
MDG-DIS: Monocular Depth Guided Dichotomous Image Segmentation

Aditya Tewari
EUIC, Amazon
aditewar@amazon.lu

Rita Meziati Sabour
EUIC, Amazon
ritamez@amazon.es

Rui Gong
EUIC, Amazon
ruigong@amazon.lu

Iaroslav Melekhov
EUIC, Amazon
imlkhv@amazon.es

Nikolay Marin
EUIC, Amazon
nikolmrn@amazon.lu

Virginia Fernandez Arguedas
EUIC, Amazon
virfer@amazon.es

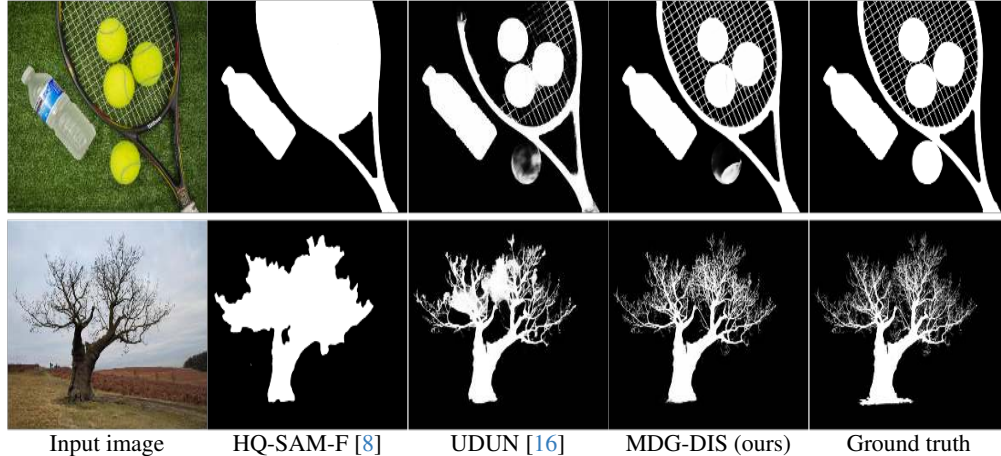
Abstract

Dichotomous Image Segmentation (DIS) aims to generate precise binary segmentation masks for objects in natural images. One of DIS applications is foreground separation. While existing class-agnostic dense prediction methods have shown promise in DIS for foreground segmentation, they struggle with overlapping objects at different depths, resulting in imprecise boundaries and misclassifications. To address these limitations, we propose MDG-DIS, a novel DIS framework that leverages depth information to resolve ambiguities in foreground separation. Our approach effectively differentiates objects at varying depths, refining object boundaries and improving segmentation accuracy in complex scenes. During training, the model learns to compensate for depth estimation errors, enhancing precision even in regions with uncertain depth information. At inference, the model operates without depth input, relying instead on learned depth-aware representations for computational efficiency. Extensive experiments on the DIS5K dataset, comprising both overlapping and non-overlapping foregrounds, demonstrate that MDG-DIS consistently outperforms existing methods of comparable size and achieves competitive results with substantially larger models. The proposed approach delivers superior segmentation accuracy while maintaining high efficiency, making it particularly suitable for real-time applications in challenging scenarios.

1 Introduction

Recent advancements in dense image prediction have significantly improved semantic segmentation [9] and class-agnostic foreground-background separation [18], powered by the robust representational capabilities of neural networks [18, 26]. This progress, coupled with precise annotations, has enabled solutions for increasingly fine-grained segmentation tasks. Among these, Dichotomous Image Segmentation (DIS) emerges as a unifying framework that addresses the fundamental challenge of separating foreground objects from their backgrounds with pixel-level precision. DIS aims to generate binary masks that precisely depict object boundaries, regardless of object category or scene complexity, encompassing challenges such as salient object detection [18] and camouflaged object detection [7]. By combining high-quality segmentation with class-agnostic foreground separation, DIS ensures robust performance across various input resolutions, making it versatile for fine-grained image segmentation applications. DIS is attracting growing interest for its potential in a range of applications, including augmented reality [19], 3D reconstruction [12] and e-commerce [17].

Existing DIS solutions have primarily focused on handling both fine-grained and coarse regions within images. Some approaches [16] incorporate structure and trunk estimation, while others utilize



(a) Comparison of different approaches on the DIS5K dataset.



(b) Qualitative results of MDG-DIS evaluated on Amazon catalogue data

Figure 1: **MDG-DIS**. The proposed approach demonstrates strong performance providing fine-grained segmentation masks on challenging public data (Top: evaluation on the DIS5K dataset) as well as on Amazon catalogue (Bottom: qualitative results of MDG-DIS on Amazon catalogue).

frequency priors [31] to distinguish fine details from coarse regions. However, two critical challenges remain in high-quality foreground segmentation: (1) existing DIS methods [19] often fail when background and salient regions share similar visual characteristics, and (2) segmentation accuracy deteriorates when fine-grained details of foreground and background overlap. To address these limitations, we introduce *MDG-DIS*, a novel monocular depth guided DIS framework that integrates depth information into the segmentation process. Unlike conventional appearance-based methods, our approach leverages depth guidance to enhance segmentation accuracy, particularly in challenging cases involving overlapping objects at different depths.

MDG-DIS extends the UDUN approach [16] by incorporating a simple yet effective depth-guided optimization scheme during training. This scheme maintains depth consistency within segmented regions, helping distinguish objects based on their spatial separation. While solutions like BAS-Net [19] and IS-Net [17] rely on intrinsic 2D properties, MDG-DIS explicitly incorporates external guidance through monocular depth estimation, introducing valuable 3D scale information. Notably, this integration incurs no additional computational cost during inference, as depth information is utilized exclusively during training. This approach enhances fine-grained segmentation performance while maintaining model efficiency.

In summary, we make the following contributions:

- We introduce MDG-DIS, a novel depth-guided DIS approach that leverages localized depth variance to enhance fine-grained segmentation, improving object delineation in complex scenes through explicit 3D spatial context;
- We propose a framework that achieves superior segmentation performance without increasing inference complexity. Extensive evaluations on the DIS5K dataset demonstrate that MDG-DIS consistently outperforms the lightweight UDUN [16] (cf. Fig. 1) and matches the performance of

the more computationally intensive BiRefNet [30], while maintaining faster inference and fewer parameters;

- We demonstrate MDG-DIS’s practical value through extensive evaluation on Amazon product images, showing its effectiveness for real-world applications in large-scale retail platforms.

