UMM: A Unified Multi-Modal Model for Low-Level Vision Tasks with Dual-Driven Prompting

Ziqi Luo[1,*]
luozq2022@mail.sustech.edu.cn
Jinxiang Lai[4,*]
jlaial@connect.ust.hk
Ruitao Chen[1,*]
chenrt2022@mail.sustech.edu.cn
Jinyu Yang[1,†]
jinyu.yang96@outlook.com
Bin-Bin Gao[2]
csgaobb@gmail.com
Qiang Nie[2]
qiangnie@hkust-gz.edu.cn
Jun Liu[2]
junsenselee@gmail.com
Jinfan Wang[1,3]
wangjf@sustech.edu.cn
Feng Zheng[1]
zfeng02@gmail.com

[1] Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, China

[2] Youtu Lab
Tencent
Shenzhen, China

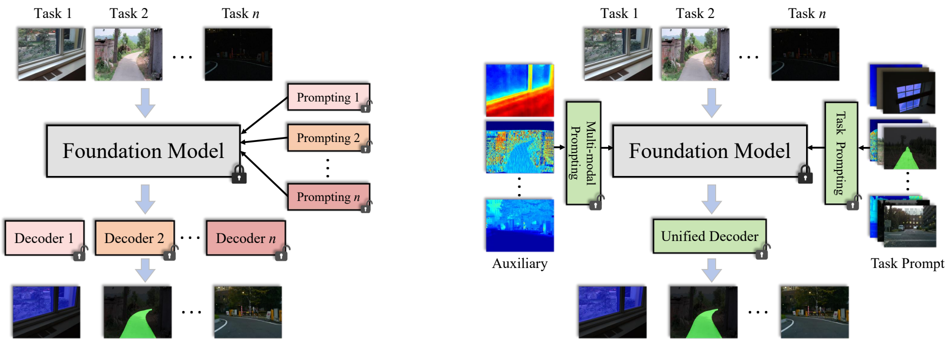
[3] Linkinsense (Anhui) Ltd.
Anhui, China

[4] Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China

Abstract

Current research lacks exploration of the unified multi-modal model for low-level vision tasks. Existing RGB-only unified models rely on multiple decoders or task-specific parameters, neglecting potential shared knowledge between tasks. In this paper, we propose a Unified Multi-modal Model (UMM) that handles diverse low-level tasks with fully shared parameters and a single decoder. The core innovation is the dual-driven prompting paradigm, combining multi-modal prompting (to enhance model robustness) and task prompting (to guide task-specific feature extraction). We also propose the Task-aware Fusion Module (TFM), which uses task prompts to direct multi-modal fusion, ensuring the model to focus more on task-relevant auxiliary information. Experiments demonstrate that UMM achieves competitive performance on various multi-modal low-level vision tasks, including RGB-T glass detection, RGB-T low-light enhancement, RGB-D salient object detection and RGB-N drivable area detection.

∗ Equally contribute to this work
† Corresponding author



(a) Traditional Visual Prompt (b) Dual-driven Prompts (Ours)

Figure 1: **Comparison of traditional methods and our dual-driven prompts.** Unlike traditional methods (a) which need to train separate prompts and decoders for each task, our method (b) combines multi-modal prompting and task prompting to introduce auxiliary features and guide task-specific feature extraction. All tasks share fully unified parameters.

1 Introduction

In recent years, multi-modal unified model has made significant progress in the field of vision-language [1, 13, 29, 38]. However, research on applying it to low-level vision tasks remains relatively scarce, which focus on processing raw pixel-level information in images, encompassing tasks such as binary image segmentation, restoration and enhancement. Although many studies [8, 15, 31] have shown that fusing RGB with visual modalities like depth and thermal can remarkably improve model performance, existing methods are computationally and storage expensive in multi-task scenarios, as specific networks need to be individually designed and fully fine-tuned end-to-end to achieve specific tasks, each of which requires maintaining an independent model copy.

Inspired by prompt learning in NLP [38], visual prompt learning [4, 12, 18, 22, 30, 55] has been explored. These methods fix large-scale pre-trained model parameters while adding learnable adapters for downstream tasks. By aligning input data with frozen model, the pre-trained knowledge and generalization of model can be retained, which is critical for multi-modal low-level vision fields where large datasets are scarce. Although these methods enable different tasks to share a backbone and reduce storage cost, separate prompt parameters and decoders need to be trained, keeping tasks isolated and preventing knowledge sharing. Moreover, most visual prompt methods rely on RGB-only data, limiting multi-modal use and model robustness. Multi-modal fusion in multi-task scenarios is another challenge. Traditional single-task methods rely on feature summation, concatenation or specially designed modules [14, 28, 40]. However, these approaches struggle in multi-task settings, as tasks focus differently on features during fusion, leading to underutilization of auxiliary modalities and potential interference with original features, degrading performance.

