve cross-modal inconsistencies, and suppress conflicting predictions during training. While such supervised paradigms have demonstrated strong performance gains, their reliance on explicit semantic labels prevents direct extension to the UMSS settings, where no annotation is available to guide modality interaction. *In fact, directly embedding supervised multi-modal fusion modules into existing USS frameworks not only fails to bring improvements, but often leads to substantial performance degradation.*

Unsupervised Semantic Segmentation. Unsupervised semantic segmentation [20, 27] aims to partition images into semantically meaningful regions without human-provided labels. Early studies [3, 6, 22] focused on discovering recurring

visual patterns using hand-crafted features or low-level spatial priors such as pixel consistency and spatial continuity. However, these approaches were limited in capturing complex semantic variations due to insufficient high-level representation capacity. The development of self-supervised Vision Transformers (ViTs), particularly the DINO family [4, 40, 50], significantly advanced USS by providing semantically structured dense representations through large-scale pre-training. Building upon these representations, STEGO [20] introduced a distillation-based framework that converts dense feature correlations into discrete semantic maps via contrastive learning [7, 26]. Subsequent works [27, 29, 42, 45] further refined this paradigm by improving clustering objectives and exploiting local structural priors to enhance boundary quality. *Despite these advances, existing USS methods are limited to RGB-only input and do not explicitly consider heterogeneous sensors* [2, 32, 35, 49, 79]. Consequently, the integration of complementary multi-modal signals in unsupervised settings remains largely unexplored.

Representation Decoupling in Multi-modal Learning. Multi-modal learning [15, 43] aims to isolate modality-specific private information from shared commonalities to establish a robust semantic space [41, 76]. In supervised scenarios, this decoupling is inherently **label-driven** [63, 65], as task-specific annotations explicitly guide the model to identify beneficial features while suppressing noise [36, 74]. *Conversely, in unsupervised settings, decoupling becomes* ***highly unstable****; the lack of guidance often leads to optimization confusion, where structural contradictions between heterogeneous sensors degrade the shared manifold.* To address this instability, our **Cross-modal Harmonizer** provides **structured decoupling** by adaptively regulating alignment strength. Unlike naive fusion mechanisms, CMH suppress unreliable relational noise while concurrently harvesting complementary cues, ensuring that modality-specific nuances are leveraged without compromising the integrity of the latent semantic structure.

Cross-modal/domain Adaptation via Distillation. Cross-modal knowledge distillation (CMKD) [21, 58, 60] and Unsupervised Domain Adaptation (UDA) [12, 25, 66] share conceptual overlap with UMSS in leveraging multiple data sources. Typically, CMKD aims to transfer complementary information from an auxiliary modality to a primary one by guiding a student to imitate teacher signals [62]. This paradigm has been widely applied in cross-sensor perception such as event, LiDAR, and thermal transfer [30, 54, 61, 64] to resolve spatial or illumination ambiguities. Similarly, UDA [24, 48] focuses on bridging different data distributions through feature alignment or adversarial learning.

By contrast, UMSS fundamentally differs from both CMKD and UDA in two key aspects. First, the **supervision paradigm**. Unlike UDA which relies on labeled source domains or CMKD which depends on pre-trained "teacher" models [62], UniM2 leverages **Cross-modal Correspondence Synergy** as an intrinsic, label-free supervisory signal. Second, the **learning objective**. While CMKD and UDA often focus on a "teacher-student" hierarchy [73] or source-to-target alignment to boost a primary modality, UniM2 treats heterogeneous signals as joint contributors to discover a shared semantic manifold.

Positioning of Our Work: UniM2 addresses the absence of labels by establishing **correspondence synergy** as an **intrinsic supervisory signal**. While previous methods rely on one-way imitation to import external guidance, UniM2 leverages a unified latent space as a mediator to enable **mutual supervision** between heterogeneous modalities, resolving structural contradictions and stabilizing the shared manifold without supervision.

3 Methodology

3.1 Preliminaries and Task Definition

Preliminaries. In the USS task, given an unlabeled image corpus $\mathcal{I} = \{I_i\}_{i=1}^N$, the objective is to learn a mapping function that assigns each pixel to one of the K latent semantic clusters without any human supervision. State-of-the-art frameworks [20, 27, 45] typically tackle this by distilling high-dimensional features from a frozen self-supervised backbone $\mathcal{F}$ [50] into a compact, low-rank embedding space via a lightweight segmentation head $\mathcal{S}$ [20]. This semantic-preserving dimensionality reduction yields low-dimensional manifolds that are more amenable to clustering [28], as they mitigate the curse of dimensionality while amplifying latent semantic correspondences.

The core of this paradigm is a novel correspondence distillation loss [14, 16, 72] that operates on three types of paired inputs: *Self*, *KNN* [18, 80], and *Random* pairs [20]. For any given pair of images (I_1, I_2), the framework extracts their dense feature maps $\{f_1, f_2\}$ from the frozen backbone $\mathcal{F}$, and generates the corresponding segmentation embeddings $\{s_1, s_2\}$ via the segmentation head $\mathcal{S}$ [20]. The underlying assumption is that the semantic correlation between f_1 and f_2 should be preserved and amplified in the embedding space of s_1 and s_2 [20]. Formally, the feature correspondence F and segmentation correspondence S are defined as pixel-wise cosine similarity [8, 70]:

$$F_{hwij} = \frac{f_{1,hw}^\top f_{2,ij}}{\|f_{1,hw}\|\|f_{2,ij}\|}, \quad S_{hwij} = \frac{s_{1,hw}^\top s_{2,ij}}{\|s_{1,hw}\|\|s_{2,ij}\|}, \tag{1}$$

where (h, w) and (i, j) denote spatial indices. For *Self* pairs, the image subscripts are omitted as the correspondence is computed within the same image ($I_1 = I_2$). The distillation process is then driven by the objective:

$$\mathcal{L} = -\sum_{hwij} (F_{hwij} - b) \odot \max(S_{hwij}, 0), \tag{2}$$

where b is a scalar bias providing negative pressure to force weakly correlated features toward orthogonality. Its value is specifically conditioned on the pair type (*Self*, *KNN*, or *Random*) to prevent representation collapse and encourage the formation of compact clusters.

