CANDIDATESHIFT: AVAILABILITY-AWARE LEARNING FOR DESCRIPTOR-FREE 2D-3D MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Descriptor-free 2D-3D matching associates image keypoints with map landmarks without using local visual descriptors. Retrieval and map sampling can remove every valid correspondence for a keypoint, changing its correct matching outcome to unmatched. We present CandidateShift, a training method that separates match availability from conditional landmark identity under candidate deletion. Geometric correspondence labels supervise whether a valid match survives. For keypoints that remain matchable, the prediction on the full candidate set is restricted and renormalized over retained landmarks to provide a soft target for the prediction on the reduced set. Both predictions come from a shared matcher and receive direct correspondence supervision. Compared with the same matcher trained with supervised candidate perturbations, CandidateShift improves conditional identity accuracy under candidate deletion and yields more accurate localization, with gains that persist under stronger outlier contamination.

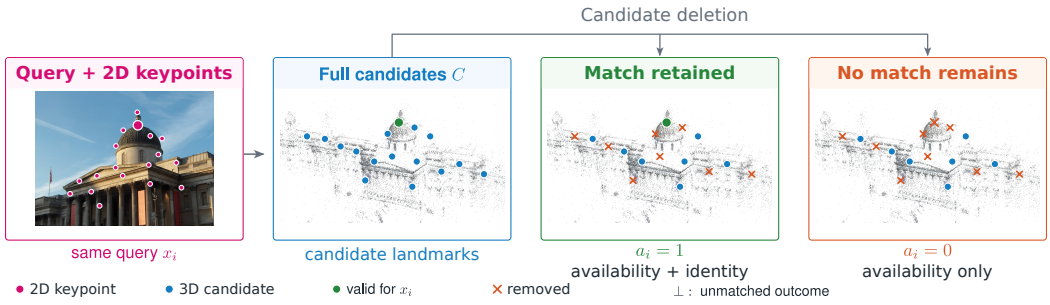

Figure 1: **Candidate deletion changes the matching outcome.** The pair objective supervises availability and identity when a valid landmark survives; otherwise, it supervises availability alone, with unmatched outcome $\perp$.

1 INTRODUCTION

Visual localization estimates a camera pose by matching image observations to a three-dimensional map. Image retrieval selects database views, whose observed landmarks provide candidates for association with query keypoints. At the matching stage, descriptor-free methods associate keypoints with landmarks using geometry and, in some models, color or semantic attributes, without local visual descriptor inputs (Zhou et al., 2022; Wang et al., 2024; Zhang et al., 2025; 2026a;b). The term descriptor-free refers to the matcher inputs; image retrieval may use global image descriptors. Predictions depend on the landmarks supplied by retrieval and sampling.

Changing the candidate set changes the matching problem itself. Removing distractors preserves a valid correspondence but alters its geometric context. Figure 1 illustrates the different outcome when every valid landmark is removed: the query must become unmatched. A model must determine both whether a match remains available and which retained landmark corresponds to the query. These decisions cannot be reduced to invariance under input perturbation. The same image keypoint can require different matching labels for two subsets of one map, even though its image coordinates and camera pose are unchanged.

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.

Existing matchers learn unmatched assignments, and supervised perturbations expose them to different candidate sets. Such supervision does not directly relate predictions for identities shared by a full set and its subset. Conversely, normalizing predictions over retained landmarks produces an identity distribution even when all valid correspondences have been removed. Applying this target to a query with no surviving match would regularize preferences among candidates that are all invalid. We introduce CandidateShift with three contributions:

- We separate match availability from conditional landmark identity under known candidate deletion, accounting for queries whose correct output changes to unmatched.
- We combine geometric availability labels with stopped-gradient identity targets obtained by restricting and renormalizing full-set predictions over retained landmarks. Both branches share a matcher and retain direct correspondence supervision.
- We show that the added supervision improves MegaDepth localization and conditional identity prediction under deletion, and transfers to A2-GNN without adding inference modules. Cross-scene evaluations examine its sensitivity to the query detector.