To address these issues, this paper proposes a novel dual-driven prompting paradigm. As illustrated in Fig. 1, unlike traditional visual prompt learning, our method considers both multi-modal and task prompting. Specifically, multi-modal prompting utilizes auxiliary modality features to generate additional embeddings to be added to the frozen foundation model, complementing RGB data. In task prompting, a set of given task-specific examples (including RGB, auxiliary modality images and labels) are used to interact with the query images, enabling model to learn associations between task prompts and queries for task-relevant feature extraction. Further, this paper proposes Task-aware Fusion Module (TFM). It guides multi-modal fusion using task prompts. Specifically, it includes two steps: task prompt alignment and cross fusion. The former aligns task prompt semantically, while the latter utilizes them to guide the interaction and fusion of RGB and auxiliary features under specific tasks. This allows model to better focus on key features in multi-task settings.

In this paper, we propose a Unified Multi-modal Model (UMM) framework for low-level vision tasks based on the proposed dual-driven prompting paradigm and TFM. Experiments show that UMM can excellently perform several multi-modal low-level vision tasks, including RGB-T (thermal) glass detection, RGB-T low-light enhancement, RGB-D (depth) salient object detection and RGB-N (surface normal) drivable area detection. The parameters for all these tasks are fully shared, no separate decoder or other extra parameters are needed.

Our contributions can be summarized as follows:

- We are the first to achieve a unified multi-modal model for low-level vision tasks, with all parameters fully shared across tasks.
- We propose TFM that guides the multi-modal fusion by utilizing task prompts, thereby achieving more efficient performance in a multi-task model.
- We experimentally demonstrate that the proposed unified model exhibits outstanding performance in multiple multi-modal low-level vision tasks.

2 Related Work

Multi-modal low-level learning. Multi-modal low-level learning aims to integrate information from different modalities to obtain richer data representations. Numerous recent studies [6, 14, 25, 42, 45, 47, 52] have confirmed its great potential. For example, DCF [21] proposes a strategy of cross reference fusion between calibrated depth map and RGB to assist salient object detection. ERDENet [37] enhances camouflaged object detection via RGB-D fusion and multi-modal refinement. DVN [23] leverages structure inconsistency to guide RGB-T fusion to improve image denoising. Moreover, in cross-modal super-resolution reconstruction, CMSR [39] adopts a coarse-to-fine strategy to achieve precise modality alignment. MaGIC [46] proposes to fuse RGB and multiple auxiliary modal via gradient guidance in the latent space, achieving flexible image completion. However, these methods usually lack flexibility for multi-modal fusion in multi-task scenarios. This paper proposes TFM, which guides multi-modal fusion by task prompts, enabling model to dynamically adjust the key features required for different low-level vision tasks.

Visual prompt learning. Prompt learning is initially proposed in NLP, with idea of guiding pre-trained models to accomplish specific tasks by providing textual prompts[27, 29, 38, 54]. This approach has been extended to computer vision recently [4, 18]. For example, VPT[22] adds a set of learnable embedding vectors as prompt to frozen encoder and achieves performance beyond full fine-tuning on multiple tasks. EVP[30] proposes explicit visual prompt, which takes high-frequency components of input as prompt and achieves excellent performance on low-level structure segmentation. ProTrack [44] firstly introduces prompt learning into multi-modal tracking by taking auxiliary modalities as prompt then simply weighting and summing RGB with auxiliary modalities. ViPT[55] improves ProTrack by performing a

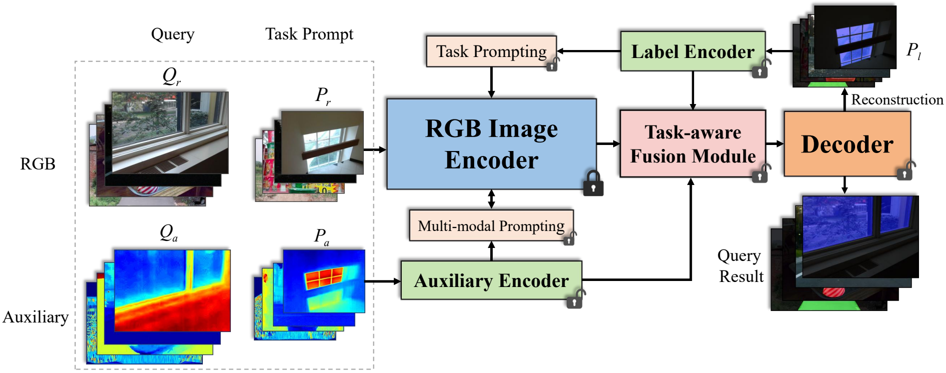


Figure 2: **UMM architecture.** The RGB Image Encoder is frozen, and is prompted with multi-modal and task information to extract task-specific features. Then, the Task-aware Fusion Module (TFM) fuses RGB and auxiliary features with the guidance of task prompt information. Finally, the query result is obtained by a unified Decoder. Our method allows different multi-modal low-level vision tasks to be implemented in a single model.

