UMSS : Towards Unsupervised Multi-modal Semantic Segmentation (Supplementary Materials)

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

Contents

1 Image Fusion Methods in UMSS

To establish a broader benchmark for UMSS, we further investigate an **image-level fusion paradigm** [13, 18]. This paradigm follows a two-stage *fusion-then-segmentation* pipeline. First, RGB and auxiliary modalities (*e.g.*, thermal images) are fused at the pixel-level to produce a single composite image using representative fusion models such as SwinFusion [13] and Mask-DiFuser [18]. The resulting fused image is then processed by existing USS frameworks (*e.g.*, STEGO [8] and EAGLE [9]) as if it were a standard RGB input.

Quantitative results are summarized in Tab. 1, and qualitative comparisons are shown in Fig. 1. Although Mask-DiFuser produces visually superior fusion results compared to SwinFusion, the resulting performance gains in semantic discovery remain modest. Specifically, the strongest image-level baseline (**Mask-DiFuser + EAGLE**) achieves **39.1%** mIoU, which still falls behind our **UniM2** (45.7%) by a substantial margin of **6.6%**. We attribute this gap to two limitations:

– **Information Loss [19]:** The two-stage design introduces an inherent constraint where the fusion stage is optimized for human perceptual quality rather than semantic discriminability. Forcing heterogeneous multi-modal

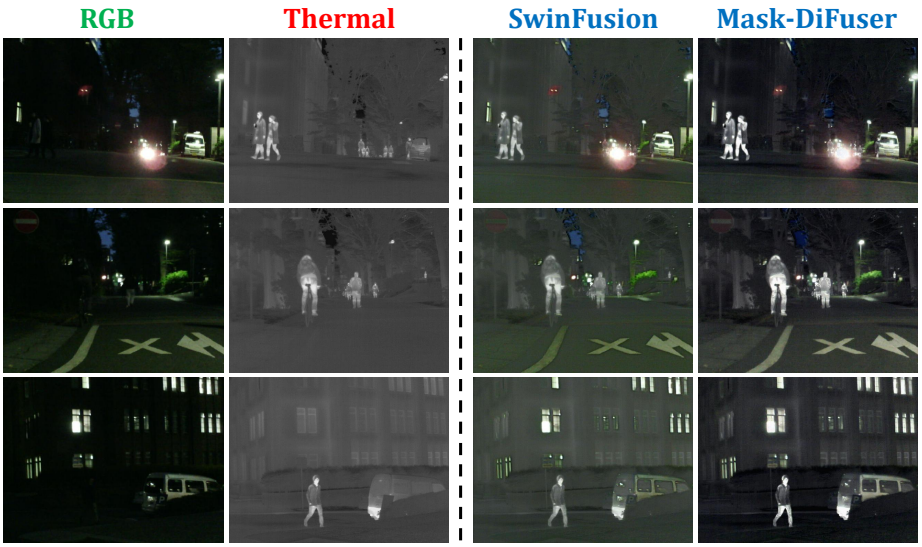

Fig. 1: Visual comparison on the MFNet [7] dataset. From left to right: RGB, Thermal, and fusion results produced by SwinFusion [13] and Mask-DiFuser [18].

signals into a compressed three-channel representation inevitably discards task-relevant modality cues before semantic grouping occurs.
– **Representation Mismatch:** While fused images may appear visually plausible, they often deviate from the natural image distribution expected by pre-trained backbones. This distributional shift leads to feature ambiguity and sub-optimal clustering during the unsupervised segmentation process.

More importantly, image-level fusion lacks scalability when extending to heterogeneous modalities such as RGB-Depth. Unlike thermal and visible images that share similar radiometric properties, RGB and Depth capture fundamentally different physical signals: radiometric reflectance versus geometric structure. Direct pixel-wise fusion between these modalities is therefore physically ill-posed and rarely adopted in RGB-D perception tasks. This intrinsic incompatibility largely confines image-level fusion to RGB-Thermal benchmarks and limits its applicability in general multi-modal perception scenarios.

In contrast, our feature-level framework (**UniM2**) preserves modality-specific characteristics within their respective representation spaces and aligns them through semantic-level interactions. This design avoids the information loss and modality mismatch inherent in pixel-level fusion, enabling more effective multi-modal semantic discovery.

