DISCARDING SENSOR DEPTH: RELIABILITY-AWARE GEOMETRY LEARNING FOR SALIENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-D salient object detection (SOD) complements RGB appearance with scene geometry, but practical depth measurements are often noisy, incomplete, or structurally misleading, making tightly coupled RGB-depth models vulnerable to negative transfer. We propose GeoDistill, a depth-free RGB-D SOD framework that learns task-oriented geometry from RGB instead of consuming sensor depth. During training, a frozen foundation depth estimation model transfers dense relative structure, multi-scale feature organization, and boundary cues to an RGB-driven geometry branch, addressing how useful geometry is acquired without dataset depth. Cross-modal enhancement then exchanges appearance and geometry across scales, while reliability-aware fusion suppresses geometric responses that disagree with RGB evidence or are irrelevant to saliency, addressing how the learned geometry should be used. The external teacher is removed after training and sensor depth is never required by the final detector. Across nine RGB-D benchmarks, GeoDistill is best or tied-best in 26 of 36 metric–dataset comparisons, with the clearest gains on cross-dataset benchmarks. Results on four RGB SOD benchmarks further confirm that the learned geometry transfers beyond RGB-D sensor domains. Overall, task-adapted geometry retains the structural benefit of depth while avoiding direct dependence on unreliable measurements.

1 INTRODUCTION

Salient object detection (SOD) aims to localize the most visually distinctive objects in a scene and supports downstream tasks such as image understanding, editing, retrieval, and segmentation. RGB SOD has advanced through multi-scale aggregation, boundary modeling, feedback refinement, and transformer-based context reasoning (Borji et al., 2015; Wang et al., 2022; Bao et al., 2026). RGB-D SOD further introduces depth as a geometric cue: when reliable, depth discontinuities can separate objects with similar appearance, recover weak RGB boundaries, and improve foreground–background discrimination (Fu et al., 2020; Pang et al., 2020; Kupyn et al., 2026).

The central difficulty is that geometric usefulness does not guarantee measurement reliability. Public RGB-D benchmarks combine heterogeneous sensors and reconstruction pipelines, and their depth maps often contain missing regions, foreground–background bleeding, weak contrast, or structured artifacts. Once these measurements are encoded and repeatedly fused with RGB features, depth errors can propagate through the detector and cause negative transfer. Existing methods mitigate this problem through uncertainty estimation, modality-quality calibration, filtering, or selective fusion (Zhang et al., 2020; Ji et al., 2021; Fan et al., 2021). Other methods remove sensor depth at inference or replace it with monocularly estimated depth. Although effective in specific settings, these approaches still treat a depth map—measured or predicted—as the representation that the SOD network must consume or reproduce.

Figure 1 motivates a different question: rather than making an unreliable depth map safer to use, can an RGB model acquire the geometry needed by SOD without depending on that map? This creates two coupled challenges. First, useful scene structure and object boundaries must be learned from RGB without sensor-depth supervision. Second, even geometrically valid structures—such as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

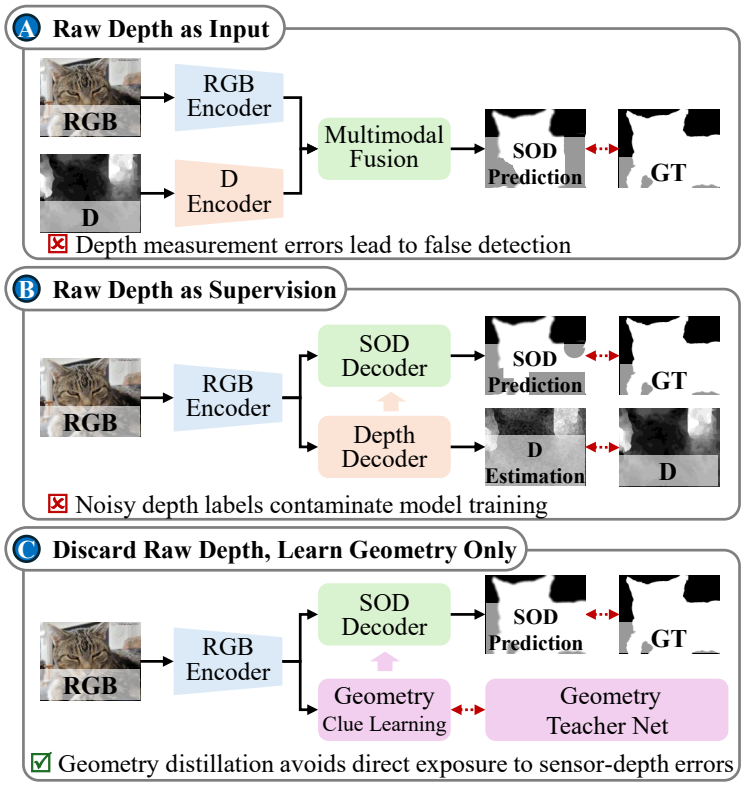

Figure 1: Three geometry-use paradigms. (A) Raw-depth fusion directly exposes predictions to sensor errors. (B) Raw-depth supervision removes depth at inference but can transfer measurement noise into geometry learning. (C) GeoDistill instead distills task-oriented geometry from a training-only foundation teacher and selectively uses the learned geometry, eliminating sensor-depth dependence at both training and inference.

walls, ground planes, or background discontinuities—should not influence saliency when they are inconsistent with foreground evidence.

