Taming Cascaded Mixture-of-Experts for Modality-missing Multi-modal Salient Object Detection

Kunpeng Wang*, **Feifan Sun, Keke Chen**

School of Computer Science and Technology, Anhui University, Hefei, 230601, China
{kp.wang, chenk1220}@foxmail.com, sunfeifan1216@163.com

Abstract

Multi-modal Salient Object Detection (SOD) shows an improvement over its uni-modal counterpart by exploiting the complementary benefits between modalities. However, this improvement relies on complete multi-modal information, which is difficult to be guaranteed in practice due to sensor failures and transmission errors. To address this issue, we propose a robust multi-modal SOD framework that enhances the adaptability to modality-missing conditions, while maintaining comparable performance in the modality-complete condition. Nevertheless, flexibly handling modality-missing and modality-complete cases and integrating their corresponding multi-modal features in a unified framework is non-trivial. To this end, we achieve this framework by designing a Cascaded Mixture-of-Experts (CMoE) network that sequentially incorporates missing-aware and multi-modal MoE. Specifically, the missing-aware MoE employs three modality-reconstruction experts with a soft router to adaptively reconstruct feature representations for both missing and available modalities, assisted by an expert modulation loss that guides the router to assign expert weights according to missing conditions. The multi-modal MoE adopts two homogeneous uni-modal experts with learned modality-specific knowledge tailored for integrating modality features, which are dynamically combined via the soft router. The cascaded architecture fully empowers CMoE with the flexibility across varying input cases. Extensive experiments on modality-missing and modality-complete conditions demonstrate the effectiveness of the proposed method.

Code — https://github.com/Angknpng/CMoE

Introduction

The availability of multiple sensors such as RGB, Depth (D), and Thermal (T) has encouraged the multi-modal (i.e., RGB-D and RGB-T) representation fusion for scene understanding tasks like Salient Object Detection (SOD) (Wu et al. 2025a,b), which aims to identify and segment the most appealing region(s) in a scene. With additional spatial structure or object shape information from depth or thermal modalities, multi-modal SOD shows an improvement over its uni-modal counterpart and plays an important role in broad applications, such as audio-visual analysis (Min et al. 2020), text-image retrieval (Delmas et al. 2022), tracking (Lu et al. 2025b,c), and medical analysis (Fan, Liu, and Zhang 2024).

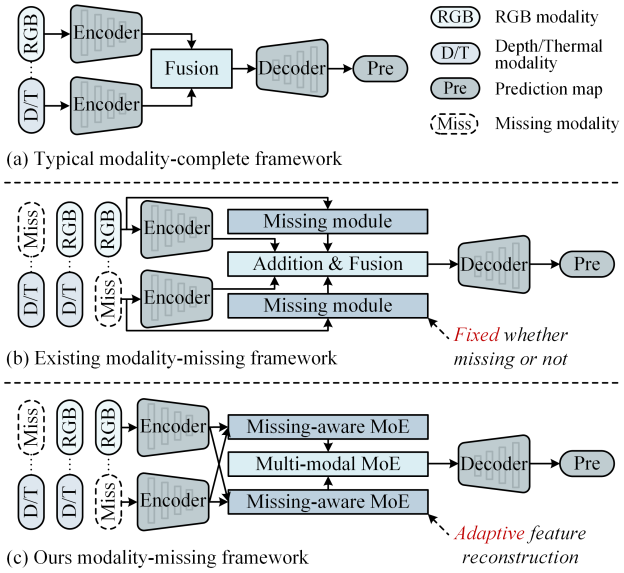

Figure 1: Workflow comparison of multi-modal SOD frameworks. (a) relies on complete modalities for fusion, but difficult to cope with missing modalities. (b) adapts to modality-missing cases, but lacks flexibility to handle missing and complete cases. (c) reconstructs modality features based on input conditions to adaptively handle both cases.

Most existing multi-modal SOD methods (Pang et al. 2023; Wang et al. 2025b) implicitly assume that all modalities are always available, as shown in Fig. 1 (a), and design diverse modality-complete fusion strategies (Zhou et al. 2021) to exploit multi-modal complementary information for accurate prediction. However, due to practical factors such as sensor failures, security or privacy concerns, and transmission errors that make data for certain modalities unavailable, the assumption may be difficult to satisfy during the inference phase, even though all modalities are available during the training phase. It turns out that in the absence of any modality, these methods struggle to maintain their original performance and even degrade drastically, as illustrated in Fig. 2. This indicates that existing methods typi-

*Corresponding author.

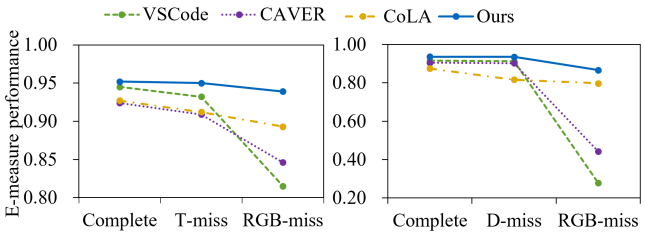

Figure 2: Comparison of the proposed CMoE with state-of-the-art methods on RGB-T (left) and RGB-D (right) datasets under modality-complete and modality-missing conditions.

cally have difficulty in dealing with modality-missing situations, which is a realistic challenge. Therefore, the development of a model capable of handling both modality-complete and modality-missing situations is of great practical significance.

Recently, CoLA (Hao, Zhong, and Tang 2024) attempts to address this problem by improving the training strategy. It freezes the trained modality-complete model and introduces additional missing modules whose parameters are learned using dropout inputs to adapt to the modality-missing situation, as shown in Fig. 1 (b). Despite improving the model performance with missing modalities, CoLA still suffers from two issues: 1) It only uses the zero strategy that creates a zero-valued matrix as the missing input for adaption, failing to effectively represent the missing modality and make full use of the available modality. 2) Its missing modules are fixed, failing to perform dynamic adjustments to meet the distinct requirements of modality-missing and modality-complete conditions. This is validated by the smoothing but moderate performance of CoLA in Fig. 2.

