CoLA: Conditional Dropout and Language-driven Robust Dual-modal Salient Object Detection

Shuang Hao[1⋆], Chunlin Zhong[1⋆], and He Tang[1✉]

School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
{shuanghao, clzhong, hetang}@hust.edu.cn

Abstract. The depth/thermal information is beneficial for detecting salient object with conventional RGB images. However, in dual-modal salient object detection (SOD) model, the robustness against noisy inputs and modality missing is crucial but rarely studied. To tackle this problem, we introduce **Co**nditional Dropout and **LA**nguage-driven(**CoLA**) framework comprising two core components. 1) Language-driven Quality Assessment (LQA): Leveraging a pretrained vision-language model with a prompt learner, the LQA recalibrates image contributions without requiring additional quality annotations. This approach effectively mitigates the impact of noisy inputs. 2) Conditional Dropout (CD): A learning method to strengthen the model's adaptability in scenarios with missing modalities, while preserving its performance under complete modalities. The CD serves as a plug-in training scheme that treats modality-missing as conditions, strengthening the overall robustness of various dual-modal SOD models. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art dual-modal SOD models, under both modality-complete and modality-missing conditions. The code is avaliable at https://github.com/ssecv/CoLA.

Keywords: Dual-modal Salient Object Detection, Modality Robustness

1 Introduction

Salient object detection (SOD) is a foundational task in computer vision so as to extract the most attractive objects/regions from a scene. Benefiting from the auxiliary inputs like depth and thermal, advanced dual-modal SOD models [6, 9, 42, 45, 47, 56, 60] are able to detect salient objects even in complex and challenging scenes. Due to the ability of redundant elimination and reduction in computational complexity, this technique has been widely applied to enhance the performance of action recognition [24], object tracking [16, 61], semantic segmentation [5, 26], and image fusion [28].

In general, the high accuracy of a dual-modal SOD model relies on high-quality and complete inputs. However, in real-world situations, *1) the input may*

⋆ Equal contribution ✉ Corresponding author

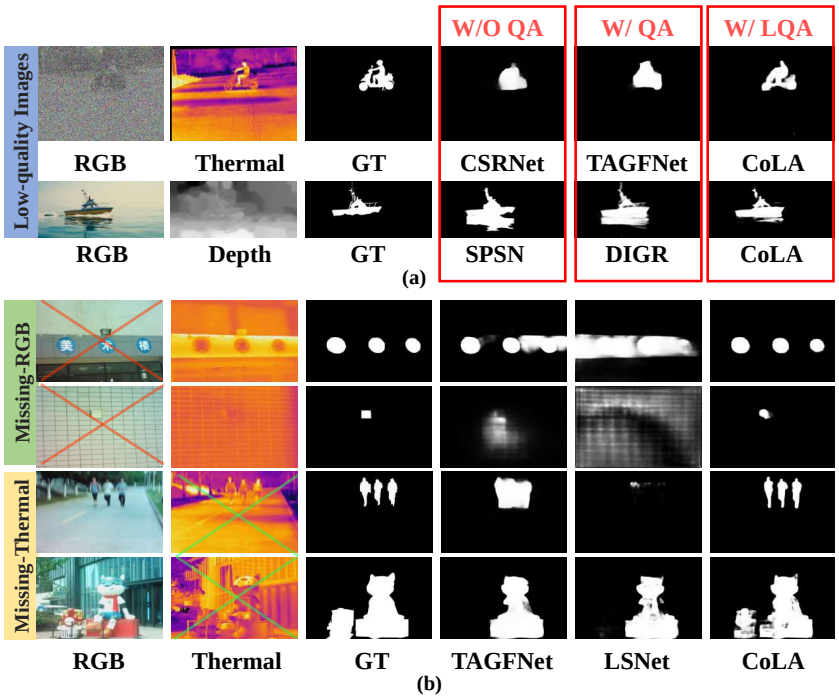

Fig. 1: (a): The two examples demonstrate the cases when the RGB or the Depth inputs are noisy, QA means quality assessment. (b): examples of two modality-missing conditions. The proposed CoLA produces robust results.

be noisy due to communication fault [1]*, 2) and one of the modalities may be missing due to device malfunction.* A SOD model trained on ideal inputs may degrade when faced with corrupted or partially missing inputs.

As shown in Fig. 1(a), the noisy RGB image or depth/thermal image leads to unsatisfactory predictions in conventional methods such as CSRNet [18] and SPSN [25]. It is necessary to reweight the contribution of each modality rather than treating all inputs as ideal. Though quality-aware SOD models have been proposed, their quality estimation relies on either an off-the-shelf estimation network [9, 47] or pseudo-labels [4, 6]. The parameters of the off-the-shelf estimation network [9] are fixed, and pseudo-labels [4, 6] are inaccurate; neither of them can adapt well to the target dataset such as TAGFNet [47] and DIGR [6]. Recent advances of the pre-trained vision-language model [40] have shown potential capability to assess the quality of an image [48]. To this end, we propose a language-driven quality assessment (LQA) module to assess the quality of each input and reweight the contribution for the network. The LQA contains a fixed prompt and a learnable prompt; only the parameters of the learnable prompt

[1] We do not consider adversarial noise in this paper.

are updated during training, while the fixed prompt and vision/language encoder are frozen. Compared with other image quality assessment methods, this design not only maintains the generalization ability of the pre-trained vision-language model but also adapts to the target dataset in a parameter-efficient fine-tuning manner. As shown in the last column of Fig. 1(a), our predictions with the proposed LQA best fit the GT in both noisy RGB or depth.