Definition of UMSS. Building upon the USS, we formally define the UMSS task. Unlike USS, which operates on a single image corpus, UMSS leverages a

paired dataset $\mathcal{D} = \{(I_i, \{X_i^{(m)}\}_{m=1}^{M})\}_{i=1}^{N}$, where each RGB image I_i is aligned with a set of M auxiliary modalities. The objective is to learn a joint representation space that captures consistent semantic categories across these heterogeneous inputs without any human annotations. Specifically, the framework aims to optimize a multi-modal mapping $\Phi(f_{rgb}, \{f_X^{(m)}\}_{m=1}^{M}) \rightarrow s$, where f_{rgb} and $f_X^{(m)}$ denote dense features extracted from the RGB and the m-th auxiliary modality, respectively. However, as demonstrated by our preliminary experiments in Fig. 1, directly adopting existing multi-modal fusion schemes in this context proves counterproductive and even degrades performance. This failure highlights the inherent difficulty of aligning heterogeneous features without proper regularization. To address this, we propose **UniM2**, a framework designed to effectively harness multi-source information and resolve modality conflicts.

3.2 The Proposed UniM2 Framework

The overall architecture of UniM2 is illustrated in Fig. 2 (b). **Taking the RGB-Depth pair as a representative case**, our framework differentiates itself from conventional USS by processing dual-modality inputs for each sample, denoted as $\{I, X\}$. As shown in the framework, both the RGB image I and the auxiliary modality X are first fed into a frozen self-supervised backbone $\mathcal{F}$ to extract their respective dense feature maps, f_{rgb} and f_X. These features are subsequently processed by Modality-Specific Networks (MSN) for refinement before being integrated through a Conv Fusion module to generate the final fused representation f_{fus}. Finally, f_{fus} is projected into a compact embedding space via $\mathcal{S}$ and subsequently clustered to yield the final semantic assignments.

Training Inputs. To optimize the framework, we adopt and extend the sampling strategy of [20] to a multi-modal context as shown in Fig. 2 (a). In the original unimodal USS setting, a training pair consists of two images (I_1, I_2), resulting in two backbone feature maps $\{f_1, f_2\}$ which are then mapped to two segmentation embeddings $\{s_1, s_2\}$ via a segmentation head S. In contrast, our UMSS approach operates on multi-modal groups, where a training pair comprises a source group $\mathcal{G}_1 = \{I_1, X_1\}$ and a target group $\mathcal{G}_2 = \{I_2, X_2\}$. Depending on the relationship between $\mathcal{G}_1$ and $\mathcal{G}_2$, the pair forms a *Self*, *KNN*, or *Random* correspondence. Consequently, each training pair in UniM2 generates four distinct backbone feature maps, $\{f_{rgb,1}, f_{X,1}, f_{rgb,2}, f_{X,2}\}$, yet still results in only two final segmentation embeddings $\{s_1, s_2\}$. Here, each s_i is derived from the multi-modal mapping Φ, which is implemented as $s_i = \Phi(f_{rgb,i}, f_{X,i}) = S(\Psi(MSN(f_{rgb,i}), MSN(f_{X,i})))$. In this formulation, Ψ represents the learnable fusion module that integrates heterogeneous features, while S denotes the segmentation head that projects the fused representation into the low-dimensional manifolds. The core objective of this framework is to optimize and activate the learnable fusion module Ψ without human supervision *by enforcing structural consistency between the fused embeddings and the heterogeneous features from frozen backbones.*

Modality Fusion. We integrate the refined features via a **learnable Conv Fusion** module Ψ: $f_{fus} = \Psi(MSN(f_{rgb}), MSN(f_X))$. While learnable fusion often converges to trivial solutions in unsupervised settings [71, 81], we demonstrate that Ψ can significantly outperform static operations (*e.g.*, sum or average)

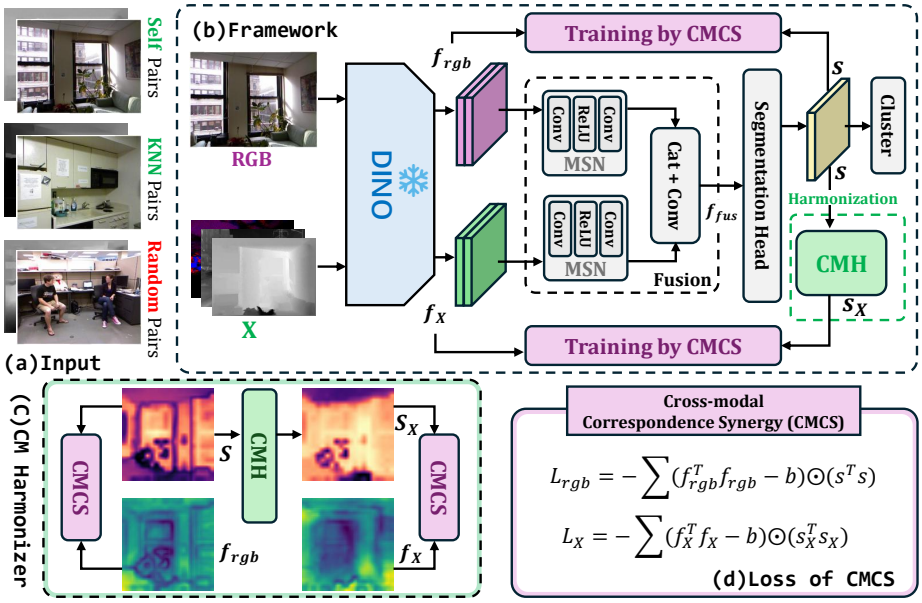


Fig. 2: UniM2 Framework Overview. (a) Training inputs for UniM2, (b) The overall architecture of UniM2, (c) The Cross-modal Harmonization process, and (d) the formulation of the CMCS loss. For illustration, (c) and (d) are depicted based on the *Self* pair scenario to resolve cross-modal structural contradictions.

when properly constrained. In UniM2, this integration is regularized by the **Cross-modal Correspondence Synergy (CMCS)**, which provides the explicit guidance necessary to harness cross-modal synergies.

