UNFOLDCRF: STRUCTURED MASK REFINEMENT WITH IMAGE-CONDITIONED LATENT REGIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned mask refiners improve segmentation accuracy, but it is hard to tell how much of the improvement comes from explicit structure rather than from extra capacity, and whether it holds up when the mask generator or its error distribution changes. UnfoldCRF treats refinement as inference in a conditional random field over pixel labels and latent region variables. Its energy has a corrected unary term, learned local pairwise interactions, and image-conditioned latent-region consistency, with a null state that lets a region with weak label agreement withdraw from the consistency term; inference unrolls damped mean-field updates on this one energy. To isolate the effect of structure, we compare against recurrent black-box refiners that read the same inputs and receive the same parameter budget, stage count, and supervision. On COD10K, UnfoldCRF beats the strongest matched control by 1.0 F_β^ω point, improves all four COD metrics, and lowers the fraction of images made worse from 11.7% to 8.5%. Under a train-once protocol over five datasets and several mask sources, the 2.6M-parameter variant gains 4.2 mean ΔIoU against 2.0 for its control, and a variant built on frozen DINOv2 features matches the strongest foundation-model refiner with about a seventh of its resident parameters while staying ahead of its own control. On mask generators never seen in training, the gain is 2.0 F_β^ω points against 0.6 for the control. Zeroing individual messages shows where the corrections come from: the pairwise messages mostly fix boundaries, the region messages mostly fix non-boundary errors. Code and supporting materials will be publicly released.

1 INTRODUCTION

Segmentation networks rarely fail as independent pixel noise. On targets entangled with their context, such as camouflaged objects, transparent surfaces, clinical structures, and thin or fragmented shapes, the failures have structure: interiors are hollowed out, components are dropped or hallucinated, and boundaries drift coherently along low-contrast contours (Fan et al., 2020; Xie et al., 2020; He et al., 2023b). Refinement methods post-process an initial mask to remove such errors, from dense CRFs with prescribed affinity functions (Krähenbühl & Koltun, 2011) to learned refiners built on cascades (Cheng et al., 2020), boundary correction (Yuan et al., 2020), diffusion (Wang et al., 2023; Shen et al., 2025), and prompted foundation models (Lin et al., 2025; Ke et al., 2023; Price et al., 2026). Deep unfolding has also been used for iterative refinement, though recent segmentation-oriented unfolding methods derive their updates from reconstruction objectives rather than from an energy over the labels themselves (He et al., 2025b; 2026).

Most learned refiners predict mask updates with generic network modules and no explicit energy over the labels. When such a refiner improves on the upstream mask, one cannot tell whether structure did the work or whether capacity, iteration, and supervision did, nor whether the improvement would survive a different mask generator or a different error distribution. We ask: **under carefully matched inputs, parameter count, iteration count, and supervision, does explicit structured inference provide more reliable refinement than black-box iterative refinement, both in-distribution and under shifts in mask generators and error distributions?** By *reliable* we mean a positive average gain, few images made worse, and a gain that survives shift. A second question is whether the corrections can be traced to particular energy messages.

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.

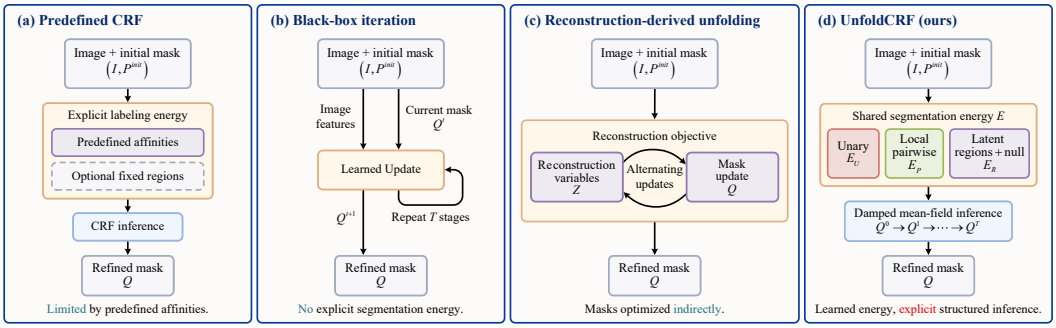

Figure 1: **Refinement formulations.** (a) Predefined CRFs use prescribed affinities and optional fixed regions. (b) Black-box refiners learn iterative mask updates without an explicit energy. (c) Reconstruction-derived unfolding alternates reconstruction and mask updates. (d) UnfoldCRF performs damped mean-field inference on a shared, image-conditioned segmentation energy with corrected unary, local pairwise, and latent-region terms.

UnfoldCRF is an image-conditioned latent-region CRF for mask refinement. Its energy has a corrected unary term, a symmetric local pairwise term in a learned feature space, and a mass-normalized consistency term over latent region variables. Learned local soft incidences connect pixels to regions at several granularities; a region variable takes either a semantic label or a null state, and the null state reduces the influence of regions whose members disagree. The construction builds on robust P^N potentials and associative hierarchical CRFs (Kohli et al., 2009; Ladickỳ et al., 2009; Arnab et al., 2016), but the pixel–region structure is image-conditioned and learned end to end through unrolled inference. For a given input, every stage runs on the same conditional energy; only a stage-dependent damping coefficient changes.

A refiner can raise in-distribution accuracy in two ways: by learning what valid masks look like, or by learning the error habits of the generators it was trained on. The pairwise and latent-region terms encode the first kind of knowledge, local boundary consistency and regional label agreement, and the unary term keeps the upstream prediction as evidence rather than re-predicting it. When the generator changes, these relations should be less tied to its particular error patterns than a learned update rule is. Section 4.4 tests this with held-out generators and message interventions.

Our main contributions are:

(1) Structural model. An image-conditioned latent-region CRF for mask refinement with local learnable soft incidences, symmetric pairwise interactions, and a mass-normalized region-consistency term whose latent state is semantic or null; its MAP reduction is a truncated region-consistency potential with a vote-fraction threshold (Proposition 1).

(2) Shared-energy unfolding. A damped mean-field refinement process in which all stages share one conditional energy and only the damping coefficient varies. Damping preserves the fixed points of the shared system (Lemma 1), and the latent-region message passing is sparse, costing $O(RNsL)$ per stage for N pixels, R levels, s candidate regions per pixel, and L labels.

(3) Empirical findings. Against parameter-matched recurrent controls, UnfoldCRF improves COD10K F_β^ω by 1.0 point and makes fewer images worse; on a five-benchmark train-once suite its S and L variants gain 4.2 and 6.3 mean ΔIoU against 2.0 and 4.6 for their controls. On unseen mask generators it keeps a 2.0-point F_β^ω gain where the control keeps 0.6, and zeroing messages shows that the pairwise messages mainly correct boundary errors while the region messages mainly correct non-boundary false negatives and false positives.

2 Related Work

Higher-order and hierarchical CRFs. Robust P^N potentials encode segment-level label consistency with a truncated penalty and admit an exact representation through auxiliary segment variables with a free label (Kohli et al., 2009); associative hierarchical CRFs stack such auxiliary variables across segmentation layers (Ladickỳ et al., 2009). Filter-based mean-field inference extends dense pairwise models to certain higher-order terms (Vineet et al., 2014). We keep the latent-variable

form of these potentials but make the pixel–region associations soft, local, image-conditioned, and trainable through unrolled inference.

Neural and trainable CRF inference. CRF-as-RNN unrolls parallel mean-field inference for dense Gaussian CRFs into a trainable recurrent module (Zheng et al., 2015), while subsequent work recasts CRF message passing using convolutional and pixel-adaptive filtering (Teichmann & Cipolla, 2018; Su et al., 2019). Higher-order detection and superpixel potentials have also been incorporated end to end (Arnab et al., 2016). Exact inference for Gaussian CRFs can instead be obtained by solving linear systems (Chandra & Kokkinos, 2016). Parameter learning and convergent inference for dense CRFs are studied in (Krähenbühl & Koltun, 2013); principled parallel variational inference with convergence guarantees is developed in (Baqué et al., 2016), while Regularized Frank–Wolfe provides a broader optimization framework that includes mean field as a special case (Lê-Huu & Alahari, 2021). Learnable high-dimensional filtering is explored in (Jampani et al., 2016). Deep unfolding has also been applied to segmentation through reconstruction-derived objectives (He et al., 2025b; 2026), whereas the unrolled updates here are those of a labeling energy.

