Disentangled Cross-modal Transformer for RGB-D Salient Object Detection and Beyond

Hao Chen, Feihong Shen, Ding Ding, Yongjian Deng and Chao Li

***Abstract*—Previous multi-modal transformers for RGB-D salient object detection (SOD) generally directly connect all patches from two modalities to model cross-modal correlation and perform multi-modal combination without differentiation, which can lead to confusing and inefficient fusion. Instead, we disentangle the cross-modal complementarity from two views to reduce cross-modal fusion ambiguity: 1) Context disentanglement. We argue that modeling long-range dependencies across modalities as done before is uninformative due to the severe modality gap. Differently, we propose to disentangle the cross-modal complementary contexts to intra-modal self-attention to explore global complementary understanding, and spatial-aligned inter-modal attention to capture local cross-modal correlations, respectively. 2) Representation disentanglement. Unlike previous undifferentiated combination of cross-modal representations, we find that cross-modal cues complement each other by enhancing common discriminative regions and mutually supplement modal-specific highlights. On top of this, we divide the tokens into consistent and private ones in the channel dimension to disentangle the multi-modal integration path and explicitly boost two complementary ways. By progressively propagate this strategy across layers, the proposed Disentangled Feature Pyramid module (DFP) enables informative cross-modal cross-level integration and better fusion adaptivity. Comprehensive experiments on a large variety of public datasets verify the efficacy of our context and representation disentanglement and the consistent improvement over state-of-the-art models. Additionally, our cross-modal attention hierarchy can be plug-and-play for different backbone architectures (both transformer and CNN) and downstream tasks, and experiments on a CNN-based model and RGB-D semantic segmentation verify this generalization ability.**

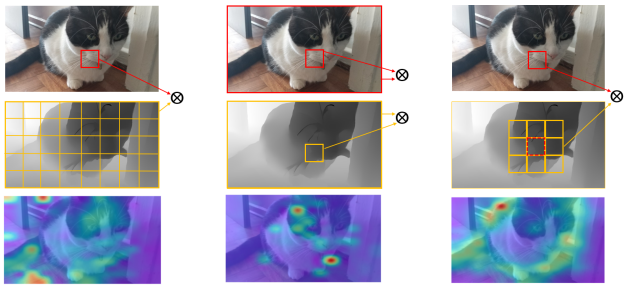
***Index Terms*—RGB-D salient object detection, cross-modal attention, disentanglement, transformer.**

I. Introduction

Salient object detection (SOD), which aims to simulate human visual systems and identify the most attractive objects in a scene, has benefited a large variety of computer vision tasks, including image compression [1], detection [2], segmentation [3], [4] and tracking [5].

RGB-D SOD has attracted increasing attention due to the additional spatial structures from depth that complement the RGB inference in challenging cases when the background and foreground hold similar appearance. Most existing RGB-D SOD methods [6], [7], [8], [9], [10], [11], [12], [13] follow the CNN-based paradigm and focus on designing various cross-modal cross-level interaction and fusion paths [14], [11], [7], [10] to explore the heterogeneous feature complementarity. They also present diverse strategies such as attention modules [15], [16], [17], dynamic convolution [18], cross-modal reconstruction [19] and knowledge distillation [20] to enhance adaptivity in selecting complementary cues. Although these methods have advanced the RGB-D SOD community, they have an intrinsic limitation in capturing global contexts, as convolutions have a natural locality. However, it is widely acknowledged [21], [22], [23], [24] that global contexts are crucial in correctly localizing the salient object. Even strategies that attempt to enhance the global understanding by appending fully connected [25] or global pooling layers [23] on restricted layers still struggle with large computational cost or limited capability in modeling global correlations.

Recently, Transformers [26], which excel in capturing long-range dependencies, have overcome CNNs' limitations and shown potential in modeling complex cross-modal complementarity and studying global contexts for SOD. Therefore, the Visual Saliency Transformer (VST) [27] model, using the transformer as the backbone, has been proposed for RGB(D) SOD. Specifically, VST adopts T2T ViT [28] as the encoder,



(a) Heterogeneous cross-modal attention (CA) (b) Global self-attention (GSA) (c) Local-aligned cross-modal attention (LCA)

Fig. 1. Examples to show the advantages of our proposed hierarchical cross-modal attention scheme, formed by (b) and (c). We visualize the attention maps generated by three types of cross-modal attention, where $\otimes$ denotes element-wise multiplication to calculate the similarity. With the modality gap and spatial discrepancy present, the heterogeneous cross-attention (a) fails to localize salient objects and instead highlights background noise. In contrast, by disentangling the heterogeneous cross-attention into global self-attention and local-aligned cross-modal attention, our HCA successfully explores global and local cross-modal complements to identify important regions and refine/complete local details.

Manuscript received 1 May 2023; revised 5 November 2023; accepted 9 January 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62102083, Natural Science Foundation of Jiangsu Province under Grant BK20210222, the National Natural Science Foundation of China (NSFC) under Grant 62261160576, 62203024 and 92167102, and the R&D Program of Beijing Municipal Education Commission (KM202310005027).

Hao Chen, Feihong Shen, Ding Ding are with the School of Computer Science and Engineering, Southeast University, Nanjing, China (e-mail: haochen303@seu.edu.cn; feihongshen@seu.edu.cn; dingding-1@seu.edu.cn). Yongjian Deng is with the College of Computer Science, Beijing University of Technology (yjdeng@bjut.edu.cn) and Chao Li is with Alibaba Group (lllcho.lc@alibaba-inc.com).

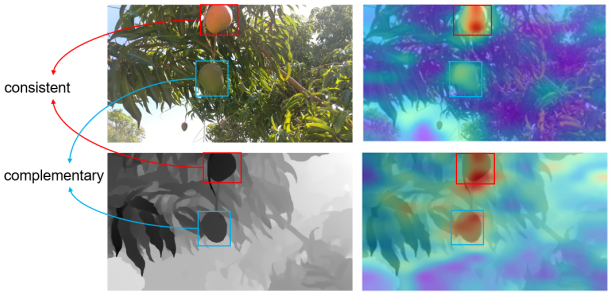


Fig. 2. Examples to show the variety of cross-modal complementarity, including (1) consistent complements represented by the yellow mango and (2) private complements illustrated by the green mango. Left: an RGB-D pair. Right: Corresponding output attention maps.

and the cross-modal fusion problem is solved by a cross-modal attention module formed by the similarity between a query from one modality and all keys from the paired modality. Compared to the CNN-based counterpart, VST catches long-range dependencies within/cross modalities, achieving state-of-the-art results on RGB-D SOD.

However, there are two key questions remaining in selecting cross-modal complements.

1. **How to model the dependencies among cross-modal cues is the key to an informative cross-modal combination.** VST answers this question by involving all RGB and depth patches to form a multi-modal token pool to model the cross-modal long/short-range dependencies jointly. As shown in Fig. 1 (a), in its heterogeneous cross-modal attention solution, each patch in RGB will be compared to all regions in the paired depth image and vice verse to generate cross-modal dependencies. However, it makes little sense to measure the dependencies between cross-modal tokens that lie distant in space (e.g., the cat in RGB and the floor in depth) due to the severe cross-modal representation gap. Hence, this strategy is not effective to model the multi-modal contexts. Additionally, such heterogeneous cross-modal fusion overlooks the inborn spatial-alignment in RGB-D pairs and computing the similarity between cross-modal spatial-distant patches will introduce noise to the query from keys in other areas.