Another limitation of existing dual-modal SOD models is that they do not consider performance degradation when the modality is partially missing. This leads to existing models overly relying on complete modal inputs, resulting in poor performance when modality-missing conditions. Furthermore, in system deployment, the phenomenon of Modality-missing occurs naturally. It should be noted that even though the RGB modality is missing from the model's input, it still exists in principle. As shown in 1(b), when one modality is missing, dual-modal SOD models like LSNet [60] and TAGF [47] can not recognize the salient object in the images, even though another modality image is easy to recognize the salient object. This indicates existing dual-modal SOD models lack consideration for the situation of modality-missing. Resulting in insufficient utilization of the image information of another existing modality when the modality is missing. To address this limitation, we explore a training strategy to enhance the performance of dual-modal SOD under both modality-complete and modality-missing conditions. Inspired by conditional controls [55], we treat the modality missing as the *condition* and replicate the encoders trained when the modality is complete. Subsequently, we freeze the original encoder and update the parameters of the copied encoder and zero convolutions. This approach, namely Conditional Dropout (CD), preserves the model's capability when dealing with complete modalities and enables it to extract modality-invariant and specific features.

We propose a new perspective on dual-modal SOD, aiming to investigate a robust framework resistant to noisy and incomplete inputs while maintaining accuracy when facing ideal inputs. In sum, the main contributions and why this work is non-trivial are as follows:

- To the best of our knowledge, this is the first robust dual-modal SOD model against both noisy image and modality-missing.
- A language-driven quality assessment (LQA) module with a learnable prompt is proposed to reweight modality contributions. LQA enhances the robustness and performance of the model with noisy images.
- A Conditional Dropout (CD) learning method is proposed to promote the performance of the model in both modality complete and missing conditions.
- Though the proposed network is simply designed for better scalability, our model outperforms the state-of-the-art in both modality complete and missing conditions.

2 Related Work

2.1 Dual-modal salient object detection

Recent years have seen increased research in dual-modal salient object detection, particularly in RGB-T [9,18,19,42–44,47,49,60] and RGB-D [6,8,25,30–33,38,56] modalities. These works have made significant progress in the field of salient object detection. For instance, [6,9] enhanced performance through various estimations, [18,19] focused on fusion methods like context-guided and early single-stream fusion, and [42,43] tackled issues like multi-interactive fusion and spatial misalignment. [44,47,49,60] introduced strategies like selective feature collection, compensation for low illumination, cross-scale fusion, and feature propagation. Additionally, [8, 25, 56] contributed with methods like Criss-Cross Dynamic Filter Networks, cross-modality interaction, and Superpixel Prototype Sampling, respectively, to advance salient object detection.

These studies have collectively enhanced dual-modal salient object detection performance. However, these methods frequently show limited adaptability to low-quality or absent modalities. Often, models trained under ideal conditions struggle with generalization in scenarios with noisy or partially missing inputs. To address this, we propose CoLA adept at handling both complete and incomplete modalities. CoLA guarantees robust performance in scenarios with noisy or unavailable modalities.

2.2 Incomplete modality learning

Incomplete modality learning aims to alleviate the modality missing problem due to sensor failures or environmental constraints [12]. Recent advantages can be categorized into reconstruction-based and fusion-based approaches. Reconstruction-based methods exploit the remaining modalities to recover or reconstruct the missing ones. For example, latent space strategies [21] can be used to generate complete multimodal inputs, and Generative Adversarial Network [3, 23, 37] can be employed to restore the absent modalities. Reconstruction-based methods require training and deploying a distinct model for each subset of missing modalities, leading to high complexity in practical applications. Recent works have explored fusion-based methods to learn joint representations directly from incomplete multimodal inputs across different applications. [10, 12, 34, 51] explore fusion methods for incomplete modalities from various perspectives such as regional perception, multi-tasking, and feature space.

These methods tend to make the model adapt to the missing modality, which can hardly ensure unaffected performance on complete modalities. We propose a Conditional Dropout method to improve performance in incomplete modalities without compromising complete modalities, alleviating the missing modality issue in dual-modal SOD tasks.

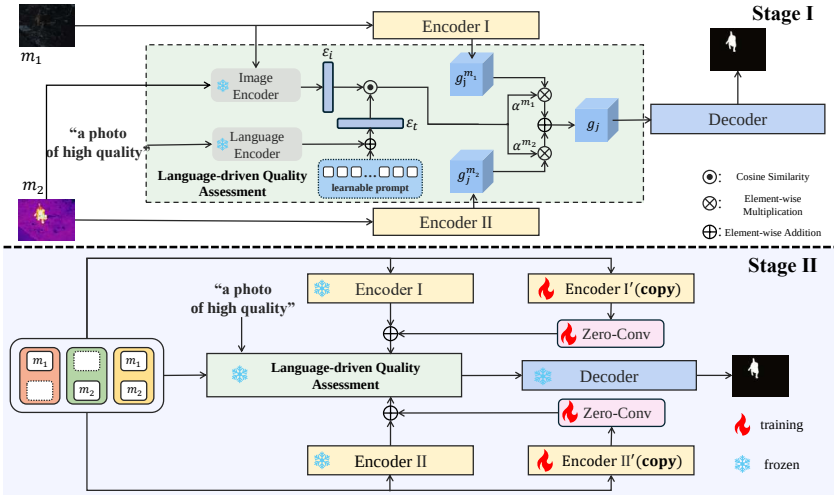

Fig. 2: The architecture of CoLA represents a two-stage neural network with Stage I training a language quality assessment (LQA) to calibrate feature fusion, and Stage II training with Conditional Dropout enhances the capabilities of both missing and complete modalities.

2.3 Vision-language model and application

Recent works research has extensively investigated a range of vision-language models, e.g., DALL-E [41], ALIGN [20] and CLIP [40], which harmonizes text and image understanding by learning to associate images with their textual descriptions effectively. These models have demonstrated remarkable capabilities across a spectrum of computer vision tasks such as crowd counting [27], pedestrian detection [29], human and object interaction detection [36], scene text detection [54], etc. Beyond recognition, CLIP is studied for monocular depth estimation in a zero-shot manner [57]. Due to the variety of downstream tasks, many fine-tuning improvements based on CLIP have been proposed [53, 58, 59], which have greatly enhanced the versatility of CLIP [50, 53, 54].