Differentiable region grouping. SLIC provides a classical hard superpixel construction (Achanta et al., 2012), while Superpixel Sampling Networks learn task-specific soft superpixels with local candidate assignment (Jampani et al., 2018); recent work further studies differentiable superpixel formation and connectivity (Juybari et al., 2026). Slot Attention (Locatello et al., 2020) learns soft pixel-to-slot associations in the same spirit. Our incidences are computed once from each image–mask pair and held fixed; what changes across stages is the label distribution of each region.

Learned mask refinement. CascadePSP combines global and local refinement (Cheng et al., 2020), SegFix corrects boundaries (Yuan et al., 2020), and SegRefiner formulates refinement as a discrete diffusion process (Wang et al., 2023). For camouflaged object detection, UMBD uses uncertainty-guided Bernoulli diffusion to selectively refine uncertain regions while exploiting features from the upstream segmenter (Shen et al., 2025). Foundation-model-based refinement builds on SAM-style promptable segmentation (Kirillov et al., 2023; Ke et al., 2023): SAMRefiner derives noise-tolerant prompts from coarse masks (Lin et al., 2025), while PromptMoE uses mixture-of-experts prompting for refinement across tasks and domains (Price et al., 2026). Phoenix trains on semantically structured mask errors generated by adversarial perturbation and corrects them with contrastive refinement learning (Kim & Hwang, 2026). We differ from these refiners in the question asked: refinement here is inference in a shared segmentation energy, and the effect of that structure is isolated with matched controls.

3 METHOD

3.1 A CONDITIONAL RANDOM FIELD OVER MASKS AND LATENT REGIONS

Let I be an image and let $\mathcal{G}$ denote the inference grid with N spatial sites (stride 4 in our implementation); the initial per-site label distribution $P^{\text{init}} \in \Delta_L^N$ is the output of an arbitrary upstream model over L classes (L=2 for binary tasks), resized to $\mathcal{G}$. All site-indexed quantities below live on $\mathcal{G}$. We model refinement as approximate inference in a conditional random field over pixel labels $\mathbf{x} = (x_1, \ldots, x_N)$ and latent region labels $\mathbf{z} = \{z_k^{(r)} : 1 \le r \le R,\ k \in \mathcal{K}_r^+\}$, one per active region k at each of R granularity levels, where the active-region sets $\mathcal{K}_r^+$ are defined in Sec. 3.2:

$$E(\mathbf{x}, \mathbf{z} \mid I, P^{\text{init}}) \;=\; E_U(\mathbf{x}) + E_P(\mathbf{x}) + E_R(\mathbf{x}, \mathbf{z}), \tag{1}$$

conditioned jointly on I and P^{init}. A lightweight encoder consumes (I, P^{init}) and predicts the input-dependent quantities in Eq. (1): unary corrections, pairwise interaction weights, and pixel–region incidences; the global energy parameters, the label compatibility μ and the per-level strengths β_r and truncations κ_r, are learned jointly and shared across stages, as is the incidence temperature τ_r of Eq. (5). The damping coefficients α_t belong to the inference procedure (Sec. 3.3), not to the energy. The input-dependent quantities are computed once and held fixed across inference stages, so every stage performs inference on the same conditional energy; they remain differentiable, so gradients reach the encoder through the unrolled computation, with the exception of the active-region mask, which is fixed and treated as stop-gradient (App. B).

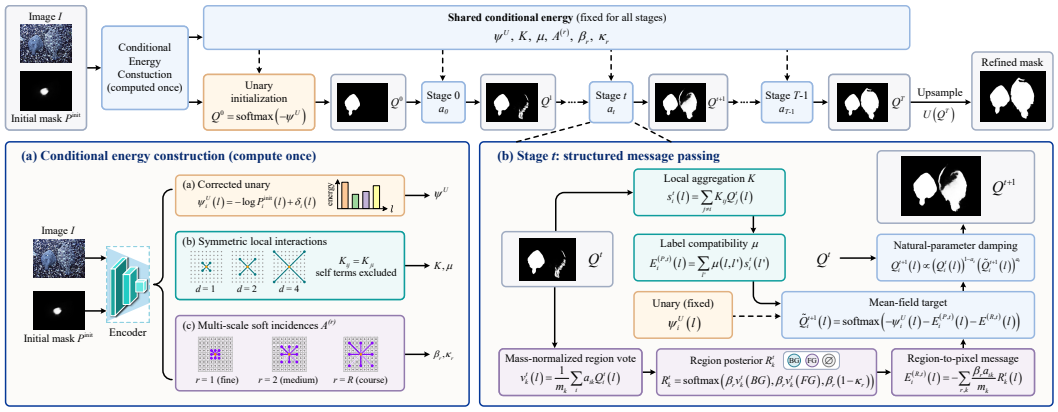

Figure 2: **Framework of UnfoldCRF.** Top: I and P^{init} define a conditional segmentation energy that is shared by all T stages and initializes Q^0; only the damping coefficient α_t varies by stage. Bottom left: the encoder constructs the corrected unary potential, symmetric local interactions, and multi-scale soft pixel–region incidences. Bottom right: stage t forms the region posterior R^t from Q^t, computes the pairwise and mass-normalized latent-region messages, and combines them with the fixed unary term to obtain $\widetilde{Q}^{t+1}$; natural-parameter damping then produces Q^{t+1}.

Corrected unary. The unary term uses the upstream prediction as its base and adds an image-conditioned residual:

$$E_U(\mathbf{x}) = \sum_i \psi_i^U(x_i), \qquad \psi_i^U(l) = -\log P_i^{\text{init}}(l) + \delta_{\theta,i}(l), \tag{2}$$

where P^{init} is clipped to $[\epsilon_0, 1-\epsilon_0]$ and renormalized ($\epsilon_0 = 10^{-4}$) so that hard upstream masks have finite log-evidence, and δ_θ is a per-pixel, per-label residual predicted by the encoder.

Symmetric local pairwise term. For neighboring sites we use a signed, image-conditioned pairwise interaction:

$$E_P(\mathbf{x}) = \sum_{i<j} K_{ij}\, \mu(x_i, x_j), \qquad K_{ij} = \tfrac{1}{2}\big(\widehat{K}_{ij} + \widehat{K}_{ji}\big)\, \mathbf{1}[j \in \Omega(i)]\,, \tag{3}$$

where $\Omega(i)$ is the union of 3×3 windows at dilations $\{1, 2, 4\}$ around pixel i and $\widehat{K}_{ij} = g_\theta(h_i, j-i)$ is a signed interaction weight predicted by a 1×1 convolution on the encoder features h_i, one output per offset $j - i \in \Omega$. The window relation is symmetric ($j \in \Omega(i)$ iff $i \in \Omega(j)$) and the explicit symmetrization gives $K_{ij} = K_{ji}$; the learnable label compatibility is symmetric by parameterization, $\mu = \frac{1}{2}(\hat{\mu} + \hat{\mu}^\top)$, so Eq. (3) is a well-defined energy over unordered pixel pairs. The pairwise term stays local; longer-range interactions go through the latent regions. Its mean-field message excludes the self contribution,

$$\mathcal{E}_i^P(l) = \sum_{l'} \mu(l, l') \sum_{j \neq i} K_{ij}\, Q_j(l'). \tag{4}$$

3.2 Latent-region robust consistency

Soft incidences and active regions. All levels share the same inference grid $\mathcal{G}$ but use different region-center densities. At level r, the encoder yields pixel embeddings $e_i^{(r)} \in \mathbb{R}^C$, K_r region centers are placed on a regular grid, and the center embedding $u_k^{(r)}$ is the average of the pixel embeddings in grid cell k. Following the local-candidate design of differentiable superpixels (Jampani et al., 2018), each pixel is softly assigned only to the s=9 centers of its 3×3 neighboring cells $\mathcal{N}_r(i)$:

$$a_{ik}^{(r)} = \frac{\exp\big(-\|e_i^{(r)} - u_k^{(r)}\|^2/\tau_r\big)}{\sum_{k' \in \mathcal{N}_r(i)} \exp\big(-\|e_i^{(r)} - u_{k'}^{(r)}\|^2/\tau_r\big)} \quad \text{for } k \in \mathcal{N}_r(i), \qquad a_{ik}^{(r)} = 0 \ \text{ otherwise}, \tag{5}$$

with a learned per-level temperature $\tau_r > 0$. Rows of $A^{(r)}$ sum to one, the scheme is the same for all tasks and label counts, and across the R levels incidence construction costs $O(RNsC)$. Because $e^{(r)}$ is computed from the encoder features of (I, P^{init}), incidences are conditioned on both the image and the initial mask. Let $m_k^{(r)} = \sum_i a_{ik}^{(r)}$ be the region mass and $\mathcal{K}_r^+ = \{k : m_k^{(r)} \geq m_{\min}\}$

the set of active regions; regions outside $\mathcal{K}_r^+$ carry no latent variable and neither enter the energy nor send messages. Since incidences are fixed before inference, $\mathcal{K}_r^+$ is fixed across stages.