2 Related work

Structure-based localization and map representations. Structure-from-motion recovers sparse maps (Schönberger & Frahm, 2016). Hierarchical localization retrieves places with global descriptors, such as NetVLAD, before establishing local correspondences (Sarlin et al., 2019; Arandjelović et al., 2016). SuperGlue and LightGlue refine sparse descriptor matches (Sarlin et al., 2020; Lindenberger et al., 2023); LoFTR learns features for detector-free matching (Sun et al., 2021). Map compression reduces storage by changing the retained scene representation (Camposeco et al., 2019); cross-descriptor localization instead learns compatibility between feature types (Dusmanu et al., 2021). Our focus is the supervision needed when the landmark identities available to a matcher change.

Descriptor-free image-to-map matching. BPnPNet learns correspondences and camera pose jointly without visual descriptors (Campbell et al., 2020). GoMatch (Zhou et al., 2022) makes sparse image and map geometry comparable by expressing landmarks as bearing coordinates in known database views. DGC-GNN (Wang et al., 2024) uses global structure to guide local matching; A2-GNN (Zhang et al., 2025) emphasizes angular relations and annular neighborhoods. GeoMix (Zhang et al., 2026b) combines local geometric embeddings with learned global context. These three methods also use color. SAG-GNN (Zhang et al., 2026a) introduces semantic distributions for feature fusion and cross-set interaction. We retain the matching architecture and study the relation between predictions on different candidate sets.

Partial assignment and candidate deletion. Correspondence filtering uses agreement among putative matches to distinguish inliers from outliers (Yi et al., 2018; Zhang et al., 2019; Zhao et al., 2021); PointDSC applies spatial consistency to 3D point-cloud registration (Bai et al., 2021). Transport-based matchers represent partial assignment with unmatched outcomes (Sarlin et al., 2020; Zhou et al., 2022). GoMatch, A2-GNN, and GeoMix already supervise unmatched keypoints and landmarks (Zhou et al., 2022; Zhang et al., 2025; 2026b). Known deletions supply updated unmatched labels and an identity map between the retained candidates. CandidateShift uses the labels to supervise availability and the identity map to transfer conditional predictions, supplementing the existing unmatched supervision of the backbone.

Prediction transfer and input variation. Knowledge distillation transfers soft predictions (Hinton et al., 2015), while Mean Teacher uses weight-averaged consistency targets (Tarvainen & Valpola, 2017). We instead restrict the full-set prediction of the shared model to retained landmarks, applying it only to queries with surviving matches. Deletion changes neighborhood context and assignment competition, so consistency acts as a regularizer. FiLM (Perez et al., 2018) conditions the perturbed input, and group DRO (Sagawa et al., 2020) reweights perturbation groups. Both are also used by the supervised-perturbation reference, isolating our additional objectives.

3 METHOD

CandidateShift processes reference and perturbed inputs with a shared matcher. Both predictions receive correspondence supervision; known landmark deletions additionally supervise match availability and conditional identity. Figure 2 summarizes the training paths and the unchanged inference procedure.

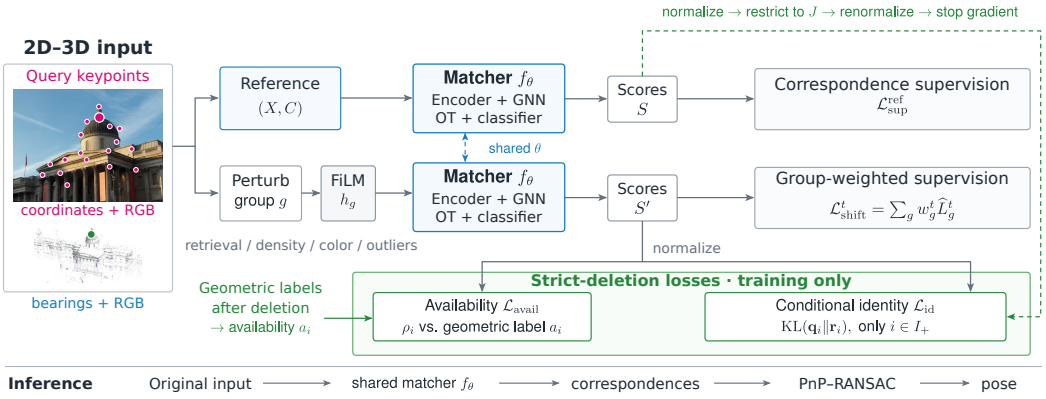

Figure 2: **Training procedure of CandidateShift.** Reference and perturbed inputs share a matcher; strict deletions add availability supervision and conditional identity transfer.

