Highly Accurate Dichotomous Image Segmentation

Xuebin Qin[1], Hang Dai[1], Xiaobin Hu[2], Deng-Ping Fan*[3], Ling Shao[4], and Luc Van Gool[3]

[1] MBZUAI, Abu Dhabi, UAE
[2] Tencent Youtu Lab, Shanghai, China
[3] ETH Zurich, Switzerland
[4] Terminus Group, China
[5] xuebin@ualberta.ca, hang.dai@mbzuai.ac.ae, xiaobin.hu@tum.de, dengpfan@gmail.com, ling.shao@ieee.org, vangool@vision.ee.ethz.ch

1 Related Work

1.1 Multi-class *vs.* Dichotomous Segmentation

Multi-class (*e.g.*, semantic [39], panoptic [32]) segmentation aims at simultaneously labeling all the pixels in an image of complex scenario [10, 77], which contains many different objects, with the pre-defined multiple categories encoded in one-hot vectors. However, the one-hot representation of the categories is memory exhaustive when the number of categories is huge (*e.g.*, 10,000 categories), especially on high-resolution images. Besides, some input images only contain objects from several categories (*e.g.*, one or two). Outputting the full-length one-hot dense predictions (10,000 categories) is not a resource-saving option. A possible alternative could be a two-step solution: "detection + segmentation", in which a bounding box and category of the certain object can be predicted first. The segmentation process can then be conducted in a dichotomous way within the bounding box region by producing a single-channel probability map (*e.g.*, similar to Mask R-CNN [24]. However, Mask R-CNN still uses the one-hot representation in the segmentation step).

Moreover, many practical applications, such as image editing, art design, shape from silhouette, robot manipulation, are usually category-agnostic, where the applications require highly accurate segmentation results of certain objects regardless of their categories. Different from the images of complex scenarios in semantic [36] or panoptic [77] segmentation, the images in these applications usually contain one or a few objects with very high resolutions, less occlusions. To this end, many related tasks have been proposed, such as salient object detection (SOD) [9,38,44,58,61,64,66], salient object in clutter (SOC) [14], high-resolution salient object detection (HRS) [68], camouflaged object detection (COD) [17, 29, 54], thin object segmentation (TOS) [34], meticulous object segmentation (MOS) [65], video object segmentation (VOS) [48], class-agnostic very high-resolution segmentation (VHRS) [8], *etc.* Most of these tasks try to solve dichotomous segmentation problems on images which are sharing specific characteristics. The

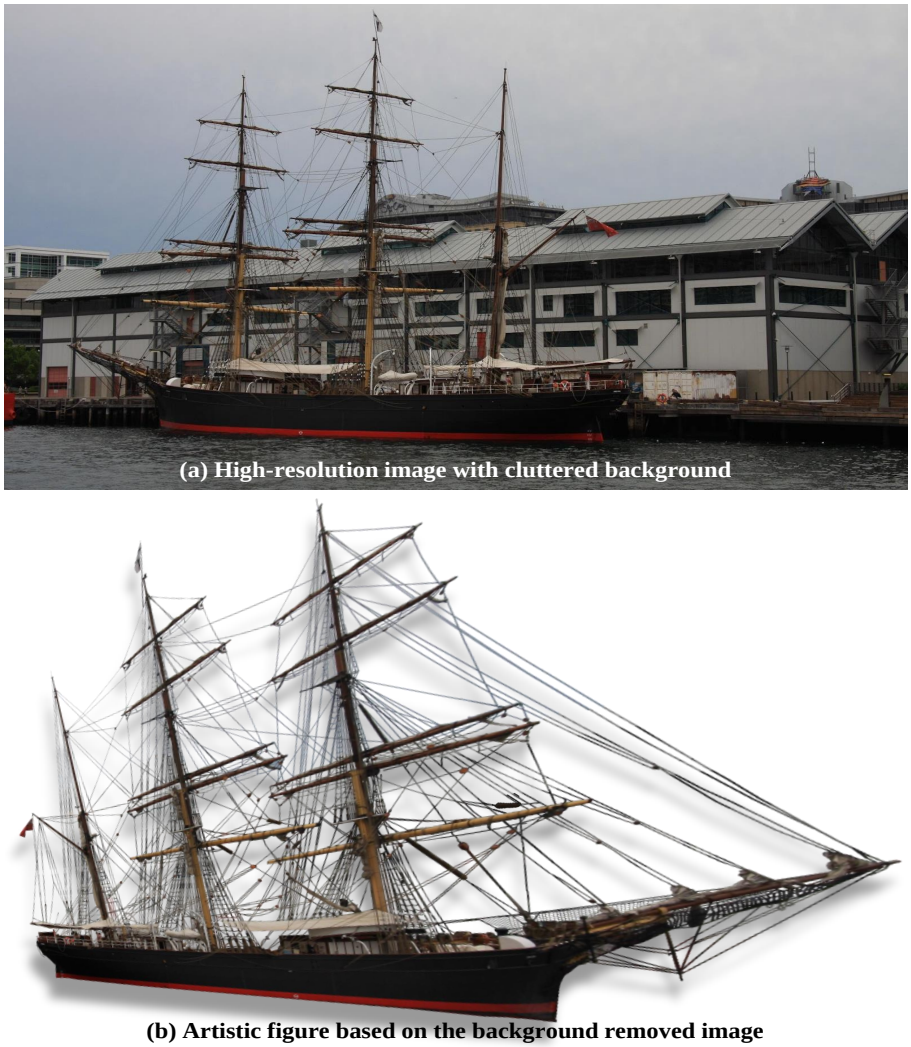


Fig. 1: Demo application: artistic figure generated based on a sample of our DIS5K dataset.

exclusive mechanisms for certain tasks are barely used so that their problem formulations are almost the same, which means most of these tasks are data-dependent. Simply combining these tasks by merging their datasets is not a decent option because these tasks' image resolutions and labeling qualities are diversified.

Considering these facts, we re-formulate a new category-agnostic dichotomous segmentation task, *highly accurate Dichotomous Image Segmentation (DIS)*, where achieving highly accurate segmentation results of objects with diversified shapes and structures is the key concern.

1.2 Datasets

Datasets are the basis of most computer vision tasks. In the past decades, many segmentation datasets for related tasks have been created. For example, semantic (PASCAL-VOC [13], MS-COCO [36]) and panoptic (Cityscapes [10], ADE20K [77]) segmentation (SMS) datasets usually contain large number of images with multiple objects from different categories in each of them. But they either have low geometrical labeling accuracy or relatively small resolutions, where details of objects are hard to be included and segmented. The entity segmentation (ES) [49] datasets proposed for class-agnostic segmentation has similar issues. Images in the salient object detection (SOD) [9, 33, 44, 58, 66] and camouflaged object detection (COD) [17] datasets are usually low-resolution ones, which contains objects with simple structures. The high-resolution salient object detection (HRS) [48,68] datasets have higher resolution, but they are built upon images with objects of simple structures similar to that in SOD and COD datasets. The meticulous object segmentation (MOS) [65] and thin object segmentation (TOS) [34] datasets show competitive resolution and object structure complexity characteristics. However, MOS is too small to enable thorough training and comprehensive evaluation, while the TOS dataset is built with synthetic images. Therefore, there is a need for a new *extendable large-scale* dataset built upon the *high-resolution* images with *diversified object structure complexities* and *highly accurate labeling*.

1.3 Existing Models

Models are the cores of vision tasks. Currently, deep models are the most popular solutions for most of the segmentation tasks. Many different deep architectures have been proposed to achieve better performance, such as FCN-based [39] feature aggregation models [5,25,41,57,62,69,70,75], Encoder-Decoder architectures [2, 6, 50, 53], Coarse-to-Fine (or Predict-Refine) models [8, 11, 35, 51, 56, 59, 60], Vision Transformers [37, 76], *etc.* Besides, many real-time models [18, 27, 31, 45, 46, 67, 72] are developed to balance the performance and time costs. To achieve highly accurate results in our DIS, the models are expected to capture fine details (and complicated structures) and large components of the diversified objects from large-size (*e.g.*, 2K, 4K or even larger) images with affordable memory, computation and time costs. These requirements are very challenging to the existing

segmentation models. Therefore, more effective, more efficient, and more stable models are needed.

1.4 Over-fitting *vs.* Regularization

Most deep segmentation models can fit the training sets very well (training accuracy close to 100%) while having different performances on the testing sets. To the best of our knowledge, there could be two main reasons. On one hand, the "distributions" between the training, validation, and testing sets are not guaranteed to be the same, which leads to performance degradation of almost all the models on testing sets. On the other hand, different model architectures have diversified capabilities of feature representations, which means they are more likely to fit the training sets in very different ways, namely, transforming the input images into other high-dimensional spaces. Most of the works are following this direction to develop more representative architectures. However, there lacks an effective way to measure the representation capabilities of these architectures before testing, so the model design is usually conducted by trial and error. Hence, some researchers turn to search for different ways for reducing over-fitting. Different supervision strategies, such as weights regularization [22], dropout [55], dense (deep) supervision [30, 50, 63], hybrid loss [40, 51, 74] and so on, have been proposed. The dense (deep) supervision [30,50,63], which imposes ground truth supervisions on the side outputs from several of the deep intermediate layers, is one of the most popular ways. However, transforming the deep intermediate features (multi-channel) into the side outputs (single-channel) in dichotomous image segmentation (DIS) is essentially a dimension reduction operation, which leads to information losses, so that weaken the supervisions. In this paper, instead of developing more complicated deep architectures, we follow the dense supervision idea but develop a simple yet more effective supervision strategy, **intermediate supervision**, to directly enforce the supervisions on high-dimensional intermediate deep features in addition to the side outputs.

