Highly Accurate Dichotomous Image Segmentation

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

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

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 DIS 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. DIS is annotated with extremely fine-grained labels. Besides, we introduce a simple intermediate supervision baseline (**IS-Net**) using both feature-level and mask-level guidance for DIS model training. IS-Net outperforms various cutting-edge baselines on the proposed DIS5K, making it a general self-learned supervision network that can 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 complexities, 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. Project page: https://xuebinqin.github.io/dis/index.html.

Keywords: Dichotomous Image Segmentation, High Resolution, Metric

1 Introduction

Currently, the annotation accuracy of 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

* Corresponding author (**Deng-Ping Fan**). We would like to thank Jiayi Zhu for his efforts in re-organizing the dataset and codes.

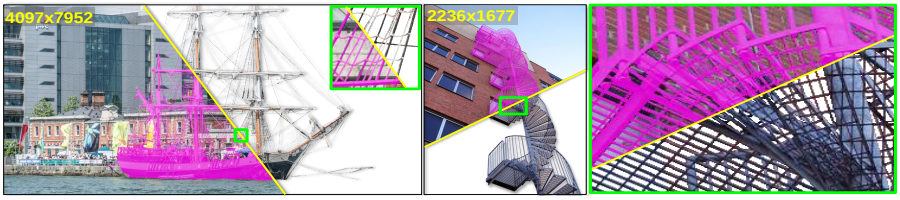


Fig. 1: Sample images from our DIS5K dataset. Zoom-in for best view.

demanding highly accurate outputs from computer vision algorithms to support delicate human-machine interaction. Compared with classification [15, 39, 73] and detection [29,30,68], segmentation can provide more geometrically accurate target descriptions for wide applications, *e.g.*, image editing [31], AR/VR [64], medical image analysis [70], robot manipulation [7], *etc.*

These applications can be grouped as "light" (*e.g.*, image editing and analysis) and "heavy" (*e.g.*, human-machine interaction), based on their immediate affects on real-world objects. The "light" ones (Fig.1), which usually allows post-corrections, are relatively tolerant to the segmentation errors. While, in the "heavy" ones, the segmentation deflects or failures are more likely to cause physic damages on objects or injuries (sometimes fatal) of humans. Hence, *highly accurate* and *robust* models are needed. Now, most of the segmentation models are still less applicable in those "heavy" applications due to the accuracy and robustness issues. Hence, **our goal** is to address the "heavy" and "light" applications in a general framework, called *dichotomous image segmentation (DIS)*, which aims to segment highly accurate objects.

Existing segmentation tasks mainly focus on objects with specific characteristics, *e.g.*, salient [79, 82, 94], camouflaged [23, 40, 74], meticulous [45, 90] or specific categories [38, 46, 55, 70, 72]. They have the same input/output formats, and the exclusive mechanisms are barely used for segmenting specific targets in their models, which means they are usually 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 [14,17,47,63,103], the proposed DIS task mainly focuses on images with single or a few targets, from which getting richer accurate details of each target is more feasible. Therefore, we provide four **contributions**:

i) A large-scale, extendable DIS dataset, **DIS5K**, contains 5,470 high-resolution images paired with highly accurate binary segmentation masks.
ii) A novel baseline **IS-Net** built with intermediate supervision reduces overfitting by enforcing direct high-dimensional feature synchronization.
iii) 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.

iv) 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 performance.

2 Related Work

Tasks and Datasets of image segmentation are closely related in deep learning era. Some of the segmentation tasks like [12,21,45,46,55,72,82,90], 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 restricts their trained models 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. **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 [49], Encoder-Decoder [3,70], Coarse-to-Fine [83], Predict-Refine [66,79], Vision Transformer [101] and so on are developed. Besides, many real-time models are designed [24,37,43,58,59,92,97] to balance the performance and the time costs. Other methods, such as weights regularization [32], dropout [75], dense supervision [41,65,87], and hybrid loss [50,66,99], 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 multi-channel deep features to single-channel side outputs. **Evaluation Metrics** can be categorized as *region-based* (*e.g.*, IoU or Jaccard index [1], F-measure [13,69] or Dice's coefficient [77], weighted F-measure [52]), *boundary-based* (*e.g.*, CM [57], boundary F-measure [16, 53, 56, 62, 66, 71, 96], boundary IoU [9], boundary displacement error (BDE) [27], Hausdorff distances [4,5,34]), *structure-based* (*e.g.*, S-measure [19], E-measure [20,22]), *confidence-based* (*e.g.*, MAE [61]), *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.