We address these challenges by separating *geometry acquisition* from *geometry utilization*. During training, a frozen foundation depth estimation model transfers dense relative structure, intermediate multi-scale organization, and boundary cues to a compact RGB-driven geometry branch. Cross-modal enhancement then exchanges geometry and appearance cues across the feature pyramid, while reliability-aware fusion estimates where the learned geometry is consistent with RGB evidence before injecting it into saliency decoding. Dataset depth is never used, and the external teacher is removed after training.

We call this setting *depth-free RGB-D SOD*: standard RGB-D training/test splits and competitors are retained, but the detector itself operates from RGB. Our contributions are threefold:

- We reformulate RGB-D SOD around RGB-driven geometry learning, avoiding depth-induced negative transfer at its source rather than calibrating sensor depth after it enters the detector.
- We develop multi-level geometry distillation, cross-modal enhancement, and reliability-aware fusion to learn task-oriented structure and regulate when it should affect appearance.
- Extensive RGB-D/RGB experiments show strong cross-dataset generalization and identify intermediate feature distillation as the primary effective geometry-supervision signal.

2 Related Work

RGB and RGB-D salient object detection. RGB SOD combines contextual reasoning with structure preservation through attention, partial decoding, boundary supervision, feedback, and object-level enhancement (Liu & Han, 2018; Wu et al., 2019; Qin et al., 2019; Zhao et al., 2019; Wei et al., 2020; Wang et al., 2023). RGB-D methods extend these designs with dual-stream or progressive cross-modal interaction, including PCFNet, JL-DCF, BBS-Net, HDFNet, and RD3D (Chen & Li, 2018; Fu et al., 2020; Fan et al., 2020; Pang et al., 2020; Chen et al., 2021). Most assume that depth is available as an input modality; our method instead learns geometry from RGB and therefore avoids dependence on sensor depth.

Unreliable or unavailable depth. Prior work models depth uncertainty, calibrates modality quality, filters unreliable measurements, distills depth for efficient inference, or removes depth at test time (Zhang et al., 2020; Ji et al., 2021; Fan et al., 2021; Piao et al., 2020; Zhang et al., 2022b). Recent transformer, self-supervised, and diffusion frameworks further strengthen multimodal interaction (Zhao et al., 2022; Pang et al., 2023; Wu et al., 2023; Zhang et al., 2025). SATNet (Duan et al., 2025) replaces sensor depth with a monocular single-channel prior, but that estimated map still enters the detector as an explicit modality. In contrast, GeoDistill internalizes geometry in an RGB-driven branch, so neither sensor depth nor an external depth estimator is required at inference.

Foundation geometry and distillation. Knowledge distillation can transfer predictions, intermediate features, or attention structure from a high-capacity teacher to a compact student (Hinton et al., 2015; Romero et al., 2015; Zagoruyko & Komodakis, 2017). Large self-supervised encoders and foundation depth estimation models provide transferable scene structure (Oquab et al., 2024; Yang et al., 2024a;b). We use such a model as a training-only geometry teacher, transferring complementary output-, feature-, and boundary-level information instead of reproducing only a final pseudo-depth map. AETP and ESC-style operators from ESCNet (Ye et al., 2025) provide established edge-aware decoding primitives; our contribution is to use them within a teacher-guided geometry student and connect that student to reliability-controlled SOD.

3 Method

3.1 Overview

As shown in Figure 2, GeoDistill follows three stages: geometry acquisition, geometry–appearance interaction, and reliability-controlled utilization. A shared RGB encoder feeds an RGB-driven geometry branch whose dense output, intermediate features, and boundaries are supervised by a frozen foundation depth estimation model during training. Cross-modal enhancement exchanges the learned geometry with appearance at multiple scales, and reliability-aware fusion regulates how much geometry enters the saliency decoder. Hence, distillation determines *what* geometry is learned, while interaction and fusion determine *where* it is useful. The teacher is removed after training; all learned RGB-driven components remain.

3.2 Geometry Distillation

Shared projection. For an RGB image $I \in \mathbb{R}^{3\times H\times W}$, the backbone produces four levels $\{X_i\}_{i=1}^4$ at strides $\{4, 8, 16, 32\}$. Independent 1×1 convolution–BN–ReLU projectors map backbone-specific channels to a common width C:

$$A_i = P_i(X_i), \qquad A_i \in \mathbb{R}^{C\times H_i\times W_i}. \tag{1}$$

The common width decouples downstream geometry/fusion modules from backbone-specific channels. We use $C = 128$ by default and ablate $C \in \{64, 128, 256\}$.

Training-only teacher. The framework requires a teacher that exposes dense geometry and intermediate representations, but its design is not tied to a specific estimator. In the main experiments, $\mathcal{T}$ is instantiated with Depth Anything V2-Small, from which we extract the DPT output and four DINOv2 features:

$$(D^T, \{T_i\}_{i=1}^4) = \mathcal{T}(I). \tag{2}$$

D^T is treated as per-image normalized relative geometry, while all teacher parameters remain frozen.