In this work, we propose a novel Cascaded Mixture-of-Experts (CMoE) framework, which sequentially incorporates a missing-aware MoE (MaMoE) and a mluti-modal MoE (MmMoE) to flexibly address multi-modal SOD under both modality-missing and modality-complete conditions, as illustrated in Fig. 1 (c). To tackle the above issues, MaMoE adaptively reconstructs the feature representation for each modality by leveraging the available modality appropriately. Specifically, we first introduce three modality-reconstruction experts, each with distinct roles: a zero expert for discarding, a copy expert for replication, and an alter expert for replacing the input feature with a trainable vector. The outputs of these experts are dynamically weighted and combined through a soft router. Second, we introduce an expert modulation loss, which uniformly guides the soft router to reconstruct the desired features for each modality according to missing conditions (i.e., missing or available). In particular, the router is directed to mainly select the copy and alter experts from the other available modality if the current modality is missing, and to mainly select the zero expert if it is available, so that the missing modality feature can be complemented without interfering with the available modality feature. Subsequently, MmMoE introduces two homogeneous uni-modal experts, each pre-trained with modality-specific knowledge, to discriminatively integrate the reconstructed modality features, which are then complementarily fused under the probabilistic guidance of the soft router. Through the collaboration of the cascaded expert groups, CMoE makes full use of the available modalities and achieves robust multi-modal representations for saliency prediction. Fig. 2 also shows the overall superior performance and slight degradation of our model, suggesting that it is little affected by modality-missing conditions.

The main contributions can be summarized as follows:

- We propose CMoE, a novel and flexible framework for multi-modal SOD, which adopts a cascaded mixture-of-experts architecture to uniformly handle both modality-missing and modality-complete input cases.
- We design a missing-aware MoE (MaMoE) that adaptively compensates for missing modality features without interfering with available modality features, thus enabling seamless feature reconstruction.
- We design a multi-modal MoE (MmMoE) that utilizes modality-specific knowledge from uni-modal experts to dynamically integrate multi-modal features.
- Experiments on six datasets under both modality-missing and modality-complete conditions demonstrate the effectiveness of CMoE, with minimum average performance degradation in modality-missing settings, emphasizing the robustness and potential of the proposed framework.

Related Work

Multi-modal Salient Object Detection

Multi-modal Salient Object Detection (SOD) aims to enhance traditional RGB-based SOD by incorporating the additional modality such as depth or thermal images, which provide additional spatial or shape information of objects. This enhancement significantly improves the detection performance in complex scenarios involving cluttered backgrounds or varying illumination. Existing methods mainly focus on exploring complementary information between modalities using elaborate fusion strategies, which can be generally categorized into early, middle, and late fusion (Zhou et al. 2021). Early fusion combines two modalities at the input stage to directly enrich the feature representation (Qu et al. 2017; Zhang et al. 2024). Middle fusion is widely used and integrates features at intermediate layers, achieving rich interactions between modalities (Hu et al. 2024; Luo et al. 2024; Wang et al. 2025a; He et al. 2025). Late fusion merges modality-specific features in the decision phase, typically combining individual modality predictions for final output (Chen et al. 2020; Han et al. 2017). However, most existing methods are designed based on the assumption of complete modality inputs, making it difficult to cope with modality-missing conditions in reality. The recent method CoLA (Hao, Zhong, and Tang 2024) alleviates this issue through a conditional dropout training strategy. Although adapted to modality-missing conditions, its fixed network structure limits the flexibility of handling both modality-missing and modality-complete cases. To this end, we propose CMoE to adaptively handle modality-missing and modality-complete conditions through a cascaded mixture-of-experts design.

Modality-missing Learning

Modality-missing learning addresses the challenge of one or more input modalities being missing or unavailable, which occurs frequently in the real world. This issue commonly arises in multi-modal tasks involving multiple modalities (e.g., RGB, thermal, audio, text), all of which may not always be accessible due to sensor failure, privacy preservation, and deployment limitation. To address this challenge, existing methods attempt to compensate for missing modalities using strategies such as zero-valued padding (Lee et al. 2023; Hao, Zhong, and Tang 2024), feature copying (Yao et al. 2024; Xu, Jiang, and Liang 2024), or feature generation (Lu et al. 2025a; Ramazanova et al. 2025). However, these methods almost rely on a single compensation strategy, limiting their capacity to reconstruct feature representations. Instead, our method incorporates these three strategies and dynamically integrates them to comprehensively represent missing modality features.

Mixture-of-Experts Architecture

The Mixture-of-Experts (MoE) architecture has emerged as an effective approach to increase the capacity of deep neural networks while maintaining computational efficiency. Initially popularized in natural language processing (Shazeer et al. 2016; Lepikhin et al. 2020), MoE has recently gained attention in computer vision tasks such as image restoration (Guo et al. 2024a; Ai, Huang, and He 2024), multi-modal fusion (Xu, Jiang, and Liang 2024; Zong et al. 2024), and representation learning (Ben-Shabat et al. 2024; Jin et al. 2024). The vanilla MoE architecture typically consists of multiple identical expert models and a router for assigning expert weights. Although effective, the identical experts struggle to facilitate the model to learn diverse representations of the same input. In this work, we extend the MoE paradigm by introducing specialized modality-reconstruction experts and uni-modal experts to achieve flexible and robust multi-modal salient object detection.

Method

In this work, we propose CMoE to address multi-modal salient object detection (SOD) in practical scenarios with missing modalities. Unlike prior methods that rely on fixed network structures to handle either modality-complete or modality-missing cases, CMoE provides a flexible framework capable of adaptively dealing with both cases through a cascaded Mixture-of-Experts (MoE) architecture.

Problem Definition. Given a multi-modal input X with RGB (i.e., x^{rgb}) and depth or thermal (i.e., $x^{d/t}$) modalities, the training process for multi-modal SOD can be summarized as $f_\theta^* \leftarrow \{f_\theta; X, Y\}$, where Y is the ground truth. Due to modality missingness, there are three distinct input cases: 1) $X = \left\{x^{rgb}, x^{d/t}\right\}$ denotes the modality-complete case, 2) $X = \left\{x^{rgb}, \emptyset\right\}$ denotes the depth or thermal modality-missing case, 3) $X = \left\{\emptyset, x^{d/t}\right\}$ denotes the RGB modality-missing case, where $\emptyset$ indicates the missing modality with a zero-valued input. Note that in general incomplete multi-modal learning settings, all modalities are available during training, and missingness typically occurs at test time.