1.5 Evaluation Metrics

The evaluation strategies and metrics are expected to provide comprehensive and practically meaningful evaluations to analyze the prediction qualities. Currently, many evaluation metrics, such as IoU, boundary IoU [7], F-measure [1], boundary F-measure [12, 51], boundary displacement error (BDE) [20], boundary IoU [7], structural measure (S_m) [15], Mean Absolute Error (M) [47], and so on, are usually defined based on consistencies (or inconsistencies) between the model predictions and the ground truth. Most of them are usually biased to certain types of structures. For example, IoU and F-measure mainly rely on the object components with large areas while neglecting the fine details with relatively small areas. To alleviate this issue, boundary F-measure, BDE, and boundary IoU are developed to focus on the boundary quality. However, these boundary-based metrics are often highly dependent on those long smooth boundary segments' qualities while failing to describe the qualities of those short jagged boundary

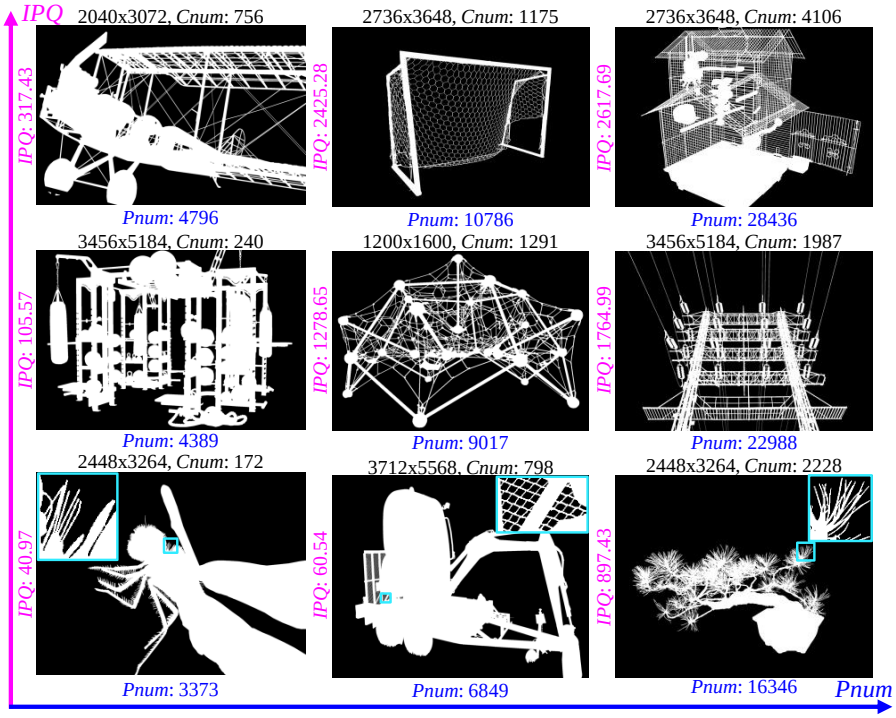


Fig. 2: GT masks of our DIS5K with diversified inter-categorical complexities. The complexity relationships are only valid within each row or column.

segments. Besides, the above metrics are mostly defined from the mathematical or cognitive perspective; none of them are able to reflect the barriers (or costs) of applying the predictions in real-world applications, where certain accuracy requirements have to be satisfied. To address these issues, we propose a novel metric, named as human correction efforts (HCE), to measure the barriers by approximating the human efforts for correcting the faulty regions of the model predictions.

2 More Details of DIS5K Dataset

2.1 Overall Complexity Analysis

The metrics used for evaluation the dataset complexities are all computed on the labeled GT masks and illustrated in Tab. 1 (in main manuscript) and Fig. 2-left (in main manuscript). It shows that DIS5K is around 20 (up to 50) times more complicated than the SOD datasets in terms of average IPQ. Although other datasets such as CHAMELEON, COD10K, BIG, COIFT, and ThinObject5K have higher average IPQ against the SOD datasets, they are still much

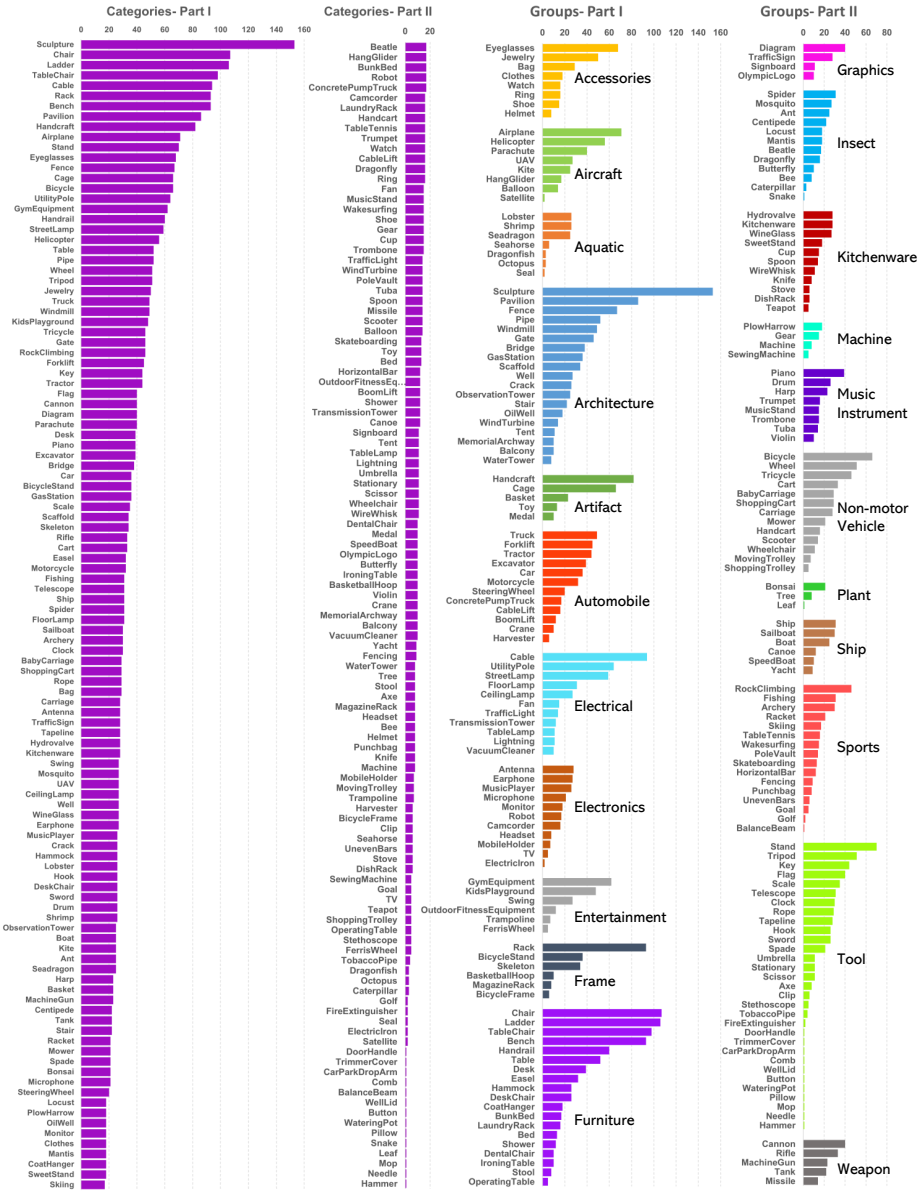


Fig. 3: Number of images per-category and per-group.

less complex than ours. The HR-SOD and HR-DAVIS-S datasets contain large-size images with accurately labeled boundaries. However, there are no significant differences between their IPQ and that of SOD datasets. Because IPQ is insensitive to the complexities of fine details as mentioned above. The average

contour-level complexities C_{num} of different datasets are almost consistent with their IPQ. The average C_{num} and its standard deviation of DIS5K are over 100 and 400, which are much higher than other datasets. This indicates the objects in DIS5K contain more detailed structures that are comprised of multiple contours. The average P_{num} of DIS5K is over 1400, which is almost five and three times greater than those of HR-SOD and the synthetic ThinObject5K, respectively. There is an interesting observation that the P_{num} of HR-SOD, HR-DAVIS-S, BIG, and ThinObject5K are not proportional to their IPQ and C_{num}, but it shows positive correlations with their image dimensions. One of the reasons is that most of the objects in these datasets are close to convex and comprised of single or a few contours, which leads to low IPQ and C_{num}. Nevertheless, their boundaries (*e.g.*, small jagged segments) are accurately labeled in high-resolution images that significantly increase the P_{num}. On the other hand, larger sizes of GT masks often directly lead to greater P_{num} because the dominant points are searched by [52], which filters out redundant boundary points based on their deviation distances (*epsilon*) against the straight lines constructed by their neighboring dominant points. For example, given two objects with the same shape comprised of smooth boundaries but different sizes, more dominant points are generated from the larger one with the same threshold of *epsilon*. That means P_{num} is determined by both the boundary complexity and the GT mask dimension. Therefore, these three complexity measurements are complementary to provide a comprehensive analysis of the object complexities.The large standard deviations in Tab. 1 (main manuscript) demonstrate the great diversities of DIS5K from different perspectives.

In Fig.2, we provide the sample masks with their complexity scores in DIS5K. The bottom-left samples with large regional components have relatively low IPQ, and the top-right samples with more thin and complicated fine structures have much higher IPQ and P_{num}.

2.2 Per-category and per-group statistics

Fig. 3 illustrates the number of images per-category and per-group. Our DIS5K contains 5,470 images from 225 categories divided into 22 groups. The average numbers of images per category and per group are around 24 and 249, respectively.