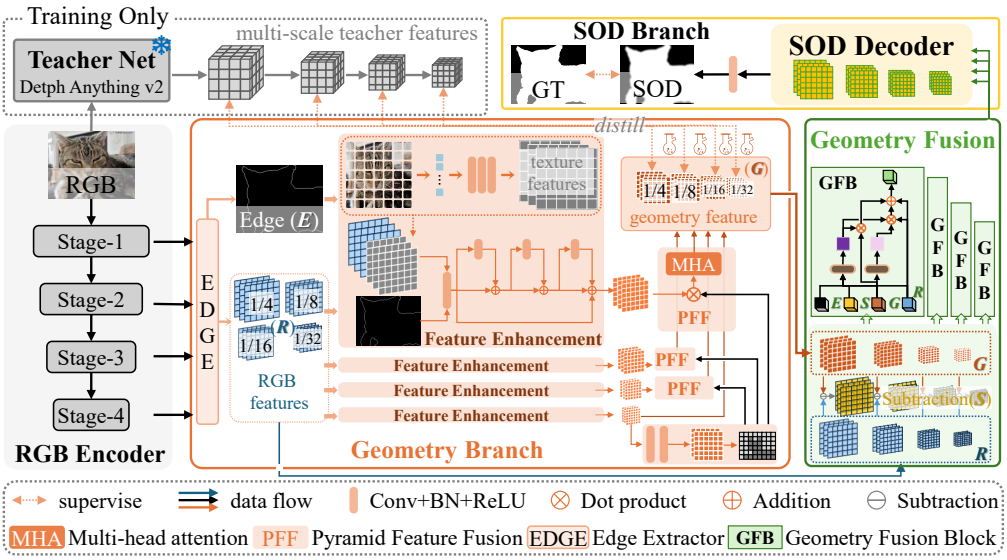

Figure 2: Overview of GeoDistill. A training-only foundation depth teacher provides output-, feature-, and boundary-level supervision to an RGB-driven geometry branch. Cross-modal enhancement exchanges geometry with appearance, while reliability-aware fusion selectively injects geometry into SOD. The teacher is removed at inference, so the final detector remains RGB-only.

Geometry branch. The student $\mathcal{G}$ predicts intermediate geometry, four geometry features, and an edge logit:

$$(\{D_k\}_{k=1}^{K}, \{G_i\}_{i=1}^{4}, E^G) = \mathcal{G}(I, \{A_i\}). \tag{3}$$

An AETP edge extractor and an ESC decoder (Ye et al., 2025) combine shallow detail with deep semantics to reconstruct task-oriented relative geometry rather than calibrated metric depth.

Dense output supervision. After min–max normalization, teacher geometry $\widetilde{D}^T$ supervises every decoder prediction through value and gradient consistency:

$$\mathcal{L}_d = \sum_{k=1}^{K} \omega_k \left(\|\sigma(D_k) - \widetilde{D}_k^T\|_1 + \eta\, \mathcal{L}_\nabla(\sigma(D_k), \widetilde{D}_k^T) \right), \tag{4}$$

where $\mathcal{L}_\nabla$ measures horizontal and vertical gradient discrepancies.

Hierarchical feature alignment. Because teacher and student channels differ, we align normalized channel-energy maps. For feature F,

$$\mathcal{A}(F) = \mathcal{N}\left(\frac{1}{C_F} \sum_{c=1}^{C_F} F_c^2 \right), \tag{5}$$

where $\mathcal{N}$ is spatial min–max normalization, and

$$\mathcal{L}_a = \frac{1}{4} \sum_{i=1}^{4} \|\mathcal{A}(G_i) - \mathcal{A}(T_i)\|_1. \tag{6}$$

This transfers the teacher's spatial organization without requiring channel-wise correspondence between different architectures.

Boundary supervision. The gradient-derived boundary of $\widetilde{D}^T$ supervises E^G:

$$\mathcal{L}_{ge} = \mathrm{BCE}(E^G, \mathrm{Edge}(\widetilde{D}^T)). \tag{7}$$

Together, $\mathcal{L}_d$, $\mathcal{L}_a$, and $\mathcal{L}_{ge}$ transfer output-, feature-, and boundary-level geometry.

3.3 Geometry–Appearance Interaction

Cross-modal enhancement. The orange blocks in Figure 2 exchange appearance and geometry at each pyramid level. Native-resolution queries attend to adaptively pooled keys and values, reducing spatial attention to $O(H_i W_i P^2)$ for pooled size $P \times P$:

$$\bar{A}_i = A_i + \alpha_i \,\mathrm{Attn}(Q_A(A_i), K_G(\Pi(G_i)), V_G(\Pi(G_i))), \tag{8}$$

$$\bar{G}_i = G_i + \beta_i \,\mathrm{Attn}(Q_G(G_i), K_A(\Pi(A_i)), V_A(\Pi(A_i))), \tag{9}$$

where Π is adaptive average pooling. Residual gates α_i, β_i start at zero, and residual coordinate attention models horizontal and vertical dependencies (Hou et al., 2021).

Reliability-aware fusion. Because valid scene geometry can still be irrelevant to saliency, the green blocks estimate a pixel-wise reliability map from appearance, geometry, their discrepancy, and the resized geometry prediction:

$$r_i = \sigma\left(\phi_i([\bar{A}_i, \bar{G}_i, |\bar{A}_i - \bar{G}_i|, \mathcal{U}_i(\sigma(D))])\right). \tag{10}$$

With channel-wise geometry attention $q_i = \sigma(\psi_i(\bar{G}_i))$, fusion is

$$F_i = \rho_i\left(\bar{A}_i \odot (1 + r_i \odot q_i) + r_i \odot \bar{G}_i\right). \tag{11}$$

The RGB-dominant form falls back toward appearance when r_i is small and injects geometry only where reliability is high; a negative reliability bias prevents immature geometry from dominating early training.

