Multi-view Aggregation Network for Dichotomous Image Segmentation

Qian Yu[†], Xiaoqi Zhao[†], Youwei Pang[†]
{ms.yuqian, zxq, lartpang}@mail.dlut.edu.cn
Lihe Zhang[*], Huchuan Lu
{zhanglihe, lhchuan}@dlut.edu.cn
Dalian University of Technology

Abstract

Dichotomous Image Segmentation (DIS) has recently emerged towards high-precision object segmentation from high-resolution natural images. When designing an effective DIS model, the main challenge is how to balance the semantic dispersion of high-resolution targets in the small receptive field and the loss of high-precision details in the large receptive field. Existing methods rely on tedious multiple encoder-decoder streams and stages to gradually complete the global localization and local refinement. Human visual system captures regions of interest by observing them from multiple views. Inspired by it, we model DIS as a multi-view object perception problem and provide a parsimonious multi-view aggregation network (MVANet), which unifies the feature fusion of the distant view and close-up view into a single stream with one encoder-decoder structure. With the help of the proposed multi-view complementary localization and refinement modules, our approach established long-range, profound visual interactions across multiple views, allowing the features of the detailed close-up view to focus on highly slender structures. Experiments on the popular DIS-5K dataset show that our MVANet significantly outperforms state-of-the-art methods in both accuracy and speed. The source code and datasets will be publicly available at MVANet.

1. Introduction

High-accuracy dichotomous image segmentation (DIS) [31] aims to accurately identify category-agnostic foreground objects within natural scenes, which is fundamental for a wide range of scene understanding applications, including AR/VR applications [30, 35], image editing [9], and 3D shape reconstruction [22]. Different from existing segmentation tasks, DIS focuses on challenging high-resolution (HR) fine-grained object segmentation. The segmentation scope encompasses a wide range of content with varying structural complexities, regardless of their characteristics. When confronted with the task of accurately segmenting HR objects, two primary challenges arise: 1) **The higher demand for segmentation capability.** Due to a larger amount of intricate details in high-accuracy HR images, accurately segmenting those objects of interest requires a more complex processing pipeline and more powerful feature modelling. And when dealing with occlusion interference, complex lighting conditions, and variable object poses, the processing of HR data also requires better adaptability and robustness compared to low-resolution (LR) data. 2) **The more need for processing efficiency.** The much larger size of HR images can result in slower processing speeds and more memory constraints. This restriction hinders the further application of existing approaches to real-world scenarios such as autonomous driving [15, 26] or real-time video processing [1, 20]. As a result, this field has higher expectations for inference efficiency in addition to ensuring algorithmic effectiveness.

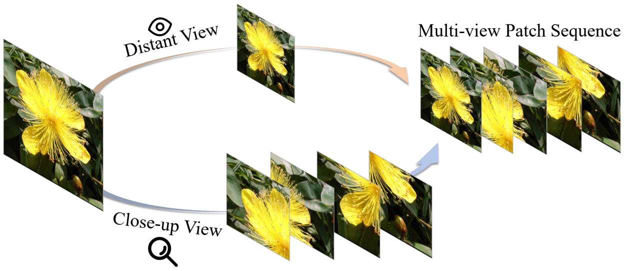


Figure 1. Process of decomposing high-resolution image into multi-view patch sequence.

Many efforts [28, 31, 47] have been made to tailor for the DIS task. Despite existing methods have demonstrated impressive performance, their reliance on CNN may pose limitations when tackling the HR image. It is be-

[†] Equal contribution.
[*] Corresponding author.

cause the increase in input resolution will result in a relatively small receptive field, which subsequently hinder the network's capacity to capture essential global semantics for the DIS task. Recently, with the introduction of the transformer [4, 23] with the global information propagation capability, the transformer-based methods [14, 40] have shown better prediction performance. However, features are extracted with global receptive field in these transformer-based methods, but they may not handle fine-grained local details as good as CNNs, which may be detrimental in high-precision segmentation tasks. Moreover, their multiple scale-independent models will increases the complexity of the feature pipeline and the redundancy of the model structure. Given that the input HR image itself contains all the information in the LR image, the multi-resolution inputs in these methods may lead to repetitive computation and information redundancy.

The core of solving the aforementioned issues is to design a parallel unified framework that can be compatible with global and local cues to avoid cascading forms of feature/model reuse. Inspired by the pattern of capturing high information content from images in the human visual system, we split the high-resolution input images from the original view into the distant view images with global information and close-up view images with local details. Thus, they can constitute a set of complementary multi-view low-resolution input patches, as shown in Fig. 1. In this paper, we make the attempt to address the HR image segmentation task by modeling it as a multi-view segmentation task. First, we design a parsimonious multi-view aggregation network (MVANet), which obtains global semantics and local features in parallel according to the characteristics of different patches. Such a design avoids the additional challenges caused by the hybridization of features in previous approaches. Second, we separately propose the novel multi-view complementary localization module (MCLM) and the multi-view complementary refinement module (MCRM). The MCLM incorporates our specially designed cross-attention mechanism driven by the global tokens and reverse attention mechanism, to enhance object localization and mitigating the local semantic gap between different patches. The MCRM aims to achieve a detailed depiction of the localized object, which is dominated by the local tokens, which is achieved through cross-attention mechanism with modeled multi-sensory global tokens. Subsequently, the enhanced local tokens are then used to refine the details in the global feature. Through the two-step process, we achieve a comprehensive representation of the scene, enabling effective object segmentation that takes into account both the overall context and the intricate details. Finally, we fuse all the patch output through a simple view rearrangement module and produce a highly accurate high-resolution prediction.