Robust region energy with a null state. Each active region has a latent state $z_k^{(r)} \in \{1, \dots, L\} \cup \{\varnothing\}$, where $\varnothing$ allows the region to opt out of consistency. With strength $\beta_r = \mathrm{softplus}(b_r) > 0$ and truncation $\kappa_r = \sigma(\gamma_r) \in (0, 1)$, the mass-normalized region energy is

$$E_R(\mathbf{x}, \mathbf{z}) = \sum_{r=1}^{R} \sum_{k \in \mathcal{K}_r^+} \frac{\beta_r}{m_k^{(r)}} \sum_i a_{ik}^{(r)}\, c_{\kappa_r}\big(x_i, z_k^{(r)}\big), \qquad c_\kappa(x_i, z_k) = \begin{cases} \mathbf{1}[x_i \neq z_k], & z_k \neq \varnothing, \\ \kappa, & z_k = \varnothing. \end{cases} \tag{6}$$

Mass normalization expresses region disagreement as a weighted label fraction; after eliminating the latent state by MAP, the region penalty is capped at $\beta_r \kappa_r$ independently of soft mass (Proposition 1).

Mean-field updates. We use the factorization $q(\mathbf{x}, \mathbf{z}) = \prod_i Q_i(x_i) \prod_r \prod_{k \in \mathcal{K}_r^+} R_k^{(r)}(z_k^{(r)})$ and suppress the level index (r) when writing the updates of a single level. With the mass-normalized vote $v_k(l) = \frac{1}{m_k} \sum_i a_{ik}\, Q_i(l)$, the exact mean-field updates of Eq. (6) are (derivation in App. A)

$$R_k(l) \;\propto\; \exp\big(\beta_r\, v_k(l)\big), \qquad R_k(\varnothing) \;\propto\; \exp\big(\beta_r(1 - \kappa_r)\big), \tag{7}$$

$$\mathcal{E}_i^R(l) \;=\; -\sum_{r=1}^{R} \sum_{k \in \mathcal{K}_r^+} \frac{\beta_r\, a_{ik}^{(r)}}{m_k^{(r)}}\, R_k^{(r)}(l) \;+\; \text{const}, \tag{8}$$

where the constant is independent of l and cancels in the pixel softmax. The region branch therefore consists of mass-normalized sparse pixel–region aggregation, a softmax over region states, and mass-normalized sparse backprojection to pixels, at $O(RNsL)$ cost per stage.

Proposition 1 (MAP reduction to normalized robust truncation)**.** *Fix a hard labeling* $\mathbf{x}$ *and an active region* k *at level* r*, suppressing the level superscript for readability, and let* $p_k(l) = \frac{1}{m_k} \sum_i a_{ik}\, \mathbf{1}[x_i = l]$ *denote the mass-weighted fraction of region* k *carrying label* l*. Then*

$$\min_{z_k} E_{R,k}(\mathbf{x}, z_k) \;=\; \beta_r\, \min\big\{\kappa_r,\; 1 - \max_l p_k(l)\big\}. \tag{9}$$

A proof is given in App. A. Eq. (9) is a robust P^N-style potential over soft regions (Kohli et al., 2009): the penalty grows with label disagreement and is truncated at $\beta_r \kappa_r$ through the null state. At finite β_r the posterior R_k is soft, so $1 - \kappa_r$ is the vote fraction at which a semantic state becomes preferred over the null state ($v_k(l) > 1 - \kappa_r$), not a deterministic gate.

3.3 Unrolled damped mean-field inference

Inference initializes $Q_i^0 = \mathrm{softmax}(-\psi_i^U)$, which differs from P_i^{init} through the learned correction, and unrolls T stages indexed $t = 0, \dots, T-1$. Stage t executes, in order,

$$R^t \leftarrow \text{Eq. (7) evaluated at } Q^t \text{ for all active regions}, \tag{10}$$

$$\widetilde{Q}_i^{t+1}(l) \propto \exp\big[-\psi_i^U(l) - \mathcal{E}_i^{P,t}(l) - \mathcal{E}_i^{R,t}(l)\big], \tag{11}$$

$$\log Q_i^{t+1}(l) = (1 - \alpha_t)\, \log Q_i^t(l) + \alpha_t \log \widetilde{Q}_i^{t+1}(l) - \log Z_i^t, \tag{12}$$

where $\mathcal{E}^{P,t}$ and $\mathcal{E}^{R,t}$ are the messages of Eqs. (4) and (8) computed from Q^t and R^t, and $\alpha_t = \sigma(\eta_t) \in (0, 1)$ is a per-stage damping coefficient with learnable η_t. All energy parameters, embeddings, and incidences are shared across stages; only α_t is stage dependent. Q^T is the final grid-level marginal, $\mathcal{U}(Q^T)$ is the final refined mask, and R^T is formed from Q^T for diagnostics.

Lemma 1 (Damping preserves fixed points)**.** *For every* $\alpha \in (0, 1]$*, the damped update in Eq. (12) and the undamped mean-field update of the conditional energy in Eq. (1) have the same set of joint fixed points* (Q^*, R^*)*, provided that* Q^* *lies in the interior of the probability simplex.*

App. A proves Lemma 1; the iterates remain in the interior because ψ^U and all messages are finite. Parallel mean field need not monotonically decrease the variational free energy (Baqué et al., 2016; Krähenbühl & Koltun, 2013). We therefore report $\mathcal{F}(Q^t, R^t)$ and the undamped fixed-point residual $\mathrm{Res}_{\mathrm{FP}}(t) = \frac{1}{N} \sum_i \|\widetilde{Q}_i^{t+1} - Q_i^t\|_1$ as empirical inference diagnostics (Fig. 5).

Method	Resident params	BIG	VOC	DAVIS585	ECSSD	MSRA-B	Mean
Unrefined (IoU / BIoU)	n/a	78.3/70.1	66.7/60.1	80.1/83.0	81.4/70.2	75.2/61.9	76.3/69.1
CascadePSP (Cheng et al., 2020)	68M	+5.0/+6.3	+1.7/+0.7	−1.3/−1.5	+0.6/+1.0	+0.9/+2.7	+1.4/+1.9
SegRefiner (Wang et al., 2023)	119M	**+9.6/+12.5**	−3.9/−3.1	−10.9/−9.1	−15.0/−21.4	−10.7/−16.2	−6.2/−7.4
DualSight (Price et al., 2025)	641M	+3.9/+4.6	+3.3/+6.3	+1.5/−0.6	+2.0/+4.5	+2.1/+6.8	+2.6/+4.3
SAMRefiner (Lin et al., 2025)	641M	+6.8/+9.5	+7.1/+9.7	+3.3/+2.0	+5.1/+9.7	+4.7/+10.4	+5.4/+8.3
PromptMoE (Price et al., 2026)	≥641M	+8.5/+11.0	+7.9/+10.4	**+3.6/+2.4**	**+6.0**/+10.7	+5.1/**+10.5**	+6.2/+9.0
Attn. recurrent-S (matched)	2.6M	+2.4/+2.9	+2.7/+3.1	+0.6/+0.2	+2.3/+3.6	+2.1/+3.9	+2.0/+2.7
UnfoldCRF-S (ours)	2.6M	+5.1/+6.9	+5.2/+7.1	+2.4/+1.6	+4.3/+7.4	+4.0/+8.1	+4.2/+6.2
Attn. recurrent-L (matched)	92M	+6.2/+8.0	+5.9/+7.4	+2.1/+1.2	+4.6/+7.6	+4.2/+7.8	+4.6/+6.4
UnfoldCRF-L (ours)	92M	+8.7/+11.4	**+8.1/+10.6**	+3.4/+2.3	**+6.0/+10.8**	**+5.3**/+10.4	**+6.3/+9.1**

Table 1: **General open refinement** on the five-benchmark suite of Price et al. (2026): ΔIoU / ΔBIoU (%) over the unrefined masks (absolute scores in the first row). Published baselines use the reported high-resolution variants under the same coarse-mask protocol. Bold: best per column. The last four rows are our runs, reported as means over 5 training seeds.