2 Distillation Paradigms in UMSS

In contrast to image-level fusion methods operating in the pixel space, we investigate a feature-level distillation paradigm that integrates cross-modal objectives

Table 1: Quantitative comparison of image-level fusion strategies on the MFNet dataset [7]. We evaluate the performance of various USS frameworks (STEGO [8] and EAGLE [9]) when using pixel-level fused images as input. All experiments utilize a DINOv3-pretrained ViT-B/16 backbone.

Method	Input Modality / Fusion Strategy	mIoU	Acc.
STEGO [8]	RGB (Baseline)	35.9	73.6
	Fused by SwinFusion [13]	34.5 (↓ 1.4)	74.6
	Fused by Mask-DiFuser [18]	38.8 (↑ 2.9)	**78.6**
EAGLE [9]	RGB (Baseline)	37.8	72.5
	Fused by SwinFusion [13]	36.2 (↓ 1.6)	73.1
	Fused by Mask-DiFuser [18]	39.1 (↑ 1.3)	74.0
UniM2 (Ours)	**RGB+Thermal**	**45.7** (↑ 9.8)	76.1

within the latent representation space. In this framework, both RGB and auxiliary modalities are processed by a DINO backbone to extract high-level features. The auxiliary features are then passed through a lightweight projection layer to facilitate alignment with the RGB feature space.

To guide the learning of this projection, a distillation objective is introduced to encourage the auxiliary representation to align with the semantic structure of the RGB branch. By minimizing the discrepancy between the projected auxiliary features and the corresponding RGB features, the model transfers the object-centric priors learned by the DINOv3 backbone to other modalities.

Since various distillation formulations commonly adopted in research areas such as Cross Modal Knowledge Distillation (CMKD) or Unsupervised Domain Adaptation (UDA) can be applied in this context, we express the general alignment objective as:

$$\mathcal{L}_{distill} = \mathcal{L}\left(f_{rgb}, \text{Proj}(f_X)\right), \tag{1}$$

where f_{rgb} and f_X denote the features extracted from the RGB and auxiliary modalities, respectively, and $\text{Proj}(\cdot)$ represents the projection function. This formulation allows the auxiliary modality to align with the semantic manifold of the RGB representation while contributing complementary information for multi-modal learning.

To instantiate Eq. 1, we adopt three representative distillation objectives that characterize feature alignment from different perspectives.

– **Correlation Alignment (CORAL) [17].** CORAL aligns the second-order statistics of feature distributions by minimizing the distance between covariance matrices of RGB features and projected auxiliary features:

Table 2: Ablation study of different modal alignment and fusion strategies on NYU-Depth-v2 dataset. All methods use DINOv3-Small/16 as the frozen backbone.

Method	Alignment	Fusion	mIoU (%)	Acc. (%)
STEGO [8]	None (Original)	Sum	23.2	43.3
		Conv	25.0	48.1
	CORAL (2nd-order)	Sum	16.1	35.7
		Conv	21.3	43.7
	MMD (Kernel)	Sum	20.1	40.2
		Conv	18.5	35.9
	Cosine (Patch-wise)	Sum	16.4	40.7
		Conv	24.5	49.6
UniM2 (Ours)	**CMCS + CMH**	**Conv**	**36.9**	**56.1**

$$\mathcal{L}_{CORAL} = \frac{1}{4d^2} \left\| C_{rgb} - C_X \right\|_F^2, \tag{2}$$

where C_{rgb} and C_X denote the covariance matrices computed from f_{rgb} and $\mathrm{Proj}(f_X)$, respectively, and d is the feature dimension.

– **Maximum Mean Discrepancy (MMD) [12].** MMD measures the distribution discrepancy in a Reproducing Kernel Hilbert Space (RKHS) [1] and encourages the projected auxiliary features to match the global distribution of RGB features:

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N} \sum_{i=1}^{N} \phi(f_{rgb,i}) - \frac{1}{N} \sum_{i=1}^{N} \phi(\mathrm{Proj}(f_{X,i})) \right\|^2, \tag{3}$$

where $\phi(\cdot)$ denotes the kernel feature mapping.

– **Cosine Similarity.** Cosine similarity enforces geometric alignment between feature vectors by maximizing their angular consistency:

$$\mathcal{L}_{cos} = 1 - \frac{f_{rgb} \cdot \mathrm{Proj}(f_X)}{\|f_{rgb}\| \, \|\mathrm{Proj}(f_X)\|}. \tag{4}$$

We conduct a comprehensive evaluation on the NYU-Depth-v2 dataset to investigate the effectiveness of the aforementioned distillation paradigms. As reported in Tab. 2, the quantitative results reveal a counter-intuitive trend where the introduction of rigid distillation objectives often leads to a significant performance degradation compared to the vanilla STEGO baseline. For instance, employing CORAL or MMD with a SUM fusion strategy results in mIoU scores of 16.1% and 20.1% respectively, both of which are inferior to the 23.2% achieved by the original single modal setup. This suggests that forcing the auxiliary depth modality to strictly mimic the RGB feature distribution can be counterproductive for unsupervised semantic segmentation.

