ICPR 2026 · 28TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION · LYON, FRANCE · AUGUST 17–22, 2026

Multi-Source Pseudo-Label Generation for Weakly Supervised Salient Object Detection

Handan Zhang (Presenter) · Tie Liu* · Yuanyuan Shang · Hui Ding · Zhuhong Shao

College of Information Engineering, Capital Normal University, Beijing, China

*Corresponding author: liutiel@163.com

SESSION
O6-B.T2

CMT PAPER ID
847

PRESENTED BY
Handan Zhang

Outline

Background: SOD and Its Annotation Cost

THE TASK

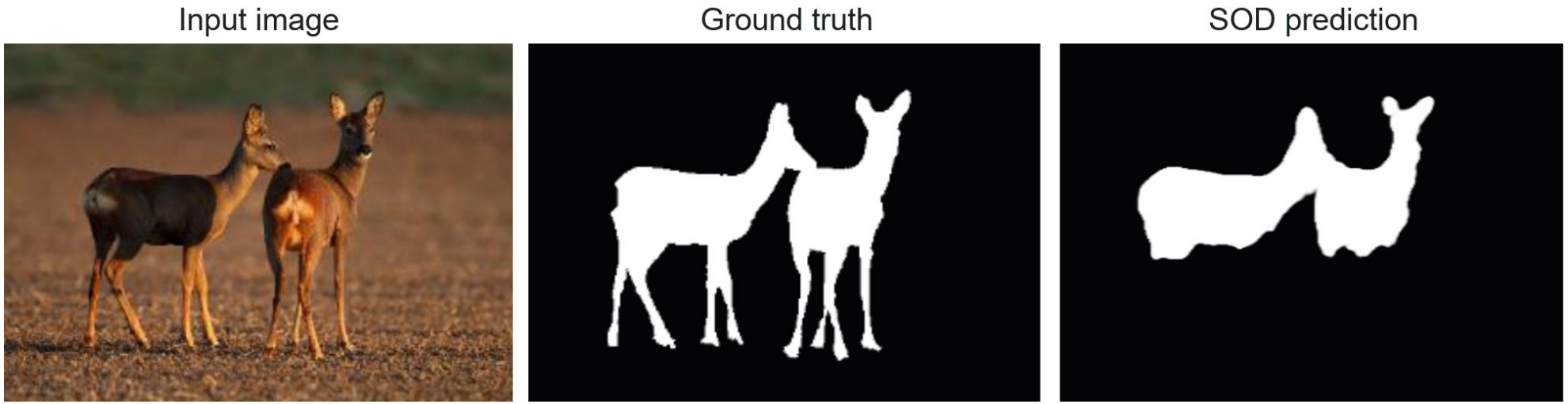
Salient Object Detection (SOD) separates the **most visually prominent objects** from the background at pixel level.

It is a core step for biological classification, pose estimation, artifact recognition, and more — from RGB to RGB-D, co-saliency, and video.

THE BOTTLENECK

Fully supervised models rely on **dense pixel-level masks** — costly to annotate, saturating in accuracy, and often generalizing poorly.

Weak supervision (image tags, bounding boxes, model consensus) is the scalable alternative.

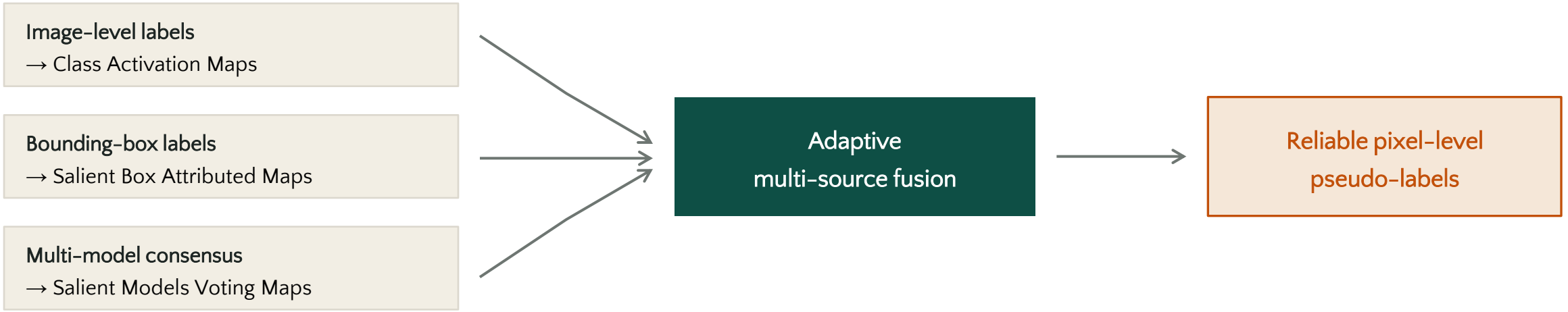


SOD in one glance: given an input image, predict a pixel-accurate mask of the salient object.