Therefore, ***our core insight is that the modality gap and spatial discrepancy should not concurrent when measuring cross-modal dependencies for complementing.*** Based on this insight, we propose a principled **H**ierarchical **C**ross-modal **A**ttention (HCA) module, where the hybird cross-modal attention adopted in VST is disentangled into two scales: I) Intra-modal global contexts. As illustrated in Fig. 1 (b), the within-modal self-attention reveals global contrasts and contexts in each modality. Combining this global understanding from two modalities will better facilitate discriminating the foreground from background. II) Cross-modal local spatial-aligned enhancement. As a local region aligned in two modalities carries the same local context (Fig. 1 (c)), the cross-modal attention restricted in the small aligned region can capture the cross-modal correlation and mapping, thereby bridging the cross-modal gap, easing multi-modal feature fusion and boosting joint refinement of local details. With the complements from these two perspectives, our HCA can exploit the cross-modal global/local contexts for complementary global reasoning and local enhancement.

2. **The fusion of heterogeneous cross-modal cues through an undifferentiated combination limits the potential gains achieved.** Cross-modal representations complement each other in a blended manner, contributing to a common understanding (e.g., they both emphasize the yellow mango in Fig. 2) while carrying private but complementary cues to complete the inference context (e.g., depth additionally highlights the green mango). However, current cross-modal fusion methods typically combine cross-modal cues by direct combination without differentiation, leading to ambiguous fusion and limited gains. To tackle this issue, we propose disentangling hybrid complementarity and explicitly dividing the tokens from two modalities to learn complementary and consistent cues, respectively. By constraining consistency and diversity between two transformer streams and progressively propagate such contribution-aware tokens across levels, we form a Disentangled Feature Pyramid (DFP), which exposes the desired versatile complementary cues.

The contributions of our paper are as follows:

1) Disentangled cross-modal attention contexts and modality representations: We propose a multi-modal network that disentangles the cross-modal attention contexts and modality representations. The disentanglement allows for clear separation of long/short-range cross-modal dependencies and ensures each modality is aware of contributing desired consistent/private complementary cues, thereby avoiding confusion and facilitating informative multi-modal fusion.

2) Achieving state-of-the-art performance on RGB-D SOD: Comprehensive experiments are conducted on 8 public RGB-D SOD datasets, which quantitatively demonstrate that the proposed methods effectively improve the SOD quality and achieve state-of-the-art performance.

3) Universality of our insights and designs: The proposed insight of disentangling cross-modal attention context proves to be applicable to various backbone architectures and downstream tasks. Additional experiments conducted on CNN-based methods and semantic segmentation further demonstrate the seamless integration of the proposed HCA module with different backbone architectures and tasks. These experiments consistently show improvement, highlighting the versatility of the HCA module.

II. Related Work

A. RGB-D Salient Object Detection

As handcrafted saliency cues [29], [30] are weak in learning global contexts and hold limited generalization ability, recent RGB-D SOD models mainly focus on designing CNN architectures and cross-modal fusion paths to explore the complements. For example, DF [31] integrates handcrafted cues from two modalities as a joint input to train a shared CNN, while PCF [32] proposes a progressive multi-scale fusion strategy and TANet [16] introduces a three-stream architecture to select complementary cues in each level. Apart from the basic fusion architectures (i.e., single stream, two-stream, and

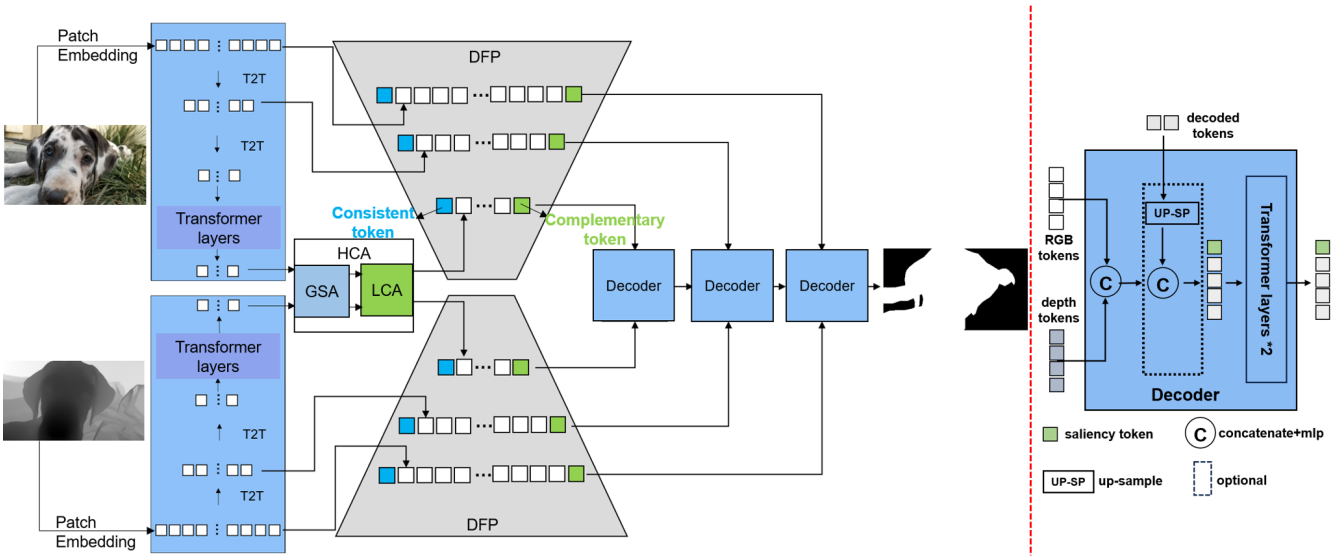


Fig. 3. Left: overall architecture of our model, HCA and DFP denotes the hierarchical cross-modal attention (as described in section III-C) and disentangled feature pyramid (as described in section III-D). Specifically, HCA is consisted of global self attention (GSA) and local-alignment cross-modal attention (LCA). The detailed architecture of GSA, LCA and DFP are illustrated in Fig. 4 (b), Fig. 4 (c) and Fig. 6 respectively. Right: the detailed architecture of the decoder.

three-stream), some other works explore feature combination strategies [11], [9], cross-modal cross-level interaction paths [33], and other strategies such as knowledge distillation [20] and dynamic convolution [34] to improve fusion sufficiency. Despite these advances, the locality nature of CNNs makes learning global contexts a challenge.

In contrast, VST [27] proposes a transformer-based architecture to extract intra/inter-modal long-range dependencies by using ViTs as encoders for each modality and stitching tokens from two modalities for cross-modal fusion. However, VST ignores the large cross-modal gap and do not differentiate the varying contributions of different tokens in inference. Similar problems are also witnessed in [35], [36], [37]. Furthermore, to improve SOD performance, many works such as [27], [35], [36], [38] have incorporated an edge detection task as auxiliary supervision. However, the process of generating edge labels can be exceedingly costly and time-consuming.

To address these limitations, we propose disentangling the cross-modal complements from both the context and representation views to respect the cross-modal gap and explicitly encourage desired consistent and private complements. .

B. Transformers in Computer Vision

ViT [39] first integrates transformers [26] into visual tasks by dividing images into patches, thereby enabling the introduction of various transformers, such as T2T ViT [28] and PVT [40], which allow for the modeling of local structures and the learning of multi-scale features flexibly. The Swin transformer [41] fully takes advantage of displacement and size invariance, leveraging them into the transformer architecture. These transformers have been widely adopted in other computer vision tasks to extract features, such as detection [42], segmentation [43] and SOD [27], [35], [44].