2.3 Typical Samples from DIS5K

Fig. 4 shows some samples from our DIS5K, which have certain characteristics similar to that of the existing dichotomous segmentation tasks, such as salient object detection (SOD) [58], salient object in clutter (SOC) [14], camouflaged object detection (COD) [17], thin object segmentation (TOS) [34], meticulous object segmentation (MOS) [65]. It is worth mentioning that "salient object", "salient object in clutter" and "camouflaged object" are mainly defined based on the contrast between foreground targets and background environments. In comparison, "thin object" and "meticulous object" are based on the geometric

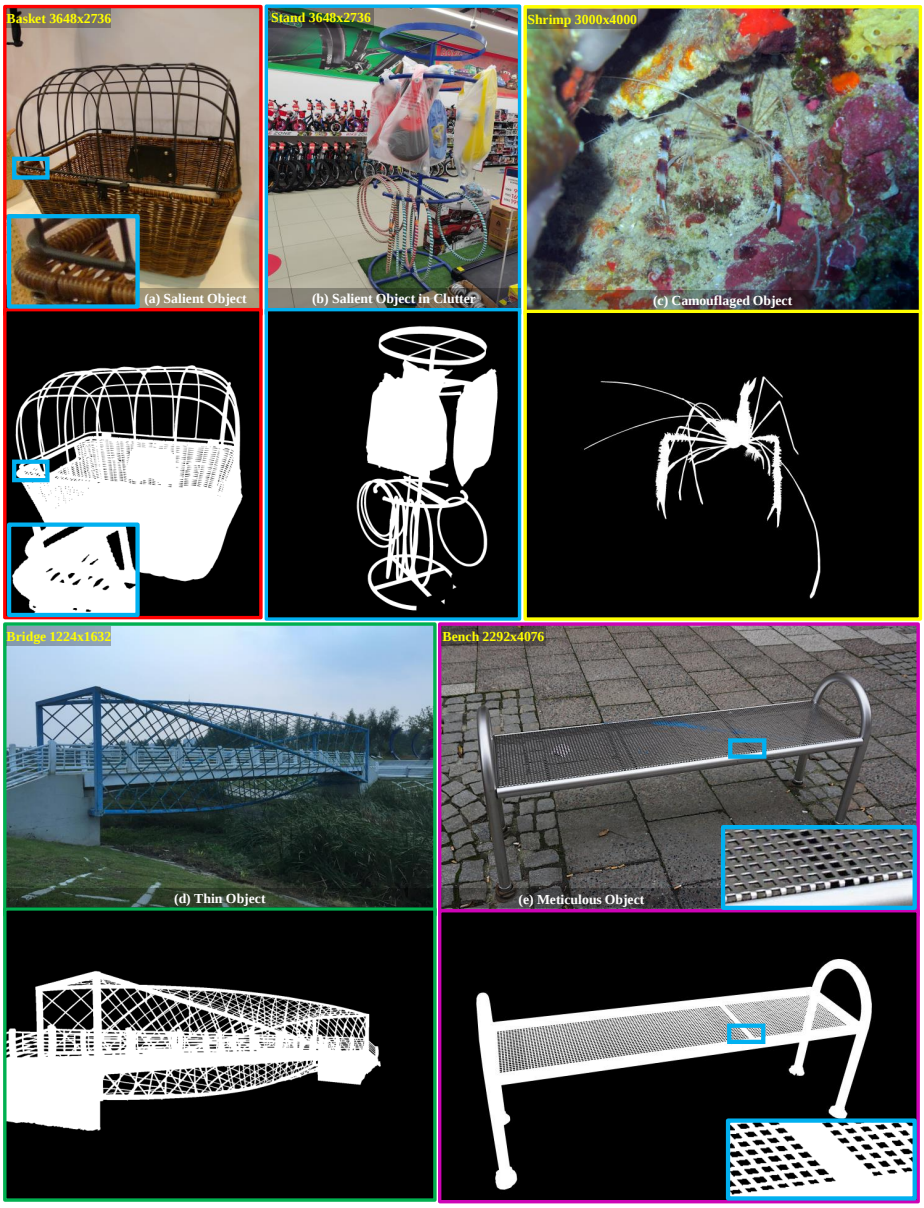


Fig. 4: Sample images and ground truth masks with objects of certain characteristics.

structure complexities of the foreground targets. Therefore, the first three types of objects and the last two types of targets are not exclusive. For example, the basket in Fig. 4 (a) and the shrimp in Fig. 4 (c) can also be taken as meticulous because the basket has many holes and the shrimp has jagged boundaries. Be-

sides, the boundaries among SOD, SOC, and COD and the boundaries between TOS and MOS are blurring. There are some overlaps between them in terms of data samples. Our DIS5K contains all the above types of images paired with highly-accurate ground truth masks.

2.4 Object Structure Analysis

In addition to the above mentioned image characteristics, there are also some interesting observations on object structures from our DIS5K, as shown in Fig. 5.
Intra-category structure similarity. As shown in Fig. 5 (a) and (b), the objects in the same categories are usually showing the same or similar structures and shapes. We call this *intra-category structure similarity*, which is one of the main cues for categorizing. However, the intra-category structure similarity is not always guaranteed. Fig. 5 (c) and (d) show two typical examples against that in different magnitudes. Fig. 5 (c) illustrates some bicycles with variant structures. Their differences are mainly caused by components absence (out-of-view imaging, incomplete architecture), variations on the design, view angle changes, co-existence of multiple targets, etc. Although the structures of these bicycles are different, they are still sharing some common features, such as wheels, frames, *etc.* However, objects in some other categories may share no structure similarities. For example, the sculptures in Fig. 5 (d) show very different structures and shapes, which indicates low intra-category similarity. Because artists or designers usually prefer to design unique architectures, which leads to very diversified object appearances and structures. Besides, compared against the relatively stable shapes and structures of the natural targets (*e.g.*, animals, plants), the structures of these human-created objects, which play vital roles in the human-environment interaction of our daily lives, are updated very fast, which further magnifies the challenges in the DIS task. These intra-category dissimilarities significantly increase the difficulty of accurate segmentation and lead to robustness risks.
Inter-category structure similarity. In contrary to the low intra-category similarity, there also exist some categories that have high *inter-category structure similarity*. Fig. 5 (e) shows some targets from different categories, such as *crack*, *lightning*, *cable*, *rope*, *pipe* and so on. These targets are mainly comprised of thin and elongated components. For example, the shapes of the crack and the lightning are very close to each other so that they are hard to be differentiated without showing the RGB images. The cable, rope, and pipe are also comprised of thin and elongated components with relatively smoother boundaries. Besides other targets like roads and rivers in satellite images, vessels in medical images also have similar structural characteristics to those mentioned above. The *inter-category structure similarities* haven't been thoroughly studied, which could be promising directions for exploring the models' explain-abilities and data augmentation strategies.

Our DIS5K dataset provides relatively richer samples for studying the *intra-category* and *inter-category* similarities and dissimilarities. More qualitative and

Fig. 5: Structure analysis of inter- and intra-category targets.

quantitative studies will be helpful to diversified vision tasks, such as image (shape) classification, segmentation, *etc.*

2.5 Attributes of Subsets in DIS5K

Tab. 1 illustrates the essential attributes of the subsets of our DIS5K dataset. As seen, the image dimensions of these subsets are close to each other. At the same time, the complexities of the four testing subsets are in ascending order. Fig. 6 shows the qualitative comparisons of the structural complexities of our four

Table 1: Image dimension and object complexity of the subsets of DIS5K. $\sigma_{(\cdot)}$ is the standard deviation of the corresponding index.

Task	Dataset	Number	Image Dimension			Object Complexity		
		I_{num}	$H \pm \sigma_H$	$W \pm \sigma_W$	$D \pm \sigma_D$	$IPQ \pm \sigma_{IPQ}$	$C_{num} \pm \sigma_C$	$P_{num} \pm \sigma_P$
	DIS5K	5470	2513.37 ± 1053.40	3111.44 ± 1359.51	4041.93 ± 1618.26	107.60 ± 320.69	106.84 ± 436.88	1427.82 ± 3326.72
DIS	DIS-TR	3000	2514.15 ± 1052.45	3091.23 ± 1356.92	4028.09 ± 1612.45	69.32 ± 261.98	73.99 ± 367.81	1153.05 ± 2893.36
	DIS-VD	470	2472.59 ± 963.43	3102.85 ± 1308.72	4006.49 ± 1526.56	156.85 ± 349.75	163.91 ± 650.42	1954.73 ± 5119.89
	DIS-TE1	500	2240.35 ± 1092.92	2678.50 ± 1291.11	3535.32 ± 1598.89	27.13 ± 29.07	6.94 ± 6.37	237.48 ± 96.27
	DIS-TE2	500	2402.09 ± 1047.89	3032.25 ± 1298.45	3904.03 ± 1583.39	50.79 ± 69.85	21.20 ± 16.30	583.04 ± 120.90
	DIS-TE3	500	2597.15 ± 988.88	3336.51 ± 1339.10	4263.78 ± 1571.21	92.68 ± 118.99	60.96 ± 40.32	1190.93 ± 255.00
	DIS-TE4	500	2847.55 ± 1069.37	3527.81 ± 1412.89	4580.93 ± 1645.86	443.32 ± 667.01	482.98 ± 843.50	4858.80 ± 5618.87