Method	Source	CHAMELEON	CAMO	COD10K	NC4K
		M / F_β^ω / E_ϕ / S_α	M / F_β^ω / E_ϕ / S_α	M / F_β^ω / E_ϕ / S_α	M / F_β^ω / E_ϕ / S_α
SINet V2	TPAMI'21	.029/.792/.922/.890	.071/.733/.875/.822	.036/.668/.867/.820	.048/.769/.898/.848
+SegRefiner	NeurIPS'23	.073/.619/.812/.777	.116/.583/.763/.736	.079/.487/.764/.714	.075/.635/.830/.781
+Phoenix	ECCV'26	.030/.851/.915/.866	.079/.768/.834/.781	.034/**.766**/.872/.818	.049/**.822**/.883/.836
+UMBD	arXiv'25	.027/**.853**/**.955**/.892	.067/.774/.885/.829	.032/.733/.909/.825	.044/.807/.911/.849
+UnfoldCRF-S	Ours	**.026**/.849/.953/**.898**	**.066**/**.775**/**.890**/**.834**	**.031**/.740/**.910**/**.832**	**.043**/.812/**.915**/**.855**
FEDER	CVPR'23	.030/.824/.941/.888	.073/.740/.866/.799	.032/.713/.899/.822	.045/.796/.909/.847
+SegRefiner	NeurIPS'23	.041/.751/.908/.850	.081/.683/.848/.785	.043/.635/.866/.794	.050/.742/.896/.832
+Phoenix	ECCV'26	.034/.846/.925/.861	.081/.760/.818/.773	.032/**.779**/.884/.824	.047/**.826**/.886/.836
+UMBD	arXiv'25	.028/.838/.949/.890	.069/.757/.871/.807	.030/.732/.905/.827	.043/.809/.912/.849
+UnfoldCRF-S	Ours	**.026**/**.848**/**.953**/**.896**	**.067**/**.763**/**.879**/**.813**	**.029**/.745/**.910**/**.834**	**.042**/.815/**.917**/**.855**

Table 2: **Concealed target refinement** on COD under the UMBD (Shen et al., 2025) protocol: $M \downarrow$ / $F_\beta^\omega \uparrow$ / $E_\phi \uparrow$ / $S_\alpha \uparrow$ on four test sets with SINetV2 (Fan et al., 2021) and FEDER (He et al., 2023a) upstreams. SegRefiner (Wang et al., 2023) and UMBD results are quoted from UMBD; Phoenix (Kim & Hwang, 2026) is evaluated zero-shot on the same coarse masks, and our predictions use the same metric implementation. UMBD additionally uses upstream features.

3.4 Training and configuration

Let $\mathcal{U}$ denote bilinear upsampling from $\mathcal{G}$ to the image resolution followed by renormalization, and define $\ell(Q, Y) = \mathcal{L}_{\text{CE}}(\mathcal{U}(Q), Y) + \mathcal{L}_{\text{Dice}}(\mathcal{U}(Q), Y)$ for ground truth Y. The refiner is trained with

$$\mathcal{L} \;=\; \ell(Q^T, Y) \;+\; \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \ell(Q^t, Y), \tag{13}$$

where Q^0 is not supervised and P^{init} is resized to $\mathcal{G}$ by bilinear interpolation (App. B). Because the refiner reads only (I, P^{init}), it applies to any upstream model whose output mask is available.

4 Experiments

4.1 Experimental setup

Settings. *General open refinement* uses the five-benchmark train-once suite of Price et al. (2026) (BIG (Cheng et al., 2020), VOC (Everingham et al., 2010), DAVIS585 (Chen et al., 2022), ECSSD (Yan et al., 2013), MSRA-B (Liu et al., 2010)) with its original coarse-mask sources: DeepLabV3 (Chen et al., 2017), FCN-R50 (Long et al., 2015), and LR-ASPP-MobileNetV3 (Howard et al., 2019) for BIG/VOC; RMBG-1.4/2.0 and BiRefNet-matting for ECSSD and MSRA-B; and the SP and STM masks of DAVIS585. UnfoldCRF and its matched control are trained once on VOC 2012 train (Everingham et al., 2010) and evaluated frozen; we report ΔIoU and ΔBIoU (Cheng et al., 2021) over the initial masks. *Concealed refinement* follows UMBD (Shen et al., 2025): training uses CAMO (Le et al., 2019) and COD10K (Fan et al., 2020), and evaluation uses CHAMELEON (Skurowski et al., 2018), CAMO, COD10K, and NC4K (Lv et al., 2021), with SINetV2 (Fan et al., 2021) and FEDER (He et al., 2023a) as upstream generators and M, F_β^ω (Margolin et al., 2014), E_ϕ (Fan et al., 2018), and S_α (Fan et al., 2017) as metrics. Unless other-

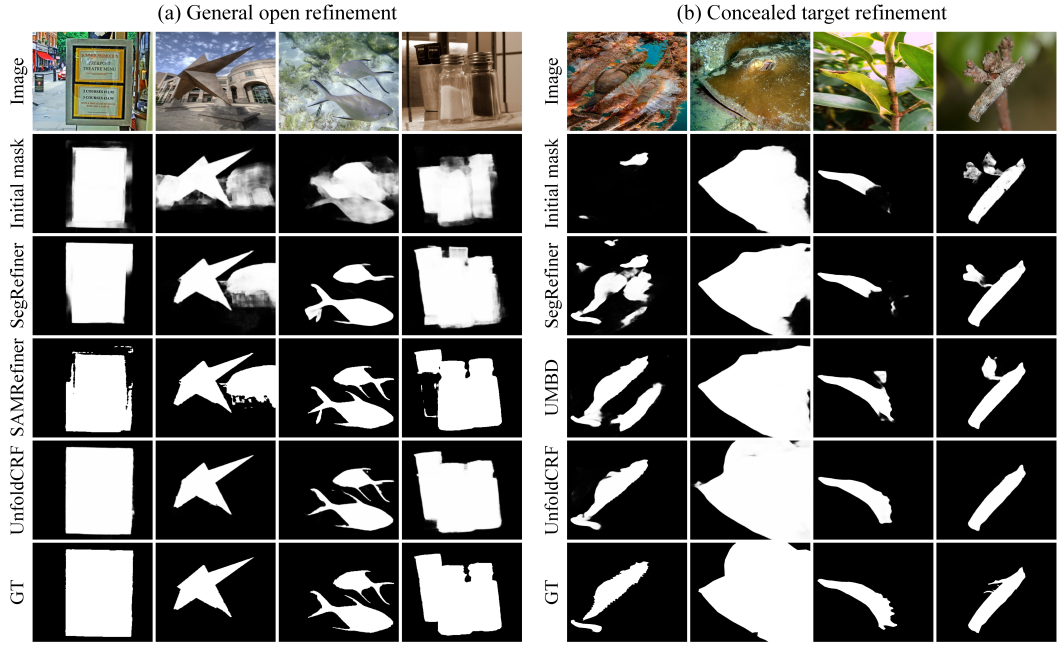

Figure 3: **Qualitative refinement.** (a) General open refinement on representative examples. (b) Concealed target refinement on representative COD examples. Columns show four examples; rows show the image, initial mask, SegRefiner, SAMRefiner/UMBD, UnfoldCRF, and ground truth.

wise stated, controlled ablations and mechanism analyses use COD10K with FEDER as upstream. Further protocol and training details are given in App. B.

