Research Article

Dongdong Zhang, Chunping Wang, Qiang Fu, Yue Cheng, Zhaorui Li and Qing Yang*

DSC: depth data quality optimization framework for RGBD camouflaged object detection

https://doi.org/10.1515/phys-2025-0236
Received March 12, 2025; accepted October 13, 2025;
published online December 4, 2025

Abstract: Camouflaged object detection (COD) faces unique challenges due to the extremely high visual similarity between objects and their surroundings, coupled with indistinct boundary features. While the introduction of depth information has provided new insights into addressing these challenges, existing methods still exhibit considerable limitations in depth data quality assessment and optimization. To address this issue, this paper proposes a depth screening and calibration (DSC) framework aimed at constructing a high-quality RGBD COD dataset. The framework first establishes a comprehensive evaluation metric that quantitatively assesses depth data generated by various monocular depth estimation (MDE) methods across multiple dimensions, including structural similarity, edge consistency, foreground smoothness, depth value utilization, and depth disparity between foreground and background. Based on these metrics, optimal depth maps are selected from those generated by multiple MDE methods for each image, forming an initial RGBD COD dataset. Subsequently, a Two-stage Depth Calibration (TDC) strategy is designed to calibrate the depth maps in the initial dataset through two consecutive phases: positive-negative sample discrimination and calibrated depth map generation, effectively enhancing the overall quality of depth maps. Experimental results on three benchmark datasets demonstrate that detection models trained with our high-quality depth data significantly outperform alternative approaches. This work provides a reliable data foundation for further exploring the role of depth information in improving COD performance.

Keywords: camouflaged object detection; monocular depth estimation; depth evaluation; depth calibration; dataset

***Corresponding author: Qing Yang**, Army Engineering University of PLA, Shijiazhuang, 050003, China, E-mail: 1042911849@qq.com
Dongdong Zhang, Chunping Wang, Qiang Fu, Yue Cheng and Zhaorui Li, Army Engineering University of PLA, Shijiazhuang, 050003, China. https://orcid.org/0000-0003-3817-5706 (D. Zhang).
https://orcid.org/0000-0002-3831-9856 (Q. Fu)

1 Introduction

Camouflage is a crucial survival skill that organisms have evolved through natural selection, enabling them to blend into their surroundings by altering their appearance, thereby reducing the probability of detection by predators [1]. From a broader perspective, camouflage occurs when target objects exhibit high visual similarity with their background environment in terms of color, texture, and other visual features, or when they cleverly utilize environmental characteristics to conceal their key features, making accurate identification by visual systems challenging. Camouflaged object detection (COD) technology aims to precisely segment these camouflaged objects from complex background environments. Compared to traditional object detection tasks, COD faces unique challenges: camouflaged objects typically share extremely high visual similarity with their backgrounds, and their boundary features are often indistinct and difficult to discern. To address these technical challenges, researchers have conducted extensive and in-depth investigations, advancing the development of COD across multiple practical applications, including agricultural pest identification [2], industrial defect detection [3], and medical image segmentation [4].

The rapid advancement of deep learning has accelerated research progress in COD, with deep learning-based COD algorithms achieving significant improvements in detection performance [1, 5–8]. However, existing methods primarily rely on RGB images for feature extraction and object detection, which often fail to achieve satisfactory results when confronting highly challenging scenarios, such as complex environmental backgrounds or cases where target objects' textures closely resemble their surroundings. As illustrated by the detection results in Figure 1, even state-of-the-art COD methods exhibit notable performance bottlenecks when processing such complex scenes, indicating substantial room for optimization and improvement in COD technology.

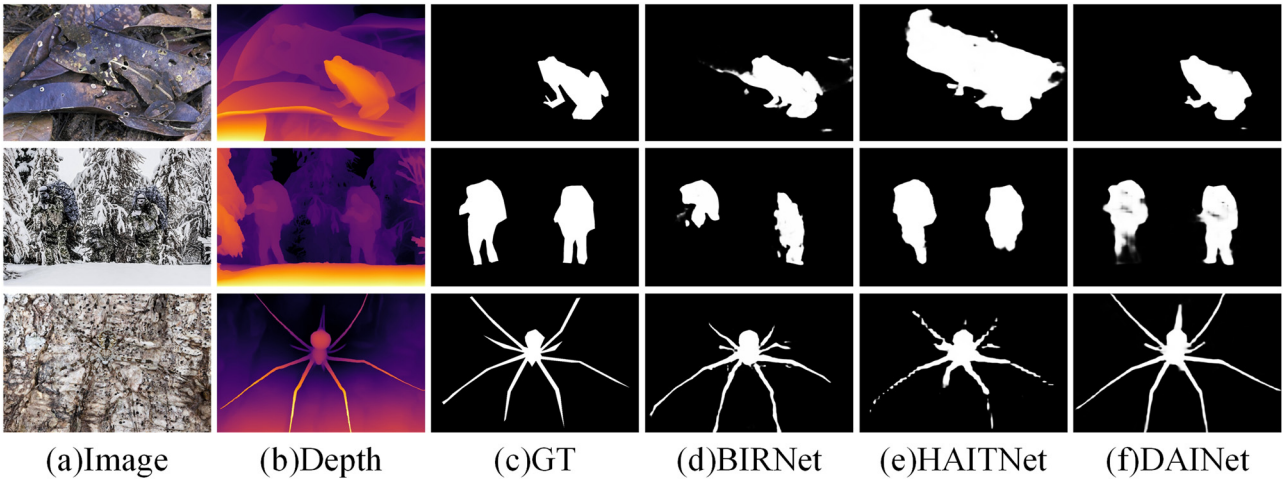

Figure 1: Visual examples of different methods. (a) RGB images. (b) Depth maps. (c) Ground truths. (d)–(f) Prediction maps produced by BIRNe [6], HAITNet [8], and DAINet [9], respectively.

Compared to RGB images that primarily provide color and texture information, depth images contain depth information that offers additional geometric and spatial cues, which are crucial for determining object position and shape. In the field of salient object detection (SOD), researchers have successfully incorporated depth information to address challenges in complex scenes [10–12]. Inspired by this progress, scholars have begun exploring the integration of depth cues into COD tasks, achieving remarkable results. Compared to methods using RGB information alone, COD approaches that incorporate depth information have shown tremendous potential, with several studies [9, 13–16] yielding encouraging outcomes in this direction. As shown in Figure 1, RGBD-based COD methods significantly outperform RGB-only approaches in complex scenarios.

However, due to the absence of dedicated real-world RGBD datasets for COD, existing methods must rely on monocular depth estimation (MDE) methods to generate depth data. This MDE-based depth data generation approach faces three major limitations. First, without unified depth quality evaluation criteria, researchers primarily select MDE methods based on visual effects, leading to significant variations in MDE method selection across different studies. For instance, Wang et al. [13] employed New CRFs [17], Wu et al. [14] opted for DPT [18], while Bi et al. [9] and Liu et al. [15] utilized MiDaS [19]. Second, the quality of depth data generated by different MDE methods varies considerably, and the same MDE method may perform inconsistently across different images. As illustrated in Figure 2, Depth-Anything-V2 [20] demonstrates superior overall performance, while DPT generally shows inferior