2 Related Work

Image segmentation has evolved significantly with the emergence of specialized datasets and tasks targeting diverse challenges in computer vision [4, 6, 23]. A notable advancement in this field is the introduction of DIS, which focuses on achieving highly accurate segmentation for objects with complex shapes and structures [17]. Our work builds upon and extends these foundations, particularly in addressing the challenges of fine-grained segmentation with depth awareness. We organize our discussion of related work into three main areas: dichotomous image segmentation, high-resolution image segmentation, and depth-aware approaches for segmentation.

Dichotomous Image Segmentation. The introduction of the DIS5K dataset by [17] was accompanied by IS-Net, a model utilizing intermediate supervision to enhance performance on fine-grained areas. Subsequent approaches have explored various strategies to improve segmentation quality. Zhou et al. [31] introduce frequency prior integration to capture high-frequency region details, while Zheng et al. [30] propose a two-part network BiRefNet leveraging multi-scale localization with inward and outward reference reconstruction. Recent advances include MVANet [27] combining distant and close object views for training acceleration, and UDUN [16] model – a Divide-and-Conquer approach that decouples low and high-level features for an improved boundary delineation. In this work, we show that incorporating monocular depth priors allows DIS approaches to obtain significantly more detailed segmentation masks while maintaining high efficiency.

High-Resolution Image Segmentation. DIS shares a significant overlap with high-resolution semantic segmentation research. Multi-path fusion approaches have been widely adopted, with works like RefineNet [11] and ICNet [29] exploiting multi-resolution inputs and low-scale representations for efficient semantic localization. Boundary refinement techniques have been explored through explicit fine-grained estimation [28] and post-processing methods such as Conditional Random Fields (CRF) and region growing [5, 10]. The field has seen significant advancement with foundation models like SAM [9], trained on billion-scale masked labels, and its high-quality variant HQ-SAM [8], which addresses limitations in thin object boundary precision while maintaining zero-shot segmentation capabilities.

Depth Awareness for Segmentation. The integration of depth information in segmentation has a rich history, with joint optimization approaches [15] demonstrating enhanced performance across multiple dense perception tasks. Depth-aware CNNs have introduced specialized mechanisms such as depth similarities and depth-aware average pooling [24] to leverage depth information effectively. Conversely, works like [3] have shown how semantic information can enhance depth learning, highlighting the symbiotic relationship between structural and semantic understanding. The DADA framework [22] further demonstrates the value of depth information in domain adaptation scenarios for semantic segmentation. Recent developments in joint depth and semantic segmentation models [20] continue to reveal the mutual benefits of simultaneous depth and semantic estimation, reinforcing the fundamental connection between these complementary tasks.

3 Method

Our approach builds upon an existing DIS architecture by incorporating depth awareness without increasing model complexity. Starting with a network [16] pre-trained on the DIS5K dataset, represented as function $\mathcal{F}_{\mathcal{W}}(\cdot)$ with parameters $\mathcal{W}$, we enhance its performance through depth-guided optimization. The key innovation lies in refining the model parameters to $\overline{\mathcal{W}}$ by leveraging local depth inconsistencies during training, while maintaining the original parameter count. Fig. 2 illustrates the MDG-DIS framework, which augments the UDUN [16] architecture with monocular depth cues to improve segmentation mask estimation. Our method introduces depth-aware learning through a carefully designed optimization strategy that preserves the model’s efficient inference characteristics while significantly improving its segmentation capabilities.

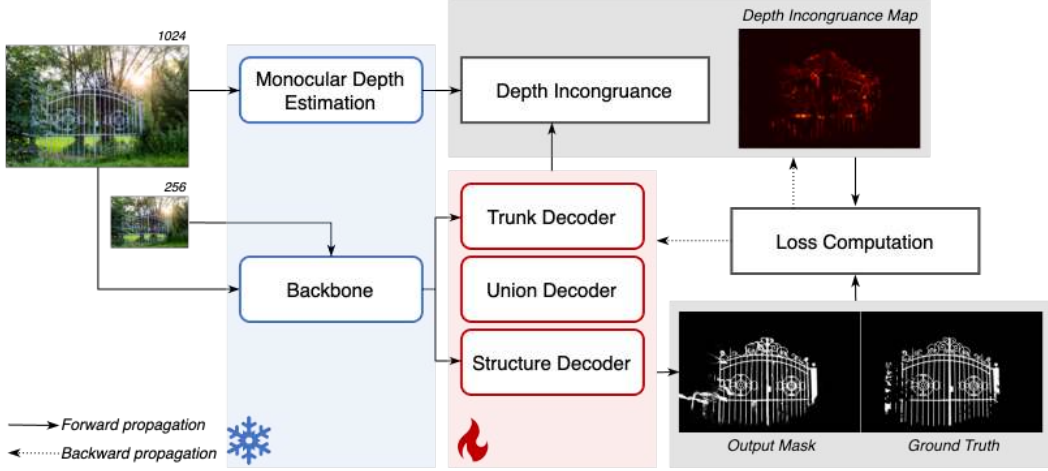


Figure 2: **Overview.** MDG-DIS leverages monocular geometric cues from a pre-trained depth estimation network [2] to guide DIS model optimization. Our approach enforces depth consistency within predicted foreground regions by computing depth inconsistency measures. The network parameters are optimized through backpropagation using both the depth inconsistency signal and traditional segmentation loss against ground-truth masks.

3.1 Multi-Scale Feature Learning with Structure-Trunk Decomposition

Our framework builds upon UDUN’s [16] dual-encoder architecture, which processes input images through parallel pathways to capture complementary object representations. This approach decomposes the segmentation task into two components: "Structure," which captures detailed object boundaries, and "Trunk," which represents the object’s skeletal framework derived from distance transforms of ground truth masks.

Structure-Trunk Formulation. The mathematical formulation for Trunk (T) and Structure (S) components is defined as:

$$T = \left(DF(\mathcal{M}_{\text{gt}}) \right)^{0.5}, \quad S = \mathcal{M}_{\text{gt}} - T \quad (1)$$

where $\mathcal{M}_{\text{gt}}$ represents the ground truth mask and DF denotes the distance transform operation.

Feature Extraction and Fusion. Using a ResNet50 backbone, we extract features from both high and low-resolution versions of the input image. These features are processed through dedicated Structure and Trunk decoders, each optimized for their respective tasks. Multi-scale feature fusion is achieved through: a) Trunk-Structure Aggregation (TSA) modules; b) Mask-Structure Aggregation (MSA) modules. These modules help cross-decoder feature fusion at corresponding scales, with the Union decoder producing the final segmentation mask.