3.4 Saliency Decoding and Optimization

A separate AETP–ESC stack maps $\{F_i\}$ to saliency logits $\{S_j\}_{j=1}^{J}$ and an SOD edge logit E^S. The geometry and saliency branches share the progressive design but not parameters. With deep supervision,

$$\mathcal{L}_s = \sum_{j=1}^{J} \nu_j \mathcal{L}_{\mathrm{str}}(S_j, Y), \qquad \mathcal{L}_e = \mathrm{BCE}(E^S, \mathrm{Edge}(Y)), \tag{12}$$

and the complete objective is

$$\mathcal{L} = \mathcal{L}_s + \lambda_e \mathcal{L}_e + \lambda_{ge} \mathcal{L}_{ge} + \lambda_d \mathcal{L}_d + \lambda_a \mathcal{L}_a. \tag{13}$$

We set $\lambda_e = 0.4$, $\lambda_{ge} = 0.1$, $\lambda_d = 0.2$, and $\lambda_a = 0.05$. At inference, $\mathcal{T}$ and teacher-dependent supervision are removed; the encoder, geometry branch, fusion modules, and SOD decoder operate from RGB alone.

4 Experiments

4.1 Experimental Protocol

Datasets. For **RGB-D SOD**, we train on 2,985 RGB-mask pairs from NJU2K (1,485) (Ju et al., 2015), NLPR (700) (Peng et al., 2014), and DUT-RGBD (800) (Piao et al., 2019); dataset depth is ignored throughout training and inference. We evaluate on NJU2K (500), NLPR (300), DUT-RGBD (400), ReDWeb-S (1,000) (Liu et al., 2022), SIP (929) (Fan et al., 2021), SSD (80) (Zhou et al., 2021), STERE (1,000) (Niu et al., 2012), COME-E (4,600), and COME-H (3,000) (Zhang et al., 2021). For **RGB SOD**, we retrain on DUTS-TR (10,553) and evaluate on DUTS-TE (5,019) (Wang et al., 2017), ECSSD (1,000) (Shi et al., 2016), HKU-IS (4,447) (Li & Yu, 2016), and PASCAL-S (850) (Li et al., 2014).

Metrics. We report S_m (Fan et al., 2017), $F_\beta^{\max}$ ($\beta^2 = 0.3$), $E_\xi^{\max}$ (Fan et al., 2018), and MAE $\mathcal{M}$; higher values are better for the first three, whereas lower values are better for $\mathcal{M}$.

Implementation details. Unless stated otherwise, we use PVT-v2-B5, a frozen Depth Anything V2-Small teacher, $C = 128$, and one ESC block per branch. Student/teacher inputs are 364×364. We train for 80 epochs with batch size 4 and AdamW, using learning rates 7.5×10^{-5} for new modules and 7.5×10^{-6} for the encoder, weight decay 1.5×10^{-4}, gradient clipping at 1.0, random cropping, and horizontal flipping. Training uses one RTX 4090 (24 GB).

Table 1: Quantitative comparison on nine RGB-D SOD benchmarks. Parentheses denote test-set sizes; bold marks the best results; ↑ / ↓ means higher/lower is better. The first three benchmarks also contribute training data and therefore involve smaller train–test dataset/depth-domain shifts; GeoDistill remains competitive there and is best or tied-best in 23/24 comparisons on the other six benchmarks, highlighting strong cross-dataset generalization without sensor depth.

Method	NJU2K (500)				NLPR (300)				DUT-RGBD (400)			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
C2DFNet	.861	.854	.912	.054	.909	.895	.953	.025	.896	.904	.942	.037
RD3D	.893	.883	.927	.047	.903	.880	.937	.033	.863	.841	.894	.060
PICRNet	.386	.274	.526	.423	.389	.157	.609	.383	.362	.247	.515	.432
HRTransNet	.917	.920	.952	.032	.931	.925	.966	.019	.918	.925	.951	.033
CAVER	.926	.928	.959	.030	.934	.929	.970	.021	.938	.944	.966	.026
CPNet	.935	**.941**	.964	.025	.940	**.936**	**.973**	**.016**	.951	**.959**	.975	.019
LAFB	.907	.912	.946	.036	.930	.921	.965	.020	.927	.934	.956	.028
CATNet	.932	.937	.961	.026	**.940**	.934	.972	.018	**.953**	.958	**.976**	**.019**
SATNet	.923	.925	.954	.030	.929	.920	.964	.021	.942	.947	.966	.022
DPPNet	.929	.932	.962	.028	.937	.927	.968	.020	.939	.946	.965	.025
GeoDistill	**.935**	.940	**.964**	**.025**	.935	.928	.967	.019	.948	.957	.973	.020

Method	ReDWeb-S (1,000)				SIP (929)				SSD (80)			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
C2DFNet	.613	.580	.708	.175	.793	.784	.855	.088	.807	.762	.872	.067
RD3D	.671	.621	.730	.163	.835	.826	.884	.074	.852	.814	.901	.058
PICRNet	.356	.311	.486	.447	.312	.246	.571	.464	.344	.253	.444	.467
HRTransNet	.724	.710	.800	.127	.860	.876	.916	.056	.848	.820	.909	.053
CAVER	.736	.737	.808	.121	.904	.915	.945	.038	.890	.884	.936	.039
CPNet	.752	.755	.822	.112	.907	.927	.946	.035	.893	.893	.935	.035
LAFB	.722	.721	.792	.129	.897	.913	.942	.041	.857	.841	.922	.045
CATNet	.748	.750	.816	.115	.911	.928	**.952**	.034	.892	.879	.927	.036
SATNet	.705	.701	.782	.131	.898	.904	.931	.042	.871	.852	.917	.044
DPPNet	.749	.746	.817	.115	.896	.911	.938	.042	.891	.885	.938	.037
GeoDistill	**.781**	**.789**	**.842**	**.097**	**.912**	**.928**	.951	**.033**	**.898**	**.894**	**.948**	**.031**

