Highly Accurate Dichotomous Image Segmentation

Xuebin Qin
MBZUAI
Abu Dhabi, UAE
xuebinua@gmail.com

Hang Dai
MBZUAI
Abu Dhabi, UAE
hang.dai@mbzuai.ac.ae

Xiaobin Hu
TUM
Munich, Germany
xiaobin.hu@tum.de

Deng-Ping Fan ∗
ETH Zurich,
Switzerland
dengpfan@gmail.com

Ling Shao
Terminus Group
China
ling.shao@ieee.org

Luc Van Gool
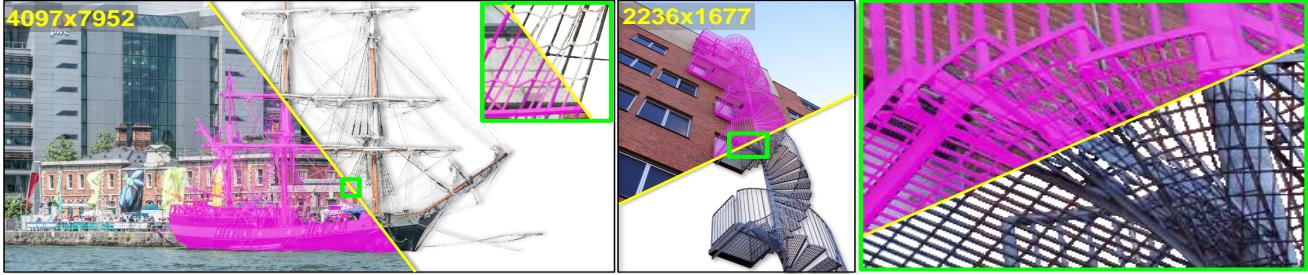ETH Zurich,
Switzerland
vangool@vision.ee.ethz.ch

Figure 1. Sample images (backgrounds partially removed by ground truth (GT) masks) from our DIS5K dataset. Zoom-in for best view.

Abstract

We present a systematic study on a new task called dichotomous image segmentation (DIS), which aims to segment highly accurate objects from natural images. To this end, we collected the first large-scale dataset, called DIS5K, which contains 5,470 high-resolution (e.g., 2K, 4K or larger) images covering camouflaged, salient, or meticulous objects in various backgrounds. All images are annotated with extremely fine-grained labels. In addition, we introduce a simple intermediate supervision baseline (IS-Net) using both feature-level and mask-level guidance for DIS model training. Without tricks, IS-Net outperforms various cutting-edge baselines on the proposed DIS5K, making it a general self-learned supervision network that can help facilitate future research in DIS. Further, we design a new metric called human correction efforts (HCE) which approximates the number of mouse clicking operations required to correct the false positives and false negatives. HCE is utilized to measure the gap between models and real-world applications and thus can complement existing metrics. Finally, we conduct the largest-scale benchmark, evaluating 16 representative segmentation models, providing a more insightful discussion regarding object complexi-

ties, and showing several potential applications (e.g., background removal, art design, 3D reconstruction). Hoping these efforts can open up promising directions for both academic and industries. Our DIS5K dataset, IS-Net baseline, HCE metric, and the complete benchmarks will be made publicly available at: https://xuebinqin.github.io/dis/index.html.

1. Introduction

In many years, the accuracy of annotations in computer vision datasets that drive a tremendous amount of Artificial Intelligence (AI) models satisfy the requirements of machine perceiving systems to some extent. However, AI has entered an era of demanding highly accurate outputs from computer vision algorithms to support delicate human-machine interaction and immersed virtual life. Image segmentation, as one of the most fundamental techniques in computer vision, plays a vital role in enabling the machines to perceive and understand the real world. Compared with image classification [17, 47, 84] and object detection [33, 34, 80], it can provide more geometrically accurate descriptions of the targets used in a wide range of applications, such as image editing [36], 3D reconstruction [57], augmented reality (AR) [76], satellite image anal-

---

* Corresponding author.

ysis [100], medical image processing [81], robot manipulation [8], *etc*. We can categorize the above applications as "light" (*e.g*., image editing and image analysis) and "heavy" (*e.g*., manufacturing and surgical robots), based on their immediate affects on real-world objects. The "light" applications (Fig.9) are relatively tolerant to the segmentation deflects and failures because these issues mainly lead to more labors and time costs, which are usually affordable. While, in the "heavy" applications, those deflects or failures are more likely to cause serious consequences, which are usually physic damages on objects or injuries, sometimes fatal for creatures, *e.g*., humans and animals. Hence, these applications require the models to be *highly accurate* and *robust*. Currently, most of the segmentation models are still less applicable in those "heavy" applications because of the accuracy and robustness issues, which restricts the segmentation techniques from playing more essential roles in broader applications. Here, **our goal** is to address the "heavy" and "light" applications in a general framework, we called this task as *dichotomous image segmentation (DIS)*, which aims to segment highly accurate objects from the nature images.

However, existing image segmentation tasks mainly focus on segmenting objects with specific characteristics, *e.g*., salient [90, 94, 109], camouflaged [26, 48, 85], meticulous [54, 105] or specific categories [45, 55, 67, 81, 83]. Most of them have the same input/output formats, and barely use exclusive mechanisms designed for segmenting targets in their models, which means almost all tasks are dataset-dependent. Thus, we propose to formulate **a category-agnostic DIS task defined on non-conflicting annotations for accurately segmenting objects with different structure complexities, regardless of their characteristics**. Compared with semantic segmentation [16, 20, 56, 75, 120], the proposed DIS task usually focuses on images with single or a few targets, from which getting richer accurate details of each target is more feasible. To this end, we provide four **contributions**:

1. A large-scale, extendable DIS dataset, **DIS5K**, contains 5,470 high-resolution images paired with highly accurate binary segmentation masks.

2. A novel baseline **IS-Net** built with our intermediate supervision reduces over-fitting by enforcing direct feature synchronization in high-dimensional feature spaces.

3. A newly designed human correction efforts (**HCE**) metric measures the barriers between model predictions and real-world applications by counting the human interventions needed to correct the faulty regions.

4. Based on the new DIS5K, we establish the complete DIS **benchmark**, making ours the most extensive DIS investigation. We compared our IS-Net with 16 cutting-edge segmentation models and

showed promising results for background removal and 3D reconstruction applications.

2. Related Work

Tasks and Datasets of image segmentation are closely related in deep learning era. Some of the segmentation tasks like [14, 24, 54, 55, 67, 83, 94, 105], are even directly built upon the datasets. Their problem formulations are exactly the same: $P = F(\theta, I)$, where I and P are the input image and the binary map output, respectively. However, the relevance between most of these tasks are rarely studied, which somehow restricts the models trained for certain tasks from being generalized to wider applications. Besides, the datasets used in different tasks are not exclusive, which shows a unified task for *dichotomous image segmentation* (DIS) is possible. However, most of the existing datasets are built on low-resolution images with objects of simple structures. There lacks an dataset built on the accurately labeled high-resolution images which contain objects with diversified shape complexities from different categories.

Models are often struggling with the conflicts between stronger representative capabilities and higher risks of over-fitting. To obtain more representative features, FCN-based models [60], Encoder-Decoder [3, 81], Coarse-to-Fine [96], Predict-Refine [78, 90], Vision Transformer [118] and so on are developed. Besides, many real-time models are designed [27, 44, 51, 70, 71, 107, 114] to balance the performance and the time costs. Other methods, such as weights regularization [37], dropout [86], dense supervision [49, 77, 102], and hybrid loss [61, 78, 116], focus on alleviating the over-fitting. Dense supervision is one of the most effective ways for reducing the over-fitting. However, supervising the side outputs from the intermediate deep features may not be the best option because the supervision is weakened by the conversion from deep features (multi-channel) to side outputs (one-channel).

Evaluation Metrics can be categorized as *region-based* (*e.g*., IoU or Jaccard index [1], F-measure [15, 92] or Dice's coefficient [88], weighted F-measure [64]), *boundary-based* (*e.g*., CM [69], boundary F-measure [19, 65, 68, 74, 78, 82, 113], boundary IoU [11], boundary displacement error (BDE) [31], Hausdorff distances [4, 5, 39]), *structure-based* (*e.g*., S-measure [22], E-measure [23, 25]), *confidence-based* (*e.g*., MAE [73]), *etc*. They mainly measure the consistencies between the predictions and the ground truth from mathematical or cognitive perspectives. But the costs of synchronizing the predictions against the requirements in real-world applications are not well studied.
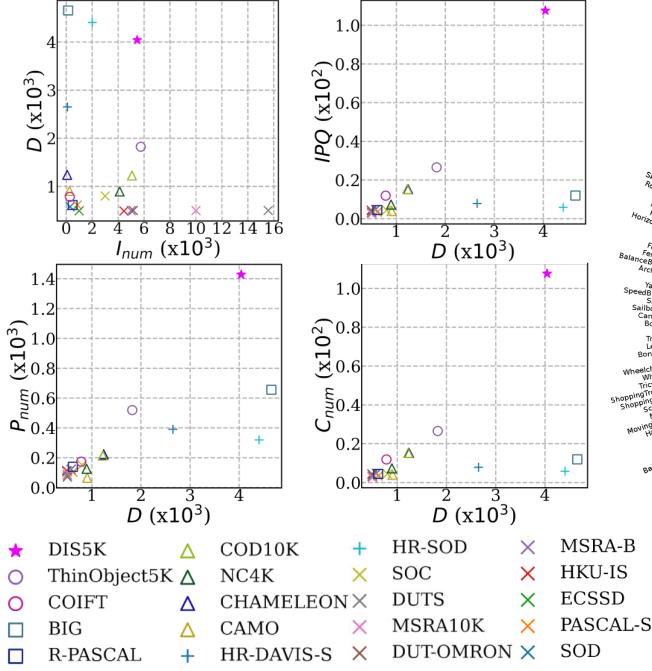
Figure 2. **Left**: Correlations between different complexities. **Right**: Categories and groups of our DIS5K dataset. Zoom-in for better view. Please refer to §3.1 for details.

3. Proposed DIS5K Dataset

3.1. Data Collection and Annotation

Data Collection. To address the dataset issue (see §2), we build a highly accurate DIS dataset named **DIS5K**. We first manually collected over 12,000 images from Flickr[1] based on our pre-designed keywords[2]. Then, according to the structural complexities of the objects, we obtained 5,470 images covering 225 categories (Fig.2) in 22 groups. Note that the adopted selection strategy is similar to Zhou *et al.* [119]. Most selected images only contain single objects to obtain rich and highly accurate structures and details. Meanwhile, the segmentation and labeling confusions caused by the co-occurrence of multiple objects from different categories are avoided to the greatest extent. Specifically, the image selection criteria can be summarized as follows:

- Cover more categories while reducing the number of "redundant" samples with simple structures, which other existing datasets have already covered.