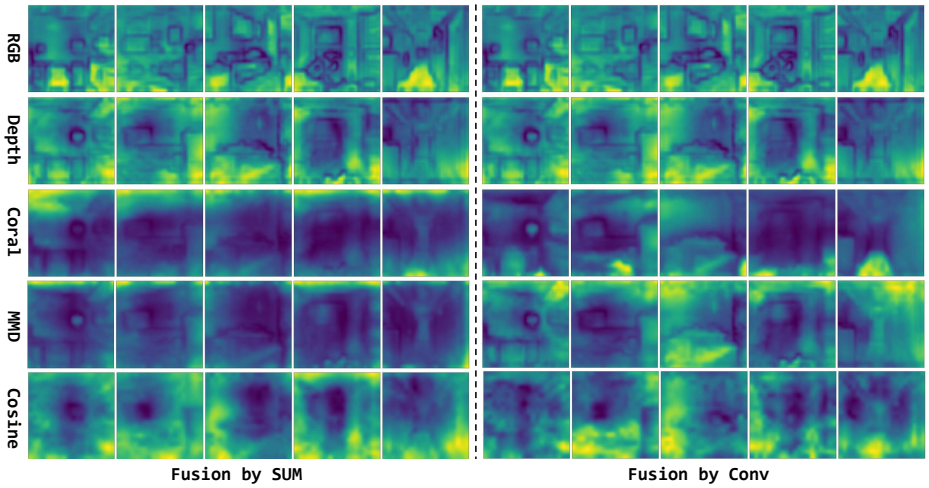

Fig. 2: Qualitative comparison of latent feature maps. The visualization illustrates the structural collapse in the auxiliary modality following distillation (Coral, MMD, and Cosine) when compared to the original Depth features. This highlights how rigid alignment objectives suppress inherent geometric cues, resulting in blurred representations that fail to capture clear object boundaries.

The qualitative visualizations in Fig. 2 further reveal the root cause of this failure. As shown in the feature maps for Coral, MMD, and Cosine, the latent representations exhibit a noticeable loss of structural clarity and object boundaries compared to the original Depth features. This degradation arises because the alignment process forces the auxiliary branch to conform to the RGB feature manifold, suppressing its inherent geometric cues. As a result, the distillation objectives behave as a destructive filter that removes high-frequency structural details, producing the blurred representations observed in the visualization. Consequently, the auxiliary branch degenerates into a weakened replica of RGB features rather than providing complementary information. In contrast, our UniM2 framework achieves a substantial performance gain to 36.9% mIoU via CMCS. Instead of enforcing feature value imitation, *UniM2 preserves the intrinsic topological structure of each modality, enabling collaborative semantic consensus while retaining high-frequency geometric details across the latent manifold.*

3 Information Bottleneck Perspective on CMH

From the perspective of the Information Bottleneck principle [19], the mapping from high-dimensional backbone features f to a compact embedding space s can be interpreted as a lossy information compression process [10]. Within the unsupervised semantic segmentation paradigm, the distillation objective serves as a fidelity constraint that encourages the bottleneck representation s to retain the most representative semantic structures encoded in the input features.