series of convolution on auxiliary modalities before adding them to foundation model. However, these methods require training separate prompt parameters. This paper proposes the dual-driven prompting paradigm, which combines multi-modal and task prompting. It enhances the robustness of model by introducing auxiliary information while enabling model to explore shared knowledge among different tasks.

3 Approach

3.1 Overall Framework

As presented in Fig. 2, the input data for UMM contains two groups: (1) Query group, includes the RGB image Q_r and auxiliary modality image Q_a to be predicted. (2) Task Prompt group, consists of the RGB image P_r, auxiliary modality image P_a, and their corresponding label P_l. The height and width of $\{Q_r, Q_a\}$ are resized to (H,W), while $\{P_r, P_a, P_l\}$ are resized to (H_2, W_2). We set $(H_2, W_2) < (H,W)$ for reducing the computation. UMM ultimately outputs the query result for the specified low-level vision task designated by task prompt.

We take RGB as the main branch and feed $\{Q_r, P_r\}$ into the frozen RGB Image Encoder. Here we use MAE [17] pre-trained ViT-Large model. Auxiliary Encoder takes $\{Q_a, P_a\}$ as input to obtain auxiliary feature $\{F_{qa}, F_{pa}\}$. Label Encoder takes P_l as input to obtain task prompt label feature F_{pl}. They are used for multi-modal and task prompting to introduce auxiliary modality and task information during feature extraction, *i.e.*, our proposed dual-driven prompting paradigm. We then fuse RGB and auxiliary features through TFM. Finally, a unified Decoder is employed to obtain query result for the task designated by the given task prompt. Moreover, we reconstruct task prompt label with the aim of preventing task prompt information from being lost during the interaction with query.

3.2 Dual-Driven Prompting

Previous visual prompt learning methods [4, 22, 30, 55] required training a set of separate additional parameters for each task. Differently, this paper proposes a dual-driven prompting paradigm that combines both multi-modal and task prompting, of which the structure is presented in Fig. 3 (a). Let $\{P_m, P_t\}$ denote the multi-modal and task prompting function, respectively. The process of dual-driven prompting is as follows:

$$\begin{aligned} \hat{F}^i_{qr} &= F^i_{qr} + P^i_m(F^i_{qr}, F_{qa}) \\ \hat{F}^i_{pr} &= F^i_{pr} + P^i_m(F^i_{pr}, F_{pa}) \cdot \beta^i + P^i_t(F_{pl}) \cdot \gamma^i \\ i &= 1, 2, ..., N \end{aligned} \tag{1}$$

where $\{F^i_{qr}, F^i_{pr}\}$ denote the query and task prompt RGB image features output by the $(i-1)^{th}$ layer transformer block of RGB Image Encoder. N denotes its number of transformer layers. β and γ are learnable scaling factors that adaptively adjust the intensity of the two prompting. Since auxiliary and label features are spatially aligned with RGB, we add the results of prompting to RGB branch in the form of residuals. Then we concatenate $\{\hat{F}^i_{qr}, \hat{F}^i_{pr}\}$ and feed them into the i^{th} layer transformer block of the frozen RGB Image Encoder IE:

$$\left[F^{i+1}_{pr}, F^{i+1}_{qr}\right] = IE^i(\left[\hat{F}^i_{pr}, \hat{F}^i_{qr}\right]), \quad i = 1, 2, ..., N \tag{2}$$

The purpose of multi-modal prompting is to introduce auxiliary to complement RGB enhancing representation robustness. To this end, we generate an attention map to enhance the crucial auxiliary features through spatial attention SA. P_m can be expressed as:

$$P^i_m(F^i_{kr}, F_{ka}) = C^i_m(SA^i(\left[F^i_{kr}, F_{ka}\right]) * F^i_{kr}), \quad i = 1, 2, ..., N \tag{3}$$

where k for p or q, referring to the feature of task prompt or query. C_m is a set of 1×1 convolution operations.