Zero-shot and fine-tune based CLIP applications present versatile approaches for diverse computer vision tasks. We utilize CLIP to access the quality of inputs by prompt learning, leveraging its transformable capabilities to tackle the challenges posed by noisy inputs.

3 Methodology

3.1 Overview of the framework

The process of training a conventional multi-modal SOD can be summarized as $f(x) \overset{t_1}{\leftarrow} \{f; x\}$ where t_1 denotes the training process, x refers to the multi-modal inputs, and $f(x)$ is the trained model with data x.

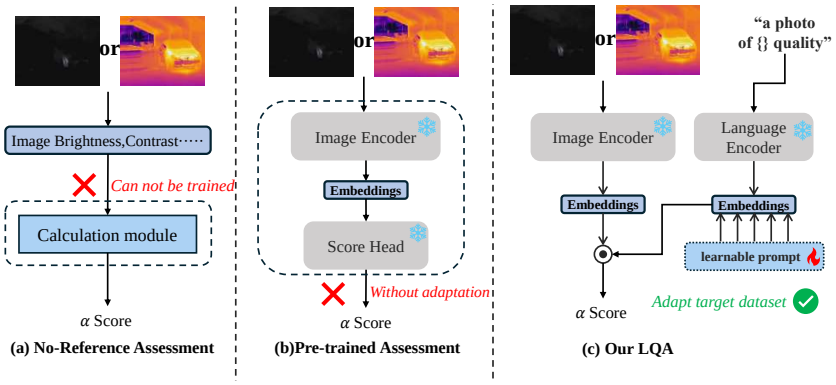

Fig. 3: Architectural comparison of (a) No-Reference Method, (b) Pre-trained Assessment and (c) Our LQA.

Unlike previous works, our CoLA consists of four components: 1) a dual-branch encoder for feature extraction; 2) an identically structured encoder with a Conditional Dropout (CD) for robust modality feature supplementation (in training stage II); 3) a Language-driven Quality Assessment module (LQA) for learning modality contribution ratios (in training stage I); 4) a decoder for feature aggregation to produce the output. Importantly, our encoder and decoder designs are kept simple without complex interactions. This allows us to demonstrate the efficacy of our proposed modules clearly. Based on the components, the training process can be expressed as:

$$f'(x) \xleftarrow{t_1} \{f'; LQA(x)\}, \tag{1}$$

$$f(x) \xleftarrow{t_2} \{f'(\rho(x)); CD(\rho(x))\}, \tag{2}$$

where x is the multi-modal inputs, $f'(x)$ represents the trained model of the first stage. The first and second stage training processes are denoted by t_1 and t_2 respectively, and $\rho(x)$ is the input with missing probability ρ.

3.2 Language-driven quality assessment for low quality image

Fig. 3 compares other image quality assessment methods with LQA. Existing assessment methods can be mainly categorized into two types. The first type (Fig. 3(a)) relies on the inherent characteristics of the image, such as brightness, contrast, and other attributes, to calculate the image quality, such as BRISQUE [35]. The second type uses neural network models pre-trained on other dataset (Fig. 3(b)), for example, GIE [9]. The first type is constrained by its incapacity for training, while the second type can not adapt to the target dataset. Inspired by CLIP-IQA [48], we found vision-language model have demonstrated superior performance in assessing the quality of an image, so we design LQA which based

on fine-tuning vision-language model (Fig. 3(c)) for modality quality assessment to reweight the contribution of each modality to extract robust features from dual-modal inputs. The process of LQA is formulated as follows:

$$g = LQA(\mathcal{M}, \mathcal{F}(\mathcal{M}; \theta), \mathcal{T}), \tag{3}$$

where $\mathcal{M} = \{m_1, m_2\} \in \mathbb{R}^{3\times H\times W}$ are a pair of dual-modal images, e.g., RGB and Thermal in Fig. 2. $\mathcal{T}$ are texts; $\mathcal{F}$ is the encoder I and II with parameters θ; g are the output of LQA and fed into the decoder. We first pass m_1 and m_2 through the image encoder of CLIP to obtain the image embedding $\varepsilon_i \in \mathbb{R}^{1\times D}$, where D is the image embedding dimension (D is set to 512). In parallel, the text encoder in CLIP receives inputs $\mathcal{T}$= {*A photo of high quality.*} and outputs text embeddings $\varepsilon_t \in \mathbb{R}^{1\times D}$. To adapt the target dataset, we add a learnable prompt ω to the text embedding ε_t:

$$\varepsilon_i = \varepsilon_i + \omega. \tag{4}$$

In this process, we explore adapting pre-trained vision-language models for quality assessment by adding a small trainable parameter to the text encoder. The parameter ω is trained to enhance the model's robustness against noisy images. To this end, we define the quality score estimated by CLIP on the two modality images as $\alpha^{m_1}, \alpha^{m_2}$:

$$\alpha^{m_1} = sim(\varepsilon_i^{m_1}, \varepsilon_t) = \frac{\varepsilon_i^{m_1} \cdot \varepsilon_t}{||\varepsilon_i||\ ||\varepsilon_t||}, \tag{5}$$

$$\alpha^{m_2} = sim(\varepsilon_i^{m_2}, \varepsilon_t) = \frac{\varepsilon_i^{m_2} \cdot \varepsilon_t}{||\varepsilon_i||\ ||\varepsilon_t||}, \tag{6}$$