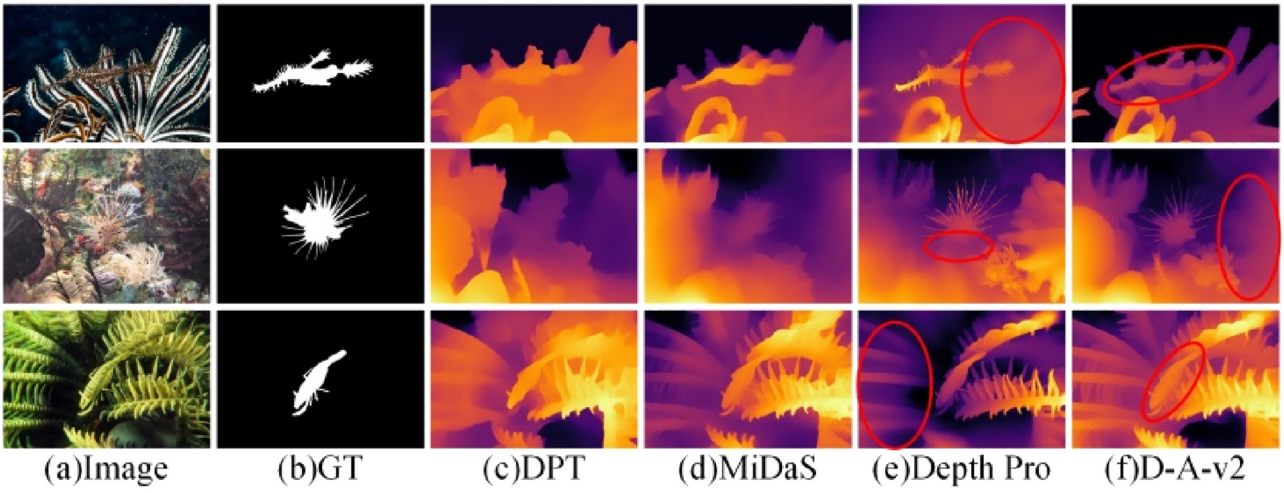

Figure 2: Visualization of some depth maps obtained by different MDE models (D-A-v2 represents Depth-Anything-V2).

results. Among all images, Depth-Anything-V2 exhibits the highest consistency with RGB images, while Depth Pro [21] provides target detail descriptions that more closely align with ground truth. Finally, due to domain gaps between MDE training datasets and COD datasets, even the best-performing MDE methods inevitably generate depth data with errors. As shown in the red-marked regions in Figure 2, the generated depth maps may exhibit poor visual quality or inconsistencies with the foreground, and directly using such data might compromise model generalization or even lead to overfitting issues.

To address these issues, researchers have conducted a series of investigations. Regarding depth quality assessment, Liu et al. [15] employed an indirect evaluation approach by comparing model performance across depth maps generated by various monocular depth estimation algorithms, including DPT [18], AdelaiDepth [22], and MiDaS [19], ultimately selecting MiDaS based on optimal performance conditions. While this performance-based evaluation method offers greater objectivity than purely visual judgment, it not only fails to fundamentally address the depth quality assessment issue but also introduces additional workload by not directly evaluating the quality of depth data itself. In contrast, this paper proposes a comprehensive depth quality metric that considers multiple factors, providing direct and reliable quantitative criteria for MDE method selection.

Concerning depth data generation, existing studies [9, 13, 14, 16] typically employ a single MDE method with good visual effects to generate all depth data. This approach overlooks a crucial fact: different MDE methods often exhibit significant performance variations when processing different images, and even the best-performing MDE method overall cannot guarantee superior results for every image. To address this issue, our paper independently evaluates multiple MDE-generated results for each image based on the proposed depth quality metrics, constructing high-quality training and testing datasets by selecting the optimal depth maps.

To mitigate the negative impact of low-quality depth maps, existing research has primarily focused on exploring effective multimodal fusion strategies [9, 13, 15, 16]. For instance, Bi et al. [9] designed a depth alignment index to evaluate depth map quality and dynamically adjust fusion weights accordingly, while Liu et al. [15] proposed a depth-weighted cross-attention fusion module that adaptively adjusts weight distribution by assessing the importance of both RGB and depth modalities. Although these methods partially suppress the interference of inaccurate depth maps on COD performance, as shown in Table 1, models still exhibit significant performance variations across

Table 1: Quantitative comparison of 4 state-of-the-art models on three benchmark datasets. "↑"/"↓" indicates that larger/smaller is better.

Data	Method	CAMO				COD10K				NC4K			
		$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$
MiDas (Q = 0.79)	DaCOD	0.805	0.879	0.769	0.071	0.791	0.865	0.691	0.039	0.824	0.886	0.782	0.053
	PopNet	0.806	0.862	0.772	0.077	0.817	0.884	0.730	0.035	0.847	0.899	0.810	0.047
	DAINet	0.802	0.860	0.761	0.077	0.807	0.884	0.712	0.037	0.838	0.895	0.796	0.049
	DAFNet	0.810	0.859	0.767	0.075	0.826	0.888	0.732	0.035	0.857	0.902	0.811	0.045
DPT (Q = 0.76)	DaCOD	0.791	0.861	0.762	0.074	0.786	0.862	0.674	0.042	0.823	0.884	0.767	0.053
	PopNet	0.794	0.843	0.755	0.083	0.815	0.883	0.725	0.035	0.847	0.896	0.807	0.047
	DAINet	0.792	0.848	0.749	0.082	0.805	0.883	0.708	0.038	0.838	0.895	0.796	0.049
	DAFNet	0.807	0.855	0.760	0.076	0.823	0.885	0.728	0.035	0.852	0.896	0.804	0.047
Depth-Anything-V2 (Q = 0.89)	DaCOD	0.851	0.914	0.829	0.048	0.823	0.898	0.740	0.029	0.864	0.922	0.833	0.035
	PopNet	0.850	0.890	0.822	0.056	0.837	0.897	0.753	0.029	0.863	0.909	0.830	0.042
	DAINet	0.854	0.905	0.833	0.053	0.830	0.897	0.746	0.031	0.861	0.912	0.825	0.041
	DAFNet	0.858	0.902	0.819	0.054	0.841	0.895	0.751	0.032	0.868	0.908	0.824	0.041
Depth pro (Q = 0.87)	DaCOD	0.834	0.898	0.810	0.060	0.807	0.877	0.713	0.034	0.841	0.907	0.799	0.042
	PopNet	0.845	0.892	0.817	0.061	0.833	0.895	0.752	0.030	0.861	0.906	0.826	0.042
	DAINet	0.852	0.905	0.832	0.055	0.829	0.897	0.742	0.031	0.853	0.907	0.818	0.044
	DAFNet	0.851	0.889	0.816	0.059	0.832	0.886	0.740	0.035	0.863	0.905	0.819	0.044
$Depth_{cal}$ (Q = 0.95)	DaCOD	0.870	0.927	0.850	0.044	0.831	0.901	0.747	0.028	0.871	0.927	0.838	0.033
	PopNet	0.865	0.908	0.842	0.050	0.854	0.912	0.783	0.026	0.879	0.921	0.850	0.036
	DAINet	0.877	0.928	0.859	0.043	0.850	0.912	0.774	0.027	0.875	0.923	0.845	0.037
	DAFNet	0.884	0.921	0.852	0.043	0.854	0.901	0.768	0.028	0.883	0.920	0.841	0.037

datasets of different quality levels, indicating that existing methods struggle to fundamentally overcome the impact of depth map quality on model performance. Unlike these approaches, we are committed to addressing depth map quality issues at their source: after screening for relatively high-quality depth maps, we further calibrate existing biases to significantly enhance depth map quality, thereby obtaining a high-quality depth dataset.

Based on the above analysis, this paper proposes a depth selection and calibration (DSC) framework aimed at constructing a high-quality RGBD COD dataset, establishing a foundation for in-depth exploration of depth information's role in enhancing COD performance. This framework optimizes depth image quality through systematic evaluation, selection, and calibration processes, building upon depth data generated by existing advanced MDE methods. Specifically, we first design comprehensive evaluation metrics that assess depth data generated by different MDE methods across multiple key dimensions, including edge consistency, structural similarity, and depth smoothness. Subsequently, we evaluate multiple MDE-generated results for each image based on these metrics, selecting the highest-scoring depth maps to form an initial dataset. Finally, through a designed Two-stage Depth Calibration (TDC) strategy, we calibrate the depth maps in the initial RGBD COD dataset and correct their biases, thereby further enhancing the overall quality of the RGBD COD dataset.

Our main contributions are summarized as:

1) We propose a depth selection and calibration (DSC) framework to construct a high-quality RGBD COD dataset, providing a reliable data benchmark for related research.
2) We introduce depth quality evaluation metrics that enable quantitative assessment of data generated by different MDE methods, offering reliable criteria for selecting high-quality depth maps.
3) We design a two-stage depth calibration strategy to calibrate depth images and correct potential biases, effectively enhancing the overall quality of depth maps.
4) We validate the effectiveness of our constructed dataset through various depth image-based detection methods. Experimental results demonstrate that models trained with our depth data significantly outperform alternative approaches.