Training by CMCS. The core philosophy of our training objective extends the principle of correspondence distillation to the multi-modal domain. While conventional USS [20] assumes that semantic similarity in a low-dimensional space should mirror the correlations of a single modality, we instead aim to achieve a **cross-modal relational consensus**. We posit that a robust unified semantic space should preserve only those structural relationships that are consistently supported across its constituent modalities. Specifically, if a semantic relationship is jointly captured by heterogeneous sensors, the shared embedding space should reflect this agreement through consistent cross-modal correspondences. Accordingly, we propose **Cross-modal Correspondence Synergy**, which encourages the fused embeddings s to be jointly constrained by the structural correlations of both f_{rgb} and f_X. By emphasizing multi-source agreement rather than one-way imitation, our framework facilitates the discovery of a shared semantic manifold while suppressing modality-specific noise.

Specifically, for a training pair $(\mathcal{G}_1, \mathcal{G}_2)$, we extract the dense feature maps $f_{rgb,1}, f_{rgb,2}$ and $f_{X,1}, f_{X,2}$ directly from the frozen backbone to serve as stable semantic anchors. The modality-specific correspondence tensors, F^{rgb} and F^X,

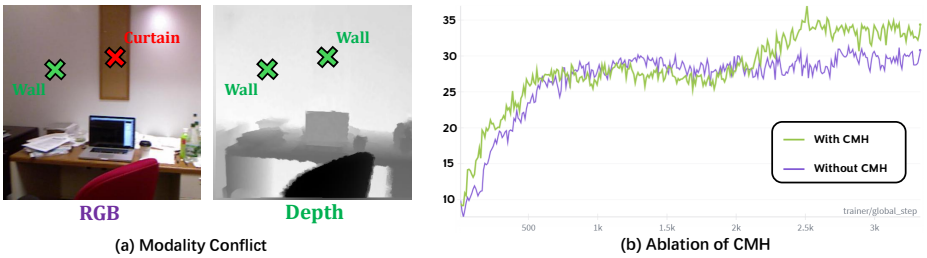


Fig. 3: Analysis of modality conflicts and CMH efficacy. (a) Example of contradictory semantic cues between RGB (semantic) and depth (geometry). (b) mIoU performance comparison with and without the proposed CMH.

are then defined as:

$$F^{rgb}_{hwij} = \cos(f_{rgb,1,hw}, f_{rgb,2,ij}), \quad F^{X}_{hwij} = \cos(f_{X,1,hw}, f_{X,2,ij}), \tag{3}$$

where (h, w) and (i, j) denote the spatial coordinates. Note that these cosine similarities are computed following the same formulation as in Eq. 1. The objective of CMCS is to distill the structural correlations from individual modalities into the unified semantic space. Similarly, the correspondence in the unified semantic space is computed as:

$$S_{hwij} = \cos(s_{1,hw}, s_{2,ij}), \tag{4}$$

where s_1 and s_2 denote the final segmentation embeddings of the two groups. To align the unified semantic space with its constituent modalities, the total CMCS loss is formulated as the weighted sum of individual modality-specific losses:

$$\mathcal{L}_{cmcs} = \mathcal{L}_{rgb} + \lambda \mathcal{L}_{X}, \tag{5}$$

where λ is a balancing hyperparameter. For the RGB and auxiliary modalities, the specific loss terms $\mathcal{L}_{rgb}$ and $\mathcal{L}_{X}$ are formulated as:

$$\mathcal{L}_{rgb} = -\sum_{h,w,i,j} (F^{rgb}_{hwij} - b) \odot \max(0, S_{hwij}), \quad \mathcal{L}_{X} = -\sum_{h,w,i,j} (F^{X}_{hwij} - b) \odot \max(0, S^{X}_{hwij}), \tag{6}$$

where S^{X}_{hwij} represents the embedding correspondence processed by the **Cross-modal Harmonizer (CMH)**, which will be detailed in the following section.

3.3 Cross-modal Harmonization for Modality Conflict

Despite its efficacy, CMCS relies on an implicit assumption of cross-modal semantic consistency, which is frequently violated by inherent physical sensor differences. As illustrated in Fig. 3 (a), RGB sensors capture sharp textural boundaries for objects like curtains, whereas depth sensors may perceive them as part of the wall due to negligible depth variance. Such contradictions introduce conflicting gradients that confuse the optimization process and degrade the quality of the unified embedding space s.

To mitigate these conflicts, we propose the **Cross-modal Harmonization** mechanism. Instead of enforcing a rigid, direct alignment between the unified embedding s and auxiliary features f_X, we designate RGB as the primary semantic reference and decouple the auxiliary supervision via a learnable buffer:

$$s^X = \mathrm{CMH}(s), \tag{7}$$

where $\mathrm{CMH}(\cdot)$ is a lightweight learnable transformation. This buffer serves as a flexible mediation layer that facilitates a *soft-alignment* between the unified space and auxiliary modalities. The harmonized correspondence S^X_{hwij} used in Eq. 6 is then formulated as:

$$S^X_{hwij} = \cos(s^X_{1,hw}, s^X_{2,ij}). \tag{8}$$

By supervising the transformed embedding s^X rather than the shared space s with the auxiliary signal f_X, CMH adaptively absorbs complementary cues while insulating the primary semantic manifold from pixel-wise modality contradictions. In practice, while CMH can be instantiated as any learnable network, we implement a lightweight two-layer convolutional structure for simplicity. This choice regulates the *strictness of alignment*: the capacity of this mediation layer determines the degree to which the primary space s is constrained by the auxiliary signal f_X. Such buffering ensures that auxiliary modalities provide structural guidance without distorting the unified semantic manifold. As evidenced in Fig. 3 (b), this mechanism achieves remarkable mIoU gains.