- Enlarge the intra-category dissimilarities (See §2.3 of the supplementary (SM)) of the selected categories by adding more diversified intra-category images (Fig.3-f).

- Include more categories with complicated structures, *e.g.*, *fence, stairs, cable, bonsai, tree, etc.*, which are common in our lives but not well-labeled (Fig.3-a) or neglected by other datasets due to labeling difficulties (Fig.4).

Therefore, the labeled targets in our DIS5K are mainly the "*foreground objects of the images defined by the pre-designed keywords*" regardless of their characteristics *e.g.*, *salient, common, camouflaged, meticulous, etc.*

Data Annotation. Each image of DIS5K is manually labeled with pixel-wise accuracy using GIMP[3]. The average per-image labeling time is ∼30 minutes and some images cost up to 10 hours. It is worth mentioning that some of our labeled ground truth (GT) masks are visually close to the image matting GT. The labeled targets, including transparent and translucent, are binary masks with one pixel's highest accuracy. Here, the DIS task is category-agnostic while our DIS5K is collected based on pre-designed keywords/categories, which seems contradictory. The reasons are threefold. (1) The keywords greatly facilitate the retrieval and organization of the large-scale dataset. (2) To achieve the goal of category-agnostic segmentation, diversified samples are needed. Collecting samples based on their categories is a reasonable way to guarantee the diversities' lower bound of the dataset. The diversities' upper bound of our DIS5K is determined by the diversified characteristics (*e.g.*, textures, structures, shapes, contrasts, complexities,

---

Table 1. Data analysis of existing datasets. See §3.2 for details.

Task	Dataset	Number	Image Dimension			Object Complexity		
		I_{num}	$H \pm \sigma_H$	$W \pm \sigma_W$	$D \pm \sigma_D$	$IPQ \pm \sigma_{IPQ}$	$C_{num} \pm \sigma_C$	$P_{num} \pm \sigma_P$
SOD	SOD [69]	300	366.87 ± 72.35	435.13 ± 72.35	578.28 ± 0.00	4.74 ± 3.89	2.25 ± 1.76	122.79 ± 62.97
	PASCAL-S [52]	850	387.63 ± 64.65	467.82 ± 61.46	613.22 ± 32.00	3.39 ± 2.46	5.14 ± 11.72	102.76 ± 70.09
	ECSSD [104]	1000	311.11 ± 56.27	375.45 ± 47.70	492.75 ± 19.78	3.26 ± 2.62	1.69 ± 1.42	107.54 ± 53.09
	HKU-IS [50]	4447	292.42 ± 51.13	386.64 ± 37.42	488.00 ± 29.44	4.41 ± 4.28	2.21 ± 2.07	114.05 ± 55.06
	MSRA-B [59]	5000	321.94 ± 56.33	370.86 ± 50.84	496.42 ± 22.53	2.89 ± 3.67	1.77 ± 2.25	102.04 ± 56.50
	DUT-OMRON [106]	5168	320.93 ± 54.35	376.78 ± 46.02	499.50 ± 22.97	4.08 ± 6.20	2.27 ± 3.54	71.09 ± 59.60
	MSRA10K [14]	10000	324.51 ± 56.26	370.27 ± 50.25	497.57 ± 22.79	2.54 ± 2.62	4.07 ± 17.94	101.95 ± 63.24
	DUTS [94]	15572	322.1 ± 53.69	375.48 ± 47.03	499.35 ± 21.95	3.37 ± 4.28	2.62 ± 4.73	84.78 ± 57.74
	SOC [21]	3000	480.00 ± 0.00	640.00 ± 0.00	800.00 ± 0.00	4.44 ± 3.57	13.69 ± 30.41	151.72 ± 154.83
HRS	HR-SOD [109]	2010	2713.12 ± 1041.7	3411.81 ± 1407.56	4405.40 ± 1631.03	5.85 ± 12.60	6.33 ± 16.65	319.32 ± 264.20
	HR-DAVIS-S [74]	92	1299.13 ± 440.77	2309.57 ± 783.59	2649.87 ± 899.05	7.84 ± 5.69	15.60 ± 29.51	389.58 ± 309.29
COD	CAMO [48]	250	564.22 ± 402.12	693.89 ± 578.53	905.51 ± 690.12	3.97 ± 4.47	1.48 ± 1.18	65.21 ± 40.99
	CHAMELEON [85]	76	741.80 ± 452.25	981.08 ± 464.88	1239.98 ± 629.19	15.25 ± 51.43	10.28 ± 48.03	222.45 ± 332.22
	NC4K [26]	4121	529.61 ± 158.16	709.19 ± 198.90	893.23 ± 223.94	7.28 ± 11.28	4.32 ± 9.44	125.43 ± 123.76
	COD10K [26]	5066	737.37 ± 185.65	963.85 ± 222.73	1224.53 ± 239.40	15.28 ± 71.84	17.18 ± 183.87	214.12 ± 857.83
SMS	R-PASCAL [13]	501	384.34 ± 64.69	469.66 ± 60.04	612.19 ± 36.32	4.44 ± 6.91	7.30 ± 8.73	139.31 ± 104.60
	BIG [13]	150	2801.11 ± 889.78	3672.43 ± 1128.90	4655.81 ± 1312.44	11.94 ± 31.43	31.69 ± 71.94	655.68 ± 710.20
TOS	COIFT [54]	280	488.27 ± 92.25	600.40 ± 78.66	782.73 ± 30.45	11.88 ± 12.5	4.01 ± 3.98	173.14 ± 74.54
	ThinObject5K [54]	5748	1185.59 ± 909.53	1325.06 ± 958.43	1823.03 ± 1258.49	26.53 ± 119.98	33.06 ± 216.07	519.14 ± 1298.54
DIS	**DIS5K (Ours)**	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

etc.) of a large number of samples, guaranteeing the robustness and generalization of the category-agnostic segmentation. (3) There are no perfect datasets, so re-organizing or further extension of the existing datasets is usually necessary for different real-world applications. The category information will significantly facilitate tracing the collected and to-be-collected samples. Therefore, the category-based data collection is not contradictory but internally consistent with the goal of DIS task.

3.2. Data Analysis

For deeper insights into DIS dataset, we compare our DIS5K against 19 other related datasets including: (1) nine salient object detection (SOD) datasets: SOD [69], PASCAL-S [52], ECSSD [104], HKU-IS [50], MSRA-B [59], DUT-OMRON [106], MSRA10K [14], DUTS [94], and SOC [21]; (2) two high-resolution salient object detection (HR-SOD) datasets: HR-SOD [109] and HR-DAVIS-S [74, 109]; (3) four camouflaged object detection (COD) datasets: CAMO [48], CHAMELEON [85], COD10K [26], and NC4K [63]; (4) two semantic segmentation (SMS)[4] datasets: R-PASCAL [13, 20] and BIG [13]; (5) two thin object segmentation (TOS) datasets: COIFT [54] and ThinObject5K [54]. The comparisons are conducted mainly from the following three perspectives: *image number*, *image dimension*, and *object complexity* as illustrated in Tab.1

Image Dimension is crucial to segmentation tasks. Because it has significant impacts on accuracy, efficiency, and computational costs. The mean (H, W, D) and their standard deviations (σ_H, σ_W, σ_D) of the image height, width

and diagonal length are provided in Tab.1. The BIG dataset has the largest average image dimensions, but it is a small-scale dataset that contains only 150 images. Although HR-SOD has slightly greater dimensions than ours, the dataset scale and complexity are less comparable. Compared with the SOD datasets, the average image dimensions of our DIS5K are almost eight times larger than theirs. The COD datasets have larger dimensions than SOD datasets, but they are still much smaller than ours. Besides, most of the targets in COD datasets are animals. Thus, it is difficult to generalize them to diversified tasks.

Object Complexity is described by three metrics including the *isoperimetric inequality quotient* (IPQ) [72, 98, 105], the *number of object contours* (C_{num}) and the *number of dominant points* P_{num}. The IPQ mainly describes the overall structure complexity as $IPQ = \frac{L^2}{4\pi A}$, where L and A denote the object perimeter and the region area, respectively. It is designed to differentiate objects with elongated components and thin concave structures from close-to-convex objects. The C_{num} is used to represent the topological complexity in contour level for observing the objects consisting of many (small) contours which usually have minor influences on the IPQ. To describe the object complexity at a finer level, we employ P_{num} to count the number of the dominant points [79] along the object boundaries. Therefore, the complexities of the small jagged segments along the boundaries, which usually cannot be accurately measured by IPQ and C_{num}, can be well-evaluated with P_{num}. Essentially, P_{num} is the total number of the polygon corners needed for approximating the segmentation masks, which also directly reflects the human labeling costs. Thus, it is then adapted to our Human Correction Efforts (HCE) metric (§5) for evaluating the prediction quality.

Discussion. The metrics above are all computed on the la-

---

[4]It is worth noting that only R-PASCAL and the BIG datasets are included here because they target highly accurate segmentation, and most of their images contain one or two objects, which is comparable to the listed tasks and datasets.
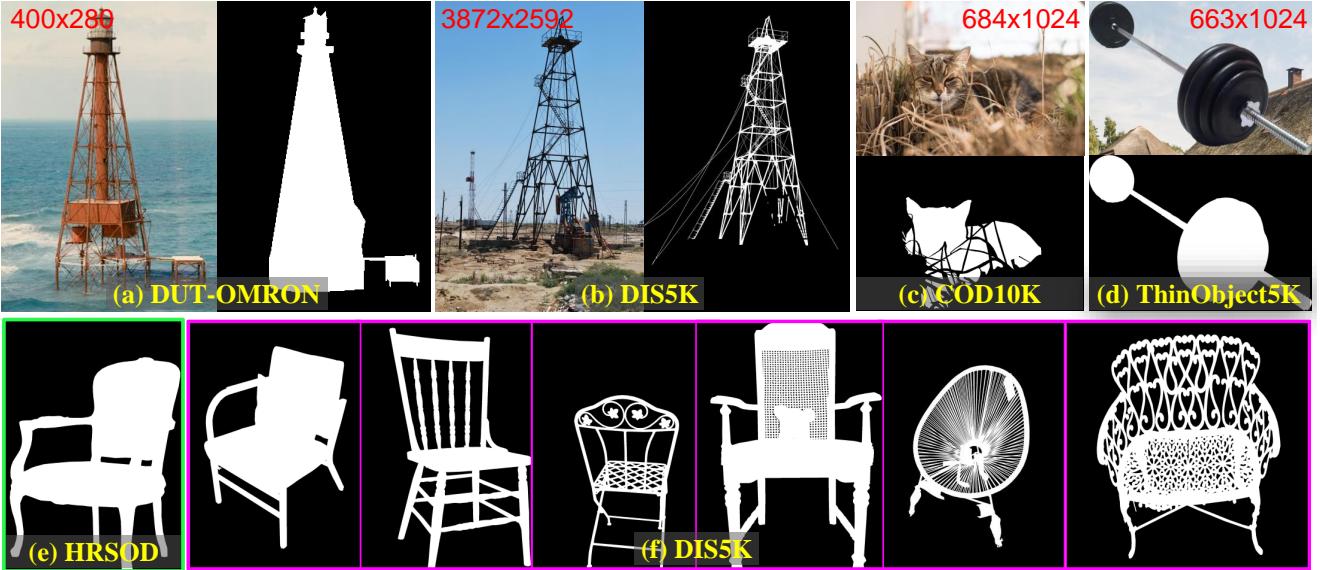
Figure 3. Qualitative comparisons of different datasets. (a) and (b) indicate that our DIS5K provides more accurate labels. (c) shows one sample from COD10K [26], of which the structural complexity is caused by occlusion. (d) illustrates the synthetic ThinObject5K [54] dataset. (e) and (f) demonstrate that DIS5K has a larger diversity of intra-categorical structure complexities.
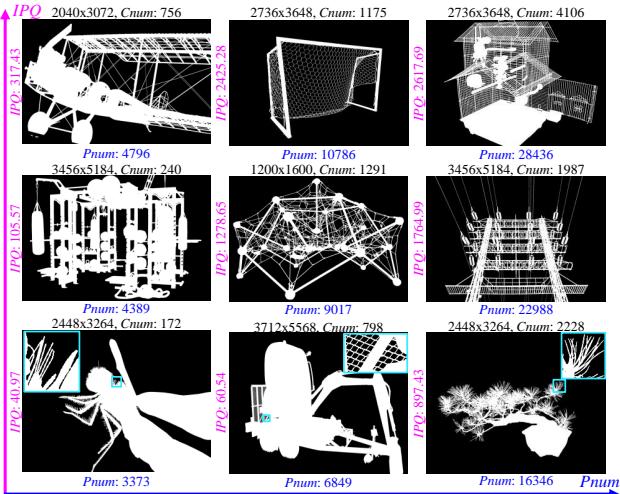

