ATLAS: Evidence-Gated Annotation Transport for Satellite Semantic Segmentation

Research Proposal

September 19, 2026

Abstract

We propose a segmentation predictor built as a per-pixel, evidence-gated mixture of two complementary experts. The first expert is an ordinary parametric head. The second is nonparametric: it classifies by weighted label transport from ground-truth-labeled, mode-indexed atoms retrieved from the training corpus itself, so its decision rule adapts to the appearance mode (urban/rural, season, sensor) of the current scene. A two-scalar *evidence gate*, driven by retrieval affinity, decides per pixel how much jurisdiction the transport expert has: where the atlas offers strong same-mode support, the prediction leans on transported annotations; where it offers none, the predictor *provably* reduces to the parametric head — the blended posterior deviates from the parametric one by at most the gate value in total variation, and the argmax can flip only where the parametric margin is already below $w/(1-w)$. The parametric expert is itself distilled from the transport expert during training, so the fallback internalizes as much of the mode-conditional structure as weights can absorb; setting the gate to zero yields a memory-free deployment that is byte-for-byte the baseline architecture. The explicit objective of this project is state-of-the-art accuracy on LoveDA and ISPRS Potsdam/Vaihingen. The route: (i) a per-pixel oracle over two complementary experts upper-bounds either expert alone, and the gate is trained to capture that headroom; (ii) the transport head plugs into frozen released state-of-the-art checkpoints, inheriting their number and adding a gated delta; (iii) the atlas ingests new-mode labels without retraining, with the gate opening automatically as coverage arrives. We give a formal core (mode-compromise gap, self-adaptation bias law, a count-normalization lemma excluding hidden class priors, a gated-fallback safety proposition, an amortization bound), a decisive spherical-versus-plain atom ablation, controls that separate the mechanism from generic ensembling and from extra training compute, and a two-week frozen-feature go/no-go gate.

1 Motivation: the case for two experts and a gate

1.1 Why one fixed rule leaves accuracy on the table

High-resolution land-cover segmentation is a mixture over appearance modes — geography, season, sensor, urbanization. A parametric head trained against the marginal mixture adopts one compromise boundary for all modes (Proposition 1 quantifies the cost), and the field's response has been to make the rule image-adaptive at test time: object-contextual aggregation [49], context-aware and spatially adaptive classifiers [42, 33], dynamic class dictionaries [53, 32], dataset-level memory heads [18, 19]. These adapt using the model's *own predictions*, and Proposition 2 shows that correction is biased toward the confuser exactly where correction is needed. The alternative with unbiased conditioning — retrieval from ground-truth-labeled support [22, 2, 28] — has the opposite failure profile: it is strong where the corpus covers the scene and unreliable where it does not, and it retrieves everywhere, including where retrieval has nothing to add.

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.

1.2 The bet

The two experts fail in different places. A parametric head is smooth, always available, and mediocre on mode-dependent classes; a labeled-transport head is sharp where same-mode support exists and undefined off-atlas. A per-pixel *oracle* choosing the better expert therefore upper-bounds both — and that headroom is measurable before any training run (Section 8). This proposal builds the deployed predictor as exactly that mixture, with the oracle replaced by a two-scalar gate driven by retrieval affinity: evidence of atlas jurisdiction. The gate is what converts retrieval from a liability under domain shift into a bounded, self-limiting correction (Proposition 5): off-atlas, the predictor *is* the parametric head. The stated goal of the project is state-of-the-art accuracy on the standard benchmarks; every design choice below is subordinated to that goal, and every claim ships with the control that protects it (ensembling control, compute-matched training, leakage exclusions).

1.3 The deployed predictor, up front

An encoder–decoder produces features h_i and parametric logits $Z^s(i)$ with posterior $P^s(i)$. A transport expert (Section 3.2) produces $P^E(i)$ from retrieved labeled atoms, together with a retrieval affinity $\bar{\rho}(i)$. The deployed posterior is

$$P(i) \;=\; \big(1 - w(i)\big)\, P^s(i) \;+\; w(i)\, P^E(i), \qquad w(i) \;=\; \sigma\Big(\tfrac{\bar{\rho}(i) - \rho_0}{T_w}\Big), \tag{1}$$

with two learned scalars ρ_0, T_w, and $w(i) \coloneqq 0$ whenever retrieval returns nothing. Setting $w \equiv 0$ ("zero-cost mode") removes atlas, retrieval, and projector entirely and leaves the distilled parametric model — architecturally identical to the baseline. Everything else in this document exists to make Eq. (1) accurate, safe, and cheap.

2 Preliminaries: modes and the cost of self-adaptation

Let X be an image with pixels i, labels $y_i \in \{1, \dots, K\}$, and a latent appearance mode $m(X) \in \{1, \dots, M\}$. An encoder f_θ produces features $h_i \in \mathbb{R}^d$; a head g_ϕ produces parametric logits $Z^s(i) \in \mathbb{R}^K$. LoveDA [44] instantiates $M = 2$ (urban/rural) with metadata; ISPRS Potsdam/Vaihingen [36] provide the high-resolution urban regime. Two results frame the problem; both are verified numerically in the accompanying script (`verify_propositions.py`).

