Modality-Balanced Contrastive Learning for RGB-D Salient Object Detection

Hao Chen, Feihong Shen, and Lichuang Zhang

***Abstract*—The scarcity of labelled data is a key obstacle to sufficient multi-modal fusion. The traditional solution that uses ImageNet-pretrained weights as initialization usually results in biased learning due to the cross-domain/modal gaps. In this work, we overcome these limitations by proposing a self-supervised learning framework. Unlike previous multi-modal contrastive learning architectures that are mainly performed by directly optimizing cross-modal agreement and always struggle with the information imbalance between modalities, especially for hard positive instances with noticeable modal-specific noises, we re-formulate the multi-modal contrastive learning with imbalanced data to a unimodal contrastive problem to remove the modality gap. To this end, we combine the features from all modalities as a new modality representation via selectively masking the discriminative features in the strong modality to rebalance two modalities. By performing contrastive learning on the synthetic modality, we enable sufficient pre-training for each modality and explicit study of cross-modal correlation for downstream multi-modal fusion. Extensive experiments verify the effectiveness of our framework and designs on the downstream RGB-D salient object detection task.**

***Index Terms*—RGB-D salient object detection, cross-modal attention, disentanglement, transformer.**

I. Introduction

Salient object detection (SOD) is to highlight the most visual-attractive objects from a scene. With rich geometry cues and additional contrast hints, the paired depth image has largely advanced the accuracy and robustness of SOD, especially for those challenging scenes with low-contrast appearance or weak lighting. Regarding the great success of deep Convolutional Neural Networks (CNNs), various RGB-D SOD methods based on CNNs have been proposed and achieved impressive improvements. These works usually follow a two-stream architecture, which typically includes two parallel encoders with popular networks (e.g., ResNet [1]) as the backbone to extract unimodal features individually. Then, the focus of these CNN-based methods is to design diverse fusion patterns [2], [3] to boost cross-modal cross-level feature interaction, varying feature integration strategies [4], [5] to fully explore spatial and semantical complements, and some enhancement policies such as the attention mechanism [6], knowledge distillation [5], [7], mutual information minimization [8] and feature disentanglement [9], to explicitly reduce the cross-modal redundancy and improve the fusion adaptivity. Given the scarcity of labelled RGB-D SOD data, these works have to use the weights pre-trained on large-scale datasets (e.g., the ImageNet [10]) as initialization to partly alleviate the data-hungry nature of CNNs. However, annotating such a million-level dataset for pre-training is time-consuming and costly, especially for dense prediction labels. More importantly, it inevitably introduces biased initialization and insufficient learning due to the domain gap between the pre-training dataset and the downstream one, as well as the modality gap between the source (i.e., RGB) and target (i.e., depth) modalities. Hence, this pre-training paradigm is cost-ineffective and holds limited generalization performance.

(a) RGB (b) Depth (c) GT

Fig. 1: The strong RGB images and the weak depth images. In examples, it's difficult to distinguish the salient objects from depth images.

Recently, the self-supervised learning scheme [11] directly tackles the deficiency of labelled data by designing various pretext tasks [12]–[14] and contrastive learning (CL) [15], [16] is one of the most popular paradigm. To date, a majority of contrastive learning frameworks are designed for unimodal data. In their domains, positive pairs are generated by different transformations from the same instance and negative pairs are samples from different scenes. By pulling positive pairs together and pushing negative pairs away, contrastive learning can learn feature representations invariant to certain transformations. Multi-modal contrastive learning is a newly rising problem and the basic philosophy is extending the pretext of within-modal instance discrimination to cross-modal agreement, that is regarding the matching multi-modal instance as the positive sample and unpaired ones as negative samples.

Noticeable works in this line include conducting audio-visual CL to learn unimodal representations [15], [16], image-text CL for cross-modal generation (i.e., image caption) [17] and the Point-RGB CL for 3D point understanding [18]. Differently, our task holds spatial-aligned multi-modal data

Manuscript under review.

Hao Chen and Feihong Shen are with the School of Computer Science and Engineering, Southeast University, Nanjing, China (e-mail: haochen303@seu.edu.cn; feihongshen@seu.edu.cn). Lichuang Zhang is with the School of Computer Science and Engineering, Nanjing University, Nanjing, China (zhanglc@lamda.nju.edu.cn).

and the downstream is multi-modal joint inference, which requires each modality to be well learned. Also, it will introduce remarkable benefits for the downstream task if cross-modal correlation can be explicitly explored in the CL process. As shown in Fig. 1, RGB and depth images hold huge differences of data distribution and structure and modal-specific noises. Also, RGB usually carries more and stronger cues than the paired depth. With such a large cross-modal gap, it is difficult to learn sufficient representation from each modality by optimizing the cross-modal agreement. The contrastive loss may be optimized by merely learning the low-level spatial alignment (such as some shared object edges), thus failing to extract high-level semantics from each modality. An alternative solution as done in [19], [20] is to combine multi-modal data as a joint instance to generate positive/negative samples by controlling the component distribution. In their settings, negative samples are crafted by carefully tuning the distribution of unaligned modalities with hyper-parameters, which is laborsome and the resulting negative samples with different distributions actually hold varying similarities to the anchor.