Model variants and controls. UnfoldCRF-S is the standalone model of Sec. 3 (custom encoder, 2.6M parameters, 352^2 input) and is used for all controlled analyses. UnfoldCRF-L keeps the energy and inference operator unchanged but uses a frozen DINOv2-B/14 encoder (Oquab et al., 2023) whose intermediate features are fused with a local high-resolution branch at 448^2 input ($\sim$92M resident and $\sim$6M trainable parameters; App. B). For each variant, the matched attention control uses the same encoder and fusion architectures, with weights trained independently, the same stage count T, losses, and supervision, and replaces only the structured operator by $Q^{t+1} = F_{\text{attn}}(h, Q^t)$. For UnfoldCRF-S, the total parameter budget is matched; for UnfoldCRF-L, the frozen DINOv2 backbone is shared and the trainable non-backbone budget is matched.

4.2 Comparison with state-of-the-art refiners

General open refinement. Table 1 reports the five-benchmark suite. Among the published refiners, the SAM-based methods gain most, CascadePSP-Slow gains little, and SegRefiner-HR transfers poorly outside BIG. UnfoldCRF-S gains 4.2 mean ΔIoU and 6.2 mean ΔBIoU with 2.6M resident parameters, 2.2 points in ΔIoU above its matched control (paired 95% CI $[+1.6, +2.8]$; App. B). With frozen DINOv2-B features, UnfoldCRF-L reaches 6.3 mean ΔIoU and 9.1 mean ΔBIoU, comparable to PromptMoE on both averages with about one-seventh of the resident parameters, and remains 1.7 ΔIoU points above its matched control (paired 95% CI $[+1.1, +2.3]$). The 1.7-point gap remains when both models use the same frozen DINOv2 backbone, which rules out backbone strength as the sole explanation. Both variants gain most on BIG and VOC and least on DAVIS585, whose SP and STM initial masks are already accurate.

Concealed target refinement. Table 2 reports all four COD metrics because they rank the refiners differently. Phoenix has the highest weighted F-measure on COD10K and NC4K but lowers E_ϕ or S_α below the upstream mask in several settings, and SegRefiner lowers all four metrics everywhere. UnfoldCRF-S improves all four metrics on all eight upstream–dataset pairs, with the best MAE on all eight, the best E_ϕ on seven, and the best S_α on all eight. Unlike UMBD, it uses only the image and the initial mask rather than upstream hidden features; Fig. 3 shows representative corrections of boundary drift, missing structures, and false positives in both settings.

Operator	Params	$M \downarrow$	$F_\beta^\omega \uparrow$	$E_\phi \uparrow$	$S_\alpha \uparrow$	Harm↓
Components						
Upstream (FEDER)	n/a	.032	.713	.899	.822	n/a
Unary only	2.4M	.031	.725 ± .003	.903	.826	15.6 ± 0.6
Unary + learned pairwise	2.5M	.030	.736 ± .002	.906	.829	11.2 ± 0.5
Unary + learned latent regions	2.5M	.030	.738 ± .003	.907	.830	10.7 ± 0.5
Full UnfoldCRF-S (T=5)	2.6M	**.029**	**.745** ± .002	**.910**	**.834**	**8.5 ± 0.4**
Without null state	2.6M	.031	.737 ± .003	.906	.829	10.1 ± 0.5
Structure controls						
Unary + fixed pairwise	2.4M	.031	.731 ± .003	.904	.827	13.0 ± 0.6
Unary + fixed SLIC regions	2.4M	.031	.732 ± .004	.905	.828	13.5 ± 0.7
Black-box and formulation controls						
Conv recurrent (matched)	2.6M	.031	.733 ± .004	.905	.828	12.3 ± 0.6
Attention recurrent (matched)	2.6M	.030	.735 ± .003	.906	.829	11.7 ± 0.5
Reconstruction-unfolding control (RUN-style)	2.6M	.032	.712 ± .005	.885	.808	10.9 ± 0.5

Table 3: **Controlled comparison** on COD10K with FEDER upstream (5 seeds; SD for F_β^ω and Harm). Component ablations share the encoder and training setting; structure controls use predefined pairwise or region structure; recurrent controls match the stage count and parameter budget. Harm is the percentage of images whose IoU decreases after refinement.

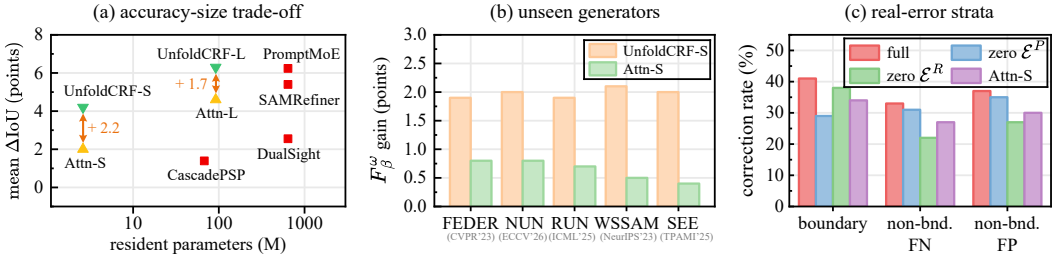

Figure 4: **Efficiency, generalization, and mechanism.** (a) Mean ΔIoU versus resident parameters; arrows indicate gains over matched recurrent controls at both scales. (b) F_β^ω gains on held-out generators. (c) Correction rates by error type under message interventions.

4.3 Ablation Study

Table 3 evaluates the contributions of the unary correction, the learned pairwise term, the latent regions, and the null state, with parameter counts reported for each variant. Structure controls replace learned interactions or regions with RGB–XY affinities (Krähenbühl & Koltun, 2011) or SLIC regions (Achanta et al., 2012; Arnab et al., 2016). Matched recurrent controls replace structured inference with generic recurrence under the same parameter budget and stage count; the reconstruction-unfolding control alternates reconstruction-state and mask updates in the style of RUN (He et al., 2025b) under the same budget (App. B). The full model improves F_β^ω by 1.0 point over the matched attention control (95% CI $[+0.6, +1.4]$, App. B) while reducing Harm from 11.7% to 8.5%; replacing learned structure with fixed structure, or removing the null state, lowers every metric.

4.4 Further Analysis

Generalization and mechanism. Figure 4a plots the accuracy–parameter trade-off of Table 1; panels (b,c) turn to generalization across generators and to mechanism. Figure 4b evaluates unseen-generator generalization on COD with FEDER, NUN (He et al., 2026), RUN (He et al., 2025b), WSSAM (He et al., 2023b), and SEE (He et al., 2025a). For each held-out generator, the refiner is trained on the remaining generators and evaluated frozen, comparing the F_β^ω gain of UnfoldCRF-S and Attn-S (paired difference 1.3 points, 95% CI $[+0.9, +1.9]$). Figure 4c partitions initially wrong pixels into boundary errors, non-boundary false negatives, and non-boundary false positives (definitions in App. B), and reports the fraction corrected after refinement. Zeroing the pairwise messages reduces boundary correction by 12 points but changes non-boundary FN/FP correction by only 2 points, whereas zeroing the region messages reduces non-boundary FN/FP correction by 11/10 points but boundary correction by only 3 points; the two branches act on different error types.

Stage-wise inference dynamics. Figure 5 follows one default checkpoint through $Q^0, \ldots, Q^T$ and, beyond the trained depth, through test-time extrapolation to Q^{2T} with $\alpha_t := \alpha_{T-1}$. The gap between

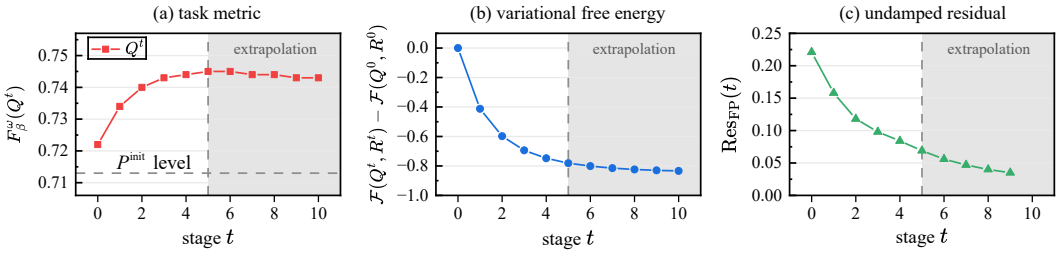

Figure 5: **Stage-wise inference dynamics** of one default checkpoint (T=5) on COD10K: (a) F_β^ω of intermediate marginals with the initial-mask level as reference, (b) variational free energy relative to Q^0 (test-set mean of the per-image difference), and (c) undamped fixed-point residual. Vertical dashed line: trained depth; shading: test-time extrapolation beyond T with $\alpha_t := \alpha_{T-1}$. Residuals are recorded before each update for $t = 0, \ldots, 2T-1$.