3.2 Depth-Guided Refinement for Fine-Grained Segmentation

We address the task of distinguishing overlapping foreground and background regions by introducing depth consistency constraints. Our approach leverages local depth variations as indicators of object boundaries, utilizing a monocular depth estimator $f_{\mathcal{D}}(\cdot)$ that generates depth maps $\mathcal{D}$ from high-resolution inputs. Given a binary segmentation mask $\mathcal{M}$ (where 1 indicates foreground and 0 background), we enforce depth similarity within local foreground regions. We compute a localized variance $\mathcal{V}$ on the depth map for estimated foreground regions, and aggregate it into $\mathcal{L}_{\mathcal{V}}$ across image dimensions (H, W) . We use $\mathcal{L}_{\mathcal{V}}$ as a depth loss:

$$\mathcal{L}_{\mathcal{V}} = \sum_W \sum_H \mathcal{V}(\mathcal{M} \odot \mathcal{D}) \quad (2)$$

The local variance $\mathcal{V}_{i,j}$ for each pixel (i,j) is computed over a neighborhood patch, considering only foreground pixels:

$$\mathcal{V}_{i,j} = \frac{\sum_{i',j'} (d_{i',j'} - \frac{\sum_{i',j'} d_{i',j'} m_{i',j'}}{\sum_{i',j'} m_{i',j'}})^2 * m_{i',j'}}{\sum_{i',j'} m_{i',j'}} \quad (3)$$

where $d_{i',j'}$ and $m_{i',j'}$ are elements from depth map D and mask M patches, respectively. Fig. 2 (upper-right corner) visualizes the local depth variance $\mathcal{V}$ computed using ZoeDepth [2] estimates.

We extend the original UDUN training framework [16], which uses:

$$\mathcal{L}_{\text{orig}} = \mathcal{L}_{\text{trunk}} + \mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{mask}} \quad (4)$$

Here, $\mathcal{L}_{\text{trunk}}$, $\mathcal{L}_{\text{struct}}$, and $\mathcal{L}_{\text{mask}}$ represent the losses computed at the outputs of the Trunk, Structure, and Union decoders, respectively. The outputs of the Trunk and Structure decoders are compared to their corresponding ground truth masks minimizing Binary Cross-Entropy (BCE) losses. The mask loss $\mathcal{L}_{\text{mask}}$ combines BCE and Intersection over Union (IoU) terms:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{iou}} \quad (5)$$

Our final loss incorporates the depth inconsistency term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{trunk}} + \mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\mathcal{V}} \quad (6)$$

During training, the depth inconsistency gradients propagate through the entire network, updating parameters in all decoders (Structure, Trunk, and Union) of the UDUN [16] architecture.

Implementation details. Our implementation builds upon the training scheme proposed in [16]. We use ResNet-50 as the backbone network for extracting both high and low-resolution features from input images, ensuring fair comparison with UDUN [16]. The network processes two input scales: a high-resolution input of 1024×1024 pixels and a low-resolution input of 256×256 pixels for feature extraction. Depth inconsistency measures are computed using monocular depth estimates from ZoeDepth [2]. The training process consists of two phases spanning 48 epochs in total. During the first 24 epochs, we train only the UDUN-based pipeline. The network head is optimized using a learning rate of 0.05. In the second phase, we incorporate the depth incongruence-based optimization using Adam optimizer with a learning rate twice that of the main network. The entire framework is implemented in PyTorch and trained on a single NVIDIA A10G GPU with 24GB memory.

4 Experiments

In this section, we perform extensive evaluations of the proposed approach on real-world data for the DIS task. Next, we provide qualitative and quantitative comparisons against strong DIS baselines. Finally, we demonstrate the practical applicability of our method by evaluating its performance on real-world e-commerce product images.

Datasets. Following the established protocol [16, 17, 19], we evaluate our approach on the DIS5K dataset [17]. The dataset comprises a training set (DIS-TR) of 3000 images and a testing set (DIS-TE) of 2000 images. DIS-TE is further divided into four equal subsets (DIS-TE- i , $i \in \{1, 2, 3, 4\}$), each containing 500 images categorized by foreground structure complexity and shape. We report both the average performance across all subsets and individual results for each DIS-TE- i .

Baselines. We compare against a) dichotomous image segmentation methods: UDUN [16], IS-Net [31], and BASNet [19]; b) Transformer-based segmentation approach: BiRefNet [30].

Metrics. We assess the performance of our method through metrics commonly used for class agnostic segmentation. Specifically, we report F-measure F_{max}^{β} [1], mean F-measure F_{mean}^{β} , and the Mean Absolute Error (MAE). To evaluate the segmentation quality on complex shapes, we also provide the Human Correction Effort (HCE_{γ}) required to correct the computed segmentation mask errors, as in [16]. The value of γ is set to 5 across all experiments.