Overall Framework. Fig. 3 presents the framework of the proposed CMoE, which randomly receives one of the three possible input cases described above. CMoE consists of four main components: a dual encoder for modality feature (i.e., $F^m = \{F_1^m, F_2^m, \ldots, F_4^m\}, m \in \{rgb, d/t\}$) extraction, a missing-aware MoE (MaMoE) for reconstructing features of each modality based on missing conditions, a multi-modal MoE (MmMoE) for integrating and combining features from both modalities, and a decoder that aggregates features across layers to predict the final saliency map S. The structure of encoder and decoder follows the recent advanced works (Liu et al. 2022; Wang et al. 2024a) without complex interactions, allowing a clear demonstration of the efficacy of the proposed MaMoE and MmMoE components.

Missing-aware MoE (MaMoE)

Existing multi-modal SOD methods are almost designed for modality-complete input pairs, while overlooking practical modality-missing conditions in real-world applications. The recent method (Hao, Zhong, and Tang 2024) attempts to adapt to modality-missing input cases, however, it lacks the flexibility to handle both missing and complete conditions. To overcome this limitation, we propose the MaMoE that adaptively reconstructs each modality by preserving the features of the available modality and compensating for the features of the missing modality.

As illustrated in Fig. 3, the MaMoE for each modality consists of three modality-reconstruction experts, $\{E_{zero}, E_{copy}, E_{alter}\}$, along with a soft router $\mathcal{G}$. Each expert is essential for coping with the missing or complete conditions, while maintaining a simple and effective structure. The soft router dynamically assigns weights to each expert conditioned on the input feature. Formally, given an extracted feature F^m, the adapted feature F_{ada}^m is computed by the MaMoE through a weighted summation of the outputs from the modality-reconstruction experts:

$$F_{ada}^m = \sum_{i \in \mathcal{E}} w_i \cdot E_i(F^m), \quad \mathcal{E} = \{zero, copy, alter\}, \quad (1)$$

where $w_i = \mathcal{G}\left(F^m\right)$ denotes the weight assigned to each expert by the soft router, ensuring $\sum_{i \in \mathcal{E}} w_i = 1$. Then, the reconstructed feature F_{rec}^m is obtained by combining the adapted feature from the other modality with the extracted feature of the current modality:

$$F_{rec}^m = F_{ada}^{\bar{m}} + F^m, \quad \bar{m} \in \{rgb, d/t\} \setminus \{m\}. \quad (2)$$

Furthermore, an expert modulation loss $\mathcal{L}_{EM}$ is introduced to guide the soft router in assigning appropriate weights to each expert based on the missing condition (i.e., missing or available) of the current modality.

Zero Expert. The zero expert is crucial in the situation where the current modality is available. It performs a discard operation by directly setting the input feature to zero, formulated as:

$$E_{zero}\left(F^m\right) = \mathbf{0}. \quad (3)$$

In essence, the presence of the zero expert can invalidate the MaMoE output from the other modality to not disturb the

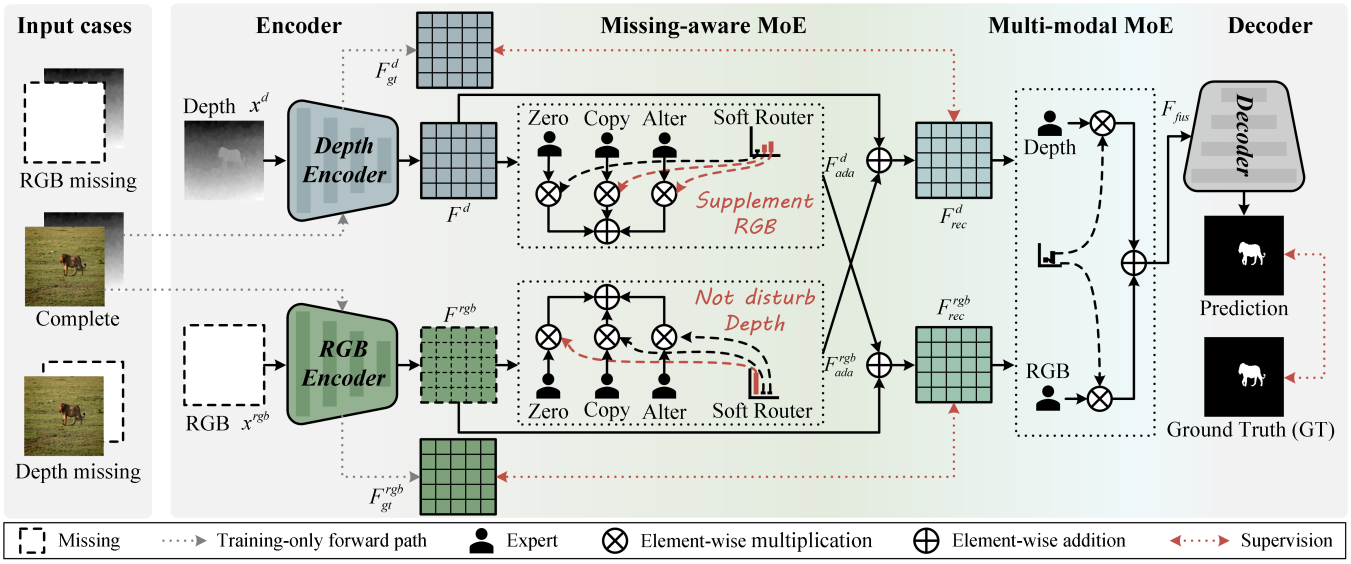

Figure 3: Overview of the proposed CMoE framework for both modality-missing and modality-complete conditions. Features from RGB and depth/thermal image pairs are first processed by the missing-aware MoE to adaptively reconstruct features for each modality based on its missing condition. The reconstructed features are supervised by the corresponding complete features to ensure consistency. Then, the multi-modal MoE dynamically integrates both modalities for final saliency prediction.

original extracted features of the current modality. Specifically, when the current modality is available, the missing condition of the other modality is uncertain (e.g., missing), which might introduce features with heavy noise as input to MaMoE. In such case, the potentially unreliable information from the other modality can be suppressed by assigning larger weights to the zero expert, as shown in the middle bottom of Fig. 3. Thus, the zero expert provides flexibility for MaMoE, so that the contributions of the two components in reconstructed features can be regulated, as in Eq. 2.