Single Weak Sources Are Noisy — but Complementary

- **Class Activation Maps (CAMs)** — coarse localization; activate only the most discriminative parts, not the whole object
- **Box-based attribution maps** — cover only part of the salient region; the full extent is never observed
- **Off-the-shelf SOD models** — any single model gives inconsistent or blurry predictions; no ground truth to trust

Our idea: these sources fail in different ways — adaptively fuse them into reliable pixel-level pseudo-labels.



Contributions

1

Voting-based weak-label generation

Fuse outputs of multiple pre-trained SOD models into Salient Models Voting Maps (SMVMs) with rank-derived, per-model weights — consensus beats any single model.

2

A two-stage multi-source WSOD framework

A three-branch network (CAMs, SBBAMs, SMVMs) plus an aggregation module produce more precise weak labels with clearer boundaries.

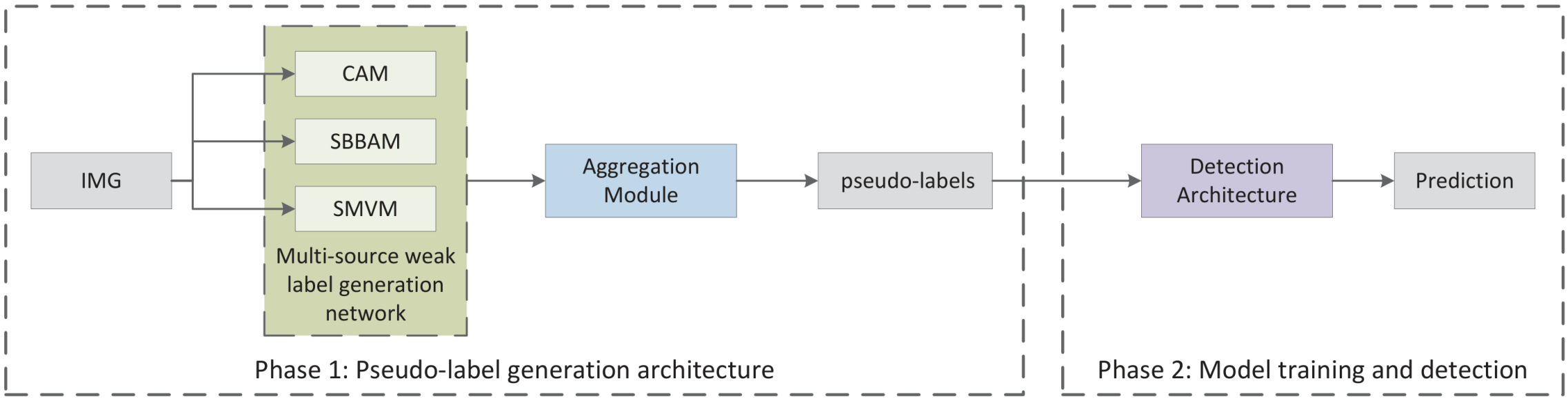
3

Stronger pseudo-labels, stronger SOD

Superior localization and segmentation over existing weakly / unsupervised methods — and competitive with several fully supervised ones.

A Two-Stage Weakly Supervised Framework

Stage 1 turns three heterogeneous weak signals into pixel-level pseudo-labels; **Stage 2** trains the final detector on them.



STAGE 1 · PSEUDO-LABEL GENERATION

Three branches capture coarse regions, spatial extents, and model consensus

STAGE 2 · DETECTION

Train the SOD network on pseudo-labels

Fig. 1 — Overview of the proposed framework.

Stage 1 — A Three-Branch Weak-Label Network

- **CAM branch**
  Coarse object localization from an image classifier
- **SBBAM branch**
  Precise spatial extents from a box-supervised detector
- **SMVM branch**
  Weighted consensus of several pre-trained SOD models