Our main contributions can be summarized as follows:

- The traditional single-view high-resolution image processing mode is upgraded to a multi-view processing mode based on multi-view learning.
- We propose the multi-view aggregation network (MVANet), which is the first single stream and single stage framework for the dichotomous image segmentation.
- Two efficient transformer-based multi-view complementary localization and refinement modules are proposed to jointly capturing the localization and restoring the boundary details of the targets.
- MVANet achieves state-of-the-art performance in terms of almost all metrics on the DIS benchmark dataset, while being twice as fast as the second-best method in terms of inference speed, demonstrating the superiority of our multi-view scheme.

2. Related works

2.1. Dichotomous Image Segmentation

Dichotomous image segmentation (DIS) is formulated as a category-agnostic task defined on non-conflicting annotations for accurately segmenting objects with various structural complexities, regardless of their characteristics. What sets it apart from classic segmentation tasks is the demand for highly precise object delineation, even down to the internal details of objects. Additionally, it addresses a broader range of objects, including salient[27, 45], camouflaged[7, 16], meticulous[19, 42], etc. Many efforts have been made to tailor for DIS, the first solution, IS-Net [31], tackles the DIS task by employing U^{2}Net as backbone and leveraging the intermediate supervision strategy. PF-DIS [47] is the first to leverage frequency priors to identify fine-grained object boundaries in DIS. Instead of using a general encoder-decoder architecture, UDUN [28] proposes a unite-divide-unite scheme to disentangle the trunk and structure segmentation for high-accuracy DIS. Although these works have achieved good performance, their reliance on CNN may pose limitations when tackling HR, high-accuracy tasks. It is because the increase in input resolution will result in a relatively small receptive field, which subsequently hinders the capacity of deep networks to capture essential global semantics necessary for the DIS task. Recently, Xie *et al.* [14] have proposed a novel architecture which enables to merge multiple results regardless of the size of the input. It is constructed to be trained with task-specified LR or HR inputs and generate HR output with a multi-resolution pyramid blending at the testing stage. However, the aforementioned methods usually relay multiple stages/streams to aggregation the global and local features, which will introduce additional drawbacks such as large parameters, low efficiency, and difficulty in optimization. In this paper, we focus on providing a parsimonious single stream and single stage baseline for the DIS task.

2.2. Multi-view Learning

Multi-view learning is an emerging direction in machine learning that leverages the use of multiple perspectives to enhance generalization performance [34]. It involves the utilization of distinct functions to model individual views, and collectively optimizes these functions to exploit other perspectives of the same input data, thereby enhancing overall learning performance [44]. In recent years, the integration of multi-view information with deep learning has garnered significant attention in many areas, such as 3D object recognition [33, 43], 3D reconstruction [17, 36, 37, 41], and feature matching [12, 32]. Su *et al.* [33] pioneered the utilization of multi-view 2D projected images as inputs, constructing a multi-view convolutional neural network to leverage information from object perspectives for 3D shape recognition. Moreover, Wang *et al.* [36]proposed a representative multi-view 3D reconstruction scheme, which encodes the relevant information amongst different views to jointly explore multi-level correspondence and associations between the 2D input views and 3D output volume with in a single unified framework. To this end, we're inspired to split the input with high-resolution image information into multi-view patch sequences to leverage complementary information for a more comprehensive understanding of visual data.

3. Method

In this section, we present the proposed approach in detail, including the overall architecture and specific components.

3.1. Overall Architecture

Multi-view Input. As illustrated in Fig. 2, the HR image input $\mathbf{I} \in \mathbb{R}^{B\times3\times H\times W}$ is resized to create the LR version $G \in \mathbb{R}^{B\times3\times h\times w}$, which simulates the distant view. Also, we evenly crop I into several non-overlapping local patches $\{L_m\}_{m=1}^{M} \in \mathbb{R}^{B\times3\times h\times w}$. Each of them can be seen as a specific close-up view focusing on the fine-grained texture. In this paper, we set M to be 4, *i.e.*, $(H, W) = (2h, 2w)$, and the corresponding discussion can be found in Sec. 4.4.
Multi-level Feature Extraction. G and $\{L_m\}_{m=1}^{M}$ together make up the multi-view patch sequence, which is fed in batches into the feature extractor to generate the multi-level feature maps, *i.e.*, $\{E_i | i = 1, 2, 3, 4, 5\}$. Each E_i includes representations of both the distant and close-up views.
Complementary Localization. The feature map E_5 from the highest level is partitioned along the batch dimension into the two different sets, *i.e.*, global and local features. They are fed into the multi-view complementary localization module (MCLM) to highlight the positional information about the object within the global representation. It is subsequently used to guide the local representation for object localization and effectively filter out erroneous information from the close-up view. After the MCLM, the updated global and local feature maps are concatenated along the batch dimension to form a single feature map $D_5 \in \mathbb{R}^{B\times3\times\frac{h}{32}\times\frac{w}{32}}$, which is sent to the well-designed top-bottom decoder.

Refinement Decoding. Our novel network differs from the classic FPN [21]-like architecture. We insert the on-the-fly multi-view complementary refinement module (MCRM) in each decoding stage as shown in Fig. 2. These models can dynamically optimize missing fine-grained details in the global representation with information from the local representations. And shallow features are also absorbed layer by layer into the upsampling path in the decoder.

Multi-view Integration. As illustrated in the bottom-right section of Fig. 2, we introduce a simple view rearrangement module to merge the positional and semantic information from the distant view with the detailed information from the close-up view, into a unified whole. After the aforementioned steps, we can obtain $D_1^{merge} \in \mathbb{R}^{B\times3\times\frac{h}{2}\times\frac{w}{2}}$, whose shape is a quarter of the shape of the original image when $M = 4$. Instead of directly upsampling it by 4 times, we incorporate shallow features[18] as low-level visual cues to further enhance the quality of image segmentation.