where $\varepsilon_i^{m_1}$ and $\varepsilon_i^{m_2}$ are the image embeddings from the CLIP image encoder, sim indicates cosine similarity between ε_i and ε_t. It should be noted that LQA is trained only in stage I, which is expected to improve the ability to recognize noisy images. Its parameters remain fixed and are not updated in the training stage II. Subsequently, we fuse the individual layers of RGB and T features using the α:

$$g_j = g_j^{m_1} * \frac{\alpha^{m_1}}{\alpha^{m_1} + \alpha^{m_2}} + g_j^{m_2} * \frac{\alpha^{m_2}}{\alpha^{m_1} + \alpha^{m_2}}, \tag{7}$$

where $g_j^{m_1}$ and $g_j^{m_2}$ represent the j-th layer features of the first and second modality respectively; $g = \{g_1, ..., g_n\}$ is the fused feature passed into the decoder; α^{m_1} and α^{m_2} are quality score for m_1 and m_2. For the decoder, we perform standard CBAM operations on g_j, then conv and upsample before adding to g_{j+1}. Based on our model, n is set to 5.

3.3 Conditional Dropout for incomplete multi-modal learning

Previous methods for handling missing modalities mostly employ direct modality dropout [51] (Fig. 4(a)) during training, which can improve model performance with missing modalities but significantly degrade performance with complete

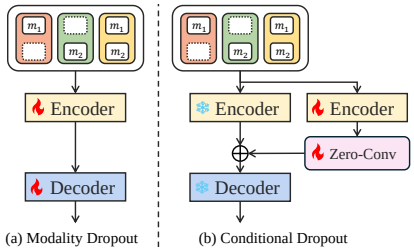

Fig. 4: Architectural comparison of (a) modality dropout [51] and (b) Conditional Dropout in dual-modal object detection.

modalities. We proposed Conditional Dropout (Fig. 4(b)) to preserve the capability in modality-complete scenarios by freezing the pre-trained model and finetuning the trainable copy and zero convolution.

In stage II, we will choose one of the input pairs $\rho(\mathcal{M})$ from the following conditions:

$$\rho(\mathcal{M}) = \begin{cases} Cond_1 := \{m_1, m_2\} \\ Cond_2 := \{m_1,\ \phi\ \}, \\ Cond_3 := \{\ \phi\ , m_2\} \end{cases} \tag{8}$$

where ϕ denotes the zero input, $Cond_1$, $Cond_2$, and $Cond_3$ denote the selection probabilities for three distinct input pairs. The original parameters θ of the encoder from the stage I are frozen and we duplicate the encoder to create a trainable copy Encoder I$'$, Encoder II$'$ with parameters θ_f. Then we connect them using zero convolutions $\mathcal{Z}$ with both weight and bias initialized to zero:

$$g = \mathcal{F}(\rho(\mathcal{M}); \theta) + \mathcal{Z}(\mathcal{F}(\rho(\mathcal{M}); \theta_f); \theta_z), \tag{9}$$

where θ_z is the parameter of the zero convolution. According to Eq. 9, the first term $\mathcal{F}(\rho(\mathcal{M}); \theta)$ represents the frozen model f trained in stage I with parameter θ, which gains the ability to recognize common features from dual modalities. The second term $\mathcal{Z}(\mathcal{F}(\rho(\mathcal{M}); \theta_f); \theta_z)$ denotes the additional parameters θ_f and θ_z introduced by the Conditional Dropout in the second stage. By using incomplete modalities as input during training, we force the model to extract more fine-grained representations from each modality. In summary, the single modality inputs compel the model to learn richer representations recorded in the second term. Meanwhile, the frozen first term preserves the previously learned knowledge. Finally, the $\mathcal{Z}(\cdot;\cdot)$ with zero initialization ensures minimal influence at the beginning of the training stage II, and the newly learned enriched features are progressively integrated into the model during training. The Conditional Dropout framework paves a way to obtain missing data robustness while maintaining accuracy on complete inputs. This framework could strengthen both single-modality and full-modality potentials concurrently.

3.4 Learning objective

Given a pair of images m_1, m_2 and an input prompt $\mathcal{T}$, in the training stage I, we input m_1 and $\mathcal{T}$ into the proposed LQA module, learning to predict scores α for evaluating image quality as shown in Eq. 5. The m_1 and m_2 are also fed into the encoder, and the output $g_i^{m_i}$ are fused by α in Eq. 7. After passing through the decoder, the loss is calculated between the output and GT. During training, We use BCE loss and IOU loss as our loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{bce}(pred, GT) + \mathcal{L}_{iou}(pred, GT). \tag{10}$$

During training stage II, parameters of the initial encoder are frozen, a trainable copy is initialized, and images are fed into both the stage I encoder and the new stage II encoders as per Eq. 9. After feeding into the decoder, the loss is calculated between the output and GT to optimize the parameters of the stage II trainable copy and zero convolution. CoLA focuses on simplicity and extensibility rather than complex model design. The encoder, decoder, and training loss are kept straightforward without meticulous engineering. Combining the above methods, our approach establishes a new evaluation in the dual-modal SOD field, capable of maintaining the model's accuracy under conditions of noisy and missing inputs.

4 Experiments

4.1 Experiment setup

Dataset. For RGB-T, we employ three widely recognized public datasets for our experiment: VT821 [46], VT1000 [45], and VT5000 [44]. In terms of training, we utilized 2500 image pairs from the VT5000 dataset for training purposes. The model's performance under modality-complete and modality-missing was evaluated using the remaining 2500 image pairs from the VT5000 dataset, along with the VT821 and VT1000 datasets.
For RGB-D, we utilize four well-known public datasets for our experiment: SIP [15], NJUK [22], DES [7], and NLPR [39]. For training purposes, we used a combination of 1485 image pairs from the NJUK dataset and 700 image pairs from the NLPR dataset.