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[6] based on our pre-designed keywords[7]. Then, according to

[6] Images with the license of "Commercial use & mods allowed"

[7] Since the long-term goal of this research is to facilitate the "safe" and "efficient" interaction between the machines and our living/working environments, these key-

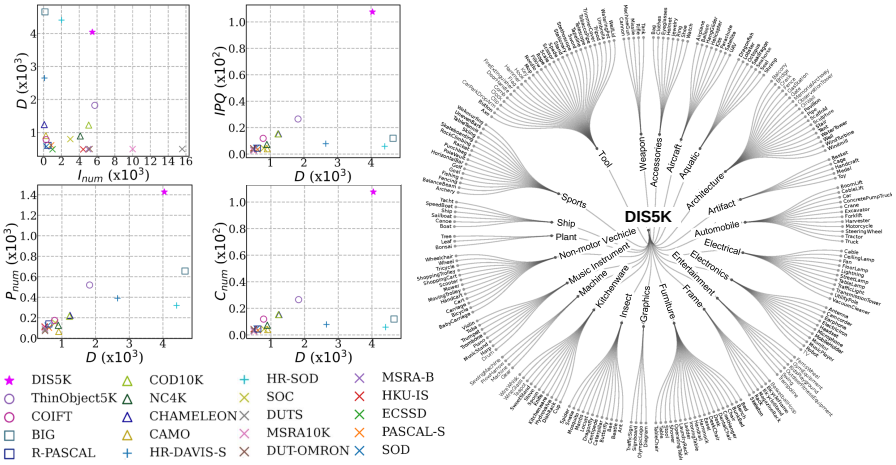


Fig. 2: **Left**: Correlations between different complexities. **Right**: Categories and groups of our DIS5K dataset. Zoom-in for better view. Please refer to §3.1 for details.

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.* [102]. 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.

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[8]. 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

words are mainly related to the common targets (*e.g.*, bicycle, chair, bag, cable, tree, *etc*) in our daily lives.

[8] `https://www.gimp.org/`

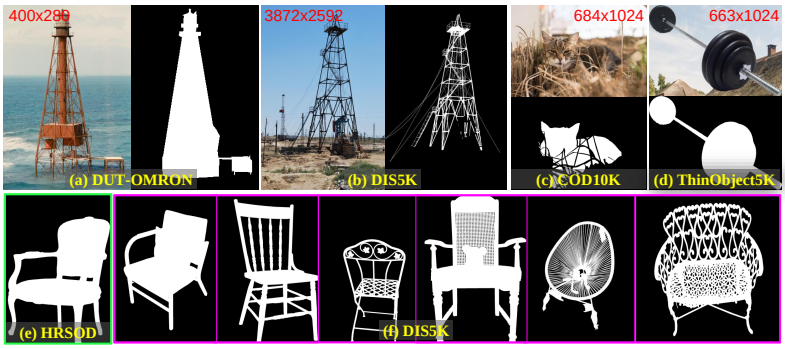


Fig. 3: Qualitative comparisons of different datasets. (a) and (b) indicate that our DIS5K provides more accurate labels. (c) shows one sample from COD10K [23], of which the structural complexity is caused by occlusion. (d) illustrates the synthetic ThinObject5K [45] dataset. (e) and (f) demonstrate that DIS5K has a larger diversity of intra-categorical structure complexities.