Scalability to More Modalities. The modularity of CMH enables UniM2 to scale seamlessly to multiple auxiliary modalities. By assigning an independent CMH branch to each source, auxiliary supervision is decoupled from the shared space s, preventing gradient interference. This allows the unified embedding to be jointly regularized without complex balancing strategies. Accordingly, the total CMCS loss generalizes to:

$$\mathcal{L}_{cmcs} = \mathcal{L}_{rgb} + \sum_{n=1}^{N} \lambda_n \mathcal{L}_{X_n}, \tag{9}$$

where $\mathcal{L}_{X_n}$ is the harmonized loss for the n-th modality. Mediating signals through independent transformations ensures stable optimization even as N increases, allowing UniM2 to scale across diverse sensing configurations while preserving the integrity of the primary semantic manifold.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate UniM2 on three representative multi-modal benchmarks: NYU-Depth-V2 [49], MFNet [19], and MCubeS [32].

NYU-Depth-V2 [49] is a standard indoor RGB-D segmentation benchmark containing 1,449 aligned RGB–depth image pairs with dense annotations. We adopt the 13 class evaluation protocol for indoor semantic segmentation.
MFNet [19] is an urban RGB–thermal segmentation dataset with 1,569 aligned image pairs captured under both daytime and nighttime. We evaluate performance on its 8 categories to assess robustness under illumination changes.
MCubeS [32] is a quad-modal dataset for semantic material segmentation, featuring aligned RGB, Near-Infrared (NIR), Degree of Linear Polarization (DoLP), and Angle of Linear Polarization (AoLP) images. We perform evaluation on its 20 categories to validate our model's effectiveness in **fusing more than two modalities** for robust material recognition.
Implementation Details. We employ **DINOv3** as the frozen backbone. Following [20], we adopt a *5-crop* strategy during pre-processing to enhance spatial resolution and feature correspondence quality. The segmentation head consists of two convolutional layers with an intermediate activation layer, consistent with standard USS frameworks. All ablation studies are conducted using the **DINOv3-Small/16** variant. We use mean Intersection over Union (**mIoU**) and Pixel Accuracy (**Acc.**) as our evaluation metrics. Benefiting from the frozen backbone, the training process remains efficient, with a total training time of **less than two hours per model** on a single NVIDIA GeForce RTX 5090 GPU.
Hyperparameter Settings and Fairness. Unsupervised semantic segmentation is generally sensitive to hyperparameter choices, and this sensitivity is not unique to UniM2. To ensure a fair and reproducible comparison, we allocate the same hyperparameter search budget to all compared methods. Specifically, each method is tuned with **200 iterations of Bayesian hyperparameter optimization** [68], rather than using a single unified configuration across different datasets and methods, which can lead to suboptimal or biased results. All models are trained using the **Adam** optimizer with a learning rate of 5×10^{-4} and a batch size of 32. Additional hyperparameter details, including those related to λ and b, are provided in the **Supplementary Material**.

4.2 Comparison Results

Comparison Methods. Our primary comparisons are based on USS extensions, including direct K-means clustering on DINO features, representative USS methods such as STEGO [20] and EAGLE [27], and their multi-modal variants. These baselines are the most relevant to UMSS, as they share the same label-free semantic segmentation objective and can be directly extended to multi-modal inputs. We also include image-level fusion and RGB-to-X distillation alternatives as supplementary comparisons. The former is constrained by modality-specific input compatibility, while the latter follows an asymmetric imitation protocol rather than joint multi-modal representation learning. The overall comparison coverage is summarized in Tab. 1, with detailed supplementary results provided in the **Supplementary Material**.
Performance on NYU-Depth-v2 and MFNet. Tab. 2 shows that directly introducing auxiliary modalities into existing USS baselines does not necessarily

Table 1: Summary of comparison coverage. We compare UniM2 with representative alternatives from image-level fusion, RGB-to-X distillation, and USS-based multi-modal extensions. The reported values are representative mIoU results: NYU-Depth-V2 uses DINOv3-Small/16, while MFNet uses DINOv3-Base/16.

Category	Method	NYU-Depth-V2	MFNet	Location
Image fusion	SwinFusion [39]	-	36.2	Supp. Chapter 1
Image fusion	Mask-DiFuser [56]	-	39.1	Supp. Chapter 1
RGB to X distillation	CORAL [53]	21.3	-	Supp. Chapter 2
RGB to X distillation	MMD [38]	20.1	-	Supp. Chapter 2
RGB to X distillation	Cosine	24.5	-	Supp. Chapter 2
USS extension	STEGO [20]	28.8	35.9	Main Text
USS extension	EAGLE [27]	27.4	37.8	Main Text
Ours	**UniM2**	**36.9**	**45.7**	Main Text

improve performance. For both STEGO and EAGLE, naive Depth/Thermal fusion often leads to clear mIoU degradation, suggesting that heterogeneous modalities introduce structural conflicts that cannot be properly resolved without label supervision. In contrast, UniM2 consistently turns auxiliary modalities into positive gains. With DINOv3-Base/16, UniM2 improves over the RGB-only STEGO baseline by 6.4 mIoU on NYU-Depth-v2 and 9.8 mIoU on MFNet. These results demonstrate that CMCS provides effective cross-modal correspondence supervision, while CMH mitigates unreliable auxiliary guidance and stabilizes multi-modal representation learning.

Per-class Analysis on NYU-Depth-v2. Tab. 3 provides a finer-grained view of how different fusion strategies affect semantic categories. Naive depth fusion improves geometry-sensitive categories such as *Sofa*, where depth offers useful structural cues, but it substantially hurts appearance-dominant categories such as *Floor* and *Wall*. This indicates that auxiliary modalities can be beneficial for some categories while being harmful for others if cross-modal conflicts are not controlled. UniM2 achieves a better balance: it obtains the best results on *Sofa*, *Table*, and *TV*, while preserving strong performance on *Floor* and *Wall*. This confirms that UniM2 can exploit complementary geometric information without sacrificing the semantic structure captured by RGB.