Copy Expert. The copy expert is to deal with the missing situation. Inspired by the residual connection (He et al. 2016), it directly passes the input feature to the output without extra operations, which can be formulated as:

$$E_{copy}\left(F^m\right) = F^m. \tag{4}$$

Intuitively, the copy expert enables the model to fully leverage the information from the available modality. Specifically, if the current modality is missing, the other modality is definitely available and can complement the missing information for further exploration, enhancing feature representation and prediction.

Alter Expert. The alter expert is also for the missing situation but is more flexible. It introduces a learnable vector v and a weight matrix W_g for the input feature to dynamically represent the missing modality:

$$E_{alter}\left(F^m\right) = \alpha \cdot F^m + \beta \cdot V, \quad [\alpha, \beta] = \sigma\left(W_g F^m\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the softmax function. Unlike the copy expert, the alter expert introduces additional information to complement the missing modality, making it possible to be generated rather than being limited to the available modality, especially when the available modality is of low quality.

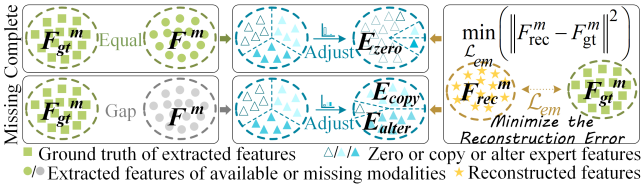

Figure 4: Illustration of the expert modulation loss.

Overall, the MaMoE compensates for the missing modality by assigning larger weights to the copy and alter experts, as shown in the upper middle of Fig. 3.

Soft Router. The soft router $\mathcal{G}$ is introduced to dynamically assign weights to the experts based on the input feature representation. Specifically, the soft router first smooths the input feature F^m and reduces its dimension through a convolutional operation. The resulting feature map is then flattened and passed through a two-layer Multilayer Perceptron (MLP) to generate the weights for each expert, as follows:

$$\mathcal{G}\left(F^m\right) = \sigma\left(\text{MLP}\left(\text{Flatten}\left(Conv\left(F^m\right)\right)\right)\right), \tag{6}$$

where $Conv(\cdot)$ denotes the 3×3 convolution.

Expert Modulation Loss. We further introduce an Expert Modulation (EM) loss to guide the soft router to select the appropriate experts in different situations, enabling adaptive processing without manual intervention. Specifically, the reconstructed feature F^m_{rec} in Eq. 2 is composed of the extracted feature F^m and the adapted feature F^m_{ada}, in which F^m varies depending on whether the current modality is missing or available. The EM loss takes advantage of this dynamic property to modulate experts. As illustrated in Fig. 4, when the current modality is available, F^m should

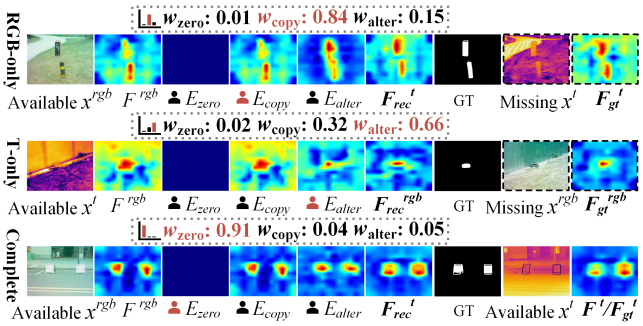

Figure 5: Visualization of router weights (w_i), available modality features (F^m), expert outputs (E_i), reconstructed features ($F^{\bar{m}}_{rec}$), and their corresponding ground truth features ($F^{\bar{m}}_{gt}$) in modality-missing and -complete cases.

ideally match its ground truth F^m_{gt}. In this case, minimizing the discrepancy between F^m_{rec} and F^m_{gt} encourages the router to assign larger weights to the zero expert in $F^{\bar{m}}_{ada}$, thereby suppressing the redundant or misleading contributions from the other modality. Conversely, when the current modality is missing (i.e., $F^m \neq F^m_{gt}$), minimizing the same discrepancy encourages the router to rely more on the copy and alter experts to effectively compensate for the missing information. The EM loss is thus defined as:

$$\mathcal{L}_{em} = \frac{1}{N}\sum_{n=1}^{N}\left(F^m_{rec}(n) - F^m_{gt}(n)\right)^2, \tag{7}$$

where N is the number of spatial positions. The Em loss uniformly satisfies the demands of missing and available cases.

In Fig. 5, the visualizations of feature maps and router weights show that each expert focuses on capturing distinct features. For each input case, the expert tailored for the situation is consistently selected with a larger weight, enabling effective feature reconstruction. Specifically, the reconstructed features closely approximate the ground truth of the missing modality features (e.g., rows 1 and 2) or avoid disrupting the available modality features (e.g., row 3). These results indicate that each expert is both effective and essential, and that adaptive expert selection is successfully achieved through the soft router guided by the EM loss.

Multi-modal MoE (MmMoE)