Instead, we propose a new multi-modal contrastive learning paradigm, namely Modality-balanced Contrastive Fusion, to carefully pre-train each modality and cross-modal correlation in RGB-D pairs. First, we combine the aligned RGB-D pairs as a new synthetic modality for instance discrimination. Such a simple scheme fills the modality gap and enables the encoders to learn high-level semantics to differentiate input scenes. More importantly, it is independent of any hyper-parameters to set the modality distribution and largely ease the work of generating negative samples. However, it will still lead to biased sub-optimization by direct concatenating representations from all modalities as the joint instance. Due to the modality imbalance shown in Fig. 1, the network may depend on the stronger modality to discriminate scenes. To avoid the ignorance of the weak depth modality, we manually weaken the strong RGB modality by masking its most discriminative features in the modality synthesis process. In doing so, the depth modality has to learn more supplementary cues for scene discrimination. In summary, the main contributions of this work are as follows:

- We propose a new contrastive framework for self-supervised pre-training on imbalanced multi-modal data.
- We reformulate the cross-modal agreement to instance discrimination scheme to remove the modality gap by introducing an attentive mask modality-synthesis method.
- Comprehensive experiments on wide public datasets of RGB-D salient object detection consistently demonstrate the efficacy of our designs and the large improvement over state-of-the-art supervised learning methods.

II. Related Work

A. RGB-D Saliency Objection Detection

Initial RGB-D SOD methods utilize handcrafts cues [21]–[24] to infer saliency by local/global contrasts. However, these handcrafts features are weak in high-level semantics and hold limited generalization ability. The great success of deep convolutional neural networks (CNNs) motivates the RGB-D SOD community to design CNN-based detectors and generate various noticeable architectures. Single-stream models [25] that fuse the modalities directly and adopt an shared encoder to extract joint features. Two-stream architecture is a dominated one, which [2] use two encoders to extract features from each modality and designs various multi-scale cross-modal interactions for informative multi-modal fusion. TANet [4] disigns a three-stream architecture for RGB-D SOD, which considers the cross-modal complementarity in the encoder additionally to solve the issue that the two-stream architecture only focuses the multi-modal representation fusion in the decoder process.

Generally speaking, these RGB-D SOD methods are equipped with diverse cross-modal cross-level combination paths. To alleviate the deficiency of labelled data in training, they resort to the models well pre-trained on the large-scale labelled dataset ImageNet [10] to initialized their encoders. However, these pre-trained encoders are not appropriate to extract depth features due to the large cross-modal gap between RGB and depth. To alleviate the dependence on ImageNet pre-trained weights, [26] proposes to use cross-modal generation and depth contour estimation as pretexts to pre-train encoders for each modality and cross-modal interactions. However, their SSL method is not end-to-end and relies on manually generating pretext labels from the source depth maps. In contrast to designing pretext labels and tasks, we adopt the contrastive learning paradigm and merely use the naturally source data to implement end-to-end SSL. Also, our motivation is to solve the modality imbalance in multi-modal contrastive learning, which significantly differs from [26].

B. Multi-modal Contrastive learning

As one of the typical self-supervised methods, contrastive learning has been widely used in tasks with single modality. For example, PIRL [27] maintains a memory bank to store previously-computed representations, therefore a positive sample can compare with negative samples quantitively. To address the expensive cost in maintaining a memory bank, MOCO [15] stores the representations with a dynamic queue. SimCLR [16] further simplifies the memory bank and designs an end-to-end framework along with a large batch-size.

For multi-modal data, several contrastive attempts have achieved noticeable results [18], [28]–[30]. Tian et al. propose multiview contrastive learning to maximize the mutual information between different views of the same scene. Morgado [29] proposes to learn audio and visual representations by optimizing audio-visual cross-modal agreement. Alayrac et al. [30] leverage visual, audio, and text by a multimodal versatile network and exploit semantic comparisons of them while CrossPoint [18] try to capture the compositional correlations between 3D objects and 2D images to benefit 3D point cloud understanding. However, these methods ignore the cross-modal gap and imbalance in contrastive learning, thereby typically resulting in partial or biased pre-training in the weak modality.

To address this problem, [19] designs a TupleInfoNCE for visual representation learning, which disturb the modality