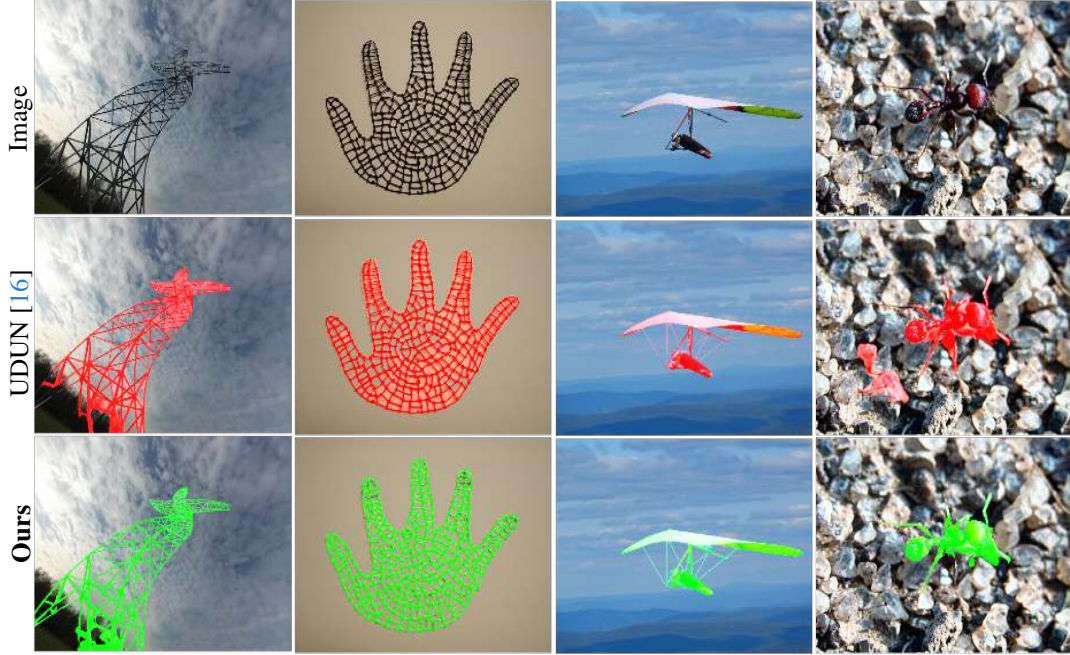


Figure 3: **Qualitative results on DIS-TE.** Comparison of segmentation results between MDG-DIS and UDUN on DIS-TE-1 (first two columns) and DIS-TE-4 (last two columns) datasets. Our approach demonstrates superior performance in preserving both structural integrity and fine details across varying complexity levels. Additional examples are provided in the supplementary material.

Method	Dataset									
	DIS-TE-1		DIS-TE-2		DIS-TE-3		DIS-TE-4		DIS-TE-all	
	$F_{\max}^{\beta} \uparrow$	$F_{\text{mean}}^{\beta} \uparrow$	$F_{\max}^{\beta} \uparrow$	$F_{\text{mean}}^{\beta} \uparrow$	$F_{\max}^{\beta} \uparrow$	$F_{\text{mean}}^{\beta} \uparrow$	$F_{\max}^{\beta} \uparrow$	$F_{\text{mean}}^{\beta} \uparrow$	$F_{\max}^{\beta} \uparrow$	$F_{\text{mean}}^{\beta} \uparrow$
IS-Net [17]	.663	.622	.799	.774	.830	.758	.827	.770	.779	.731
BASNet [19]	.740	.560	.738	.662	.790	.733	.785	.706	.761	.664
UDUN [16]	.766	.590	.807	.689	.829	.827	.807	.724	.791	.677
MDG-DIS (ours)	.833	.710	.856	.774	.871	.869	.857	.795	.854	.767

(a) $F_{\max}^{\beta}$ and F_{mean}^{β} metrics (the higher the better)

Method	Dataset									
	DIS-TE-1		DIS-TE-2		DIS-TE-3		DIS-TE-4		DIS-TE-all	
	MAE $\downarrow$	HCE $_{\gamma}$ $\downarrow$	MAE $\downarrow$	HCE $_{\gamma}$ $\downarrow$	MAE $\downarrow$	HCE $_{\gamma}$ $\downarrow$	MAE $\downarrow$	HCE $_{\gamma}$ $\downarrow$	MAE $\downarrow$	HCE $_{\gamma}$ $\downarrow$
IS-Net [17]	.074	155	.070	341	.072	681	.070	2888	.072	1016
BASNet [19]	.105	149	.096	340	.087	687	.092	2852	.095	1007
UDUN [16]	.063	188	.057	327	.053	676	.053	2505	.057	924
MDG-DIS (ours)	.041	142	.040	371	.038	558	.038	2669	.039	934

(a) MAE and HCE $_{\gamma}$ metrics (the lower the better)

Table 1: **Quantitative results on DIS-TE.** Performance evaluation using standard DIS metrics (see Section 4) across individual and combined splits of the DIS-TE dataset. MDG-DIS demonstrates consistent improvement in segmentation quality across all scene complexity levels.

4.1 Performance Analysis: DIS Approaches

To show the effectiveness of our method for dichotomous image segmentation, we compare against various baselines on the challenging DIS5K dataset.

Quantitative Results. We evaluate MDG-DIS against UDUN [16] and IS-Net [17] using input images of size 1024×1024 and a ResNet-50 backbone. As shown in Table 1, the proposed approach demonstrates significant improvements across multiple metrics. On the complete DIS-TE dataset,

MDG-DIS achieves an 8% higher $F_{\max}^{\beta}$ compared to UDUN, while maintaining superior performance over IS-Net and BASNet [19]. Notably, on the challenging DIS-TE-4 subset, MDG-DIS outperforms UDUN by 6% in $F_{\max}^{\beta}$ and reduces MAE by 30%. These results consistently demonstrate the effectiveness of our depth-guided approach across varying levels of segmentation complexity.

Qualitative Analysis. The integration of depth-guided gradients significantly improves segmentation quality, particularly in regions with overlapping objects. Our method produces sharper edges and more precise corner delineation across both simple (DIS-TE-1) and complex (DIS-TE-4) scenarios. Fig. 3 illustrates this improvement, comparing MDG-DIS outputs with UDUN baseline results. The enhancement is particularly evident in architectural objects with distinct depth variations, such as electrical poles, where our method maintains structural integrity while capturing fine details.

Performance Analysis on Complex Cases. While MDG-DIS achieves superior $F_{\max}^{\beta}$ and MAE scores across all test sets, we observe slightly higher HCE values on the DIS-TE-2 and DIS-TE-4 subsets. Further investigation reveals that this increase in Human Correction cost arises from specific challenging scenarios where objects have intricate overlapping components. For instance, as shown in Fig. 4, images containing intersecting structures (e.g. bridge cables) with high local depth variances can lead to increased False Negative predictions. However, even in these challenging cases, MDG-DIS maintains better fine-grained detail preservation compared to UDUN [16] (cf. Fig. 4). These examples highlight the trade-off between precise boundary detection and handling complex overlapping structures.

4.2 Efficiency and Performance Trade-offs

We compare MDG-DIS against BiRefNet [30], a recent transformer-based approach that employs inward scale reference and outward gradient priors, in contrast to our strategy of utilizing depth-based gradient priors. While the SWIN [13] transformer-based BiRefNet achieves superior segmentation quality compared to our method, it comes with significant computational overhead. As shown in Table 2, BiRefNet with SWIN backbone requires an average runtime of 83.3ms, which is 2.37 times slower than our ResNet50-based solution. BiRefNet’s alternative implementation using PVTv2 [25] backbone achieves more competitive inference speed (47.8ms). However, MDG-DIS maintains comparable segmentation quality while offering faster inference time (33ms), fewer model parameters, and competitive segmentation accuracy. These results, detailed in Table 2, demonstrate that MDG-DIS achieves an effective balance between performance and computational efficiency, making it particularly suitable for real-world applications where both accuracy and speed are crucial.