Proposition 1 (Mode-compromise gap)**.** *In a two-mode Gaussian shift model with inter-mode shift b and class separation Δ, for $0 \le b < \Delta/2$ the risk of the best marginal threshold rule exceeds the mode-conditional Bayes risk by $G(b)$, with $G(0) = 0$ and G strictly increasing in b. For $b \ge \Delta/2$ the modes interleave and no single threshold separates the classes; the gap grows further.*

Proposition 1 locates the transport expert's headroom: gains should concentrate on classes with large empirical inter-mode shift $\hat{b}_k$ (distance between mode-conditional class-k feature means), and in the interleaved regime a fixed-features linear rule *cannot* close the gap — the formal argument for keeping a conditional expert at inference rather than only distilling it away (Proposition 6(iii)).

Proposition 2 (Pseudo-label prototype bias)**.** *For class centres self-adapted from the model's own posteriors $\hat{p}$,*

$$\hat{c}_k \;=\; \frac{\sum_i \hat{p}_k(i)\, h_i}{\sum_i \hat{p}_k(i)}, \qquad \hat{c}_k - c_k^\star \;=\; \frac{\sum_{j \neq k} \alpha_{jk}\, \big(c_j^\star - c_k^\star\big)}{1 + \sum_{j \neq k} \alpha_{jk}}, \tag{2}$$

where α_{jk} is the confusion mass that class-j pixels contribute to the class-k posterior, relative to the mass from true class-k pixels. The bias points toward the confuser and grows with exactly the error it is meant to fix.

Proposition 2 is why the second expert's memory carries *ground-truth* labels rather than model predictions: its error is then retrieval mismatch (Proposition 4) — measurable, independent of the model's own error rate, and shrinkable with coverage — rather than a bias that grows with the error it is meant to fix [41].

3 Method

Two experts, one gate. The nonparametric expert consults the labeled corpus through a mode-indexed atom memory; the gate assigns it per-pixel jurisdiction; the parametric expert is distilled from it so the fallback is as strong as weights allow. We frame the method throughout as *what the predictor consults* — transported annotations where the corpus has jurisdiction — never as "a memory module plus an ensemble."

3.1 Labeled atom memory

The memory is a set of ground-truth-labeled region atoms

$$\mathcal{M} \;=\; \big\{(c_a,\; y_a,\; m_a,\; \iota_a)\big\}, \tag{3}$$

where $c_a \in \mathbb{R}^d$ is the mean feature of a ground-truth-connected region of class y_a in training image ι_a of mode m_a. During training, atoms are computed with an EMA copy $f_{\bar\theta}$ of the encoder ($\bar\theta \leftarrow \eta\bar\theta + (1-\eta)\theta$, as in momentum-queue contrastive learning [11]) and organized as per-(k, m) FIFO queues of capacity Q, refreshed every T_{ref} iterations (Lemma 2 sets T_{ref}). After training, the atlas is rebuilt once with the final encoder and frozen; deployment needs no labels at test time and, for fixed areas of interest, the atlas can be pre-resolved per geography. The mode index comes from LoveDA metadata (urban/rural) in the primary setting, with k-means scene clusters as the no-metadata generality ablation. Engineering envelope (to be measured, not asserted): a LoveDA-scale atlas is on the order of $4{\times}10^4$ atoms (≈21 MB fp16), queried per *region* rather than per pixel via ANN search [20].

Leakage control. Training-time retrieval for pixel i excludes atoms with $\iota_a = \text{image}(i)$ and, on Potsdam/Vaihingen (crops of single acquisitions), atoms within a geographic exclusion radius; otherwise the expert is near-duplicate ground-truth lookup and its training-time accuracy is memorization. We treat this as non-negotiable; the un-excluded number appears only as a labeled diagnostic.

3.2 The transport expert

A projector $\psi_e : \mathbb{R}^d \to \mathbb{R}^{d_e}$ (2-layer MLP) maps features into the transport space; queries and atoms are normalized to the unit sphere,

$$\tilde e_i = \frac{\psi_e(h_i)}{\|\psi_e(h_i)\|_2 + \varepsilon}, \qquad \tilde c_a = \frac{\psi_e(c_a)}{\|\psi_e(c_a)\|_2 + \varepsilon}. \tag{4}$$

Section 4 argues why the sphere is a theorem hypothesis rather than a styling choice — and, new in this design, why the gate's threshold ρ_0 is meaningless without it. For pixel i, retrieval returns a mode-matched candidate set $R(i) \subset \mathcal{M}$, subject to the exclusions above; write $R_k(i) =$

$\{a \in R(i) : y_a = k\}$. With within-class-normalized weights $w_a(i) = \mathrm{softmax}_{a \in R_k(i)}\big(\rho_a(i)/\tau_r\big)$, where $\rho_a(i)$ is the retrieval affinity, the expert logit is

$$Z_k^E(i) \;=\; b_k \;+\; \tau \log \sum_{a \in R_k(i)} w_a(i)\, \exp\Big(\tfrac{1}{\tau}\,\langle \tilde{e}_i, \tilde{c}_a \rangle\Big), \qquad Z_k^E(i) = b_k - 1 \text{ if } R_k(i) = \varnothing, \tag{5}$$

with a single global temperature τ and one learned scalar bias b_k per class — the only class-indexed parameters anywhere in the method. $P^E = \mathrm{softmax}(Z^E)$. Two lemmas govern this rule.