Implementation details. We use a single NVIDIA GeForce RTX 3090 GPU. The backbone of all methods utilizes a ResNet-50 [17] model pre-trained on the ImageNet [11] dataset. We use the Adam optimizer with an initial learning rate of 1e-4. In stage I, we train the model for a total of 100 epochs with a batch size of 8 and divide the learning rate by 10 every 45 epochs. In stage II, we conduct a training period covering 60 epochs, starting with an initial learning rate set at 1e-4. This learning rate is then decreased by a factor of 10 after every 35 epochs.

Evaluation metrics. We employ four widely used evaluation metrics to validate

Table 1: Quantitative experiments of different RGB-T models in modality-complete and modality-missing conditions. ↑ denotes that a larger value is better. The best results are in red, the second-best results are in blue, and the third-best results are in green.

Datasets	Conditions		Metric	CSRNet	ADF	TNet	DCNet	LSNet	MIDD	TAGFNet	Ours
	RGB	T		[18]	[44]	[9]	[43]	[60]	[42]	[47]	
VT821	●	●	E_m↑	0.909	0.841	0.919	0.911	0.910	0.901	0.912	**0.922**
			F_β↑	0.837	0.718	0.823	0.841	0.829	0.819	0.825	**0.849**
	○	●	E_m↑	0.720	0.640	0.834	0.750	0.787	0.799	0.855	**0.864**
			F_β↑	0.540	0.465	0.713	0.644	0.687	0.680	0.727	**0.752**
	●	○	E_m↑	0.726	0.765	0.813	0.872	0.841	0.840	0.861	**0.900**
			F_β↑	0.551	0.655	0.740	0.780	0.749	0.737	0.771	**0.817**
	Average Drop		E_m↑	-0.186	-0.139	-0.096	-0.100	-0.096	-0.082	-0.049	**-0.040**
			F_β↑	-0.291	-0.158	-0.097	-0.129	-0.111	-0.110	-0.076	**-0.065**
	Average		E_m↑	0.785	0.749	0.855	0.844	0.846	0.847	0.876	**0.895**
			F_β↑	0.643	0.613	0.759	0.755	0.755	0.745	0.774	**0.806**
VT1000	●	●	E_m↑	0.952	0.928	0.937	0.949	0.951	0.946	0.952	**0.955**
			F_β↑	0.900	0.873	0.885	0.902	0.882	0.884	0.890	**0.904**
	○	●	E_m↑	0.789	0.761	0.896	0.788	0.858	0.886	0.919	**0.930**
			F_β↑	0.634	0.656	0.810	0.723	0.783	0.807	0.834	**0.855**
	●	○	E_m↑	0.781	0.872	0.891	**0.943**	0.913	0.927	0.936	**0.943**
			F_β↑	0.652	0.801	0.839	0.884	0.853	0.860	0.871	**0.887**
	Average Drop		E_m↑	-0.167	-0.112	-0.044	-0.083	-0.065	-0.040	-0.025	**-0.018**
			F_β↑	-0.257	-0.145	-0.061	-0.099	-0.064	-0.050	-0.038	**-0.027**
	Average		E_m↑	0.841	0.854	0.908	0.893	0.907	0.920	0.936	**0.943**
			F_β↑	0.729	0.777	0.845	0.836	0.839	0.851	0.865	**0.882**
VT5000	●	●	E_m↑	0.901	0.887	**0.927**	0.922	0.914	0.906	0.913	**0.927**
			F_β↑	0.818	0.802	0.827	0.822	0.820	0.807	0.819	**0.843**
	○	●	E_m↑	0.744	0.751	0.854	0.693	0.801	0.819	0.869	**0.887**
			F_β↑	0.523	0.603	0.743	0.572	0.689	0.699	0.742	**0.774**
	●	○	E_m↑	0.733	0.820	0.786	0.895	0.847	0.874	0.895	**0.913**
			F_β↑	0.550	0.707	0.717	0.792	0.752	0.764	0.791	**0.822**
	Average Drop		E_m↑	-0.163	-0.102	-0.107	-0.128	-0.090	-0.060	-0.031	**-0.027**
			F_β↑	-0.281	-0.147	-0.097	-0.140	-0.100	-0.076	-0.052	**-0.045**
	Average		E_m↑	0.793	0.819	0.856	0.837	0.854	0.866	0.892	**0.909**
			F_β↑	0.630	0.704	0.762	0.729	0.754	0.757	0.784	**0.813**

the performance of our model and other existing SOTA methods: S-measure(S_α) [13], Mean F-measure (F_β) [1], Mean E-measure(E_m) [14] and Mean Square Error(MAE) [2]. **Average Drop** and **Average**, to measure the robustness and overall capability of a dual-modal SOD model respectively. **Average Drop** refers to the average performance drop when modalities are missing compared to when modalities are complete. **Average** refers to the average performance across the three conditions.

Table 2: Ablation experiments for each component on the VT5000 dataset. The Baseline is the network with a value of α set to 0.5.

	Modality Complete				Missing RGB				Missing Thermal				Average			
	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE$\downarrow$
(a)Baseline	.859	.892	.801	.052	.820	.866	.727	.064	.845	.883	.793	.058	.841	.880	.774	.058
(b)Baseline+LQA	.887	.923	.834	.039	.828	.876	.754	.059	.849	.884	.789	.061	.855	.894	.792	.053
(c)Baseline+CD	.880	.908	.818	.043	.833	.870	.750	.061	.868	.902	.811	.044	.860	.892	.793	.049
(d)Baseline+LQA+CD	**.892**	**.927**	**.843**	**.037**	**.840**	**.887**	**.774**	**.052**	**.874**	**.913**	**.822**	**.042**	**.869**	**.909**	**.813**	**.044**

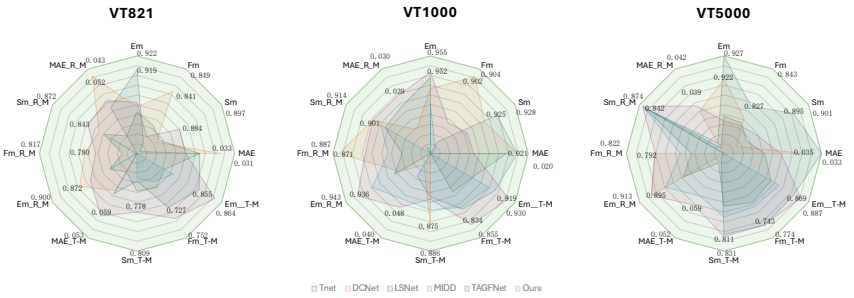

Fig. 5: Radar chart of the proposed model and some existing state-of-the-art methods in three datasets. R-M means the performance of the model under RGB-Missing. T-M means the performance of the model under Thermal-Missing.