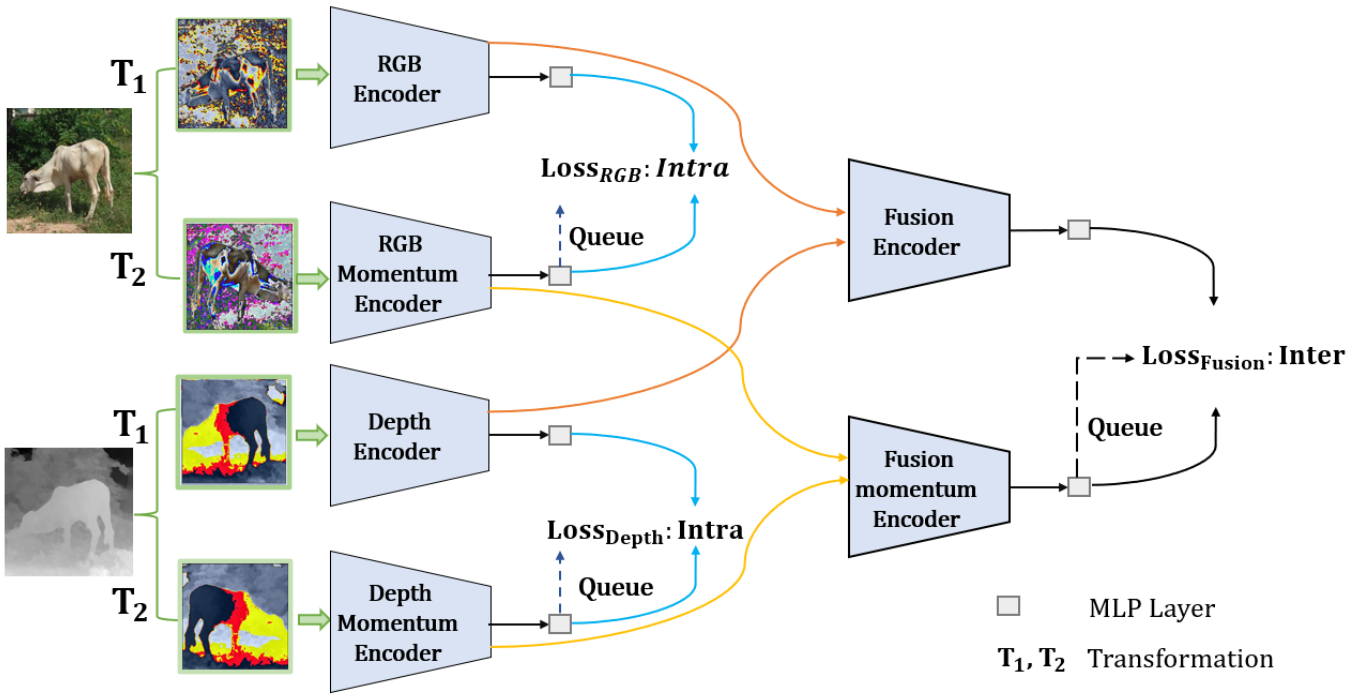

Fig. 2: The overall architecture of our proposed method.

combination to avoid the domination of the strong modality. However, TupleInfoNCE remains two problems to be solved. 1. TupleInfoNCE requires to careful tune the hyper-parameters for tuple disturbing and augmentation, thus being difficult to well balance modal-shared and modal-specific representations; 2. Their negative samples are heterogeneous and hold varying distances to the anchor, while they fail to take it into account. Instead, we reformulate the unbalanced cross-modal agreement objective to synthetic unimodal instance discrimination by designing a fusion encoder to integrate paired multi-modal data. Besides, we alleviate the domination of the strong modality by proposing a attentive mask to explicitly weaken the contribution of the strong modality when combining multiple modalities in the fusion encoder, forcing the contrastive loss to pay more attention on exploring the representations from the weak modality.

III. Method

A. The Overall Architecture

The overall architecture of our proposed Modality-balanced Contrastive Fusion framework is shown in Fig. 2. Compare with the traditional unimodal contrastive learning methods, we use two encoders to extra features from RGB and depth respectively and the corresponding representations of positive samples are generated by momentum encoders [15]. The pre-training model contains two significant contrastive learning parts: intra-modal contrastive learning module (IntraCLM) and inter-modal contrastive learning module (InterCLM) with different objectives. The IntraCLM adopts MoCo [15] as the baseline to learn modal-specific features while the InterCLM, using a synthetic modality combined by paired multi-modal images as the contrastive instance is performed to remove the cross-modal gap and capture cross-modal correlations via a fusion encoder, where a channel attentive masking module is inserted to encourage the learning of the weak modality.

B. Intra-modal Contrastive Learning Module

The Intra-modal Contrastive Learning Module (IntraCLM) involves two separate contrastive learning stages: intra-RGB and intra-depth. We denote the multi-modal RGB-D dataset as $D = \{(r_1, ..., r_K, d_1, ..., d_K)\}$ where r represents RGB and d represents depth. Two random data augmentations are adopted for each anchor to generate two augmented views of the same sample. Then, they are input into the encoders θ_q and θ_k (denoted as momentum encoder) and additional MLPs ϕ_{r_q}, ϕ_{r_k} are taken to make projections which could benefit the learning progress [16]. In the dictionary look-up [15], we describe the former as query and the latter as key. The whole process can be expressed by the following equation:

$$q_j = f_q(I_j; \theta_q, \phi_q), \tag{1}$$

$$k_j = f_k(I_j; \theta_k, \phi_k), \tag{2}$$

where I_j is the input of the encoder (RGB or depth).

After that, InfoNCE [31] is adopted as our intra-modal loss:

$$\mathcal{L}(q, k_+, k_i) = -log\frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} exp(q \cdot k_i/\tau)}, \tag{3}$$

where K represents the size of the memory bank which consists of negative samples.

The total intra-modal loss includes the part in both the RGB and depth modalities and can be formulated as:

$$\mathcal{L}_{intra} = \mathcal{L}(q^r, k_+^r, k_i^r) + \mathcal{L}(q^d, k_+^d, k_i^d). \tag{4}$$

C. Inter-modal Contrastive Learning Module

After our pre-training model have learned modal-specific features by IntraCLM, we append the Inter-modal Contrastive Learning Module (InterCLM) to learn cross-modal correlations and fuse the features from two modalities through a specifically designed fusion encoder f_{F_q}. Accordingly, the keys are fused through a fusion momentum encoder f_{F_k} which have the same structure as the fusion encoder. In doing so, the complementarity and cross-modal high-level semantic features of the two modalities can be well explored. For a paired sample of RGB-D data (r_j, d_j), we first pass it through two basic encoders f_q and f_k to generate intra-modal specific features q and k, then transfer feature maps to generate fused modalities features via a fusion encoder. As same as the IntraCLM, additional MLPs ϕ_{F_q}, ϕ_{F_k} are taken to make projections in the latent space. The formulation is shown by the following equation:

$$q_j^F = f_{F_q}((q^r, q^d); \theta_{F_q}, \phi_{F_q}), \tag{5}$$

$$k_j^F = f_{F_k}((k^r, k^d); \theta_{F_k}, \phi_{F_k}), \tag{6}$$

For each query q^F, and key k_+^F from the same anchor, we can formulate the inter-modal contrastive loss as follows:

$$\mathcal{L}_{inter} = -log\frac{exp(q^F \cdot k_+^F/\tau)}{\sum_{i=0}^{K} exp(q^F \cdot k_i^F/\tau)}, \tag{7}$$

The total loss combined with InterCLM and IntraCLM can be formulated as:

$$\mathcal{J} = \lambda_1 * \mathcal{L}_{intra} + \lambda_2 * \mathcal{L}_{inter} \tag{8}$$

where λ_1 and λ_2 are trade-off hyper-parameters and we set them as 0.5 equally without further tuning.