Performance on MCubeS. Tab. 4 further evaluates UniM2 under a more challenging multi-modal setting with RGB, NIR, DoLP, and AoLP inputs. UniM2 achieves consistent improvements when informative modalities are added, such as $I \rightarrow IN$ and $ID \rightarrow IND$, demonstrating that the proposed CMH design can naturally extend beyond bi-modal fusion. Meanwhile, the slight drops observed in settings involving AoLP suggest that weak or noisy modalities may still limit the final performance. This observation is consistent with supervised multi-modal segmentation, where sensor quality and modality reliability remain important factors [33, 83, 84].

Visualization Results. Fig. 4 visually compares UniM2 with RGB-only and naive multi-modal baselines on NYU-Depth-v2 and MFNet. While baseline

Table 2: Quantitative comparison on NYU-Depth-v2 [49] and MFNET [19] datasets. Note that performance variations (↑, ↓) for UniM2 are reported relative to the RGB-only baseline of STEGO [20].

Method	Modality	Backbone	NYU-Depth-v2 [49]		MFNET [19]	
			mIoU ↑	Acc. ↑	mIoU ↑	Acc. ↑
DINOv3 [50]	RGB	ViT-S/16	11.1	26.2	20.1	67.3
	+ Depth/Thermal		9.4 (↓ 1.7)	25.5 (↓ 0.7)	19.8 (↓ 0.3)	66.5 (↓ 0.8)
+ STEGO [20]	RGB		28.8	52.0	32.2	72.1
	+ Depth/Thermal		25.3 (↓ 3.5)	45.9 (↓ 6.1)	31.3 (↓ 0.9)	74.9 (↑ 2.8)
+ EAGLE [27]	RGB		27.4	51.6	34.7	79.6
	+ Depth/Thermal		20.1 (↓ 7.3)	40.0 (↓ 11.6)	31.1 (↓ 3.6)	77.4 (↓ 2.2)
+ UniM2 (Ours)	**+ Depth/Thermal**		**36.9** (↑ 8.1)	**56.1** (↑ 4.1)	**35.2** (↑ 3.0)	**81.5** (↑ 9.4)
DINOv3 [50]	RGB	ViT-B/16	14.3	32.8	20.0	72.1
	+ Depth/Thermal		10.4 (↓ 3.9)	23.3 (↓ 9.5)	21.2 (↑ 1.2)	72.4 (↑ 0.3)
+ STEGO [20]	RGB		31.7	55.1	35.9	73.6
	+ Depth/Thermal		31.1 (↓ 0.6)	49.7 (↓ 5.4)	32.5 (↓ 3.4)	74.1 (↑ 0.5)
+ EAGLE [27]	RGB		30.9	49.5	37.8	72.5
	+ Depth/Thermal		25.8 (↓ 5.1)	46.9 (↓ 2.6)	33.5 (↓ 4.3)	74.5 (↑ 2.0)
+ UniM2 (Ours)	**+ Depth/Thermal**		**38.1** (↑ 6.4)	**58.8** (↑ 3.7)	**45.7** (↑ 9.8)	**76.1** (↑ 3.7)

Table 3: Per-class IoU comparison on the NYU-Depth-v2 dataset. The **best** and second-best results are highlighted in **bold** and underline, respectively.

Method	Modality	Bed	Book	Ceil	Chair	Floor	Furn	Obj	Pic	Sofa	Table	TV	Wall	Wind	**mIoU**
STEGO [20]	RGB	**54.2**	1.1	25.1	34.6	58.4	44.7	19.2	19.8	1.0	15.8	13.6	**55.8**	**68.8**	31.7
	+ Depth	51.2	11.4	21.8	41.4	23.3	41.5	15.5	29.8	46.7	14.5	9.4	40.9	56.9	31.1
EAGLE [27]	RGB	44.9	15.5	**46.0**	41.5	45.7	43.8	**23.3**	**31.9**	0.0	9.1	0.0	43.0	57.0	30.9
	+ Depth	47.3	**16.3**	2.1	30.4	**68.4**	**45.5**	19.2	6.4	1.6	4.7	5.8	39.6	48.1	25.8
UniM2 (Ours)	**+ Depth**	52.8	12.9	22.8	**41.7**	58.2	44.5	21.8	30.8	**55.7**	**18.9**	**20.7**	53.5	61.0	**38.1**

methods often produce fragmented masks or incorrect regions after introducing auxiliary modalities, UniM2 generates cleaner semantic maps with sharper object boundaries. Fig. 5 further shows that the fused representation f_{fus} is more spatially coherent than individual modality features, qualitatively supporting the effectiveness of CMCS and CMH in resolving modality conflicts.

4.3 Ablation Studies

We conduct ablation studies on NYU-Depth-v2 with the **DINOv3-Small/16** model, covering component effectiveness, CMH placement, and fusion strategies.

Component Effectiveness. As shown in Tab. 5, the baseline without the proposed modules obtains 25.0% mIoU, corresponding to multi-modal STEGO [20] with Conv Fusion. Adding CMCS and MSN improves the result to 31.3%, already surpassing the RGB-only baseline of 28.8%. Introducing CMH further boosts the performance from 31.3% to 36.9%, highlighting its importance in harmonizing modality conflicts. The gain from 34.2% to 36.9% also confirms the benefit of MSN for feature refinement before fusion.

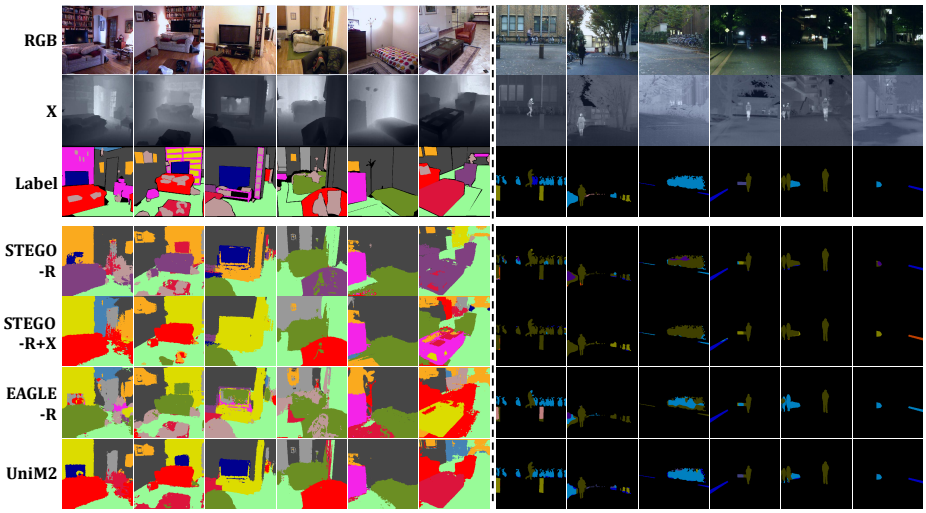


Fig. 4: Qualitative Comparison. Visual results on the NYU-Depth-v2 and MFNet datasets, comparing UniM2 against RGB-only and multi-modal (R+X) variants of STEGO [20] and RGB-only EAGLE [27] baselines.