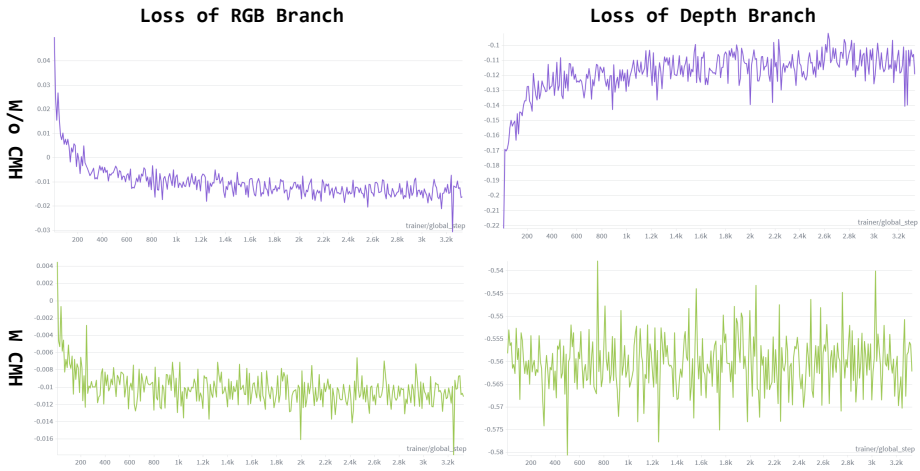

Fig. 3: Training loss analysis on NYU-Depth-v2 on both RGB and Depth branch. Without CMH, the auxiliary branch loss increases continuously, revealing severe modality conflicts that force the model to neglect auxiliary cues. With the proposed CMH, the auxiliary loss stabilizes throughout the training process, enabling effective and synergistic multi-modal representation learning.

Consequently, the learning objective can be interpreted through the IB trade-off between compression and semantic sufficiency:

$$\min_{\theta} \mathcal{L}_{IB} = \mathcal{I}(f; s) - \beta \mathcal{I}(s; Y_{proxy}), \tag{5}$$

where β controls the balance between compressing redundant feature information and preserving task-relevant semantics represented by the proxy labels Y_{proxy}.

In the multi-modal setting, however, the proposed **Cross-modal Correspondence Synergy** objective introduces multiple modality-dependent constraints. From an optimization perspective, the shared latent representation s is required to simultaneously accommodate information from heterogeneous feature sources:

$$\mathcal{L}_{joint} = \mathcal{I}(f_{rgb}; s) + \lambda \mathcal{I}(f_X; s) - \beta \mathcal{I}(s; Y_{proxy}), \tag{6}$$

where f_{rgb} and f_X denote backbone features extracted from RGB and auxiliary modalities respectively. This formulation implicitly imposes dual mutual-information constraints on the shared embedding space, which may introduce competing optimization pressures during training.

Under a rigorous Bayesian hyper-parameter optimization process (200 iterations) [20], we empirically observe that the framework without CMH tends to converge to an RGB-dominated equilibrium. As illustrated by the loss curves in Fig. 3, the auxiliary depth loss without CMH exhibits a continuous upward trend, suggesting that the optimization tends to suppress auxiliary signals in order to maintain the stability of the primary RGB-driven semantic manifold. In contrast, with the introduction of CMH, the auxiliary loss stabilizes within a

consistent range without significant increasing or decreasing trends, indicating that the conflicting gradients between modalities are effectively mitigated.

To alleviate this limitation, CMH introduces a learnable transformation $s^X = \Phi(s)$ that re-parameterizes the auxiliary branch and converts the joint objective into a conditional bottleneck formulation:

$$\mathcal{L}_{CMH} = \mathcal{L}_{rgb}(s) + \lambda \mathcal{L}_X(s^X). \tag{7}$$

Importantly, this design preserves the original RGB optimization pathway while allowing the auxiliary modality to interact with the shared representation through a dedicated transformation. Applying the chain rule, the gradient propagated from the auxiliary branch becomes

$$\nabla_s \mathcal{L}_X = \frac{\partial \mathcal{L}_X}{\partial s^X} \cdot \frac{\partial \Phi}{\partial s}, \tag{8}$$

where the Jacobian $\frac{\partial \Phi}{\partial s}$ functions as a learnable gating mechanism that selectively modulates the influence of auxiliary signals. This mechanism filters task-irrelevant modal noise while preserving complementary information, enabling both modalities to be optimized simultaneously. *Consequently, CMH relaxes the conventional single-bottleneck constraint into a conditional bottleneck structure, effectively decoupling the optimization trajectories of heterogeneous modalities while preserving a shared semantic representation. This transformation converts potential inter-modal interference into synergistic representation learning.*

4 Hyperparameter analysis in UMSS

In the unsupervised learning paradigm [5, 6], the absence of ground-truth labels makes it inherently difficult to determine the optimal latent structure during training. Consequently, USS frameworks often exhibit a high sensitivity to hyperparameter configurations [8, 9]. Without meticulously selected hyperparameters, the performance of USS is prone to a complete collapse, where the model fails to learn any meaningful semantic boundaries; this critical vulnerability remains equally prevalent in the UMSS setting.