The purpose of task prompting is to introduce task information, enabling the model to continuously learn the association between given task prompt and query, so as to help model to better understand current task and extract its specific features. P_t can be expressed as:

$$P^i_t(F_{pl}) = C^i_t(F_{pl}), \quad i = 1, 2, ..., N \tag{4}$$

where C_t is the same convolution operation as C_m. This dual-driven prompting paradigm introduces multi-modal and task information while adjusting the input data to adapt the frozen RGB image encoder, thus preserving its knowledge and generalization. Since the parameters of all tasks is shared, this facilitates the exploitation of their shared knowledge. Finally, the RGB Image Encoder outputs the query and task prompt RGB image features $\{F_{qr}, F_{pr}\}$.

3.3 Task-Aware Fusion Module

Different tasks focus on distinct features in multi-modal fusion process. Unlike previous works [6, 7, 21, 28] for single-task multi-modal fusion, we propose Task-aware Fusion Module (TFM) to integrate RGB feature $\{F_{qr}, F_{pr}\}$ and auxiliary feature $\{F_{qa}, F_{pa}\}$. TFM use task prompt to guide model to fusion based on specific downstream low-level vision tasks, which consists of two steps, including task prompt alignment and cross fusion.

Task Prompt Alignment (TPA). Although task prompt feature $\{F_{pl}, F_{pa}, F_{pr}\}$ are spatially aligned, they are not semantically aligned because they are obtained from different encoders.

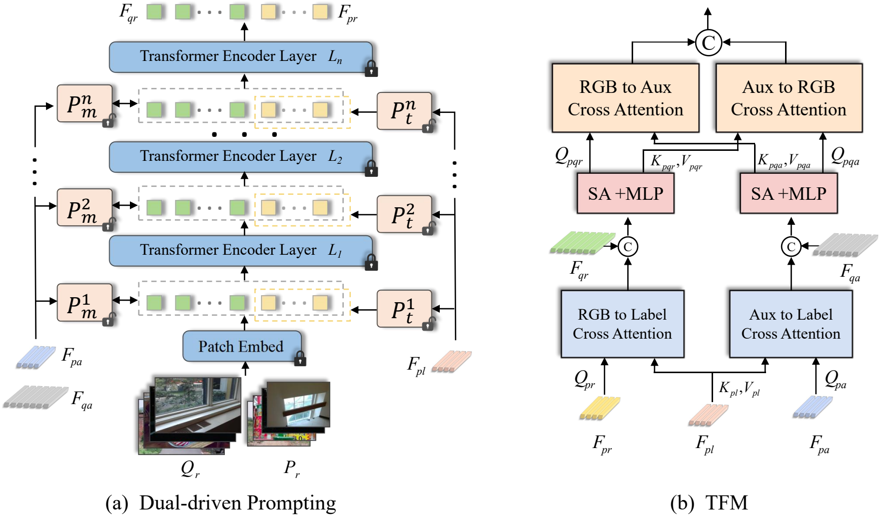


Figure 3: **The architecture of Dual-driven Prompting and TFM.** (a) The multi-modal and task information is introduced into RGB branch through dual-driven prompting. (b) TFM guides the multi-modal fusion of RGB and auxiliary modality through task prompts (Aux means auxiliary).

We take F_{pl} as the reference feature and utilize cross attention to establish semantic associations between them:

$$
\begin{aligned}
G_{pl} &= K_{pl}^T V_{pl} \\
F_{pr} &= Q_{pr} Softmax(G_{pl}) \\
F_{pa} &= Q_{pa} Softmax(G_{pl})
\end{aligned}
\tag{5}
$$

where $\{K_{pl}, V_{pl}\}$ are generated by F_{pl}. $\{Q_{pr}, Q_{pa}\}$ are generated by $\{F_{pr}, F_{pa}\}$ respectively. G_{pl} is the global context vector.

Cross Fusion (CF). After semantic alignment, we obtain RGB feature F_{pqr} and auxiliary feature F_{pqa} by concatenating $\{F_{pr}, F_{qr}\}$ and $\{F_{pa}, F_{qa}\}$ respectively. We enhance their contextual connections through Self Attention (SA) and MLP layer, with the aim of reinforcing the model's understanding of specific tasks. Then we interact RGB and auxiliary feature through cross attention:

$$
\begin{aligned}
G_{pqr} = K_{pqr}^T V_{pqr}, &\quad G_{pqa} = K_{pqa}^T V_{pqa} \\
\hat{F}_{pqr} &= Q_{pqr} Softmax(G_{pqa}) \\
\hat{F}_{pqa} &= Q_{pqa} Softmax(G_{pqr})
\end{aligned}
\tag{6}
$$

Finally, we employ a linear layer to integrate $\{\hat{F}_{pqr}, \hat{F}_{pqa}\}$, obtaining multi-modal feature that will be fed into the unified Decoder to get the query results.