Method	STERE (1,000)				COME-E (4,600)				COME-H (3,000)			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
C2DFNet	.868	.863	.919	.048	.779	.776	.848	.094	.724	.726	.800	.130
RD3D	.889	.868	.920	.048	.836	.818	.877	.073	.782	.764	.824	.109
PICRNet	.386	.253	.539	.421	.362	.295	.489	.427	.361	.329	.483	.434
HRTransNet	.915	.912	.954	.032	.858	.856	.909	.056	.816	.818	.869	.084
CAVER	.917	.916	.955	.033	.870	.874	.918	.052	.822	.831	.872	.082
CPNet	.920	.923	.960	.029	.884	.889	.928	.045	.843	.854	.889	.071
LAFB	.908	.906	.945	.037	.864	.864	.908	.056	.814	.817	.860	.088
CATNet	.921	.922	.958	.030	.892	.897	.932	.043	.847	.855	.890	.071
SATNet	.919	.913	.951	.032	.864	.855	.901	.056	.814	.809	.857	.087
DPPNet	.922	.919	.957	.032	.878	.876	.917	.052	.839	.840	.879	.078
GeoDistill	**.927**	**.925**	**.960**	**.028**	**.896**	**.902**	**.935**	**.041**	**.856**	**.869**	**.899**	**.065**

Table 2: Quantitative comparison on four RGB SOD benchmarks. Parentheses denote test-set sizes; bold marks the best results; ↑ means higher is better.GeoDistill clearly outperforms all specialist RGB SOD models and is only slightly behind generalist models on ECSSD, despite the latter being trained on substantially more data than the roughly 10K images used by GeoDistill. It still surpasses them on the other three benchmarks, highlighting strong transferability. * denotes generalist models.

Method	DUTS-TE (5,019)			ECSSD (1,000)			HKU-IS (4,447)			PASCAL-S (850)		
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
VST	.896	.877	.939	.932	.944	.964	.928	.937	.968	.873	.850	.900
ICON	.890	.876	.931	.928	.943	.960	.920	.931	.960	.862	.844	.888
VST-T++	.901	.887	.943	.937	.949	.968	.930	.939	.968	.878	.855	.901
MENet	.905	.895	.943	.927	.938	.956	.927	.939	.965	.871	.848	.892
VSCode-T*	.917	.910	.954	**.945**	**.957**	**.971**	.935	.946	.970	.878	.852	.900
VSCode-v2-T*	.922	.917	.957	.940	.950	.965	.929	.936	.962	.875	.847	.891
GeoDistill	**.926**	**.920**	**.959**	.943	.956	.970	**.940**	**.951**	**.975**	**.889**	**.891**	**.931**

4.2 Quantitative and Qualitative Comparison

RGB-D SOD. We compare GeoDistill with ten recent RGB-D SOD methods (Zhang et al., 2022a; Chen et al., 2022; Cong et al., 2023; Tang et al., 2023; Pang et al., 2023; Hu et al., 2024; Wang

et al., 2024; Sun et al., 2024; Duan et al., 2025; Yuan et al., 2025). Table 1 first reveals a clear split between training-source and cross-dataset benchmarks. On NJU2K, NLPR, and DUT-RGBD, **whose training splits contribute to the conventional training mixture**, GeoDistill is not uniformly best but remains near the top: on NJU2K it ties the best S_m, E_ξ, and MAE and trails the best F_β by only 0.001. **In contrast, on the six benchmarks not used for training**, GeoDistill is best or tied-best in 23 of 24 metric–dataset comparisons. The margin is particularly large on ReDWeb-S, where the strongest prior S_m, F_β, and E_ξ improve by 3.9%, 4.5%, and 2.4%, respectively, while MAE decreases by 13.4%. MAE also drops by 11.4% on SSD, 8.5% on COME-H, 4.7% on COME-E, and 3.5% on STERE.

This distribution is more informative than a uniform average gain. The first three test sets share source datasets with training and therefore exhibit smaller train–test dataset/depth-domain shifts; RGB-D competitors can also directly exploit sensor-depth statistics that recur within those datasets. Once evaluation moves away from that mixture, the advantage shifts strongly toward GeoDistill. **This supports the central hypothesis of the paper:** replacing dataset-specific sensor depth with task-adapted RGB geometry sacrifices little in-distribution performance while substantially improving robustness to dataset and depth-quality variation.

Pseudo-depth Substitution. A direct comparison with SATNet (Duan et al., 2025) isolates whether replacing sensor depth with a predicted depth map is sufficient. Although SATNet also removes sensor depth, it feeds a monocular single-channel depth prior to the detector. GeoDistill outperforms SATNet in all 36 metric–dataset comparisons; representative MAE reductions reach 29.5% on SSD, 26.8% on COME-E, 26.0% on ReDWeb-S, and 25.3% on COME-H. The qualitative examples likewise show fewer incomplete foregrounds and depth-induced background responses. Because both methods avoid sensor depth at inference, these gains indicate that the key advantage comes from internalizing multi-level teacher structure—rather than substituting one explicit depth map for another—which directly supports our geometry-distillation design.