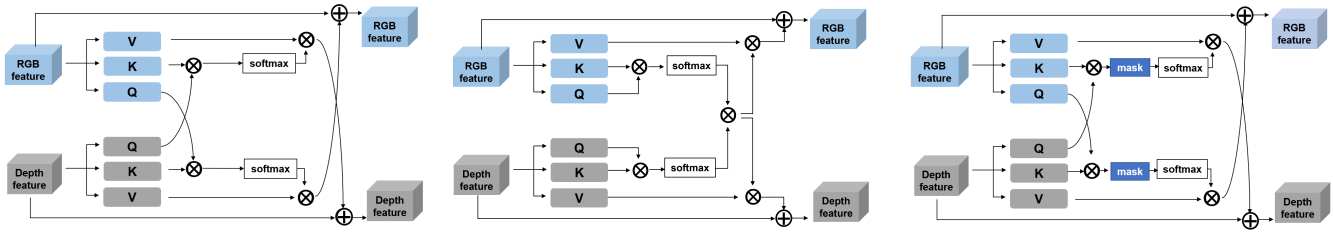
Transformers have demonstrated their efficacy in the multi-modal learning community due to their flexible input designs. For instance, the Multimodal Bottle-neck Transformer (MBT) [45] directly feeds latent units into a shared transformer backbone to combine cross-modal information, facilitating audio-visual fusion. Similarly, TriTransNet [46] concatenats RGB-depth features across various scales as an additional transformer stream for RGB-D SOD inference. Nevertheless, these methods typically concatenate multi-modal features directly prior to applying self-attention mechanisms[26], leading to uninformative fusion and potentially noisy features. SwinNet [35] combines the two modalities through direct element-wise multiplication, followed by general spatial attention and channel attention to enhance the features. In contrast, VST [27] and SiaTrans [44] extract unimodal features separately and employs a cross-modal attention module to address this issue. Subsequent transformer-based multi-modal models [47], [48], [36], [38] mainly follow this paradigm. However, none of these methods explicitly study the disentanglement of cross-modal attention context and representations, thus limiting their fusion adaptivity and sufficiency.

III. The Proposed Method

In this section, we begin by presenting the overall architecture of our model. Subsequently, we delve into our principal designs, namely the hierarchical cross-modal attention module (HCA) and disentangled feature pyramids (DFP).

A. Overall Architecture

The architecture of our proposed model is illustrated in Fig. 3. Specifically, two T2T ViT [28] backbones pre-trained on ImageNet are applied to extract RGB and depth features, respectively. Then, the HCA module is customized to capitalize on the cross-modal spatial alignment explicitly, thereby enhancing global/local contextual fusion. Later, we introduce the disentangled feature pyramid to bifurcate heterogeneous complementary cues into modal-shared and modal-specific



(a) Heterogeneous cross-modal attention (b) Global self-attention (c) Local-aligned cross-modal attention

Fig. 4. Detailed architecture of different cross-modal attention blocks: (a) Heterogeneous cross-modal attention and the components of the hierarchical cross-modal attention module, including (b) global self-attention (GSA) and (c) local-aligned cross-attention (LCA).

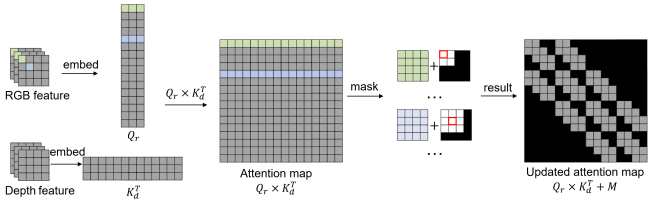


Fig. 5. The masking process for local-aligned cross-modal attention in HCA.

ones, thus bolstering fusion sufficiency. A combination of these cues is utilized to generate the final joint prediction. Lastly, we leverage RT2T modules [27] to up-sample the fused features and predict the salient maps more accurately.

B. Previous Heterogeneous Cross-modal Attention

VST [27] extends intra-modal self-attention to cross-modal by directly connecting all tokens from two modalities. As shown in Fig. 4 (a), this process can be formulated as follows:

$$CA(Q_r, K_r, V_d) = softmax(\frac{Q_r K_d^T}{\sqrt{d}})V_d, \tag{1}$$

$$CA(Q_d, K_d, V_r) = softmax(\frac{Q_d K_r^T}{\sqrt{d}})V_r. \tag{2}$$

Here Q_r, K_r, V_r denotes query, key and value generated from RGB features respectively, while Q_d, K_d, V_d come from depth features.

The result of this cross-modal attention is visualized in Fig. 1 (a). However, we observe that the attention map tends to make the model messy and fails to highlight the most discriminative parts. This is because this method simultaneously involves modality gaps and spatial discrepancies in measuring cross-modal dependencies.

C. Our Hierarchical Cross-Modal Attention

GSA. To account for cross-modal gaps and spatial discrepancies, we model the cross-modal attention in a disentangled manner with varying scales, including global self-attention and local-aligned cross attention. As is shown in Fig. 4, our hierarchical cross-modal attention (HCA) module consists of two stages: Global self-attention (GSA) in Fig. 4 (b) and Local-aligned cross-modal attention (LCA) in Fig. 4 (c). In the GSA stage, RGB and depth features are processed individually to obtain their intra-modal self-attention maps, which are then dot-producted to obtain the multi-modal global attention Att_{mm}^G, as follows:

$$Att_{mm}^G = softmax(\frac{Q_r K_r^T}{\sqrt{d}}) * softmax(\frac{Q_d K_d^T}{\sqrt{d}}). \tag{3}$$

The shared global attention map Att_{mm}^G is then applied to each modality to extract complementary global contexts:

$$GSA(Q_r, K_r, V_d) = Att_{mm}^G * V_d, \tag{4}$$

$$GSA(Q_d, K_d, V_r) = Att_{mm}^G * V_r. \tag{5}$$

The resulting $GSA(Q_d, K_d, V_r)$ and $GSA(Q_r, K_r, V_d)$ are added back to the original RGB and depth features respectively as residual components.

LCA. After sharing structural information, we utilize a local-aligned cross-modal attention (LCA) module to strengthen the fusion of cross-modal local semantics. The primary difference between LCA and traditional cross-modal attention lies in the mask operation, which retains attention similarity $Q \times K^T$ in adjacent areas and set similarity to 0 in remote areas. The mathematical formula is as follows:

$$M(i,j) = \begin{cases} 0, & \textit{if isnear(i,j)} \\ -100, & \textit{else} \end{cases}, \tag{6}$$

where i and j are the pixel position at the i^{th} row and j^{th} column in the RGB feature and depth feature, *isnear(i,j)* is a function to assess whether i and j are adjacent in position. Specifically, the global cross-attention map $Q \times K^T$ is added with a mask matrix M, where 0 represents adjacent areas and -100 represents remote areas. After the softmax operation, the attention scores from remote areas become close to 0, and remote areas make no further contribution to the fusion process. This process can be expressed as follows:

$$LCA(Q_r, K_d, V_d) = softmax(\frac{Q_r K_d^T + M}{\sqrt{d}})V_d, \tag{7}$$

$$LCA(Q_d, K_r, V_r) = softmax(\frac{Q_d K_r^T + M}{\sqrt{d}})V_r. \tag{8}$$

Fig. 5 illustrates the process of mask generation. After embedding the RGB and depth features, the first (green) line in Q_r represents the features of the patch at position [0,0]

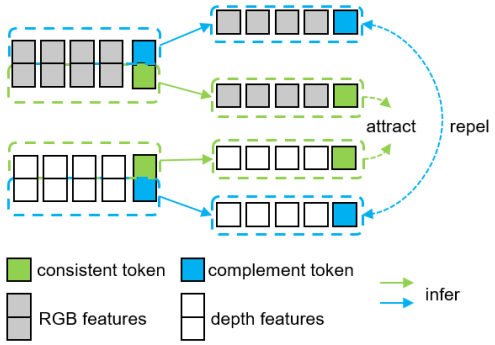


Fig. 6. Details of the disentangled feature pyramid module under a layer.