To systematically address this dependency, we analyze the feature correspondence distillation objective established by STEGO [8], which serves as the foundation for our UMSS framework. As shown in Eq. 9, the optimization process relies on three distinct types of correspondence pairs inherited from the standard USS paradigm: **Self**, **KNN**, and **Random**.

$$\mathcal{L} = -\sum_{hwij} (F_{hwij} - b) \odot \max(S_{hwij}, 0), \tag{9}$$

where F_{hwij} denotes the backbone feature similarity, S_{hwij} represents the similarity in the latent space, and b acts as a bias term. In practical implementation, each of these three pair categories is governed by two fundamental parameters including a loss weight (λ) and a bias (b), resulting in a total of six critical hyperparameters that define the optimization landscape.

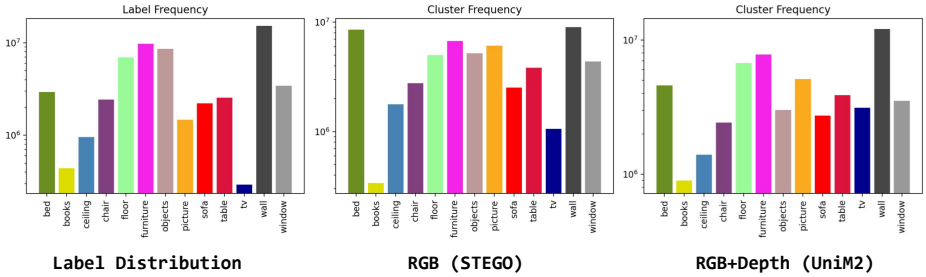

Fig. 4: Comparison of Category Distributions. The plots showcase the label frequency (GT) alongside the cluster frequencies of the STEGO (RGB) baseline and our UniM2 (RGB + Depth) framework to highlight the improved class balance.

- **Self-correlation** (λ_{self}, b_{self}): This pair reinforces nearby pixel relationships to preserve the local spatial consistency of backbone features.
- **KNN-correlation** (λ_{knn}, b_{knn}): This pair clusters semantically similar neighbors from the backbone feature space into a compact latent manifold.
- **Random-correlation** (λ_{rand}, b_{rand}): This pair introduces negative pressure on random samples to act as a contrastive regularizer against feature collapse.

The bias terms b are particularly sensitive because they define the neutral point for semantic affinity; a poorly calibrated bias will lead to either excessive feature grouping or an inability to form coherent clusters.

Beyond these distillation parameters, our UniM2 framework introduces the CMCS loss to synchronize semantic patterns across diverse data streams. Specifically, for each additional modality m incorporated into the system, we introduce a corresponding weight parameter $\lambda_{cmcs}^{(m)}$ to regulate the influence of cross-modal consistency relative to individual modality refinement. Although the inclusion of these modality specific weights expands the high dimensional search space, we still utilize the same 200 iterations of Bayesian hyperparameter optimization to ensure the stability of the learned semantic manifold.

5 Category Distribution and Confusion Matrix

To investigate the impact of multi-modal integration on semantic consistency, we visualize the cluster frequency and the confusion matrix for both the STEGO baseline and our UniM2 framework. The class distribution analysis in Fig. 4 reveals that the single modal RGB model (STEGO) tends to exhibit a biased frequency distribution that deviates from the ground truth label frequency. In contrast, the cluster distribution of UniM2 (RGB + Depth) aligns more closely with the actual label frequency, demonstrating a superior capability in capturing diverse categories such as sofa, table, and tv.

The confusion matrices presented in Fig. 5 provide a more granular view of the performance gains. As illustrated, the STEGO baseline suffers from significant

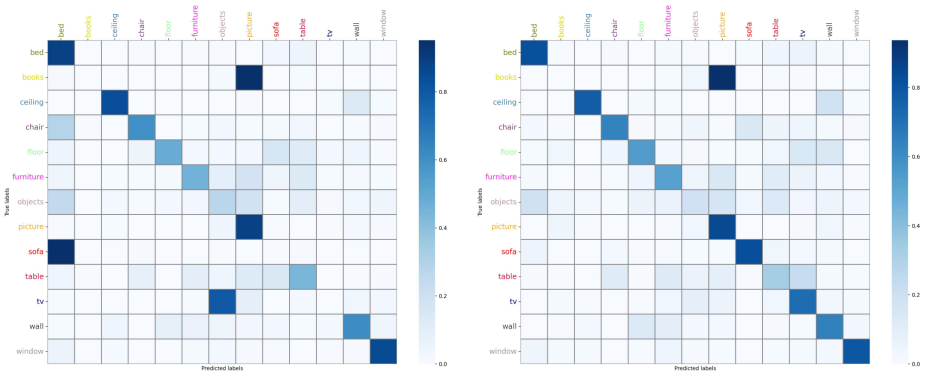

Fig. 5: Confusion Matrix Analysis. The matrices illustrate the prediction accuracy for the STEGO baseline and UniM2, demonstrating that the inclusion of depth information reduces misclassifications between photometrically similar categories.