3.1 MATCHING REPRESENTATION

Let $X = \{x_i \in \mathbb{R}^2\}_{i=1}^n$ be query keypoints and $C = \{c_j \in \mathbb{R}^3\}_{j=1}^m$ candidate map landmarks. The matcher receives normalized query coordinates, landmark bearing coordinates in a known database camera, and RGB vectors on both sides. World coordinates and camera geometry determine correspondence labels and support pose estimation. We retain the GeoMix architecture, including its geometric encoders, within-set and cross-set attention, global context nodes, transport assignment, and correspondence classifier (Zhang et al., 2026b). The transport layer follows entropically regularized assignment (Cuturi, 2013).

The matcher f_θ, with shared parameters θ, produces nonnegative scores $S \in \mathbb{R}^{(m+1)\times(n+1)}$, with landmark rows and query columns. The last row represents an unmatched query, denoted by $\perp$; the last column represents unmatched landmarks. For each real query column, we form

$$\mathbf{p}_i = \mathcal{N}_\epsilon(S_{\cdot i}), \qquad [\mathcal{N}_\epsilon(z)]_k = \frac{\max\{z_k, \epsilon\}}{\sum_l \max\{z_l, \epsilon\}}, \tag{1}$$

where $\epsilon > 0$ is a numerical floor and the sum runs over all entries of z. Thus $p_i(j)$ is the probability of landmark j and $p_i(\perp)$ is the unmatched probability. The unmatched-landmark column is excluded. These distributions define the additional training objective; inference retains the original correspondence selection. We denote the inherited transport and correspondence-classification loss by $\mathcal{L}_{\text{sup}}$. Numerical constants and optimization settings appear in Section 4.1.

3.2 CONSTRUCTING RELATED CANDIDATE SETS

Each training access samples one of four perturbation groups uniformly to transform a reference sample. Both inputs must contain query keypoints and candidate landmarks; otherwise, the next example is tried with a new perturbation. The proposals are renewed online.

Retrieval. We construct the perturbed input from an alternative retrieved database image. For a list of length $k \geq 2$, its zero-based rank is sampled uniformly from 1 to $\min\{k_r, k-1\}$. The new view changes the candidate context and supplies its own geometric correspondence labels. Since this construction does not provide a fixed map of query and landmark identities between the two inputs, it receives direct supervision without a cross-set identity target.

Density and outlier composition. Density perturbation retains a uniformly sampled subset of $m' = \min\{m, \max\{m_{\min}, \mathrm{round}(\alpha_d m)\}\}$ landmarks. Outlier perturbation instead keeps all landmarks without a positive correspondence and subsamples those that have at least one. If m_+ landmarks have a positive correspondence, we retain $\min\{m_+, \max\{m_{+,\min}, \mathrm{round}(\alpha_o m_+)\}\}$ of them uniformly. When the resulting set is smaller than $m_{\min}$, additional landmarks are sampled from the complement up to the available total. Labels are recomputed from the final retained set, since added landmarks can restore a match. Query identities are preserved, and retained landmark indices align the two score matrices for identity comparison.

Color. One RGB channel is selected independently on each side and set to zero. The normalized colors are then quantized to L_c uniformly spaced levels on $[-1, 1]$. This perturbation changes appearance while retaining all candidate identities. Its correspondence supervision remains active, but it does not receive the deletion objective.