(here [x,y] denotes the x^{th} row and y^{th} column) in the RGB feature. We then perform matrix multiplication between Q_r and K_d^T to produce the unified attention map. For example, the first (green) line in the attention map captures the similarity between the first patch in RGB features and all the patches in the paired depth feature. Subsequently, we design a mask for each line based on its distance to the target patch, and derive the final attention map. More specifically, we convert each line in the attention map into the matrix form, the position [0,0] in RGB feature is adjacent to [0,0], [0,1], [1,0] and [1,1] in depth, therefore we retain these positions while mask the other patches. Once all the masks are computed, we reshape them from matrices back to vectors and concatenate them together to obtain the final mask M. The masked attention map, which contains complementary local contexts and cross-modal correlation, is then multiplied by V_d to generate the final depth features. Additionally, a symmetrical interaction line is also performed to enhance RGB features.

D. Disentangled Feature Pyramid

In this section, we address the issue of heterogeneity in cross-modal feature complementarity by proposing the Disentangled Feature Pyramid (DFP) module, inspired by previous works [8], [19]. The goal of the DFP module is to explicitly disentangle the cross-modal complements into consistent and complementary parts to improve the adaptivity of feature fusion.

The DFP module, depicted in Fig. 6, comprises several steps. Firstly, the input RGB features f_{rgb} and depth features f_{depth} are segregated into consistent and complementary parts in the channel dimension, denoted as f_{rgb}^{cons}, f_{rgb}^{com}, f_{depth}^{cons}, and f_{depth}^{com}. Subsequently, a learnable token is appended to the disentangled RGB and depth features, respectively, acting as a task-guided token as used in [28], [27] to distill consistent/complementary representations in each modality and propagate them to shallower layers, as shown in the dashed boxes. Next, a patch-task-attention [27] is applied to infer the disentangled features, f'^{cons}_{rgb}, f'^{com}_{rgb}, f'^{cons}_{depth} and f'^{com}_{depth}, thereby explicitly disentangling the RGB and depth features.

To push apart the RGB and depth features, we propose a complementary loss that aims to minimize the cosine similarity between f'^{com}_{rgb} and f'^{com}_{depth}, as follows:

$$L_i^{com} = sim_{cos}(f'^{com}_{rgb}, f'^{com}_{depth}), \tag{9}$$

where L_i^{com} denotes the complementary loss in i^{th} level. The function $sim_{cos}(.)$ computes the cosine similarity. To minimize the gap between RGB and depth features, we force them to predict the same ground-truth by minimizing:

$$L_i^{cons} = bce(pred(f'^{cons}_{rgb}), y) + bce((pred(f'^{cons}_{depth}), y)), \tag{10}$$

where L_i^{cons} is the consistent loss in the i^{th} level, $bce(.)$ denotes the binary cross-entropy loss, $pred(.)$ represents the prediction of the saliency map and y is the ground-truth.

After disentangling the RGB and depth features, we concatenate them and then feed them into the decoder module. Moreover, the consistent and complementary tokens are reused in the next level. In this way, our DFP enables the model to pass consistent features and complementary information across modalities and layers.

E. Model Optimization

Our decoder is similar to that in VST; however, we remove its boundary token. In each decoder layer, we combine the RGB-D features across adjacent layers and predict the salient maps, optimizing the saliency loss as follows:

$$L_i^{sal} = bce(P_i, y), \tag{11}$$

where L_i^{sal} is the salient loss in the i^{th} stage, P_i denotes the predicted salient map by the i^{th} decoder layer. As discussed in section III-D, we employ a consistent loss L_i^{cons} to pull half channels of the RGB and depth features together, and a complementary loss L_i^{com} to push the other halves apart. Consequently, the final objective function is minimized as follows:

$$L_{total} = min \sum_{i=1}^{3} (\lambda_i * L_i^{sal} + \gamma_i * L_i^{com} + \eta_i * L_i^{cons}), \tag{12}$$

where λ_i, γ_i and η_i are weights that control the importance of each loss. We use these weights to make different losses approximate without further tuning. By balancing the weights of losses, our model successfully learns salient cues and bifurcates the consistent and complementary features to enhance modality complementarity.

IV. Experiments

A. Datasets

We evaluate the proposed model on eight public RGB-D SOD datasets NJUD [54] (1985 image pairs), NLPR [30] (1000 image pairs), DUTLF [55](1200 image pairs), STERE [56] (1000 image pairs), RGBD135 [57] (135 image pairs), SIP [58] (929 image pairs), RedWeb-S [59] (3,179 image pairs) and COME15K [60] (15625 image pairs). We follow the consistent setting in previous works [55], [61], [27], [52] and choose 1485 image pairs in NJUD, 700 in NLPR, 800 in DUTLF and 8025 in COME15K as the training set, while the remaining images are used for testing. We also apply data

TABLE I
COMPARISON WITH STATE-OF-THE-ART RGB-D SOD METHODS. ↑ AND ↓ INDICATE THAT LARGER AND SMALLER VALUES, RESPECTIVELY, ARE BETTER. THE BEST RESULTS ARE LABELED IN **BOLD**. "-" INDICATES THAT THE CODE OR RESULT IS NOT AVAILABLE.