misclassifications between photometrically similar objects, most notably mistaking sofa for bed due to their shared texture patterns. By incorporating geometric depth cues, our UniM2 framework successfully resolves these ambiguities, resulting in a cleaner diagonal across the matrix. Specifically, the prediction accuracy for categories including tv, table, and sofa is substantially improved, confirming that the fusion of depth effectively stabilizes the learned semantic manifold.

6 Visualization of MCubeS

MCubeS [11] is a quad-modal dataset designed for semantic material segmentation. It features aligned RGB, Near-Infrared (NIR), Angle of Linear Polarization (AoLP), and Degree of Linear Polarization (DoLP) images. In our qualitative analysis, the single-modal baseline (I) represents the results obtained by STEGO [8], while the subsequent multi-modal combinations showcase the performance of our proposed UniM2 framework. We perform evaluation on its 20 categories to validate the effectiveness of our model in fusing these diverse modalities for robust material recognition.

As shown in the visual comparison in Fig. 6, the STEGO baseline (I) often fails to distinguish between materials with similar photometric appearances due to the lack of auxiliary physical cues. By incorporating additional modalities through our UniM2 framework, the model captures supplementary spectral and polarimetric properties that are essential for material discrimination. The integration of diverse physical information leads to a progressive refinement of semantic boundaries. Specifically, the tri-modal (IND) and the full quad-modal configuration (IAND) yield segmentation masks that are highly consistent with the ground truth, demonstrating that the inclusion of more modalities generally enhances the discriminative power of the learned manifold.

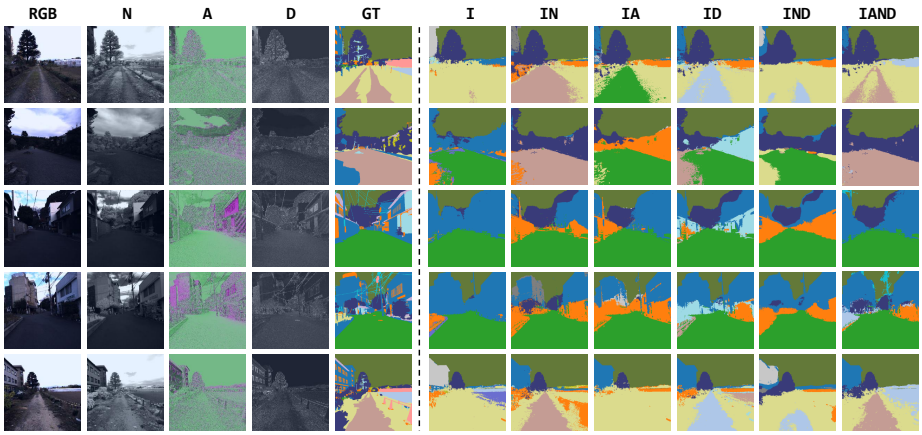

Fig. 6: Qualitative visualization and comparison on the MCubeS [11] dataset. The left columns display input modalities including RGB, Near-Infrared (N), Angle of Linear Polarization (A), and Degree of Linear Polarization (D) alongside the ground truth (GT). The right columns present segmentation results where (I) denotes the STEGO baseline, while the remaining columns (IN, IA, ID, IND, IAND) represent various fusion strategies within our UniM2 framework to highlight how the addition of multiple modalities leads to superior material recognition.

7 Visualization of MSN, Fusion, CMH Process

7.1 Modality-Specific Refinement and Fusion Process

To elucidate the internal mechanism of UniM2, we visualize the transition from raw backbone features to the final fused representation. As illustrated in Fig. 7, while the raw features extracted from the frozen DINOv3 backbone provide a semantically rich foundation, they are not inherently optimized for cross-modal alignment or synergistic integration. The Modality-Specific Networks (MSN) serve as a critical refinement stage, which sharpens structural patterns and enhances the contrast of semantic regions within each individual modality. The final fused feature map effectively integrates these refined complementary cues, exhibiting superior spatial coherence compared to the individual raw branches.