2 Proposed method

2.1 Method overview

Figure 3 presents the overall architecture of the depth selection and calibration (DSC) framework. This framework comprises two key components: evaluation-based selection and depth calibration. Specifically, we first generate depth data using various MDE methods to establish an initial database.

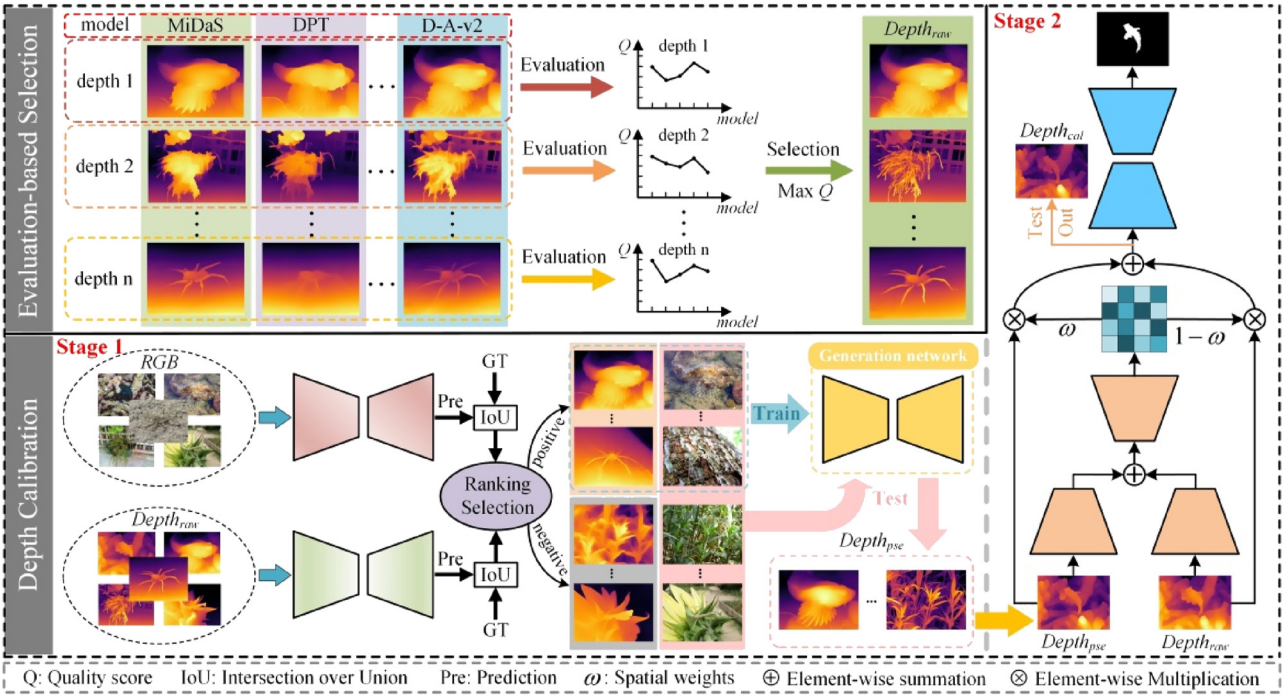

Figure 3: Detailed architecture of our depth selection and calibration (DSC) network (D-A-v2 represents Depth-Anything-V2).

Subsequently, based on our proposed depth quality evaluation metric (Quality score, represented by Q), we select the optimal depth map for each image from the initial database to construct a preliminary RGBD COD dataset. Finally, we employ the designed Two-stage Depth Calibration (TDC) strategy to calibrate the depth maps in the preliminary RGBD COD dataset, correcting potential noise introduced by unreliable original depth maps, thereby constructing a high-quality RGBD COD dataset. In the following sections, we will discuss these two components in detail.

2.2 Evaluation-based selection

As previously discussed, depth data generated by different MDE methods exhibits significant variations, and researchers typically rely on subjective observations to assess depth quality, which is both time-consuming and difficult to quantify. Moreover, due to the varying complexity of camouflaged image scenes, single MDE methods have limited generalization capability and struggle to generate high-quality depth maps consistently across all images. Therefore, relying solely on depth data generated by a single MDE method as a training benchmark cannot guarantee optimal model performance.

To address the challenges of quantitative depth quality assessment and data variability, we construct depth quality evaluation metrics from multiple dimensions to achieve quantitative assessment of depth map quality and selection of high-quality depth maps. Our proposed quality assessment metric (quality score, represented by Q) comprises five key components: structural similarity, edge consistency, foreground smoothness, depth value utilization rate, and depth difference between foreground and background.

Structural Similarity (SSIM) [23] is a widely used metric for measuring similarity between images. We compare the depth map with the grayscale version of the RGB image, using SSIM to evaluate their consistency. SSIM models similarity as a combination of three factors: luminance similarity, contrast similarity, and structural similarity. The definition of SSIM is as follows:

$$\mathrm{SSIM} = l_{\mathrm{xy}} \cdot c_{\mathrm{xy}} \cdot s_{\mathrm{xy}} \tag{1}$$

where

$$l_{\mathrm{xy}} = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, c_{\mathrm{xy}} = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \quad s_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}} + c_3}{\sigma_x\sigma_y + c_3} \tag{2}$$

where $\{c_1, c_2, c_3\}$ is a constant that prevents instability when the denominator approaches zero. $\{x, y\}$ represents the input image, while $\{\mu_x, \mu_y\}, \{\sigma_x, \sigma_y\}$ and σ_{xy} denote the corresponding means, variances, and covariances, respectively.

The edge structure of high-quality depth maps should maintain consistency with the original RGB images. However, RGB images contain rich color details that are not present in depth maps. As shown in Figure 4, although Canny edge detection results from RGB images and corresponding depth maps are difficult to compare directly, we observe that high-quality depth map edges demonstrate good consistency with the ground truth edges of camouflaged objects. Based on this key finding, we propose using the mean absolute error between depth map edges and the ground truth edges of camouflaged objects to evaluate edge consistency. The specific steps are as follows: first, normalize the depth map, then use the Canny operator to extract

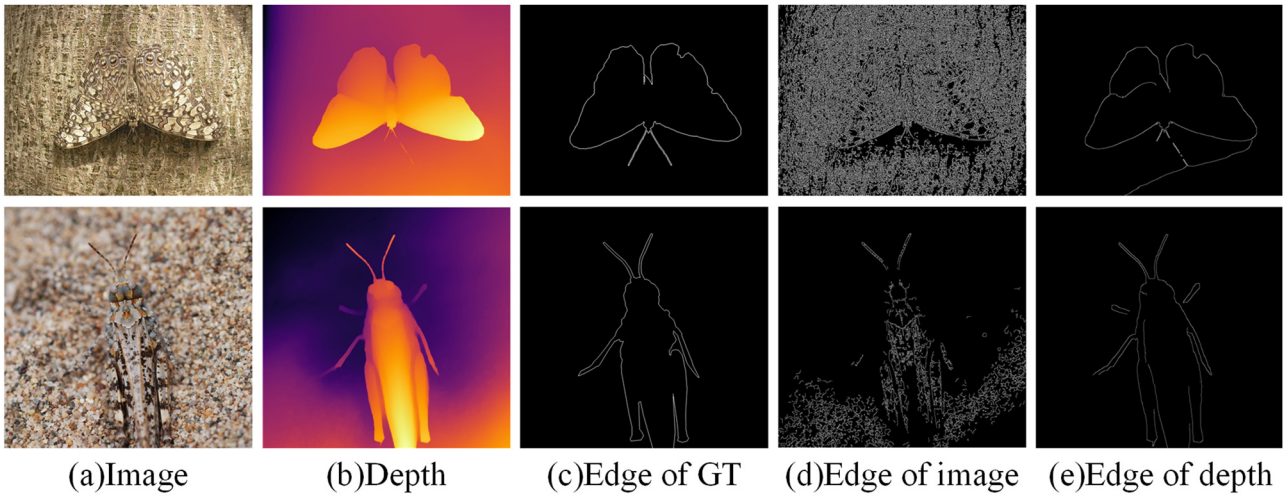

Figure 4: Comparative example of edge extraction results.

the edge map E_D, and finally calculate the edge consistency score E between the edge map E_D and the edge ground truth E_G. The specific formula is:

$$E = \text{mean}\left(\sum_{(x,y)\in E_G} \left| E_G(x,y) - E_D(x,y) \right| \right) \tag{3}$$

where mean(·) represents the mean operation.