RGB SOD generalization. We retrain GeoDistill on DUTS-TR, with the depth teacher used only during training. As shown in Table 2, GeoDistill outperforms all specialist RGB SOD models (Liu et al., 2021; Zhuge et al., 2023; Liu et al., 2024; Wang et al., 2023) across the four benchmarks and achieves the best results in 9 of 12 comparisons. On PASCAL-S, it improves the strongest prior S_m, F_β, and E_ξ by 1.3%, 4.2%, and 3.3%, respectively, while the gains on DUTS-TE remain consistent (0.2–0.4%). **The only exception is ECSSD**, where GeoDistill is slightly behind **generalist models** VSCode (Luo et al., 2024; 2026), which use substantially broader multi-dataset supervision. Despite this advantage, GeoDistill still surpasses them on DUTS-TE, HKU-IS, and PASCAL-S. Since no RGB-D sensor data are used in this protocol, these results show that the distilled geometry acts as a transferable structural prior rather than being tied to a specific depth distribution.

Qualitative Comparison. Figure 3 provides visual evidence. When RGB appearance is ambiguous but scene structure separates the target, GeoDistill recovers more complete foregrounds. When sensor depth contains strong background responses, several RGB-D competitors introduce false positives or lose object regions, whereas GeoDistill remains stable because the displayed depth is never consumed. With incomplete or blurred depth boundaries, its contours also remain more coherent. These cases mirror the quantitative pattern: the main benefit is not merely sharper masks, but reduced sensitivity to sensor-depth failure modes while retaining useful geometric structure.

4.3 Ablation Studies

All ablations follow the same protocol and isolate four questions: projector capacity, geometry supervision (geometry branch and feature distillation), teacher choice, and feature-transfer strategy.

Projector capacity. The Projector maps multi-level encoder features into a shared C-dimensional representation space, providing a unified feature basis for subsequent geometry learning and appearance–geometry interaction. Table 3 shows clear capacity saturation. For example, increasing C from 64 to 128 reduces SIP MAE from 0.052 to 0.033, indicating that sufficient projection capacity is important for preserving structural information. Further increasing C to 256 yields no consistent accuracy gain while substantially increasing computation, with FLOPs rising from 311.34G to 989.56G ($+218\%$). We therefore adopt $C = 128$ as the best accuracy–efficiency trade-off.

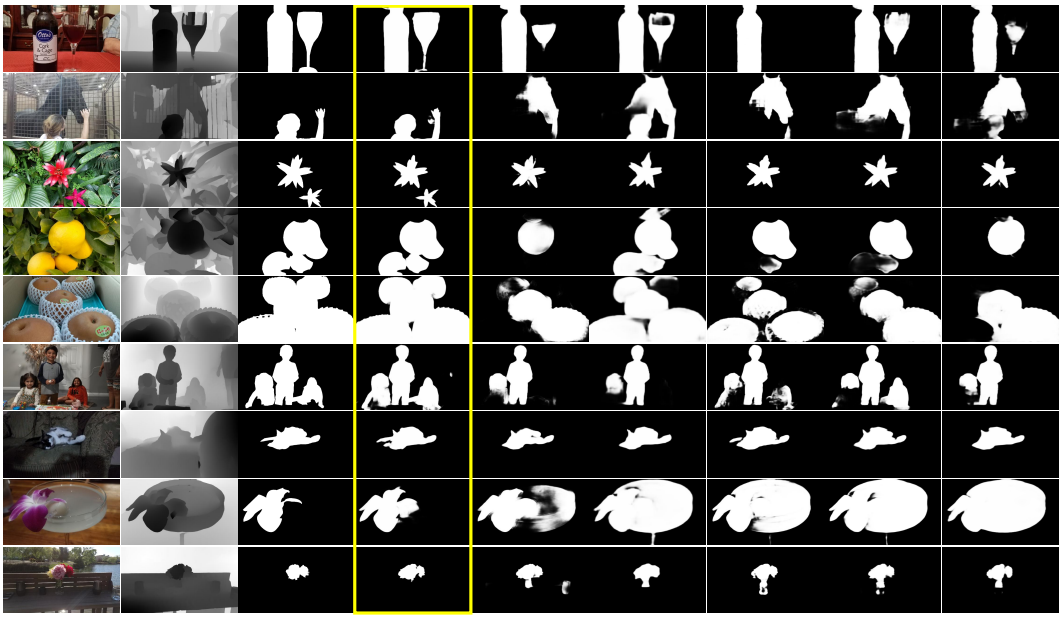

Figure 3: Qualitative comparison with SOTA RGB-D methods. Sensor depth (D) is shown only for diagnosis and is never used by GeoDistill. Across ambiguous RGB appearance, structured depth noise, and incomplete depth boundaries, GeoDistill better preserves salient regions while avoiding depth-induced false responses, illustrating the robustness gained by discarding sensor depth.

Table 3: Shared projection width C denotes the channel dimension of the projected multi-level encoder features. Bold marks the best results; $\uparrow$ / $\downarrow$ means higher/lower is better. Among the tested widths, $C = 128$ achieves the best accuracy–efficiency trade-off.

Channels C	Param. (M)	FLOPs (G)	DUT-RGBD (400)				SIP (929)				COME-H (3,000)			
			$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
64	89.68	139.46	.945	.955	.971	.038	.906	.921	.944	.052	.855	.865	.898	.086
128	111.76	311.34	**.948**	**.957**	**.973**	**.020**	**.912**	**.928**	**.951**	**.033**	**.856**	**.869**	**.899**	**.065**
256	199.03	989.56	.947	.954	.972	**.020**	.911	.926	.949	.034	.854	.866	.897	.066

Geometry branch supervision. Table 4(a) first separates architectural benefit from supervision quality. Adding the geometry branch with SOD supervision alone improves COME-H and slightly reduces SIP MAE, but degrades DUT-RGBD and does not consistently benefit SIP, indicating that the structure-oriented pathway is useful yet unstable when learned only indirectly from saliency labels. Supervising its final output with DA-V2 stabilizes the branch and improves all 12 comparisons; for example, on SIP, S_m increases from 0.886 to 0.893 while MAE decreases from 0.045 to 0.042. In contrast, GT-depth supervision is weaker than DA-V2 in 11 of 12 comparisons, despite using physical depth measurements. These results directly support our motivation: geometry is beneficial, but sensor depth is not necessarily the most effective supervision target; a stronger foundation geometry signal provides more reliable guidance for learning task-relevant structure.