D. Fusion Encoder

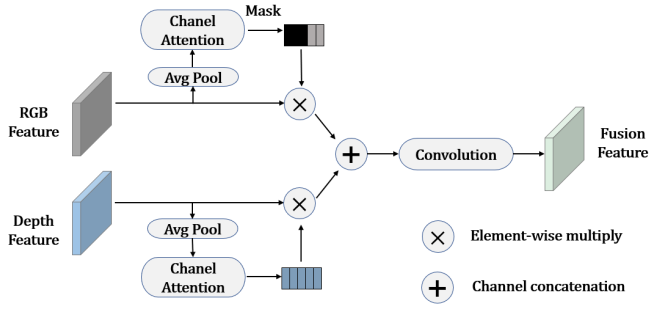

Fig. 3: The architecture of fusion encoder in the inter-modal contrastive learning module.

To remove the cross-modal gap and encourage sufficient pre-training for both the strong and weak modalities in the contrastive learning process, we propose a fusion encoder in the InterCLM to convert cross-modal agreement to instance discrimination in a synthetic modality. Figure. 3 shows the structure of the fusion encoder. Firstly, the global average pooling layers are taken to remove the redundant noises while reducing the computational costs. After that, we adopt the channel attention operations to enforce our module's focus on meaningful areas of the input feature maps. At last, the features are concatenated in the latent space and passed through the convolution layers to obtain the fused representation.

E. Attentive Mask for the Strong Modality

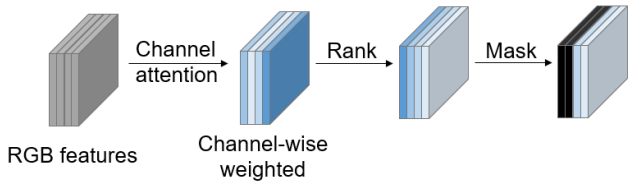

Fig. 4: The details of mask for the strong modality.

To make full use of the weak modality (depth), we design an attentive mask for the strong modality. The specific details of the mask process is shown in Fig. 4, when computing the RGB channel attention, we have c channels of RGB features $f_r^1, f_r^2, ..., f_r^c$. We first sort the channels by attention scores and the sorted features are $f_r^{1'}, f_r^{2'}, ..., f_r^{c'}$. Then we mask a certain proportion of the features with the largest weights, i.e., set the important RGB features to 0. The remaining features $f_r^{m'}, f_r^{m+1'}, ..., f_r c'$ denotes as $mask(f_r')$. The fusion features can be calculated as follow formulas:

$$F_{fused} = Conv(F_r \otimes mask(f_r') \oplus F_d \otimes F_d'), \tag{9}$$

where $\otimes$ denotes element-wise multiplication, $\oplus$ means channel concatenation, Conv is the convolution layer. In this way, the depth modality is forced to learn more discriminative features to complement the masked RGB features in the fusion process to guarantee the representation ability learned from the fusion encoder.

F. Suppression and enhancement on learning gradient

In addition to masking the dominant modality, adjusting the learning gradient on downstream tasks [32], [33] has been demonstrated to be effective in mitigating modality imbalance. Specifically, we adopt two encoders $\phi^r(\theta^r)$ and $\phi^d(\theta^d)$, where θ^r and $theta^d$ denotes the parameters of rgb encoder and depth encoder respectively. To simplify the proof, we assume that after concatenating the features of two modalities, they are directly passed through a linear layer to output the final logits output. This process can be represented as:

$$f(x_i) = W(\phi^r(\theta^r, x_i^r); \phi^d(\theta^d, x_i^d)) + b, \tag{10}$$

After dividing W into [W^r,W^d], Equation 10 can be rewritten as:

$$f(x_i) = W^r \cdot \phi^r(\theta^r, x_i^r) + W^d \cdot \phi^d(\theta^d, x_i^d) + b, \tag{11}$$

With the Gradient Descent optimization method, W^r and the parameters of encoder $\phi^r(\theta^r)$ are updated as:

$$\begin{aligned} W_{t+1}^r &= W_t^r - \eta \nabla_{W^r} L(W_t^r) \\ &= W_t^r - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \phi^r(\theta^r, x_i^r), \end{aligned} \tag{12}$$

$$\begin{aligned}\theta_{t+1}^r &= \theta_t^r - \eta \nabla_{\theta^r} L(\theta_t^r) \\ &= W_t^r - \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial f(x_i)} \frac{\partial (W_t^r \cdot \phi_t^r(\theta^r, x_i^r))}{\partial \theta_t^r},\end{aligned} \tag{13}$$

where η is the learning rate. Observe the Equation 12 and 13, the optimization of W and ϕ is only related to the training loss $\frac{\partial L}{\partial f(x_i)}$, which can be rewritten as:

$$\frac{\partial L}{\partial f(x_i)_c} = \frac{e^{(W^r \cdot \phi^r(\theta^r, x_i^r) + W^d \cdot \phi^d(\theta^d, x_i^d) + b)_c}}{\sum_{k=1}^{M} e^{(W^r \cdot \phi^r(\theta^r, x_i^r) + W^d \cdot \phi^d(\theta^d, x_i^d) + b)_k}} - 1_{c=y_i}. \tag{14}$$

As shown in Fig. 1, rgb modality is more confident to contribute to $\frac{\partial L}{\partial f(x_i)_{y_i}}$ by $W^r \cdot \phi_i^r$, leading to domination of rgb modality in downstream tasks. Based on the observations above, we speculate that providing weak-modality-dominated pre-training parameters for the encoders will alleviate the problem of downstream modality imbalance. We propose a method focus on pre-training process to suppress the gradient for strong modality and enhance the gradient for weak modality. Specifically, We separate the parameters of RGB backbone and depth backbone from all the parameters. To prevent RGB modality from dominating the downstream task and to promote the comprehensive utilization of the depth modality, we maintain the base learning rate η=0.007 for other parameters, and adjust the RGB learning rate to $\frac{1}{2}\eta$ and depth learning rate to 2η in pre-training process.