In the real world, object surfaces typically exhibit smooth characteristics, which should be reflected as continuous depth variations in depth maps. However, MDE methods may introduce interference, such as noise and depth discontinuities, when generating depth maps. To evaluate the smoothness of depth in foreground regions, we multiply the depth map with the ground truth mask to obtain the foreground depth map and describe the spatial distribution smoothness by calculating the depth variance in the object region. The foreground smoothness S_f is defined as:

$$S_f = \sigma_f^2 \tag{4}$$

where σ_f^2 represents the depth variance in the foreground region.

As a measure of object distances in a scene, depth map quality depends not only on local matching with the original image but more importantly on the richness and effective utilization of depth information. When depth values are overly concentrated (e.g., most pixels having similar or identical depth values), the depth map will struggle to accurately characterize the three-dimensional structural features of the scene. Therefore, we introduce the depth value utilization rate as an evaluation metric to quantify the distribution characteristics of depth values in depth maps, ensuring that depth maps can fully utilize the available depth range and effectively express the scene's depth hierarchy information. This metric is based on information entropy theory and is used to evaluate the uniformity and diversity of depth value distribution. Specifically, we first divide the depth values in the depth map into N intervals (typically $N = 256$, corresponding to the grayscale levels of an 8 bit depth map) to obtain a depth value histogram $\{n_i\}(i = 1,2,\ldots,N)$. Then, we normalize the histogram to calculate the probability p_i for each depth value interval:

$$p_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \tag{5}$$

Subsequently, based on the definition of information entropy, we calculate the entropy of depth value distribution:

$$H = -\sum_{i=1}^{N} p_i \log(p_i + \varepsilon) \tag{6}$$

where ε is a small exponent to prevent log0 cases.

Finally, we calculate the maximum entropy value $H_{\max}$ and compute the normalized depth value utilization rate U:

$$\begin{cases} H_{\max} = \log N \\ U = \dfrac{H}{H_{\max}} \end{cases} \tag{7}$$

The value of U ranges from [0,1], with values closer to 1 indicating more uniform depth value distribution and higher utilization rate.

Furthermore, although camouflaged objects exhibit high similarity with their backgrounds in RGB images, they still maintain significant spatial differences. High-quality depth maps should effectively capture and reflect these spatial differential characteristics. Based on this observation, we propose a depth difference metric between foreground and background. To ensure assessment accuracy, rather than directly calculating the average depth difference between foreground and background, we compute the average depth difference within a 5×5 neighborhood around the target boundary, thereby quantifying the depth map's ability to highlight camouflaged targets. The mathematical definition of the foreground-background depth difference D is as follows:

$$D = \text{mean}\left(\sum_{(x,y)\in E_G} \left| \mu_f(x,y) - \mu_g(x,y) \right| \right) \tag{8}$$

where $\mu_f(\cdot)$ and $\mu_g(\cdot)$ represent the mean depth values of foreground and background regions within a 5×5 neighborhood around point (x,y) on the target boundary, respectively.

Integrating the above five dimensions, our depth quality evaluation metric (Q) is calculated as:

$$Q = w_1 \text{SSIM} + w_2(1 - E) + w_3 S_f + w_4 U + w_5 D \tag{9}$$

where $\{w_1, w_2, w_3, w_4, w_5\}$ represents weight parameters used to balance the contribution ratios of different metrics, which in this paper is set to $\{0.7, 1, 0.2, 0.3, 2\}$. The determination process of the weight parameters is as follows: first, all weight parameters are initially set to 1, and the mean value of each metric across all depth images is calculated; then, each metric's mean is divided into 0.2 to obtain the corresponding weight parameter. This method ensures that each metric has equal influence in the evaluation of depth quality. By comparing quality scores of different depth maps, we can select the optimal depth map for each image from results generated by different MDE methods, thereby constructing a high-quality initial RGBD COD dataset.

2.3 Depth calibration

Spatial information provided by depth maps and the texture-free separation of foreground and background play crucial roles in breaking camouflage. However, due to domain gaps, depth maps generated by monocular depth estimation methods contain substantial noise, resulting in unreliable depth data. Direct use of such depth data may significantly degrade the performance of RGBD COD models.

To address the performance bottleneck caused by noise, similar to [24, 25], we attempt to calibrate the original depth maps to obtain high-quality depth maps consistent with foreground objects. Two key issues need to be addressed: 1) how to distinguish between good-quality (positive samples) and poor-quality (negative samples) depth maps in the initial RGBD COD dataset. (2) how to generate calibrated depth maps that preserve high-quality portions while correcting low-quality regions. Therefore, we design a Two-stage Depth Calibration (TDC) strategy, which forms the core component of DSC. The two consecutive stages involve distinguishing positive and negative samples and generating calibrated depth maps. Figure 3 illustrates the proposed TDC strategy, with specific details as follows:

Stage 1: Positive and negative samples are separated based on the IoU [26] between prediction results and their ground truth (GT). This is based on the consideration that IoU can measure the consistency between prediction results and their corresponding GT, thereby reflecting the reliability of information contained in depth images to some extent.

Specifically, first, under-ground truth supervision, we train two encoder-decoder networks with identical architectures for RGB data and depth data separately. Here, the Resnet50 network [27] serves as the encoder, while the decoder part of U-Net [28] serves as the decoder. Then, RGB data and depth data are input separately into their respective pre-trained networks to generate camouflaged object prediction results, and IoU values between each prediction result and its corresponding ground truth are calculated, denoted as $\mathrm{IoU}_{\mathrm{depth}}$ and $\mathrm{IoU}_{\mathrm{RGB}}$ respectively. Finally, depth images that provide reliable cues, namely samples ranking in the top 20 % of $\mathrm{IoU}_{\mathrm{depth}}$ and samples where $\mathrm{IoU}_{\mathrm{depth}} > \mathrm{IoU}_{\mathrm{RGB}}$, are selected from the initial RGBD COD dataset as the positive sample set, with the remaining images forming the negative sample set. Compared to the negative sample set, depth images in the positive sample set are more beneficial for COD, with more acceptable depth quality. The middle position of the lower half of Figure 3 shows typical examples from both positive and negative sample sets.

Stage 2: Utilize a generation network to generate depth images and calibrate original depth images using the generated images.

Specifically, we first retrain the image generation network from [29] with RGB images as input and depth maps from the positive sample set as supervision information to reduce noise in the original depth data. Subsequently, RGB images from the initial RGBD COD dataset are input into the trained generation network to obtain high-quality pseudo depth images. These pseudo depth images do not directly replace the initial depth maps (depth maps in the initial RGBD COD dataset) but are used for calibration. During calibration, we adopt a spatial weighted sum of initial depth maps and pseudo depth maps to replace the original depth maps, with weights determined by depth's contribution to detection. As shown on the right side of Figure 3, initial depth maps and pseudo depth maps are separately input into the encoder (Resnet50) for feature extraction, followed by feature fusion through the decoder to generate spatial weights. These weights are applied to both types of depth maps to obtain calibrated depth maps, which are then input into the same encoder-decoder network used in Stage 1 for camouflaged object detection. Spatial weights are dynamically adjusted during network training, and upon completion of training, optimal weights and calibrated depth maps $Depth_{cal}$ are obtained, calculated as follows:

$$Depth_{cal} = \omega * Depth_{raw} + (1 - \omega) Depth_{pse} \tag{10}$$

where $Depth_{raw}$ and $Depth_{pse}$ represent the initial depth map and pseudo depth map respectively, and ω represents spatial weights. For better understanding, we visualize intermediate results of the depth calibration process in Figure 5. Through comparative analysis of different depth maps in Figure 5, we can draw the following observations: First, comparing the third and fourth columns, $Depth_{pse}$ provides richer three-dimensional spatial layout information compared to $Depth_{raw}$. Second, comparing the third and fifth columns clearly shows that $Depth_{cal}$ presents more complete scene structure than $Depth_{raw}$, with clearer target structural details. Finally, comprehensive comparison of results in the third, fourth, and fifth columns demonstrates that $Depth_{cal}$ exhibits significant advantages in overall visual quality.