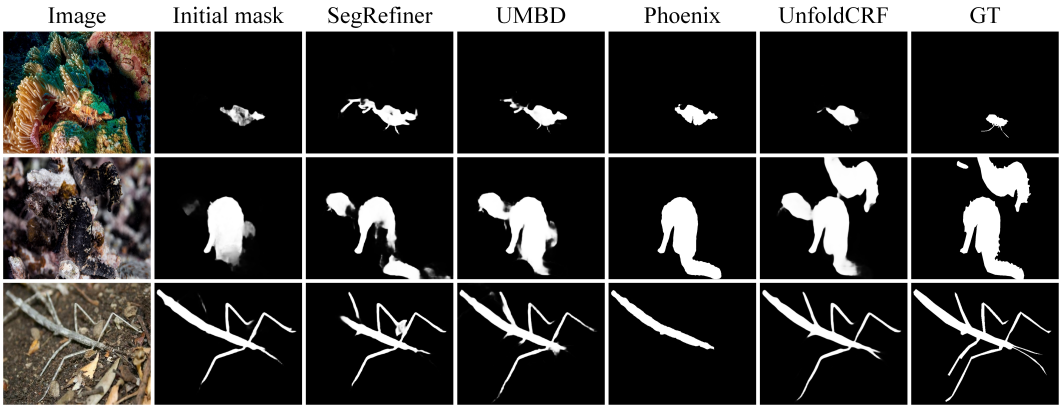

Figure 6: **Representative failure cases** with FEDER upstream. Columns show the image, initial mask, SegRefiner, UMBD, Phoenix, UnfoldCRF-S, and ground truth. Rows show a confidently mislabeled region, a missing object part, and a thin structure that remains difficult to recover.

P^{init} and $Q^0 = \text{softmax}(-\psi^U)$ is the learned unary correction, before any structured iteration. The task metric peaks near the trained depth and drifts slightly downward under further extrapolation, while the free energy and the undamped fixed-point residual keep decreasing beyond it.

5 Discussion and Limitations

Structured inference beyond representation strength. The controlled comparisons separate two things that a black-box refiner entangles: which pixels and regions interact, which the encoder decides, and how labels change under those interactions, which the shared energy decides. A 1.7-point gap survives when both models sit on the same frozen DINOv2 backbone, so the backbone cannot be the whole story. Zeroing the pairwise or region messages affects different error types. For the checkpoint examined in Fig. 5, segmentation accuracy peaks near the trained depth despite continued decreases in variational free energy.

Limitations. UnfoldCRF still depends on the quality of the upstream mask and on the resolution of its stride-4 inference grid: confidently wrong regions, missing object parts, and very thin structures are often not recovered (Fig. 6). UnfoldCRF-L also inherits the resident parameter cost of its DINOv2 backbone. Except for the train-once setting in Sec. 4.2, we train a separate refiner for each task, and our experiments are limited to binary masks although the formulation supports L labels; multi-class, instance-level, and temporal settings remain future work.

6 Conclusion

UnfoldCRF refines masks by unrolling damped mean-field inference in one image-conditioned segmentation energy. Matched recurrent controls, held-out generators, and message interventions separate its gain from parameter count and backbone strength, and they show that the pairwise term mostly corrects boundary errors while the region term mostly corrects non-boundary ones.

AI USE STATEMENT

Generative AI tools were used to assist with editing and revising the manuscript. The authors reviewed all AI-assisted changes and take full responsibility for the final content.

ETHICS STATEMENT

This work uses publicly available benchmark datasets under their respective terms of use and involves no human subjects, personal data collection, or annotation by the authors. As a mask refiner, UnfoldCRF can inherit errors from the upstream segmenter and may also worsen individual predictions; we therefore report image-harm rates alongside aggregate accuracy. The method should be validated for the target domain before use in safety-relevant applications.

REPRODUCIBILITY STATEMENT

The conditional energy, mean-field updates, and damping scheme are specified in Sec. 3, with derivations and proofs in App. A. We also provide input provenance, evaluation conventions, implementation and training details, and the construction of the matched controls in App. B.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, pp. 524–540. Springer, 2016.

Pierre Baqué, Timur Bagautdinov, François Fleuret, and Pascal Fua. Principled parallel mean-field inference for discrete random fields. In *CVPR*, pp. 5848–5857, 2016.

Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, pp. 402–418. Springer, 2016.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *CVPR*, pp. 1290–1299. IEEE, 2022.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, pp. 15329–15337. IEEE, 2021.

Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, pp. 8890–8899, 2020.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88:303–338, 2010.

Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, and Tao Li. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 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. *arXiv preprint arXiv:1805.10421*, 2018.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, and Jianbing Shen. Camouflaged object detection. In *CVPR*, pp. 2777–2787, 2020.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021.

Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pp. 22046–22055. IEEE, 2023a.

Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *NeurIPS*, 36, 2023b.

Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Longxiang Tang, Yulun Zhang, Linghe Kong, and Sina Farsiu. Segment concealed object with incomplete supervision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025a.

Chunming He, Rihan Zhang, Fengyang Xiao, Chengyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. Run: Reversible unfolding network for concealed object segmentation. *ICML*, 2025b.

Chunming He, Rihan Zhang, Dingming Zhang, Fengyang Xiao, Deng-Ping Fan, and Sina Farsiu. Nested unfolding network for real-world concealed object segmentation. *ECCV*, 2026.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, pp. 1314–1324, 2019.

Varun Jampani, Martin Kiefel, and Peter V Gehler. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks. In *CVPR*, pp. 4452–4461, 2016.

Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *ECCV*, pp. 363–380. Springer, 2018.

Jeremy Juybari, Josh Hamilton, Shuvra Das, Chaofan Chen, Andre Khalil, and Yifeng Zhu. Differentiable laplacian matrix guided superpixel segmentation. In *CVPR*, pp. 36168–36178, 2026.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NeurIPS*, 36:29914–29934, 2023.

Beomyoung Kim and Sung Ju Hwang. Learning from adversity: Semantic-aware mask refinement through adversarial perturbation. *ECCV*, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 3992–4003. IEEE, 2023.

Pushmeet Kohli, L'ubor Ladickỳ, and Philip HS Torr. Robust higher order potentials for enforcing label consistency. *Int. J. Comput. Vis.*, 82(3):302–324, 2009.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24, 2011.

Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *ICML*, pp. 513–521. PMLR, 2013.

L'ubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, pp. 739–746. IEEE, 2009.

Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.*, 184: 45–56, 2019.

Đ Khuê Lê-Huu and Karteek Alahari. Regularized frank-wolfe for dense crfs: Generalizing mean field and beyond. *NeurIPS*, 34:1453–1467, 2021.

Yuqi Lin, Hengjia Li, Wenqi Shao, Zheng Yang, Jun Zhao, Xiaofei He, Ping Luo, and Kaipeng Zhang. Samrefiner: Taming segment anything model for universal mask refinement. In *ICLR*, pp. 57368–57392, 2025.

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):353–367, 2010.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 33:11525–11538, 2020.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pp. 11591–11601, 2021.

Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pp. 248–255, 2014.

Maxime Oquab, Timothée Darcet, Théo 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. *arXiv preprint arXiv:2304.07193*, 2023.

Stephen Price, Kiran Judd, Kyle Tsaknopoulos, Rodica Neamtu, and Danielle L Cote. Dualsight: multi-stage instance segmentation framework for improved precision. *Sci. Rep.*, 15(1):27521, 2025.