3.2. Multi-view Complementary Localization

To tackle the challenge of jointly localizing objects through distant view and close-up views, we propose the multi-view complementary localization module (MCLM). The well-constructed process facilitates the proposed model in attaining the holistic scene understanding and effectively identifying potential areas of interest, thereby accomplishing the goal of jointly localizing close-up views and distant views. First, we divide the E_5 into global feature $E_5^G \in \mathbb{R}^{B\times C\times\frac{H}{32}\times\frac{W}{32}}$ and local features $\{E_5^{L_m}\}_{m=1}^{M}$ where $E_5^{L_m} \in \mathbb{R}^{B\times C\times\frac{h}{32}\times\frac{w}{32}}$ and M denotes the number of local features. Subsequently, the local features are assembled into a unified global feature $E_5^{L_g} \in \mathbb{R}^{B\times C\times\frac{H}{32}\times\frac{W}{32}}$ by aligning with their respective positions in the original image. To simultaneously obtain the rich visual representation and capture important contextual feature cues, we embed the multi-granularity pooling operation into the vanilla transformer block [39, 49], which reduces the computational cost of `MHCA` by 56.25% while facilitating deeper interaction between multiple views. Specifically, we apply multiple average pooling layers with various receptive fields onto the aforementioned unified global feature, thereby generating

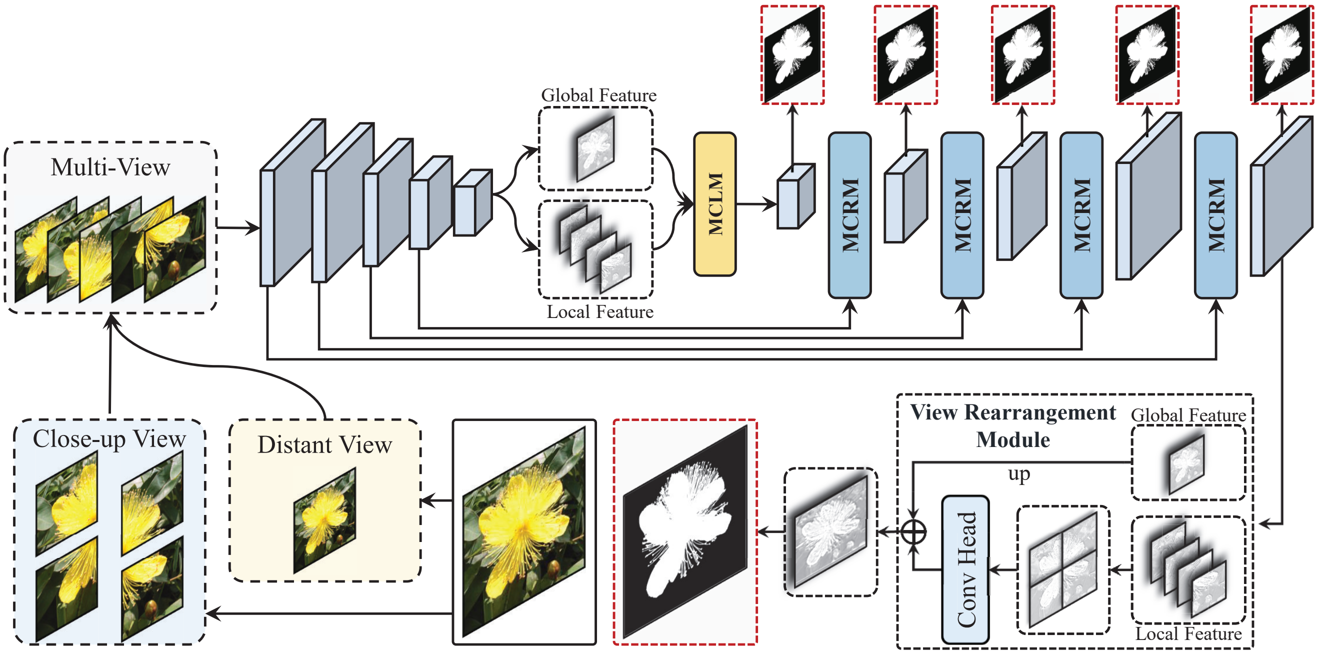
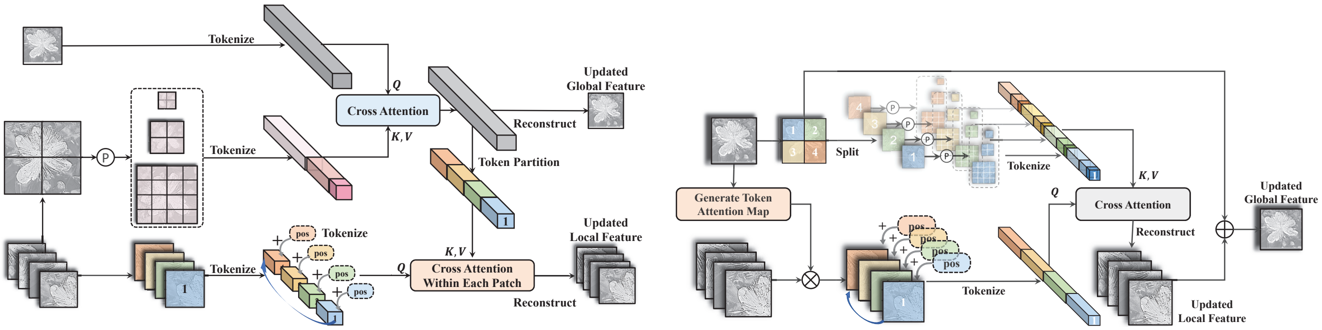


Figure 2. Overall framework of the proposed MVANet. The downsampled original image and non-overlapping local patches are adopted as inputs for the global context and detailed cues, representing distant and close-up views, respectively. To enhance object localization and achieve detailed depiction, we propose multi-view complementary localization module (MCLM) and refinement module (MCRM), respectively. Besides, a view rearrangement module is introduced to integrate multiple views, thereby generating predictions with highly accurate dominant areas while preserving detailed object structures. The red dashed box indicates the location that is deeply supervised.