Figure 4. GT masks of our DIS5K with diversified inter-categorical complexities. The complexity relationships are only valid within each row or column.

beled GT masks and illustrated in Tab.1 and Fig.2 (Left). It shows that DIS5K is around 20 (up to 50) times more complicated than the SOD datasets in terms of average structure complexity IPQ. Although other datasets such as CHAMELEON, COD10K, BIG, COIFT, and ThinObject5K have higher average IPQ against the SOD datasets, their complexities are still much less 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 [79], 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

demonstrate the great diversities of DIS5K from different perspectives.

Fig.3-a shows an observation tower from DUT-OMRON. Similar object (b) has also been included in our DIS5K, which has higher labeling accuracy and structural complexity. Fig.3-c shows a sample from COD10K where the relatively higher structure complexity than that of SOD datasets is partially caused by the labeled occlusions, which are not the structural complexity of the target itself. A sample, where a set of the barbell is floating in the sky, from the synthesized ThinObject5K dataset is shown in Fig.3-d. Synthesizing images is a common way for generating training sets in image matting [103, 108], where training samples are difficult to be labeled. However, the synthesized images usually show different characteristics from the real ones, which leads to biases in both training and evaluation. Fig.3-e and Fig.3-f demonstrate the larger diversity of intra-categorical structure complexities of our DIS5K. In Fig.4, 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}.

3.3. Dataset Splitting

We split 5,470 images in DIS5K into three subsets: DIS-TR (3,000), DIS-VD (470), and DIS-TE (2,000) for training, validation, and testing. The categories in DIS-TR and those in DIS-VD and DIS-TE are mainly consistent. Since our dataset's object shapes and structure complexities are diversified, the 2000 images of DIS-TE are further split into four subsets with ascending shape complexities for a more comprehensive evaluation. Specifically, we first rank the 2,000 testing images in ascending order according to the multiplication ($IPQ \times P_{num}$) of their structure complexities IPQ and boundary complexities P_{num}. Then, DIS-TE is split into four subsets (DIS-TE1∼DIS-TE4) with 500 images in each subset to represent four testing difficulty levels.

4. Proposed IS-Net Baseline

4.1. Overview

As shown in Fig.5, our IS-Net consists of a ground truth (GT) encoder, a image segmentation component, and a newly proposed intermediate supervision strategy. The **GT encoder** (27.7 MB) is designed to encode the GT masks into high-dimensional spaces and then used to enforce intermediate supervision on the segmentation component. While, the **image segmentation component** (176.6 MB) is expected to have the capability of capturing fine structures and handle large size *e.g.*, 1024×1024, inputs with affordable memory and time costs. In the following experiment, we choose U^2-Net [77] as the image segmentation component

because of its strong capability in capturing fine structures. Note that other segmentation models, such as transformer backbone, are also compatible with our strategy.

Technique Details. U^2-Net was originally designed for small size (320×320) SOD image. Because of its GPU memory costs, it cannot be used directly for handling large size (*e.g.*, 1024×1024) inputs. We adapt the architecture of U^2-Net by adding an input convolution layer before its first encoder stage. The input convolution layer is set as a plain convolution layer with a kernel size of 3×3 and stride of 2. Given an input image with a shape of $I^{1024 \times 1024 \times 3}$, the input convolution layer first transforms it to a feature map $f^{512 \times 512 \times 64}$ and this feature map is then directly fed to the original U^2-Net, where the input channel is changed to 64 correspondingly. Compared with directly feeding $I^{1024 \times 1024 \times 3}$ to U^2-Net, the input convolution layer helps the whole network reduce three quarters of the overall GPU memory overhead while maintaining spatial information in feature channels.

4.2. Intermediate Supervision

DIS can be seen as a mapping in segmentation models from image domain $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ to segmentation GT domain $\mathcal{G} \in \mathbb{R}^{H \times W \times 1}$: $\mathcal{G} = F(\theta, \mathcal{I})$, where F indicates the model that uses learnable weights θ to map inputs from image to mask domain. Most of the models are easy to over-fit on the training set. Thus, the deep supervision has been proposed to supervise the intermediate outputs of a given deep network [49]. In [77, 102], the dense supervisions are usually applied to the side outputs, which are single-channel probability maps produced by convolving the last feature maps of particular deep layers. However, transforming high-dimensional features to single-channel probability maps is essentially a dimension reduction operation, inevitably losing critical cues.

To avoid this issue, we propose a novel intermediate supervision training strategy. Given an input image $I^{H \times W \times 3}$ and its corresponding segmentation mask $G^{W \times H \times 1}$, we first train a self-supervised GT encoder to extract the high-dimensional features using a lightweight deep model F_{gt}, Fig.5-b, as: $\underset{\theta_{gt}}{\operatorname{argmin}} \sum_{d=1}^{D} BCE(F_{gt}(\theta_{gt}, G)_d, G)$, where θ_{gt} indicates the model weights, BCE is the binary cross entropy loss and D denotes the number of the intermediate feature maps.

After obtaining the GT encoder F_{gt}, its weights θ_{gt} are frozen for generating the "ground truth" high-dimensional intermediate deep features by: $f_D^G = F_{gt}^-(\theta_{gt}, G), D = \{1, 2, 3, 4, 5, 6\}$, where F_{gt}^- represents the F_{gt} without the last convolution layers for generating the probability maps. F_{gt}^- is to supervise those corresponding features f_D^I from the segmentation model F_{sg}. In the image segmentation component F_{sg} (Fig.5-a), the image I is transformed to a set

Figure 5. Proposed IS-Net baseline: (a) shows the image segmentation component, (b) illustrates the ground truth encoder built upon the intermediate supervision (IS) component.

Figure 6. Feature maps produced by the last layer of the EN_2 stage of our GT encoder. "21", "23", "29" and "37" are the indices (start with 1) of the corresponding channels in the feature map.

of high-dimensional intermediate feature maps f_D^I before producing the probability maps. Each feature map f_d^I has the same dimension with its corresponding GT intermediate feature map f_d^G: $f_D^I = F_{sg}^-(\theta_{sg}, I), D = \{1, 2, 3, 4, 5, 6\}$, where θ_{sg} denotes the weights of the segmentation model. Then, the intermediate supervision (IS) via *feature synchronization* on the deep intermediate features can be conducted by the following high-dimensional feature consistency loss: $L_{\text{fs}} = \sum_{d=1}^{D} \lambda_d^{fs} \left\| f_d^I - f_d^G \right\|^2$, where λ_d^{fs} denotes the weight of each FS loss. The training process of the segmentation model F_{sg} can be formulated as the following optimization problem: $\underset{\theta_{sg}}{\operatorname{argmin}}(L_{\text{fs}} + L_{\text{sg}})$, where L_{sg} indicates the BCE loss of the side outputs of F_{sg}: $L_{\text{sg}} = \sum_{d=1}^{D} \lambda_d^{sg} BCE(F_{sg}(\theta_{sg}, I), G)$, where λ_d^{sg} repre-

sents the hyperparameter to weight each side output loss.

Fig.6 illustrates the feature maps from the stage 2 in Fig.5, EN_2, of the GT encoder. We can see the diversified characteristics of the input mask are encoded into different channels. For example, the 21st channel encodes both the fine and large structures close to the original mask. While the 23rd, 29th, and 37th channels encode the middle size structures (frame, seat, wheels), delicate structures (brake cables and spokes), large size region (the overall shape of the bicycle), respectively. These diversified features of the GT can provide stronger regularizations and more comprehensive supervisions for reducing the risks of over-fitting.

| (a) Error Map | (b) FN$_N$ | (c) FN$_{TP}$ | (d) FP$_P$ | (e) FP$_{TN}$ |

Figure 7. Faulty regions to be corrected. Refer to §5 for details.

5. Proposed HCE Metric

Given a predicted segmentation probability map $P \in \mathbb{R}^{W \times H \times 1}$ and its corresponding GT mask $G \in \mathbb{R}^{W \times H \times 1}$, the existing metrics, *e.g.*, IoU, boundary IoU [12], F-measure [2], boundary F-measure [19, 78], and MAE [73], usually evaluate the quality of the prediction P by calculating the scores based on the mathematical or cognitive consistency (or in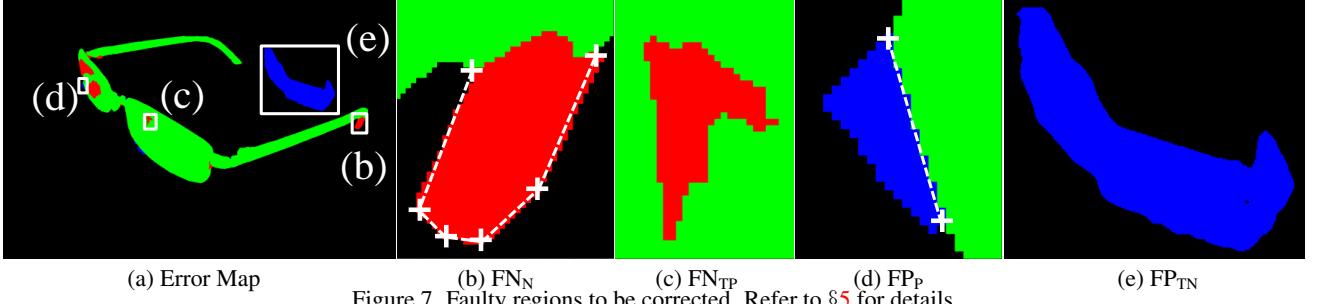consistency) between P and G. In other words, these metrics describe how significant the "gap" is between P and M from different perspectives. However, measuring the magnitude of the "gap" is insufficient when applying the models in many real-world applications, where evaluating the costs of filling the "gap" is more important.