Table 4: Quantitative comparison on the MCubeS [32] dataset. All methods utilize the ViT-S/16 backbone. I, N, A, and D denote RGB, NIR, AoLP, and DoLP modalities.

Method	I		IN		IA		ID		IND		INAD	
	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.
DINOv3 [50]	13.9	50.7	14.6	52.7	11.2	35.5	12.9	38.3	10.9	43.3	11.1	36.5
+ STEGO [20]	19.1	55.2	18.5	54.3	17.4	54.6	17.6	52.1	18.8	56.7	18.3	56.4
+ EAGLE [27]	18.1	57.7	18.7	58.2	16.5	57.2	17.0	51.1	18.6	57.2	16.5	53.2
+ UniM2 (Ours)	-	-	**21.5**	**59.1**	**18.5**	**61.7**	**20.9**	**57.8**	**21.8**	**64.5**	**20.7**	**62.6**

Placement of CMH. CMH provides learnable flexibility for mitigating structural conflicts, while the modality without CMH serves as a fixed anchor. Tab. 6 studies this anchoring effect by applying CMH to different branches. The *No anchors* setting drops to 19.8% mIoU, showing that excessive flexibility makes training unstable without a reliable reference. Keeping *RGB only* as the anchor achieves the best result of 36.9%, clearly outperforming the *Depth only* anchor setting of 27.6%. This suggests that RGB offers a more reliable semantic structure for unsupervised clustering, while depth is better used as complementary guidance. This supports our asymmetric design, where RGB provides a stable semantic reference and the auxiliary modality adapts through CMH.

Fusion Strategy. Tab. 7 compares different fusion operators. Conv Fusion achieves the highest mIoU of 36.9%, outperforming static operations such as Max, Mean, and Sum. Unlike static fusion, which applies fixed aggregation rules, Conv Fusion can adaptively select useful cues across pixels and channels. These results show that learnable fusion can be effective in UMSS when modality conflicts are properly regulated by CMH.

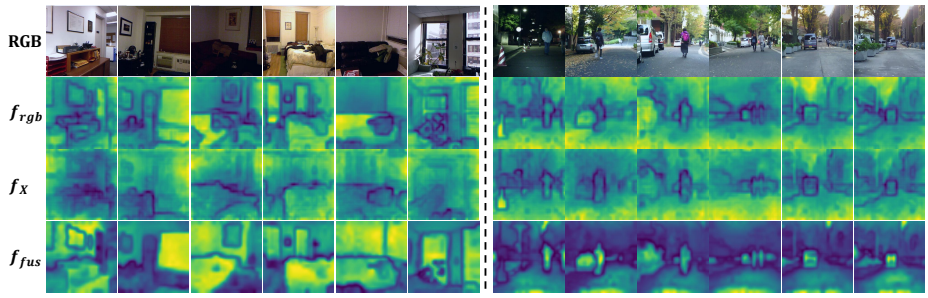


Fig. 5: Feature Visualization. Comparative visualization of feature maps from the RGB modality, the auxiliary modality, and the proposed fused representation.

Table 5: Ablation of each component in UniM2.

CMCS	MSN	CMH	mIoU ↑	Acc. ↑
			25.0	48.1
✓	✓		31.3	50.7
✓		✓	34.2	54.3
✓	✓	✓	**36.9**	**56.1**

Table 6: Ablation of anchor position.

Position	mIoU ↑	Acc. ↑
Both	31.3	50.7
Depth only	27.6	45.3
RGB only	**36.9**	**56.1**
No anchors	19.8	50.8

Table 7: Ablation of fusion strategies.

Strategy	mIoU ↑	Acc. ↑
Max	30.5	50.2
Mean	32.5	53.2
Sum	33.0	53.6
Conv	**36.9**	**56.1**

More analyses are provided in the **Supplementary Material**, including: (1) **additional fusion baselines** [39,56,57] in UMSS; (2) in-depth **theoretical analysis of CMH**; (3) distillation paradigms used in CMKD/UDA for UMSS; (4) hyperparameter analysis in UMSS; and (5) per-category distribution and confusion matrices. Extensive visualizations are also provided.

5 Conclusion and Future Work

In this paper, we have defined the task of **Unsupervised Multi-modal Semantic Segmentation** and proposed **UniM2**, a novel framework for leveraging heterogeneous modalities without any human annotations. We achieve effective multi-modal integration through **Cross-modal Correspondence Synergy**, which enforces structural consistency between a unified latent space and its constituent modalities. To address the inherent inter-modal conflicts arising from the diverse physical properties of different sensors, we further introduce the **Cross-modal Harmonizer**. By designating RGB as a stable semantic reference, CMH facilitates the absorption of complementary cues while mitigating contradictory supervision. **Crucially, UniM2 transforms unsupervised multi-modal semantic segmentation from performance degradation to consistent and substantial gains**, reversing the common failure of conventional fusion schemes in label-free settings. Notably, our modular design ensures high **Scalability**, allowing UniM2 to effectively extend to multiple auxiliary modalities. We hope that our UniM2 framework and the established evaluation benchmark can serve as a valuable baseline and inspire further advancements in the field of UMSS.