(a) Multi-view Complementary Localization Module.

(b) Multi-view Complementary Refinement Module.

Figure 3. Pipeline of the proposed multi-view complementary localization and refinement modules. Ⓟ represents the multi-granularity pooling operation.

pyramid feature maps:

$$\begin{aligned} P_1 &= \texttt{AvgPool}_1(E_5^{L_g}), \\ P_2 &= \texttt{AvgPool}_2(E_5^{L_g}), \\ &\dots, \\ P_n &= \texttt{AvgPool}_n(E_5^{L_g}), \end{aligned} \tag{1}$$

where n denotes the number of parallel pooling branches. And we set the respective receptive fields to be 4, 8, 16 in practice. These maps are then tokenized and concatenated to be K and V for the MHCA block:

$$K, V = [T(P_1), T(P_2), \dots, T(P_n)]W^{K,V}, \tag{2}$$

where $W^{K,V} \in \mathbb{R}^{C \times 2C}$ is used to transform all branches. And $T(\cdot)$ indicates the tokenization operation which is achieved with a flattening process for simplicity. The operation $[\cdot]$ concatenates all sequences into a single one. Besides, the global feature is also tokenized directly to be Q for MHCA:

$$Q = T(E_5^G)W^Q, \tag{3}$$

where $W^Q \in \mathbb{R}^{C\times C}$ is a projection matrix. Then, Q, K and V are fed into MHCA, followed by LN and FFN as shown in the "Cross-Attention" part of Fig. 3a:

$$T^G = T(E_5^G) + \texttt{LN}(\texttt{MHCA}(Q,K,V)), \tag{4}$$

$$T^G = T^G + \texttt{LN}(\texttt{FFN}(T^G)). \tag{5}$$

The updated global token $T^G \in \mathbb{R}^{\frac{HW}{32^2}\times B\times C}$ can be utilized to reconstruct and update the global feature, where we can obtain $F^{G'}$ for further processing. Besides, in order to utilize it to further assist in the activation of object-related cues in the local field of view, we also rearrange and partition $T^G \in \mathbb{R}^{\frac{HW}{32^2}\times B\times C}$ according to the order of patch tokens, referred as $\{T^{G_m}\}_{m=1}^M$. To effectively remain the positional correlation between different views, we supplement the position encoding into these local features as shown in Fig. 3a. Subsequently, we tokenize them and apply MHCA within each patch, where the local tokens is used as Q and the rearranged global tokens as K and V:

$$Q_m = T(E_5^{L_m})W^{Q_m}, \tag{6}$$

$$T^{L'_m} = \texttt{MHCA}(Q_m, T^{G_m}, T^{G_m}). \tag{7}$$

Finally, based on the updated local tokens $\{T^{L'_m}\}_{m=1}^M$, a straightforward unflatten and reshape procedure is applied to generate the reconstructed local features $\{E_5^{L'_m}\}_{m=1}^M$, which are then simply concatenated in batches with the updated global feature $E_5^{G'}$ to form the feature map D_5 for subsequent processing.

3.3. Multi-view Complementary Refinement

After the LR global feature provides a broader context aiding in coarse-level identification, we introduce the multi-view complementary refinement module (MCRM) as shown in Fig. 3b. In this module, local features provide localized and detailed views to enhance the accuracy and robustness of segmentation. To be specific, the input feature is denoted as D_i, where $i \in \{1,2,3,4,5\}$ represents the layer number of the decoder. Similar to the MCLM, we partition the feature D_i into global feature D_i^G and local features $\{D_i^{L_m}\}_{m=1}^M$ along the batch dimension. To filter out background noise from the local features, a one-channel token attention map A is initially generated using a 1×1 convolution layer followed by a sigmoid function. A is subsequently utilized to the modulate feature map and eliminate the background noise, thereby obtaining a purer representation for the object segmentation. And the aforementioned operations can be formulated as:

$$A = \texttt{sigmoid}(\texttt{conv}(D_i^G)), \tag{8}$$

$$\{D_i^{L_m}\}_{m=1}^M = \texttt{split}(A \odot \texttt{assemble}(\{D_i^{L_m}\}_{m=1}^M)), \tag{9}$$

where $\odot$ is the Hadamard product. assemble and split are a pair of opposite operations. The former rearranges the independent patches into to the original image form, while the latter reverses the process to the patch sequence. After that, as in the MCLM, the individual position encoding is also added to each local feature to model their positional relationships. We then tokenize and concatenate these features to serve as Q for the cross attention:

$$T_i^{L_m} = [T(D_i^{L_1}), T(D_i^{L_2}), \dots, T(D_i^{L_m})], \tag{10}$$

$$Q_i = [T_i^{L_1}, T_i^{L_2}, \dots, T_i^{L_m}]W^{Q_i}. \tag{11}$$

Besides, we partition the global feature into corresponding regions based on the original positions of each local feature:

$$\{D_i^{G_m}\}_{m=1}^M = \texttt{split}(D_i^G). \tag{12}$$

And then, a similar multi-granularity pooling process as in the MCLM is imposed in these patch-wise features $\{D_i^{G_m}\}_{m=1}^M$ as shown in Fig. 3b, which involves the extraction of contextual information through the utilization of multiple branches with varying receptive fields. After the transformation and concatenation, we can obtain the multi-sensory tokens $T_i^{G_m}$ with different contextual abstraction levels in the m^{th} patch. These tokens are then concatenated into a unified whole, as K and V for the cross attention:

$$K_i, V_i = [T_i^{G_1}, T_i^{G_2}, \dots, T_i^{G_m}]W^{K_i,V_i}. \tag{13}$$

During the cross attention operation, we employ a vanilla transformer block to facilitate interaction between informative local tokens and multi-sensory tokens from corresponding regions in the global context. Then, we reconstruct the updated local tokens to the local features $\{D_i^{L'_m}\}_{m=1}^M$ by adjusting the shape, which are then integrated into the original global feature by the addition operation to obtain globally optimized features $D_i^{G'}$ with enhanced details. Finally, these two sets of features are concatenated along the batch dimension, resulting in a detail-enhanced feature map:

$$D'_i = [\{D_i^{L'_m}\}_{m=1}^M, D_i^{G'}]. \tag{14}$$

After repetitively stacking the decoding components while continuously integrating the multi-level features from the encoder, we can obtain output features D'_1 with a higher resolution, which incorporates the broader context and the fine-grained locality.

3.4. View Rearrangement

The patch-based local enhancement strategy can retain sufficient texture details for the model, but also introduces the problem of misalignment between neighboring patch boundaries when reorganizing the patches into the image. In our decoder embedded with the refinement module, the

Datasets	*Metric*	F³Net	GCPANet	PFNet	BSANet	ISDNet	IFA	PGNet	IS-Net	FP-DIS	UDUN	InSPyReNet	Ours
		[38]	[2]	[25]	[48]	[10]	[13]	[40]	[31]	[47]	[28]	[14]	
		R-50	R-50	R-50	R2-50	R-50	R-50	S+R	-	R-50	R-50	Swin-B	Swin-B
DIS-TE1	$F_\beta^{max}\uparrow$	0.726	0.741	0.740	0.683	0.717	0.673	0.754	0.740	0.784	0.784	0.845	**0.893**
	$F_\beta^{\omega}\uparrow$	0.655	0.676	0.665	0.545	0.643	0.573	0.680	0.662	0.713	0.720	0.788	**0.823**
	$E_\phi^{m}\uparrow$	0.820	0.834	0.830	0.773	0.824	0.785	0.848	0.820	0.860	0.864	0.874	**0.911**
	$S_m\uparrow$	0.783	0.797	0.791	0.754	0.782	0.746	0.800	0.787	0.821	0.817	0.873	**0.879**
	$\mathcal{M}\downarrow$	0.074	0.070	0.075	0.098	0.077	0.088	0.067	0.074	0.060	0.059	0.043	**0.037**
DIS-TE2	$F_\beta^{max}\uparrow$	0.789	0.799	0.796	0.752	0.783	0.758	0.807	0.799	0.827	0.829	0.894	**0.925**
	$F_\beta^{\omega}\uparrow$	0.719	0.741	0.729	0.628	0.714	0.666	0.743	0.728	0.767	0.768	0.846	**0.874**
	$E_\phi^{m}\uparrow$	0.860	0.874	0.866	0.815	0.865	0.835	0.880	0.858	0.893	0.886	0.916	**0.944**
	$S_m\uparrow$	0.814	0.830	0.821	0.794	0.817	0.793	0.833	0.823	0.845	0.843	0.905	**0.915**
	$\mathcal{M}\downarrow$	0.075	0.068	0.073	0.098	0.072	0.085	0.065	0.070	0.059	0.058	0.036	**0.030**
DIS-TE3	$F_\beta^{max}\uparrow$	0.824	0.844	0.835	0.783	0.817	0.797	0.843	0.830	0.868	0.865	0.919	**0.936**
	$F_\beta^{\omega}\uparrow$	0.762	0.789	0.771	0.660	0.747	0.705	0.785	0.758	0.811	0.809	0.871	**0.890**
	$E_\phi^{m}\uparrow$	0.892	0.909	0.901	0.840	0.893	0.861	0.911	0.883	0.922	0.917	0.940	**0.954**
	$S_m\uparrow$	0.841	0.855	0.847	0.814	0.834	0.815	0.844	0.836	0.871	0.865	0.918	**0.920**
	$\mathcal{M}\downarrow$	0.063	0.068	0.062	0.090	0.065	0.077	0.056	0.064	0.049	0.050	0.034	**0.031**
DIS-TE4	$F_\beta^{max}\uparrow$	0.815	0.831	0.816	0.757	0.794	0.790	0.831	0.827	0.846	0.846	0.905	**0.911**
	$F_\beta^{\omega}\uparrow$	0.753	0.776	0.755	0.640	0.725	0.700	0.774	0.753	0.788	0.792	0.848	**0.857**
	$E_\phi^{m}\uparrow$	0.883	0.898	0.885	0.815	0.873	0.847	0.899	0.870	0.906	0.901	0.936	**0.944**
	$S_m\uparrow$	0.826	0.841	0.831	0.794	0.815	0.841	0.811	0.830	0.852	0.849	**0.905**	0.903
	$\mathcal{M}\downarrow$	0.070	0.064	0.072	0.107	0.079	0.085	0.065	0.072	0.061	0.059	0.042	**0.041**
Overall	$F_\beta^{max}\uparrow$	0.789	0.804	0.797	0.744	0.778	0.755	0.809	0.799	0.831	0.831	0.891	**0.916**
	$F_\beta^{\omega}\uparrow$	0.722	0.746	0.730	0.618	0.707	0.661	0.746	0.726	0.770	0.772	0.838	**0.855**
	$E_\phi^{m}\uparrow$	0.864	0.879	0.871	0.811	0.864	0.832	0.885	0.858	0.895	0.892	0.917	**0.938**
	$S_m\uparrow$	0.816	0.831	0.823	0.789	0.812	0.791	0.830	0.819	0.847	0.844	0.900	**0.905**
	$\mathcal{M}\downarrow$	0.071	0.065	0.071	0.098	0.073	0.084	0.063	0.070	0.057	0.057	0.039	**0.035**