Therefore, we propose a novel evaluation metric, Human Correction Efforts (HCE), which approximates the human efforts required in correcting faulty predictions to satisfy specific accuracy requirements in real-world applications. According to our labeling experiences, there are mainly two frequently used operations: (1) points selection along target boundaries to formulate polygons and (2) region selection based on similar pixel intensities inside the region. Both operations correspond to one mouse click by the human operator. Therefore, the HCE here is quantified by the approximated number of mouse clicking numbers. Particularly, to correct a faulty predicted mask, the operators need to manually sample dominant points along the erroneously predicted targets' boundaries or regions for correcting both False Positive (FP) and False Negative (FN) regions. As shown in Fig.7, the FNs and FPs can be categorized into two classes, respectively, according to their adjacent regions: FN$_N$ (N=TN+FP), FN$_{TP}$, FP$_P$ (P=TP+FN) and FP$_{TN}$. To correct the FN$_N$ regions, its boundaries adjacent to the TN need to be manually labeled with dominant points (Fig.7-b). Similarly, to correct the FP$_P$ regions, we only need to label its boundaries adjacent to the TP regions (Fig.7-d). The FN$_{TP}$ regions (Fig.7-c) enclosed by TP and the FP$_{TN}$ regions (Fig.7-e) enclosed by TN can be easily corrected by one-click region selection. Therefore, the HCE for correcting the faulty regions in Fig.7 (b-e) is 10 (six and two clicks needed in (b) and (d), one click needed in (c) and one click needed in (e)). The dominant point selection operations and the region selection operations are approximated by DP al-

---

Input: $P, G, \gamma = 5, epsilon = 2.0$
Output: HCE_γ
1 G_{ske} = skeletonize (G);
2 $P_{or}G, TP$ = or (P, G), and (P,G);
3 FN, FP= $(G$ - $TP)$, $(P$ - $TP)$;
4 **for** $(i = 0; i \leq \gamma; i + +)$ **do**
5      $P_{or}G$ = erode $(P_{or}G, disk$ (1));
6 **end**
7 FN', FP' = and $(FN, P_{or}G)$, and $(FP, P_{or}G)$;
8 **for** $(i = 0; i \leq \gamma; i + +)$ **do**
9   FN' = dilate $(FN', disk$ (1));
10   FN' = and $(FN',$ not $P)$;
11   FP' = dilate $(FP', disk$ (1));
12   FP' = and $(FP',$ not $G)$;
13 **end**
14 FN', FP' = and (FN, FN'), and (FP, FP');
15 FN' = or $(FN',$ xor $(G_{ske},$ and $(TP, G_{ske})))$;
16 HCE_γ = compute_HCE $(FN', FP', TP,$ epsilon)

Algorithm 1: Relax HCE.

gorithm [79] based on the contours obtained by OpenCV findContours [87] function and the connected regions labeling algorithm [30,101], respectively, in the evaluation stage. **Relax HCE.** Practically, some applications may be tolerant to certain minor prediction errors. Therefore, we extend the computation of HCE by taking the error tolerance γ into consideration (HCE_γ). The key idea is to relax the FP and FN regions by excluding the small FP and FN components using erosion [38] and dilation [38] operations. Given a segmentation map P, its corresponding GT mask G, the error tolerance (*e.g.*, $\gamma = 5$, which denotes the size of the to-be-ignored small faulty regions), the *epsilon* of DP algorithm, the computation of the HCE_γ can be summarized in Alg. 1. Note that the erosion operation (Line5 of Alg. 1) can remove all thin and fine components of $P_{or}G$. However, some of the thin components (*e.g.*, thin cables, nets) are critical in representing the targets, and they need to be retained regardless of their sizes. To address this, the skeleton of the GT mask is extracted by [112] and combined with the relaxed FN' mask for retaining these structures.

6. DIS5K Benchmark

As discussed above, our DIS5K is built from scratch to cover highly diversified objects with very different geometrical structures and image characteristics. One of the most important reasons is to exclude the existing datasets'

Table 2. Quantitative evaluation on DIS5K validation and test sets. R = ResNet [41]. R2 = Res2Net [32]. S-813 = STDC813 [27], E-B1 = EffinetB1 [89].

Dataset	Metric	UNet [81]	BASNet [78]	GateNet [117]	F³Net [99]	GCPANet [10]	U²Net [77]	SINetV2 [24]	PFNet [66]	PSPNet [115]	DLV3+ [7]	HRNet [93]	BSV1 [107]	ICNet [114]	MBV3 [43]	STDC [27]	HySM [70]	IS-Net
Attr.	Backbone	-	R-34	R-50	R-50	R-50	-	R2-50	R-50	R-50	R-50	-	R-18	R-18	MBV3	S-813	E-B1	-
	Size (MB)	121.4	348.6	515.0	102.6	268.7	176.3	108.5	186.6	196.1	161.8	264.4	47.6	46.5	21.5	48.4	49.6	176.6
	Time (ms)	3.87	10.71	12.69	14.23	11.04	19.73	18.69	17.16	8.08	8.68	40.5	6.07	4.93	8.86	6.17	24.06	19.49
	Input Size	512^2	320^2	384^2	352^2	320^2	320^2	352^2	416^2	512^2	513^2	1024^2	1024×2048	1024×2048	1024^2	512×1024	512×1024	1024^2
DIS-VD	$maxF_\beta \uparrow$	0.692	0.731	0.678	0.685	0.648	0.748	0.665	0.691	0.691	0.660	0.726	0.662	0.697	0.714	0.696	0.734	**0.791**
	$F_\beta^w \uparrow$	0.586	0.641	0.574	0.595	0.542	0.656	0.584	0.604	0.603	0.568	0.641	0.548	0.609	0.642	0.613	0.640	**0.717**
	$M \downarrow$	0.113	0.094	0.110	0.107	0.118	0.090	0.110	0.106	0.102	0.114	0.095	0.116	0.102	0.092	0.103	0.096	**0.074**
	$S_\alpha \uparrow$	0.745	0.768	0.723	0.733	0.718	0.781	0.727	0.740	0.744	0.716	0.767	0.728	0.747	0.758	0.740	0.773	**0.813**
	$E_\phi^m \uparrow$	0.785	0.816	0.783	0.800	0.765	0.823	0.798	0.811	0.802	0.796	0.824	0.767	0.811	0.841	0.817	0.814	**0.856**
	$HCE_\gamma \downarrow$	1337	1402	1493	1567	1555	1413	1568	1606	1588	1520	1560	1660	1503	1625	1598	1324	**1116**
DIS-TE1	$maxF_\beta \uparrow$	0.625	0.688	0.620	0.640	0.598	0.694	0.644	0.646	0.645	0.601	0.668	0.595	0.631	0.669	0.648	0.695	**0.740**
	$F_\beta^w \uparrow$	0.514	0.595	0.517	0.549	0.495	0.601	0.558	0.552	0.557	0.506	0.579	0.474	0.535	0.595	0.562	0.597	**0.662**
	$M \downarrow$	0.106	0.084	0.099	0.095	0.103	0.083	0.094	0.094	0.089	0.102	0.088	0.108	0.095	0.083	0.090	0.082	**0.074**
	$S_\alpha \uparrow$	0.716	0.754	0.701	0.721	0.705	0.760	0.727	0.722	0.725	0.694	0.742	0.695	0.716	0.740	0.723	0.761	**0.787**
	$E_\phi^m \uparrow$	0.750	0.801	0.766	0.783	0.750	0.801	0.791	0.786	0.791	0.772	0.797	0.741	0.784	0.818	0.798	0.803	**0.820**
	$HCE_\gamma \downarrow$	233	220	230	244	271	224	274	253	267	234	262	288	234	274	249	205	**149**
DIS-TE2	$maxF_\beta \uparrow$	0.703	0.755	0.702	0.712	0.673	0.756	0.700	0.720	0.724	0.681	0.747	0.680	0.716	0.743	0.720	0.759	**0.799**
	$F_\beta^w \uparrow$	0.597	0.668	0.598	0.620	0.570	0.668	0.618	0.633	0.636	0.587	0.664	0.564	0.627	0.672	0.636	0.667	**0.728**
	$M \downarrow$	0.107	0.084	0.102	0.097	0.109	0.085	0.099	0.096	0.092	0.105	0.087	0.111	0.095	0.083	0.092	0.085	**0.070**
	$S_\alpha \uparrow$	0.755	0.786	0.737	0.755	0.735	0.788	0.753	0.761	0.763	0.729	0.784	0.740	0.759	0.777	0.759	0.794	**0.823**
	$E_\phi^m \uparrow$	0.796	0.836	0.804	0.820	0.786	0.833	0.823	0.829	0.828	0.813	0.840	0.781	0.826	0.856	0.834	0.832	**0.858**
	$HCE_\gamma \downarrow$	474	480	501	542	574	490	593	567	586	516	555	621	512	600	556	451	**340**
DIS-TE3	$maxF_\beta \uparrow$	0.748	0.785	0.726	0.743	0.699	0.798	0.730	0.751	0.747	0.717	0.784	0.710	0.752	0.772	0.745	0.792	**0.830**
	$F_\beta^w \uparrow$	0.644	0.696	0.620	0.656	0.590	0.707	0.641	0.664	0.657	0.623	0.700	0.595	0.664	0.702	0.662	0.701	**0.758**
	$M \downarrow$	0.098	0.083	0.103	0.092	0.109	0.079	0.096	0.092	0.092	0.102	0.080	0.109	0.091	0.078	0.090	0.079	**0.064**
	$S_\alpha \uparrow$	0.780	0.798	0.747	0.773	0.748	0.809	0.766	0.777	0.774	0.749	0.805	0.757	0.780	0.794	0.771	0.811	**0.836**
	$E_\phi^m \uparrow$	0.827	0.856	0.815	0.848	0.801	0.858	0.849	0.854	0.843	0.833	0.869	0.801	0.852	0.880	0.855	0.857	**0.883**
	$HCE_\gamma \downarrow$	883	948	972	1059	1058	965	1096	1082	1111	999	1049	1146	1001	1136	1081	887	**687**
DIS-TE4	$maxF_\beta \uparrow$	0.759	0.780	0.729	0.721	0.670	0.795	0.699	0.731	0.725	0.715	0.772	0.710	0.749	0.736	0.731	0.782	**0.827**
	$F_\beta^w \uparrow$	0.659	0.693	0.625	0.633	0.559	0.705	0.616	0.647	0.630	0.621	0.687	0.598	0.663	0.664	0.652	0.693	**0.753**
	$M \downarrow$	0.102	0.091	0.109	0.107	0.127	0.087	0.113	0.107	0.107	0.111	0.092	0.114	0.099	0.098	0.102	0.091	**0.072**
	$S_\alpha \uparrow$	0.784	0.794	0.743	0.752	0.723	0.807	0.744	0.763	0.758	0.744	0.792	0.755	0.776	0.770	0.762	0.802	**0.830**
	$E_\phi^m \uparrow$	0.821	0.848	0.803	0.825	0.767	0.847	0.824	0.838	0.815	0.820	0.854	0.788	0.837	0.848	0.841	0.842	**0.870**
	$HCE_\gamma \downarrow$	3218	3601	3654	3760	3678	3653	3683	3803	3806	3709	3864	3999	3690	3817	3819	3331	**2888**
Overall DIS-TE (1-4)	$maxF_\beta \uparrow$	0.708	0.752	0.694	0.704	0.660	0.761	0.693	0.712	0.710	0.678	0.743	0.674	0.711	0.729	0.710	0.757	**0.799**
	$F_\beta^w \uparrow$	0.603	0.663	0.590	0.614	0.554	0.670	0.608	0.624	0.620	0.584	0.658	0.558	0.622	0.658	0.628	0.665	**0.726**
	$M \downarrow$	0.103	0.086	0.103	0.098	0.112	0.083	0.101	0.097	0.095	0.105	0.087	0.110	0.095	0.085	0.094	0.084	**0.070**
	$S_\alpha \uparrow$	0.759	0.783	0.732	0.750	0.728	0.791	0.747	0.756	0.755	0.729	0.781	0.737	0.758	0.770	0.754	0.792	**0.819**
	$E_\phi^m \uparrow$	0.798	0.835	0.797	0.819	0.776	0.835	0.822	0.827	0.819	0.810	0.840	0.778	0.825	0.850	0.832	0.834	**0.858**
	$HCE_\gamma \downarrow$	1202	1313	1339	1401	1395	1333	1411	1427	1442	1365	1432	1513	1359	1457	1426	1218	**1016**