Lemma 1 (Count/prior separation)**.** *If all similarities lie in* $[-1, 1]$ *and* $\sum_{a \in R_k(i)} w_a(i) = 1$*, then* $Z_k^E(i) - b_k \in [-1, 1]$ *uniformly in* $|R_k(i)|$*, and* Z^E *is invariant to replication of the whole retrieved multiset* $R_k(i)$*. Hence per-class atlas coverage cannot act as a hidden class prior;* b_k *is the only class prior.*

Lemma 2 (Staleness)**.** *The map* $x \mapsto \tau \log \sum_a w_a e^{x_a/\tau}$ *is* 1*-Lipschitz in* $\|\cdot\|_\infty$*. If every atom similarity drifts by at most* ε *between memory refreshes, every expert logit drifts by at most* ε*. The refresh period* T_{ref} *is therefore a principled dial for the training-time queue schedule.*

Proposition 3 (Containment)**.** *With a single mode-independent atom per class and uniform weights, Eq.* (5) *reduces to a cosine-normalized linear classifier* $Z_k = b_k + \langle \tilde{e}_i, \tilde{c}_k \rangle$*. Frozen-prototype heads [52, 26] and dataset-level memory classifiers [18, 19] are further special cases: the transport family strictly contains the rules it competes with.*

Proposition 4 (Expert bias is retrieval mismatch)**.** *If a fraction* $\beta = 1 -$ *(retrieval precision) of retrieved same-class atoms come from the wrong mode* m'*, the class-conditional density the expert effectively uses is* $(1 - \beta)\, p(h \mid k, m) + \beta\, p(h \mid k, m')$*. The expert's error is independent of the parametric head's error rate and non-increasing in atlas coverage — the structural opposite of Proposition 2.*

3.3 The per-pixel evidence gate

Retrieval affinity is the gate's evidence of jurisdiction:

$$\bar{\rho}(i) \;=\; \max_{a \in R(i)} \langle \tilde{e}_i, \tilde{c}_a \rangle \quad \text{(class-agnostic; top-}k\text{ mean as an ablation)}, \qquad w(i) = \sigma\Big(\tfrac{\bar{\rho}(i) - \rho_0}{T_w}\Big), \tag{6}$$

with $w(i) = 0$ if $R(i) = \varnothing$, and the deployed posterior given by Eq. (1). The gate has exactly two learned scalars; a per-class gate $w_k(i)$ from $\bar{\rho}_k(i)$ is an ablation, not the default. Blending posteriors (rather than logits) is the default because it yields the guarantees below; the logit-space blend is an ablation arm.

Proposition 5 (Gated fallback safety)**.** *For the predictor of Eq.* (1)*:*

(i) ***Bounded deviation.*** $\|P(i) - P^s(i)\|_{\mathrm{TV}} \;=\; w(i)\, \|P^E(i) - P^s(i)\|_{\mathrm{TV}} \;\le\; w(i)$*.*

(ii) ***Flip condition.*** *The argmax of* $P(i)$ *can differ from the argmax of* $P^s(i)$ *only if the parametric top-1/top-2 margin satisfies* $P^s_{(1)}(i) - P^s_{(2)}(i) < \frac{w(i)}{1 - w(i)}$ *(in particular, for* $w(i) \le \frac{1}{2}$*, only where the margin is below* $2w(i)$*). The transport expert has jurisdiction exactly where the parametric head is already uncertain.*

(iii) ***Off-atlas reduction.*** *If retrieval returns nothing, or* $\bar{\rho}(i) \le \rho_0 - cT_w$*, then* $w(i) \to 0$ *and* $P(i) \to P^s(i)$*: under total domain novelty (a mode absent from the atlas), the deployed model* is *the distilled parametric model. Proofs are one line each from Eq.* (1)*; numerical checks go into the verification script.*

Proposition 5 is the formal answer to the domain-shift objection against test-time retrieval ("a river trained in one continent, tested in another"): the gate turns unfamiliarity into abstention, so the retrieval path can help but — at low affinity — provably cannot hurt beyond the parametric fallback. The residual risk the gate cannot see is *high-affinity, class-inverted* retrieval (a wrong-class atom that genuinely resembles the query); this is exactly per-class cross-domain retrieval precision, which we audit directly (Section 7.3) rather than assume.