Table 1. Quantitative comparison of DIS5K with 11 representative methods. ↓ represents the lower value is better, while ↑ represents the higher value is better. The best score is highlighted in **bold**. R-50, R2-50, and SwinB respectively denote the utilization of ResNet-50[11], Res2Net-50[8], and Swin-B[23] as backbones, while S+R represents the combination of Swin-B and ResNet-50 as a new backbone.

iterative dense interactions between patches and global features alleviate this problem. And we make further optimizations in the cascaded view rearrangement module as shown in Fig. 2 Specifically, we split the local features in D_1' along batch dimension and `assemble` them into a global form. To address aforementioned issue, we introduce a convolutional head that consists of three convolutional layers interspersed with `BN` and `ReLU` layers, to smooth the features as in Fig. 2. This architecture is purposefully tailored to prioritize the enhancement of patch alignment. Subsequently, the aligned feature is added to the global feature split from D_1' to further enhance image quality, and used to generate the final segmentation map.

3.5. Loss Function

As illustrated in the Fig. 2, we incorporate supervision at each layer output of the decoder and the final prediction. Specifically, the former consists of three components: l_l, l_g and l_a for the assembled local representation, the global representation, and the token attention map in the refinement module, respectively. Note that the side outputs here each require a separate convolutional layer to obtain a single-channel prediction. And the latter is represented as l_f. These components employ the combination of the binary cross-entropy (BCE) loss and the weighted IoU loss, following the common practice in most segmentation tasks [28, 46, 47]:

$$l = l_{BCE} + l_{IoU}. \tag{15}$$

To this end, our total loss can be written as:

$$L = l_f + \sum_{i-1}^{5} (l_l^i + \lambda_g l_g^i + \lambda_a l_a^i), \tag{16}$$

where λ_g and λ_h are set to 0.3 in our experiment.

4. Experiments

4.1. Datasets and Evaluation Metrics

Data Settings. We conduct extensive experiments on the DIS5K [31] benchmark dataset, comprising $5,470$ HR images (e.g., 2K, 4K or larger) across 225 categories. The dataset is partitioned into three subsets: DIS-TR, DIS-VD, and DIS-TE. DIS-TR and DIS-VD consist of $3,000$ training images and 470 validation images, respectively. DIS-TE is further divided into four subsets (DIS-TE1, 2, 3, 4) with increasing shape complexities, each containing 500 images. With its diverse objects featuring varying geometric structures and appearances, the DIS5K dataset presents higher resolution images, intricate structural details, and superior annotation accuracy compared to existing object segmentation datasets. As a result, segmentation on DIS5K proves to

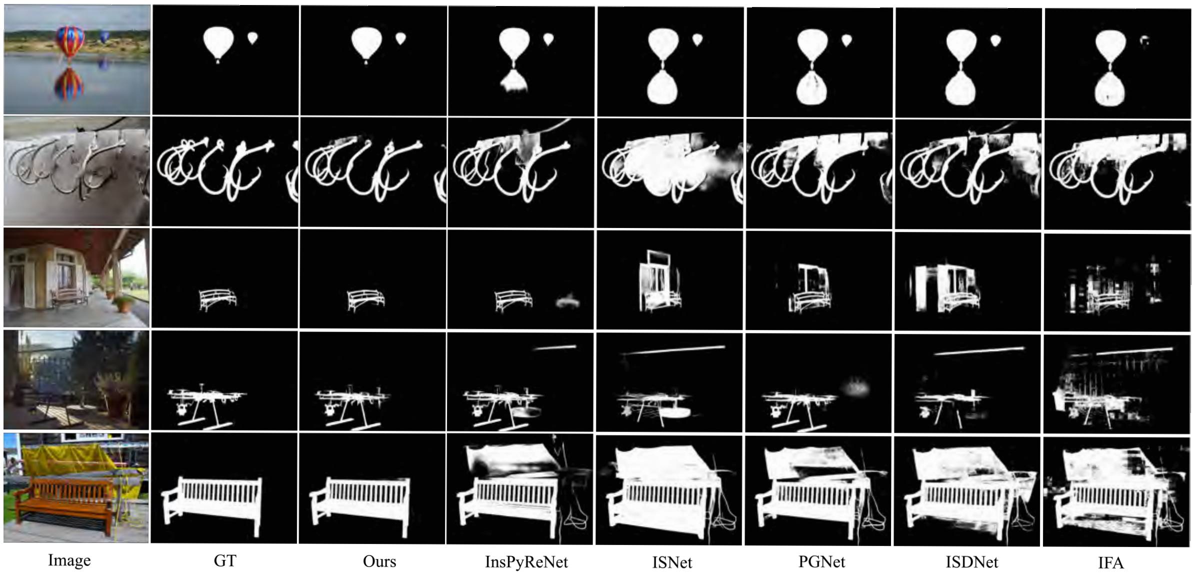


Figure 4. Visual comparison of different DIS methods.

be challenging and necessitates models with robust capabilities in identifying structural details.