possible biases (to specific image or object characteristics). Therefore, its diversities (*e.g.*, resolutions, image characteristics, object complexities, labeling accuracy) and distributions differ from the existing datasets. All models are trained, validated, and tested on DIS-TR, DIS-VD, and DIS-TE, respectively, to provide a fair comparison. Currently, cross-dataset evaluations [91] are not conducted mainly because their labeling accuracy is not consistent with ours.

Metrics. To provide relatively comprehensive and unbiased evaluations, six different metrics, including maximal F-measure ($F_\beta^{mx} \uparrow$) [2], weighted F-measure ($F_\beta^w \uparrow$) [64], mean absolute error ($M \downarrow$) [73], structural measure ($S_\alpha \uparrow$) [22], mean enhanced alignment measure ($E_\phi^m \uparrow$) [23, 25] and our human correction efforts ($HCE\gamma \downarrow$), are used to evaluate the performance from different perspectives.

Competitors. To provide comprehensive evaluations, we compared our IS-Net with 16 popular networks designed for different segmentation tasks, including (i) popular medical image segmentation model, U-Net [81]; (ii) salient object detection models such as BASNet [78], GateNet [117], F³Net [99], GCPA [10] and U²-Net [77]; (iii) models designed for COD like SINet-V2 [24] and PFNet [66]; (iv)

semantic segmentation models: PSPNet [115], DeepLab-V3+ [7] and HRNet [93]; (v) real-time semantic segmentation models: BiSeNetV1 [107], ICNet [114], MobileNet-V3-Large [43], STDC [28] and HyperSegM [70]. All models are re-trained using DIS-TR set (on Tesla V100 or RTX A6000) and the time costs in Tab.2 are all tested on RTX A6000.

6.1. Quantitative Evaluation

From Tab.2, compared with the 16 SOTA models, our IS-Net achieves the most competitive performance across all metrics. According to our observations, the performance of different models may be partially related to the model input size and the spatial size of their feature maps. As we know, most of the segmentation models introduce the existing image classification backbones to construct their encoder-decoder architectures for image segmentation tasks. However, some of the backbones like ResNet-50 [41] starts with an input convolution layer (stride of two) followed by a pooling operation (stride of two) to reduce the spatial size of the feature maps to a quarter of the input size, which leads to the loss of much spatial information and significant performance degradation. When the shape of the to-
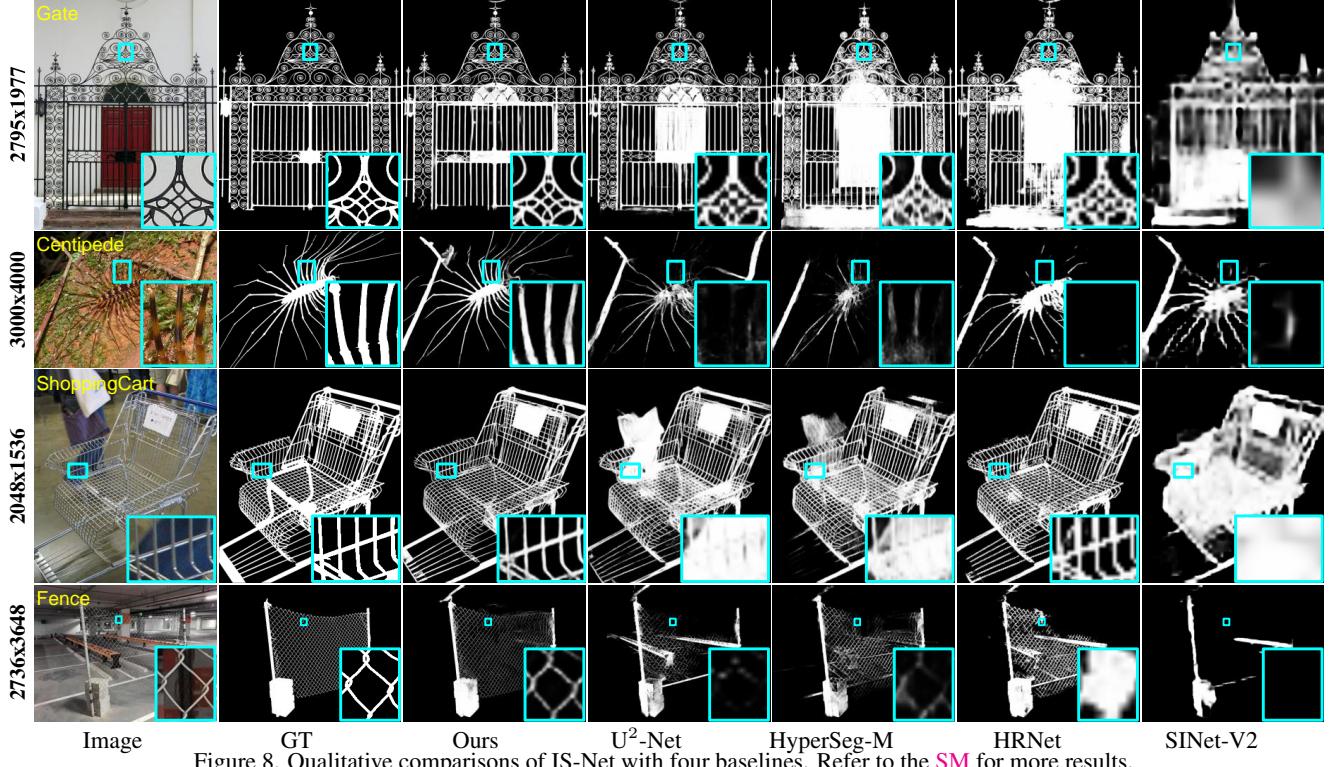
Figure 8. Qualitative comparisons of IS-Net with four baselines. Refer to the SM for more results.

be-segmented target is close to convex, the information lost and performance degradation is less significant. However, many objects in DIS5K are non-convex, and they have very complicated and fine structures. Therefore, DIS5K requires the models to keep the spatial information as much as possible, which is challenging to most models.

6.2. Qualitative Evaluation

Fig.8 presents qualitative comparisons between our approach and four SOTA baselines. Our model achieves promising results on the diverse scenes no matter that they are salient (gate), camouflaged (centipede), thin (shopping cart) or meticulous (fence) objects, demonstrating the generalization capability of our IS-Net baseline.

6.3. Ablation Study

To validate the effectiveness of our adaptation on recent SOTA model *e.g.*, U^2-Net and our newly proposed intermediate supervision strategy, we conduct comprehensive ablation studies.

Input Size. As can be seen in Tab.3, a larger input size can improve the performance of U^2-Net. However, it also increases the GPU memory costs so that we need to reduce the batch size (3 on Tesla V100, 32 GB) when the input size is 1024×1024, which degrades the performance. Our simple and effective variant (*i.e.*, Adp, 4rd row) addresses this memory issue and improves the performance.

Supervision on Different Decoder Stages. In Tab.3, Last-

Table 3. Ablation studies on our DIS-VD set.

Settings	$F_\beta^{mx} \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi^m \uparrow$	$HCE_\gamma \downarrow$
U^2-Net 320^2 (baseline)	.748	.656	.090	.781	.823	1413
U^2-Net 512^2	.769	.677	.085	.789	.826	1146
U^2-Net 1024^2	.764	.667	.088	.792	.820	1085
U^2-Net 1024^2 (**Adp**)	**.776**	**.695**	**.080**	**.804**	**.844**	1076
Adp+Last-1(L_2)	.777	.695	.080	.799	.840	1115
Adp+Last-2(L_2)	.778	.704	.079	.803	.847	**1049**
Adp+Last-3(L_2)	.788	.708	.079	**.812**	.845	1078
Adp+Last-4(L_2)	.782	.703	.079	.807	.849	1063
Adp+Last-5(L_2)	.788	**.715**	**.074**	.811	**.853**	1059
Adp+Last-6(L_2)	**.790**	.710	**.074**	.810	.852	1056
Adp+Last-6(KL)	.770	.684	.084	.794	.837	1092
Adp+Last-6(L_1)	.770	.686	.080	.797	.837	1144
Adp+Last-6(L_2) (shared outconv)	.745	.646	.094	.779	.813	1191
Adp+Last-6(L_2,sd(1))	.786	.706	.076	.807	.844	1086
Adp+Last-6(L_2,sd(58))	.790	.709	.078	.812	.848	**1085**
Adp+Last-6(L_2,sd(472))	.790	.712	.075	.812	.852	1071
Adp+Last-6(L_2,sd(5289)) (**IS-Net**)	**.791**	**.717**	**.074**	**.813**	**.856**	1116

S means the intermediate supervision is applied on the last S decoder stages. As shown, applying intermediate supervisions on the Last-6 stage gives relatively better performance, which is used as our default setting.