3.4 Training objective and lifecycle

Let $\bar{P}^E$ denote the expert computed entirely on the EMA encoder/projector and EMA atoms (targets that do not chase the student). With distillation temperature T_d [13]:

$$\begin{aligned}\mathcal{L} \;=\; \mathcal{L}_{\mathrm{NLL}}\big(P,\,Y\big) + \lambda_s\,\mathcal{L}_{\mathrm{CE}}\big(Z^s, Y\big) \;+\; \lambda_e\,\mathcal{L}_{\mathrm{CE}}\big(Z^E, Y\big) \\ + \lambda_T(t)\;T_d^2\,\mathrm{KL}\Big(\mathrm{sg}\big[\bar{P}^E_{T_d}\big]\;\Big\|\;P^s_{T_d}\Big).\end{aligned} \tag{7}$$

The first term trains the deployed object itself, including the gate scalars. The two auxiliaries keep each expert individually strong: λ_s guarantees the fallback is a competent standalone model, λ_e trains ψ_e, b, τ. The KL term distills the expert into the parametric head; its deployment payoff is direct — wherever the two experts agree, the gate is harmless, and the KL term explicitly shrinks the disagreement mass $\mathbb{P}(g \neq E)$ that the gate must adjudicate (Proposition 6(ii)). $\lambda_T(t) = 0$ during a CE-only warm-up (early features produce early garbage atoms; risk R5). An optional Lovász term [3] may be added symmetrically.

Gate calibration by construction. Two augmentations prevent the degenerate always-on gate: *atlas dropout* (with probability p_d a batch sees an empty atlas, forcing $w = 0$ and keeping the fallback path trained) and *mode holdout* (with probability p_h retrieval for a batch is restricted to off-mode atoms, teaching ρ_0 where affinity stops being evidence). Hyperparameters are fixed once on Vaihingen validation with ResNet-18 and reused unchanged for every other encoder, precluding per-backbone tuning.

3.5 Operating points

ATLAS (gated) — Eq. (1); the headline configuration and the state-of-the-art candidate. **ATLAS-zero** ($w \equiv 0$) — atlas, retrieval, and projector deleted; byte-for-byte the baseline architecture at exactly zero overhead; always reported alongside the gated row. **Ingestion** — new-mode ground-truth regions are appended to the atlas without any retraining; the gate opens automatically as coverage arrives. The gap between gated and zero rows is the measured value of keeping the corpus at inference, reported per class.

4 Why spherical atoms (and the ablation that settles it)

Normalization in Eq. (5) is a hypothesis of Lemma 1 and of the gate's semantics, not a cosmetic choice. Five arguments, in decreasing order of strength.

1. The count lemma is false without it. With raw dot products, $\langle h, c\rangle$ scales with $\|c\|$, and deep feature and classifier norms are known to track class frequency — the observation behind τ-normalized classifiers in long-tailed recognition [21], weight imprinting [35], cosine classifiers in few-shot learning [8], and the cosine fix for magnitude bias in incremental learning [14]. Off the sphere, atom norms become hidden per-class confidence knobs that compound with per-class atlas coverage — exactly the pathology Lemma 1 excludes. On the sphere, every similarity lies in $[-1, 1]$ uniformly in $|R_k|$ and in class frequency, and b_k remains the only class prior.

Arm	Similarity	Purpose
A (default)	cosine, both normalized, temperature τ	full spherical rule
B	raw dot product, none normalized, tuned global scale s	is the sphere cosmetic?
C	atoms normalized only	isolates atom-norm effects
D	features normalized only	isolates query-norm effects
E	$-\|h-c\|^2/\tau$ (NCM-style Euclidean)	the third natural metric
F	cosine with $\tau = 1$	isolates the temperature from the geometry

Table 1: Sphere-versus-plain ablation arms. All other components of Eq. (5), the gate, and the training recipe are held fixed. The gate uses each arm's own affinity, so arm B also tests argument 2 directly.

2. The gate threshold needs a bounded, scale-free affinity. The gate compares $\bar{\rho}(i)$ to a learned threshold ρ_0. Cosine affinity is bounded in $[-1, 1]$ with fixed semantics across training, modes, and backbones, so ρ_0 means something. Raw dot-product affinity drifts with feature norms over training and across modes; a threshold on it silently degrades into an always-on or always-off gate, and Proposition 5 loses its bite.

3. The norm is a mode-correlated nuisance channel. Feature norms are modulated by illumination, season, sensor gain, and augmentation strength — nuisances that are mode-correlated in Earth-observation imagery. Projecting to the sphere quotients out precisely this direction; a plain-vector dictionary would let the expert encode "brightness" as class evidence.