Task		RGB-T glass detection	RGB-T low-light enhancement		RGB-D salient object detection				RGB-N drivable area detection
Metrics		MAE↓	PSNR↑	SSIM↑	E_ξ ↑	F_β^ω ↑	F_β ↑	MAE↓	mIoU↑
RGB-only	VPT [22]	0.066	27.01	0.8439	0.947	0.912	0.924	0.030	49.3
	Adaptformer [4]	0.060	27.32	0.8526	0.955	0.926	0.937	0.024	47.1
Multi-modal	ProTrack [44]	0.075	25.13	0.8233	0.941	0.890	0.909	0.037	79.2
	ViPT [53]	0.053	**27.85**	**0.8537**	0.954	0.927	0.938	0.023	67.0
	Ours(ST)	0.053	27.26	0.8451	0.959	0.930	0.938	0.021	69.2
	Ours(JT)	**0.035**	27.45	0.8468	**0.967**	**0.948**	**0.956**	**0.019**	**93.5**

Table 1: Comparison of our model with other visual prompt learning models on four representative low-level vision tasks. (ST: Separate Training, JT: Joint Training)

4 Experiments

4.1 Dataset

RGB-T glass detection. The dataset proposed by Hou *et al.* [19] contains 5551 RGB-T image pairs at the size of 640×480, covering 40 scenes (7 without glass), split into 4427/1124 for training/testing. We adopt MAE metric for evaluation.
RGB-T low-light enhancement. MFNet [14] dataset containing 1569 RGB-T image pairs at the size of 640×480, 820 captured in daylight. We adjust their brightness by gamma transform and random Gaussian noise to simulate low-light environments. They are split into 574/246 for training/testing. We use PSNR and SSIM as evaluation metrics.
RGB-D salient object detection. DUTLF-Depth [35] dataset contains 1200 RGB-D image pairs at the size of 600×400, splitting into 800/400 for training/testing. We use E-measure (E_ξ), weighed F-measure (F_β^ω), F-measure (F_β) and MAE as evaluation metrics.
RGB-N drivable area detection. ORFD [32] dataset includes 12198 frames of LiDAR point cloud and image pairs at the size of 1280×720, splitting into 8398/1245/2555 for training/validation/testing. We use the method proposed by Fan *et al.* [11] to calculate the surface normal from depth map that obtained from point cloud projection. We adopt mIoU metric for evaluation.

4.2 Implementation Details

We train our model on 8 Tesla V100 GPUs. For data augmentation, we use random resize cropping with a scale range of [0.5, 1.5] and horizontal flipping during training. The query and task prompt images are resized to 592×448 and 448×336 respectively. The optimizer is AdamW [24] with momentum terms (0.9, 0.999), weight decay 0.01. The initial learning rate is set to 1e-4 and is scheduled by the poly strategy with power 0.95. We train the model for 500 epochs and warm up the first 20 epochs. The batch size is set to 8. The sampling weight for each task is 0.35(RGB-T glass detection), 0.15(RGB-T low-light enhancement), 0.1(RGB-D salient object detection), 0.4(RGB-N drivable area detection) respectively.

4.3 Main Results

Comparison with joint and separate training. As shown in Tab. 1, we compare the results of separate and joint training for each task in the same experiment settings. The results show that the four low-level vision tasks perform better under joint training than separate

Method	Images with glass	Images without glass	All images
RTFNet [41]	0.068	0.188	0.058
ShapeConv [2]	0.059	0.019	0.054
CMX [28]	0.031	0.004	0.029
Segformer [43]	0.052	0.063	0.053
UCNet [48]	0.079	0.028	0.071
HDFNet [33]	0.055	0.019	0.048
RD3D [3]	0.064	0.037	0.045
DCF [21]	0.059	0.036	0.056
GSNet [19]	**0.027**	**0.003**	**0.024**
Ours	0.037	0.011	0.035

Table 2: Quantitative evaluation on the RGB-T glass detection dataset on MAE.

Method	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$
BBSNet [9]	0.833	0.663	0.774	0.120
DMRA [34]	0.927	0.858	0.883	0.048
SSF [50]	0.946	0.894	0.914	0.034
A2dele [36]	0.924	0.864	0.890	0.043
FRDT [51]	0.941	0.878	0.902	0.039
DANet [53]	0.925	0.847	0.884	0.047
HDFNet [33]	0.934	0.865	0.892	0.040
CoNet [20]	0.947	0.896	0.908	0.034
PGAR [8]	0.944	0.889	0.914	0.035
ATSA [49]	0.947	0.901	0.918	0.032
DCF [21]	0.952	0.909	0.926	0.030
Ours	**0.967**	**0.948**	**0.956**	**0.019**

Table 3: Quantitative evaluation on DUTLF-Depth dataset.