Given the reconstructed modality features, we propose the MmMoE to effectively integrate and combine them. Unlike prior works (Guo et al. 2024b; Wang et al. 2024b), MmMoE preserves modality-specific information before performing complementary multi-modal fusion, thereby fully leveraging the advantage of each modality. In specific, MmMoE introduces two uni-modal experts with modality-specific knowledge, $\{E_{rgb}, E_{d/t}\}$, tailored for feature integration within the modality. Each expert has a homogeneous structure and implemented using the attention mechanism (Vaswani 2017), with separate modality pre-training, formulated as:

$$F^m_{out} = E_m\left(F^m_{rec}\right) = \sigma\left(\frac{Q^m(K^m)^\top}{\sqrt{d_k}}\right)V^m, \tag{8}$$

where $m \in \{rgb, d/t\}$, F^m_{out} denotes the integrated uni-modal feature, and Q^m, K^m, and V^m are query, key, and value matrices derived from F^m_{rec} via linear projection. The soft router $\mathcal{G}$ is also used here to measure the importance of each uni-modal expert. Specifically, the features F^{rgb}_{out} and $F^{d/t}_{out}$ are concatenated along the channel dimension and passed through a Channel Attention (CA) block (Woo et al. 2018) to capture inter-modal dependencies. The resulting representation is then processed by $\mathcal{G}$ to produce modality-specific weights w_m:

$$w_m = \mathcal{G}\left(\mathrm{CA}\left([F^{rgb}_{out}, F^{d/t}_{out}]\right)\right), \tag{9}$$

where $[,]$ denotes channel-wise concatenation. Finally, the fused multi-modal feature F_{fus} is computed as:

$$F_{fus} = w_{rgb}\cdot F^{rgb}_{out} + w_{d/t}\cdot F^{d/t}_{out}, \tag{10}$$

which will be fed into the decoder (Liu et al. 2022) for saliency prediction.

Training Paradigm

CMoE follows a pre-training and fine-tuning training flow and is optimized by a combination of loss functions.

Pre-training. In the pre-training phase, we utilize data from RGB and depth or thermal modalities to independently train each uni-modal expert with modality-specific knowledge. Each expert is trained within a encoder-decoder architecture consistent with CMoE. Following previous works (Wu et al. 2021; Tu et al. 2021), we adopt a saliency loss $\mathcal{L}_{sal}$ composed of a binary cross-entropy loss $\mathcal{L}_{bce}$ and a dice loss $\mathcal{L}_{dice}$ (Milletari, Navab, and Ahmadi 2016):

$$\mathcal{L}_{sal} = \mathcal{L}_{bce}(S, Y) + \mathcal{L}_{dice}(S, Y). \tag{11}$$

Fine-tuning. In the fine-tuning phase, we initialize and freeze the uni-modal experts and encoders using the pre-trained parameters to preserve their modality-specific capabilities and focus on optimizing the remaining modules in CMoE. Since this stage involves both modality-complete and modality-missing inputs, we jointly optimize the model using the proposed EM loss and the saliency loss:

$$\mathcal{L} = \mathcal{L}_{em} + \mathcal{L}_{sal}. \tag{12}$$

Experiments

We implement our method on two main multi-modal SOD tasks: RGB-T and RGB-D SOD. The model for each task is evaluated in both modality-missing and -complete cases.

Datasets and Evaluation Metrics

For RGB-T SOD, we utilize three prevalent datasets: VT821 (Tang et al. 2019), VT1000 (Tu et al. 2019), and VT5000 (Tu et al. 2022). For RGB-D SOD, experiments are conducted on three relatively large-scale datasets: STERE (Niu et al. 2012), SIP (Fan et al. 2020), and ReDWeb-S (Liu et al. 2021a). As the setup in prior work (He et al. 2025), we use the training set of VT5000, and the training sets of NJUD (Ju et al. 2014), NLPR (Peng et al. 2014), DUTLF-Depth (Piao et al. 2019) to train RGB-T and RGB-D SOD models, respectively. Besides, we adopt two widely used evaluation metrics E-measure (E_m) (Fan et al. 2018) and F-measure (F_m) (Achanta et al. 2009) to access model.

Dataset	VT5000								VT1000								VT821							
Missing condition	RGB ○	T ●	RGB ●	T ○	RGB ●	T ●	**Average**		RGB ○	T ●	RGB ●	T ○	RGB ●	T ●	**Average**		RGB ○	T ●	RGB ●	T ○	RGB ●	T ●	**Average**	
Metric	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$
SwinNet	.866	.743	.934	.845	.940	.861	.913	.816	.889	.812	.943	.889	.949	.897	.927	.866	.853	.731	.922	.824	.928	.851	.901	.802
HRTrans	.863	.725	.937	.856	.945	.871	.915	.817	.900	.808	.941	.893	.945	.900	.929	.867	.876	.747	.914	.823	.929	.853	.906	.808
CAVER	.846	.694	.909	.823	.924	.841	.893	.786	.884	.788	.933	.889	.946	.903	.921	.860	.809	.671	.897	.783	.919	.839	.875	.764
UniTR	.809	.728	.903	.839	.927	.854	.880	.807	.881	.837	.940	.914	.949	.912	.923	.888	.793	.708	.872	.802	.902	.831	.856	.780
LAFB	.858	.741	.908	.823	.931	.857	.899	.807	.888	.812	.932	.885	.945	.905	.922	.867	.842	.696	.895	.780	.915	.843	.884	.773
VSCode	.815	.720	.932	.865	.945	.884	.897	.823	.927	.861	.945	.909	.958	.924	.943	.898	.807	.721	.864	.811	.948	.887	.873	.806
CoLA	.893	.784	.912	.824	.927	.849	.911	.819	.925	.857	.926	.877	.940	.893	.930	.876	.868	.754	.896	.809	.916	.841	.893	.801
Samba	.891	.754	**.948**	**.884**	.950	**.886**	.930	.841	.921	.841	.958	**.926**	.958	.926	.946	.898	.884	.757	.939	**.879**	.940	.884	.921	.840
CMoE	**.939**	**.845**	**.948**	.873	**.954**	.885	**.947**	**.868**	**.947**	**.899**	**.959**	.925	**.963**	**.930**	**.956**	**.918**	**.931**	**.838**	**.942**	.876	**.955**	**.897**	**.943**	**.870**

Table 1: Quantitative comparison with multi-modal SOD methods on RGB-T SOD datasets in modality-missing and modality-complete conditions. "○ / ●" denotes that a modality is missing / available. The best two results are marked in **bold** and underline.