Feature-distillation supervision. Table 4(b) examines which teacher signal contributes most to geometry learning. Replacing output-only supervision with feature-only distillation improves all 12 metrics. For example, on SIP, MAE decreases from 0.042 to 0.036, a 14.3% reduction. Adding output supervision to feature distillation further yields a better overall balance: it achieves the best SIP results and improves COME-H F_β/MAE, although feature-only distillation remains slightly better on a few metrics. We therefore adopt the combined supervision in the final model for more consistent cross-dataset performance. More importantly, the clear gain from output-only supervision to feature distillation shows that hierarchical intermediate features, rather than regression to a single depth-like map, provide the primary geometry-learning signal, while output supervision serves as a complementary constraint.

Table 4: Geometry-learning ablations. Bold marks the best results; ↑ / ↓ means higher/lower is better. The four studies show that (a) geometry needs reliable foundation-model supervision, (b) intermediate feature distillation is the main supervisory signal while output regression is complementary, (c) teacher quality affects transfer, and (d) task-adapted distillation is stronger than direct teacher-feature fusion. Rows marked with † denote the final configuration used in the main experiments.

(a) Geometry branch supervision	DUT-RGBD				SIP				COME-H			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
RGB-only	.940	.948	.967	**.022**	.883	.902	.931	.046	.831	.841	.875	.080
Geometry branch	.939	.945	.966	.024	.886	.901	.930	.045	.838	.843	.880	.076
DA-V2 output supervision	**.941**	**.949**	**.968**	**.022**	**.893**	**.908**	**.937**	**.042**	**.841**	**.848**	**.885**	**.073**
GT-depth output supervision	.939	.946	.967	**.022**	.887	.906	.934	.045	.835	.844	.881	.076
(b) Feature-distillation supervision	DUT-RGBD				SIP				COME-H			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
Output supervision only	.941	.949	.968	.022	.893	.908	.937	.042	.841	.848	.885	.073
Feature distillation only	**.949**	**.957**	**.974**	**.020**	.907	.922	.945	.036	**.857**	.867	**.899**	.066
Output + feature distillation†	.948	**.957**	.973	**.020**	**.912**	**.928**	**.951**	**.033**	.856	**.869**	**.899**	**.065**
(c) Distillation teacher	DUT-RGBD				SIP				COME-H			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
YOLO26-X	.937	.944	.966	.024	.887	.903	.933	.044	.838	.845	.881	.076
Depth Anything V2†	**.948**	**.957**	**.973**	**.020**	**.912**	**.928**	**.951**	**.033**	**.856**	**.869**	**.899**	**.065**
(d) Feature transfer strategy	DUT-RGBD				SIP				COME-H			
	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
Direct DA-V2 feature fusion	.942	.955	.972	**.020**	.898	.918	.941	.038	.843	.845	.885	.070
Feature distillation†	**.948**	**.957**	**.973**	**.020**	**.912**	**.928**	**.951**	**.033**	**.856**	**.869**	**.899**	**.065**

Distillation teacher. With the transfer mechanism fixed, Table 4(c) isolates teacher choice. Depth Anything V2 improves every reported metric over YOLO26-X: S_m rises from $0.937/0.887/0.838$ to $0.948/0.912/0.856$, while MAE falls by 16.7%, 25.0%, and 14.5% on DUT-RGBD/SIP/COME-H. The architecture therefore does not require a specific teacher interface, but it is not quality-agnostic: a teacher whose intermediate representation better organizes dense scene geometry transfers a more useful structural prior. This also explains why the method is described generically in the formulation while Depth Anything V2 is selected for the final implementation.

Feature transfer strategy. Finally, Table 4(d) tests whether distillation is necessary at all by directly replacing student geometry features with Depth Anything V2 features. Despite giving the detector direct access to teacher representations, direct fusion is weaker: distillation raises S_m from $0.942/0.898/0.843$ to $0.948/0.912/0.856$ and matches or improves all 12 metrics, with only DUT-RGBD MAE tied. The gap shows that teacher features are not automatically aligned with saliency; learning an RGB-driven student allows geometry to be adapted to the downstream objective before reliability-controlled fusion. Distillation therefore provides both better task compatibility and teacher-free inference, rather than merely compressing an otherwise identical feature source.

5 Conclusion and Discussion

We presented GeoDistill, a depth-free RGB-D SOD framework that replaces sensor-depth dependence with RGB-driven, task-adapted geometry. A training-only foundation teacher distills output-, feature-, and boundary-level structure, while cross-modal enhancement and reliability-aware fusion exchange and selectively inject geometry before SOD decoding. The results show competitive performance on training-source benchmarks but markedly stronger cross-dataset generalization, transfer to RGB SOD, and a clear ablation advantage for intermediate feature distillation and task-adapted transfer over raw-depth supervision or direct teacher-feature fusion. This supports our central claim that the useful component of depth for SOD is transferable geometric structure rather than the sensor measurement itself. Limitations include extra teacher computation, dependence on teacher quality, and relative rather than metric geometry; future work will study lighter teachers and more teacher-agnostic transfer.