4. The probabilistic reading makes the ablation predictable in advance. On the sphere, $\langle \tilde{e}, \tilde{c} \rangle / \tau$ is the von Mises–Fisher log-likelihood up to constants [34], so Eq. (5) is a vMF-mixture posterior with concentration $1/\tau$. The Euclidean alternative decomposes as

$$-\|h-c\|^2 \;=\; 2\langle h, c\rangle \;-\; \|h\|^2 \;-\; \|c\|^2, \tag{8}$$

re-introducing a class-dependent bias ($\|c\|^2$) and an input-dependent bias ($\|h\|^2$): Euclidean scoring is the dot product plus the same hidden knobs in disguise. On unit vectors all three coincide up to monotone transforms; off the sphere they diverge in exactly the ways the lemma forbids.

5. Distillation stability. Bounded expert logits give T_d fixed semantics across training, and the staleness bound of Lemma 2 becomes bounded angular drift.

4.1 Ablation specification

Tier-2 recipe (Section 7.1); everything fixed except the similarity. Table 1 lists the arms; arm B receives its own global scale search on validation for fairness.

Diagnostics (mean mIoU alone is not the point). (i) Per-class mean atom norm versus class pixel frequency (correlation and plot); expected strong in B/C/D, absent by construction in A. (ii) *Replication-invariance test*: at evaluation time, duplicate one tail class's atom multiset $\times\{2, 5, 10\}$ and measure the change in that class's predicted share and IoU. Lemma 1 implies arm A is exactly invariant, while the raw log-sum-exp grows by approximately $\tau \log K$ under K-fold duplication. The test runs in minutes and doubles as a numerical unit test of the lemma. (iii) Gate reliability per arm: AUROC of $\bar{\rho}$ for atlas-coverage detection, and the stability of the learned ρ_0 across seeds. (iv) Tail-versus-head class F1 split; (v) calibration (ECE).

Pre-registered prediction. Scale-tuned arm B may match arm A on mean in-distribution mIoU; it will fail replication invariance, exhibit the norm–frequency correlation, destabilize ρ_0, and degrade tail-class F1 and calibration. If B matches A on *all* diagnostics, normalization is cosmetic and we will report exactly that.

5 Formal core

Roles at a glance. Proposition 1 locates the transport expert's headroom and, in the interleaved regime, the argument for keeping it at inference. Proposition 2 is why the memory is labeled. Proposition 3 guarantees the expert is never less expressive than the parametric rules it competes with. Proposition 4 makes the expert's bias measurable and shrinkable. Lemma 1 excludes hidden class priors; Lemma 2 disciplines the training memory. Proposition 5 bounds what the gate can do to the fallback. What remains is the division of labor between weights and memory:

Proposition 6 (Amortization)**.** *Work with* 0–1 *loss. Let* $g^\star_{\mathrm{cond}}(h, m)$ *denote the mode-conditional Bayes rule with risk* $R^\star_{\mathrm{cond}}$*, and let* E *denote the transport expert.*

(i) ***Realizability.*** *If there exists a probe* q *with* $\mathbb{P}\big(q(h) \neq m\big) \leq \delta$*, then* $\tilde{g}(h) := g^\star_{\mathrm{cond}}(h, q(h))$ *is a function of* h *alone and satisfies* $R(\tilde{g}) \leq R^\star_{\mathrm{cond}} + \delta$*. Distillation can therefore push the parametric fallback toward the conditional rule whenever the mode is decodable from features; the obstruction is learnability under cross-entropy, not capacity.*

(ii) ***Gap chain.*** *For any parametric head* g*:* $R(g) \leq R(E) + \mathbb{P}(g \neq E)$ *and* $R(E) \leq R^\star_{\mathrm{cond}} + \varepsilon_{\mathrm{retr}}(\beta)$*, with every term estimable (held-out disagreement; the mixture of Proposition 4). The KL term of Eq.* (7) *directly minimizes* $\mathbb{P}(g \neq E)$ *— which is simultaneously the mass the gate must adjudicate at inference, so distillation quality and gate robustness improve together.*

(iii) ***Limit (why the expert stays).*** *If the deployed rule is restricted to threshold or linear rules on fixed features, then in the interleaved regime* $b \geq \Delta/2$ *of Proposition 1,* $\inf_g R(g) - R^\star_{\mathrm{cond}} \geq G(b) > 0$*: no amount of distillation closes the gap there, and per-class gains of the gated model over the zero-cost model are predicted to concentrate exactly on those classes.*

Proofs will be completed and added to the verification script before any claim ships; statements are final, proofs are sketched.