IV. Experiments

A. Datasets

For pre-training, we evaluate the proposed model on five public RGB-D SOD datasets which are NJUD [37] (1985 image pairs), NLPR [38] (1000 image pairs), DUT [39] (1200 image pairs), ReDWeb-S [40] (3179 image pairs) and COME [8] (15626 image pairs). We follow the consistent setting in previous works and choose 1485 image pairs in NJUD, 700 image pairs in NLPR, 800 image pairs in DUT, 2179 image pairs in ReDWeb-S and 8025 pairs image in COME15K as the training set and the remaining samples form the testing set. Labels are not used in this stage.

For the downstream task, we adopt the same training set as previous methods [5], [26], i.e., 800 samples from the DUT, 1485 samples from the NJUD and 700 samples from the NLPR are used for training. The test set involves NJUD, NLPR, DUT, STERE [41] (1000 image pairs), SIP [42] (929 image pairs), COME-E [8] (3000 image pairs) and COME-H [8] (2000 image pairs).

B. Evaluation Metrics

We adopt three widely used evaluation metrics to evaluate our model performance. Specifically, Maximum F-measure [43] jointly considers precision and recall under the optimal threshold:

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{15}$$

where β is set to 0.3 as suggested by Liu [44]. Maximum enhanced-alignment measure Mean Absolute Error (MAE) computes pixel-wise average absolute error:

$$MAE = \frac{1}{W * H} \sum_{i=1}^{W} \sum_{j=1}^{H} \mid S_{i,j} - G_{i,j} \mid, \tag{16}$$

where y_i and $y_i^{'}$ represent the ground truth pixels and the pixels of the predicted saliency graph respectively. E_ρ^{max} [45] simultaneously considers pixel-level errors and image-level errors:

$$E_{max} = \frac{1}{W * H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_{FM}(i,j). \tag{17}$$

C. Implementation Details

The experiment is composed of upstream pre-training and downstream fine-tuning for the RGB-D SOD task. The parameters of RGB encoder and depth encoder are obtained by pre-training learning, and then applied to the downstream tasks. To clearly show the advantages of our pre-training scheme, the model parameters (RGB and depth encoders) obtained by pre-training are frozen, and other parameters of the downstream model are updated by reverse propagation when the downstream task is fine-tuned.

Upstream contrastive learning experiment: for the intra-modal contrastive learning module, we adopt ResNet50 as the encoder for two modalities. The data augmentation methods for input samples include random clipping, grayscale transformation and random flip. The intra-modal temperature coefficient τ_{intra} is set to 0.1. We use the stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0002 for optimization. The batch size is set to 64, epoch is 50, queue size is 16384. We train the model on 2 Tesla V100.

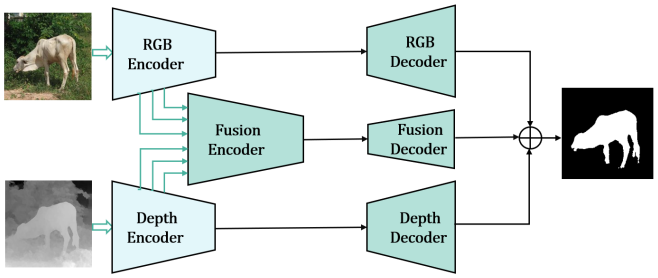

Fig. 5: The architecture of downstream RGB-D SOD model.

Downstream RGB-D SOD task: the architecture of downstream RGB-D SOD model is shown in Fig. 5. We follow [3] to set the batch size to 16, learning rate to 0.0001, and train it on Tesla V100 with 100 epochs.

D. Comparison with State-of-the-art

In order to verify the effectiveness of the proposed multi-modal contrastive learning framework, we compare it with four supervised SOTA models and two other popular training strategies random initialization (denoted by "Random") and

TABLE I: Comparisions with state-of-art methods on 5 datasets. 'Supervised' denotes the backbones are pre-trained on Imagenet, 'Scratch' denotes the backbones are initialized with random parameters, 'Self-supervised' means the backbones are trained with our schema without any annotations.