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, *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 [57], PASCAL-S [44], ECSSD [89], HKU-IS [42], MSRA-B [48], DUT-OMRON [91], MSRA10K [12], DUTS [82], and SOC [18]; (2) two high-resolution salient object detection (HR-SOD) datasets: HR-SOD [94] and HR-DAVIS-S [62, 94]; (3) four camouflaged object detection (COD) datasets: CAMO [40], CHAMELEON [74], COD10K [23], and NC4K [51]; (4) two semantic segmen-

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 [57]	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 [44]	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 [89]	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 [42]	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 [48]	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 [91]	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 [12]	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 [82]	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 [18]	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 [94]	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 [62]	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 [40]	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 [74]	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 [23]	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 [23]	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 [11]	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 [11]	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 [45]	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 [45]	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

tation (SMS)[9] datasets: R-PASCAL [11, 17] and BIG [11]; (5) two thin object segmentation (TOS) datasets: COIFT [45] and ThinObject5K [45]. 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 of its significant impacts on accuracy, efficiency, and computational costs. The mean (H, W, D) and 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 only contains 150 images. HR-SOD has slightly greater dimensions than ours, its complexity is low. The average dimensions of our DIS5K are almost eight times larger than those of the SOD and COD datasets. Besides, the targets in COD datasets are mainly animals and insects, which restricts its applications in diversified tasks.

Object Complexity is described by three metrics including the *isoperimetric inequality quotient* (IPQ) [60,84,90], 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 [67] 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.

[9] 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.

Discussion. Tab.1 and Fig.2 (Left) illustrate the computed metrics. Our DIS5K is around 20 (up to 50) times more complicated than the SOD datasets in terms of average IPQ. Although CHAMELEON, COD10K, BIG, COIFT, and ThinObject5K have higher average IPQ against the SOD datasets, they are still much less complicated than ours. The average C_{num} and its standard deviation of DIS5K are over 100 and 400. 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. 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. Refer to the SM for more results. 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 [88, 93]. But the synthesized images are usually different from the real ones, which leads to biases in predictions. Fig.3-e & f demonstrate the larger diversity of intra-categorical structure complexities of our DIS5K.

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 (*i.e.*, DIS-TE1∼DIS-TE4) with 500 images in each subset to represent four testing difficulty levels.

4 Proposed IS-Net Baseline

Overview. As shown in Fig.4, 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

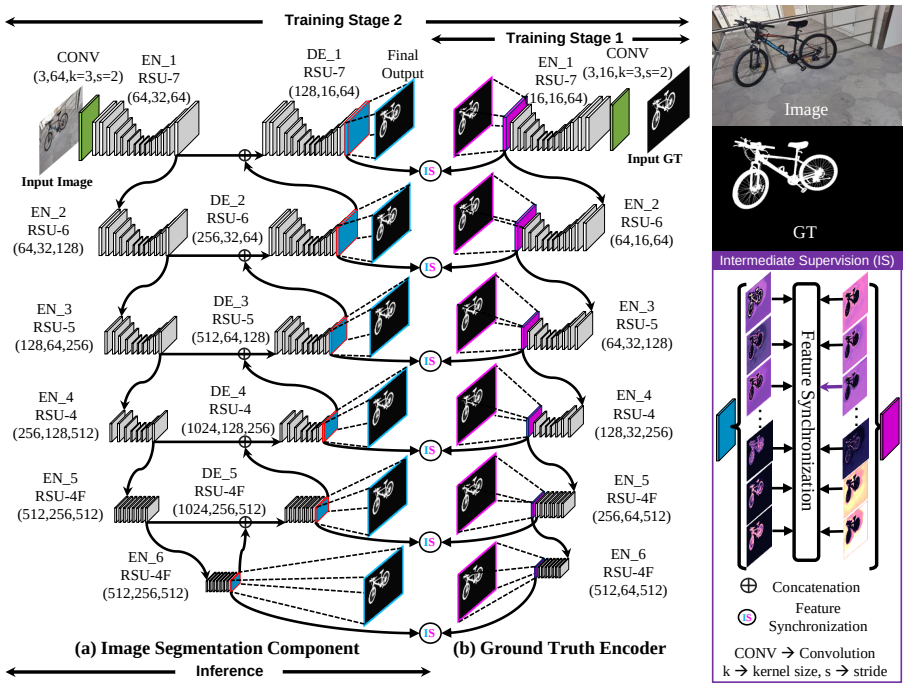


Fig. 4: Our IS-Net: (a) shows the image segmentation component, (b) illustrates the ground truth encoder built upon the intermediate supervision (IS) component.