Dataset	STERE								SIP								ReDWeb-S							
Missing condition	RGB ○	D ●	RGB ●	D ○	RGB ●	D ●	**Average**		RGB ○	D ●	RGB ●	D ○	RGB ●	D ●	**Average**		RGB ○	D ●	RGB ●	D ○	RGB ●	D ●	**Average**	
Metric	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$F_m\uparrow$
SwinNet	.628	.455	.929	.891	.929	.894	.829	.747	.680	.543	.925	.880	.923	.872	.843	.765	.541	.318	.766	.717	.764	.711	.690	.582
HRTrans	.602	.252	.768	.732	.930	.894	.767	.626	.660	.349	.915	.872	.893	.834	.823	.685	.597	.333	.745	.686	.745	.686	.696	.568
CAVER	.717	.582	.919	.881	.928	.893	.855	.785	.793	.710	.903	.847	.930	.894	.875	.817	.602	.495	.755	.700	.760	.724	.706	.640
UniTR	.423	.342	.925	.892	.927	.891	.758	.708	.703	.679	.901	.849	.909	.864	.838	.797	.431	.339	.724	.700	.728	.695	.628	.578
LAFB	.648	.433	.923	.885	.927	.886	.833	.735	.556	.478	.925	.883	.937	.905	.806	.755	.564	.406	.755	.705	.763	.719	.694	.610
VSCode	.306	.290	.927	.898	.929	.903	.721	.697	.473	.188	.936	.892	.943	.911	.784	.664	.267	.320	.752	.725	.776	.763	.598	.603
CoLA	.653	.310	.895	.828	.859	.745	.802	.628	.576	.112	.884	.792	.802	.651	.754	.518	.568	.303	.722	.629	.686	.552	.659	.495
Samba	.697	.448	.934	**.911**	.935	**.915**	.855	.758	.696	.677	.948	.916	.950	**.925**	.865	.839	.575	.417	.775	.753	.774	.759	.708	.643
CMoE	**.764**	**.621**	**.935**	**.911**	**.938**	.913	**.879**	**.815**	**.869**	**.765**	**.954**	**.919**	**.954**	.916	**.926**	**.867**	**.678**	**.539**	**.800**	**.791**	**.798**	**.788**	**.759**	**.706**

Table 2: Quantitative comparison on RGB-D SOD datasets in modality-missing and modality-complete conditions.

Implementation Details

The proposed network is trained on two RTX 3090 Ti GPUs. Both pre-training and fine-tuning stages employ the AdamW optimizer (Loshchilov and Hutter 2019) with learning rate of 1e-5, weight decay of 1e-4, and batch size of 4. The model is trained around 120 epochs. Following previous works (Liu et al. 2022; Luo et al. 2024), we resize each image to 384 $\times$ 384, as specified by the Swin-B (Liu et al. 2021b) encoder.

Comparison with State-of-the-Art Methods

For fairness, we compare CMoE with 8 state-of-the-art multi-modal SOD methods that are also designed for both RGB-T and RGB-D tasks, including SwinNet (Liu et al. 2022), HRTrans (Tang et al. 2023), CAVER (Pang et al. 2023), UniTR (Guo et al. 2024b), LAFB (Wang et al. 2024b), VSCode (Luo et al. 2024), CoLA (Hao, Zhong, and Tang 2024), and Samba (He et al. 2025). Tables 1 and 2 present the quantitative results on six multi-modal SOD datasets. CMoE consistently achieves overall superior performance in both modality-missing and modality-complete conditions. While CMoE slightly underperforms the suboptimal method Samba on the F_m metric in certain cases, its average performance (i.e., the "Average" column) across all datasets demonstrates notable improvements. For instance, although CMoE lags by 0.2% on F_m of the STERE dataset

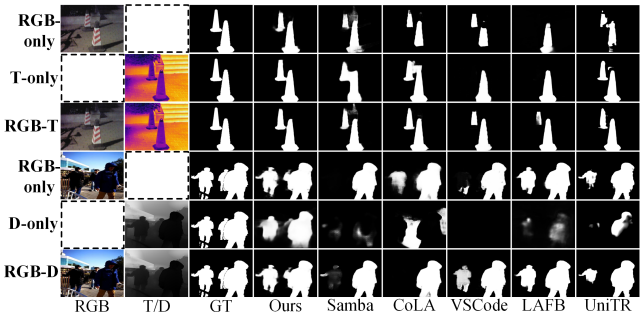

Figure 6: Visual comparison in missing and complete cases.

under the modality-complete condition, it achieves an average improvement of 7.5% across all three conditions. This improvement is mainly attributed to the cascaded design of missing-aware MoE (MaMoE) and multi-modal MoE (MmMoE), which effectively leverage the available modality information and reduce reliance on complete input modalities.

In addition, Fig. 6 vividly illustrates the visual comparison between CMoE and recent advanced methods for RGB-T and RGB-D inputs in the challenging low-illumination scene. For modality-missing cases (i.e., rows 1, 2, 4, and 5), CMoE locates object regions more accurately without being

ID	Model	VT5000						STERE					
		RGB ○	T ●	RGB ●	T ○	RGB ●	T ●	RGB ○	D ●	RGB ●	D ○	RGB ●	D ●
		E_m	F_m	E_m	F_m	E_m	$F_m \uparrow$	E_m	F_m	E_m	F_m	E_m	F_m
0	CMoE	**.939**	**.845**	**.948**	**.873**	**.954**	**.885**	**.764**	**.621**	**.935**	**.911**	**.938**	**.913**
1	w/o MaMoE	.923	.818	.937	.844	.946	.867	.753	.590	.925	.889	.930	.898
2	w/o EM loss	.932	.835	.939	.857	.947	.871	.755	.596	.928	.895	.931	.901
3	w/o MmMoE	.933	.831	.939	.852	.946	.869	.757	.603	.926	.893	.928	.899
4	w/o Pre-train	.935	.839	.943	.865	.948	.876	.761	.616	.931	.903	.934	.904
5	w/o Router	.933	.836	.942	.862	.948	.873	.758	.605	.930	.899	.933	.903

Table 3: Ablation study of CMoE on RGB-T and RGB-D SOD datasets. "w/o" denotes the removal of a component.

interfered by too much background noise, indicating that it fully leverages the available modality information and effectively adapts to modality-missing conditions. For modality-complete cases (i.e., rows 3 and 6), the results of CMoE are visually closer to the ground truth (GT), indicating that it successfully captures the complementary information between modalities.

Ablation Study

To evaluate the contribution of key components in CMoE, we conduct ablation studies on two representative multi-modal SOD datasets. The results are reported in Table 3.