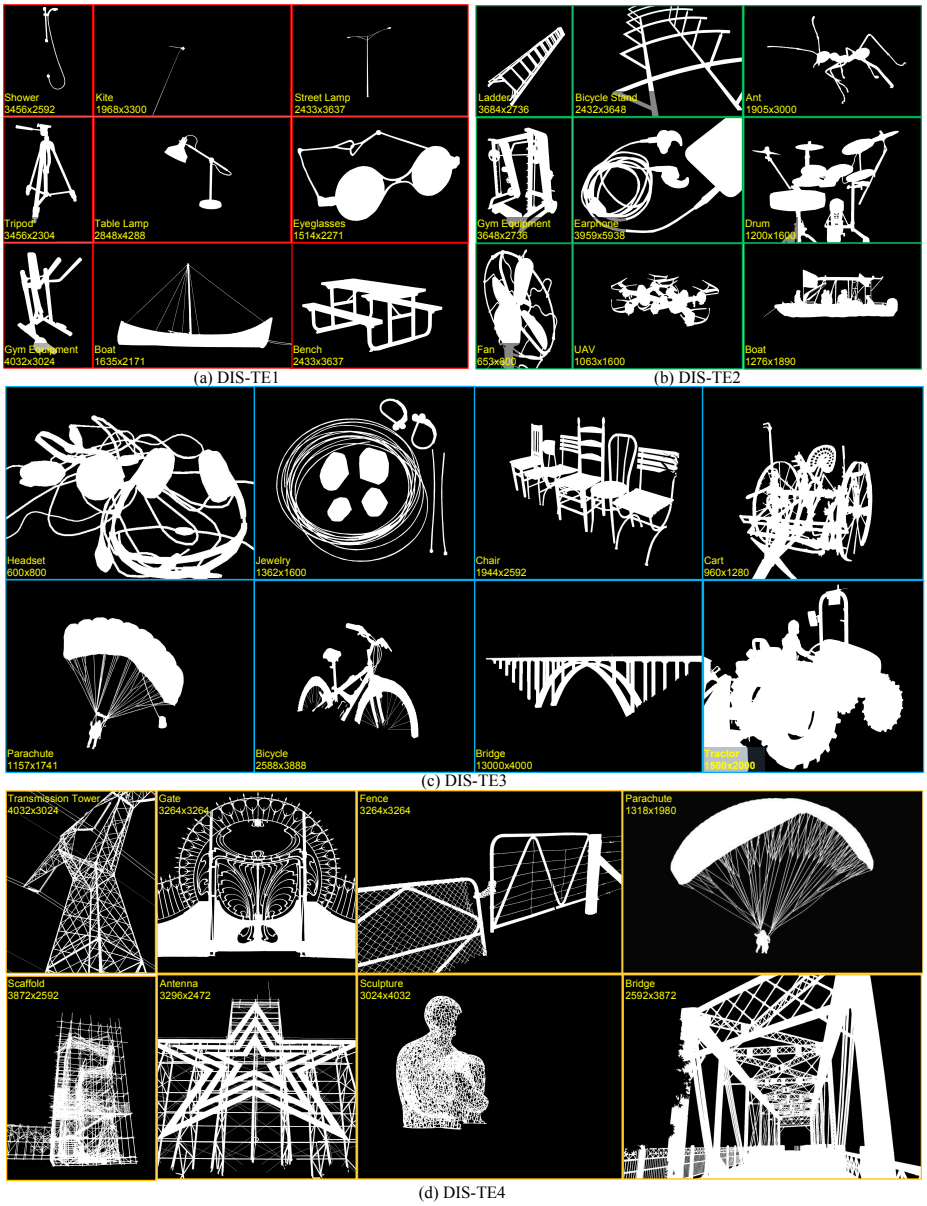
testing subsets, DIS-T1∼DIS-TE4. Their structure complexities in ascending order can be visually perceived.

3 More Details of Experiments

3.1 Implementation details

Our models and other baseline models are trained with our DIS-TR (3,000 images) and validated on DIS-VD (470 images). The input size of our model is set to 1024×1024. It is worth noting that there are many large-size images in our dataset so that the image loading operations in the training and validation are very time-consuming. To address this issue and boost the speed of training and validation, we resize all the input images and their corresponding ground truth to 1024×1024 off-line and store them as Pytorch tensor files on the hard disk drive. Although this strategy requires relatively more storage space, it dramatically reduces the time costs for the data loading process in the training and validation stages. Our training process consists of two training stages: (i) the training stage of the ground truth encoder and (ii) the training stage of the image segmentation component. In both training stages, these three-channel inputs (GT masks are repeated to have three channels) are normalized to [-0.5, 0.5] and only augmented with horizontal flipping. The models weights are initialized by Xavier [21] and optimized with Adam [28] optimizer with the default settings (initial learning rate lr=1e-3, betas=(0.9, 0.999), eps=1e-8, weight decay=0) for both the ground truth encoder and the segmentation component. The batch size of each training step is set to eight, and the validation on DIS-VD is conducted every 1,000 iterations. If the validation results (in terms of $maxF$ and M) are improved, the hard disk drive saves the model weights. It is worth mentioning that the loss weights of the dense supervision in the ground truth encoder training and intermediate supervision of the segmentation component training are all set to 1.0.

According to our experiments, the training process of our ground truth encoder is easy to converge, and it usually takes only 1,000 iterations (stop training when the valid $maxF$ is greater than 0.99). While the segmentation component of our model usually converges after around 100k iterations, and the whole training process takes less than 48 hours. Besides, all the models are implemented



(a) DIS-TE1 (b) DIS-TE2

(c) DIS-TE3

(d) DIS-TE4

Fig. 6: Sample ground truth (GT) masks from DIS-TE1, DIS-TE2, DIS-TE3, and DIS-TE4.

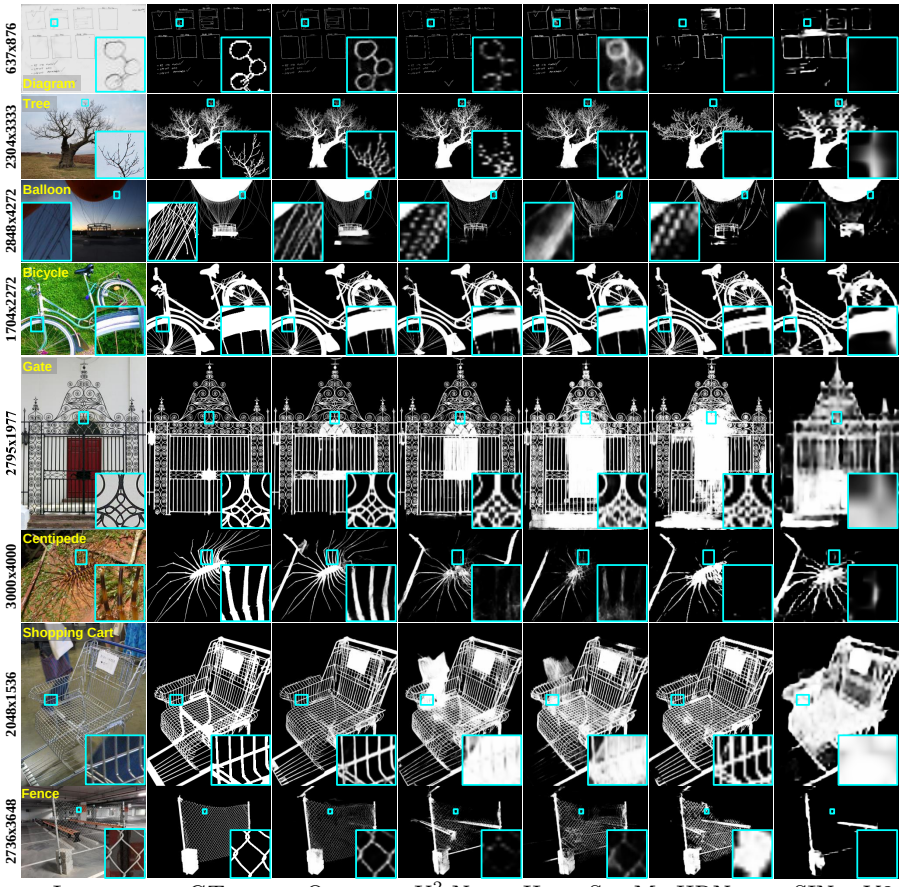


Fig. 7: Qualitative comparisons of our model and four cutting-edge baselines.

using Pytorch 1.8.0. Some experiments are conducted on a desktop that has a 2.9 GHz CPU (128 cores AMD Ryzen Threadripper 3990X), 256 GB RAM and a NVIDIA RTX A6000 GPU. Some other models are trained on NVIDIA TESLA V100 GPU (32 GB).

3.2 More Analysis of the Experimental Results

Performance comparisons among different models. As shown in Tab. 2 in the main manuscript, our model achieves the most competitive performance against other existing models in terms of almost all the evaluation metrics on different datasets. Among the dichotomous segmentation models, U-Net [53], BASNet [51], U^2-Net [50] and PFNet [43] performs relatively better against other SOD and COD models. Among the semantic segmentation and real-time semantic segmentation models, the results of HRNet [57] and HyperSeg-M [45] show more competitive performance. Among all the existing models, the performance

Table 2: PART-I: Quantitative evaluation on our validation, DIS-VD, and test sets, DIS-TE (1-4), based on groups. ResNet18=R-18. ResNet34=R-34. ResNet50=R-50. Res2Net50=R2-50. DeepLab-V3+=DLV3+. BiseNetV1=BSV1. STDC813=S-813. EffiNetB1=E-B1. MobileNetV3-Large=MBV3. HyperSeg-M=HySM.