Different Loss. The results of different losses show that L_2 is better than KL divergence and $L1$. Besides, sharing the "outconvs", which transform the deep feature maps to the segmentation probability maps, of the GT encoders and the segmentation decoders leads to negative impacts.

Random Seeds. To study the influences of random weights initialization, we trained the same GT encoder multiple times with weights initialized by different random seeds.

As seen, although the performance produced by different random seeds are different, their variations are minor, and all of them are better than that of the models (U^2-Net and Adp) trained without our intermediate supervision strategy. Since the model from seed 5289 ranks the 1st on five out of six overall metrics, we use this model as our IS-Net.

7. Conclusions

We have systematically studied the highly accurate dichotomous image segmentation (DIS) task from both the application and the research perspective. To prove that the task is solvable, we have built a new challenging **DIS5K** dataset, introduced a simple and effective intermediate supervision network, called IS-Net, to achieve high-quality segmentation results in real-time, and designed a novel Human Correction Efforts (**HCE**) metric by considering the shape complexities for applications. With an extensive ablation study and comprehensive benchmarking, we obtained that our newly formulated DIS task is solvable.

Broader impacts. This work may greatly facilitate the applications of segmentation techniques in both academia and industry, and hereby invite all the researchers in related fields to collaborate and improve the whole eco-system.

References

[1] Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index. Accessed: 2021-09-21. 2

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 8, 9, 17

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 2, 16

[4] T. Birsan and D. Tiba. One hundred years since the introduction of the set distance by dimitrie pompeiu. In *IFIP SMO*, 2005. 2

[5] H. Blumberg. *Hausdorff's Grundzüge der Mengenlehre*. Bulletin of the American Mathematical Society, 27 (3): 116–129, American, 1920. 2

[6] G. Borgefors. Distance transformations in digital images. *Comput. Vis. Graph. Image Process.*, 34(3):344–371, 1986. 29

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 9, 25, 26

[8] S. Chen, X. Ma, Y. Lu, and D. Hsu. Ab initio particle-based object manipulation. In D. A. Shell, M. Toussaint, and M. A. Hsieh, editors, *RSS*, 2021. 2

[9] S. Chen, X. Tan, B. Wang, and X. Hu. Reverse attention for salient object detection. In *ECCV*, 2018. 16

[10] Z. Chen, Q. Xu, R. Cong, and Q. Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, 2020. 9, 16, 25, 26

[11] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 2

[12] B. Cheng, R. B. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 8, 17

[13] H. K. Cheng, J. Chung, Y.-W. Tai, and C.-K. Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, 2020. 4, 16

[14] M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 2, 4, 16

[15] N. Chinchor. MUC-4 evaluation metrics. In *MUC*, 1992. 2

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 16

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[18] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng. R3net: Recurrent residual refinement network for saliency detection. In *IJCAI*, 2018. 16

[19] M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In *K-CapW*, 2005. 2, 8, 17

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 2, 4, 16

[21] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018. 4, 16, 17

[22] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017. 2, 9, 17

[23] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018. 2, 9

[24] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE TPAMI*, 2021. 2, 9, 25, 26

[25] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 6, 2021. 2, 9

[26] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *CVPR*, 2020. 2, 4, 5, 16, 17

[27] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. Rethinking bisenet for real-time semantic segmentation. In *CVPR*, 2021. 2, 9, 16, 25, 26

[28] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. Rethinking bisenet for real-time semantic segmentation. In *CVPR*, 2021. 9

[29] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. *Theory Comput.*, 8(1):415–428, 2012. 29

[30] C. Fiorio and J. Gustedt. Two linear time union-find strategies for image processing. *TCS*, 154(2):165–181, 1996. 8

[31] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí. Yet another survey on image segmentation: Region and boundary information integration. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *ECCV*, 2002. 2, 17

[32] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2019. 9

[33] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[34] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[35] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 20

[36] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10):1915–1926, 2012. 1

[37] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 2, 17

[38] R. M. Haralick, S. R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *IEEE TPAMI*, PAMI-9(4):532–550, 1987. 8, 29

[39] F. Hausdorff. *Grundzüge der Mengenlehre*. Leipzig: Veit, ISBN 978-0-8284-0061-9Reprinted by Chelsea Publishing Company in 1949, Germany, 1914. 2

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 16

[41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 9

[42] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017. 16

[43] A. Howard, R. Pang, H. Adam, Q. V. Le, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, and Y. Zhu. Searching for mobilenetv3. In *ECCV*, 2019. 9, 25, 26

[44] P. Hu, F. Caba, O. Wang, Z. Lin, S. Sclaroff, and F. Perazzi. Temporally distributed networks for fast video semantic segmentation. In *CVPR*, 2020. 2, 16

[45] Z. Ke, K. Li, Y. Zhou, Q. Wu, X. Mao, Q. Yan, and R. W. Lau. Is a green screen really necessary for real-time portrait matting? *ArXiv*, abs/2011.11961, 2020. 2

[46] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 20

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[48] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019. 2, 4, 16

[49] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 2, 6, 17

[50] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015. 4

[51] H. Li, P. Xiong, H. Fan, and J. Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *CVPR*, 2019. 2, 16

[52] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 4, 16

[53] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia. Fully convolutional networks for panoptic segmentation. In *CVPR*, 2021. 16

[54] J. H. Liew, S. Cohen, B. Price, L. Mai, and J. Feng. Deep interactive thin object selection. In *WACV*, 2021. 2, 4, 5, 16, 17

[55] S. Lin, L. Yang, I. Saleemi, and S. Sengupta. Robust high-resolution video matting with temporal guidance. *CoRR*, abs/2108.11515, 2021. 2, 16

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 16

[57] F. Liu, L. Tran, and X. Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *CVPR*, 2021. 1

[58] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han. Visual saliency transformer. In *ICCV*, 2021. 16

[59] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011. 4, 16

[60] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 16

[61] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. 2, 17

[62] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 16

[63] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, 2021. 4

[64] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps. *CVPR*, 2014. 2, 9, 29

[65] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE TPAMI*, 26(5):530–549, 2004. 2

[66] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021. 9, 24, 25, 26

[67] V. Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 2

[68] V. Mnih and G. E. Hinton. Learning to detect roads in high-resolution aerial images. In *ECCV*, 2010. 2

[69] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, 2010. 2, 4, 16

[70] Y. Nirkin, L. Wolf, and T. Hassner. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. *arXiv preprint arXiv:2012.11582*, 2020. 2, 9, 16, 24, 25, 26

[71] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *CVPR*, 2019. 2, 16

[72] R. Osserman. The isoperimetric inequality. *BAM*, 84(6):1182–1238, 1978. 4

[73] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 2, 8, 9, 17

[74] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 4, 16

[75] L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, Z. Lin, P. Torr, and J. Jia. Open-world entity segmentation. *arXiv preprint arXiv:2107.14228*, 2021. 2, 16

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

[77] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020. 2, 6, 9, 16, 17, 24, 25, 26

[78] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 2, 8, 9, 16, 17, 24, 25, 26

[79] U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *CGIP*, 1(3):244–256, 1972. 4, 5, 8

[80] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 1

[81] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2, 9, 16, 24, 25, 26

[82] S. Saito, T. Yamashita, and Y. Aoki. Multiple object extraction from aerial imagery with convolutional neural networks. *EI*, 2016(10):1–9, 2016. 2

[83] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs. Automatic portrait segmentation for image stylization. In *CGF*, 2016. 2

[84] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 1

[85] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished Manuscript*, 2018. 2, 4, 16

[86] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 2, 17

[87] S. Suzuki and K. Abe. Topological structural analysis of digitized binary images by border following. *CVGIP*, 30(1):32–46, 1985. 8

[88] T. J. Sørensen. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. København, I kommission hos E. Munksgaard, Denmark, 1948. 2

[89] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 9

[90] L. Tang, B. Li, Y. Zhong, S. Ding, and M. Song. Disentangled high quality salient object detection. In *ICCV*, 2021. 2, 16

[91] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 9

[92] C. J. van Rijsbergen. Information retrieval. London:Butterworths, 1979.http://www.dcs.gla.ac.uk/Keith/Preface.html, 1979. 2

[93] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2019. 9, 16, 24, 25, 26

[94] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 4, 16, 17

[95] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stage-wise refinement model for detecting salient objects in images. In *ICCV*, 2017. 16

[96] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, 2018. 2, 16

[97] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE TPAMI*, 2021. 16

[98] A. B. Watson. Perimetric complexity of binary digital images. *Math J*, 14:1–40, 2012. 4

[99] J. Wei, S. Wang, and Q. Huang. F³net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020. 9, 16, 25, 26

[100] X. Wei, X. Li, W. Liu, L. Zhang, D. Cheng, H. Ji, W. Zhang, and K. Yuan. Building outline extraction directly using the u2-net semantic segmentation model from high-resolution aerial images and a comparison study. *RS*, 13(16):3187, 2021. 2

[101] K. Wu, E. J. Otoo, and A. Shoshani. Optimizing connected component labeling algorithms. In J. M. Fitzpatrick and J. M. Reinhardt, editors, *MI*, 2005. 8

[102] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015. 2, 6, 17

[103] N. Xu, B. Price, S. Cohen, and T. Huang. Deep image matting. In *CVPR*, 2017. 6

[104] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013. 4, 16

[105] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Lin, and A. Yuille. Meticulous object segmentation. *arXiv preprint arXiv:2012.07181*, 2020. 2, 4, 16, 17

[106] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 4, 16

[107] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 2, 9, 16, 25, 26

[108] H. Yu, N. Xu, Z. Huang, Y. Zhou, and H. Shi. High-resolution deep image matting. *arXiv preprint arXiv:2009.06613*, 2020. 6

[109] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu. Towards high-resolution salient object detection. In *CVPR*, pages 7234–7243, 2019. 2, 4, 16

[110] P. Zhang, W. Liu, H. Lu, and C. Shen. Salient object detection by lossless feature reflection. In *IJCAI*, 2018. 16

[111] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017. 16

[112] T. Y. Zhang and C. Y. Suen. A fast parallel algorithm for thinning digital patterns. *Commun. ACM*, 27(3):236–239, 1984. 8, 29

[113] Z. Zhang, Q. Liu, and Y. Wang. Road extraction by deep residual u-net. *GRSL*, 15(5):749–753, 2018. 2

[114] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018. 2, 9, 16, 25, 26

[115] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 9, 25, 26

[116] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, 2019. 2, 17

[117] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, 2020. 9, 16, 25, 26

[118] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2, 16

[119] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017. 3

[120] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 2, 16

Highly Accurate Dichotomous Image Segmentation

-Supplementary Material

(a) High-resolution image with cluttered background

(b) Artistic figure based on the background removed image

Figure 9. Demo application: artistic figure generated based on a sample of our DIS5K dataset.