For a density or outlier perturbation, let $J \subseteq \{1, \ldots, m\}$ be the retained indices and $C' = \{c_j : j \in J\}$. The pair objective applies only when J is nonempty and strict. A small candidate set may remain unchanged after the minimum-size constraint; it still contributes supervised loss. Let $Y_{ji} \in \{0, 1\}$ denote a geometric correspondence in the reference input. Multiple valid landmarks per query are allowed. We retain $Y'_{ji} = Y_{ji}$ for $j \in J$ and recompute the unmatched entries:

$$Y'_{\perp i} = \mathbf{1}\left[\sum_{j \in J} Y_{ji} = 0\right], \qquad Y'_{j\perp} = \mathbf{1}\left[\sum_{i=1}^{n} Y_{ji} = 0\right], \qquad Y'_{\perp\perp} = 0. \tag{2}$$

This updates the matched status of each query when its last valid candidate is deleted. Slicing the original unmatched row without recomputing it would leave that query with no positive label.

3.3 Supervising match availability

A conditional distribution over retained landmarks sums to one even when all are invalid. We therefore supervise availability separately, using identity consistency only when a ground-truth match survives.

To describe the distinction, let T_J denote the candidate-deletion operator on probability distributions. It maps a distribution on $\{1, \ldots, m\} \cup \{\perp\}$ to one on $J \cup \{\perp\}$, preserving retained identities and collecting removed probability mass in the unmatched outcome:

$$(T_J \mathbf{p}_i)(j) = p_i(j) \quad (j \in J), \qquad (T_J \mathbf{p}_i)(\perp) = p_i(\perp) + \sum_{j \notin J} p_i(j), \tag{3}$$

where $T_J \mathbf{p}_i$ is the resulting distribution and the sum ranges over removed real landmarks. This operator preserves total probability, but errors in the reference prediction can give high availability to an entirely invalid subset. We instead determine availability from geometric labels and use the reference prediction for conditional identity.

Let $\mathbf{p}'_i$ be the normalized perturbed prediction, indexed by the original landmark identities in J. Its availability target and predicted probability are

$$a_i = \mathbf{1}\left[\sum_{j \in J} Y_{ji} > 0\right], \qquad \rho_i = \sum_{j \in J} p'_i(j) = 1 - p'_i(\perp). \tag{4}$$

Define $I_+ = \{i : a_i = 1\}$ and $I_- = \{i : a_i = 0\}$, and let $\mathcal{H}$ contain the nonempty sets among I_+ and I_-. We use class-balanced binary cross-entropy:

$$\mathcal{L}_{\mathrm{avail}} = \frac{1}{|\mathcal{H}|} \sum_{H \in \mathcal{H}} \frac{1}{|H|} \sum_{i \in H} \left[-a_i \log \bar{\rho}_i - (1 - a_i) \log(1 - \bar{\rho}_i)\right], \tag{5}$$

where $\bar{\rho}_i = \mathrm{clip}(\rho_i, \delta, 1 - \delta)$ and $0 < \delta < 1/2$ prevents logarithms of zero. Availability refers to a surviving geometric correspondence under the label rule, rather than physical visibility. The two nonempty classes receive equal weight; if only one occurs, its mean is used. This term directly supervises the aggregate real-candidate probability alongside the existing unmatched supervision.

3.4 TRANSFERRING CONDITIONAL LANDMARK IDENTITY

For a query in I_+, we transfer reference preferences among all retained real candidates by excluding the unmatched outcome and renormalizing over J:

$$\mathbf{q}_i = \mathrm{stopgrad}[\mathcal{N}_\epsilon(\mathbf{p}_i|_J)], \qquad \mathbf{r}_i = \mathcal{N}_\epsilon(\mathbf{p}'_i|_J), \qquad i \in I_+, \tag{6}$$

where $|_J$ restricts a vector to retained landmarks and $\mathrm{stopgrad}$ stops its gradient. The target uses the full input; the subset prediction is recomputed by the same matcher. Both distributions include every retained candidate, not only ground-truth matches.