Dataset	Metrics	CoNet [34]	BBS-Net [35]	DCF [3]	SPSN [36]	ours Supervised	ours Scratch	ours Self-supervised	BBS-Net Self-supervised
NJUD	$maxF\uparrow$	0.893	**0.919**	0.913	**0.920**	0.905	0.862	0.894	0.913
	$E^{max}\uparrow$	0.937	**0.949**	**0.949**	**0.950**	0.942	0.914	0.936	0.945
	$MAE\downarrow$	0.046	**0.035**	0.038	**0.032**	0.041	0.057	0.043	0.041
NLPR	$maxF\uparrow$	0.893	**0.918**	**0.923**	0.910	0.899	0.841	0.887	0.909
	$E^{max}\uparrow$	0.948	**0.961**	**0.961**	0.958	0.955	0.896	0.946	0.957
	$MAE\downarrow$	0.027	**0.023**	**0.023**	0.024	0.026	0.039	0.029	0.026
DUT	$maxF\uparrow$	**0.932**	0.870	0.920	0.858	0.914	0.828	0.900	**0.931**
	$E^{max}\uparrow$	**0.959**	0.912	**0.947**	0.907	0.944	0.896	0.935	0.946
	$MAE\downarrow$	**0.029**	0.058	0.032	0.053	0.038	0.067	0.043	**0.032**
STERE	$maxF\uparrow$	0.901	0.903	**0.907**	0.900	0.892	0.872	0.886	**0.904**
	$E^{max}\uparrow$	**0.947**	0.942	0.931	0.943	0.938	0.916	0.936	**0.945**
	$MAE\downarrow$	**0.037**	0.041	**0.037**	**0.035**	0.046	0.048	0.046	0.042
SIP	$maxF\uparrow$	0.873	0.884	**0.887**	0.886	0.882	0.842	0.874	**0.890**
	$E^{max}\uparrow$	0.917	**0.922**	0.920	**0.924**	0.919	0.895	0.915	**0.922**
	$MAE\downarrow$	**0.048**	0.055	**0.051**	0.052	0.056	0.074	0.059	**0.051**
COME-E	$maxF\uparrow$	0.831	**0.837**	-	0.831	0.742	0.613	0.722	**0.839**
	$E^{max}\uparrow$	0.883	**0.889**	-	**0.889**	0.822	0.739	0.810	**0.885**
	$MAE\downarrow$	0.071	0.070	-	**0.062**	0.126	0.163	0.127	**0.069**
COME-H	$maxF\uparrow$	**0.795**	0.789	-	0.791	0.710	0.599	0.693	**0.794**
	$E^{max}\uparrow$	**0.840**	0.839	-	**0.844**	0.781	0.709	0.775	0.839
	$MAE\downarrow$	**0.102**	0.105	-	0.094	0.156	0.200	0.162	**0.102**

TABLE II: Ablation study of training intra-modal only and training intra-modal and inter-modal together.

Settings	NLPR		DUT		COME-E		COME-H	
	$maxF\uparrow$	$MAE\downarrow$	$maxF\uparrow$	$MAE\downarrow$	$maxF\uparrow$	$MAE\downarrow$	$maxF\uparrow$	$MAE\downarrow$
Intra-modal	0.845	0.039	0.828	0.069	0.531	0.200	0.523	0.231
Intra-modal+Inter-modal	0.867	0.033	0.845	0.061	0.579	0.180	0.567	0.213

Imagenet-pretraining (denoted by "Supervised") on five RGB-D SOD datasets. Our downstream is similar to DCF but without the pre-process depth calibration block. Tab. I shows that the performance of our proposed method with an simple downstream network is close to the current supervised SOTA models, especially on the NJUD datasets. Fig. 7 shows the superiority of our detection results in localizing the salient object and highlight its details.

Compare with the random initialization, our self-supervised learning brings significant improvement to the downstream RGB-D SOD task, which illustrates that our self-supervised learning framework can effectively learn the intra-modal and inter-modal features.

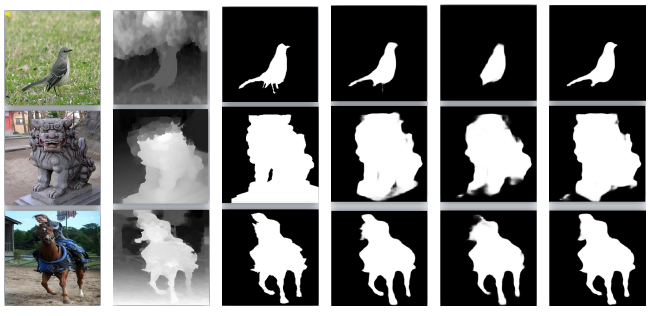

Fig. 6: The comparision of saliency maps generated by our proposed model, intra-modal encoder and supervised model.

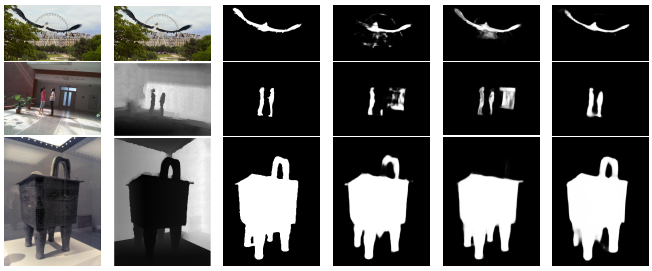

Fig. 7: Comparison with other SOTA methods. "Our" denotes the model adopt two-stage strategy, fusion encoder and mask for strong modality. From the figure we can easily draw the conclusion that with our proposed methods, the downstream network has stronger comprehension ability to catch the saliency objects, no matter in locating or in details.

E. Ablation Study

In this section, we verify the effectiveness of each proposed structure by ablation experiments, including intra-modal and inter-modal contrastive learning paths, the attentive mask module for the strong modality and the two-stage pre-raining strategy.

Effectiveness of our contrastive learning framework. In order to verify the effectiveness of the proposed contrastive learning framework, we compare the traditional contrastive

TABLE III: Ablation experiments of two-stage training strategy and mask for strong modality.