3 Experiments

3.1 Experimental setup

Datasets: To evaluate the effectiveness of the proposed DSC framework, we conducted experiments on three widely used and challenging COD datasets: CAMO [30], COD10K [31], and NC4K [32]. CAMO contains 1,250 images, with 1,000 for training and 250 for testing. COD10K is currently the

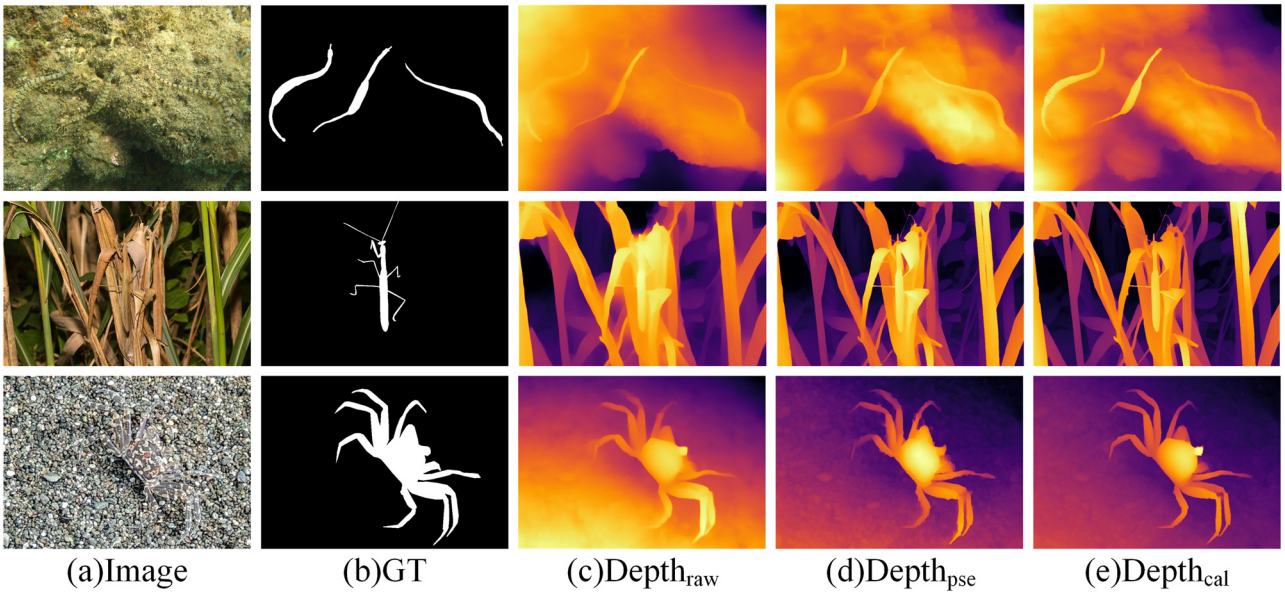

Figure 5: The internal inspections of depth calibration: examples of initial depth map $Depth_{raw}$, pseudo depth map $Depth_{pse}$ and the calibrated depth map $Depth_{cal}$.

largest camouflaged object dataset with high-quality pixel-level annotations, comprising 5,066 camouflaged images, of which 3,040 are used for training and 2,026 for testing. NC4K is the largest camouflaged object test set to date, consisting of 4,121 images downloaded from the internet, providing a rigorous test of model generalization capability. Following mainstream training configurations [1, 5–7, 31], we use 3,040 samples from COD10K and 1,000 samples from CAMO as the training set, while the remaining samples and the NC4K dataset are used for testing.

Evaluation Metrics: To quantitatively assess the impact of different depth data on model performance, we employ four widely-used evaluation metrics: S-measure (S_α), F-measure (F_β), E-measure (E_ϕ), and Mean Absolute Error (MAE, M). Among these metrics, higher values of S_α, F_β, and E_ϕ indicate better performance, while the opposite holds true for M. For detailed definitions of these evaluation metrics, please refer to [31].

Implementation details: The framework is implemented in PyTorch and trained using an NVIDIA GeForce RTX 3090 GPU with 24 GB memory. The backbone network employs Resnet50 with parameters pre-trained on ImageNet. Input images are uniformly resized to 352 × 352 pixels, and various data augmentation techniques are applied, including random flipping, rotation, and cropping. During training, the initial learning rate is set to 1e-4 using the Adam optimizer with a batch size of 16. During inference, the encoder-decoder architecture predicts camouflaged objects in an end-to-end manner without requiring any post-processing operations.

3.2 Comparative experiments

We compare Depthcal with depth data generated by four MDE methods: DPT [18], MiDaS [19], Depth-Anything-V2 [20], and Depth Pro [21]. DPT leverages Vision Transformers for dense prediction tasks and is known for its high accuracy and strong generalization across diverse visual scenes. MiDaS emphasizes robustness by mixing multiple datasets for training, enabling zero-shot cross-dataset transfer; however, its depth maps may be less sharp in fine-structured regions. Depth-Anything-V2 is designed for scalability with large-scale data and achieves impressive performance on both indoor and outdoor scenarios, but it may require substantial computational resources. Depth Pro focuses on delivering sharp and metrically accurate depth maps with fast inference speed, though its performance can fluctuate in highly complex or ambiguous scenes. By utilizing these diverse models, we are able to comprehensively evaluate the effectiveness of the proposed framework. In addition, the performance of four advanced RGBD COD methods (DaCOD [13], PopNet [14], DAINet [9], and DAFNet [15]) are evaluated on different depth data. All methods are implemented using author-provided open-source code, with MDE methods utilizing original weights and detection methods retrained using default parameters. To ensure fairness, we employ unified evaluation

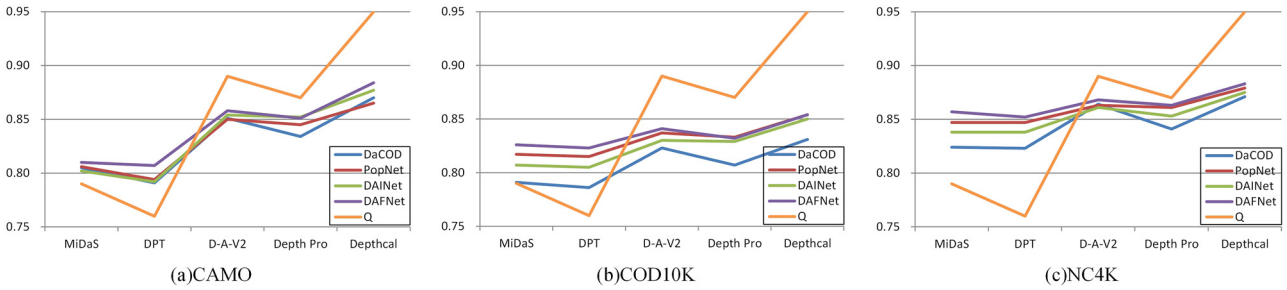

Figure 6: Comparison of S-measure metrics across different methods on three benchmark datasets (D-A-v2 represents Depth-Anything-V2).

protocols and code for objective assessment of prediction results.

Quantitative evaluation: Table 1 presents the quantitative comparison results. As shown in Table 1, Depthcal demonstrates the most outstanding performance in overall quality assessment, achieving a quality score Q of 0.95, significantly outperforming depth data generated by other MDE methods. Compared to other depth data sources, all detection methods achieve optimal performance when using Depthcal. Specifically, taking DAFNet as an example, compared to Depth-Anything-V2 (the second-best performing depth data), using Depthcal results in average improvements of 2.5 %, 1.2 %, and 2.2 % in Sα, Fβ, and Eφ metrics respectively, while reducing M by 0.3 %. To clearly illustrate the experimental results, we visualize part of the quantitative data from Table 1 as line graphs (as shown in Figure 6). Through systematic analysis, we find a significant positive correlation between depth data quality score Q and model performance: higher quality scores correspond to better model performance. This finding strongly validates the rationality of our proposed quality assessment metrics. Additionally, the performance variations of different detection models across various depth data sources further confirm the effectiveness of the Depthcal.