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 [65] 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 $I^{1024\times1024\times3}$, the input convolution layer first transforms it to a feature map $f^{512\times512\times64}$ 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 $I^{1024\times1024\times3}$ 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.

4.1 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 [41]. In [65, 87], 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 by "over-fitting" the training ground truth using a lightweight deep model F_{gt}, Fig.4-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.4-a), the image I is transformed to a set 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} represents the hyperparameter to weight each side output loss.

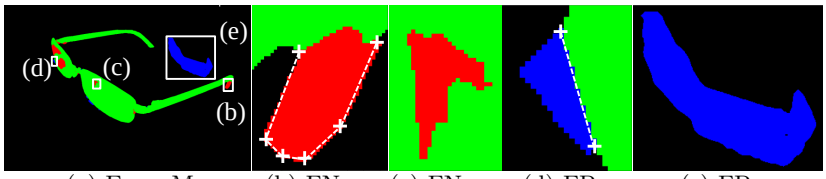
Fig.5 illustrates the feature maps from the stage 2 in Fig.4, EN_2, of the GT encoder. We can see the diversified characteristics of the input mask are encoded into different channels. For example, the 21^{st} channel encodes both the fine and large structures close to the original mask. While the 23^{rd}, 29^{th}, and 37^{th} channels encode the middle size structures (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.

Fig. 5: 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.

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 [10], F-measure [2], boundary F-measure [16, 66], and MAE [61], usually evaluate the quality of the prediction P by calculating the scores based on the mathematical or cognitive consistency (or inconsistency) between P and G. In other words, these metrics describe how significant the "gap" is between P and G. However, evaluating the costs of filling the "gap" is more important than measuring the magnitude of the "gap" in many applications.

Therefore, we propose a novel evaluation metric, Human Correction Efforts (HCE), which measures 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 clicking. Therefore, the HCE here is quantified by the number of mouse clicking. 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.6, the FNs and FPs can be categorized into two classes, according to their adjacent regions: $\mathrm{FN_N}$ (N=TN+FP), $\mathrm{FN_{TP}}$, $\mathrm{FP_P}$ (P=TP+FN) and $\mathrm{FP_{TN}}$. To correct the $\mathrm{FN_N}$ regions, its boundaries adjacent to the TN need to be manually labeled with dominant points (Fig.6-b). Similarly, to correct the $\mathrm{FP_P}$ regions, we only need to label its boundaries adjacent to the TP regions (Fig.6-d). The $\mathrm{FN_{TP}}$ regions (Fig.6-c) enclosed by TP and the $\mathrm{FP_{TN}}$ regions (Fig.6-e) enclosed by TN can be easily corrected by one-click region selection. Therefore, the HCE for correcting the faulty regions in Fig.6 (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 algorithm [67] based on the contours obtained by OpenCV findContours [76] function and the connected regions labeling algorithm [26, 86], respectively, in the evaluation stage.



(a) Error Map (b) FN_N (c) FN_{TP} (d) FP_P (e) FP_{TN}

Fig. 6: Faulty regions to be corrected. Refer to §5 for details.

Input: P, G, $\gamma = 5$, $epsilon = 2.0$
Output: HCE_γ

G_{ske} = skeletonize (G);
$P_{or}G$, TP = or (P, G), and (P,G);
FN, FP= (G - TP), (P - TP);
for $(i = 0; i \leq \gamma; i++)$ **do**
  $P_{or}G$ = erode ($P_{or}G$, $disk$ (1));
end
FN', FP' = and (FN,$P_{or}G$), and (FP,$P_{or}G$);
for $(i = 0; i \leq \gamma; i++)$ **do**
  FN' = dilate (FN', $disk$ (1));
  FN' = and (FN', not P);
  FP' = dilate (FP', $disk$ (1));
  FP' = and (FP', not G);
end
FN', FP' = and (FN, FN'), and (FP, FP');
FN' = or (FN', xor (G_{ske}, and (TP, G_{ske})));
HCE_γ = compute_HCE (FN', FP', TP, epsilon)

Algorithm 1: Relax HCE.