Node A — aggregation

Fuses the three maps into refined pixel-level pseudo-labels

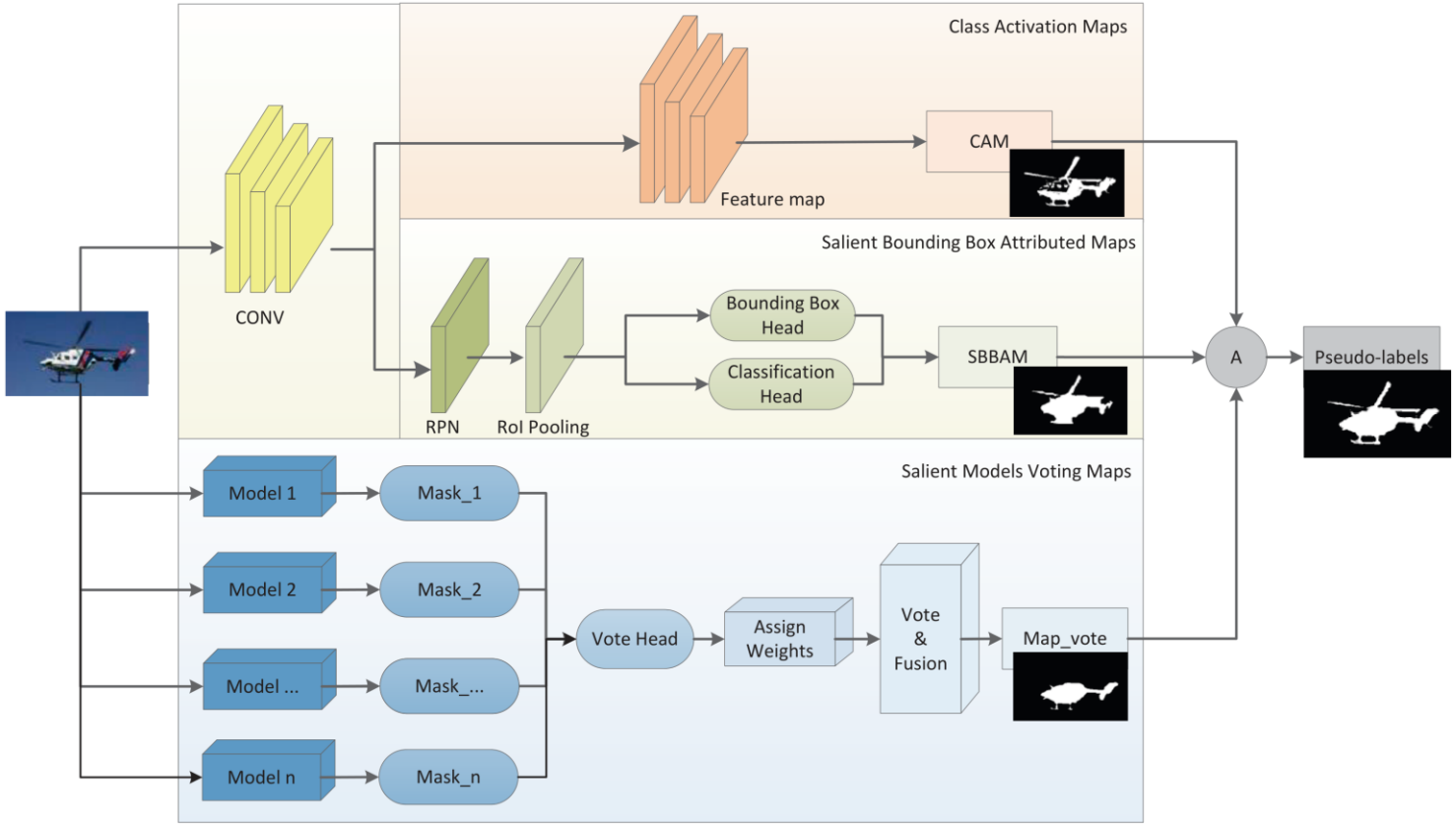


Fig. 2 — Three-branch network of stage 1 (CAM / SBBAM / SMVM → aggregation).

Branch A — Class Activation Maps (CAMs)

- **ResNet-50 classifier** trained with image-level labels; downsampling stride reduced to 1 → CAMs at **1/16 input resolution**
- GAP class weights re-weight the last conv feature maps
- Pixels unrelated to salient objects are **zeroed out**
- SOD is category-agnostic — CAMs serve as a **coarse localization prior**