Dataset	Metric	CNN-based							Transformer-based			Ours
		CMW [49]	HDFNet [34]	CoNet [50]	BBS-Net [51]	SSP [8]	SPSN [52]	MVSalNet [53]	VST [27]	SwinNet [35]	SiaTrans [44]	
	Backbone	Vgg16	Resnet50	Resnet50	Resnet50	Vgg16	Vgg16	Resnet50	T2T-14	T2T-14	T2T-14	T2T-14
	$MACs(G)$	208.03	91.77	20.89	31.2	-	-	-	30.99	124.3	10.91	24.52
	$Params(M)$	85.65	44.15	43.66	49.77	74.14	-	83.00	83.83	198.7	22.24	80.04
NJUD	$S_m \uparrow$	0.870	0.908	0.896	0.921	0.909	0.918	0.910	0.922	0.915	0.923	**0.932**
	$maxF \uparrow$	0.871	0.911	0.893	0.919	0.923	0.920	0.922	0.920	-	0.921	**0.934**
	$E^{max} \uparrow$	0.927	0.944	0.937	0.949	0.951	0.950	0.939	0.939	0.919	0.956	**0.959**
	$MAE \downarrow$	0.061	0.039	0.046	0.035	0.039	0.032	0.035	0.035	0.034	0.035	**0.031**
NLPR	$S_m \uparrow$	0.917	0.923	0.912	0.931	0.922	0.923	0.927	0.932	0.929	0.928	**0.934**
	$maxF \uparrow$	0.903	0.917	0.893	0.918	0.889	0.910	**0.929**	0.920	-	0.918	0.923
	$E^{max} \uparrow$	0.951	0.963	0.948	0.961	0.960	0.958	0.959	0.962	0.958	0.964	**0.965**
	$MAE \downarrow$	0.027	0.027	0.027	0.023	0.025	0.024	**0.021**	0.023	0.022	0.024	0.023
DUTLF	$S_m \uparrow$	0.797	0.908	0.923	0.882	0.929	0.871	0.929	0.943	-	0.940	**0.948**
	$maxF \uparrow$	0.779	0.915	0.932	0.870	0.947	0.858	0.935	0.948	-	0.944	**0.952**
	$E^{max} \uparrow$	0.864	0.945	0.959	0.912	0.958	0.907	0.958	**0.969**	-	0.968	**0.969**
	$MAE \downarrow$	0.098	0.041	0.029	0.058	0.029	0.053	0.029	0.024	-	0.025	**0.023**
STERE	$S_m \uparrow$	0.852	0.900	0.905	0.908	0.904	0.913	0.914	0.907	0.894	0.911	**0.922**
	$maxF \uparrow$	0.837	0.900	0.901	0.903	0.914	0.900	**0.920**	0.907	-	0.907	0.919
	$E^{max} \uparrow$	0.907	0.943	0.947	0.942	0.939	0.943	0.946	0.951	0.918	0.951	**0.955**
	$MAE \downarrow$	0.067	0.042	0.037	0.041	0.039	**0.035**	**0.035**	0.038	0.044	0.038	**0.035**
RGBD135	$S_m \uparrow$	0.934	0.926	0.914	0.934	0.936	0.908	0.931	0.943	-	0.936	**0.948**
	$maxF \uparrow$	0.931	0.921	0.902	0.928	**0.944**	0.900	0.934	0.940	-	0.932	**0.944**
	$E^{max} \uparrow$	0.99	0.970	0.948	0.966	0.978	0.947	0.971	0.978	-	0.975	**0.978**
	$MAE \downarrow$	0.02	0.022	0.024	0.021	**0.017**	0.024	0.019	**0.017**	-	0.020	**0.017**
SIP	$S_m \uparrow$	0.705	0.886	0.860	0.879	0.888	0.892	0.875	0.904	0.897	0.899	**0.924**
	$maxF \uparrow$	0.677	0.894	0.873	0.884	0.909	0.899	0.886	0.915	-	0.913	**0.940**
	$E^{max} \uparrow$	0.804	0.930	0.917	0.922	0.927	0.934	0.924	0.944	0.931	0.945	**0.962**
	$MAE \downarrow$	0.141	0.048	0.048	0.055	0.046	0.042	0.052	0.040	0.045	0.041	**0.031**
ReDWeb-S	$S_m \uparrow$	0.634	0.728	0.696	0.693	-	0.698	0.733	0.759	-	0.753	**0.769**
	$maxF \uparrow$	0.607	0.717	0.693	0.680	-	0.689	0.724	**0.763**	-	0.759	0.761
	$E^{max} \uparrow$	0.714	0.804	0.782	0.763	-	0.756	0.803	0.826	-	0.830	**0.835**
	$MAE \downarrow$	0.195	0.129	0.147	0.150	-	0.113	0.128	0.113	-	0.113	**0.106**
COME-E	$S_m \uparrow$	-	-	0.838	0.847	0.852	0.845	0.854	0.871	-	-	**0.910**
	$maxF \uparrow$	-	-	0.831	0.837	0.843	0.831	0.854	0.862	-	-	**0.909**
	$E^{max} \uparrow$	-	-	0.883	0.889	0.891	0.889	0.899	0.913	-	-	**0.944**
	$MAE \downarrow$	-	-	0.071	0.070	0.066	0.062	0.062	0.054	-	-	**0.036**
COME-H	$S_m \uparrow$	-	-	0.795	0.793	0.799	0.797	0.810	0.830	-	-	**0.872**
	$maxF \uparrow$	-	-	0.795	0.789	0.796	0.791	0.807	0.827	-	-	**0.875**
	$E^{max} \uparrow$	-	-	0.840	0.839	0.846	0.844	0.853	0.872	-	-	**0.909**
	$MAE \downarrow$	-	-	0.102	0.105	0.099	0.094	0.093	0.083	-	-	**0.060**

augmentation techniques such as resizing, random cropping and random flipping to avoid overfitting, as done in previous works.

B. Evaluation Metrics

We adopt four widely used evaluation metrics to evaluate our model. Specifically, Structure-measure S_m [62] evaluates region-aware and object-aware structural similarity. E-measure E_ρ^{max} [63] simultaneously considers pixel-level errors and image-level errors. Maximum F-measure [64] jointly considers precision and recall under the optimal threshold. Mean Absolute Error (MAE) computes pixel-wise average absolute error.

C. Implementation Details

We follow VST [27] to use the T2T-ViTt-14 [28] as our backbone. Our model is trained using Pytorch on a RTX 3090 with a batch size of 8, and the training process runs for 50 epochs. We use the Adam optimizer with a learning rate that gradually decays from 10^{-4} to 10^{-6}.

D. Comparisons with State-of-the-art

To quantitatively evaluate the performance of our proposed model, we compare it with 10 SOTA RGB-D SOD methods, including seven CNN-based methods (CMW [49], HDFNet [34], CoNet [50], BBS-Net [51], SSP [8], SPSN [52], and MVSalNet [53]) and three transformer-based methods (VST [27], SiaTrans [44] and SwinNet [35]). Table. I shows the comparison in terms of the S-measure, F-measure, E-measure and MAE scores. The quantitative results demonstrate that VST achieves improvement over CNN-based methods on most of datasets and metrics, denoting the superiority of the transformer. Our model outperforms previous RGB-D SOD models, including VST, on all datasets, highlighting the advantages of our disentangled cross-modal fusion designs.

We also provide visualized results on some representative challenging scenes in Fig. 7. In scenarios with large intra-difference in the foreground (i.e., 1^{st}, 2^{nd} and 4^{th} rows), previous models often fail to detect the correct salient objects completely, while our model can achieve more accurate and uniform detection. Similarly, when the foreground and