Dataset	Metric	UNet [53]	BASNet [51]	GateNet [75]	F^3Net [62]	GCPANet [6]	U^2Net [50]	SINetV2 [16]	PFNet [43]	PSPNet [73]	DLV3+ [4]	HRNet [57]	BSV1 [67]	ICNet [72]	MBV3 [26]	STDC [18]	HySM [45]	Ours
1 Accessories	$maxF_\beta \uparrow$	0.680	0.735	0.677	0.700	0.664	0.749	0.684	0.703	0.701	0.659	0.733	0.655	0.681	0.723	0.714	0.749	0.788
	$F_\beta^w \uparrow$	0.576	0.641	0.572	0.608	0.560	0.658	0.606	0.619	0.614	0.565	0.652	0.535	0.590	0.651	0.631	0.657	0.716
	$M \downarrow$	0.133	0.109	0.130	0.121	0.135	0.110	0.124	0.117	0.116	0.131	0.106	0.144	0.123	0.108	0.115	0.106	0.093
	$S_\alpha \uparrow$	0.714	0.746	0.700	0.721	0.706	0.757	0.720	0.730	0.725	0.694	0.755	0.698	0.711	0.742	0.734	0.767	0.788
	$E_\phi^m \uparrow$	0.761	0.806	0.770	0.800	0.759	0.804	0.794	0.810	0.800	0.777	0.818	0.738	0.786	0.829	0.814	0.809	0.837
	$HCE_\gamma \downarrow$	547	549	571	612	682	562	679	634	662	580	581	688	585	684	630	547	432
2 Aircraft	$maxF_\beta \uparrow$	0.823	0.846	0.788	0.800	0.781	0.847	0.804	0.811	0.816	0.788	0.831	0.798	0.814	0.825	0.791	0.835	0.886
	$F_\beta^w \uparrow$	0.732	0.756	0.683	0.715	0.667	0.757	0.717	0.729	0.722	0.691	0.746	0.686	0.727	0.757	0.712	0.750	0.821
	$M \downarrow$	0.068	0.063	0.079	0.069	0.075	0.064	0.064	0.067	0.065	0.076	0.062	0.075	0.068	0.056	0.070	0.062	0.047
	$S_\alpha \uparrow$	0.829	0.828	0.779	0.803	0.791	0.830	0.807	0.810	0.814	0.791	0.830	0.810	0.817	0.827	0.800	0.835	0.872
	$E_\phi^m \uparrow$	0.875	0.875	0.828	0.865	0.840	0.871	0.884	0.879	0.869	0.851	0.890	0.851	0.873	0.906	0.874	0.875	0.911
	$HCE_\gamma \downarrow$	1185	1248	1153	1258	1222	1242	1241	1243	1229	1190	1448	1296	1159	1315	1314	1123	1066
3 Aquatic	$maxF_\beta \uparrow$	0.612	0.681	0.613	0.613	0.581	0.691	0.571	0.649	0.615	0.601	0.700	0.563	0.617	0.672	0.604	0.654	0.715
	$F_\beta^w \uparrow$	0.489	0.576	0.481	0.510	0.464	0.581	0.481	0.550	0.511	0.492	0.603	0.424	0.519	0.591	0.505	0.542	0.624
	$M \downarrow$	0.119	0.093	0.109	0.107	0.124	0.090	0.124	0.099	0.103	0.113	0.085	0.119	0.103	0.085	0.104	0.106	0.080
	$S_\alpha \uparrow$	0.692	0.728	0.665	0.687	0.670	0.738	0.676	0.716	0.692	0.673	0.748	0.658	0.695	0.729	0.681	0.713	0.759
	$E_\phi^m \uparrow$	0.732	0.779	0.732	0.743	0.704	0.786	0.735	0.796	0.755	0.758	0.832	0.678	0.781	0.822	0.735	0.747	0.799
	$HCE_\gamma \downarrow$	879	867	867	905	916	872	945	937	988	906	926	984	899	1009	938	839	710
4 Architecture	$maxF_\beta \uparrow$	0.720	0.742	0.678	0.685	0.638	0.751	0.671	0.702	0.694	0.674	0.739	0.681	0.710	0.706	0.704	0.756	0.792
	$F_\beta^w \uparrow$	0.610	0.649	0.570	0.595	0.528	0.657	0.587	0.612	0.601	0.576	0.649	0.563	0.621	0.633	0.622	0.661	0.713
	$M \downarrow$	0.099	0.087	0.106	0.101	0.115	0.084	0.105	0.100	0.097	0.106	0.087	0.103	0.095	0.091	0.093	0.084	0.070
	$S_\alpha \uparrow$	0.769	0.779	0.725	0.741	0.716	0.790	0.739	0.752	0.751	0.729	0.780	0.747	0.761	0.759	0.756	0.794	0.814
	$E_\phi^m \uparrow$	0.803	0.828	0.779	0.806	0.759	0.828	0.813	0.824	0.808	0.803	0.841	0.781	0.821	0.842	0.829	0.835	0.849
	$HCE_\gamma \downarrow$	1949	2180	2263	2368	2322	2217	2362	2418	2409	2331	2342	2525	2329	2413	2424	2053	1746
5 Artifact	$maxF_\beta \uparrow$	0.721	0.736	0.687	0.678	0.640	0.767	0.648	0.696	0.702	0.664	0.741	0.658	0.713	0.717	0.693	0.750	0.805
	$F_\beta^w \uparrow$	0.622	0.657	0.594	0.598	0.543	0.683	0.575	0.621	0.619	0.578	0.666	0.543	0.630	0.647	0.618	0.670	0.733
	$M \downarrow$	0.125	0.107	0.125	0.128	0.147	0.100	0.141	0.125	0.117	0.134	0.107	0.144	0.118	0.114	0.122	0.107	0.080
	$S_\alpha \uparrow$	0.758	0.770	0.725	0.727	0.708	0.794	0.712	0.744	0.747	0.713	0.777	0.718	0.751	0.757	0.735	0.784	0.822
	$E_\phi^m \uparrow$	0.795	0.833	0.797	0.795	0.755	0.834	0.781	0.812	0.806	0.792	0.834	0.748	0.815	0.831	0.809	0.824	0.854
	$HCE_\gamma \downarrow$	2126	2248	2572	2607	2508	2326	2454	2601	2647	2534	2494	2789	2517	2554	2613	2223	1821
6 Automobile	$maxF_\beta \uparrow$	0.773	0.816	0.781	0.787	0.765	0.825	0.789	0.794	0.790	0.761	0.801	0.756	0.796	0.809	0.789	0.824	0.844
	$F_\beta^w \uparrow$	0.683	0.741	0.687	0.708	0.676	0.752	0.715	0.719	0.718	0.680	0.734	0.659	0.717	0.748	0.715	0.745	0.785
	$M \downarrow$	0.113	0.088	0.109	0.100	0.109	0.084	0.097	0.098	0.096	0.111	0.092	0.118	0.096	0.083	0.098	0.084	0.076
	$S_\alpha \uparrow$	0.780	0.813	0.770	0.786	0.776	0.822	0.794	0.792	0.792	0.765	0.808	0.776	0.795	0.806	0.786	0.823	0.836
	$E_\phi^m \uparrow$	0.824	0.865	0.832	0.850	0.829	0.868	0.858	0.862	0.857	0.842	0.861	0.820	0.860	0.879	0.859	0.868	0.881
	$HCE_\gamma \downarrow$	860	896	955	994	1026	911	1037	1016	1043	959	967	1102	974	1056	1006	860	703
7 Electrical	$maxF_\beta \uparrow$	0.625	0.716	0.656	0.653	0.584	0.731	0.625	0.638	0.638	0.610	0.691	0.593	0.662	0.658	0.653	0.700	0.778
	$F_\beta^w \uparrow$	0.512	0.614	0.551	0.554	0.469	0.626	0.529	0.538	0.543	0.512	0.592	0.472	0.562	0.578	0.561	0.598	0.700
	$M \downarrow$	0.091	0.065	0.074	0.073	0.089	0.064	0.081	0.082	0.076	0.083	0.069	0.090	0.074	0.072	0.074	0.070	0.053
	$S_\alpha \uparrow$	0.730	0.771	0.728	0.732	0.701	0.782	0.715	0.722	0.728	0.709	0.760	0.706	0.742	0.737	0.731	0.769	0.808
	$E_\phi^m \uparrow$	0.766	0.830	0.804	0.819	0.750	0.826	0.804	0.808	0.800	0.804	0.826	0.758	0.822	0.838	0.826	0.816	0.853
	$HCE_\gamma \downarrow$	1104	1368	1333	1398	1335	1380	1358	1428	1409	1376	1501	1501	1336	1435	1421	1149	911
8 Electronics	$maxF_\beta \uparrow$	0.721	0.740	0.688	0.718	0.658	0.769	0.712	0.714	0.715	0.665	0.733	0.682	0.723	0.723	0.712	0.760	0.801
	$F_\beta^w \uparrow$	0.629	0.660	0.592	0.637	0.563	0.692	0.638	0.637	0.634	0.577	0.658	0.572	0.642	0.665	0.636	0.678	0.744
	$M \downarrow$	0.094	0.089	0.106	0.098	0.112	0.080	0.091	0.096	0.092	0.108	0.087	0.108	0.089	0.086	0.092	0.084	0.063
	$S_\alpha \uparrow$	0.780	0.780	0.737	0.766	0.739	0.808	0.769	0.766	0.769	0.730	0.784	0.752	0.771	0.783	0.764	0.805	0.834
	$E_\phi^m \uparrow$	0.808	0.819	0.782	0.816	0.774	0.841	0.826	0.820	0.812	0.793	0.826	0.781	0.816	0.834	0.823	0.832	0.872
	$HCE_\gamma \downarrow$	804	857	842	924	953	861	965	947	985	902	956	1019	868	995	958	781	622
9 Entertainment	$maxF_\beta \uparrow$	0.747	0.784	0.718	0.716	0.654	0.774	0.704	0.738	0.722	0.699	0.768	0.727	0.746	0.746	0.730	0.791	0.831
	$F_\beta^w \uparrow$	0.628	0.681	0.603	0.615	0.532	0.671	0.605	0.639	0.615	0.592	0.671	0.600	0.648	0.663	0.640	0.688	0.748
	$M \downarrow$	0.110	0.093	0.111	0.111	0.126	0.095	0.110	0.105	0.106	0.117	0.094	0.112	0.100	0.097	0.103	0.093	0.071
	$S_\alpha \uparrow$	0.768	0.786	0.737	0.743	0.713	0.783	0.742	0.761	0.745	0.729	0.781	0.759	0.769	0.767	0.754	0.799	0.827
	$E_\phi^m \uparrow$	0.802	0.839	0.798	0.821	0.760	0.834	0.830	0.837	0.816	0.814	0.850	0.801	0.836	0.852	0.840	0.838	0.872
	$HCE_\gamma \downarrow$	1644	1793	1837	1862	1834	1872	1849	1904	1907	1838	1969	2029	1819	1920	1870	1643	1369
10 Frame	$maxF_\beta \uparrow$	0.681	0.718	0.678	0.651	0.596	0.742	0.629	0.671	0.680	0.638	0.687	0.643	0.675	0.696	0.695	0.724	0.783
	$F_\beta^w \uparrow$	0.564	0.625	0.573	0.561	0.482	0.639	0.543	0.576	0.586	0.547	0.597	0.513	0.581	0.621	0.610	0.619	0.702
	$M \downarrow$	0.097	0.080	0.086	0.093	0.113	0.075	0.104	0.093	0.088	0.097	0.087	0.104	0.087	0.082	0.083	0.082	0.064
	$S_\alpha \uparrow$	0.757	0.787	0.750	0.745	0.717	0.800	0.732	0.754	0.758	0.735	0.767	0.735	0.758	0.761	0.759	0.786	0.826
	$E_\phi^m \uparrow$	0.791	0.832	0.818	0.810	0.752	0.843	0.796	0.819	0.821	0.812	0.819	0.777	0.827	0.845	0.842	0.824	0.863
	$HCE_\gamma \downarrow$	1066	1169	1248	1317	1311	1187	1318	1354	1380	1266	1294	1425	1258	1371	1318	1122	850
11 Furniture	$maxF_\beta \uparrow$	0.655	0.721	0.662	0.670	0.629	0.725	0.664	0.680	0.675	0.644	0.706	0.623	0.670	0.702	0.680	0.718	0.773
	$F_\beta^w \uparrow$	0.549	0.636	0.558	0.580	0.525	0.636	0.583	0.593	0.586	0.553	0.622	0.506	0.583	0.629	0.597	0.626	0.695
	$M \downarrow$	0.119	0.090	0.109	0.106	0.121	0.089	0.109	0.103	0.103	0.111	0.095	0.126	0.103	0.090	0.102	0.095	0.076
	$S_\alpha \uparrow$	0.725	0.768	0.715	0.730	0.711	0.773	0.733	0.741	0.736	0.710	0.761	0.705	0.734	0.754	0.736	0.768	0.804
	$E_\phi^m \uparrow$	0.764	0.822	0.787	0.796	0.761	0.819	0.803	0.805	0.799	0.794	0.813	0.750	0.803	0.834	0.811	0.811	0.842
	$HCE_\gamma \downarrow$	871	904	951	1001	1012	914	1012	1035	1044	959	1018	1120	978	1048	1020	862	671

of HyperSeg-M and U^2-Net are close and perform better than other models in both validation and testing sets. Although HRNet and BASNet show slightly inferior performance against HyperSeg-M and U^2-Net, they are still more competitive than others. Fig. 7 provides the qualitative comparisons of our model and other four competitive baseline models. As can be seen, our model achieves the best overall performance on different objects. Surprisingly, other models like U^2-Net, HyperSeg-M, and HRNet also obtain encouraging results on certain targets, such as the *tree*, the *gate* and the *shopping cart*, after training on our DIS-TR dataset, which further proves the value of DIS5K.