7.2 Visualization of Learning Dynamics of Fusion

To investigate the learning dynamics of UniM2, we visualize the evolution of the fused feature maps across different training iterations ($t = 1, 30, 90, 200, 1000, 4000, 6000$). As shown in Fig. 8, the initial representations at $t \leq 90$ are predominantly noisy and lack clear semantic structure, reflecting the disorganized state of the shared latent space before sufficient optimization. Driven by the CMCS loss, the framework gradually discovers shared semantic patterns, with meaningful object boundaries and consistent region-level clusters beginning to

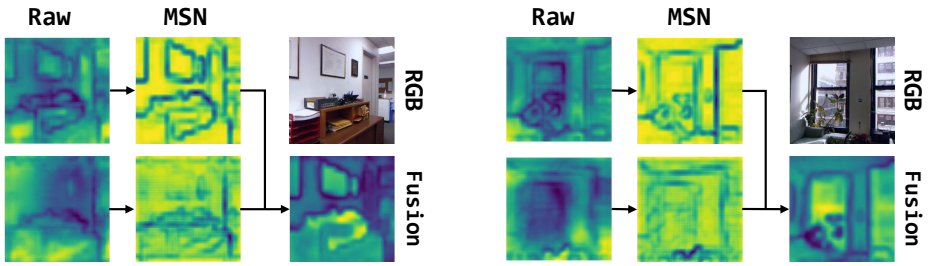

Fig. 7: Visualization of the feature refinement and fusion process. The MSN modules enhance the structural discriminativeness of raw backbone features, while the final fusion stage integrates complementary cues into a coherent semantic map.

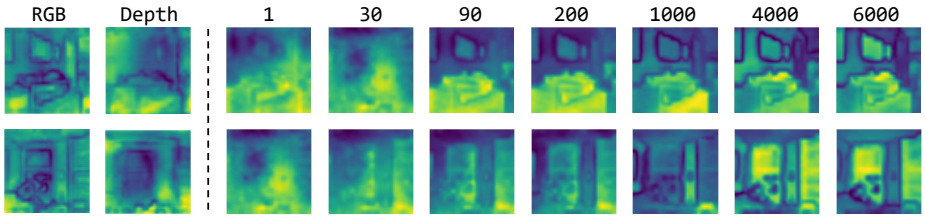

Fig. 8: Visual evolution of fused semantic manifolds across training iterations. The feature maps transition from disorganized noise to spatially coherent semantic clusters as training progresses from $t = 1$ to $t = 6000$.

emerge around $t = 1000$. By $t = 6000$, the manifold has converged into a highly discriminative and spatially coherent state, demonstrating that UniM2 effectively distills and refines structural knowledge from the frozen backbones into a unified semantic space without any human supervision.

7.3 Visualization of Cross-modal Harmonization

To further investigate the internal mechanism of multi-modal synchronization, we provide a qualitative visualization of the Cross-modal Harmonization (CMH) in Fig. 9. The fundamental objective of the CMH is to facilitate a flexible alignment strategy that avoids the structural degradation inherent in rigid feature imitation. As illustrated in the visualization, the shared representation s captures rich semantic information inherited from the RGB branch, maintaining a high level of discriminative detail. If the framework were to enforce a strict alignment between s and the auxiliary Depth modality directly, the high density information in s would inevitably be suppressed to conform to the geometric constraints of the depth data. To resolve this conflict, UniM2 introduces s_X as a modality specific transformation that is tailor made for the auxiliary input. By calculating the CMCS loss between s_X and the Depth modality rather than using s directly, our framework achieves a soft alignment. This mechanism allows the shared representation s to remain strictly aligned with the RGB manifold

Fig. 9: Qualitative analysis of the Cross-modal Harmonization. The figure displays the RGB image, the shared representation s, the auxiliary Depth modality, and the modality specific proxy s_X. This comparison highlights how s_X serves as a tailor made bridge for CMCS loss calculation, ensuring that s maintains rich semantic information while achieving robust alignment with the depth structure.

while simultaneously leveraging the unique structural cues provided by the depth information through the s_X proxy. The visual contrast between s and s_X confirms that CMH effectively preserves the semantic integrity of the master modality while successfully integrating auxiliary physical priors.