Stephen Price, Danielle L Cote, and Elke A Rundensteiner. Promptmoe: A segmentation refinement framework leveraging mixture of experts for improved prompting. In *CVPR*, pp. 6325–6335, 2026.

Yuqi Shen, Fengyang Xiao, Sujie Hu, Youwei Pang, Yifan Pu, Chengyu Fang, Xiu Li, and Chunming He. Uncertainty-masked bernoulli diffusion for camouflaged object detection refinement. *arXiv preprint arXiv:2506.10712*, 2025.

Przemysław Skurowski, Hassan Abdulameer, and J Błaszczyk. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, pp. 7, 2018.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pp. 11158–11167. IEEE, 2019.

Marvin TT Teichmann and Roberto Cipolla. Convolutional crfs for semantic segmentation. *BMVC*, 2018.

Vibhav Vineet, Jonathan Warrell, and Philip HS Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *Int. J. Comput. Vis*, 110(3):290–307, 2014.

Mengyu Wang, Henghui Ding, Jun Hao Liew, Jiajun Liu, Yao Zhao, and Yunchao Wei. Segrefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *NeurIPS*, 36: 79761–79780, 2023.

Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, pp. 696–711, 2020.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pp. 1155–1162, 2013.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *ECCV*, pp. 489–506, 2020.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *ICCV*, pp. 1529–1537, 2015.

A Derivations

Notation. Throughout, expectations are taken under the mean-field factorization $q(\mathbf{x}, \mathbf{z}) = \prod_i Q_i(x_i) \prod_r \prod_{k \in \mathcal{K}_r^+} R_k^{(r)}(z_k^{(r)})$. For fixed (I, P^{init}), all quantities defining the conditional energy, including the incidences a_{ik}, region masses m_k, and active-region sets $\mathcal{K}_r^+$, are held fixed throughout inference. Since only active regions are considered, $m_k \geq m_{\min} > 0$. The level index (r) is suppressed when a single level is considered. For coordinate updates, conditional energies are written up to additive constants independent of the conditioned label, which cancel in the softmax; the full expectation is retained when evaluating the variational free energy. All derivations use the mass-normalized energy of Eq. (6), with $v_k(l) = \frac{1}{m_k} \sum_i a_{ik} Q_i(l)$.

Region update. Conditioning the region energy of an active region k on $z_k = l$ for a semantic label l gives

$$\mathbb{E}[E_{R,k} \mid z_k = l] = \frac{\beta_r}{m_k} \sum_i a_{ik} \big(1 - Q_i(l)\big) = \beta_r \big(1 - v_k(l)\big), \tag{S1}$$

while conditioning on $z_k = \varnothing$ gives

$$\mathbb{E}[E_{R,k} \mid z_k = \varnothing] = \frac{\beta_r}{m_k} \kappa_r \sum_i a_{ik} = \beta_r \kappa_r. \tag{S2}$$

Thus $R_k(\cdot) \propto \exp(-\mathbb{E}[E_{R,k} \mid \cdot])$. Removing the common factor $\exp(-\beta_r)$ from all unnormalized state probabilities yields $R_k(l) \propto \exp\big(\beta_r v_k(l)\big)$ and $R_k(\varnothing) \propto \exp\big(\beta_r (1 - \kappa_r)\big)$, which is Eq. (7).

Pixel message. Conditioning on $x_i = l$ and using $\sum_{s=1}^{L} R_k(s) \mathbf{1}[l \neq s] = 1 - R_k(\varnothing) - R_k(l)$, we obtain

$$\mathbb{E}[E_{R,k} \mid x_i = l] = \frac{\beta_r a_{ik}}{m_k} \Big[1 - R_k(\varnothing) - R_k(l) + \kappa_r R_k(\varnothing)\Big] + \text{const}. \tag{S3}$$

The only l-dependent term is $-\frac{\beta_r a_{ik}}{m_k} R_k(l)$; summing over levels and active regions gives Eq. (8). The null state's direct contribution is independent of the pixel label and therefore cancels in the pixel softmax. Its effect on the region message is mediated through the semantic probabilities $R_k(l)$, whose total mass is $1 - R_k(\varnothing)$.

Expected region energy. For one active region,

$$\begin{aligned} \mathbb{E}_q[E_{R,k}] &= \beta_r \left[\sum_{l=1}^{L} R_k(l) \big(1 - v_k(l)\big) + \kappa_r R_k(\varnothing) \right] \\ &= \beta_r \left[1 - \sum_{l=1}^{L} R_k(l) v_k(l) - (1 - \kappa_r) R_k(\varnothing) \right]. \end{aligned} \tag{S4}$$

Hence, over all levels and active regions,

$$\mathbb{E}_q[E_R] = \sum_{r=1}^{R} \sum_{k \in \mathcal{K}_r^+} \beta_r \left[1 - \sum_{l=1}^{L} R_k^{(r)}(l) v_k^{(r)}(l) - (1 - \kappa_r) R_k^{(r)}(\varnothing) \right]. \tag{S5}$$

We retain the full expectation so that the reported free energy corresponds to the conditional energy in Eq. (1).

Expected pairwise energy. The pairwise expectation is

$$\mathbb{E}_q[E_P] = \frac{1}{2} \sum_{i \neq j} K_{ij} \sum_{l,l'} Q_i(l)\, \mu(l, l')\, Q_j(l'). \tag{S6}$$

The factor $\frac{1}{2}$ converts the unordered-pair summation of Eq. (3) into a double sum, while the self term is excluded. The symmetries $K_{ij} = K_{ji}$ and $\mu(l, l') = \mu(l', l)$ ensure that the unordered-pair energy yields the full-neighborhood message in Eq. (4).

Proof of Proposition 1. For hard $\mathbf{x}$, $E_{R,k}(\mathbf{x}, z_k{=}l) = \frac{\beta_r}{m_k} \sum_i a_{ik} \mathbf{1}[x_i \neq l] = \beta_r \big(1 - p_k(l)\big)$, whereas $E_{R,k}(\mathbf{x}, z_k{=}\varnothing) = \beta_r \kappa_r$. Minimizing over z_k therefore selects the smaller of κ_r and $1 - \max_l p_k(l)$, scaled by β_r, which gives Eq. (9). □

Proof of Lemma 1. Write one undamped mean-field iteration as $R = \mathcal{R}(Q)$, using Eq. (7), followed by $\widetilde{Q} = \mathcal{T}(Q, R)$. The damped iteration uses the same two maps and then applies Eq. (12). Since the region map is identical in both systems, a joint fixed point (Q^*, R^*) satisfies $R^* = \mathcal{R}(Q^*)$, so it remains to compare the pixel maps at R^*.

If $\widetilde{Q} = Q$, then Eq. (12) returns $Q^{t+1} = Q$ for every $\alpha \in (0, 1]$. Conversely, let Q lie in the interior of the simplex and fix $\alpha \in (0, 1]$. Suppose Eq. (12) returns $Q^{t+1} = Q$. Then, for every pixel i, $\alpha\big(\log \widetilde{Q}_i(l) - \log Q_i(l)\big)$ is constant in l. Hence $\widetilde{Q}_i \propto Q_i$, and because both are normalized distributions, $\widetilde{Q}_i = Q_i$. The damped and undamped systems therefore have the same interior joint fixed points. The argument applies to each stage-specific coefficient $\alpha_t \in (0, 1]$. □

Variational free energy. The variational free energy is

$$\mathcal{F}(Q, R) = \mathbb{E}_q[E_U] + \mathbb{E}_q[E_P] + \mathbb{E}_q[E_R] + \sum_i \sum_l Q_i(l) \log Q_i(l) + \sum_{r=1}^{R} \sum_{k \in \mathcal{K}_r^+} \sum_{u \in \{1,\dots,L,\varnothing\}} R_k^{(r)}(u) \log R_k^{(r)}(u), \tag{S7}$$

where $\mathbb{E}_q[E_U] = \sum_i \sum_l Q_i(l)\psi_i^U(l)$ and the pairwise and region expectations are given above. For the Gibbs distribution $\mathcal{P}$ induced by the conditional energy, $\mathcal{F}(Q, R) = \mathrm{KL}(q \,\|\, \mathcal{P}) - \log Z$.