Qualitative evaluation: Figure 7 shows typical samples generated by $Depth_{cal}$ and various advanced MDE methods along with their corresponding quality scores. Comparative analysis indicates that calibrated depth ($Depth_{cal}$) provides richer 3D scene information and target structural details. Meanwhile, depth maps with higher quality scores typically exhibit superior visual quality and better foreground consistency with corresponding RGB images. Figure 8 demonstrates the prediction results of the advanced detection method DAFNet based on different depth data. For simple scenes like Image 1, depth maps generated by various monocular depth estimation methods show similar and relatively high quality, enabling DAFNet to accurately identify camouflaged targets. However, for complex scenes like Image 2, some generated depth maps exhibit poor quality or inconsistency with RGB image foregrounds, leading to suboptimal detection results. For instance, DAFNet encounters incomplete target detection issues when using depth maps

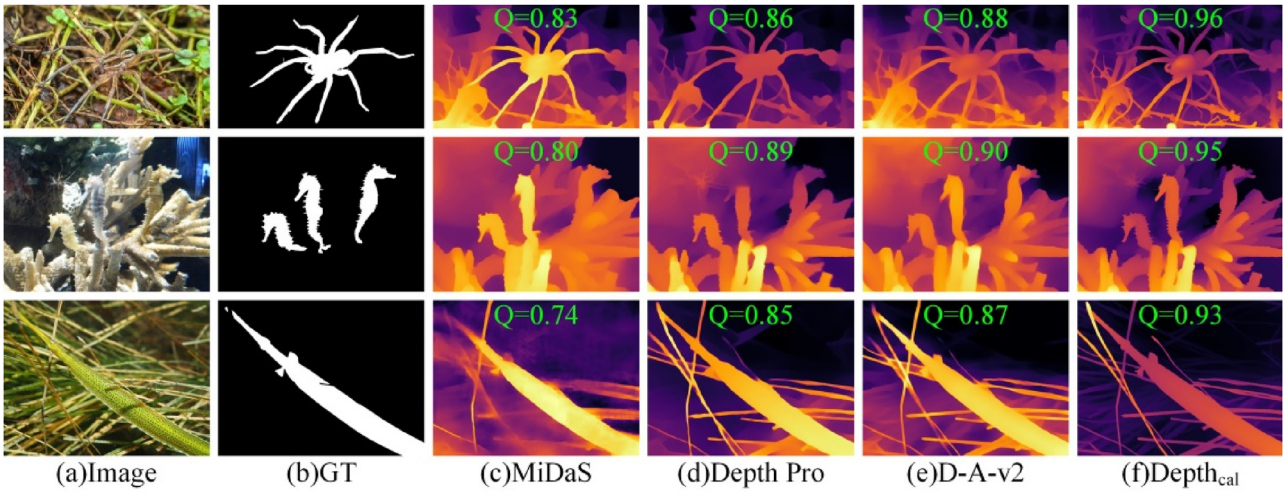

Figure 7: Typical examples of different depth data (D-A-v2 represents Depth-Anything-V2).

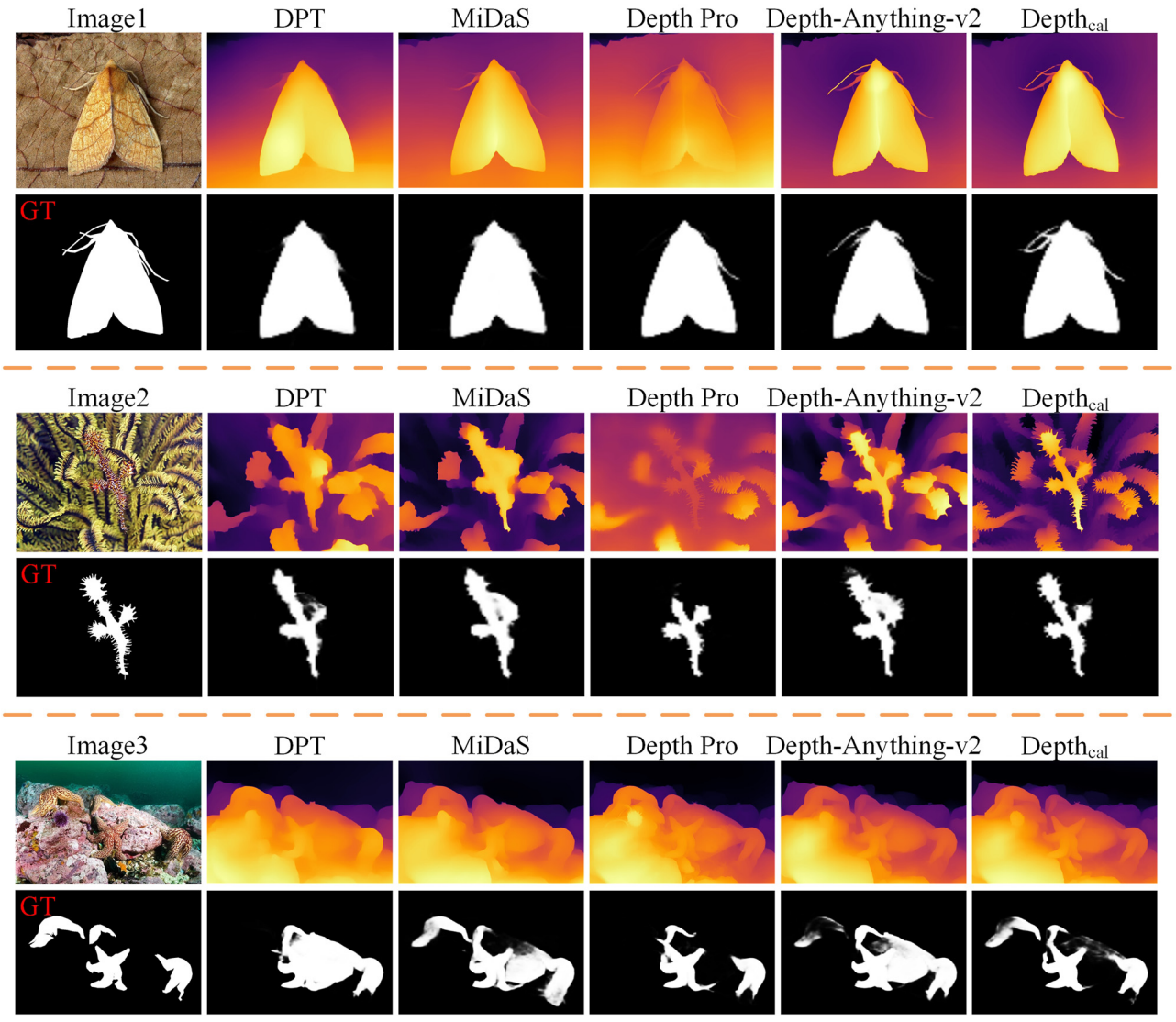

Figure 8: Visual comparison of detection results obtained by models using different depth data (D-A-v2 represents Depth-Anything-V2).

generated by Depth Pro, while depth maps from Depth-Anything-V2 cause background regions to be misidentified as foreground. Furthermore, for multi-target images like Image 3, most MDE methods struggle to completely estimate depth information for all targets, preventing detection models from achieving comprehensive identification of all camouflaged targets. In contrast, detection methods using $Depth_{cal}$ can effectively identify camouflaged targets in these challenging scenarios, primarily benefiting from $Depth_{cal}$'s higher depth quality and significantly improved foreground consistency after calibration.

3.3 Ablation studies

Our DSC framework primarily consists of two core modules: evaluation-based selection and depth calibration. To systematically validate the effectiveness of each module and its components, we conducted two groups of ablation experiments on three benchmark datasets.

1) Effectiveness of evaluation-based selection.