Evaluation Metrics. For evaluating the results, we adopt some widely used metrics, including max F-measure (F_β^{max}) [29], weighted F-measure (F_β^ω) [24], structural similarity measure (S_m) [5], E-measure (E_ϕ^m) [6] and mean absolute error (MAE, $\mathcal{M}$) [29]. F_β^{max} and F_β^ω are the maximum and weighted scores of the precision and recall, respectively, where β^2 is set to 0.3. S_m simultaneously evaluates region-aware and object-aware structural similarity between the prediction and mask. E_ϕ^m is widely used for evaluating pixel-level and image-level matching. MAE measures the average error of the prediction maps.

4.2. Implementation Details

Experiments are implemented in PyTorch on a single RTX 3090 GPU. During the training phase, the original images are first resized to 1024 × 1024. Then, both the With the number of patches set to 4, the resulting patch size is 512 × 512. Consequently, the low-resolution (LR) global image is also resized to 512 × 512. Swin-B [23] is used as the backbone with the pre-trained weights on the ImageNet [3], while other parameters are initialized randomly. To avoid overfitting, we adopt some data augmentation techniques, including random horizontal flipping, cropping and rotating. We use the Adam optimizer with an initial learning rate of 0.00001. The batch size is set to 1, and the maximum number of epochs is set to 80.

4.3. Comparison with State-of-the-arts

Quantitative Evaluation. In Tab. 1, we compare our proposed MVANet with other 11 well-known task-related models, including F^3Net [38], GCPANet [2], PFNet [25], BSANet [48], ISDNet [10], IFA [13], IS-Net [31], FP-DIS [47], UDUN [28], PGNet [40], InSPyNet [14]. For a fair comparison, we standardize the input size of the comparison models to 1024 × 1024. It can be seen that MVANet significantly outperforms the other models on all the datasets under different metrics. In particular, ours outperforms the second-best model (InSPyReNet) with the gain of 2.5% , 2.1% , 0.5% , 0.4% in terms of the F_β^{max}, E_ϕ^m, S_m and MAE, respectively. Besides, we evaluate the inference speed for the InSPyReNet and ours. Both of them are tested under the same NVIDIA RTX 3090 GPU. Benefiting from the parsimonious single stream design, MVANet achieves the 4.6 FPS over the InSPyReNet with the 2.2 FPS.

Qualitative Evaluation. To demonstrate the highly accurate prediction of our model in an intuitive perspective, we visualize the output of some images selected from the test set. As shown in Fig. 4, our model can capture both the accurate object localization and edge details under different complex scenes. In particular, other methods suffer from interference from the salient yellow gauze and shadows, whereas our model allows for a complete segmentation of the chair and accurate differentiation of the interior for each grille (see the last row).

4.4. Ablation Study

In this section, we analyze the effects of each component. All the results are tested on the DIS-TE1.

Diverse Views Inputs. To investigate the effectiveness of our multi-view input strategy, we conduct a series of experiments involving different inputs, as shown in Tab. 2. First, we separately list the in-dependent results of the "HR-Ori",

HR-Ori	LR-Dis	HR-Clo	$F_\beta^{max} \uparrow$	$E_\phi^m \uparrow$	$S_m \uparrow$	$\mathcal{M} \downarrow$
✓			0.822	0.869	0.812	0.058
	✓		0.815	0.858	0.801	0.058
		✓	0.801	0.814	0.759	0.069
✓	✓		0.875	0.897	0.862	0.041
	✓	✓	**0.893**	**0.911**	**0.879**	**0.037**

Table 2. Ablation experiments of diverse views inputs. ”HR-Ori” is the high-resolution original images. ”LR-Dis” is the low-resolution global images generated by resizing the ”HR-Ori”. ”HR-Clo” is the 4 non-overlapping patches evenly cropped from the ”HR-Ori”.

MCLM	MCRM	VRM	$F_\beta^{max} \uparrow$	$E_\phi^m \uparrow$	$S_m \uparrow$	$\mathcal{M}\downarrow$	FPS ↑
			0.822	0.869	0.812	0.058	**9.2**
	✓	✓	0.880	0.889	0.866	0.038	5.71
✓		✓	0.884	0.903	0.870	0.041	5.38
✓	✓		0.888	0.897	0.874	0.039	4.76
✓	✓	✓	**0.893**	**0.911**	**0.879**	**0.037**	4.62

Table 3. Ablation experiments of each component.

Number	$F_\beta^{max} \uparrow$	$E_\phi^m \uparrow$	$S_m \uparrow$	$\mathcal{M} \downarrow$
4	**0.893**	**0.911**	**0.879**	**0.037**
9	0.821	0.856	0.799	0.058
16	0.717	0.751	0.745	0.081

Table 4. Ablation experiments of the number of the patches in the sequence for the close-up view.

”LR-Dis”, ”HR-Clo” models. The ”HR-Orr” will as the performance anchor to show the effectiveness of the multi-view inputs strategy. Next, the gap between the first and fourth rows show the effectiveness of low-resolution global images in providing the distant view. Then, the gap between the fourth and last rows can illustrate the necessity of the local patches in providing the close-up view. Finally, the combined global and local multi-view inputs provide a complete set of target perceptual cues with no mutually exclusive effects on each other.

Effectiveness of MCLM. In Tab. 3, the comparison of the second row with last row shows the effectiveness of the proposed multi-view complementary localization module (MCLM). The utilization of pyramid pooling for generating multi-sensory tokens enables the identification of targets within tokens at minimal cost and facilitates long-range visual interactions across multiple local views. Therefore, the MCLM guarantees an increase in performance while decreasing speed by only 1.09 FPS.