6 Positioning

Neighbors and deltas. *Memory heads in segmentation*: MCIBI/MCIBI++ [18, 19] keep a dataset-level (marginal) memory at inference with no jurisdiction control; ECAC [40] distills a ground-truth-center teacher but keeps marginal memory and coarse-segmentation machinery at inference. Ours is conditional (class × mode), count-normalized (Lemma 1), and gated with a provable fallback. *Training-only memories*: cross-image pixel contrast [47, 1] and relational distillation [48] discard the memory at inference and use it as pairwise or contrastive signal, not as a deployable labeled expert; they are controlled arms, not just citations. *Retrieval-augmented prediction*: kNN-LM interpolates a parametric model with retrieval at a fixed global coefficient [22]; learned per-token interpolation exists in NLP [10], and REALM/RETRO keep the datastore always-on [9, 4]. Our delta: the gate is per-pixel, two-parameter, affinity-driven, and comes with the fallback guarantee of Proposition 5; the retrieved objects are ground-truth *labels* (annotation transport [28, 43, 16, 2]), not text tokens; and the parametric side is distilled from the expert it is blended with, so the mixture components cohere [23, 51]. *Mixtures and selective prediction*: the predictor is an input-gated mixture in the lineage of adaptive mixtures of experts [17] and per-pixel selective prediction [7]; the specific content — one expert nonparametric over

GT-labeled mode-indexed atoms, gate driven by retrieval affinity with a total-variation fallback bound — is, to our knowledge, new for dense prediction. Prototype and generative heads [52, 26] and adaptive classifiers [49, 42, 33, 53, 32] remain the special cases identified by Proposition 3.

Novelty sentence. A per-pixel, evidence-gated fusion of a distilled parametric head with a ground-truth labeled, mode-conditional transport expert — provably reducing to the parametric head off-atlas, provably free of coverage-induced class priors, and able to ingest new-mode labels without retraining.

7 Experimental plan (state of the art is the objective)

7.1 Benchmarks, protocol parity, and tiers

Benchmarks: LoveDA [44] (headline; headroom at ∼55 mIoU, native mode structure, official online test) and ISPRS Potsdam/Vaihingen [36] under the SFA-Net split protocol [15]. Protocol parity with published rows includes their test-time augmentation; all Potsdam/Vaihingen rows report mean±std over at least three seeds because baselines near 90 mIoU imply sub-point deltas. Every headline table carries *two* rows per configuration: **ATLAS (gated)** — the state-of-the-art candidate — and **ATLAS-zero** — identical architecture to the baseline at zero overhead.

- **Tier 1 (matched backbone).** ResNet-18 [12] against UNetFormer, ABCNet, MSGCNet, RS3Mamba, and CFENet [46, 25, 50, 31, 37], and Swin-T [29] against DC-Swin [45], plain FPN neck [27]. The zero row keeps the attribution airtight (same deployed graph); the gated row goes for the top of the table.
- **Tier 2 (controlled).** One encoder (ResNet-18+FPN), one recipe; the classifier/loss is the only variable (Section 7.2).
- **Tier 3 (best config).** DINOv3 ViT-B [39, 6] (ConvNeXt-T [30] in the scaling table), always paired with its own same-backbone softmax control, never cited against weaker-backbone methods.

State-of-the-art plug-in (highest-probability route). Freeze the strongest released checkpoints [46, 37, 53], build the atlas from each model's own features, train only ψ_e, b, τ, and the gate, and deploy the gated blend on top. This inherits the published number and adds a gated delta at small overhead; by Proposition 5 the blend cannot fall meaningfully below the frozen model where evidence is weak. It is also the cheapest experiment in the plan (≈2–3 weeks) and therefore runs early.

Confound control: training compute. Matched iterations across arms; wall-clock and GPU-hours reported; and a budget-matched baseline arm (the plain recipe given the same extra training budget) so that "you just trained more" is pre-empted rather than rebutted.

7.2 Controlled comparison (Tier 2)

Arms on the identical ResNet-18+FPN recipe: plain CE; CE + cross-image contrastive memory [47, 1]; CE + CIRKD [48]; CE + ECAC-style ground-truth-center distillation [40] (if reimplementable); MCIBI++ with its memory at inference [19]; **a deep-ensemble control** — two independently trained parametric heads blended with a learned scalar at matched inference cost [24]; ours zero; ours hard-attach ($w \equiv 1$); **ours gated**. This table answers, at matched everything: (a) does the gated model beat generic ensembling — the sharpest confound; (b) does it beat inference-memory and training-memory neighbors; (c) what the gate adds over hard attachment; (d) what the zero row gets for free.

7.3 Oracle headroom and gate analysis

Oracle-gate headroom: per-pixel best-of-two-experts using ground truth — the upper bound of Eq. (1) — reported next to the learned gate's capture fraction. *Gate reliability*: AUROC of $\bar{\rho}$ for atlas-coverage detection (in-mode versus held-out-mode queries); distribution of w in-domain versus under shift; per-class cross-domain retrieval precision (the pooled number hides class-order inversions, which are the failures the gate cannot see).

7.4 Structured-gain falsifiers

Per-class gain (gated over zero, and zero over CE) regressed on measured inter-mode shift $\hat{b}_k$ — Propositions 1 and 6(iii) predict positive slopes, with the gated-over-zero gain concentrated on interleaved classes; a flat regression falsifies the mechanism. Boundary-band versus interior split; ECE; mode-identifiability probe before/after training.