Dataset	Metrics	One-stage				Two-stage			
		No mask	25%mask	50%mask	50%mask+GA	No mask	25%mask	50%mask	50%mask+GA
DUT [39]	$maxF \uparrow$	0.845	0.853	0.864	0.861	0.900	0.899	0.902	0.900
	$E^{max} \uparrow$	0.905	0.910	0.919	0.917	0.935	0.937	0.935	0.935
	$MAE \downarrow$	0.061	0.058	0.054	0.055	0.045	0.043	0.043	0.043
COME-E [8]	$maxF \uparrow$	0.579	0.581	0.594	0.603	0.699	0.705	0.710	0.722
	$E^{max} \uparrow$	0.717	0.720	0.727	0.736	0.790	0.795	0.803	0.810
	$MAE \downarrow$	0.180	0.180	0.169	0.176	0.136	0.13	0.128	0.127
COME-H [8]	$maxF \uparrow$	0.567	0.571	0.582	0.589	0.656	0.676	0.684	0.693
	$E^{max} \uparrow$	0.692	0.695	0.701	0.707	0.760	0.762	0.771	0.775
	$MAE \downarrow$	0.213	0.210	0.201	0.208	0.170	0.167	0.162	0.162

method with ours. Specifically, traditional contrastive learning only focuses on intra-modal and extracts unimodal features respectively, which is denoted as "Intra-modal" in Tab. II. Our method consider the complementarity in multi-modal encoders and extracts their joint features, which is denoted as "Intra-modal+Inter-modal" in Tab. II. We can easily find that training the intra-modal and inter-modal modules together brings significant improvement over using the intra-modal module only. Fig. 6 well shows the benefits of our (Intra+Inter) pre-training scheme, which achieves better details than Intra-model only and comparable to thesupervised learning model. Hence, for multi-modal representation learning, it is highly recommended to additionally learn the complementary cross-modal features rather than learning the representation of each single modal respectively. The proposed model combineing the intra-modal and inter-modal contrastive learning modules, can successfully learn modal-specific representations and the cross-modal complementary representations and their combinations at the same time.

Effectiveness of two-stage training strategy. To effectively explore the complementarity between two modalities, we propose a two-stage training strategy, i.e., we first train the intraCML to keep the independence of each modality and then train the intraCML and interCML together to catch the correlation between two modalities. Specifically, line "one-stage" in Tab. III means the variant that we trains the whole model together with 150 epoches at a learning rate of 0.07, line "two-stage" denotes the variant that we train the intraCML first with 200 epoch at a learning rate of 0.03 and train both the intraCML and interCML with 100 epoch at a learning rate of 0.05. These parameters perform best in their corresponding training processes. Experiment results in Table 3 show that two-stage training strategy is largely better than one-stage, which illustrates that learning well the unimodal features plays a decisive role in multi-modal representation learning. The cross-modal representations can be effectively learned only if the single model representation is strong enough.

Effectiveness of mask for strong modality. To prevent the strong modality from dominating the Inter-modal CL process and sufficiently explore the weak modality as well as the complements between two modalities, we propose an attentive mask operation for the strong modality. Specifically, we mask 0%, 25% and 50% most discriminative channels in RGB features and compare their performance on the downstream RGB-D SOD task. Tab III shows that the attentive mask operation improves the performance consistently in one-stage or two-stage training settings, which verifies its efficacy in encouraging the weak modality to learn more complementary information and promoting the downstream performance. Also, we find that masking with 50% performs better than 25%.

Effectiveness of gradient adjustment. Similar to mask for strong modality, to alleviate the dominance of the strong modality in downstream tasks, we propose adjusting gradients during the pretraining process. Tab III shows that gradient adjustment promote the SOD performance, especially in two large datasets COME-E and COME-H. Gradient adjustment balance the multi-modal learning process from the aspect of gradient, and achieve significant improvement for RGB-D SOD task.

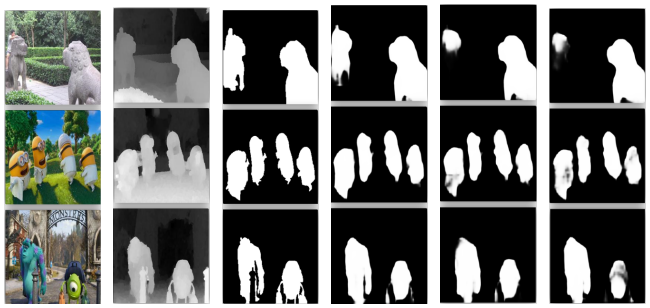

Fig. 8: The visualization of ablation experiments about fusion encoder, two stage strategy and mask for strong modality.

F. Compare with Traditional Alignment Methods

Traditional multi-modal contrasitive learning methods extract the features respectively and adopt cross-modal alignment [17], [19], [47] as contrastive objective by shortening the distance of multi-modal features from the same anchor scene. Line "Alignment" in Tab. IV denotes that the alignment method and "Fusion+50%mask" denotes our proposed modality-balanced CL with a attentive fusion encoder. They have the same settings in intraCLM with 0.1 temperature coefficient. For interCLM, the best temperature coefficient is 0.1 for alignment, and 0.04 for ours. Tab. IV shows that our proposed modality-balanced contrastive fusion scheme outperforms the alignment largely on all datasets.

Although the alignment reduces the distance of similar representations from different modalities, it also closes the

TABLE IV: Comparison between the traditional alignment and our fusion contrastive schemes.

Datasets	Metrics	One-stage		Two-stage	
		Alignment	Our scheme	Alignment	Our scheme
DUT	$maxF \uparrow$	0.859	0.861	0.895	0.900
	$E^{max} \uparrow$	0.916	0.917	0.931	0.935
	$MAE \downarrow$	0.056	0.055	0.045	0.043
COME-E	$maxF \uparrow$	0.589	0.603	0.698	0.722
	$E^{max} \uparrow$	0.724	0.736	0.799	0.810
	$MAE \downarrow$	0.178	0.176	0.130	0.127
COME-H	$maxF \uparrow$	0.578	0.589	0.680	0.693
	$E^{max} \uparrow$	0.699	0.707	0.769	0.775
	$MAE \downarrow$	0.210	0.208	0.166	0.162

TABLE V: Ablation study of our designs on RGB-T Salient Object Detection. Scratch means the backbone is trained from scratch, GA denotes the gradient adjustment operation, AMSM denotes the Attentive Mask for the Strong Modality, Ours denotes all designs we proposed.