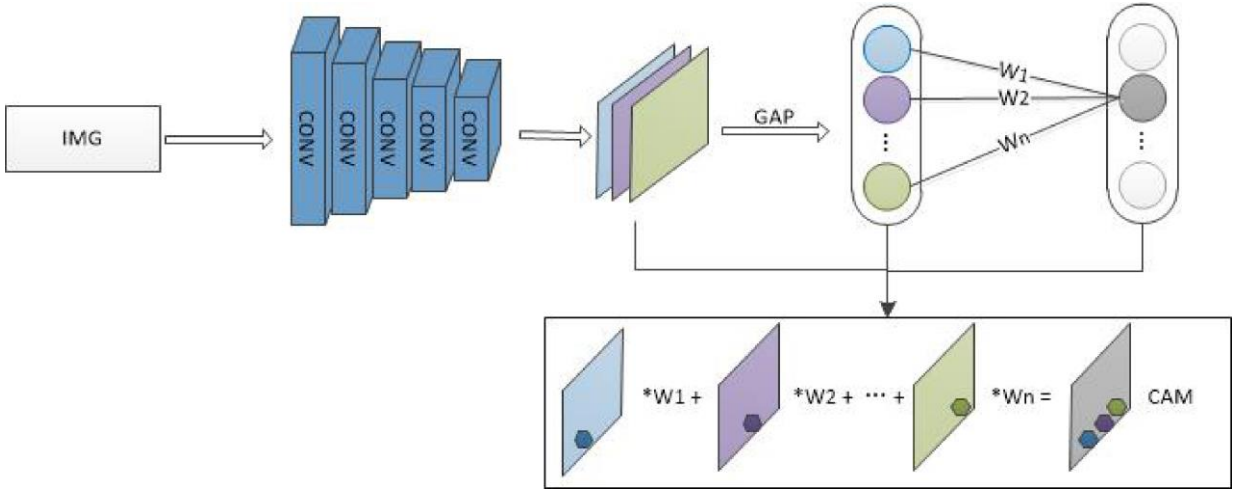


Fig. 3 — CAM generation branch.

$$Cam(x) = w_c \cdot f(x) \,/\, \max_x W_c \cdot f(x)$$

w: GAP class weights · f: last conv feature map

Role: rough but useful localization of the salient object.

Branch B — Salient Bounding Box Attributed Maps

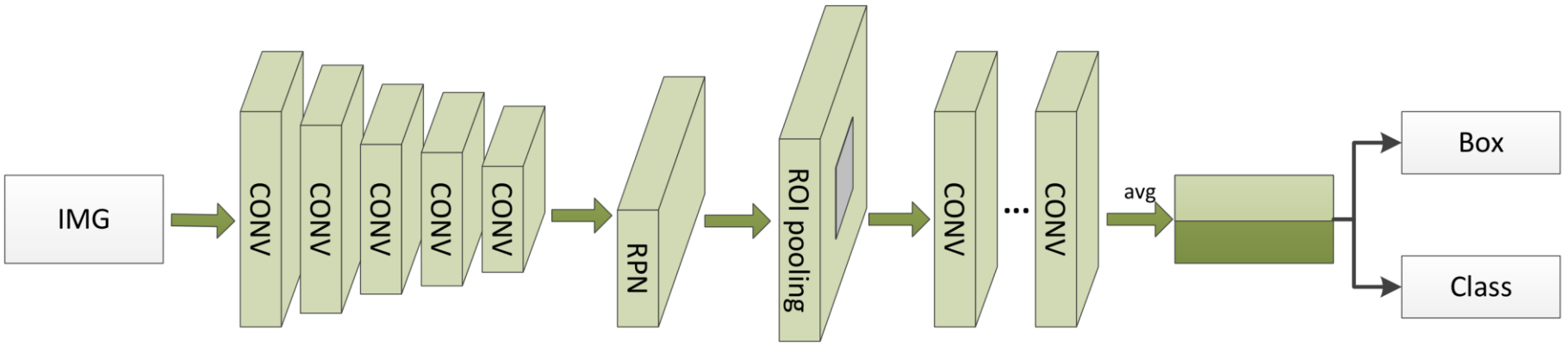


Fig. 4 — Faster R-CNN detector behind the SBBAM branch.

- **Minimal-mask perturbation:** the smallest region that keeps the detector's prediction unchanged is the salient evidence
- Candidate boxes perturbed by up to **±30%**; kept if correctly classified and **IoU > 0.8**, then merged
- Classification + box heads give **complementary** localization
- Dual-threshold CRF refinement; annotation cost: **seconds per image**

$$\phi(I, M) = I \circ M + \mu \circ (1 - M)$$

$$M^* = \arg\max_M \ \lambda \|M\| + L_{perturb}$$

$\lambda = 0.07 \cdot \mu$: dataset mean pixel · both heads contribute to $L_{perturb}$

Branch C — Salient Models Voting Maps (SMVMs)