References

1. Bojanowski, P., Joulin, A.: Unsupervised learning by predicting noise. In: International conference on machine learning. pp. 517–526. PMLR (2017) 2
2. Brödermann, T., Bruggemann, D., Sakaridis, C., Ta, K., Liagouris, O., Corkill, J., Van Gool, L.: Muses: The multi-sensor semantic perception dataset for driving under uncertainty. In: ECCV. pp. 21–38. Springer (2024) 4
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV. pp. 132–149 (2018) 3
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021) 2, 4
5. Chib, P.S., Singh, P.: Recent advancements in end-to-end autonomous driving using deep learning: A survey. IEEE Transactions on Intelligent Vehicles **9**(1), 103–118 (2023) 1
6. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR. pp. 16794–16804 (2021) 3
7. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. NeurIPS **33**, 8765–8775 (2020) 4
8. Chung, I., Kim, D., Kwak, N.: Maximizing cosine similarity between spatial features for unsupervised domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1351–1360 (2022) 5
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 2
10. Dike, H.U., Zhou, Y., Deveerasetty, K.K., Wu, Q.: Unsupervised learning based on artificial neural network: A review. In: 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). pp. 322–327. IEEE (2018) 2
11. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(2), 230–244 (2022) 1
12. Fang, Y., Yap, P.T., Lin, W., Zhu, H., Liu, M.: Source-free unsupervised domain adaptation: A survey. Neural Networks **174**, 106230 (2024) 4
13. Feng, Z., Xue, R., Yuan, L., Yu, Y., Ding, N., Liu, M., Gao, B., Sun, J., Zheng, X., Wang, G.: Multi-agent embodied ai: Advances and future directions. arXiv preprint arXiv:2505.05108 (2025) 1
14. Fundel, F., Schusterbauer, J., Hu, V.T., Ommer, B.: Distillation of diffusion features for semantic correspondence. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6762–6774. IEEE (2025) 5
15. Gao, L., Chen, W., Wang, D., Guo, F., Liang, C.: Disentangled cross-modal representation learning with enhanced mutual supervision. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 4
16. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International journal of computer vision **129**(6), 1789–1819 (2021) 5
17. Greene, D., Cunningham, P., Mayer, R.: Unsupervised learning and clustering. In: Machine learning techniques for multimedia: Case studies on organization and retrieval, pp. 51–90. Springer (2008) 2

18. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: Knn model-based approach in classification. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". pp. 986–996. Springer (2003) 5
19. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T.: Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5108–5115. IEEE (2017) 2, 3, 9, 10, 12
20. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. ICLR (2022) 2, 3, 4, 5, 6, 7, 10, 11, 12, 13
21. Hu, H., Xie, L., Hong, R., Tian, Q.: Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: CVPR. pp. 3123–3132 (2020) 4
22. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: ICCV. pp. 9865–9874 (2019) 3
23. Jia, D., Guo, J., Han, K., Wu, H., Zhang, C., Xu, C., Chen, X.: Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. ICML (2024) 1
24. Kang, B., Mithun, N.C., Rajvanshi, A., Chiu, H.P., Samarasekera, S.: Duda: Distilled unsupervised domain adaptation for lightweight semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8124–8135 (2026) 4
25. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. pp. 4893–4902 (2019) 4
26. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS **33**, 18661–18673 (2020) 4
27. Kim, C., Han, W., Ju, D., Hwang, S.J.: Eagle: Eigen aggregation learning for object-centric unsupervised semantic segmentation. In: CVPR. pp. 3523–3533 (2024) 2, 3, 4, 5, 10, 11, 12, 13
28. Koenig, A., Schambach, M., Otterbach, J.: Uncovering the inner workings of stego for safe unsupervised semantic segmentation. In: CVPRW. pp. 3789–3798 (2023) 5
29. Lan, M., Wang, X., Ke, Y., Xu, J., Feng, L., Zhang, W.: Smooseg: smoothness prior for unsupervised semantic segmentation. NeurIPS **36**, 11353–11373 (2023) 2, 4
30. Li, B., Wang, S., Ye, H., Gong, X., Xiang, Z.: Cross-modal knowledge distillation for depth privileged monocular visual odometry. IEEE Robotics and Automation Letters **7**(3), 6171–6178 (2022) 4
31. Li, B., Zhang, D., Zhao, Z., Gao, J., Li, X.: Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. pp. 1308–1317 (2025) 1, 2, 3
32. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal material segmentation. In: CVPR. pp. 19800–19808 (2022) 3, 4, 9, 10, 13
33. Liao, C., Lei, K., Zheng, X., Moon, J., Wang, Z., Wang, Y., Paudel, D.P., Van Gool, L., Hu, X.: Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In: CVPRW. pp. 1576–1586 (2025) 11
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 2
35. Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., Fan, X.: Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In: ICCV. pp. 8115–8124 (2023) 2, 4

36. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4823–4833 (2021) 4
37. Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., Lin, L.: Aligning cyber space with physical world: A comprehensive survey on embodied ai. IEEE/ASME Transactions on Mechatronics (2025) 1
38. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105. PMLR (2015) 11
39. Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y.: Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA Journal of Automatica Sinica **9**(7), 1200–1217 (2022) 11, 14
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 4
41. Qian, C., Xing, S., Li, S., Zhao, Y., Tu, Z.: Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. arXiv preprint arXiv:2503.11892 (2025) 4
42. Qing, Y., Zeng, D., Xie, S., Huang, K., Wang, Y.: Integrating low-level visual cues for enhanced unsupervised semantic segmentation. In: AAAI. vol. 39, pp. 6603–6611 (2025) 2, 4
43. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine **34**(6), 96–108 (2017) 4
44. Rolf, B., Beier, A., Jackson, I., Müller, M., Reggelin, T., Stuckenschmidt, H., Lang, S.: A review on unsupervised learning algorithms and applications in supply chain management. International Journal of Production Research **63**(5), 1933–1983 (2025) 2
45. Seong, H.S., Moon, W., Lee, S., Heo, J.P.: Leveraging hidden positives for unsupervised semantic segmentation. In: CVPR. pp. 19540–19549 (2023) 2, 4, 5
46. Shah, D., Osiński, B., Levine, S., et al.: Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In: Conference on robot learning. pp. 492–504. pmlr (2023) 1
47. Shin, U., Park, J., Kweon, I.S.: Deep depth estimation from thermal image. In: CVPR. pp. 1043–1053 (2023) 3
48. Si, X., Zhang, C., Li, S., Liang, J.: Source-free domain adaptation for unsupervised radar-based human activity recognition. Pattern Recognition **169**, 111866 (2026) 4
49. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. pp. 746–760. Springer (2012) 2, 3, 4, 9, 10, 12
50. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 2, 3, 4, 5, 12, 13
51. Sinaga, K.P., Yang, M.S.: Unsupervised k-means clustering algorithm. IEEE access **8**, 80716–80727 (2020) 2
52. Singamaneni, P.T., Bachiller-Burgos, P., Manso, L.J., Garrell, A., Sanfeliu, A., Spalanzani, A., Alami, R.: A survey on socially aware robot navigation: Taxonomy and future challenges. The International Journal of Robotics Research **43**(10), 1533–1572 (2024) 1
53. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV. pp. 443–450. Springer (2016) 11