Table 3: PART-II: Quantitative evaluation on our validation, DIS-VD, and test sets, DIS-TE (1-4), based on groups. ResNet18=R-18. ResNet34=R-34. ResNet50=R-50. Res2Net50=R2-50. DeepLab-V3+=DLV3+. BiseNetV1=BSV1. STDC813=S-813. EffiNetB1=E-B1. MobileNetV3-Large=MBV3. HyperSeg-M=HySM.

Dataset	Metric	UNet [53]	BASNet [51]	GateNet [75]	F^3Net [62]	GCPANet [6]	U^2Net [50]	SINetV2 [16]	PFNet [43]	PSPNet [73]	DLV3+ [4]	HRNet [57]	BSV1 [67]	ICNet [72]	MBV3 [26]	STDC [18]	HySM [45]	Ours
12 Graphics	$maxF_\beta \uparrow$	0.750	0.719	0.685	0.663	0.524	0.746	0.568	0.645	0.646	0.621	0.671	0.575	0.681	0.616	0.647	0.732	0.780
	$F_\beta^w \uparrow$	0.654	0.628	0.598	0.584	0.431	0.653	0.496	0.569	0.566	0.540	0.585	0.473	0.606	0.566	0.578	0.647	0.706
	$M \downarrow$	0.061	0.064	0.066	0.069	0.094	0.057	0.088	0.078	0.067	0.073	0.074	0.096	0.064	0.065	0.068	0.059	0.049
	$S_\alpha \uparrow$	0.825	0.800	0.784	0.772	0.703	0.823	0.717	0.754	0.772	0.752	0.772	0.719	0.790	0.750	0.763	0.814	0.839
	$E_\phi^m \uparrow$	0.835	0.831	0.835	0.843	0.726	0.834	0.795	0.827	0.798	0.817	0.819	0.740	0.828	0.847	0.865	0.836	0.873
	$HCE_\gamma \downarrow$	670	976	1009	1268	1403	938	1423	1294	1447	1201	990	1425	1122	1457	1331	824	621
13 Insect	$maxF_\beta \uparrow$	0.673	0.681	0.641	0.627	0.554	0.718	0.608	0.634	0.637	0.617	0.706	0.620	0.650	0.700	0.643	0.692	0.762
	$F_\beta^w \uparrow$	0.552	0.586	0.530	0.537	0.442	0.617	0.523	0.541	0.541	0.522	0.617	0.482	0.552	0.629	0.557	0.592	0.683
	$M \downarrow$	0.073	0.065	0.071	0.070	0.089	0.058	0.076	0.075	0.069	0.074	0.061	0.075	0.068	0.058	0.069	0.062	0.049
	$S_\alpha \uparrow$	0.766	0.766	0.733	0.738	0.694	0.786	0.724	0.737	0.740	0.728	0.785	0.725	0.747	0.776	0.743	0.783	0.820
	$E_\phi^m \uparrow$	0.804	0.821	0.781	0.804	0.753	0.827	0.810	0.803	0.817	0.817	0.844	0.748	0.825	0.863	0.820	0.809	0.860
	$HCE_\gamma \downarrow$	570	595	604	656	683	592	701	663	714	636	622	700	609	713	667	574	488
14 Kitchenware	$maxF_\beta \uparrow$	0.704	0.754	0.678	0.697	0.688	0.734	0.713	0.708	0.692	0.661	0.739	0.667	0.689	0.730	0.702	0.749	0.771
	$F_\beta^w \uparrow$	0.588	0.654	0.555	0.596	0.578	0.633	0.620	0.608	0.587	0.550	0.649	0.545	0.587	0.647	0.606	0.657	0.685
	$M \downarrow$	0.167	0.144	0.174	0.163	0.170	0.151	0.152	0.160	0.167	0.178	0.143	0.178	0.166	0.144	0.159	0.140	0.128
	$S_\alpha \uparrow$	0.704	0.733	0.662	0.690	0.691	0.723	0.712	0.698	0.680	0.653	0.729	0.679	0.688	0.729	0.697	0.743	0.763
	$E_\phi^m \uparrow$	0.737	0.777	0.721	0.754	0.742	0.761	0.777	0.764	0.736	0.731	0.798	0.725	0.753	0.795	0.764	0.786	0.798
	$HCE_\gamma \downarrow$	541	536	554	574	579	536	602	583	588	543	608	637	540	608	571	484	367
15 Machine	$maxF_\beta \uparrow$	0.798	0.807	0.744	0.777	0.746	0.845	0.778	0.767	0.800	0.766	0.842	0.755	0.812	0.812	0.782	0.818	0.869
	$F_\beta^w \uparrow$	0.692	0.713	0.629	0.676	0.638	0.755	0.695	0.676	0.710	0.666	0.760	0.639	0.722	0.738	0.694	0.727	0.801
	$M \downarrow$	0.126	0.119	0.147	0.131	0.145	0.100	0.124	0.131	0.118	0.138	0.100	0.147	0.116	0.111	0.123	0.116	0.089
	$S_\alpha \uparrow$	0.764	0.771	0.701	0.739	0.728	0.809	0.761	0.736	0.770	0.729	0.802	0.739	0.773	0.780	0.747	0.783	0.842
	$E_\phi^m \uparrow$	0.812	0.833	0.781	0.816	0.786	0.851	0.844	0.824	0.843	0.821	0.870	0.779	0.848	0.857	0.835	0.835	0.881
	$HCE_\gamma \downarrow$	1544	1687	1728	1846	1849	1693	1910	1860	1925	1787	1937	1987	1799	1957	1899	1589	1322
16 Music Instrument	$maxF_\beta \uparrow$	0.748	0.809	0.740	0.777	0.756	0.817	0.775	0.777	0.777	0.752	0.808	0.748	0.774	0.811	0.777	0.829	0.852
	$F_\beta^w \uparrow$	0.643	0.726	0.636	0.691	0.660	0.734	0.699	0.698	0.690	0.656	0.730	0.640	0.689	0.739	0.698	0.745	0.783
	$M \downarrow$	0.159	0.123	0.163	0.137	0.145	0.115	0.127	0.133	0.139	0.154	0.117	0.156	0.140	0.113	0.135	0.114	0.101
	$S_\alpha \uparrow$	0.732	0.781	0.706	0.753	0.749	0.790	0.767	0.761	0.749	0.722	0.787	0.736	0.750	0.782	0.755	0.799	0.820
	$E_\phi^m \uparrow$	0.775	0.825	0.764	0.811	0.792	0.834	0.826	0.818	0.809	0.796	0.842	0.771	0.809	0.848	0.814	0.828	0.853
	$HCE_\gamma \downarrow$	671	683	653	693	708	705	735	713	732	678	791	796	687	771	748	598	492
17 Non-motor Vehicle	$maxF_\beta \uparrow$	0.762	0.800	0.755	0.761	0.718	0.803	0.740	0.755	0.774	0.748	0.791	0.731	0.764	0.779	0.768	0.794	0.840
	$F_\beta^w \uparrow$	0.662	0.719	0.658	0.674	0.612	0.722	0.654	0.673	0.687	0.660	0.713	0.620	0.683	0.709	0.691	0.710	0.774
	$M \downarrow$	0.100	0.086	0.103	0.100	0.118	0.086	0.107	0.101	0.095	0.101	0.086	0.113	0.095	0.087	0.093	0.088	0.068
	$S_\alpha \uparrow$	0.788	0.816	0.767	0.784	0.759	0.817	0.770	0.781	0.791	0.769	0.812	0.768	0.790	0.800	0.787	0.815	0.846
	$E_\phi^m \uparrow$	0.839	0.870	0.836	0.853	0.807	0.866	0.845	0.852	0.857	0.852	0.870	0.811	0.857	0.874	0.863	0.859	0.891
	$HCE_\gamma \downarrow$	1956	2098	2134	2219	2217	2121	2269	2293	2274	2169	2314	2319	2161	2334	2245	1971	1623
18 Plant	$maxF_\beta \uparrow$	0.685	0.745	0.690	0.685	0.680	0.771	0.696	0.701	0.723	0.703	0.755	0.642	0.718	0.743	0.706	0.785	0.766
	$F_\beta^w \uparrow$	0.566	0.637	0.569	0.576	0.564	0.665	0.589	0.602	0.623	0.595	0.658	0.500	0.621	0.654	0.597	0.689	0.665
	$M \downarrow$	0.144	0.119	0.138	0.138	0.145	0.111	0.141	0.134	0.126	0.131	0.111	0.153	0.125	0.112	0.136	0.104	0.109
	$S_\alpha \uparrow$	0.697	0.730	0.689	0.695	0.685	0.761	0.703	0.696	0.727	0.700	0.752	0.662	0.720	0.737	0.693	0.779	0.764
	$E_\phi^m \uparrow$	0.749	0.778	0.749	0.755	0.748	0.787	0.758	0.774	0.790	0.783	0.810	0.707	0.801	0.804	0.762	0.804	0.779
	$HCE_\gamma \downarrow$	9194	9174	10036	10164	10488	9062	10268	10137	10231	9910	9615	10444	9798	10309	10230	8334	8563
19 Ship	$maxF_\beta \uparrow$	0.773	0.793	0.739	0.747	0.726	0.792	0.730	0.760	0.769	0.756	0.779	0.761	0.772	0.785	0.744	0.791	0.834
	$F_\beta^w \uparrow$	0.686	0.705	0.632	0.660	0.614	0.713	0.648	0.672	0.676	0.657	0.698	0.653	0.690	0.711	0.659	0.711	0.766
	$M \downarrow$	0.095	0.095	0.114	0.107	0.116	0.089	0.108	0.103	0.103	0.107	0.098	0.104	0.098	0.085	0.108	0.091	0.069
	$S_\alpha \uparrow$	0.796	0.796	0.742	0.760	0.741	0.804	0.753	0.770	0.775	0.758	0.784	0.772	0.787	0.790	0.757	0.806	0.840
	$E_\phi^m \uparrow$	0.840	0.842	0.793	0.823	0.785	0.849	0.838	0.837	0.828	0.826	0.846	0.811	0.846	0.870	0.831	0.848	0.880
	$HCE_\gamma \downarrow$	3193	3341	3233	3242	3225	3355	3183	3265	3189	3178	3443	3454	3134	3381	3334	3046	2951
20 Sports	$maxF_\beta \uparrow$	0.699	0.721	0.674	0.675	0.637	0.745	0.661	0.687	0.685	0.639	0.724	0.679	0.676	0.727	0.684	0.744	0.788
	$F_\beta^w \uparrow$	0.596	0.629	0.572	0.583	0.526	0.651	0.573	0.590	0.594	0.547	0.637	0.554	0.583	0.654	0.597	0.647	0.714
	$M \downarrow$	0.076	0.065	0.074	0.074	0.081	0.059	0.077	0.075	0.072	0.081	0.064	0.078	0.072	0.059	0.072	0.062	0.051
	$S_\alpha \uparrow$	0.766	0.778	0.743	0.747	0.728	0.797	0.740	0.748	0.748	0.724	0.784	0.751	0.752	0.780	0.750	0.795	0.827
	$E_\phi^m \uparrow$	0.807	0.822	0.805	0.825	0.777	0.832	0.821	0.825	0.820	0.803	0.831	0.801	0.816	0.868	0.836	0.838	0.860
	$HCE_\gamma \downarrow$	1137	1283	1274	1329	1247	1315	1274	1355	1323	1297	1450	1401	1306	1352	1343	1180	934
21 Tool	$maxF_\beta \uparrow$	0.656	0.714	0.649	0.678	0.643	0.719	0.670	0.683	0.679	0.628	0.700	0.628	0.670	0.697	0.680	0.717	0.757
	$F_\beta^w \uparrow$	0.538	0.622	0.543	0.582	0.533	0.624	0.581	0.589	0.588	0.532	0.612	0.505	0.573	0.623	0.592	0.611	0.676
	$M \downarrow$	0.100	0.082	0.095	0.089	0.100	0.080	0.094	0.090	0.086	0.101	0.086	0.104	0.091	0.081	0.087	0.082	0.071
	$S_\alpha \uparrow$	0.733	0.771	0.721	0.743	0.727	0.773	0.739	0.746	0.752	0.708	0.759	0.719	0.740	0.758	0.739	0.771	0.797
	$E_\phi^m \uparrow$	0.784	0.829	0.797	0.822	0.785	0.831	0.815	0.822	0.820	0.802	0.827	0.769	0.815	0.842	0.836	0.823	0.844
	$HCE_\gamma \downarrow$	568	589	620	659	673	602	689	673	689	632	662	724	625	707	660	554	433
22 Weapon	$maxF_\beta \uparrow$	0.763	0.805	0.728	0.787	0.765	0.816	0.780	0.799	0.798	0.747	0.812	0.757	0.773	0.794	0.785	0.806	0.848
	$F_\beta^w \uparrow$	0.672	0.726	0.616	0.706	0.668	0.737	0.706	0.717	0.718	0.654	0.743	0.657	0.689	0.728	0.707	0.730	0.794
	$M \downarrow$	0.108	0.090	0.124	0.097	0.106	0.087	0.096	0.093	0.091	0.113	0.084	0.108	0.099	0.088	0.096	0.085	0.071
	$S_\alpha \uparrow$	0.784	0.802	0.718	0.788	0.775	0.814	0.790	0.797	0.794	0.750	0.816	0.778	0.776	0.802	0.784	0.820	0.850
	$E_\phi^m \uparrow$	0.822	0.861	0.790	0.843	0.833	0.855	0.857	0.861	0.856	0.828	0.870	0.826	0.838	0.864	0.854	0.867	0.899
	$HCE_\gamma \downarrow$	793	826	849	888	899	845	923	902	928	864	914	969	861	937	899	779	649
All VD+TE(1-4)	$maxF_\beta \uparrow$	0.705	0.748	0.691	0.700	0.658	0.758	0.687	0.708	0.706	0.675	0.739	0.671	0.709	0.726	0.708	0.752	0.798
	$F_\beta^w \uparrow$	0.600	0.659	0.587	0.611	0.551	0.668	0.604	0.620	0.617	0.581	0.654	0.556	0.620	0.655	0.625	0.660	0.724
	$M \downarrow$	0.105	0.087	0.104	0.099	0.113	0.085	0.102	0.099	0.096	0.107	0.088	0.112	0.096	0.087	0.095	0.087	0.071
	$S_\alpha \uparrow$	0.756	0.780	0.730	0.747	0.726	0.789	0.743	0.753	0.753	0.727	0.778	0.735	0.756	0.768	0.751	0.788	0.818
	$E_\phi^m \uparrow$	0.796	0.832	0.794	0.815	0.774	0.833	0.817	0.824	0.816	0.807	0.837	0.776	0.822	0.849	0.829	0.830	0.857
	$HCE_\gamma \downarrow$	1228	1330	1368	1433	1425	1348	1441	1461	1470	1394	1457	1541	1387	1489	1459	1239	1035