1. Related Work

1.1. Multi-class *vs.* Dichotomous Segmentation

Multi-class (*e.g.* semantic [60], panoptic [53]) segmentation aims at simultaneously labeling all the pixels in an image of complex scenario [16, 120], 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 [40]. 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 [56] or panoptic [120] 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) [14, 59, 69, 94, 97, 104, 106], salient object in clutter (SOC) [21], high-resolution salient object detection (HRS) [109], camouflaged object detection (COD) [26, 48, 85], thin object segmentation (TOS) [54], meticulous object segmentation (MOS) [105], video object segmentation (VOS) [74], class-agnostic very high-resolution segmentation (VHRS) [13], *etc*. Most of these tasks try to solve dichotomous segmentation problems on images which are sharing specific characteristics. The 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 [20], MS-COCO [56]) and panoptic (Cityscapes [16], ADE20K [120]) 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) [75] datasets proposed for class-agnostic segmentation has similar issues. Images in the salient object detection (SOD) [14, 52, 69, 94, 106] and camouflaged object detection (COD) [26] datasets are usually low-resolution ones, which contains objects with simple structures. The high-resolution salient object detection (HRS) [74, 109] 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) [105] and thin object segmentation (TOS) [54] 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 [60] feature aggregation models [9, 42, 62, 93, 99, 110, 111, 117], Encoder-Decoder architectures [3, 10, 77, 81], Coarse-to-Fine (or Predict-Refine) models [13, 18, 55, 78, 90, 95, 96], Vision Transformers [58, 118], *etc*. Besides, many real-time models [27, 44, 51, 70, 71, 107, 114] 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 [37], dropout [86], dense (deep) supervision [49, 77, 102], hybrid loss [61, 78, 116] and so on, have been proposed. The dense (deep) supervision [49, 77, 102], 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 [12], F-measure [2], boundary F-measure [19, 78], boundary displacement error (BDE) [31], boundary IoU [12], structural measure (S_m) [22], Mean Absolute Error (MAE) [73], 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 segments. Besides, the above metrics are mostly de-

fined 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. Per-category and per-group statistics

Fig. 10 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.

2.2. Typical Samples from DIS5K

Fig. 11 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) [94], salient object in clutter (SOC) [21], camouflaged object detection (COD) [26], thin object segmentation (TOS) [54], meticulous object segmentation (MOS) [105]. 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 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. 11 (a) and the shrimp in Fig. 11 (c) can also be taken as meticulous because the basket has many holes and the shrimp has jagged boundaries. Besides, 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.

2.3. 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. 12.
Intra-category structure similarity. As shown in Fig. 12 (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. 12 (c) and (d) show two typical examples against that in different magnitudes. Fig. 12 (c) illustrates some bicycles with
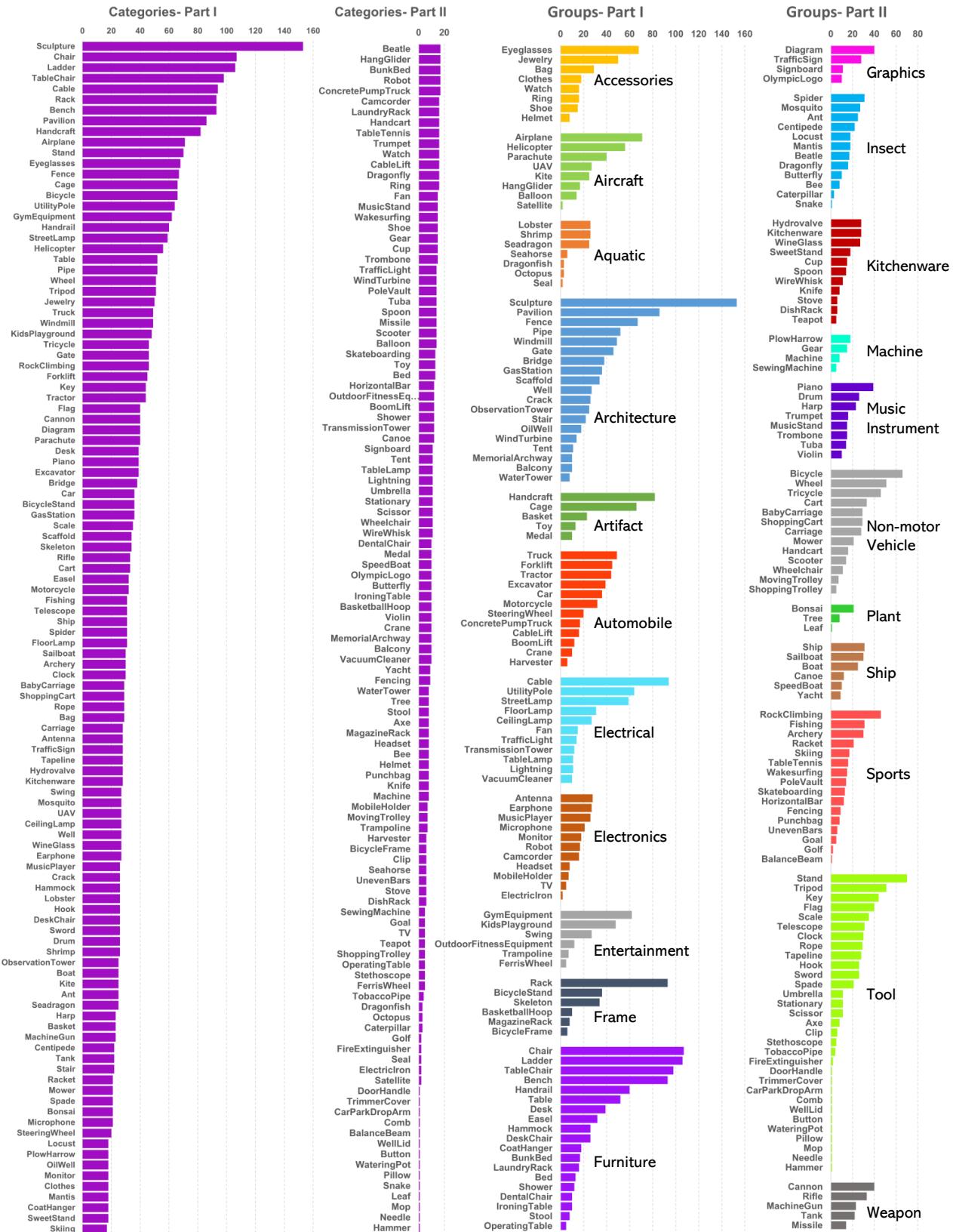
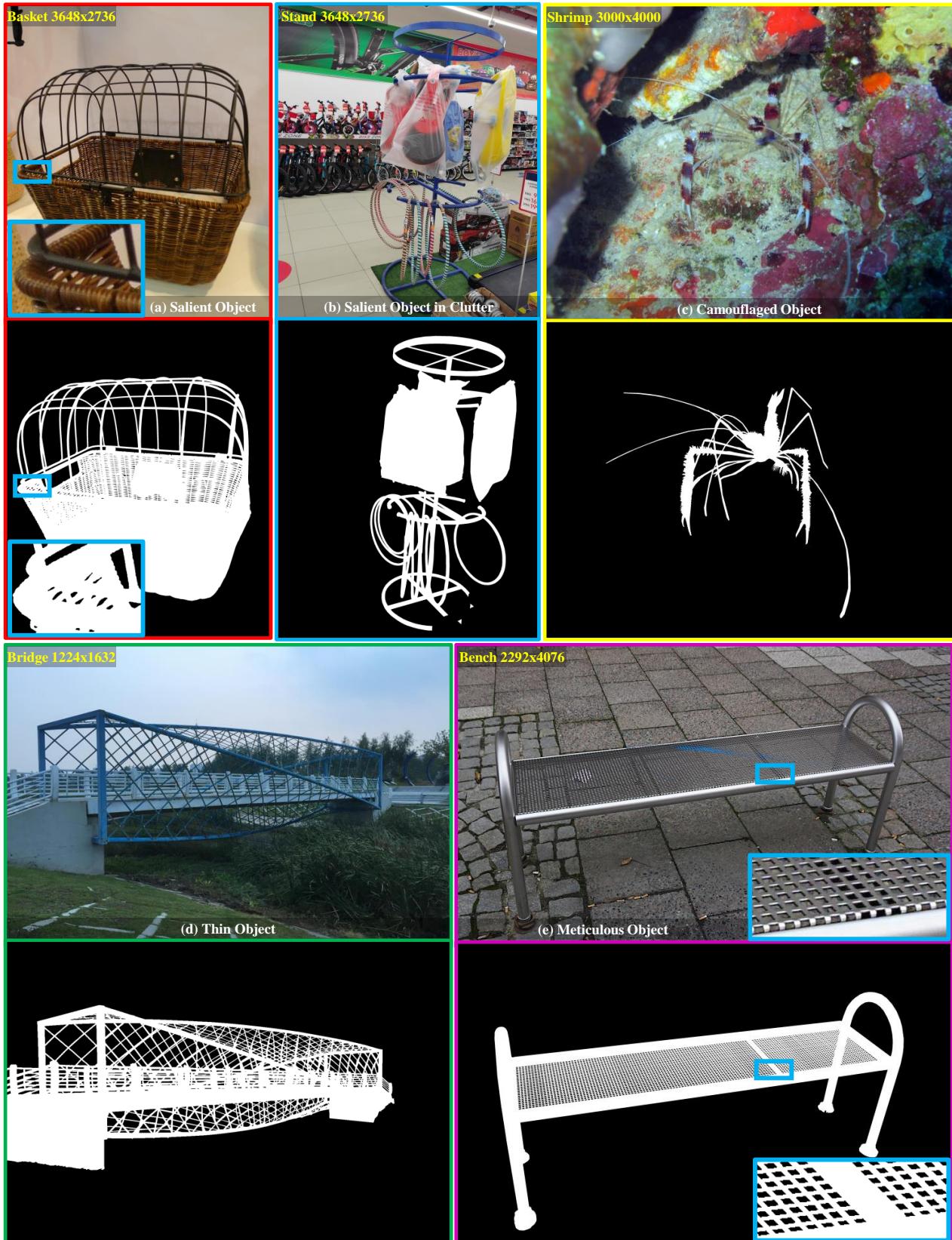Figure 10. Number of images per-category and per-group.

Figure 11. Sample images and ground truth masks with objects of certain characteristics.

Table 4. Image dimension and object complexity of the subsets of DIS5K. $\sigma_{(.)}$ 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 ± 988.85	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

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. 12 (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. 12 (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 quantitative studies will be helpful to diversified vision tasks, such as image (shape) classification, segmentation, *etc*.