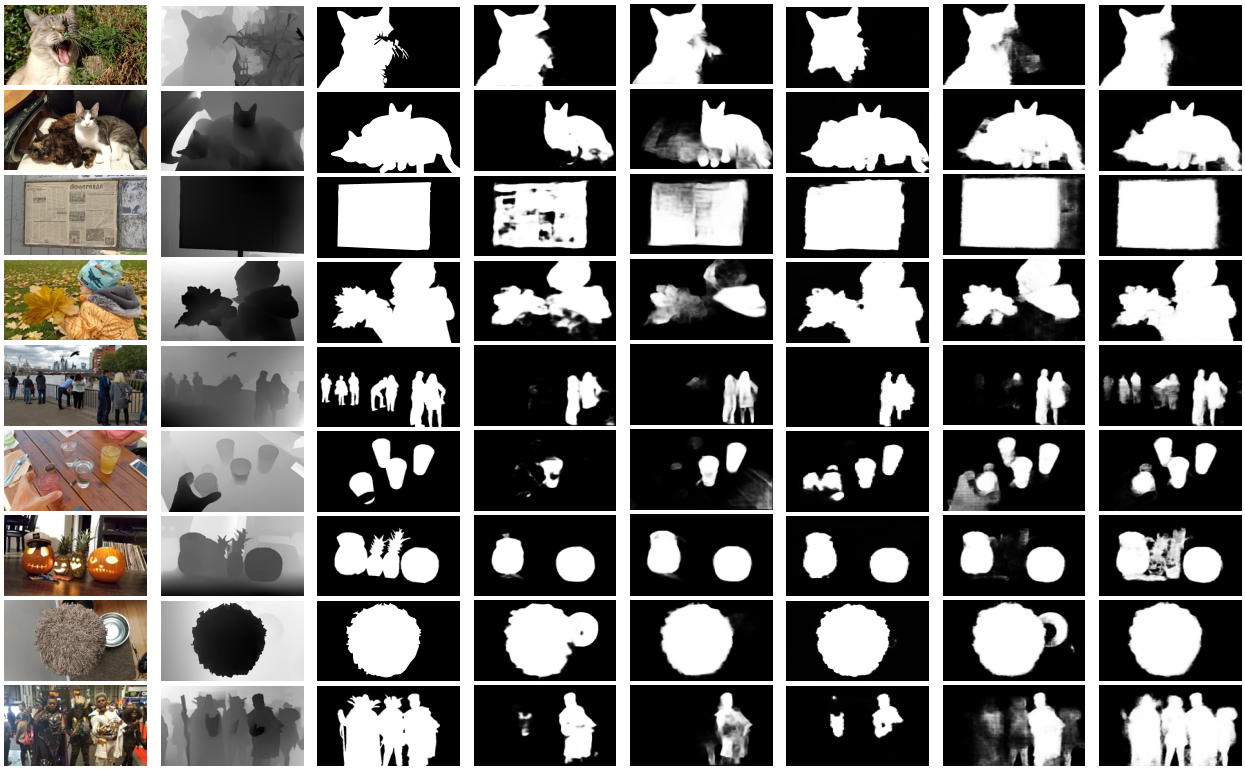


Fig. 7. Visual comparison to SOTA RGB-D SOD methods

background hold a similar appearance or depth (i.e., 1^{st}, 3^{rd}, 6^{th} and 9^{th} rows), previous methods may mistake some background areas as foreground, while our model successfully handles this confusion by leveraging the discriminating modality. For scenes having multiple salient objects (i.e., 2^{nd}, 5^{th}, 6^{th}, 7^{th} and 9^{th} rows), other models tend to overlook some regions, while our method can highlight all salient objects. The success of our model on these difficult cases demonstrates our advantages in modeling global/local cross-modal correlations and exploring consistent/private complementary features.

E. Ablation Study

In this section, we comprehensively conduct ablation experiments on the largest COME-EASY and COME-HARD datasets to verify the effectiveness of each design in our proposed model. To demonstrate the advantages of our cross-modal fusion scheme fairly, we remove the auxiliary edge detection task in VST [27] and retrain it with our training set as the baseline model (denoted by "Base"). We argue that edge detection requires additional edge labels, which are unavailable in SOD datasets, and is not closely related to the multi-modal fusion process.

Effectiveness of hierarchical cross-modal attention. As shown in Table II, replacing the heterogeneous cross-modal attention layer in the VST cross-modal fusion baseline ("Base") with our global self-attention layer ("+GSA") improves the performance, indicating the irrationality of using spatial-distant cross-modal dependencies as complementary contexts. The improvement of GSA can potentially be attributed to its effectiveness on exploring global complementary contexts and mitigating the cross-modal gap. Adding the local-aligned cross-modal attention ("+GSA+LCA") further captures local complements and contributes additional gains. This stage-wise improvement well supports our motivation that cross-modal gap and spatial discrepancy should not be concurrent when measuring cross-modal correlations and verifies the efficacy of our hierarchical global-local attention combination.

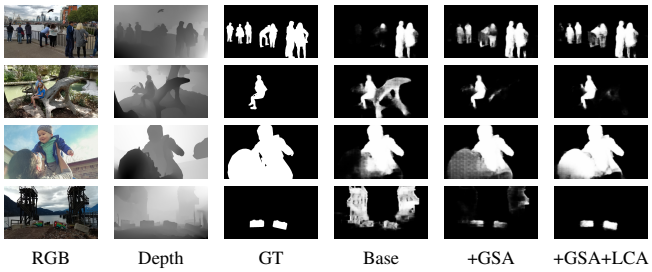


Fig. 8. Visualization to demonstrate the effectiveness of each component in HCA. "+GSA" denotes adding the global self-attention shown in Fig. 4 (b) additionally. "+GSA+LCA" represents adding the entire hierarchical cross-modal attention block, including the modules shown in Fig. 4 (b) and (c).

Fig. 8 visually demonstrates the effectiveness of the GSA and LCA components in HCA. The heterogeneous cross-attention in VST ("Base") introduces noises from other distant areas, leading to incorrect identification of the background as foreground. With the GSA module complementing global contexts, the model has stronger global reasoning ability to localize the salient object. The LCA module takes advantage

TABLE II
ABLATION STUDY ON OUR DESIGNS. GSA REFERS TO THE GLOBAL SELF-ATTENTION MODULE, LCA DENOTES THE LOCAL-ALIGNMENT CROSS-MODAL ATTENTION MODULE, AND DFP STANDS FOR THE DISENTANGLED FEATURE PYRAMID MODULE.

Settings	COME-E [60]				COME-H [60]			
	$S_m \uparrow$	$maxF \uparrow$	$E^{max} \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$maxF \uparrow$	$E^{max} \uparrow$	$MAE \downarrow$
Base	0.902	0.899	0.940	0.042	0.865	0.865	0.903	0.067
+GSA	0.906	0.904	0.942	0.038	0.868	0.869	0.907	0.062
+GSA+LCA	0.908	0.906	0.943	0.037	0.871	0.871	0.908	0.061
+GSA+LCA+DFP	**0.910**	**0.909**	**0.944**	**0.036**	**0.872**	**0.875**	**0.909**	**0.060**

of local cross-modal correlations to remove noise and enhance object details.

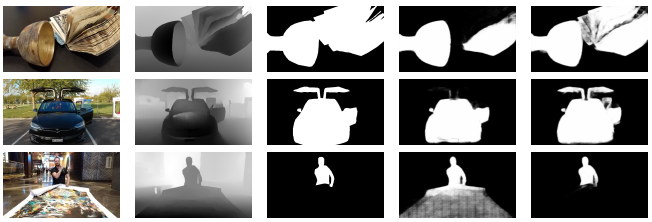


Fig. 9. Visualization to show the effectiveness of DFP. "+HCA" refers to the model merely with HCA. "+HCA+DFP" means additionally adding the disentangled feature pyramid module.

Effectiveness of the disentangled feature pyramid. The comparison between "+GSA+LCA" and "+GSA+LCA+DFP" in Table II shows consistent improvement by DFP, indicating that diversifying complex cross-modal complements into consistent and complementary ones boosts the fusion adaptivity. The visualization results in Fig. 9 show that undifferentiated concatenation is too ambiguous to fully explore cross-modal complements, resulting in incomplete maps (e.g., 1^{st} and 2^{nd} row) or wrongly highlighted background (e.g., the 3^{rd} row). Instead, DFP explicitly decouples the difference and consistency between modalities and combines them adaptively for varying scenes to localize salient objects accurately and highlight salient regions uniformly.

F. Visualization of the Hierarchical Cross-modal Attention

In this section, we visualize the attention maps generated by the heterogeneous cross-modal attention and our hierarchical cross-modal attention modules depicted in Fig. 4 according to attention rollout [65]. We calculate the mean attention map of all the heads and patches and present the final maps in Fig. 10, where brighter areas indicate higher attention scores.

From the column "CA", we observe that the heterogeneous cross-modal attention scheme mistakes some background areas as high confident discriminative ones (e.g., 1^{st}, 2^{nd} and 4^{th} rows). we attribute this to the involvement of spatially-distant cross-modal similarities, which makes the attention value of each token unable to truly reflect its discrimination. In contrast, by disentangling the heterogeneous cross-modal attention to different contextual scales, our global self-attention module ("GSA" in Fig. 10) successfully integrates the global understanding from RGB and depth to jointly identify discriminative regions and suppress noises. Afterwards, our "LCA" module leverages the spatial alignment information between RGB and depth to explore local correlations to complete the salient objects well and refine their boundaries. Combining the two modules together, the "HCA" simultaneously possesses the capability of precise localization and boundary refinement, effectively tackling the issue of wrongly highlighting backgrounds resulted from "CA", as illustrated in Fig. 10.