We minimize the safeguarded forward KL divergence

$$\mathcal{L}_{\mathrm{id}} = \frac{1}{|I_+|} \sum_{i \in I_+} \sum_{j \in J} q_i(j) \log \frac{\max\{q_i(j), \epsilon\}}{\max\{r_i(j), \epsilon\}}, \tag{7}$$

and set it to zero when I_+ is empty. The floor is applied to both normalizations and separately to the logarithm, leaving its multiplier $q_i(j)$ unchanged. With inactive logarithm floors, Eq. (7) is ordinary KL. The pair loss is

$$\mathcal{L}_{\mathrm{pair}} = \mathcal{L}_{\mathrm{avail}} + \mathcal{L}_{\mathrm{id}}. \tag{8}$$

Availability uses a binary geometric target, whereas identity uses a soft prediction over surviving landmarks. Consequently, queries that lose all valid matches contribute only to availability. For those that remain matchable, identity transfer does not depend on whether the predicted availability exceeds a threshold. This uses the output partition in Eq. (3), but replaces its predicted availability with geometric supervision.

Both branches retain direct geometric supervision. Stopping gradients blocks updates through the reference target, but does not alter its predicted probabilities. The pair weight increases gradually during early training.

3.5 CONDITIONING AND ADAPTIVE GROUP SUPERVISION

FiLM (Perez et al., 2018) conditions a network through feature-wise scaling and shifting. We use a bounded input-level form, conditioned on the sampled perturbation group g. The reference branch processes its original input; before the shared matcher, the perturbed branch transforms each coordinate or color vector z:

$$h_g(z) = z \odot [1 + \alpha_h \tanh(s_g)] + \alpha_h \tanh(b_g), \tag{9}$$

where $\odot$ denotes elementwise multiplication, α_h bounds the modulation amplitude, and s_g, b_g are learned vectors with the dimension of the corresponding input. A group-indexed lookup provides separate scale and bias vectors for query coordinates, landmark bearings, query RGB, and landmark RGB. Zero initialization makes h_g the identity at the start of training. Modulation is applied before feature encoding; geometric labels and pose targets remain in the original scene coordinates. Its parameters receive gradients only through the perturbed path and are unused at inference.

We apply group DRO (Sagawa et al., 2020) to the four perturbation groups, increasing the weight of groups with higher supervised loss. Let $\widehat{L}_g^t$ be the mean supervised loss of the examples in group g at minibatch t, and M_g^t indicate whether that group is present. The mean of an absent group is set to zero. For iterations $t = 1, 2, \ldots$, starting from uniform persistent weights u_g^0, we update

$$\widetilde{u}_g^t = u_g^{t-1} \exp\Big(\eta M_g^t \min\{\mathrm{stopgrad}(\widehat{L}_g^t), L_{\mathrm{cap}}\}\Big), \qquad u_g^t = \frac{\widetilde{u}_g^t}{\sum_h \widetilde{u}_h^t}, \tag{10}$$

$$w_g^t = \frac{M_g^t u_g^t}{\sum_h M_h^t u_h^t}, \qquad \mathcal{L}_{\mathrm{shift}}^t = \sum_g w_g^t \widehat{L}_g^t, \tag{11}$$

where $\eta > 0$ is the group update step, L_{cap} caps the detached loss in the exponential, and the sums over h range over all four groups. The denominators are floored at ϵ. Persistent weights favor higher-loss groups; minibatch weights normalize over present groups. The cap affects only the detached weight update, which precedes formation of the loss for the same minibatch. Conditioning and group weighting are fixed across loss ablations; the deletion-pair loss is averaged separately.

For reference minibatch $\mathcal{B}$ and its valid strict-deletion pairs $\mathcal{B}_{\text{del}}$, the full objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}}^{\text{ref}}(\mathcal{B}) + \lambda_{\text{shift}}\mathcal{L}_{\text{shift}} + \frac{\lambda_{\text{pair}}(\tau)}{|\mathcal{B}_{\text{del}}|} \sum_{b \in \mathcal{B}_{\text{del}}} \mathcal{L}_{\text{pair}}^{(b)}, \tag{12}$$

where $\mathcal{B}_{\text{del}}$ is the subset of the minibatch with known, nonempty strict deletions, λ_{shift} weights perturbed supervision, and $\lambda_{\text{pair}}(\tau)$ weights the additional pair loss at epoch $\tau \in \{1, 2, \ldots\}$. The latter increases linearly to its maximum during warm-up. The final term is zero if no valid deletion pair occurs. The index b denotes a deletion pair. Its class-averaged availability loss and positive-query identity loss are summed, then averaged over pairs without group reweighting.