54. Sun, J., Zhang, L., Zha, Y., Gonzalez-Garcia, A., Zhang, P., Huang, W., Zhang, Y.: Unsupervised cross-modal distillation for thermal infrared tracking. pp. 2262–2270 (2021) 4
55. Szeliski, R.: Computer vision: algorithms and applications. Springer Nature (2022) 2
56. Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. IEEE TPAMI (2025) 11, 14
57. Tang, L., Wang, Y., Cai, Z., Jiang, J., Ma, J.: Controlfusion: A controllable image fusion framework with language-vision degradation prompts. arXiv preprint arXiv:2503.23356 (2025) 14
58. Thoker, F.M., Gall, J.: Cross-modal knowledge distillation for action recognition. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 6–10. IEEE (2019) 4
59. Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: A brief review. Computational intelligence and neuroscience **2018**(1), 7068349 (2018) 2
60. Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 216–226. Springer (2023) 4
61. Wang, X., Zhang, H., Yu, H., Wan, X.: Evlsd-ied: Event-based line segment detection with image-to-event distillation. IEEE Transactions on Instrumentation and Measurement **73**, 1–12 (2024) 4
62. Wang, Y., Yu, H., Li, X.: Teacher-student consistent distillation for source-free domain adaptation object detection. In: NeurIPS. pp. 230–245. Springer (2025) 4
63. Wang, Y., Albrecht, C.M., Braham, N.A.A., Liu, C., Xiong, Z., Zhu, X.X.: Decoupling common and unique representations for multimodal self-supervised learning. In: ECCV. pp. 286–303. Springer (2024) 4
64. Wang, Z., Li, D., Luo, C., Xie, C., Yang, X.: Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In: ICCV. pp. 8637–8646 (2023) 4
65. Wei, S., Luo, Y., Wang, Y., Luo, C.: Robust multimodal learning via representation decoupling. In: ECCV. pp. 38–54. Springer (2024) 4
66. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(5), 1–46 (2020) 4
67. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018) 2, 3
68. Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H.: Hyperparameter optimization for machine learning models based on bayesian optimization. Journal of Electronic Science and Technology **17**(1), 26–40 (2019) 10
69. Wu, Y., Wang, Y., Zhang, S., Ogai, H.: Deep 3d object detection networks using lidar data: A review. IEEE Sensors Journal **21**(2), 1152–1171 (2020) 2
70. Xia, P., Zhang, L., Li, F.: Learning similarity with cosine similarity ensemble. Information sciences **307**, 39–52 (2015) 5
71. Xu, H., Ma, J., Yuan, J., Le, Z., Liu, W.: Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In: CVPR. pp. 19679–19688 (2022) 6
72. Xu, R., Wang, C., Sun, J., Xu, S., Meng, W., Zhang, X.: Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3045–3053 (2023) 5

73. Yang, C., Yu, X., Yang, H., An, Z., Yu, C., Huang, L., Xu, Y.: Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9148–9156 (2025) 4
74. Yuan, Y., Li, Z., Zhao, B.: A survey of multimodal learning: Methods, applications, and future. ACM Computing Surveys **57**(7), 1–34 (2025) 4
75. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access **8**, 58443–58469 (2020) 1
76. Zang, X., Zhang, J., Tang, B.: Molecular representation learning via multimodal fusion and decoupling. Information Fusion p. 103493 (2025) 4
77. Zhang, H., Wang, X., Xu, C., Wang, X., Xu, F., Yu, H., Yu, L., Yang, W.: Frequency-adaptive low-latency object detection using events and frames. arXiv preprint arXiv:2412.04149 (2024) 3
78. Zhang, H., Xu, C., Wang, X., Liu, B., Hua, G., Yu, L., Yang, W.: Detecting every object from events. IEEE TPAMI (2025) 2
79. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: CVPR. pp. 1136–1147 (2023) 1, 2, 3, 4
80. Zhang, S., Li, X., Zong, M., Zhu, X., Cheng, D.: Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST) **8**(3), 1–19 (2017) 5
81. Zhang, Y., Chen, Y., Gao, C.: Deep unsupervised multi-modal fusion network for detecting driver distraction. Neurocomputing **421**, 26–38 (2021) 6
82. Zhao, J., Wu, Y., Deng, R., Xu, S., Gao, J., Burke, A.: A survey of autonomous driving from a deep learning perspective. ACM Computing Surveys **57**(10), 1–60 (2025) 1
83. Zheng, X., Lyu, Y., Jiang, L., et al.: Reducing unimodal bias in multi-modal semantic segmentation with multi-scale functional entropy regularization. ICCV (2025) 11
84. Zheng, X., Lyu, Y., Zhou, J., Wang, L.: Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation. In: ECCV. pp. 192–212. Springer (2024) 1, 3, 11
85. Zhou, Q., Shi, Y., Yang, X., Xian, X., Liao, L., Zhang, R., Lin, L.: Dfvo: Learning darkness-free visible and infrared image disentanglement and fusion all at once. IEEE Transactions on Instrumentation and Measurement (2025) 3