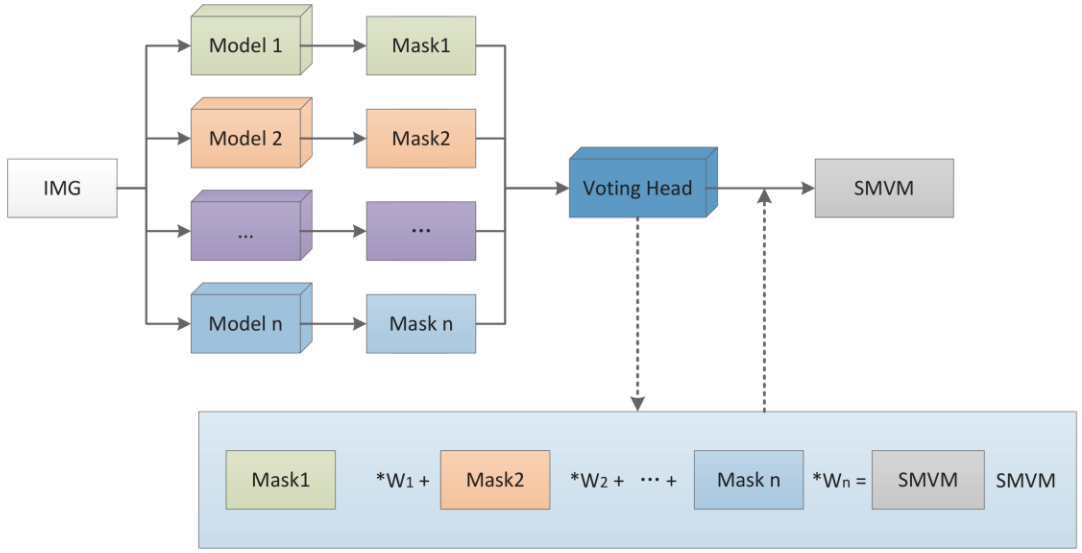


Fig. 5 — Voting branch: masks → weights → fusion.

Expert models: TRACER · SOC · BASNet — top-3 of 9 candidates under a composite score (IoU, correlation, KL divergence)

Rank-based weights: rank each model per dataset by MAE, average the ranks → M_n

$$w_n = \tfrac{1}{2} \cdot \log((1 - M_n) / M_n), \quad W_n = e^{w_n} / \Sigma_n e^{w_n}$$

$$S_\phi = \Sigma_n f_{ij}^n \times W_n$$

pixel ϕ is salient when $S_\phi > 0.5$

Role: model consensus — suppresses inconsistent or blurry individual predictions through weighted voting.

Aggregation — From Three Maps to Pseudo-Labels

STEP 1 · WEIGHTED FUSION

- CAM refined by **dense CRF** with an adaptive threshold $\gamma = 2 \times$ mean pixel value
- Fused with SBBAM (S) and SMVM (V) into a coarse **CSVMap**

$$\text{CSVMap} = \theta_1 \cdot \hat{C} + \theta_2 \cdot S + (1 - \theta_1 - \theta_2) \cdot V$$

θ_1, θ_2: trade-off factors set by experiment

STEP 2 · GRAPH-REGULARIZED REFINEMENT

- CSVMap segmented into **superpixels**; RBF affinity over mean Lab color
- **Graph Laplacian-regularized nonlinear regression** spreads saliency to unlabeled superpixels

$$\delta^* = \left(JK + \gamma_A \cdot uI + \gamma_I \cdot u \,/\, (u+v)^2 \cdot LK \right)^{-1} \cdot y$$

closed-form solution; $\gamma\_A = 10^{-6}$, $\gamma\_I = 1$

Stage 2 — detection architecture. Self-Reformer (Transformer encoder + context refinement module) is trained with the generated pseudo-labels as ground truth; a refined attention map guides fine-grained segmentation, and pixel-shuffle up/down-sampling preserves structure in the final saliency map.

Experiments — Quantitative Comparison

Setup — train: DUTS-TR / ImageNet (CAM), 60k+ box-annotated images (SBBAM), DUTS-TR (SMVM) · test: 6 benchmarks · metrics: $F_\beta \uparrow$ and MAE $\downarrow$