Inference diagnostics. Because the input-conditioned energy terms, incidences, and active-region sets are fixed across stages, $\mathcal{F}(Q^t, R^t)$ evaluates the same variational objective throughout refinement and is therefore directly comparable across stages, with $R^t = \mathcal{R}(Q^t)$. In particular, after obtaining the final marginals Q^T, the diagnostic state R^T is recomputed from Q^T.

The fixed-point residual $\mathrm{Res}_{\mathrm{FP}}(t)$ compares the undamped target $\widetilde{Q}^{t+1}$ with the current marginals Q^t, rather than Q^{t+1} with Q^t. Therefore its magnitude cannot be reduced merely by choosing a small damping coefficient.

B Protocols, Implementation, and Statistics

Input provenance. Table S1 records, for each setting, the coarse-mask source, the checkpoint or prediction provenance, the UnfoldCRF training data, and the source of the baseline numbers. Published numbers are quoted when official predictions are available or the corresponding mask-generation protocol can be matched; agreement of the aggregate unrefined scores serves only as a sanity check. Per-sample quantities are computed only from predictions available to us.

Setting	Coarse-mask source	Checkpoint or prediction source	UnfoldCRF training data	Baseline numbers
Five-benchmark suite	DeepLabV3, FCN-R50, LR-ASPP (BIG, VOC); RMBG-1.4/2.0, BiRefNet-matting (ECSSD, MSRA-B); SP/STM (DAVIS585)	official checkpoints (torchvision, Hugging Face); DAVIS585 masks from the dataset	VOC 2012 train	published where input provenance is matched; otherwise rerun
COD	SINetV2, FEDER	predictions regenerated from the official upstream checkpoints following the UMBD protocol	CAMO + COD10K train	published (Shen et al., 2025); Phoenix rerun zero-shot
Controlled (COD10K, FEDER)	FEDER	as for COD	CAMO + COD10K train	our runs

Table S1: **Provenance of inputs and baseline numbers.**

Evaluation. IoU and boundary IoU on the five-benchmark suite use the boundary-iou-api with dilation ratio 0.02 and are averaged per image. COD metrics follow the PySODMetrics implementation used by the UMBD protocol, reporting M, F_β^ω, E_ϕ, and S_α with the same normalization and threshold conventions. All predictions are resized back to the original benchmark resolution before evaluation. The harm rate is computed independently of any benchmark-specific normalization, by thresholding the initial and refined probability maps at 0.5 and comparing per-image IoU. For the error-stratified analysis of Fig. 4c, initially wrong pixels are identified at the original mask resolution after thresholding P^{init} at 0.5; boundary errors are those whose Euclidean distance to the ground-truth contour is at most 2% of the image diagonal, measured on both sides of the contour, and the

remaining errors are divided into non-boundary false negatives and non-boundary false positives, so that the three sets partition the initial errors. The relative width matches the bandwidth parameter of boundary IoU, although the band here is symmetric around the contour rather than the interior strip used by that metric. The correction rate of a set is the fraction of its pixels that are correct after thresholding $\mathcal{U}(Q^T)$ at 0.5, pooled over the test set; the damage rate, the fraction of initially correct pixels that become wrong, changes by less than 0.5 points under either intervention. Message interventions zero $\mathcal{E}^P$ or $\mathcal{E}^R$ at every stage of the same trained checkpoint at inference time, without retraining.

Architecture and training. Unless otherwise stated, configurations refer to UnfoldCRF-S. A lightweight convolutional encoder produces stride-4 features that define the inference grid $\mathcal{G}$. Pairwise interactions use multi-scale local neighborhoods, while latent regions are constructed at multiple spatial granularities with learned incidence embeddings and per-level temperatures. We use $T{=}5$ damped inference stages and train the model end to end with the loss in Eq. (13). Optimization uses AdamW with a cosine learning-rate schedule on 352^2 inputs, together with standard geometric and photometric augmentations and occasional synthetic perturbations of the training masks. Initial masks are resized and renormalized as required by the inference grid; refined predictions are mapped back to the original mask resolution before evaluation.

UnfoldCRF-L. UnfoldCRF-L keeps the energy formulation and inference operator unchanged while replacing the default encoder with a frozen DINOv2-B/14. Multi-level pretrained features are fused with a lightweight high-resolution branch conditioned on $[I; P^{\text{init}}]$ to produce the stride-4 feature map h used by the same unary, pairwise, and region modules as UnfoldCRF-S. The model uses a higher input resolution and proportionally adjusted region scales. Only the non-backbone modules are trained, while DINOv2-B/14 remains frozen. Additional architectural details are provided with the implementation. An S variant trained at 448^2 separates the contribution of the input resolution from that of the representation: it raises the mean ΔIoU on the five-benchmark suite from 4.2 to 4.6 points, against 6.3 for UnfoldCRF-L at the same resolution, so most of the S-to-L improvement is associated with the representation change rather than with resolution alone.

Matched controls. The recurrent controls match the encoder and unary-head architectures, training data, stage count T, loss, supervision, optimization protocol, and trainable parameter budget of their structured counterparts, while replacing structured inference with a stage-shared recurrent update from (h, Q^t) to Q^{t+1}. The attention control uses windowed multi-head self-attention on the concatenated $[h, Q^t]$ with $8{\times}8$ windows and four heads for S and six heads for L, whereas the convolutional control uses three $3{\times}3$ convolutional layers with GroupNorm and GELU; channel widths are chosen to match the trainable parameter budget. All models are trained independently and use the same initialization rule $Q^0 = \mathrm{softmax}(-\psi^U)$. For the L variants, the same frozen DINOv2-B/14 checkpoint and the same local-branch and fusion architectures are used, with matched trainable non-backbone budgets. The RUN-style control alternates an image-reconstruction state and a mask state for the same T stages: the reconstruction state minimizes an ℓ_1 reconstruction error of the input from foreground and background layers, and the mask is updated from the resulting reconstruction state, with encoder architecture, supervision, stage count, and parameter budget matched.

Statistics. For principal matched comparisons, we use a hierarchical bootstrap with 2,000 replicates. Static-image benchmarks are resampled at the image level, while DAVIS585 is resampled at the video–object-sequence level; all predictions of the compared methods and all upstream outputs associated with a sampled cluster are kept together to preserve test-sample pairing. Clusters are resampled separately within each benchmark. Training seeds are additionally resampled when multiple independently trained runs are available; unless the runs were explicitly paired during training, seeds are resampled independently across methods. Each replicate recomputes the statistic of interest: the dataset-level paired difference for COD10K, the macro average of per-benchmark differences for the five-benchmark suite, or the average difference across held-out generators. We report percentile 95% confidence intervals. When only one training seed is available, the interval reflects test-set resampling only and is reported as such. The COD10K and general-suite matched comparisons use 5 training seeds.

C COMPLETE RESULTS

Concealed refinement. Table S2 extends the controlled comparison of Table 3 to the remaining COD test sets.

Variant	CHAMELEON	CAMO	NC4K
Unary only	.029/.833/.945/.891	.071/.748/.869/.804	.044/.803/.911/.850
Unary + learned pairwise	.028/.842/.948/.892	.069/.758/.872/.807	.043/.810/.913/.851
Unary + learned latent regions	.028/.840/.951/.894	.069/.755/.877/.811	.043/.808/.915/.853
Full UnfoldCRF-S	.026/.848/.953/.896	.067/.763/.879/.813	.042/.815/.917/.855
Without null state	.027/.841/.949/.892	.069/.754/.874/.809	.043/.809/.914/.852
Unary + fixed pairwise	.029/.836/.946/.891	.070/.752/.870/.806	.044/.806/.912/.851
Unary + fixed SLIC regions	.028/.838/.947/.893	.070/.751/.873/.808	.044/.805/.914/.852
Conv recurrent (matched)	.028/.838/.947/.891	.070/.754/.872/.807	.044/.807/.913/.851
Attention recurrent (matched)	.027/.839/.949/.893	.069/.757/.873/.808	.043/.809/.913/.852
Reconstruction-unfolding control (RUN-style)	.028/.841/.950/.893	.069/.755/.875/.810	.043/.810/.915/.853

Table S2: **Controlled variants on the other COD test sets** (M / F_β^ω / E_ϕ / S_α; FEDER upstream; mean over 5 seeds). Results on COD10K for the same variants are reported in Table 3.