G. Plug-and-play: Generalize HCA to Other Tasks and Backbones

Our proposed HCA can serve as a plug-and-play solution to enhance cross-modal fusion adaptivity and sufficiency. To verify its generalization, we insert our HCA into state-of-the-art RGB-D semantic segmentation models with both transformer and CNN backbones. The experiments are performed on the NYUD-v2 dataset [66] and SunRGBD dataset [67]. NYUD-v2 contains 1,449 indoor RGB-D images (795 pairs for training and 654 pairs for testing) with a common 40-class label setting. SunRGBD has 37 classes and consists of 10,335 indoor RGB-D images, with 5,285 pairs for training and 5,050 pairs left for testing, as used in previous works [68], [69].

TABLE III
AN ABLATION STUDY TO DEMONSTRATE THE EFFICACY OF HCA ON TRANSFORMER-BASED SEMANTIC SEGMENTATION.

Settings	NYUD-v2 [66] $mIoU \uparrow$	SunRGBD [67] $mIoU \uparrow$
InvPT [68]	52.8	41.2
InvPT + CA (base)	53.2	41.9
InvPT + GSA	53.7	42.5
InvPT + GSA + LCA	54.2	42.8

Generalized to RGB-D Semantic Segmentation. InvPT [68] is a transformer-based model for RGB scene understanding. We modify the model to receive RGB and depth input and adopt ViT-Base as the backbone. Simple concatenation is chosen for the basic model to fuse RGB-D features. To illustrate the advantage of our proposed HCA, we add the heterogeneous cross-modal attention module ("CA") used in VST [27] to the InvPT model as the benchmark method for comparison (denoted by "InvPT+CA (base)"). GSA and LCA are added to the original InvPT model in turn. Specifically, we first reduce the embedding dimension of extracted features from 768 to 384 to reduce parameters, and then apply GSA

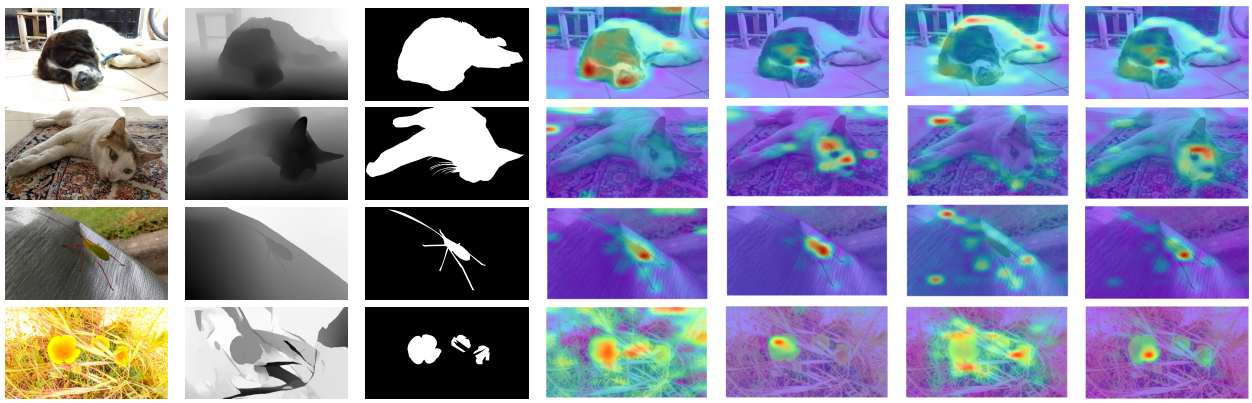


Fig. 10. Comparison of the attention maps generated by heterogeneous cross-modal attention (denoted as "CA") and our "HCA" (comprised of "GSA" and "LCA") to exhibit the advantages of our proposed HCA.

and LCA modules to paired RGB-depth features in three scales to achieve sufficient feature fusion. Finally, the fused features are recovered to 768 dimensions and fed into the decoder. Table III clearly shows stage-wise improvement from adding the GSA and LCA modules, demonstrating the effectiveness of HCA on transformer-based RGB-D semantic segmentation tasks.

TABLE IV
ABLATION STUDY ON SEMANTIC SEGMENTATION DATASET "NYUD-V2" AND "SUNRGBD". "ESANET" IS A CNN-BASED MODEL FOR RGB-D SEMANTIC SEGMENTATION.

Settings	NYUD-v2 [66] $mIoU \uparrow$	SunRGBD [67] $mIoU \uparrow$
ESANet [69]	48.3	47.1
ESANe + CA (base)	48.8	47.6
ESANet + GSA	49.1	47.9
ESANet + GSA + LCA	49.5	48.4

Generalized to CNN-based models. Our hierarchical cross-modal Attention also benefits CNN-based models by endowing them with both long and short cross-modal dependencies. To verify this, we insert our HCA into the ESANet [69]. The fusion module is applied to the deepest RGB and depth features. We reshape the convolutional features extracted by ResNet50 from B*C*H*W to B*(H*W)*C to cater to the input of attention. To reduce computational cost, the deepest RGB and depth features are reduced from 2048 dimensions to 64 dimensions. After HCA, the features are restored to 2048 dimensions, recovered to convolutional features, and finally fed into the decoder. To show the advantages of our designs, we also apply the heterogeneous cross-modal attention (denoted by "CA") to ESANet for comparison. The quantitative results in Table IV demonstrate the progressive gains from GSA and LCA.

These experiments on CNN-based and transformer-based RGB-D semantic segmentation methods demonstrate the promising future of our proposed HCA for various multi-modal data and tasks.

Fig. 11 visually illustrates the effectiveness of our proposed HCA on RGB-D semantic segmentation. With HCA, the model can better capture and combine global/local semantics from both modalities to distinguish objects in different categories accurately and completely (e.g., 1^{st}, 2^{nd} and 4^{th} rows) and promote segmentation details (e.g., the 3^{rd} row).

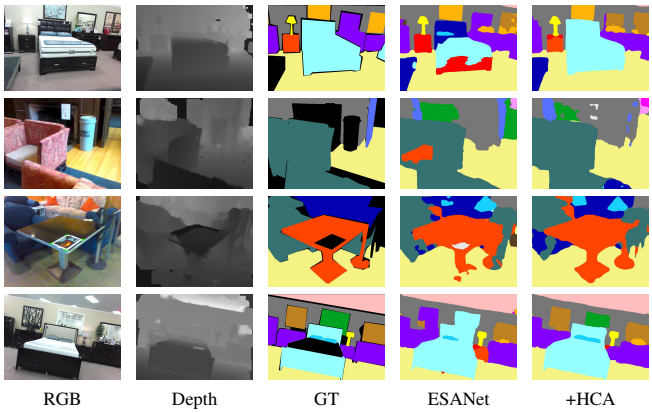


Fig. 11. Visual results on RGB-D semantic segmentation to demonstrate effectiveness of HCA qualitatively.

These generalization experiments on semantic segmentation and a CNN-based backbone confirm the importance of disentangling the context scale when modeling cross-modal dependencies and the high universality of our HCA block.

Discussion of the varying improvement on SOD and semantic segmentation. Based on the observation in Table II, III and IV, there are cases where our HCA demonstrates more substantial enhancements on semantic segmentation, while the improvements on SOD datasets are relatively minor. We speculate that this may be attributed to two factors: (a) Most SOD datasets already exhibit high metrics (greater than 0.9), making further improvement challenging. (b) SOD, compared to semantic segmentation, is a relatively simpler task. While both tasks require global semantics to infer the most salient object, SOD does not demand the same level of semantic concept understanding for each region as semantic segmentation. Our global self-attention (GSA) and local-aligned cross-modal

attention (LCA) can fuse multi-modal semantics at different scales, hence resulting in more noticeable improvements on semantic segmentation.