4.2 Compare with state-of-the-art

To validate the effectiveness of our CoLA in RGB-T SOD under modality-complete and modality-missing, we compared it with nine state-of-the-art methods from the past three years. The compared methods include ADF [44], MIDD [42], CSRNet [18], DCNet [43], TNet [9], TAGFNet [47] and LSNet [60]. We also compared our CoLA with six state-of-the-art methods in RGB-D from the past three years. The compared methods include D3Net [15], DIGR [6], C^{2}DFNet [56], CIRNet [8], SPSN [25], HiDAnet [52].

Our quantitative comparison results are shown in Table 1. The compared models can be classified into two categories: models like CSRNet [18] and ADF [44] exhibit significant performance degradation when either RGB or Thermal modality is missing, indicating poor robustness in handling modality-missing conditions. Some methods like DCNet [43] perform well when one modality is missing but suffer greatly when the other is absent, indicating an excessive reliance on a particular modality.

Our method excels by achieving the best results in both Average Drop and Average across all datasets, demonstrating unparalleled robustness and overall performance. CoLA also achieved the best performance when dealing with modality-complete as well as when dealing with modality-missing. As shown in Fig. 5, we compared the best five models: MIDD [42], DCNet [43], TNet [9],

Table 3: Compare with other image quality assessment methods for training stage I . The Baseline is the network with a value of α set to 0.5. CLIP-IQA [48] and CLIP-IQA$^+$ [48] refer to the use of frozen CLIP and CLIP fine-tuned with the CoOp [59] method, respectively.

		Baseline	+BRISQUE [35]	+GIE [9]	+CLIP-IQA [48]	+CLIP-IQA$^+$ [48]	+LQA
VT821	$S_\alpha\uparrow$	.844	.854	.870	.869	.878	**.888**
	$E_m\uparrow$	.873	.893	.910	.898	.910	**.915**
	$F_\beta\uparrow$	.805	.797	.812	.796	.830	**.839**
	MAE$\downarrow$	.055	.046	.044	.050	.042	**.038**
VT1000	$S_\alpha\uparrow$	.897	.909	.920	.913	.918	**.924**
	$E_m\uparrow$	.924	.939	.944	.945	.949	**.955**
	$F_\beta\uparrow$	.854	.867	.880	.876	.874	**.904**
	MAE$\downarrow$	.038	.034	.030	.031	.030	**.024**
VT5000	$S_\alpha\uparrow$	.859	.864	.866	.878	.882	**.887**
	$E_m\uparrow$	.892	.901	.914	.916	.915	**.923**
	$F_\beta\uparrow$	.801	.810	.819	.822	.825	**.834**
	MAE$\downarrow$	.052	.048	.047	.042	.040	**.039**

TAGFNet [47], LSNet [60], and our model to establish a radar chart. From the chart, it can be seen that CoLA achieved state-of-the-art performance across all metrics under the three conditions on all datasets.

4.3 Ablation studies

Effectiveness of each component in our model. Table 2 summarizes how progressively incorporating our proposed modules improves model performance under modality-complete and modality-missing. Comparing (a) and (b) shows that adding the LQA module helps CoLA better fuse information from the two modalities to handle noisy images, thus achieving better results under modality-complete condition. Contrasting (c) and (d), the model trained without LQA in stage I will be weaker than the model trained with LQA, which means employing only CD in stage II results in weaker performance compared to (d). By incorporating both LQA and Conditional Dropout, the model significantly enhances its ability to extract valuable information from each individual modality. This fundamentally strengthens modal robustness, enabling CoLA to minimize performance loss when modalities are missing by better utilizing the available single modalities.

The efficiency of LQA. Table 3 compares other image quality assessment methods with LQA. In addition to the No-Reference [35] and Pre-trained quality assessment networks [9]. We also compared the CLIP-IQA [48], which is based on CLIP [40], to assess image quality. The results indicate that compared to using a fixed threshold of 0.5, employing image quality assessment methods such as BRISQUE [35], GIE [9], CLIP-IQA [48], etc., can lead to some performance improvement. However, due to the lack of inherent learning capability and generalization to datasets, these methods have limited impact.

Table 4: Ablation experiments of the Conditional Dropout module on the VT5000 dataset, “Copy" denotes duplicating encoder, “Freeze" denotes freezing modules except the copied encoder, and “MD" denotes Modality Dropout. “Z-Conv” denotes Zero-Convolution.