Effect of Missing-aware MoE. We assess the impact of the Missing-aware MoE (MaMoE) by directly removing it (i.e., ID1), which means that the model struggles to flexibly handle modality-missing and modality-complete cases. Compared with the complete model (i.e., ID0), the E_m and F_m metrics on the two datasets across all conditions decrease by an average of 1.2% and 2.9%, respectively. This is due to the collaboration of modality-reconstruction experts and the constraints of EM loss, which jointly enhance the adaptability to different conditions. To this end, ID2 removes the constraints imposed by EM loss, and the comparison (ID2 vs. ID0) shows its positive effect. Furthermore, Table 4 analyzes the necessity of the three modality-reconstruction experts (i.e., zero, copy, and alter). Specifically, we successively train each variant using only a single expert in MaMoE, and perform evaluation under different conditions on the VT5000 dataset. It can be observed that, with the assistance of different modality-reconstruction experts, each variant achieves superior performance in the condition that it should deal with. In particular, although both copy and alter experts excel at modality-missing conditions, the alter expert is especially effective in RGB-missing cases by modifying and enhancing thermal features, which are typically less informative than RGB features. Nevertheless, all single-expert variants are inferior to CMoE, proving that combining all experts allows CMoE to fully exploit their advantages.

Effect of Multi-modal MoE. To verify the effectiveness of the Multi-modal MoE (MmMoE), we replace it with a simple feature summation strategy, where the reconstructed multi-modal features are directly fused without specific integration and dynamic weighting. The obvious performance decline in ID3 confirms that MmMoE enables more effective multi-modal feature processing. In addition, the comparison

ID	Model	RGB-missing		T-missing		Complete	
		$E_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$F_m \uparrow$
0	CMoE (all experts)	**.939**	**.845**	**.948**	**.873**	**.954**	**.885**
1	only zero expert	.924	.817	.936	.846	<u>.948</u>	<u>.870</u>
2	only copy expert	.927	.824	<u>.941</u>	<u>.856</u>	.945	.866
3	only alter expert	<u>.932</u>	<u>.832</u>	.938	.847	.944	.865

Table 4: Comparison of single-expert variants on VT5000 in modality-missing and modality-complete conditions.

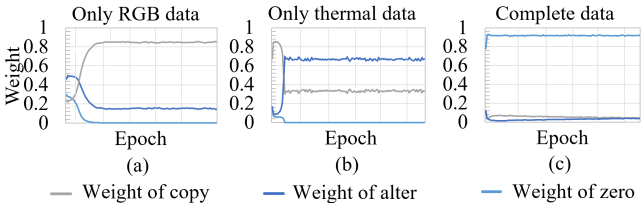

Figure 7: Visualization of weight assignment during training in modality-missing (a, b) and modality-complete (c) cases.

between ID4 and ID0 shows that discarding the pre-training of uni-modal experts in MmMoE consistently degrades performance, validating the benefit of modality-specific knowledge learned during pre-training.

Effect of Soft Router. Since the soft router is embedded into both MaMoE and MmMoE, we also remove it to verify its effectiveness. In this setting, expert features are combined with equal weights without dynamic selection. The clear performance drop in ID5 shows the importance of soft router. Furthermore, Fig. 7 records the average weights assigned by the soft router to each modality-reconstruction expert when training under modality-missing and modality-complete conditions. With the iteration of training, the soft router gradually assigns larger weights to the expert tailored to the current input condition. This adaptive behavior confirms the effectiveness of the soft router in dynamically scheduling expert contributions.

Conclusion

In this paper, we present CMoE, a flexible framework that introduces cascaded mixture-of-experts to achieve robust multi-modal salient object detection (SOD) in both modality-missing and modality-complete conditions. Specifically, a missing-aware MoE comprising three modality-reconstruction experts is proposed to adaptively reconstruct the features of each modality under the constraints of an expert modulation loss, and a multi-modal MoE containing two pre-trained uni-modal experts is proposed to fully integrate multi-modal features. With the guidance of the soft router, experts within the two modules are dynamically selected and combined to enhance the adaptability of CMoE across varying input cases. Extensive experiments on six representative multi-modal SOD datasets demonstrate the effectiveness and generalization capability of the proposed framework.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1597–1604.

Ai, Y.; Huang, H.; and He, R. 2024. LoRA-IR: Taming Low-Rank Experts for Efficient All-in-One Image Restoration. *arXiv preprint arXiv:2410.15385*.

Ben-Shabat, Y.; Hewa Koneputugodage, C.; Ramasinghe, S.; and Gould, S. 2024. Neural experts: Mixture of experts for implicit neural representations. *Advances in Neural Information Processing Systems*, 37: 101641–101670.

Chen, H.; Deng, Y.; Li, Y.; Hung, T.-Y.; and Lin, G. 2020. RGBD salient object detection via disentangled cross-modal fusion. *IEEE Transactions on Image Processing*, 29: 8407–8416.

Delmas, G.; Rezende, R. S.; Csurka, G.; and Larlus, D. 2022. ARTEMIS: Attention-based Retrieval with Text-Explicit Matching and Implicit Similarity. In *Proceedings of the The International Conference on Learning Representations*.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 698–704.

Fan, D.-P.; Lin, Z.; Zhang, Z.; Zhu, M.; and Cheng, M.-M. 2020. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5): 2075–2089.

Fan, X.; Liu, L.; and Zhang, H. 2024. Multimodal Information Interaction for Medical Image Segmentation. *arXiv preprint arXiv:2404.16371*.

Guo, H.; Dai, T.; Bai, Y.; Chen, B.; Ren, X.; Zhu, Z.; and Xia, S.-T. 2024a. Parameter efficient adaptation for image restoration with heterogeneous mixture-of-experts. *Advances in Neural Information Processing Systems*, 37: 13522–13547.

Guo, R.; Ying, X.; Qi, Y.; and Qu, L. 2024b. UniTR: A Unified TRansformer-Based Framework for Co-Object and Multi-Modal Saliency Detection. *IEEE Transactions on Multimedia*, 26: 7622–7635.

Han, J.; Chen, H.; Liu, N.; Yan, C.; and Li, X. 2017. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics*, 48(11): 3171–3183.