Performance comparisons among different test sets. performance analysis based on the targets' complexities for demonstrating the importance of our newly proposed $HCE_\gamma \downarrow$ metric. As shown in Tab. 2 in the main manuscript, our model achieves different performances on the four testing sets, obtained by ordering (ascending) and splitting the whole test set according to the structural complexities of the to-be-segmented objects. However, except for our newly proposed $HCE_\gamma \downarrow$, other metrics, such as $maxF_\beta \uparrow$, $F_\beta^w \uparrow$, $M \downarrow$, $S_\alpha \uparrow$ and $E_\phi^m \uparrow$, of DIS-TE1, DIS-TE2, DIS-TE3, and DIS-TE4 show no strong (negative or positive) correlations with respect to the shape complexities. For example, M of

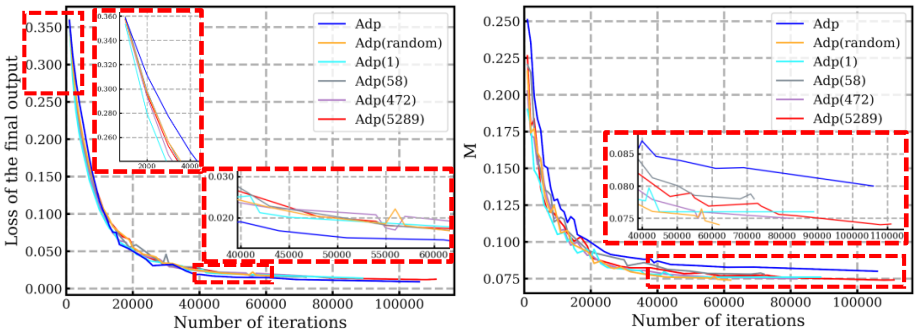


Fig. 8: Curves of the training loss computed on the last prediction probability map and the Mean Absolute Error (M) on our validation set (DIS-VD).

our model on these DIS-TE1 (0.074) and DIS-TE4 (0.072) are very close. The $maxF_\beta \uparrow$, $F_\beta^w \uparrow$, $S_\alpha \uparrow$ and $E_\phi^m \uparrow$ of DIS-TE4 are even greater than those of DIS-TE1, which probably provides misleading information that DIS-TE4 is less challenging than DIS-TE1. On the contrary, the $HCE_\gamma \downarrow$ of our model on DIS-TE1 and DIS-TE4 are 149 and 2,888, respectively. That indicates the cost for correcting the predictions of DIS-TE4 is around 20 times more than that of correcting predictions on DIS-TE1, which is consistent with the complexities illustrated in Tab. 1. It means our $HCE_\gamma \downarrow$ can correctly describe the correlations between prediction quality and the shape complexities. Thus, it can assess the human interventions needed when applying the models to real-world applications. We can get similar observations from the evaluation scores of other models on different test sets, which further proves the importance of our $HCE_\gamma \downarrow$ in evaluating highly accurate dichotomous image segmentation results. It is worth noting that the weak correlations between the conventional metrics and the shape complexities of different test sets are partial because image context complexity also plays a vital role in determining the segmentation difficulties. But this factor is hard to be quantified and has relatively less impact on the labeling workloads. Therefore, it is not considered in this work and will be studied in the future. In addition, performance comparisons of different models based on different groups are illustrated in Tab. 2 and 3, from which the per-group segmentation difficulties and performance can be found.