AI USE STATEMENT

Generative AI tools were used to polish the language of selected manuscript passages and to assist with diagnosing and fixing bugs in the source code. All AI-assisted text was reviewed and edited by the authors, and all AI-suggested code changes were manually inspected and tested before use. The authors take responsibility for the final content of the paper, including text and code produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete algorithmic and architectural specifications, including loss definitions and hyperparameters, to ensure end-to-end reproducibility. After the review process, we will release the full source code, training/testing scripts, and model checkpoints to enable transparent verification.

REFERENCES

Shilong Bao, Qianqian Xu, Feiran Li, Boyu Han, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Towards size-invariant salient object detection: A generic evaluation and optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:1115–1132, 2026.

Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3051–3060, 2018.

Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 1063–1071, 2021.

Qian Chen, Zhenxi Zhang, Yanye Lu, Keren Fu, and Qijun Zhao. 3-d convolutional neural networks for rgb-d salient object detection and beyond. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):4309–4323, 2022.

Runmin Cong, Hongyu Liu, Chen Zhang, Wei Zhang, Feng Zheng, Ran Song, and Sam Kwong. Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 406–416, 2023.

Songsong Duan, Xi Yang, Nannan Wang, and Xinbo Gao. Lightweight rgb-d salient object detection from a speed-accuracy tradeoff perspective. *IEEE Transactions on Image Processing*, 34:2529–2543, 2025. doi: 10.1109/TIP.2025.3560488.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision*, pp. 4548–4557, 2017.

Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conference on Artificial Intelligence*, pp. 698–704, 2018.

Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *European Conference on Computer Vision*, pp. 275–292, 2020.

Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2075–2089, 2021.

Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3052–3062, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.

Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. *International Journal of Computer Vision*, 132(8):3067–3085, 2024.

Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Calibrated rgb-d salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9471–9481, 2021.

Ran Ju, Libo Ge, Weisheng Geng, Tao Ren, and Gangyi Wu. Depth-aware salient object detection using anisotropic center-surround difference. *Signal Processing: Image Communication*, 38:115–126, 2015.

Orest Kupyn, Hirokatsu Kataoka, and Christian Rupprecht. S3od: towards generalizable salient object detection with synthetic data. In *International Conference on Learning Representations*, 2026.

Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing*, 25(11):5012–5024, 2016.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014.

Nian Liu and Junwei Han. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089–3098, 2018.

Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *IEEE International Conference on Computer Vision*, pp. 4722–4732, 2021.

Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9026–9042, 2022.

Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, 2024.

Ziyang Luo, Nian Liu, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Shahbaz Khan, and Junwei Han. Vscode-v2: Dynamic prompt learning for general visual salient and camouflaged object detection with two-stage optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Ziyang Luo et al. VSCode: General visual salient and camouflaged object detection with 2d prompt learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, 2012.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. HDFNet: Hierarchical dynamic filtering network for rgb-d salient object detection. In *European Conference on Computer Vision*, 2020.

Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32:892–904, 2023.

Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: A benchmark and algorithms. In *European Conference on Computer Vision*, pp. 92–109, 2014.

Yongri Piao, Zan Rong, Miao Zhang, Wenqi Ren, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *IEEE International Conference on Computer Vision*, pp. 7254–7263, 2019.

Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9060–9069, 2020.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7479–7489, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016.

Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. *IEEE Transactions on Multimedia*, 26:2249–2262, 2024.

Bin Tang, Zhengyi Liu, Yacheng Tan, and Qian He. Hrtransnet: Hrformer-driven two-modality salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 728–742, 2023.

Kunpeng Wang, Zhengzheng Tu, Chenglong Li, Cheng Zhang, and Bin Luo. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7344–7358, 2024.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, 2017.

Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3239–3259, 2022.

Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Jun Wei, Shuhui Wang, and Qingming Huang. F3Net: Fusion, feedback and focus for salient object detection. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 12321–12328, 2020.

Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2019.

Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cedric Demonceaux. HiDAnet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Transactions on Image Processing*, 32, 2023.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Sheng Ye, Xin Chen, Yan Zhang, Xianming Lin, and Liujuan Cao. ESCNet: Edge-semantic collaborative network for camouflaged object detection. In *IEEE International Conference on Computer Vision*, 2025.

Junbin Yuan, Yiqi Wang, Zhoutao Wang, Qingzhen Xu, Bharadwaj Veeravalli, and Xulei Yang. Dppnet: a depth pixel-wise potential-aware network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 27:4256–4268, 2025. doi: 10.1109/TMM.2025.3535386.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8582–8591, 2020.

Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *IEEE International Conference on Computer Vision*, pp. 4338–4347, 2021.

Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Trans. Multimed.*, 25:1–13, 2022a.

Shuo Zhang, Jiaming Huang, Wenbing Tang, Yan Wu, Terrence Hu, Xiaogang Xu, and Jing Liu. DiMSOD: A diffusion-based framework for multi-modal salient object detection. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 10103–10111, 2025.

Yuan-Fang Zhang, Jiangbin Zheng, Wenjing Jia, Wenfeng Huang, Long Li, Nian Liu, Fei Li, and Xiangjian He. Deep rgb-d saliency detection without depth. *IEEE Transactions on Multimedia*, 24:755–767, 2022b.

Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *IEEE International Conference on Computer Vision*, pp. 8779–8788, 2019.

Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Xiang Ruan. Self-supervised pretraining for rgb-d salient object detection. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 3463–3471, 2022.

Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *Computational Visual Media*, 7:37–69, 2021.

Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2023.