Method	SUP	ECSSD $F_\beta\uparrow$	ECSSD MAE↓	DUTS-TE $F_\beta\uparrow$	DUTS-TE MAE↓	HKU-IS $F_\beta\uparrow$	HKU-IS MAE↓	DUT-O $F_\beta\uparrow$	DUT-O MAE↓
MLMSNet	F	.928	.045	.840	.053	.921	.039	.764	.070
PAGENet	F	.932	.042	.834	.052	.917	.037	.789	.062
JointCRF	F	.927	.049	.828	.063	.920	.039	.792	.065
TSD	U	.916	.044	.810	.047	.902	.037	.745	.061
DCFD	U	.888	.059	.764	.064	.889	.042	.710	.070
SCW	W	.900	.049	.823	.049	.896	.038	.758	.060
MWS+	W	.909	.071	.809	.072	.899	.054	.783	.073
MWS	W	.878	.096	.745	.107	.856	.084	.718	.114
WSS	W	.856	.104	.736	.105	.859	.079	.687	.118
Ours	**W**	**.938**	**.033**	**.855**	**.033**	**.928**	**.027**	**.790**	**.049**

Best among all weakly / unsupervised methods on every dataset — and it surpasses fully supervised baselines on ECSSD, DUTS-TE and HKU-IS.

SUP: F = fully supervised, W = weakly supervised, U = unsupervised. Full six-dataset results (incl. PASCAL-S, SOD) in the paper.

Experiments — PR and F-Measure Curves

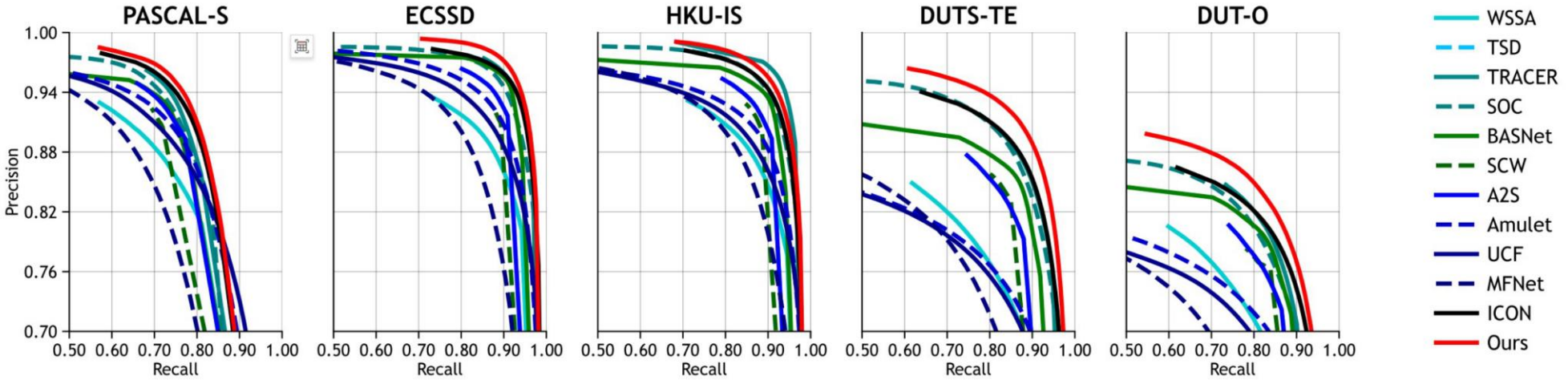


Fig. 6a — Precision–recall curves on six benchmarks.

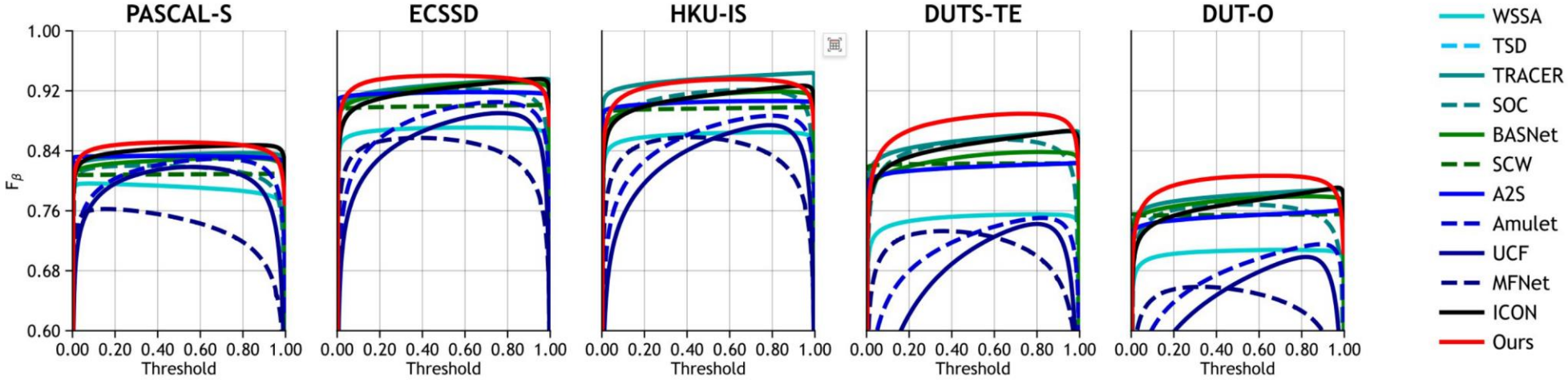


Fig. 6b — F-measure curves; E-measure curves (paper) show the same trend.