Effectiveness of Our Intermediate Supervision To further demonstrate the effectiveness of our intermediate supervision, we show the training loss and validation mean absolute error $M \downarrow$ curves of our adapted U^2-Net with and without our intermediate supervisions in Fig.8. The top part of Fig.8 shows the training loss of the last side output, which is taken as the final result in the inference stage. As can be seen, the models with intermediate supervisions converge faster before around 10,000 iterations. Later, the model without intermediate supervisions gradually produces a lower loss. These curves demonstrate that our intermediate supervision plays a typical role of regularizer for reducing the probability of over-fitting. The bottom plot of Fig.8 shows that our interme-

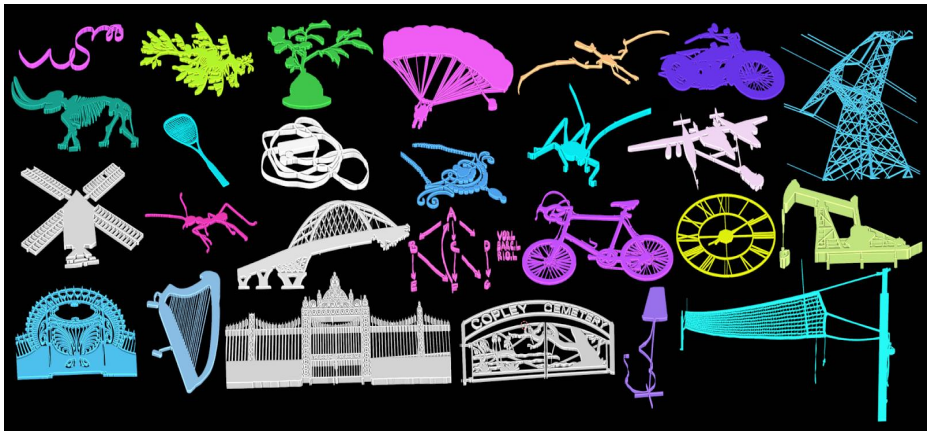


Fig. 9: 3D models built upon the ground truth masks sampled from DIS5K by the "Extrude" operation in Blender.

diate supervision significantly decreases the $M \downarrow$ on the validation set, which validates its effectiveness in performance improvement.

4 Applications

Our DIS task will benefit both academia and industrious. In addition to the DIS task, we believe that our highly accurate large-scale DIS5K dataset can also be used in various related research fields, such as:

- providing pre-trained segmentation models for other specific object segmentation tasks as well as facilitating the downstream tasks, such as image matting, editing, and so on;
- the subsets of DIS5K can be used for fast prototyping of different segmentation tasks;
- providing materials and examples for shape and structure analysis in graphics and topology;
- high resolution fine-grained image classification;
- segmentation guided super-resolution and image processing;
- synthesizing more composite images with diversified backgrounds for more robust image segmentation;
- edge, boundary or contour detection, *etc.*

Thanks to the high resolution and accurate labeling, many samples in our DIS5K show high artistic and aesthetic values. Fig. 1 shows the comparison between the original ship image with cluttered background and the background-removed image with perspective transforms (See more samples in Fig. 10). As can be seen, compared with the original image, the background-removed image shows higher aesthetic values and good usability, which can even be directly used as:

Fig. 10: Comparisons between the original images and their backgrounds-removed correspondences generated from our DIS5K.

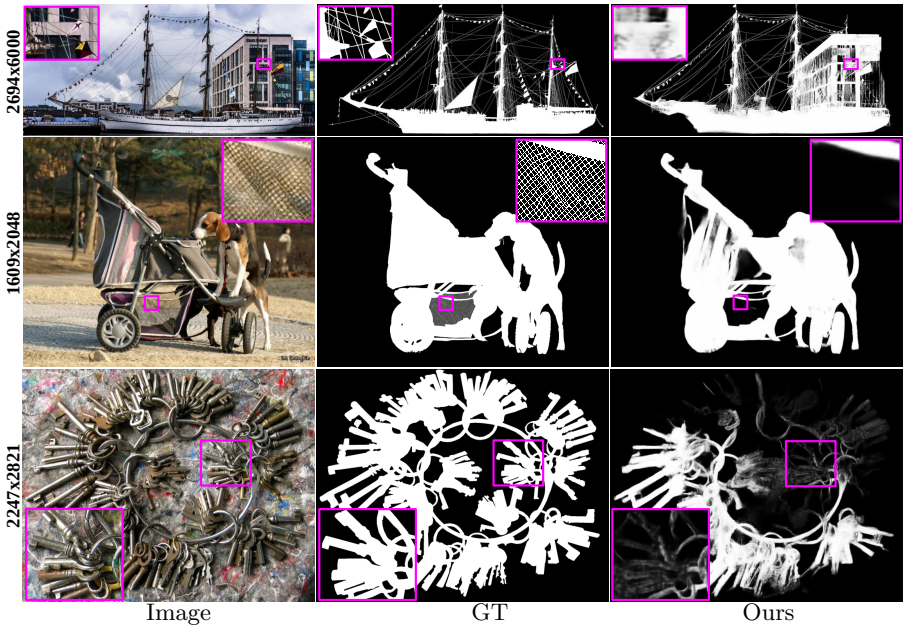


Fig. 11: Typical failure cases.

– materials of art design, image and video editing;
– backgrounds of posters or slides, wall papers of cellphones, tablets, desktops;
– materials for 3D modeling, as shown in Fig. 9 (A demo video is also attached).

5 Limitations and Future Works

Failure Cases of Our Model. Fig.11 shows some typical failure cases of our model. The first row shows the result of a sail ship image. Our model fails in segment two of the masts and the ropes because this region has a cluttered background (a building). The second row shows the segmentation result of a baby carriage. Our model fails in segmenting the mesh-like structure of the carriage since it is too meticulous (just one-pixel width), so that it is hard to be segmented by our model from the input images with the size of 1024×1024. The third row illustrates the segmentation result of a key chain with a cluttered background. As can be seen, the color differences between the critical chain and the background are small, which significantly increases the difficulty of the segmentation. In summary, the highly accurate DIS is a highly challenging task. There is still a large room for improvement. Therefore, more powerful deep segmentation models are needed to handle larger size input for obtaining very detailed object structures. In contrast, the model size, memory occupation, training, and inference time costs are expected to be affordable on the mainstream GPUs.
Limitations of Our DIS5K dataset. Although our DIS5K is currently the most complex dichotomous segmentation dataset, there is still a large room for

improvement. For example, compared with the vast number of categories and the diversified general object classes in the real-world, 225 categories in our DIS5K dataset are far from enough. Therefore, more categories, more samples of specific categories, and more diversified image qualities are needed to further improve the diversity of this dataset. Besides, semi-automatic and highly accurate annotation tools are expected to simplify and boost the ground truth labeling processes. We will explore semi-supervised and weakly supervised methods for further reducing the labeling workloads. In addition, it also requires a set of standard criteria to control the labeling accuracy.

Limitations of Our HCE metric. Our HCE metric provides direct measures of the human correction efforts needed for fixing faulty predictions under certain accuracy requirements. To leverage different accuracy requirements, the erosion [23] and dilation [23] operations are used to remove small false positive and false negative regions, while the skeleton extraction algorithm [71] is used to preserve the structural information of the thin components in the ground truth masks. However, the skeleton extraction algorithm is slow when processing the large-size masks. Therefore, the evaluation of large-scale datasets takes a long time. This issue also happens when computing the weighted F-measure [42], which uses a distance transform algorithm [3,19] to calculate the weights. Therefore, more works need to be conducted on these conventional algorithms, such as skeleton extraction, distance transform, etc., to handle larger and more complicated inputs.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39**(12), 2481–2495 (2017)
3. Borgefors, G.: Distance transformations in digital images. Comput. Vis. Graph. Image Process. **34**(3), 344–371 (1986)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
5. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV (2018)
6. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI (2020)
7. Cheng, B., Girshick, R.B., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: Improving object-centric image segmentation evaluation. In: CVPR (2021)
8. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR (2020)
9. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.: Global contrast based salient region detection. IEEE TPAMI **37**(3), 569–582 (2015)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
11. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: IJCAI (2018)
12. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: K-CapW (2005)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
14. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV (2018)
15. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV (2017)
16. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE TPAMI (2021)
17. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR (2020)
18. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR (2021)
19. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Theory Comput. **8**(1), 415–428 (2012)
20. Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet another survey on image segmentation: Region and boundary information integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV (2002)
21. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS (2010)
22. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`

23. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE TPAMI **PAMI-9**(4), 532–550 (1987)
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
25. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: CVPR (2017)
26. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: ECCV (2019)
27. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
29. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU **184**, 45–56 (2019)
30. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
31. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: CVPR (2019)
32. Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z., Sun, J., Jia, J.: Fully convolutional networks for panoptic segmentation. In: CVPR (2021)
33. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR (2014)
34. Liew, J.H., Cohen, S., Price, B., Mai, L., Feng, J.: Deep interactive thin object selection. In: WACV (2021)
35. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. CoRR **abs/2108.11515** (2021)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
37. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: ICCV (2021)
38. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE TPAMI **33**(2), 353–367 (2011)
39. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
40. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
41. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: CVPR (2017)
42. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. CVPR (2014)
43. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: CVPR (2021)
44. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: CVPRW (2010)
45. Nirkin, Y., Wolf, L., Hassner, T.: Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. arXiv preprint arXiv:2012.11582 (2020)
46. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: CVPR (2019)
47. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)

48. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
49. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Lin, Z., Torr, P., Jia, J.: Open-world entity segmentation. arXiv preprint arXiv:2107.14228 (2021)
50. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. PR **106**, 107404 (2020)
51. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR (2019)
52. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. CGIP **1**(3), 244–256 (1972)
53. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
54. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished Manuscript (2018)
55. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**(1), 1929–1958 (2014)
56. Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: ICCV (2021)
57. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2019)
58. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
59. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: ICCV (2017)
60. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: CVPR (2018)
61. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE TPAMI (2021). https://doi.org/10.1109/TPAMI.2021.3051099
62. Wei, J., Wang, S., Huang, Q.: F^3net: Fusion, feedback and focus for salient object detection. In: AAAI (2020)
63. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
64. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)
65. Yang, C., Wang, Y., Zhang, J., Zhang, H., Lin, Z., Yuille, A.: Meticulous object segmentation. arXiv preprint arXiv:2012.07181 (2020)
66. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013)
67. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (2018)
68. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: CVPR. pp. 7234–7243 (2019)
69. Zhang, P., Liu, W., Lu, H., Shen, C.: Salient object detection by lossless feature reflection. In: IJCAI (2018)
70. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: ICCV (2017)
71. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (1984)

72. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV (2018)
73. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
74. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)
75. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: ECCV (2020)
76. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)
77. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)