2.4. Attributes of Subsets in DIS5K

Table 4 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. 13 shows the qualitative comparisons of the structural complexities of our four 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 [35] and optimized with Adam [46] optimizer with the default settings

(a) Objects with stable and less variant structures   (b) Objects sharing common structures

(c) Objects with different structures from the SAME category

(d) Objects with very different structures from the SAME category

(e) Targets with similar structure characteristics across DIFFERENT categories

Figure 12. Structure analysis of inter- and intra-category targets.

(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 con-

Figure 13. Sample ground truth (GT) masks from DIS-TE1, DIS-TE2, DIS-TE3, and DIS-TE4.

Figure 14. Qualitative comparisons of our model and four cutting-edge baselines.

ducted 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 using Pytorch 1.8.0. Some experiments are conducted on a desktop that has a 2.9GHz 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).

Figure 15. Curves of the training loss computed on the last prediction probability map and the Mean Absolute Error (M) on our validation set (DIS-VD).

3.2. More Analysis of the Experimental Results

Performance comparisons among different models. As shown in Table 2, 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 [81], BASNet [78], U^2-Net [77] and PFNet [66] performs relatively better against other SOD and COD models. Among the semantic segmentation and real-time semantic segmentation models, the results of HRNet [93] and HyperSeg-M [70] show more competitive performance. Among all the existing models, the performance 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. 14 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.

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 Table 2, 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 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 Table 4. 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 Table 5 and 6, from which the per-group segmentation difficulties and performance can be found.

Effectiveness of Our Intermediate Supervision To fur-

Table 5. 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 [81]	BASNet [78]	GateNet [117]	F³Net [99]	GCPANet [10]	U²Net [77]	SINetV2 [24]	PFNet [66]	PSPNet [115]	DLV3+ [7]	HRNet [93]	BSV1 [107]	ICNet [114]	MBV3 [43]	STDC [27]	HySM [70]	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

Table 6. 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.

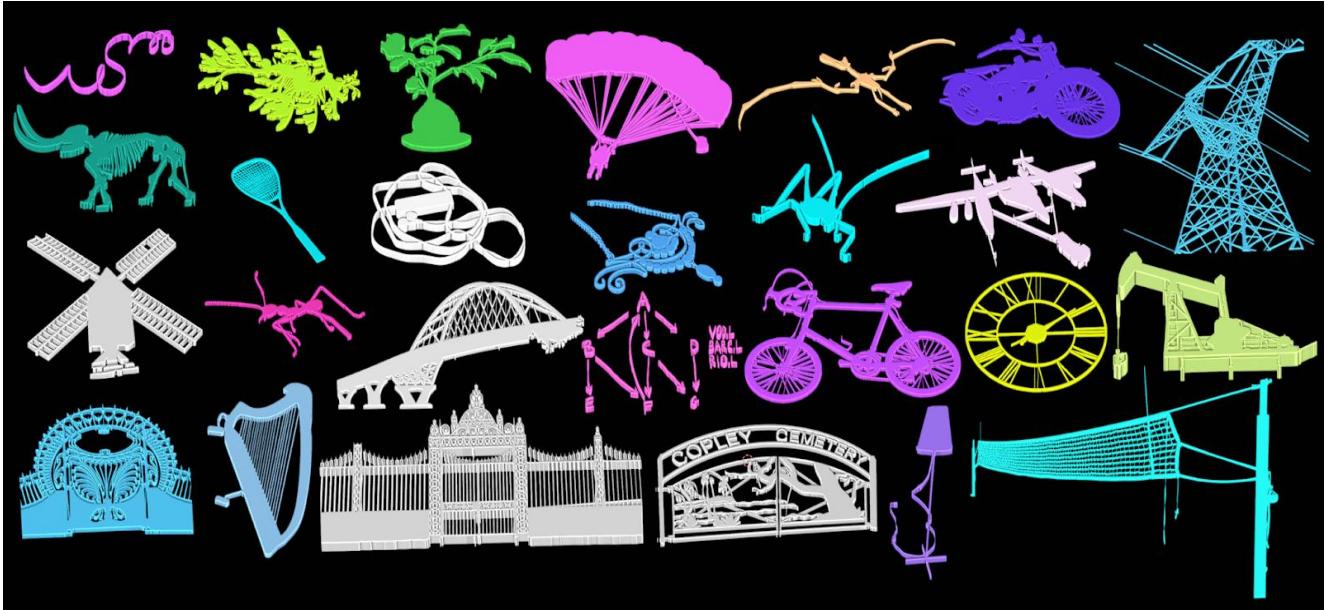| Dataset	Metric	UNet [81]	BASNet [78]	GateNet [117]	F³Net [99]	GCPANet [10]	U²Net [77]	SINetV2 [24]	PFNet [66]	PSPNet [115]	DLV3+ [7]	HRNet [93]	BSV1 [107]	ICNet [114]	MBV3 [43]	STDC [27]	HySM [70]	Ours
12 Graphics	$maxF_\beta$ ↑	0.750	0.719	0.685	0.663	0.524	0.746	0.568	0.645	0.646	0.621	0.671	0.575	0.681	0.616	0.647	0.732	0.780
	F^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	670	976	1009	1268	1403	938	1423	1294	1447	1201	990	1425	1122	1457	1331	824	621
13 Insect	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	570	595	604	656	683	592	701	663	714	636	622	700	609	713	667	574	488
14 Kitchenware	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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	0.797
	HCE_γ ↓	541	536	554	574	579	536	602	583	588	543	608	637	540	608	571	484	367
15 Machine	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	1544	1687	1728	1846	1849	1693	1910	1860	1925	1787	1937	1987	1799	1957	1899	1589	1322
16 Music Instrument	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	671	683	653	693	708	705	735	713	732	678	791	796	687	771	748	598	492
17 Non-motor Vehicle	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	1956	2098	2134	2219	2217	2121	2269	2293	2274	2169	2314	2319	2161	2334	2245	1971	1623
18 Plant	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	9194	9174	10036	10164	10488	9062	10268	10137	10231	9910	9615	10444	9798	10309	10230	8334	8563
19 Ship	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	3193	3341	3233	3242	3225	3355	3183	3265	3189	3178	3443	3454	3134	3381	3334	3046	2951
20 Sports	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	0.807	0.822	0.805	0.825	0.777	0.832	0.821	0.820	0.820	0.803	0.831	0.801	0.816	0.868	0.836	0.838	0.860
	HCE_γ ↓	1137	1283	1274	1329	1247	1315	1274	1355	1323	1297	1450	1401	1306	1352	1343	1180	934
21 Tool	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	568	589	620	659	673	602	689	673	689	632	662	724	625	707	660	554	433
22 Weapon	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	793	826	849	888	899	845	923	902	928	864	914	969	861	937	899	779	649
All VD+TE(1-4)	$maxF_\beta$ ↑	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^w_β ↑	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 ↓	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_α ↑	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^m_ϕ ↑	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_γ ↓	1228	1330	1368	1433	1425	1348	1441	1461	1470	1394	1457	1541	1387	1489	1459	1239	1035

Figure 16. 3D models built upon the ground truth masks sampled from DIS5K by the "Extrude" operation in Blender.

ther 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.15. The top part of Fig.15 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.15 shows that our intermediate 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. 9 shows the comparison between the original ship image with cluttered background and the background-removed image with perspective transforms (See more samples in Fig. 17). 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:

- materials of art design, image and video editing;

- backgrounds of posters or slides, wall papers of cell-phones, tablets, desktops;

- materials for 3D modeling, as shown in Fig. 16 (A demo video is also attached).

5. Limitations and Future Works

Failure Cases of Our Model. Fig.18 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

Figure 17. Comparisons between the original images and their backgrounds-removed correspondences generated from our DIS5K.
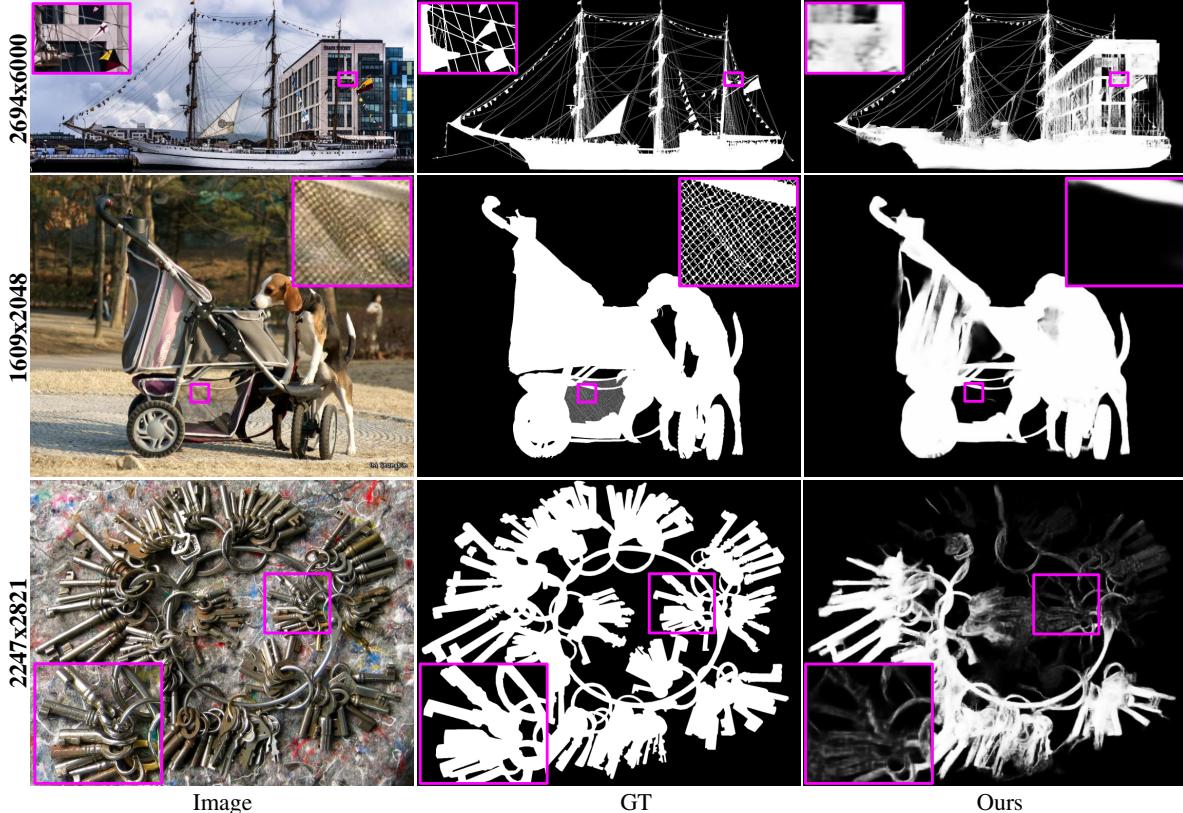
Figure 18. Typical failure cases.

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 dichotomous image segmentation (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 [38] and dilation [38] operations are used to remove small false positive and false negative regions, while the skeleton extraction algorithm [112] 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 [64], which uses a distance transform algorithm [6,29] 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.