	Copy	Z-Conv	Freeze	MD	Modality Complete				Missing RGB				Missing Thermal				Average			
					$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓
(a)					.887	.923	.834	.039	.828	.876	.754	.059	.849	.884	.789	.061	.855	.894	.792	.053
(b)	✓				.872	.900	.813	.049	.815	.860	.735	.066	.851	.877	.788	.062	.857	.894	.786	.054
(c)	✓	✓			.878	.910	.819	.046	.820	.863	.733	.062	.847	.875	.785	.066	.857	.894	.786	.054
(d)				✓	.867	.902	.805	.051	.839	.880	.759	.057	.865	.899	.795	.054	.857	.894	.786	.054
(e)	✓	✓		✓	.872	.913	.822	.046	.834	.875	.749	.061	.868	.901	.803	.049	.858	.896	.791	.052
(f)	✓	✓	✓		.890	.923	.836	.038	.822	.869	.743	.064	.855	.889	.795	.056	.856	.894	.791	.053
(g)	✓		✓	✓	.889	.923	.832	.039	.837	.883	.765	.054	.868	.905	.805	.046	.865	.904	.801	.046
(h)	✓	✓	✓	✓	**.892**	**.927**	**.843**	**.037**	**.840**	**.887**	**.774**	**.052**	**.874**	**.913**	**.822**	**.042**	**.869**	**.909**	**.813**	**.044**

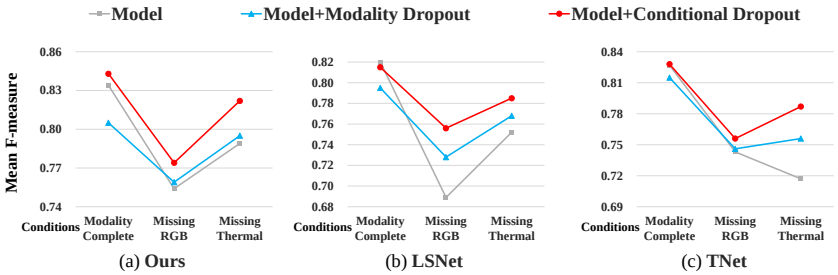

Fig. 6: Comparison of Models Trained with Conditional Dropout, Modality Dropout, and Original Model in VT5000 dataset.

For better demonstrate the effectiveness of LQA, we extract all noisy images like 1(a) line 1 from original VT821 to create a new dataset VT821-noisy with 76 images. Experiments were conducted on this dataset using all compared methods, with results provided in the supplementary materials.

Ablation study for CD. Ablation experiments were conducted in this module to validate the effectiveness of the proposed Conditional Dropout. As shown in Table 4, observing (a) indicates that the original model lacks modality robustness, as evidenced by the significant performance reduction under modality-missing. Comparing (a) and (d) shows that simply applying Modality Dropout can improve model performance under modality-missing, but this leads to worse performance under modality-complete. Contrasting (d) and (e) demonstrates that adding the Copy operation can effectively reduce the performance decline caused by Modality Dropout under modality-complete. A comparison between (e) and (h) reveals that the absence of the freeze operation impacts the encoder's ability to effectively extract features, leading to diminished performance under modality-complete. Analyzing (f) and (h) suggests that employing Copy and Freeze without Modality Dropout does not enhance model robustness.

Table 5: Extended experiments under modality-complete. "AT" denotes additional training for 60 epochs.

	AT	Copy	Z-Conv	Freeze	MD	VT821				VT1000				VT5000			
						$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓
(a)						.888	.915	.839	.038	.924	.955	.893	.024	.887	.923	.834	.039
(b)	✓					.870	.881	.806	.051	.915	.944	.879	.031	.871	.910	.810	.046
(c)	✓				✓	.851	.872	.788	.055	.910	.935	.870	.034	.867	.902	.805	.051
(d)	✓	✓	✓			.878	.905	.832	.043	.922	.948	.879	.029	.878	.910	.819	.042
(e)	✓	✓	✓	✓	✓	**.897**	**.922**	**.849**	**.031**	**.928**	**.955**	**.904**	**.024**	**.892**	**.927**	**.843**	**.037**

Conditional Dropout as a plug-in. We conducted generalization experiments to validate the effectiveness of Conditional Dropout. The results in Fig. 6 show that LSNet [60] and T-Net [9] augmented with Conditional Dropout achieves significantly improved robustness under modality-missing while the performance under modality-complete remains almost unaffected. These experiments validate the generalization of Conditional Dropout in advancing the robustness of dual-modal methods.

Extended experiments under modality-complete with CD. As shown in Table 5, the model demonstrates improved performance under modality-complete after applying Conditional Dropout, especially on the VT821 dataset. Comparing (a) with (b) and (d) reveals that simply extending training epochs or copying parameters fails to enhance performance under modality-complete. Contrasting (c) to (a) indicates that directly utilizing Modality Dropout damages performance under complete modalities. Comparing (c) and (e), it can be observed that Conditional Dropout, as opposed to Modality Dropout, can significantly enhance the model's performance in modality-complete.

Overall, simply increasing model parameters (copying encoders), training epochs, or using Modality Dropout does not improve the model's performance under modality-complete. Conditional Dropout enhances the model's ability to extract information from individual modality, allowing the model to maintain or even improve its performance under modality-complete.

5 Conclusion

In this paper, we proposed a robust dual-modal SOD method to enhance model performance in noisy and modality-missing environments. We adeptly adjust reweighted image contributions by the proposed LQA, which is crucial for noise handling. This method bolsters noise robustness across diverse environments. We further introduce a Conditional Dropout training scheme, effective in processing missing modalities, thus improving model performance with incomplete and complete inputs. Our approach sets a new evaluation in dual-modal SOD, prompting discussions on handling robustness in the SOD field.