Model	$F_{\max}^{\beta} \uparrow$	Params, MB $\downarrow$	Runtime, ms $\downarrow$
BiRefNet _{swinT}	0.890	215	81.3
BiRefNet _{pvtv2}	0.878	35	47.8
UDUN	0.791	25	33
MDG-DIS	0.854	25	33

Table 2: **Efficiency-Performance Trade-off.** We report $F_{\max}^{\beta}$ over the DIS-TE-all split.

4.3 Application to E-commerce Product Segmentation

MDG-DIS demonstrates significant potential for enhancing several Amazon customer experiences (CXs), particularly in applications such as Color Swatches¹ and Visual Cues². To validate its practical utility, we evaluated MDG-DIS on a diverse dataset of 2000 product images from the Amazon Catalog, encompassing various product categories and challenging scenarios such as complex backgrounds, multiple objects, and fine details. Our experiments show (cf. Fig. 1 and Section C.2 in the supplementary material) that MDG-DIS achieves good segmentation quality on Amazon images. Notably, MDG-DIS excels in handling challenging cases common in e-commerce imagery, such as products with intricate details (e.g., jewelry and fashion items). The model’s ability to leverage depth information proves particularly valuable in accurately separating products from visually similar backgrounds, a common challenge in product photography.

¹<https://w.amazon.com/bin/view/EUHIP/ML/LaserTag/>

²<https://w.amazon.com/bin/view/InTech/NintAI/Projects/VisualCues/>

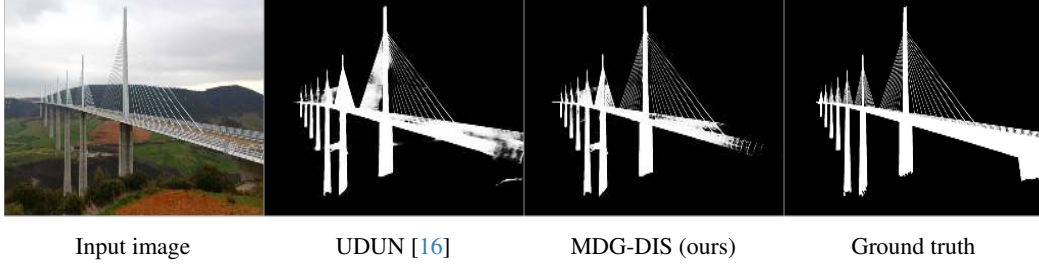


Figure 4: **Qualitative results: challenging cases.** MDG-DIS leverages monocular depth cues to enhance segmentation quality, producing fine-grained masks with sharper details compared to UDUN [16]. While it generally improves boundary precision, it may generate false negatives in regions with minimal depth variation.

5 Conclusion

We present MDG-DIS, a novel framework that enhances dichotomous image foreground segmentation by incorporating monocular depth guidance into multi-scale 2D feature fusion. Our comprehensive evaluation of the DIS task demonstrates significant improvements over existing approaches, achieving an 8% increase in $F_{\max}^{\beta}$ compared to the baseline UDUN model on the DIS5K dataset. The success of MDG-DIS in leveraging depth information for improved segmentation accuracy validates the effectiveness of integrating multiple dense perception signals for complex vision tasks implying foreground separation. The demonstrated improvements in segmentation quality, particularly in challenging scenarios with overlapping objects and complex depth relationships, suggest promising directions for future research. Specifically, we envision extending this work toward unified architectures that simultaneously address high-quality segmentation and fine-grained depth estimation while maintaining computational efficiency through shared-backbone networks.

6 Customer Problem Statement

MDG-DIS addresses critical challenges in real-world e-commerce applications, particularly in content moderation, product evaluation and product catalog management. In large-scale marketplaces like Amazon, ensuring high-quality product presentations requires accurate identification and segmentation of multiple salient objects within product images. Traditional approaches struggle with complex scenarios where products overlap or share similar visual characteristics with backgrounds, leading to suboptimal customer experiences. MDG-DIS’s ability to leverage depth information for precise object delineation makes it particularly valuable for automated content moderation systems. The framework can effectively identify and segment multiple salient objects in product images, leading to a better quality control of seller-uploaded content and improving the overall customer shopping experience. Moreover, its computational efficiency makes it suitable for large-scale deployment, where processing millions of images daily requires both accuracy and speed. This practical application demonstrates how advances in computer vision can directly translate into enhanced customer experiences in e-commerce platforms.

References

- [1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 5
- [2] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4, 5
- [3] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

- [4] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 3
- [5] P. A. Dias and H. Medeiros. Semantic segmentation refinement by monte carlo region growing of high confidence detections. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 131–146. Springer, 2019. 3
- [6] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021. 3
- [7] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1
- [8] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 1
- [9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3
- [10] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 3
- [11] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3
- [12] F. Liu, L. Tran, and X. Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433, 2021. 1
- [13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7
- [14] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang. Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer, 2022. 1, 2
- [15] A. Mousavian, H. Pirsiavash, and J. Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619, 2016. 3
- [16] J. Pei, Z. Zhou, Y. Jin, H. Tang, and P.-A. Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2139–2147, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [17] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 1, 2, 3, 5, 6
- [18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 1
- [19] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019. 1, 2, 5, 6, 7
- [20] M. Shvets, D. Zhao, M. Niethammer, R. Sengupta, and A. C. Berg. Joint depth prediction and semantic segmentation with multi-view sam. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1328–1338, 2024. 3
- [21] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 1
- [22] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 3
- [23] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 3

- [24] W. Wang and U. Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018. [3](#)
- [25] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022. [7](#)
- [26] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. [1](#)
- [27] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu. Multi-view aggregation network for dichotomous image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3921–3930, June 2024. [3](#)
- [28] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 489–506. Springer, 2020. [3](#)
- [29] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018. [3](#)
- [30] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. [3](#), [5](#), [7](#)
- [31] Y. Zhou, B. Dong, Y. Wu, W. Zhu, G. Chen, and Y. Zhang. Dichotomous image segmentation with frequency priors. In *IJCAI*, volume 1, page 3, 2023. [2](#), [3](#), [5](#)

Supplementary Material for MDG-DIS: Monocular Depth Guided Dichotomous Image Segmentation

In this **supplementary material**, we provide an additional ablation study for another type of geometric guidance based on edge detection in Section A, assess the performance of a zero-shot prompt-based segmentation framework we define as HQ-SAM-F in Section B and report additional qualitative segmentation results on the DIS5K dataset and Amazon catalogue data in Section C.