V. Conclusion

In this paper, we propose a new multi-modal transformer for spatial-aligned multi-modal pairs. Considering the modality gap, we disassemble the heterogeneous cross-modal attention and tailor a hierarchical cross-modal attention to explore the cross-modal complements successively in terms of global contexts and local correlations. We also introduce a disentangled complementing module to disentangle complex complements into consistent and complementary ones to boost cross-modal fusion adaptivity and reduce fusion ambiguity. Extensive experiments verify the advantages of our multi-modal transformer, the efficacy of our designs, and the high universality of our HCA on various tasks and backbones.

References

[1] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *TIP*, vol. 19, no. 1, pp. 185–198, 2009.

[2] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *IJCV*, vol. 128, pp. 261–318, 2020.

[3] H. Li, D. Zhang, N. Liu, L. Cheng, Y. Dai, C. Zhang, X. Wang, and J. Han, "Boosting low-data instance segmentation by unsupervised pre-training with saliency prompt," in *CVPR*, 2023, pp. 15 485–15 494.

[4] X. Zhou, T. Tong, Z. Zhong, H. Fan, and Z. Li, "Saliency-cce: exploiting colour contextual extractor and saliency-based biomedical image segmentation," *CIBM*, p. 106551, 2023.

[5] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *TPAMI*, vol. 35, no. 3, pp. 541–554, 2012.

[6] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features," in *ICCV*, 2017, pp. 2749–2757.

[7] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *CVPR*, 2021, pp. 1407–1417.

[8] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-Supervised Pretraining for RGB-D Salient Object Detection," in *AAAI*, vol. 3, 2022.

[9] G. Feng, J. Meng, L. Zhang, and H. Lu, "Encoder deep interleaved network with multi-scale aggregation for rgb-d salient object detection," *PR*, vol. 128, p. 108666, 2022.

[10] J. Li, W. Ji, M. Zhang, Y. Piao, H. Lu, and L. Cheng, "Delving into calibrated depth for accurate rgb-d salient object detection," *IJCV*, vol. 131, no. 4, pp. 855–876, 2023.

[11] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "Rgb-d salient object detection with cross-modality modulation and selection," in *ECCV*. Springer, 2020, pp. 225–241.

[12] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection," *TIP*, vol. 30, pp. 7012–7024, 2020.

[13] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection," *TIP*, vol. 31, pp. 6800–6815, 2022.

[14] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *CVPR*, 2019, pp. 3927–3936.

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

[16] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *TIP*, vol. 28, no. 6, pp. 2825–2835, 2019.

[17] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, and Y. Yu, "Cross-modality deep feature learning for brain tumor segmentation," *PR*, vol. 110, p. 107562, 2021.

[18] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *CVPR*, 2020, pp. 11 030–11 039.

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

[20] H. Chen, Y. Li, Y. Deng, and G. Lin, "Cnn-based rgb-d salient object detection: Learn, select, and fuse," *IJCV*, 2021.

[21] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *TPAMI*, vol. 34, no. 10, pp. 1915–1926, 2011.

[22] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[23] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *CVPR*, 2018, pp. 3089–3098.

[24] C. Fang, Q. Wang, L. Cheng, Z. Gao, C. Pan, Z. Cao, Z. Zheng, and D. Zhang, "Reliable mutual distillation for medical image segmentation under imperfect annotations," *TMI*, 2023.

[25] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," *CVPR*, 2016.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[27] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *ICCV*, 2021, pp. 4722–4732.

[28] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," in *ICCV*, 2021, pp. 538–547.

[29] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *BMVC*. Georgia Institute of Technology, 2013.

[30] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *ECCV*. Springer, 2014, pp. 92–109.

[31] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *TIP*, vol. 26, no. 5, pp. 2274–2285, 2017.

[32] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *CVPR*, 2018, pp. 3051–3060.

[33] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," *CVPR*, 2020.

[34] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *ECCV*. Springer, 2020, pp. 235–252.

[35] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *TCSVT*, vol. 32, no. 7, pp. 4486–4497, 2021.

[36] C. Liu, G. Yang, S. Wang, H. Wang, Y. Zhang, and Y. Wang, "Tanet: Transformer-based asymmetric network for rgb-d salient object detection," *arXiv preprint arXiv:2207.01172*, 2022.

[37] N. Zhang, J. Han, and N. Liu, "Learning implicit class knowledge for rgb-d co-salient object detection with transformers," *TIP*, vol. 31, pp. 4556–4570, 2022.

[38] C. Zeng and S. Kwong, "Dual swin-transformer based mutual interactive network for rgb-d salient object detection," *arXiv preprint arXiv:2206.03105*, 2022.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[40] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[43] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019, pp. 9404–9413.

[44] X. Jia, C. DongYe, and Y. Peng, "Siatrans: Siamese transformer network for rgb-d salient object detection with depth image classification," *IVC*, vol. 127, p. 104549, 2022.

[45] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *NeurIPS*, vol. 34, pp. 14 200–14 213, 2021.

[46] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network," in *ACMMM*, 2021, pp. 4481–4490.
[47] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *CVPR*, 2022, pp. 20 020–20 029.
[48] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *CVPR*, 2022, pp. 908–917.
[49] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *ECCV*. Springer, 2020, pp. 665–681.
[50] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *ECCV*. Springer, 2020, pp. 52–69.
[51] 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 *ECCV*. Springer, 2020, pp. 275–292.
[52] 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.
[53] J. Zhou, L. Wang, H. Lu, K. Huang, X. Shi, and B. Liu, "Mvsalnet: Multi-view augmentation for rgb-d salient object detection," in *ECCV*. Springer, 2022, pp. 270–287.
[54] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*. IEEE, 2014, pp. 1115–1119.
[55] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *ICCV*, 2019, pp. 7254–7263.
[56] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*. IEEE, 2012, pp. 454–461.
[57] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *ICIMCS*, 2014, pp. 23–27.
[58] 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," *TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2020.
[59] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for rgb-d saliency detection," *TPAMI*, vol. 44, no. 12, pp. 9026–9042, 2021.
[60] 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 *ICCV*, 2021, pp. 4338–4347.
[61] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *CVPR*, 2020, pp. 3472–3481.
[62] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.
[63] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.
[64] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*. IEEE, 2009, pp. 1597–1604.
[65] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.
[66] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images." *ECCV*, vol. 7576, pp. 746–760, 2012.
[67] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576.
[68] H. Ye and D. Xu, "Inverted pyramid multi-task transformer for dense scene understanding," in *ECCV*. Springer, 2022, pp. 514–530.
[69] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *ICRA*. IEEE, 2021, pp. 13 525–13 531.

Hao Chen received the Ph.D. degree from the City University of Hong Kong in 2019. From 2019 to 2020, he worked as a Research Fellow with Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Computer Science and Engineering, Southeast University.

His research interests include human-inspired computer vision models and machine learning algorithms.

Feihong Shen received the B.S. degree in computer science and engineering from Southeast University, Nanjing, China, in 2022. He is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Southeast University, Nanjing, China.

His research interests include computer vision, multi-modal fusion, self-supervised learning and deep learning.

Ding Ding received his Ph.D. degree in Interactive Intelligence Group at the Delft University of Technology, the Netherlands. He is currently an associate professor in the school of computer science and engineering at Southeast University, China.

His research focuses on Interactive intelligence, human–computer interaction, and Human–Machine Symbiosis.

Yongjian Deng received the Ph.D. degree from the City University of Hong Kong in 2021. He is currently an Assistant Professor in the College of Computer Science, Beijing University of Technology.

His research interests include pattern recognition and machine learning with event cameras.

Chao Li received his MS Degree from Northwestern Polytechnical University in 2016. He currently works at Alibaba Group as a senior research engineer.

His main research interests include machine learning, computer vision and their applications.