Ablation — Every Supervision Source Contributes

- The full three-branch model wins on all five datasets in both F_β and MAE
- SMVM brings the **largest single gain** over CAM alone
- SBBAM + SMVM already outperforms CAM-only by a wide margin
- ECSSD MAE: 0.143 → **0.033** (−77%)

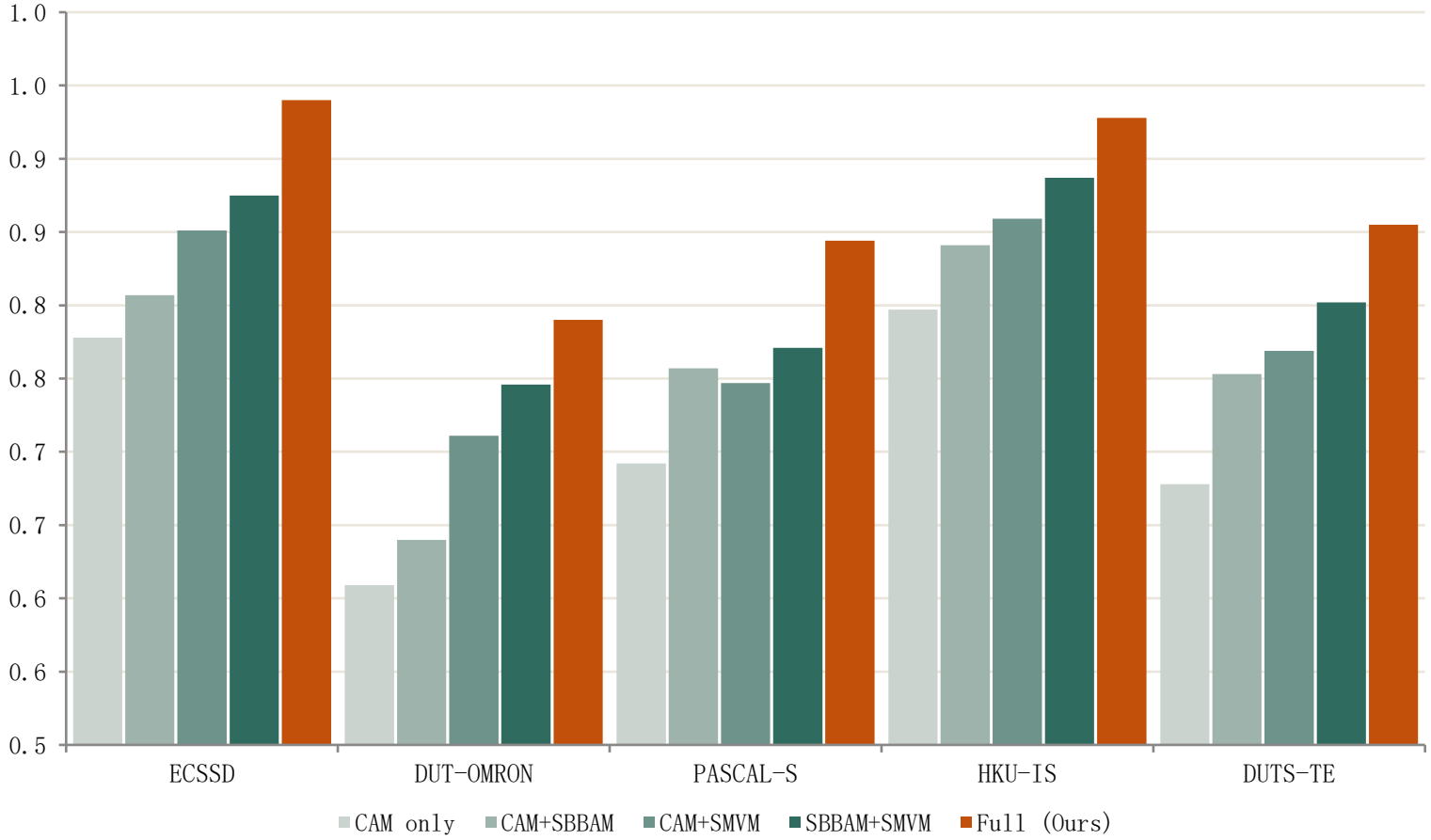


Table 1 — F-measure (Fβ) of branch combinations on five benchmarks; MAE follows the same trend.