Relax HCE. Some applications may be tolerant to certain minor prediction errors. Therefore, the HCE is extended 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 [33] and dilation [33]. Given a segmentation map P, its 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 HCE_γ is calculated as Alg. 1. Note that the erosion operation can remove all the thin and fine components of $P_{or}G$. However, some thin components (*e.g.*, thin cables, nets) are critical in describing the targets, and they need to be retained. To this end, the skeleton of the GT mask is extracted by [95] 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' 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

Table 2: Quantitative evaluation on DIS5K validation and test sets. R = ResNet [35]. R2 = Res2Net [28]. S-813 = STDC813 [24], E-B1 = EffinetB1 [78].

Dataset	Metric	UNet [70]	BASNet [66]	GateNet [100]	F^3Net [85]	GCPANet [8]	U^2Net [65]	SINetV2 [21]	PFNet [54]	PSPNet [98]	DLV3+ [6]	HRNet [81]	BSV1 [92]	ICNet [97]	MBV3 [36]	STDC [24]	HySM [58]	**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	1024x2048	1024x2048	1024^2	512x1024	512x1024	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**

comparison. Currently, cross-dataset evaluations [80] 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 ($F_\beta^w \uparrow$) [52], mean absolute error ($M \downarrow$) [61], structural measure ($S_\alpha \uparrow$) [19], mean enhanced alignment measure ($E_\phi^m \uparrow$) [20, 22] and our human correction efforts ($HCE\gamma \downarrow$), are used to evaluate the performance from different perspectives.

Competitors. We compared our IS-Net with 16 popular models designed for different segmentation tasks, including (i) popular medical image segmentation model, U-Net [70]; (ii) salient object detection models such as BASNet [66], GateNet [100], F^3Net [85], GCPA [8] and U^2-Net [65]; (iii) models designed for COD like SINet-V2 [21] and PFNet [54]; (iv) semantic segmentation models: PSPNet [98], DeepLab-V3+ [6] and HRNet [81]; (v) real-time semantic segmentation models: BiSeNetV1 [92], ICNet [97], MobileNet-V3-Large [36], STDC [25] and HyperSegM [58]. 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

Compared with the 16 SOTA models, our IS-Net achieves the most competitive performance across all metrics (see Tab.2). We observe that the performance of

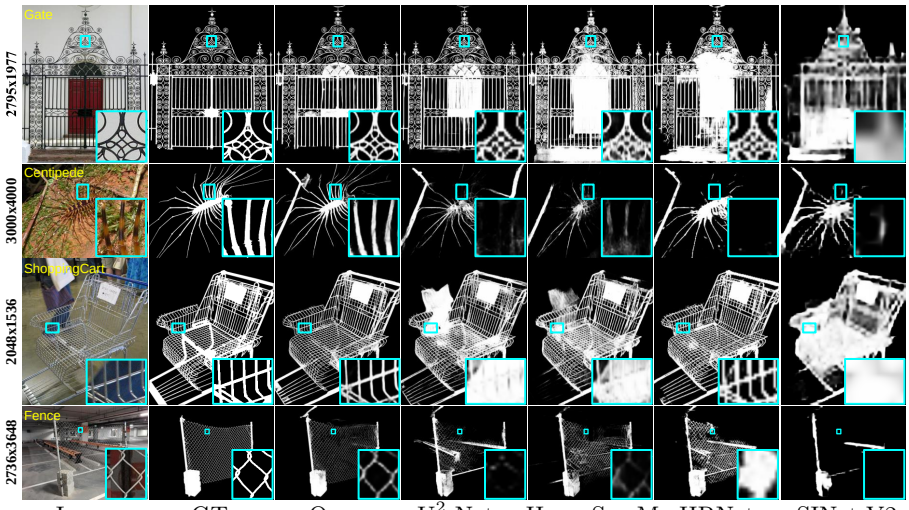


Fig. 7: Qualitative comparisons of IS-Net with four baselines.