Effectiveness of MCRM. The gap between the third row and last row verify the effectiveness of the proposed multi-view complementary refinement module (MCRM). Although the pooling operation is used in MCRM, it is only embed in the global features. Once applying cross-attention on shallow features, the computational cost will increases significantly due to the larger feature sizes. By generating multi-sensory tokens, we can reduce the sequence length and thereby greatly reduce the computational burden.

Effectiveness of VRM. Thanks to the sufficient feature fusion among local patch features in the decoder, we only need to perform a simple convolution operation at the tail of the MVANet to accomplish an effective view rearrange and obtain a complete high-resolution prediction. Finally, we can see that the combination of MCLM, MCRM and VRM can separately achieve more than 8%, 4%, 8%, 36% performance gain in terms of the F_β^{max}, E_ϕ^m, S_m and MAE, compared to the FPN baseline (the first row) with only the original view image input.

Patch Quantity. To thoroughly investigate the impact of patch quantity on our work, we crop the original image into 4, 9, and 16 patches, serving as the close-up view inputs. As shown in Tab. 4, performance degrades as the number of patches increases. It may be attributed to two reasons: 1) While an increase in patch quantity and a decrease in resolution may enhance processing speed, it also results in reduced information within each close-up view, weakened connectivity between patches, and even instances where diminished receptive fields lead to noise being mistaken for foreground objects. 2) As the resolution decreases for each patch, the resolution of the LR global image correspondingly diminishes, leading to significant information loss that is detrimental to accurate object localization.

5. Conclusion

In this paper, we tackle the high-accuracy DIS by modeling it as a multi-view object perception problem and provide a parsimonious, streamlined multi-view aggregation network, aiming at making a better trade-off among model designs, accuracy, and the inference speed. To address the target alignment problem for multiple views, we propose the multi-view complementary localization module to jointly calculate the co-attention region of the target. Besides, the proposed multi-view complementary refinement module are embed into each decoder block to fully integrate complementary local information and mitigate the semantic deficit of a single view patch, thus the final view rearrangement can be accomplished with only a single convolutional layer. Extensive experiments show that our model performs well on the DIS dataset.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 62276046 and by Dalian Science and Technology Innovation Foundation under Grant 2023JJ12GX015.

References

[1] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Strpm: A spatiotemporal residual predictive model for high-resolution video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13946–13955, 2022. 1

[2] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10599–10606, 2020. 6, 7

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 7

[6] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 7

[7] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 2

[8] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 6

[9] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1915–1926, 2011. 1

[10] Shaohua Guo, Liang Liu, Zhenye Gan, Yabiao Wang, Wuhao Zhang, Chengjie Wang, Guannan Jiang, Wei Zhang, Ran Yi, Lizhuang Ma, et al. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4361–4370, 2022. 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 3

[13] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Learning implicit feature alignment function for semantic segmentation. In *European Conference on Computer Vision*, pages 487–505. Springer, 2022. 6, 7

[14] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124, 2022. 2, 6, 7

[15] Marvin Klingner, Konstantin Müller, Mona Mirzaie, Jasmin Breitenstein, Jan-Aike Termöhlen, and Tim Fingscheidt. On the choice of data for efficient training and validation of end-to-end driving models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4803–4812, 2022. 1

[16] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019. 2

[17] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 3

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 3

[19] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 2

[20] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[22] Feng Liu, Luan Tran, and Xiaoming Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433, 2021. 1

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6, 7

[24] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 248–255, 2014. 7

[25] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021. 6, 7

[26] Ferdinand Mütsch, Helen Gremmelmaier, Nicolas Becker, Daniel Bogdoll, Marc René Zofka, and J Marius Zöllner. From model-based to data-driven simulation: Challenges and trends in autonomous driving. *arXiv preprint arXiv:2305.13960*, 2023. 1

[27] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9413–9422, 2020. 2

[28] Jialun Pei, Zhangjun Zhou, Yueming Jin, He Tang, and Pheng-Ann Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2139–2147, 2023. 1, 2, 6, 7

[29] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012. 7

[30] X Qin, DP Fan, C Huang, C Diagne, Z Zhang, AC Sant'Anna, A Suarez, M Jagersand, and L Shao. Boundary-aware segmentation network for mobile and web applications. arxiv 2021. *arXiv preprint arXiv:2101.04704*. 1

[31] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 1, 2, 6, 7

[32] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 3

[33] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 3

[34] Shiliang Sun. A survey of multi-view machine learning. *Neural computing and applications*, 23:2031–2038, 2013. 3

[35] Yang Tian, Hualong Bai, Shengdong Zhao, Chi-Wing Fu, Chun Yu, Haozhao Qin, Qiong Wang, and Pheng-Ann Heng. Kine-appendage: Enhancing freehand vr interaction through transformations of virtual appendages. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1

[36] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5722–5731, 2021. 3

[37] Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun Bao, and Xiaowei Zhou. Autorecon: Automated 3d object discovery and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21382–21391, 2023. 3

[38] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12321–12328, 2020. 6, 7

[39] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[40] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11717–11726, 2022. 2, 6, 7

[41] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2690–2698, 2019. 3

[42] Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zhe Lin, and Alan Yuille. Meticulous object segmentation. *arXiv preprint arXiv:2012.07181*, 2020. 2

[43] Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 186–194, 2018. 3

[44] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017. 3

[45] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 35–51. Springer, 2020. 2

[46] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3463–3471, 2022. 6

[47] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. 1, 2, 6, 7

[48] Hongwei Zhu, Peng Li, Haoran Xie, Xuefeng Yan, Dong Liang, Dapeng Chen, Mingqiang Wei, and Jing Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3608–3616, 2022. 6, 7

[49] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 593–602, 2019. 3