Method	FuseNet [16]	SNERoadSeg [10]	OFFNet [52]	RTFNet [41]	MFNet [14]	RoadFormer [26]	Ours
mIoU$\uparrow$	66.0	81.2	82.3	90.7	81.7	92.5	**93.5**

Table 4: Quantitative evaluation on ORFD dataset.

training. In particular, the improvement of drivable area detection and glass detection under joint training is significant. This indicates that under our framework, instead of interfering with each other, joint training of different tasks can explore the potential connection and share knowledge among them, thus facilitating their learning effects.

Comparison with other visual prompting models. As shown in Tab. 1, our method outperforms other existing visual prompt learning methods on almost all tasks. This highlights the excellence of our proposed dual-driven prompting paradigm, which not only enhances the robustness of model by introducing auxiliary modalities, but also avoids isolation between tasks and independent training of prompt parameters. It enables the model to understand and adapt to specific tasks more precisely by learning the association between queries and task prompts under a unified framework.

Comparison with task-specific methods. Our method is quantitatively compared to task-specific methods on RGB-T glass detection, RGB-D salient object detection and RGB-N drivable area detection, which is as shown in Tab. 2, Tab. 3 and Tab. 4. Our method achieves competitive results on each dataset. In particular, our method achieves state-of-the-art performance on DUTLF-Depth and ORFD datasets. Although the best results are not achieved on the RGB-T glass detection dataset, it has outperformed the most task-specific models. Given that our model aims to achieve multi-task coordination, which undoubtedly increases complexity compared with models dedicated to a single task, we believe that achieving such results under a unified model framework is acceptable.

Visualization results. To demonstrate the performance of our model more intuitively, we provide visualization results for each task, as shown in Fig. 4. Our method achieves very accurate predictions on all four low-level vision tasks.

4.4 Ablation Study

Ablation of joint training. In Tab. 1, we observe significant performance improvements through joint training compared to separate training with the same number of iterations, particularly in drivable area detection and glass detection. To explore it further, we conduct joint training with various combinations of the three low-level segmentation tasks and report

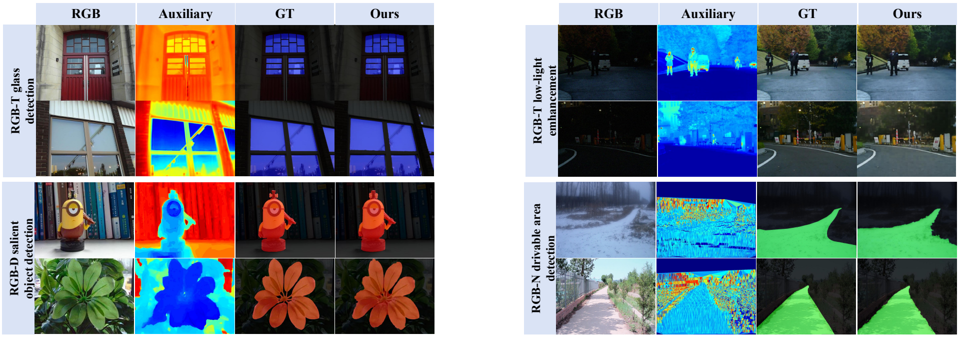


Figure 4: **Visualisation.**

Task	RGB-T glass detection	RGB-D salient object detection				RGB-N drivable area detection
Metrics	MAE↓	$E_\xi\uparrow$	$F^\omega_\beta\uparrow$	$F_\beta\uparrow$	MAE↓	mIoU↑
DAD	-	-	-	-	-	69.2
GD	0.053	-	-	-	-	-
DAD+SOD	-	0.964	0.942	0.951	0.020	88.5
GD+SOD	0.042	**0.970**	0.946	0.954	**0.019**	-
DAD+GD	**0.041**	-	-	-	-	93.1
DAD+GD+SOD	**0.041**	0.966	**0.947**	**0.956**	**0.019**	**93.5**

Table 5: Ablation study of joint training. (DAD: Drivable Area Detection, SOD: Salient Object Detection, GD: Glass Detection)

the performance changes in Tab. 5. The results indicate that when they are jointly trained in pairs, the model's performance exceeds that of training separately, with the combination of drivable area detection and glass detection showing the most substantial improvement. More notably, when all three tasks are jointly trained, the model's performance experiences further enhancement. This can be attributed to the mutual learning and knowledge sharing among different tasks within our unified model framework, compelling model to learn more generalizable feature representations. These findings show the efficacy of joint training strategy in enhancing model's comprehensive performance across multiple tasks.