different models may be partially related to the spatial size of the model input and their feature maps. Most of the segmentation models introduce the classification backbones to construct their encoder-decoder architectures. However, backbones like ResNet-50 [35] 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, which leads to the loss of much spatial information and significant performance degradation. When the shape of the to-be-segmented target is close to convex, the degradation is less significant. However, many objects in DIS5K are non-convex, and they have very complicated and fine structures. It requires the models to keep the spatial information as much as possible, which is challenging to most models.

6.2 Qualitative Evaluation

Fig.7 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

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

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, 4^{rd} row) addresses this memory issue and improves the performance.

Supervision on Different Decoder Stages. In Tab.3, Last-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 1^{st} 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.

References

1. Jaccard index. https://en.wikipedia.org/wiki/Jaccard_index, accessed: 2021-09-21
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39**(12), 2481–2495 (2017)
4. Birsan, T., Tiba, D.: One hundred years since the introduction of the set distance by dimitrie pompeiu. In: IFIP SMO (2005)
5. Blumberg, H.: Hausdorff's Grundzüge der Mengenlehre. Bulletin of the American Mathematical Society, 27 (3): 116–129, American (1920)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
7. Chen, S., Ma, X., Lu, Y., Hsu, D.: Ab initio particle-based object manipulation. In: Shell, D.A., Toussaint, M., Hsieh, M.A. (eds.) RSS (2021)
8. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI (2020)
9. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary IoU: Improving object-centric image segmentation evaluation. In: CVPR (2021)
10. Cheng, B., Girshick, R.B., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: Improving object-centric image segmentation evaluation. In: CVPR (2021)
11. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR (2020)
12. Cheng, M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.: Global contrast based salient region detection. IEEE TPAMI **37**(3), 569–582 (2015)
13. Chinchor, N.: MUC-4 evaluation metrics. In: MUC (1992)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
16. Ehrig, M., Euzenat, J.: Relaxed precision and recall for ontology matching. In: K-CapW (2005)
17. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2), 303–338 (2010)
18. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV (2018)
19. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV (2017)
20. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI (2018)
21. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE TPAMI (2021)
22. Fan, D.P., Ji, G.P., Qin, X., Cheng, M.M.: Cognitive vision inspired object segmentation metric and loss function. SSI **6** (2021)

23. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: CVPR (2020)
24. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR (2021)
25. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: CVPR (2021)
26. Fiorio, C., Gustedt, J.: Two linear time union-find strategies for image processing. TCS **154**(2), 165–181 (1996)
27. Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet another survey on image segmentation: Region and boundary information integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV (2002)
28. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE TPAMI **43**(2), 652–662 (2019)
29. Girshick, R.: Fast r-cnn. In: ICCV (2015)
30. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
31. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE TPAMI **34**(10), 1915–1926 (2012)
32. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
33. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE TPAMI **PAMI-9**(4), 532–550 (1987)
34. Hausdorff, F.: Grundzüge der Mengenlehre. Leipzig: Veit, ISBN 978-0-8284-0061-9Reprinted by Chelsea Publishing Company in 1949, Germany (1914)
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
36. Howard, A., Pang, R., Adam, H., Le, Q.V., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y.: Searching for mobilenetv3. In: ECCV (2019)
37. Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F.: Temporally distributed networks for fast video semantic segmentation. In: CVPR (2020)
38. Ke, Z., Li, K., Zhou, Y., Wu, Q., Mao, X., Yan, Q., Lau, R.W.: Is a green screen really necessary for real-time portrait matting? ArXiv **abs/2011.11961** (2020)
39. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
40. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU **184**, 45–56 (2019)
41. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)
42. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR (2015)
43. Li, H., Xiong, P., Fan, H., Sun, J.: Dfanet: Deep feature aggregation for real-time semantic segmentation. In: CVPR (2019)
44. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR (2014)
45. Liew, J.H., Cohen, S., Price, B., Mai, L., Feng, J.: Deep interactive thin object selection. In: WACV (2021)
46. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. CoRR **abs/2108.11515** (2021)
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

48. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE TPAMI **33**(2), 353–367 (2011)
49. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
50. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. arXiv preprint arXiv:1611.08408 (2016)
51. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR (2021)
52. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps. CVPR (2014)
53. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE TPAMI **26**(5), 530–549 (2004)
54. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: CVPR (2021)
55. Mnih, V.: Machine Learning for Aerial Image Labeling. Ph.D. thesis, University of Toronto (2013)
56. Mnih, V., Hinton, G.E.: Learning to detect roads in high-resolution aerial images. In: ECCV (2010)
57. Movahedi, V., Elder, J.H.: Design and perceptual validation of performance measures for salient object segmentation. In: CVPRW (2010)
58. Nirkin, Y., Wolf, L., Hassner, T.: Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. arXiv preprint arXiv:2012.11582 (2020)
59. Orsic, M., Kreso, I., Bevandic, P., Segvic, S.: In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: CVPR (2019)
60. Osserman, R.: The isoperimetric inequality. BAM **84**(6), 1182–1238 (1978)
61. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)
62. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
63. Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Lin, Z., Torr, P., Jia, J.: Open-world entity segmentation. arXiv preprint arXiv:2107.14228 (2021)
64. Qin, X., Fan, D.P., Huang, C., Diagne, C., Zhang, Z., Sant'Anna, A.C., Suàrez, A., Jagersand, M., Shao, L.: Boundary-aware segmentation network for mobile and web applications. arXiv preprint arXiv:2101.04704 (2021)
65. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. PR **106**, 107404 (2020)
66. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR (2019)
67. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. CGIP **1**(3), 244–256 (1972)
68. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS (2015)
69. van Rijsbergen, C.J.: Information retrieval. London:Butterworths, `1979.http://www.dcs.gla.ac.uk/Keith/Preface.html` (1979)
70. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)

71. Saito, S., Yamashita, T., Aoki, Y.: Multiple object extraction from aerial imagery with convolutional neural networks. EI **2016**(10), 1–9 (2016)
72. Shen, X., Hertzmann, A., Jia, J., Paris, S., Price, B., Shechtman, E., Sachs, I.: Automatic portrait segmentation for image stylization. In: CGF (2016)
73. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
74. Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished Manuscript (2018)
75. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**(1), 1929–1958 (2014)
76. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. CVGIP **30**(1), 32–46 (1985)
77. Sørensen, T.J.: 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)
78. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114 (2019)
79. Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: ICCV (2021)
80. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
81. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2019)
82. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
83. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: CVPR (2018)
84. Watson, A.B.: Perimetric complexity of binary digital images. Math J **14**, 1–40 (2012)
85. Wei, J., Wang, S., Huang, Q.: F^3net: Fusion, feedback and focus for salient object detection. In: AAAI (2020)
86. Wu, K., Otoo, E.J., Shoshani, A.: Optimizing connected component labeling algorithms. In: Fitzpatrick, J.M., Reinhardt, J.M. (eds.) MI (2005)
87. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
88. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR (2017)
89. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR (2013)
90. Yang, C., Wang, Y., Zhang, J., Zhang, H., Lin, Z., Yuille, A.: Meticulous object segmentation. arXiv preprint arXiv:2012.07181 (2020)
91. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013)
92. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (2018)
93. Yu, H., Xu, N., Huang, Z., Zhou, Y., Shi, H.: High-resolution deep image matting. arXiv preprint arXiv:2009.06613 (2020)
94. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: CVPR. pp. 7234–7243 (2019)
95. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (1984)

96. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. GRSL **15**(5), 749–753 (2018)
97. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: ECCV (2018)
98. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
99. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)
100. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: ECCV (2020)
101. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR (2021)
102. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE TPAMI **40**(6), 1452–1464 (2017)
103. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)