7.5 Domain shift and ingestion

LoveDA urban↔rural with zero target-mode atoms: the gate should close and the gated row should match the zero row — *no degradation is the success criterion*, per Proposition 5(iii). Then the ingestion curve: append target-mode ground-truth atoms (no retraining) and plot mIoU versus atoms ingested — the one capability no parametric competitor has at any price, and the figure that replaces a single attach-versus-detach bar.

7.6 Backbone scaling and generality

Paired zero-versus-gated mIoU, gain, retrieval precision, and gate statistics per rung (R18 / R50 / Swin-T / ViT-S / ViT-B); if gains collapse on CNN rungs, the mechanism is a foundation-feature artifact and the text will say so. Stretch arm for generality beyond remote sensing: Cityscapes [5] jointly with ACDC [38], condition (clear, fog, night, rain, snow) as the mode; the gate should close on conditions absent from the atlas and the ingestion curve should open it.

7.7 Efficiency, stated plainly

The gated model is not free: report parameters, GFLOPs, latency, and atlas memory against the baseline, alongside the zero row (exactly baseline cost) and an accuracy-versus-latency Pareto plot. Engineering estimates to be replaced by measurements: per-image regional ANN queries against a ≈21 MB atlas; the transport head on the order of one decoder layer.

7.8 Ablation discipline (pre-registration)

Every ablation arm ships with a pre-registered predicted sign and rough magnitude, following Section 4.1 and this section. An ablation table whose removals all move aggregates by hundredths of a point is worse than no table.

8 Go/no-go gate (two weeks, frozen features, no end-to-end training)

All on frozen DINOv3 features over LoveDA's urban/rural partition, reusing open-source dense nearest-neighbour evaluation code[1]:

[1] `https://github.com/vpariza/open-hummingbird-eval`

1. **Transport-versus-probe.** Training-free transport minus a linear probe on identical features, with per-class retrieval precision in-domain and under shift. Training-free transport on strong frozen features is known to be positive [2]; only the delta and its concentration on high-$\hat{b}_k$ classes count.
2. **Oracle-blend headroom (decisive).** Per-pixel best of {probe, transport} minus the better single head. If the oracle headroom is negligible, the gated mixture has nothing to capture and the thesis is dead before any training run.
3. **Affinity as evidence.** AUROC of $\bar{\rho}$ separating in-atlas-mode from held-out-mode queries must exceed a preset threshold; below it the gate cannot work and we fall back to distillation-only (zero mode as the sole deliverable) or stop.
4. **Mode identifiability.** A probe on frozen features must reach a preset accuracy — the necessary condition of Proposition 6(i).

9 Risks and kill criteria

R1 — Headroom capture. The oracle gap may be real but the learned two-scalar gate may capture little of it. Mitigations: the gate-analysis suite; per-class and per-mode gates as ablations; the plug-in route, whose baseline is already the table top. *Kill:* if gated $\leq$ deep-ensemble control at matched cost, the transport expert adds nothing beyond generic ensembling.

R2 — Ensembling confound. The sharpest reviewer attack on any mixture. Fixed by protocol: the deep-ensemble control [24] ships in the controlled table and is referenced next to every headline claim.

R3 — Adjacency. kNN-style interpolation [22, 10], mixtures of experts [17], memory heads [19, 40]. Mitigations: head-to-heads in Section 7.2; the fallback guarantee and count lemma, which none of the neighbors state; framing as what the predictor consults, never as a module.

R4 — Training-compute confound. Fixed by protocol (Section 7.1).

R5 — Early-training circularity. Bad features $\Rightarrow$ bad atoms $\Rightarrow$ bad expert $\Rightarrow$ bad gate. Mitigations: CE-only warm-up; EMA-encoder atoms [11]; the frozen-checkpoint plug-in as the de-risked first deliverable.

R6 — Gate miscalibration under shift. High-affinity class-inverted retrieval is invisible to the gate. Mitigations: per-class cross-domain precision audits; mode-holdout training; conservative ρ_0 selected on a held-out mode; Proposition 5 is honest that its guarantee covers low-affinity novelty only.

R7 — Overhead criticism. The gated model adds inference cost. Mitigations: the zero row is published in every table at exactly baseline cost; overhead is measured and plotted, not estimated; the Pareto framing makes the trade explicit rather than hidden.

10 Timeline and deliverables

Order of work, with GeoSeg[2] as the training trunk — protocol-native for all three benchmarks and the codebase family behind several comparison rows; nothing proceeds until UNetFormer R18 on Vaihingen reproduces to within $\sim$0.2 mIoU of its published 83.35: (0) trunk reproduction; (1) the frozen-feature gate of Section 8, $\approx$2 weeks; (2) the plug-in experiment on released

[2] `https://github.com/WangLibo1995/GeoSeg`

checkpoints — the earliest state-of-the-art candidate — ≈2–3 weeks; (3) end-to-end training on ResNet-18 (queues, EMA encoder, gate calibration augmentations), 3–4 weeks; (4) Swin-T and DINOv3 tiers, from which the scaling table falls out; (5) domain-shift, ingestion-curve, oracle-headroom, and efficiency audits. Working back from the expected CVPR 2027 submission deadline (mid-November 2026 on the conference's established cadence; the official call for papers had not been posted at the time of writing), stages (0)–(3) must complete by late October 2026; the Cityscapes→ACDC stretch arm and the full Tier-3 rows are contingent on stage-(3) results. Deliverables: the verification script extended to Propositions 5 and 6 and the replication-invariance unit test of Lemma 1; the controlled table with the ensemble control; the sphere ablation; the ingestion curve.