6 Acknowledgement

This work was supported by the Natural Science Foundation of Hubei Province of China (No.2024AFB545) and the Fundamental Research Funds for the Central Universities (Grand No.YCJJ20242406). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE transactions on image processing **24**(12), 5706–5722 (2015)
3. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1158–1166 (2018)
4. Chen, C., Wei, J., Peng, C., Qin, H.: Depth-quality-aware salient object detection. IEEE Transactions on Image Processing **30**, 2350–2363 (2021)
5. Chen, T., Yao, Y., Zhang, L., Wang, Q., Xie, G., Shen, F.: Saliency guided inter- and intra-class relation constraints for weakly supervised semantic segmentation. IEEE Transactions on Multimedia (2022)
6. Cheng, X., Zheng, X., Pei, J., Tang, H., Lyu, Z., Chen, C.: Depth-induced gap-reducing network for rgb-d salient object detection: an interaction, guidance and refinement approach. IEEE Transactions on Multimedia (2022)
7. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: Proceedings of international conference on internet multimedia computing and service. pp. 23–27 (2014)
8. Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q., Zhao, Y.: Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. IEEE Transactions on Image Processing **31**, 6800–6815 (2022)
9. Cong, R., Zhang, K., Zhang, C., Zheng, F., Zhao, Y., Huang, Q., Kwong, S.: Does thermal really always matter for rgb-t salient object detection? IEEE Transactions on Multimedia (2022)
10. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y.: Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 626–635. Springer (2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)
13. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)

14. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
15. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: 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 (2020)
16. Gurkan, F., Cerkezi, L., Cirakman, O., Gunsel, B.: Tdiot: Target-driven inference for deep video object tracking. IEEE Transactions on Image Processing **30**, 7938–7951 (2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
18. Huo, F., Zhu, X., Zhang, L., Liu, Q., Shu, Y.: Efficient context-guided stacked refinement network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 3111–3124 (2021)
19. Huo, F., Zhu, X., Zhang, Q., Liu, Z., Yu, W.: Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022)
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
21. John, V., Kawanishi, Y.: Multimodal cascaded framework with metric learning robust to missing modalities for person classification. In: Proceedings of the 14th Conference on ACM Multimedia Systems. pp. 257–265 (2023)
22. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP). pp. 1115–1119. IEEE (2014)
23. Jue, J., Jason, H., Neelam, T., Andreas, R., Sean, B.L., Joseph, D.O., Harini, V.: Integrating cross-modality hallucinated mri with ct to aid mediastinal lung tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22. pp. 221–229. Springer (2019)
24. Kong, Y., Wang, Y., Li, A.: Spatiotemporal saliency representation learning for video action recognition. IEEE Transactions on Multimedia **24**, 1515–1528 (2021)
25. Lee, M., Park, C., Cho, S., Lee, S.: Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 630–647. Springer (2022)
26. Lee, M., Lee, S., Lee, J., Shim, H.: Saliency as pseudo-pixel supervision for weakly and semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
27. Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., Bai, X.: Crowdclip: Unsupervised crowd counting via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2893–2903 (2023)
28. Liu, J., Dian, R., Li, S., Liu, H.: Sgfusion: A saliency guided deep-learning framework for pixel-level image fusion. Information Fusion **91**, 205–214 (2023)
29. Liu, M., Jiang, J., Zhu, C., Yin, X.C.: Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6662–6671 (2023)

30. Liu, N., Luo, Z., Zhang, N., Han, J.: Vst++: Efficient and stronger visual saliency transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
31. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13756–13765 (2020)
32. Liu, N., Zhang, N., Shao, L., Han, J.: Learning selective mutual attention and contrast for rgb-d saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9026–9042 (2021)
33. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4722–4732 (2021)
34. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18177–18186 (2022)
35. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
36. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23507–23517 (2023)
37. Pan, Y., Liu, M., Lian, C., Xia, Y., Shen, D.: Spatially-constrained fisher representation for brain disease identification with incomplete multi-modal neuroimages. IEEE transactions on medical imaging **39**(9), 2965–2975 (2020)
38. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing **32**, 892–904 (2023)
39. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13. pp. 92–109. Springer (2014)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
42. Tu, Z., Li, Z., Li, C., Lang, Y., Tang, J.: Multi-interactive dual-decoder for rgb-thermal salient object detection. IEEE Transactions on Image Processing **30**, 5678–5691 (2021)
43. Tu, Z., Li, Z., Li, C., Tang, J.: Weakly alignment-free rgbt salient object detection with deep correlation network. IEEE Transactions on Image Processing **31**, 3752–3764 (2022)
44. Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y.: Rgbt salient object detection: A large-scale dataset and benchmark. IEEE Transactions on Multimedia (2022)
45. Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J.: Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia **22**(1), 160–173 (2020). https://doi.org/10.1109/TMM.2019.2924578

46. Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B.: Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: Image and Graphics Technologies and Applications: 13th Conference on Image and Graphics Technologies and Applications, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers 13. pp. 359–369. Springer (2018)
47. Wang, H., Song, K., Huang, L., Wen, H., Yan, Y.: Thermal images-aware guided early fusion network for cross-illumination rgb-t salient object detection. Engineering Applications of Artificial Intelligence **118**, 105640 (2023)
48. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
49. Wang, J., Song, K., Bao, Y., Huang, L., Yan, Y.: Cgfnet: Cross-guided fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2949–2961 (2021)
50. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23034–23044 (2023)
51. Wei, S., Luo, C., Luo, Y.: Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20039–20049 (2023)
52. Wu, Z., Allibert, G., Meriaudeau, F., Ma, C., Demonceaux, C.: Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE Transactions on Image Processing **32**, 2160–2173 (2023)
53. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10899–10909 (2023)
54. Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., Bai, X.: Turning a clip model into a scene text detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6978–6988 (2023)
55. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
56. Zhang, M., Yao, S., Hu, B., Piao, Y., Ji, W.: C^2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. IEEE Transactions on Multimedia (2022)
57. Zhang, R., Zeng, Z., Guo, Z., Li, Y.: Can language understand depth? In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6868–6874 (2022)
58. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
59. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
60. Zhou, W., Zhu, Y., Lei, J., Yang, R., Yu, L.: Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. IEEE Transactions on Image Processing **32**, 1329–1340 (2023)
61. Zhou, Z., Pei, W., Li, X., Wang, H., Zheng, F., He, Z.: Saliency-associated object tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9866–9875 (2021)