8 Superiority of STEGO within DINOv3

STEGO [8] is the most classic method in the field of unsupervised semantic segmentation, and many subsequent methods are developed on its basis. As the pioneer of the feature correspondence distillation paradigm, STEGO has established the fundamental architectural design followed by nearly all state-of-the-art USS frameworks [8, 9, 14, 15]. However, when migrating these methodologies from DINOv1 [3] to the more advanced DINOv3 [16] backbone, we observe a significant performance leap across all baseline methods, accompanied by a narrowing performance gap between them.

Specifically, under the powerful representations provided by DINOv3, the relatively simpler STEGO framework exhibits remarkably competitive and even superior results compared to more complex subsequent methods. As summarized in Tab. 3, on the Cityscapes dataset with a ViT-S/16 backbone, STEGO achieves

Table 3: Quantitative comparisons on COCO-Stuff [2] (left) and Cityscapes [4] (right). Upon migrating from DINOv1 to DINOv3, the classic STEGO [8] framework consistently achieves superior performance over more recent methods.

Method	Backbone	mIoU
DINOv1 [3]	ViT-S/8	14.4
+ TransFGU [21]	ViT-S/8	16.8
+ STEGO [8]	ViT-S/8	24.5
+ EAGLE [9]	ViT-S/8	27.2
+ IL2Vseg [14]	ViT-S/8	27.6
DINOv3	ViT-S/16	19.7
+ **STEGO** [8]	ViT-S/16	29.2
+ HP [15]	ViT-S/16	27.5
+ EAGLE [9]	ViT-S/16	**29.4**
DINOv3	ViT-B/16	21.5
+ **STEGO** [8]	ViT-B/16	**30.9**
+ HP [15]	ViT-B/16	29.6
+ EAGLE [9]	ViT-B/16	30.5

Method	Backbone	mIoU
DINOv1 [3]	ViT-S/8	13.7
+ STEGO [8]	ViT-S/8	17.6
+ HP [15]	ViT-S/8	18.4
+ EAGLE [9]	ViT-S/8	19.7
DINOv3	ViT-S/16	12.3
+ **STEGO** [8]	ViT-S/16	**19.0**
+ HP [15]	ViT-S/16	18.0
+ EAGLE [9]	ViT-S/16	18.3
DINOv3	ViT-B/16	12.6
+ **STEGO** [8]	ViT-B/16	**20.1**
+ HP [15]	ViT-B/16	18.2
+ EAGLE [9]	ViT-B/16	19.1

19.0% mIoU, which slightly outperforms HP (18.0%) and EAGLE (18.3%). A similar trend is observed on the COCO-Stuff dataset, where STEGO (29.2% mIoU) exceeds HP (27.5%) and remains highly comparable to EAGLE (29.4%).

This phenomenon suggests that DINOv3's inherent object-centric discriminative power and superior spatial granularity render certain sophisticated distillation heuristics redundant. For instance, while EAGLE [9] introduces object-centric structural priors to compensate for the limitations of earlier backbones, the latest DINOv3 already possesses inherent object-centric discriminative power. Consequently, adding additional structural guidance yields limited performance gain in this regime. **Since STEGO provides a more transparent and efficient foundation while achieving state-of-the-art results on advanced backbones, it serves as the most effective baseline for our study.**

References

1. Berlinet, A., Thomas-Agnan, C.: Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media (2011) 4
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR. pp. 1209–1218 (2018) 13
3. 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) 12, 13
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016) 13

5. 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) 7
6. 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) 7
7. 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
8. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. ICLR (2022) 1, 3, 4, 7, 9, 12, 13
9. 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) 1, 3, 7, 12, 13
10. Koenig, A., Schambach, M., Otterbach, J.: Uncovering the inner workings of stego for safe unsupervised semantic segmentation. In: CVPRW. pp. 3789–3798 (2023) 5
11. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal material segmentation. In: CVPR. pp. 19800–19808 (2022) 9, 10
12. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105. PMLR (2015) 4
13. 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) 1, 2, 3
14. 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) 12, 13
15. Seong, H.S., Moon, W., Lee, S., Heo, J.P.: Leveraging hidden positives for unsupervised semantic segmentation. In: CVPR. pp. 19540–19549 (2023) 12, 13
16. 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) 12
17. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV. pp. 443–450. Springer (2016) 3
18. Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. IEEE TPAMI (2025) 1, 2, 3
19. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000) 1, 5
20. 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) 6
21. Yin, Z., Wang, P., Wang, F., Xu, X., Zhang, H., Li, H., Jin, R.: Transfgu: a top-down approach to fine-grained unsupervised semantic segmentation. In: ECCV. pp. 73–89. Springer (2022) 13