A Ablation Study

In this section, we present an analysis of edge detection guidance for the task of DIS.

Analysis of External Scale Guidance. We investigate the potential benefits of incorporating external scale information into MDG-DIS through edge-aware guidance. For this analysis, we utilize PiDiNet [21] to generate edge maps $\mathcal{E}$ using pixel difference convolution-based edge detection. To integrate edge information into the segmentation process, we define an edge-guided correction term that focuses on boundary regions of the output mask $\mathcal{M}$. The correction signal is computed as a weighted combination of the edge map and mask:

$$\mathcal{L}_{\mathcal{E}} = \sum_W \sum_H (\mathcal{M} \odot \mathcal{E} \odot \mathbf{w}) \quad (7)$$

We implement an interleaved training approach where edge guidance is applied through the "Structure" encoder and batches alternate between depth guidance and edge guidance. Our experiments reveal that external scale guidance does not enhance the UDUN model's [16] performance, as evidenced by the learnable weights converging to zero during training.

B Zero-shot Prompt-based Segmentations

In this section, we propose a zero-shot DIS pipeline based on the HQ-SAM foundational model, and evaluate its performance against MDG-DIS.

To benchmark MDG-DIS against modern foundation models, we developed HQ-SAM-F, a pipeline combining HQ-SAM [8] with class-agnostic object detection for zero-shot foreground segmentation. We evaluate this approach on the DIS5K test datasets.

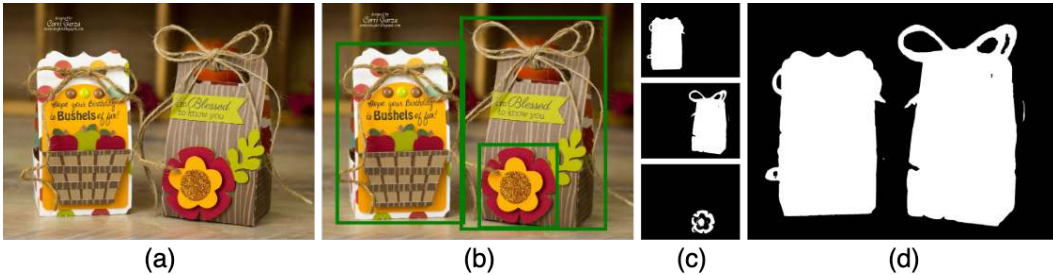


Figure 5: **HQ-SAM Foreground Pipeline.** Our baseline combines class-agnostic object detection [14] and high-quality segmentation through a two-stage process: (a) input image is processed by MAVL [14] for foreground object detection, (b) detected objects are represented as bounding boxes, (c) these boxes serve as prompts for HQ-SAM [8] to generate individual segmentation masks, and (d) the final mask is obtained through logical OR operation over all individual masks.

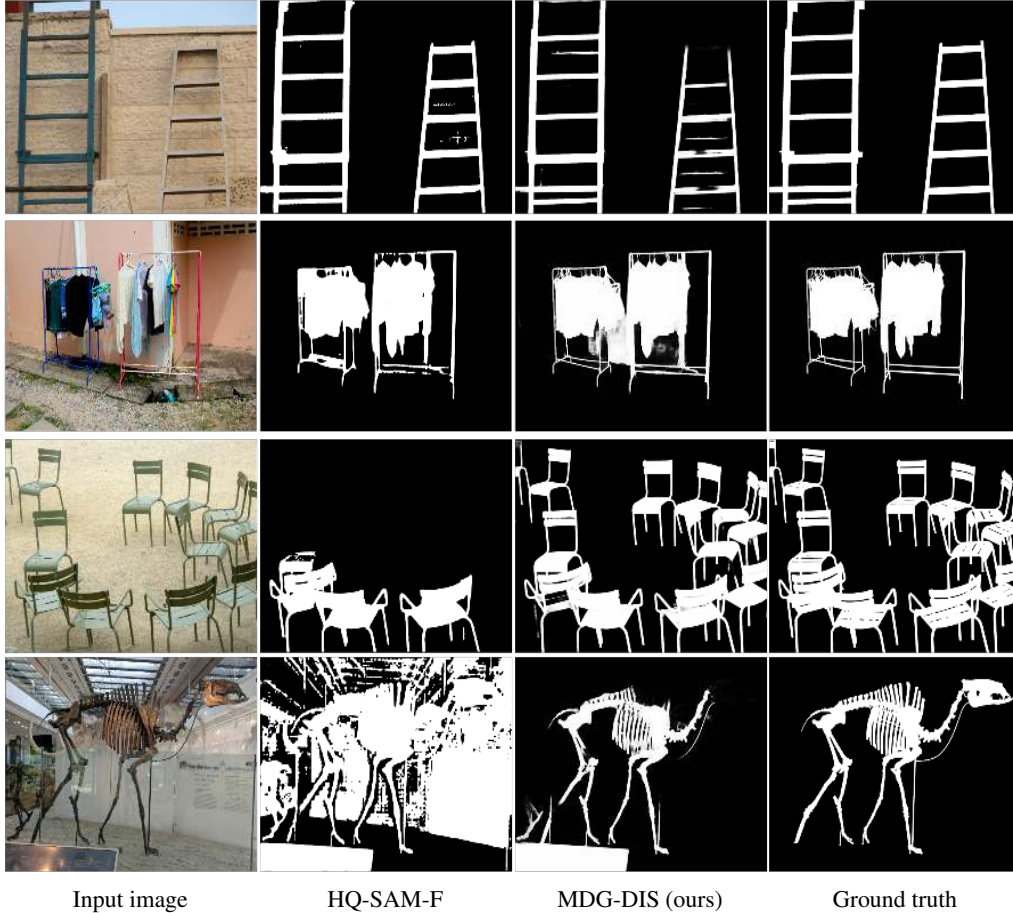


Figure 6: **Qualitative results.** Qualitative comparison between masks generated by MDG-DIS and HQ-SAM-F. The proposed approach provides fine-grained and more accurate segmentation results.