To assess the effectiveness of the evaluation-based selection module, we compared the detection performance when models use $Depth_{raw}$ (depth data from the initial RGBD COD dataset) versus depth maps generated by various MDE methods. Quantitative results are shown in Tables 1 and 2. Across three datasets, compared to Depth Pro, $Depth_{raw}$ improved DAFNet's S_α, F_β, and E_φ metrics by an average of 1.3 %, 1.2 %, and 1.5 % respectively, while reducing M by 0.5 %. $Depth_{raw}$, a collection of high-quality depth maps selected through evaluation from those generated by various MDE methods, provides reliable depth information

Table 2: Quantitative results of ablation experiments. "↑"/"↓" indicates that larger/smaller is better.

Data	Method	CAMO				COD10K				NC4K			
		$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta\uparrow$	$M\downarrow$
$Depth_{raw}$ (Q = 0.91)	DaCOD	0.856	0.919	0.835	0.047	0.830	0.900	0.743	0.029	0.869	0.924	0.835	0.035
	PopNet	0.855	0.902	0.830	0.053	0.839	0.898	0.759	0.029	0.868	0.913	0.833	0.039
	DAINet	0.861	0.911	0.838	0.048	0.834	0.898	0.752	0.031	0.862	0.915	0.828	0.041
	DAFNet	0.866	0.903	0.834	0.052	0.844	0.898	0.754	0.030	0.874	0.914	0.832	0.040
$Depth'_{cal}$ (Q = 0.85)	DaCOD	0.828	0.893	0.801	0.063	0.803	0.874	0.708	0.035	0.837	0.902	0.795	0.045
	PopNet	0.835	0.885	0.806	0.065	0.829	0.892	0.747	0.031	0.858	0.904	0.822	0.043
	DAINet	0.840	0.894	0.814	0.061	0.824	0.894	0.735	0.033	0.849	0.904	0.813	0.045
	DAFNet	0.841	0.882	0.804	0.063	0.831	0.887	0.738	0.035	0.862	0.904	0.817	0.044

for the COD task. Notably, compared to other depth data, detection models showed smaller performance improvements when using $Depth_{raw}$ versus depth maps generated by Depth-Anything-V2. This phenomenon is reasonable since depth maps generated by Depth-Anything-V2 inherently possess relatively high quality.

(2) Effectiveness of depth calibration.

To evaluate the effectiveness of the depth calibration module, we compared the performance of detection models using $Depth_{raw}$, $Depth_{cal}$, and $Depth'_{cal}$ respectively. Here, $Depth'_{cal}$ represents depth data obtained by removing Stage 1 from the TDC strategy, where $Depth_{raw}$ is directly used as supervision information for the image generation network. It should be noted that removing Stage 2 alone or removing both stages from the TDC strategy would not generate new depth data; in these cases, the depth data after TDC strategy processing remains as $Depth_{raw}$. As shown in Table 2, experimental results indicate that detection models experience performance degradation when using $Depth'_{cal}$ compared to using $Depth_{raw}$. This performance deterioration primarily occurs because $Depth_{raw}$ contains some low-quality depth maps, and directly using all $Depth_{raw}$ data as supervision information reduces the quality of generated pseudo depth images, leading to suboptimal detection results. This phenomenon confirms the necessity of Stage 1 in the TDC strategy, which plays a crucial role in generating high-quality calibrated depth maps. Further analysis reveals that all detection models achieve optimal performance when using $Depth_{cal}$ compared to using $Depth_{raw}$ and $Depth'_{cal}$, fully validating the effectiveness of the TDC strategy. To intuitively demonstrate the effectiveness of the TDC strategy, we provide visual comparison results of $Depth_{raw}$, $Depth_{cal}$, and $Depth'_{cal}$ in Figure 9. As shown in

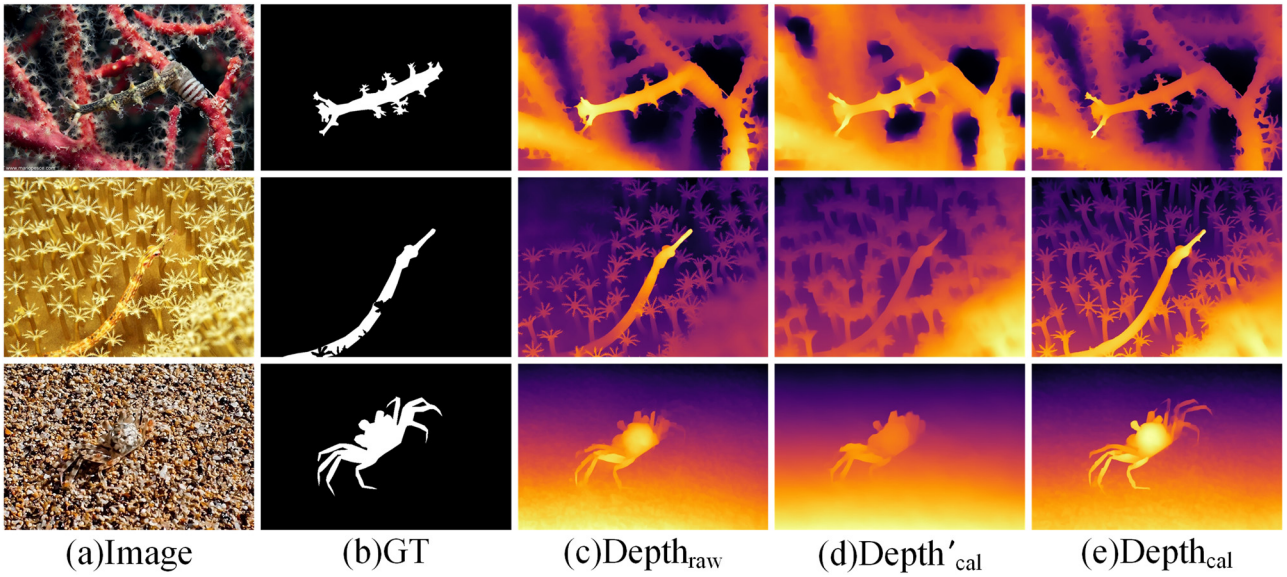

Figure 9: Visualization results for validating the effectiveness of TDC strategy.

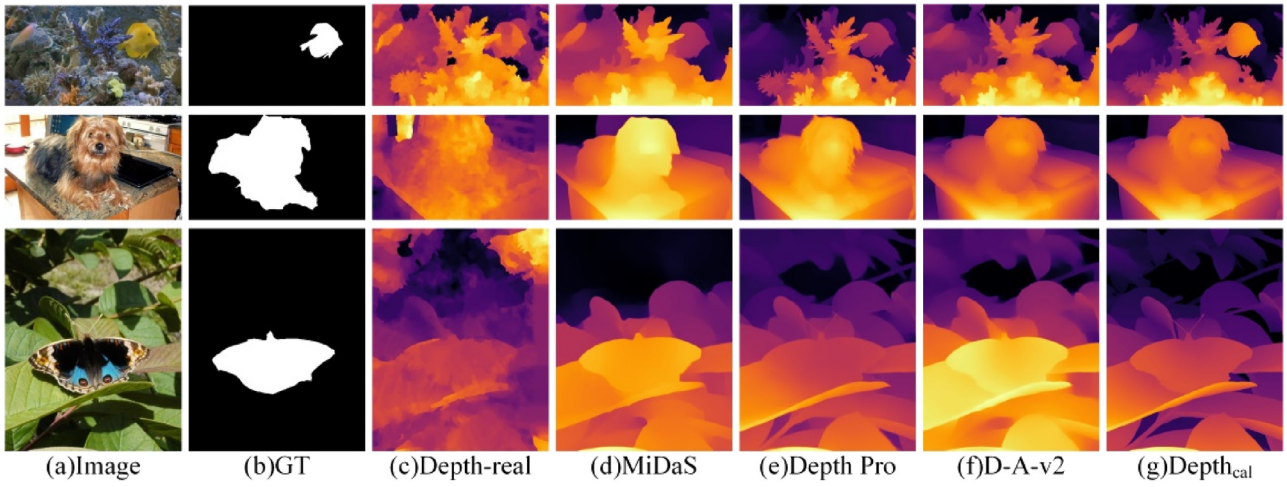

Figure 10: Examples of different depth data on the NJU2K dataset (D-A-v2 represents Depth-Anything-V2).