Inference. The trained matcher receives the original coordinates and colors for each retrieved view. It estimates correspondences, which are combined for camera-pose estimation using PnP-RANSAC (Fischler & Bolles, 1981). Inference uses the shared matcher alone; perturbation sampling, input conditioning, group weights, and the extra losses introduce no additional inference operations.

4 Experiments

4.1 Experimental setup

Data and pose estimation. We use the processed MegaDepth (Li & Snavely, 2018) split of 99 training, 16 validation, and 49 test scenes. All MegaDepth tests use SIFT (Lowe, 2004); the main target-scene comparison uses SIFT on Cambridge and SuperPoint (DeTone et al., 2018) on 7Scenes. CandidateShift combines per-view correspondences before PnP-RANSAC. Testing uses $10 \sim 1024$ points per view and outlier range $[0, 1]$. The classifier threshold is 0.5; RANSAC uses normalized threshold 0.001, 1000 iterations, and confidence 0.99.

Evaluation metrics. Following GoMatch (Zhou et al., 2022), reprojection error is the mean pixel distance between projections of ground-truth inlier landmarks under the estimated and true poses. Reprojection AUC integrates the recall curve up to 1, 5, or 10 pixels and divides by that threshold, with failures in the denominator. Pose-error quantiles use valid estimates only, with linear interpolation at 25/50/75%; rotation and translation retain the evaluator units of degrees and meters. Finite high-error poses are not classified as failures.

Implementation and training details. CandidateShift and its loss variants use random initialization, 50 epochs, Adam (Kingma & Ba, 2015), batch size 16, and FP32. A cosine schedule reduces the learning rate from 5×10^{-4} to 10^{-6}; model selection minimizes unperturbed validation loss. The matcher uses self–cross–self attention and four global context nodes. Training uses $100 \sim 1024$ points and base outlier setting 0.5; validation uses $10 \sim 1024$ points and range $[0, 1]$. Geometric labels use normalized threshold 0.001; the classification loss uses its separate threshold 0.01.

The perturbation constants are $k_r = 3$, $\alpha_d = 0.65$, $\alpha_o = 0.5$, $m_{\min} = 10$, $m_{+,\min} = 4$, and $L_c = 16$. FiLM uses $\alpha_h = 0.1$; group DRO uses $\eta = 0.05$ and $L_{\text{cap}} = 20$, with initially uniform group weights. We set $\epsilon = 10^{-8}$, $\delta = 10^{-7}$, $\lambda_{\text{shift}} = 0.5$, and $\lambda_{\text{pair}}(\tau) = 0.25 \min\{1, \tau/5\}$ for one-based epoch τ. These settings are shared by the component ablations.

Compared methods. Table 1 compares GoMatch, DGC-GNN, A2-GNN, SAG-GNN, GeoMix, and CandidateShift. SAG-GNN additionally uses semantic attributes. GeoMix (SIFT) denotes SIFT-only training; GeoMix denotes multi-detector training with SIFT, SuperPoint, and DISK (Tyszkiewicz et al., 2020). CandidateShift and CandidateShift (Mix) distinguish the same two settings. The supervised-perturbation reference shares our matcher, correspondence labels, FiLM conditioning, and group weighting, but omits $\mathcal{L}_{\text{pair}}$.

4.2 Localization and outlier sensitivity

CandidateShift has the highest AUC among SIFT-only models at all three thresholds and both retrieval depths in Table 1. Compared with GeoMix (SIFT), its advantage extends from median pose errors to the upper quartile, where large rotation and translation errors are reduced. Additional retrieved views improve every method, yet CandidateShift retains its advantage within the SIFT-only group.