Ablation of dual-driven prompting. We remove multi-modal or task prompting from our model and report results on joint training on RGB-T glass detection and RGB-D salient object detection. As shown in Tab. 6, regardless of the absence of multi-modal or task prompting, there is a significant decrease in the performance of our model. When both are removed, the decrease is more pronounced. This experimental result fully demonstrates the effectiveness of the dual-driven prompting paradigm we proposed. The combination of multi-modal and task prompting has significant potential in improving model performance. The former compensates for RGB and enhances model robustness while the latter allows the model to be more flexible in extracting features that are closely related to specific tasks.

Ablation of TFM. We remove TFM from our model and report results on joint training on RGB-T glass detection and RGB-N drivable area detection, as shown in Tab. 7. The results indicate that TFM improves model performance and is more effective compared to other fusion methods. Furthermore, we observe that certain fusion modules proposed in previous works [21] perform even worse than simple concatenation or summation. This may be due

Task	RGB-D salient object detection				RGB-T glass detection
Metrics	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	MAE↓	MAE↓
w/o Multi-modal prompting	0.962	0.937	0.945	0.021	0.045
w/o Task prompting	0.960	0.925	0.937	0.025	0.049
w/o Dual-driven prompting	0.927	0.876	0.900	0.041	0.061
w/ Dual-driven prompting(Ours)	**0.970**	**0.946**	**0.954**	**0.019**	**0.042**

Table 6: Ablation study of our dual-driven prompting paradigm.

Task	RGB-N drivable area detection	RGB-T glass detection
Metrics	mIoU ↑	MAE↓
Concatenation	91.1	0.047
Summation	91.9	0.045
CRM [21]	90.7	0.053
SA-Gate [6]	92.3	0.045
w/o TPA	92.6	0.043
w/o CF	87.8	0.045
w/o TFM	87.5	0.046
w/ TFM (Ours)	**93.1**	**0.041**

Table 7: Ablation study of TFM.

to the fact that they are designed only for single-task. In multi-task models, the focused features of different tasks vary considerably, thus interfering with fusion. By introducing task prompts to guide the multi-modal fusion process, we can effectively improve the fusion performance of the unified model.

5 Conclusion

In this work, we propose UMM, a unified multi-modal model for low-level vision tasks in which different tasks share all parameters. It adopts the proposed dual-driven prompting paradigm that combines multi-modal and task prompting to obtain complementary representations of RGB while allowing model to focus on task-specific features. Furthermore, we propose the TFM, which guides the multi-modal fusion of different tasks via task prompts. The final experiments show that UMM achieves competitive results in various multi-modal low-level vision tasks. We expect this study to receive more attention in the future.

Acknowledgement

This work is supported in part by Key-Area Research and Development Program of Guangdong Province (2020B0101130003) and the National Natural Science Foundation of China (Grant NO. 62122035).

References

[1] Eman Ali and Muhammad Haris Khan. Noise-tolerant few-shot unsupervised adapter for vision-language models. In *35th British Machine Vision Conference*, 2024.

[2] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. *IEEE/CVF International Conference on Computer Vision*, pages 7068–7077, 2021.

[3] Qian Chen, Ze Liu, Yi Xiong Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. *ArXiv*, abs/2101.10241, 2021.

[4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Neural Information Processing Systems Conference*, 2022.

[5] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *European Conference on Computer Vision*, volume 12353, pages 520–538. Springer, 2020.

[6] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *European Conference on Computer Vision*, volume 12356, pages 561–577. Springer, 2020.

[7] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4467–4473, 2021.

[8] Yu-Chun Ding, Chia-Yu Chang, Pei-Rong Li, Chao-Tsung Huang, Yung-Chen Lin, Tsung Chen, Wei-Lun Lin, Cheng-Ting Lee, Fan-Yi Lin, and Yuan-Hao Huang. High-fidelity depth map reconstruction system with rgb-guided super resolution CNN and cross-calibrated chaos lidar. *IEEE Access*, 13:19118–19131, 2025.

[9] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: RGB-D salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*, volume 12357, pages 275–292. Springer, 2020.

[10] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *European Conference on Computer Vision*, volume 12375, pages 340–356. Springer, 2020.

[11] Rui Fan, Hengli Wang, Bohuan Xue, Huaiyang Huang, Yuan Wang, Ming Liu, and Ioannis Pitas. Three-filters-to-normal: An accurate and ultrafast surface normal estimator. *IEEE Robotics and Automation Letters*, 6(3):5405–5412, 2021.