Figure 9(d), due to using initial depth maps containing low quality and foreground inconsistencies as supervision information, calibrated depth maps generated by the TDC strategy (without Stage 1) still exhibit poor visual quality or inconsistency with corresponding RGB image foregrounds. By comparing Figure 9(c)–(e), we can clearly observe that the complete TDC strategy generates depth images with significantly improved quality. This indicates that our proposed TDC strategy can effectively generate high-quality depth maps with foreground consistency, providing more reliable depth information support for the COD task.

3.4 Generalization experiments

To verify the generalization capability of the proposed framework, we applied it to the RGB-D SOD field. The widely used RGB-D SOD dataset – NJU2K dataset [33] – was selected for experiments. Following the experimental procedure described in this paper, we obtained depth maps generated by various MDE methods as well as the depth calibration results. Relevant examples are shown in Figure 10. By comparing the first six columns in Figure 10, it can be observed that, compared with the ground truth depth maps provided by the dataset (third column), the depth maps generated by MDE methods display better visual effects. Further comparison of the last four columns in Figure 10 shows that the depth calibration results produced by the proposed framework are of higher quality than the original outputs of each MDE method, not only containing more complete 3D scene information and object structural features, but also exhibiting better consistency with the RGB images in terms of foreground alignment. These experimental results indicate that the proposed framework is not only applicable to camouflaged object detection, but can also be effectively used to optimize depth data for other visual tasks, fully demonstrating the method's strong domain transferability and generalization performance.

4 Conclusions

This paper proposes the DSC framework to address depth data quality assessment and optimization issues in RGBD COD tasks. Through the design of multi-dimensional depth quality evaluation metrics, the framework enables quantitative assessment of depth data generated by different MDE methods, providing reliable criteria for selecting high-quality depth maps. Meanwhile, the designed TDC strategy effectively calibrates depth maps and corrects potential biases, significantly improving the overall quality of depth data. Experimental results demonstrate that the high-quality RGBD COD dataset constructed in this paper can provide more reliable depth information support for detection models, effectively enhancing model performance. In the future, our research will focus on improving the performance of RGBD COD detection. We will dedicate efforts to deeply exploring the synergistic mechanisms between depth information and RGB information, developing efficient multimodal feature fusion strategies to further enhance detection performance.

Funding information: Intra-military research project.
Author contribution: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Conflict of interest: The authors state no conflict of interest.
Data availability statement: The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Fan DP, Ji GP, Cheng MM, Shao L. Concealed object detection. IEEE Trans Pattern Anal Mach Intell 2022;44:6024—42.
2. Dai SL, Man H. A convolutional Riemannian texture model with differential entropic active contours for unsupervised pest detection. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans, LA, USA: IEEE; 2017:1028—32 pp.
3. Xiang XY, Liu MQ, Zhang SL, Wei P, Chen B. Multi-scale attention and dilation network for small defect detection. Pattern Recognit Lett 2023;172:82—8.
4. Guo XY, Lin X, Yang X, Yu L, Cheng KT, Yan Z. UCTNet: Uncertainty-guided CNN-Transformer hybrid networks for medical image segmentation. Pattern Recogn 2024;152:110491.
5. Liang WY, Wu JS, Mu XY, Hao F, Du J, Xu J, et al. Weighted dense semantic aggregation and explicit boundary modeling for camouflaged object detection. IEEE Sens J 2024;24:21108—22.
6. Yang JC, Zhang Q, Zhao YL, Li YT, Liu ZM. Bi-directional Boundary-object interaction and refinement network for camouflaged object detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, ON, Canada: IEEE; 2024:1—6 pp.
7. Tang Z, Tang J, Zou DP, Rao J, Qi F. Two guidance joint network based on coarse map and edge map for camouflaged object detection. Appl Intell 2024;54:7531—44.
8. Phung TH, Chen HJ, Shuai HH. Hierarchically aggregated identification transformer network for camouflaged object detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, ON, Canada: IEEE; 2024:1—6 pp.
9. Bi HB, Tong YY, Zhang JY, Zhang C, Tong J, Jin W. Depth alignment interaction network for camouflaged object detection. Multimed Syst 2024;30:51.
10. Chen TY, Xiao J, Hu XG, Zhang G, Wang S. Adaptive fusion network for RGB-D salient object detection. Neurocomputing 2023;522:152—64.
11. Sun HD, Wang Y, Ma XP. An adaptive guidance fusion network for RGB-D salient object detection. Signal, Image Video Proces 2024;18:1683—93.
12. Zhong MY, Sun J, Ren P, Wang F, Sun F. MAGNet: multi-scale awareness and global fusion network for RGB-D salient object detection. Knowl Base Syst 2024;299:112126.
13. Wang QW, Yang JY, Yu XS, Wang FY, Chen P, Zheng F. Depth-aided camouflaged object detection. In: Proceedings of the 31st ACM International Conference on Multimedia. New York, NY, USA: ACM; 2023:3297—306 pp.
14. Wu ZW, Paudel DP, Fan DP, Wang JJ, Wang S, Demonceaux C, et al. Source-free depth for object Pop-out. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE; 2023:1032—42 pp.
15. Liu XR, Qi L, Song YX, Wen Q. Depth awakens: a depth-perceptual attention fusion network for RGB-D camouflaged object detection. Image Vis Comput 2024;143:104924.
16. Xiang MC, Zhang J, Lv YQ, Li AX, Zhong YR, Dai YC. Exploring depth contribution for camouflaged object detection. arXiv preprint arXiv:2106.13217 2021.
17. Yuan WH, Gu XD, Dai ZZ, Zhu SY, Tan P. Neural window fully-connected CRFs for monocular depth estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE; 2022:3906—15 pp.
18. Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE; 2021: 12159—68 pp.
19. Ranftl RE, Lasinger K, Hafner D, Schindler K, Koltun V. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell 2020;44:1623—37.
20. Yang LH, Kang BY, Huang ZL, Zhao Z, Xu XG, Feng JS, et al. Depth anything v2. Adv Neural Inf Process Syst 2025;37:21875—911.
21. Bochkovskii A, Delaunoy AEL, Germain H, Santos M, Zhou YC, Richter SR, et al. Depth pro: sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073. 2024. https://doi.org/10.48550/arXiv.2410.02073.
22. Yin W, Zhang JM, Wang O, Niklaus S, Chen SM, Liu YF. Towards accurate reconstruction of 3d scene shape from a single monocular image. IEEE Trans Pattern Anal Mach Intell 2022;45:6480—94.
23. Wang Z, Bovik AC, Sheikh HR, Simoncelli E. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13:600—12.
24. Zhang Q, Qin Q, Yang Y, Jiao Q, Han JG. Feature calibrating and fusing network for RGB-D salient object detection. IEEE Trans Circ Syst Video Technol 2023;34:1493—507.
25. Ji W, Li JJ, Yu S, Zhang M, Piao YR, Yao SY. Calibrated RGB-D salient object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE; 2021:9466—76 pp.
26. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vis 2010;88:303—38.
27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
28. Ronneberger OAFP. U-Net: Convolutional networks for biomedical image segmentation[C]. In: Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015: 18th international conference. Munich, Germany: Springer; 2015:234—41 pp.
29. Cui JY, Zhang H, Han H, Shan SG, Chen XL. Improving 2D face recognition via discriminative face depth estimation[C]. In: 2018 International Conference on Biometrics (ICB). Gold Coast, QLD, Australia: IEEE; 2018:140—7 pp.
30. Le T, Nguyen TV, Nie ZL, Tran MT, Sugimoto A. Anabranch network for camouflaged object segmentation. Comput Vis Image Understand 2019;184:45—56.
31. Fan DP, Ji GP, Sun GL, Cheng MM, Shen JB, Shao L. Camouflaged object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2020:2774—84 pp.
32. Lv YQ, Zhang J, Dai YC, Li A, Barnes N, Fan DP. Toward deeper understanding of camouflaged object detection. IEEE Trans Circ Syst Video Technol 2023;33:3462—76.
33. Ju R, Ge L, Geng WJ, Wu GS. Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing (ICIP). Paris, France: IEEE; 2014:1115—19 pp.