Qualitative Comparison

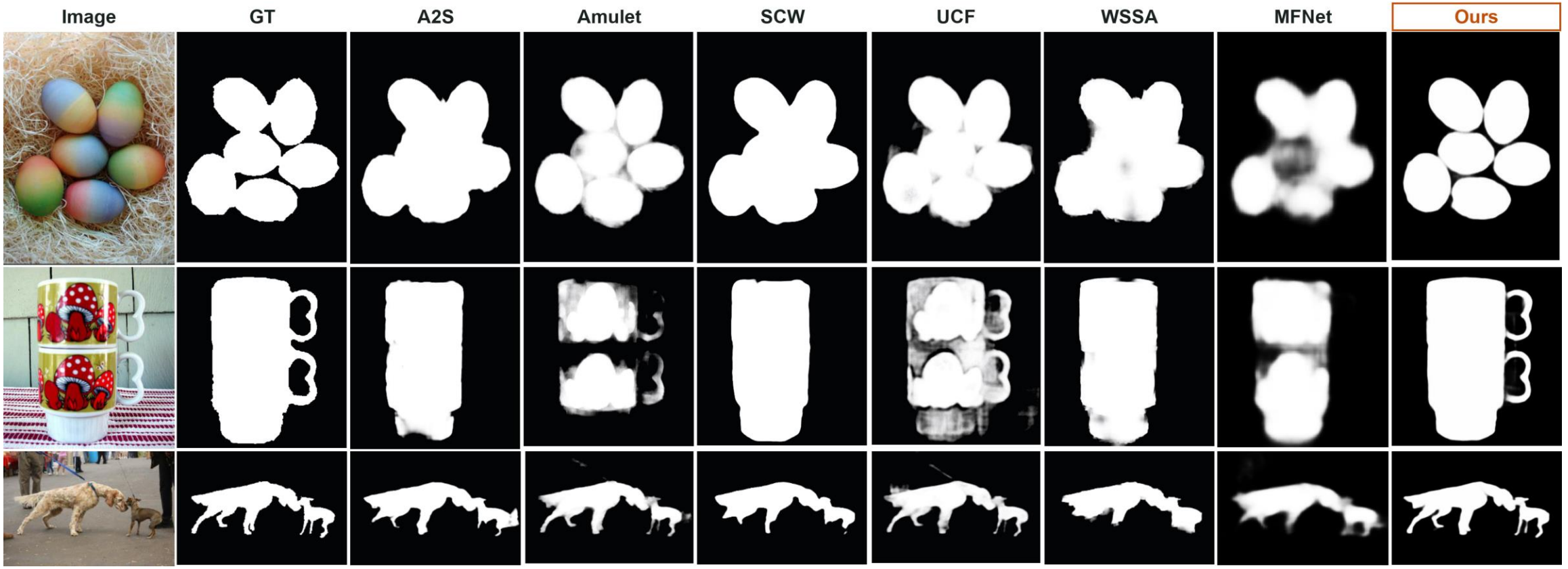


Fig. 7 — Rows: multi-object scene · fine structure (cup handle) · challenging edges. Columns: input, GT, six competing methods, ours.

Complete objects, separated instances, fine details. Only ours separates the individual eggs, keeps the cup handle, and stays sharp on cluttered edges.

Conclusion & Future Work

WHAT WE SHOWED

1 · **Multi-source beats single-source.** CAMs, box-attributed maps, and model consensus fail differently — fusing them yields reliable pixel-level pseudo-labels.

2 · **New state of the art** among weakly / unsupervised SOD on six benchmarks — competitive with fully supervised models.

3 · **Cheaper annotation:** from dense pixel masks to image tags, boxes, and existing models.

WHAT'S NEXT

- More weak supervision sources and fusion strategies
- Optimizing weight allocation of the voting mechanism
- Transferring the framework to related vision tasks — generalizability and scalability

Supported by the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence (HMHAI-202407), the National Natural Science Foundation of China (62476178), and the Beijing Natural Science Foundation (4242034, L252009).

Thank you

Questions & discussion are welcome

Multi-Source Pseudo-Label Generation for Weakly Supervised Salient Object Detection

Handan Zhang · Tie Liu* · Yuanyuan Shang · Hui Ding · Zhuhong Shao — Capital Normal University

SESSION · PAPER

O6-B.T2 · 847

CONTACT (PRESENTER)

2221002023@cnu.edu.cn