Protocol commitments (pre-registered). Same-image and geographic exclusion apply to training-time retrieval; the deep-ensemble control accompanies every headline claim; gate statistics (w distribution, affinity AUROC, per-class cross-domain precision) accompany every domain-shift row; ECAC [40] is cited and differentiated head-to-head; every ablation arm carries a pre-registered predicted sign and magnitude; Tier-3 rows are never cited against weaker-backbone methods; and the stated scope of Proposition 1 ($b < \Delta/2$ for the closed-form gap) and of Proposition 5 (low-affinity novelty only) is carried through all claims.

References

[1] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C. Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021.

[2] Ivana Balažević, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier J. Hénaff. Towards in-context scene understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[3] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240, 2022.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[7] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[8] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[9] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938, 2020.

[10] Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. Efficient nearest neighbor language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[15] Gyutae Hwang, Jiwoo Jeong, and Sang Jun Lee. SFA-Net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sensing*, 16(17):3278, 2024.

[16] Juan Eugenio Iglesias and Mert R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015.

[17] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

[18] Zhenchao Jin, Tao Gong, Dongdong Yu, Qi Chu, Jian Wang, Changhu Wang, and Jie Shao. Mining contextual information beyond image for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7231–7241, 2021.

[19] Zhenchao Jin, Dongdong Yu, Zehuan Yuan, and Lequan Yu. MCIBI++: Soft mining contextual information beyond image for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5988–6005, 2023.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

[21] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[22] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

[23] Kalle Kujanpää, Pekka Marttinen, Harri Valpola, and Alexander Ilin. Efficient knowledge injection in LLMs via self-distillation, 2024. arXiv:2412.14964.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[25] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M. Atkinson. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181:84–98, 2021.

[26] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. GMMSeg: Gaussian mixture based generative semantic segmentation models. In *Advances in Neural Information Processing Systems*, volume 35, pages 31360–31375, 2022.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[28] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2368–2382, 2011.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[31] Xianping Ma, Xiaokang Zhang, and Man-On Pun. RS3Mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

[32] Xiaowen Ma, Rongrong Lian, Zhenkai Wu, Hongbo Guo, Fan Yang, Mengting Ma, Sensen Wu, Zhenhong Du, Wei Zhang, and Siyang Song. LOGCAN++: Adaptive local-global class-aware network for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025.

[33] Xiaowen Ma, Zhenliang Ni, and Xinghao Chen. SSA-Seg: Semantic and spatial adaptive pixel-level classifier for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 37, pages 93690–93713, 2024.

[34] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000.

[35] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5822–5830, 2018.

[36] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sergio Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(3):293–298, 2012.

[37] Shufen Ruan, Quan Wan, Ruijuan Chen, Mengyang Hu, Xiuya Guo, and Kunfang Song. Context-aware feature enhancement network for remote sensing image semantic segmentation. *Remote Sensing*, 18(4):543, 2026.

[38] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[39] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3, 2025. arXiv:2508.10104.

[40] Huadong Tang, Youpeng Zhao, Min Xu, Jun Wang, and Qiang Wu. Classifier enhancement using extended context and domain experts for semantic segmentation, 2025. arXiv:2510.25174.

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[42] Zhuotao Tian, Jiequan Cui, Li Jiang, Xiaojuan Qi, Xin Lai, Yixin Chen, Shu Liu, and Jiaya Jia. Learning context-aware classifier for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[43] Joseph Tighe and Svetlana Lazebnik. SuperParsing: Scalable nonparametric image parsing with superpixels. In *Proceedings of the European Conference on Computer Vision*, pages 352–365, 2010.

[44] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[45] Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng, and Shenghui Fang. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[46] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[47] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.

[48] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022.

[49] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 173–190, 2020.

[50] Qingyu Zeng, Junshi Zhou, Jing Tao, Lei Chen, Xutong Niu, and Yun Zhang. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:5622913, 2024.

[51] Jianyi Zhang, Aashiq Muhamed, Aditya Anantharaman, Guoyin Wang, Changyou Chen, Kai Zhong, Qingjun Cui, Yi Xu, Belinda Zeng, Trishul Chilimbi, and Yiran Chen. ReAugKD: Retrieval-augmented knowledge distillation for pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[52] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022.

[53] Xuechao Zou, Yue Li, Shun Zhang, Kai Li, Shiying Wang, Pin Tao, Junliang Xing, and Congyan Lang. Dynamic dictionary learning for remote sensing image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.