Method	DIS-TE-1	DIS-TE-2	Dataset		
			DIS-TE-3	DIS-TE-4	DIS-TE-all
HQ-SAM-F [8]	0.170	0.150	0.130	0.130	0.140
MDG-DIS	0.041	0.040	0.038	0.038	0.039

Table 3: **Quantitative results.** We report the MAE metric for foreground detection obtained by HQ-SAM-F and MDG-DIS on the separate and combined splits of the DIS-TE dataset.

HQ-SAM-F Pipeline. Our implementation utilizes HQ-SAM’s multi-box prompt functionality, which accepts bounding box coordinates $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ as segmentation guidance. To generate class-agnostic bounding boxes, we employ the MAVL [14] model with the text prompt “*only foreground objects*”. The final mask for each image is created by combining individual HQ-SAM segmentation outputs for detected foreground objects. The complete pipeline is illustrated in Fig. 5.

Performance Analysis. Fig. 6 presents a comparative analysis of masks generated by HQ-SAM-F and MDG-DIS against the ground truth, showcasing both high-quality and challenging cases. The results demonstrate that HQ-SAM-F’s performance is heavily dependent on the accuracy of the initial class-agnostic object detection step. As shown in Table 3, MDG-DIS consistently outperforms HQ-SAM-F across all DIS-TE- i datasets in terms of MAE. We attribute this superior performance to several factors: a) MDG-DIS’s effective utilization of depth information for boundary refinement; b) HQ-SAM-F’s reliance on pre-trained weights [8] without DIS5K-specific fine-tuning; c) Limitations

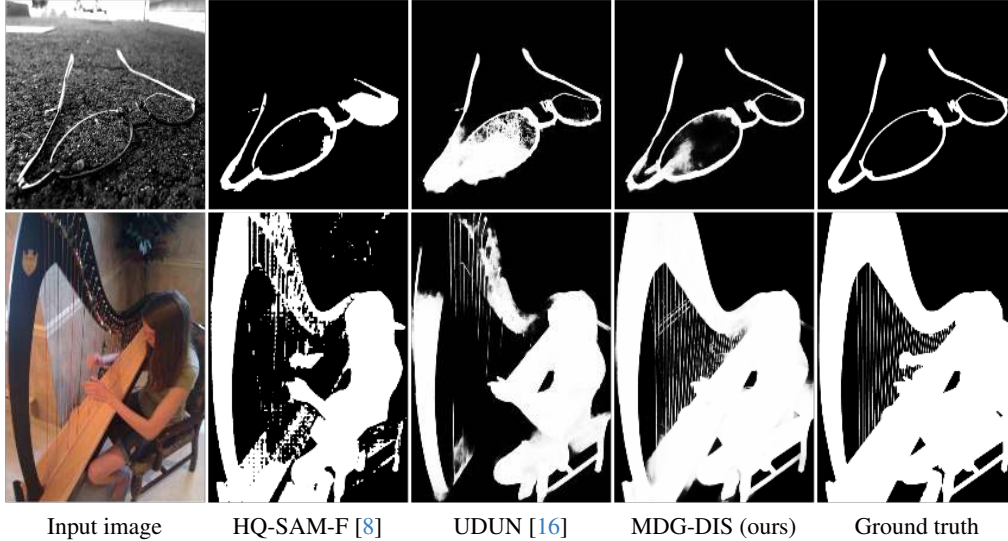


Figure 7: **Qualitative results on DIS5K.** Comparison between masks generated by HQ-SAM-F, UDUN, and MDG-DIS. Our approach provides fine-grained and more accurate segmentation results.

of MAVL-based detection in complex scenarios, particularly with multiple objects at varying spatial locations and foreground objects overlapping with detailed backgrounds.

C Additional Results

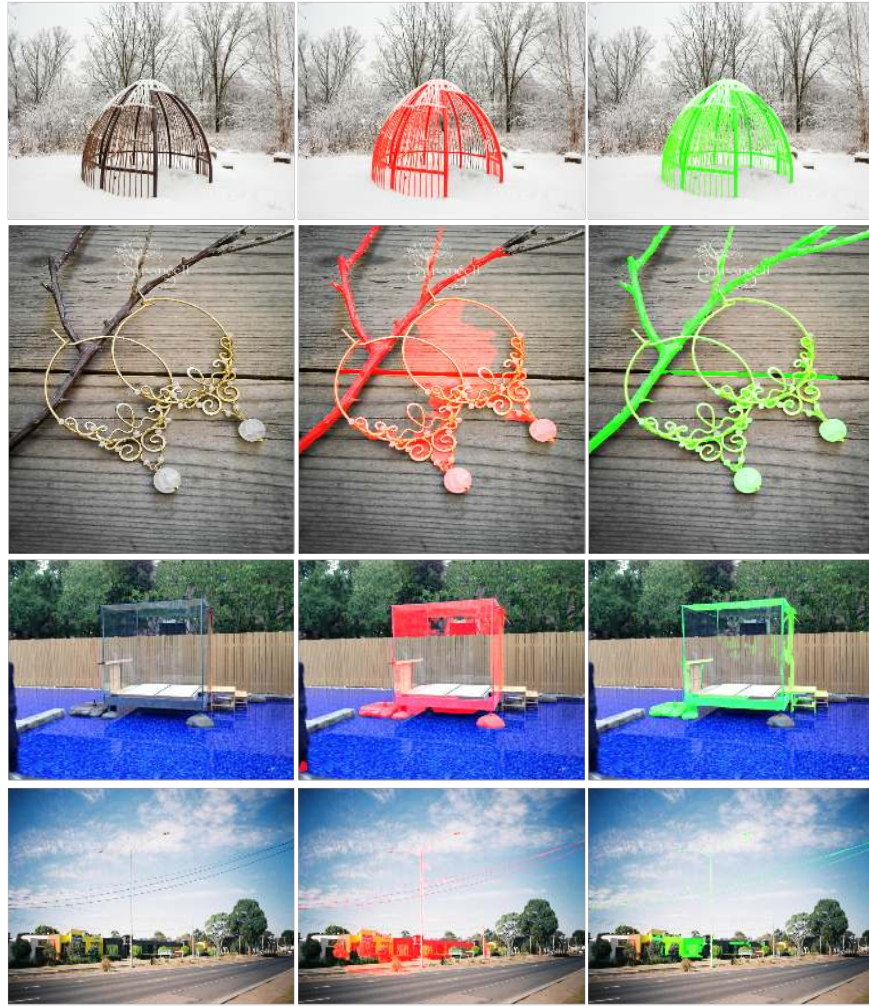
In this section, we present more qualitative results for the DIS5K dataset (Section C.1) and Amazon catalogue data (Section C.2).