Settings	VT5000 [46]		
	$maxF \uparrow$	$E^{max} \uparrow$	$MAE \downarrow$
Scratch	0.802	0.904	0.052
Ours-GA-AMSM	0.807	0.906	0.051
Ours-GA	0.811	0.908	0.050
Ours	**0.818**	**0.912**	**0.048**

distance between two nearest but unpaired samples, especially for hard negative samples, which further lowers the discriminative ability of model. Therefore, we adopt fusion operation to combine two modal representations, and drive the fused representation close to positive samples and away from negative samples, which prevents hard negetive samples from resistance to reduce inter-modal differences. Fig. 8 also shows the advantages of our CL framework, which generates more uniform and correct salient object detection.

G. Extend to RGB-T Salient Object Detection

Thermal infrared spectrum provides complementary cues as depth does in salient object detection task. To illustrate the scalability of our framework, we extend the methods to RGB-Thermal modality. The experiments are performed on the VT821 [48] and VT5000 [46] dataset. VT821 includes 821 spatially aligned RGB-T image pairs and VT5000 has 5000 spatially aligned RGB-T image pairs with ground truth annotations (2500 pairs for training and 2500 pairs for testing). For pretraining, we select 821 RGB-T pairs in VT821 and 2500 RGB-T piars in VT5000 as training dataset. For downstream RGB-T SOD, we select 2500 pairs from VT5000 for finetuning and the other 2500 piars for testing. Other super-parameters are consistent with RGB-D experiment as discussed in Section IV-C. The base RGB-T SOD model in our experiment is TNet [49].

Tab V shows the ablation study of our designs on RGB-T Salient Object Detection. The evaluation metrics are the same with RGB-D SOD tasks which is discussed in Section IV-B. We find our proposed methods also remains effective in RGB-T SOD tasks.

V. Conclusion

In this work, we propose a new CL for multi-modal dense prediction. We propose to combine multi-modal pairs as a synthetic modality to remove the modality gap, and an attentive masking module to rebalance modalities for sufficient training of each modality and exploration of cross-modal correlations and informative cross-modal fusion.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3051–3060.

[3] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu *et al.*, "Calibrated rgb-d salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9471–9481.

[4] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 6, pp. 2825–2835, 2019.

[5] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9060–9069.

[6] Y. Zhou, S. Huo, W. Xiang, C. Hou, and S.-Y. Kung, "Semi-supervised salient object detection using a linear feedback control system model," *IEEE Transactions on Cybernetics*, vol. 49, no. 4, pp. 1173–1185, 2018.

[7] H. Chen, Y. Li, Y. Deng, and G. Lin, "Cnn-based rgb-d salient object detection: Learn, select, and fuse," *International Journal of Computer Vision (IJCV)*, 2021.

[8] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "Rgb-d saliency detection via cascaded mutual information minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4338–4347.

[9] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "Rgbd salient object detection via disentangled cross-modal fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 8407–8416, 2020.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *International Conference on Machine Learning (ICML)*, 2008.

[12] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *Computer Vision and Pattern Recognition*, 2016.

[13] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," *European Conference on Computer Vision (ECCV)*, 2018.

[14] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," *European Conference on Computer Vision*, 2016.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[17] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[18] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9902–9912.

[19] Y. Liu, Q. Fan, S. Zhang, H. Dong, T. Funkhouser, and L. Yi, "Contrastive multimodal fusion with tupleinfonce," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 754–763.

[20] Y. Liu, L. Yi, S. Zhang, Q. Fan, T. Funkhouser, and H. Dong, "P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding," *arXiv preprint arXiv:2012.13089*, 2020.

[21] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 101–115.

[22] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," *Procedings of the British Machine Vision Conference 2013*, 2013.

[23] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 92–109.

[24] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 819–823, 2016.

[25] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[26] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised pretraining for rgb-d salient object detection," in *AAAI Conference on Artificial Intelligence*, vol. 3, 2022.

[27] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[28] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *European Conference on Computer Vision (ECCV)*, 2019.

[29] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 475–12 486.

[30] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. D. Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] A. Van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv e-prints*, pp. arXiv–1807, 2018.

[32] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8238–8247.

[33] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22 157–22 167, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260900398

[34] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate rgb-d salient object detection via collaborative learning," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 52–69.

[35] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 275–292.

[36] M. Lee, C. Park, S. Cho, and S. Lee, "Spsn: Superpixel prototype sampling network for rgb-d salient object detection," in *ECCV*. Springer, 2022, pp. 630–647.

[37] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for rgb-d salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2343–2350.

[38] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 92–109.

[39] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[40] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for rgb-d saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9026–9042, 2022.

[41] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.

[42] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2020.

[43] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.

[44] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 678–686.

[45] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018.

[46] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, vol. 25, pp. 4163–4176, 2023.

[47] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 17 612–17 625. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/702f4db7543a7432431df588d57bc7c9-Paper-Conference.pdf

[48] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4421–4433, 2020.

[49] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong, "Does thermal really always matter for rgb-t salient object detection?" *IEEE Transactions on Multimedia*, vol. 25, pp. 6971–6982, 2023.