Hao, S.; Zhong, C.; and Tang, H. 2024. Cola: Conditional dropout and language-driven robust dual-modal salient object detection. In *Proceedings of the European Conference on Computer Vision*, 354–371.

He, J.; Fu, K.; Liu, X.; and Zhao, Q. 2025. Samba: A Unified Mamba-based Framework for General Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25314–25324.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hu, X.; Sun, F.; Sun, J.; Wang, F.; and Li, H. 2024. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision*, 132(8): 3067–3085.

Jin, P.; Zhu, B.; Yuan, L.; and Yan, S. 2024. Moe++: Accelerating mixture-of-experts methods with zero-computation experts. *arXiv preprint arXiv:2410.07348*.

Ju, R.; Ge, L.; Geng, W.; Ren, T.; and Wu, G. 2014. Depth saliency based on anisotropic center-surround difference. In *Proceedings of the International Conference on Image Processing*, 1115–1119.

Lee, Y.-L.; Tsai, Y.-H.; Chiu, W.-C.; and Lee, C.-Y. 2023. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14943–14952.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Liu, N.; Zhang, N.; Shao, L.; and Han, J. 2021a. Learning selective mutual attention and contrast for RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9026–9042.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Liu, Z.; Tan, Y.; He, Q.; and Xiao, Y. 2022. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4486–4497.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the The International Conference on Learning Representations*.

Lu, A.; Li, C.; Zhao, J.; Tang, J.; and Luo, B. 2025a. Modality-missing RGBT tracking: Invertible prompt learning and high-quality benchmarks. *International Journal of Computer Vision*, 133(5): 2599–2619.

Lu, A.; Qian, C.; Li, C.; Tang, J.; and Wang, L. 2025b. Duality-Gated Mutual Condition Network for RGBT Tracking. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3): 4118–4131.

Lu, A.; Wang, W.; Li, C.; Tang, J.; and Luo, B. 2025c. AFTER: Attention-Based Fusion Router for RGBT Tracking. *IEEE Transactions on Image Processing*, 34: 4386–4401.

Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.-P.; Khan, F.; and Han, J. 2024. VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17169–17180.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision*, 565–571.

Min, X.; Zhai, G.; Zhou, J.; Zhang, X.-P.; Yang, X.; and Guan, X. 2020. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29: 3805–3819.

Niu, Y.; Geng, Y.; Li, X.; and Liu, F. 2012. Leveraging stereopsis for saliency analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 454–461.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2023. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32: 892–904.

Peng, H.; Li, B.; Xiong, W.; Hu, W.; and Ji, R. 2014. RGBD salient object detection: A benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision*, 92–109.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7254–7263.

Qu, L.; He, S.; Zhang, J.; Tian, J.; Tang, Y.; and Yang, Q. 2017. RGBD salient object detection via deep fusion. *IEEE transactions on image processing*, 26(5): 2274–2285.

Ramazanova, M.; Pardo, A.; Alwassel, H.; and Ghanem, B. 2025. Exploring missing modality in multimodal egocentric datasets. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 75–85.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2016. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *Proceedings of the International Conference on Learning Representations*.

Tang, B.; Liu, Z.; Tan, Y.; and He, Q. 2023. HRTransNet: HRFormer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 728–742.

Tang, J.; Fan, D.; Wang, X.; Tu, Z.; and Li, C. 2019. RGBT salient object detection: benchmark and a novel cooperative ranking approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12): 4421–4433.

Tu, Z.; Li, Z.; Li, C.; Lang, Y.; and Tang, J. 2021. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing*, 30: 5678–5691.

Tu, Z.; Ma, Y.; Li, Z.; Li, C.; Xu, J.; and Liu, Y. 2022. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25: 4163–4176.

Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; and Tang, J. 2019. RGB-T image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1): 160–173.

Vaswani, A. 2017. Attention is All You Need. *arXiv preprint arXiv:1706.03762*.

Wang, K.; Chen, K.; Li, C.; Tu, Z.; and Luo, B. 2025a. Alignment-Free RGB-T Salient Object Detection: A Large-scale Dataset and Progressive Correlation Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7780–7788.

Wang, K.; Lin, D.; Li, C.; Tu, Z.; and Luo, B. 2024a. Alignment-Free RGBT Salient Object Detection: Semantics-guided Asymmetric Correlation Network and A Unified Benchmark. *IEEE Transactions on Multimedia*, 26: 10692–10707.

Wang, K.; Tu, Z.; Li, C.; Liu, Z.; and Luo, B. 2025b. Unified-modal salient object detection via adaptive prompt learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, K.; Tu, Z.; Li, C.; Zhang, C.; and Luo, B. 2024b. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7344–7358.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, Y.-H.; Liu, Y.; Xu, J.; Bian, J.-W.; Gu, Y.-C.; and Cheng, M.-M. 2021. MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 10261–10269.

Wu, Z.; Liu, C.; Wen, J.; Xu, Y.; Yang, J.; and Li, X. 2025a. Spatial Continuity and Nonequal Importance in Salient Object Detection With Image-Category Supervision. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5): 8565–8576.

Wu, Z.; Xu, Y.; Yang, J.; and Zhang, D. 2025b. Weakly Supervised Salient Object Detection With Oversize Bounding Box: Z. Wu et al. *International Journal of Computer Vision*, 133(9): 6558–6577.

Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the ACM International Conference on Multimedia*, 438–446.

Yao, W.; Yin, K.; Cheung, W. K.; Liu, J.; and Qin, J. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16416–16424.

Zhang, J.; Zhang, R.; Xu, L.; Lu, X.; Yu, Y.; Xu, M.; and Zhao, H. 2024. FasterSal: Robust and Real-time Single-Stream Architecture for RGB-D Salient Object Detection. *IEEE Transactions on Multimedia*.

Zhou, T.; Fan, D.-P.; Cheng, M.-M.; Shen, J.; and Shao, L. 2021. RGB-D salient object detection: A survey. *Computational Visual Media*, 7: 37–69.

Zong, Z.; Ma, B.; Shen, D.; Song, G.; Shao, H.; Jiang, D.; Li, H.; and Liu, Y. 2024. Mova: Adapting mixture of vision experts to multimodal context. *Advances in Neural Information Processing Systems*, 37: 103305–103333.