C.1 DIS5K

We report qualitative results for DIS5K and show more visualizations in Fig. 7 and Fig. 8. Compared to state-of-the-art methods, our approach with monocular depth guidance produces significantly better segmentation results both visually as well as quantitatively.

C.2 Amazon Catalogue

MDG-DIS was tested on images from Amazon catalogue for the following common scenarios: product images with a white background and a single foreground object in Fig. 9, product images with multiple objects in Fig. 10, lifestyle images showcasing a product in a realistic environment in Fig. 11, and more challenging cases in Fig. 12. The latest category comprises images with objects having reflective surfaces such as mirrors and metallic objects, images with objects having coloured areas similar to the background’s color or transparent materials, and media covers with intricate visual and textual elements such as book covers and movie posters. Overall, MDG-DIS gives high-quality segmentation masks even on complex-detailed objects, but particularly struggles in segmenting objects with reflective areas. Improvements such as mask refinement and retraining on additional difficult images are future work perspectives that can lead to better results.



Input image UDUN [16] MDG-DIS (ours)

Figure 8: **UDUN vs. MDG-DIS.** MDG-DIS produces fewer errors on edges than UDUN.



Figure 9: **Qualitative results of MDG-DIS on white-background images with a single main object.** We tested MDG-DIS on a range of objects from simple-shaped products (eg. shoes) to objects with more complex structures such as thin metallic bars, and obtained qualitative segmentation masks as shown in the figure.



Figure 10: **Qualitative results of MDG-DIS on images with multiple objects.** We tested MDG-DIS on images with more than one foreground element, including cases of overlaying objects and/or with intricate details. Overall, MDG-DIS gives reasonable-quality masks and improvements to refine the masks is one of our future perspectives.

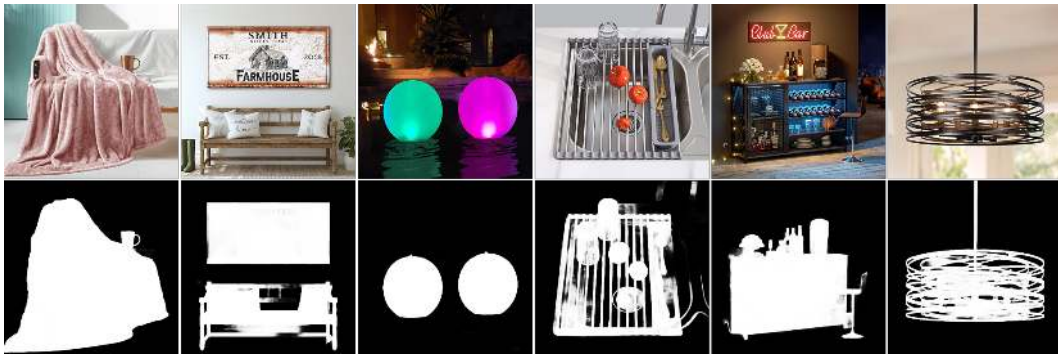


Figure 11: **Qualitative results of MDG-DIS on lifestyle images.** We tested MDG-DIS on lifestyle images where a product is shown in an in-use scenario. This type of images have non-uniform backgrounds, highlighting the capability of MDG-DIS to correctly separate foreground objects, even with complex structures and details.

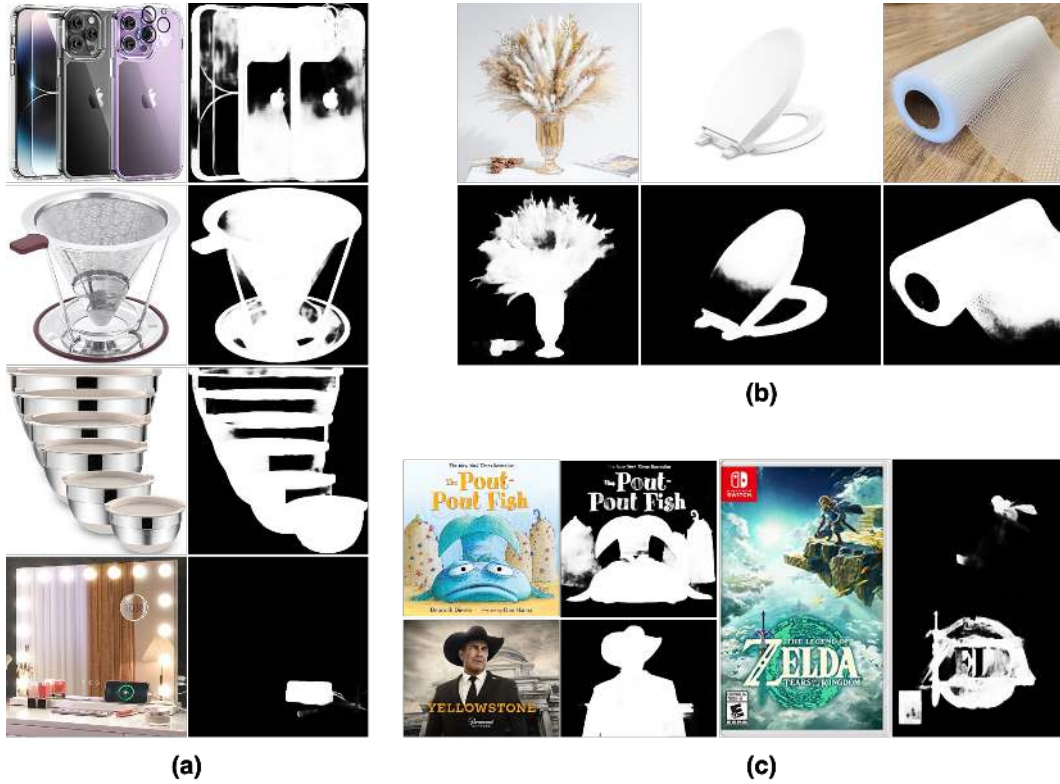


Figure 12: **Qualitative results of MDG-DIS on challenging images.** We tested MDG-DIS on images with more difficult elements to segment: **(a)** objects with reflective surfaces, **(b)** objects with transparent areas or with parts similarly coloured as the background, and **(c)** media covers with overlaying graphics and texts. The most challenging segmentation tasks are related to reflective surfaces such as metals and mirrors, creating a discontinuity in the foreground area. Retraining MDG-DIS on additional data of this particular type can lead to better results.