[12] Yufei Gao, Bin Fu, Lei Shi, Chengming Liu, and yucheng shi. Glpi: A global layered prompt integration approach for explicit visual prompt. In *35th British Machine Vision Conference*, 2024.

[13] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790, 2023.

[14] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Y. Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. *International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017.

[15] Jinyu Han, Mengyin Wang, Weiyi Wu, and Xu Jia. Perceptual localization and focus refinement network for RGB-D salient object detection. *Expert Syst. Appl.*, 259: 125278, 2025.

[16] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Asian Conference on Computer Vision*, volume 10111, pages 213–228. Springer, 2016.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15979–15988. IEEE, 2022.

[18] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10878–10887. IEEE, 2023.

[19] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *IEEE Transactions on Image Processing*, 32:1911–1926, 2022.

[20] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *European Conference on Computer Vision*, volume 12363, pages 52–69. Springer, 2020.

[21] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated RGB-D salient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9471–9481. Computer Vision Foundation / IEEE, 2021.

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, volume 13693, pages 709–727. Springer, 2022.

[23] Shuangping Jin, Bingbing Yu, Minhao Jing, Yi Zhou, Jiajun Liang, and Renhe Ji. Darkvisionnet: Low-light imaging via RGB-NIR fusion with deep inconsistency prior. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 1104–1112. AAAI Press, 2022.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[25] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[26] Jiahang Li, Yikang Zhang, Peng Yun, Guangliang Zhou, Qijun Chen, and Rui Fan. Roadformer: Duplex transformer for rgb-normal semantic road scene parsing. *CoRR*, abs/2309.10356, 2023.

[27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10965. IEEE, 2022.

[28] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *ArXiv*, abs/2203.04838, 2022.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499, 2023.

[30] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445. IEEE, 2023.

[31] Jincheng Luo, Yongjun Li, Bo Li, Xinru Zhang, Chaoyue Li, Zhimin Chenjin, Jingyi He, and Yifei Liang. Transformer-based cross-modality interaction guidance network for RGB-T salient object detection. *Neurocomputing*, 600:128149, 2024.

[32] Chen Min, Weizhong Jiang, Dawei Zhao, Jiaolong Xu, Liang Xiao, Yiming Nie, and Bin Dai. ORFD: A dataset and benchmark for off-road freespace detection. In *IEEE International Conference on Robotics and Automation*, pages 2532–2538. IEEE, 2022.

[33] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *European Conference on Computer Vision*, 2020.

[34] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. *IEEE/CVF International Conference on Computer Vision*, pages 7253–7262, 2019.

[35] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *International Conference on Computer Vision* , pages 7253–7262. IEEE, 2019.

[36] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9057–9066, 2020.

[37] Miao Qi, Zheng Wang, Cheng Liu, Yan Zhou, and Meijun Sun. Joint edge and regional depth enhancement network for camouflaged object detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[39] Guy Shacht, Dov Danon, Sharon Fogel, and Daniel Cohen-Or. Single pair cross-modality super resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6374–6383, 2021.

[40] Ukcheol Shin, Kyunghyun Lee, and In-So Kweon. Complementary random masking for rgb-thermal semantic segmentation. *ArXiv*, abs/2303.17386, 2023.

[41] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4: 2576–2583, 2019.

[42] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021.

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *ArXiv*, abs/2105.15203, 2021.

[44] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *ACM International Conference on Multimedia*, pages 3492–3500. ACM, 2022.

[45] Xiaowen Ying and Mooi Choo Chuah. Uctnet: Uncertainty-aware cross-modal transformer network for indoor RGB-D semantic segmentation. In *European Conference on Computer Vision*, volume 13690 of *Lecture Notes in Computer Science*, pages 20–37. Springer, 2022.

[46] Yongsheng Yu, Hao Wang, Tiejian Luo, Heng Fan, and Libo Zhang. Magic: Multi-modality guided image completion. *arXiv preprint arXiv:2305.11818*, 2023.

[47] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147. IEEE, 2023.

[48] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8588, 2020.

[49] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *European Conference on Computer Vision*, 2020.

[50] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3469–3478, 2020.

[51] Miao Zhang, Yu Zhang, Yongri Piao, Beiqi Hu, and Huchuan Lu. Feature reintegration over differential treatment: A top-down and adaptive fusion network for rgb-d salient object detection. *ACM International Conference on Multimedia*, 2020.

[52] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2633–2642. Computer Vision Foundation / IEEE, 2021.

[53] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *European Conference on Computer Vision*, 2020.

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021.

[55] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526. IEEE, 2023.