Table 1: MegaDepth localization with SIFT queries. SIFT-only/Mix groups are ranked separately in green/blue; darker/lighter cells indicate first/second rank. Pose errors use valid estimates.

	Method	Reproj. AUC (%) @1/5/10px (↑)	Rotation (°) Quantile@25/50/75% (↓)	Translation (m) Quantile@25/50/75% (↓)	Time (ms) (↓)
Top-1	GoMatch (Zhou et al., 2022)	5.88 / 23.26 / 28.83	0.54 / 9.75 / 34.32	0.0509 / 0.9983 / 3.5903	**33.50**
	DGC-GNN (Wang et al., 2024)	10.73 / 38.10 / 44.43	0.14 / 1.30 / 29.30	0.0131 / 0.1281 / 3.0712	112.13
	A2-GNN (Zhang et al., 2025)	12.43 / 41.46 / 47.61	0.12 / 0.91 / 26.71	0.0114 / 0.0880 / 2.8771	52.15
	SAG-GNN (Zhang et al., 2026a)	15.93 / 52.41 / 59.71	0.09 / 0.27 / 11.00	0.0079 / 0.0259 / 1.1439	67.13
	GeoMix (SIFT) (Zhang et al., 2026b)	14.99 / 47.35 / 54.11	0.10 / 0.39 / 20.43	0.0087 / 0.0388 / 2.1545	70.81
	CandidateShift	**17.69 / 54.78 / 61.68**	**0.08 / 0.23 / 8.68**	**0.0071 / 0.0220 / 0.8866**	70.46
	GeoMix (Zhang et al., 2026b)	17.35 / 54.60 / 61.82	**0.08** / 0.24 / 8.09	0.0074 / 0.0234 / 0.8544	**70.56**
	CandidateShift (Mix)	**18.35 / 57.05 / 64.27**	**0.08 / 0.21 / 4.60**	**0.0070 / 0.0201 / 0.4786**	71.63
Top-10	GoMatch (Zhou et al., 2022)	8.64 / 34.98 / 44.12	0.18 / 1.39 / 18.44	0.0171 / 0.1300 / 2.1538	**242.73**
	DGC-GNN (Wang et al., 2024)	16.53 / 54.13 / 62.28	0.07 / 0.21 / 4.28	0.0060 / 0.0194 / 0.4262	797.46
	A2-GNN (Zhang et al., 2025)	16.59 / 52.86 / 60.82	0.06 / 0.22 / 5.88	0.0057 / 0.0207 / 0.6141	369.70
	SAG-GNN (Zhang et al., 2026a)	20.13 / 63.97 / 72.40	**0.05** / 0.13 / 0.66	0.0044 / 0.0118 / 0.0633	485.61
	GeoMix (SIFT) (Zhang et al., 2026b)	18.44 / 57.73 / 65.68	0.06 / 0.16 / 2.37	0.0047 / 0.0148 / 0.2318	498.64
	CandidateShift	**21.59 / 64.90 / 72.75**	**0.05 / 0.11 / 0.56**	**0.0039 / 0.0101 / 0.0543**	497.99
	GeoMix (Zhang et al., 2026b)	21.57 / 65.72 / 73.78	**0.05** / 0.11 / 0.51	0.0040 / 0.0104 / 0.0514	511.27
	CandidateShift (Mix)	**22.47 / 67.45 / 75.41**	**0.05 / 0.10 / 0.40**	**0.0038 / 0.0097 / 0.0399**	**506.85**

Multi-detector training further improves CandidateShift. Its Mix variant attains the highest AUC in the full comparison and lower upper-quartile pose errors than GeoMix at both retrieval depths. At the strict one-pixel threshold, the AUC advantage over GeoMix is 1.00 percentage point for top-1 and 0.90 for top-10. The benefit of candidate-set supervision is therefore present within both training settings.

We vary the maximum outlier rate r from zero to one in steps of 0.1, merging candidates before matching for this experiment. The realized fraction varies by image; all 12,399 records, including input and pose failures, enter the AUC denominator. The curves in Figure 3 show that GeoMix (SIFT) is competitive at low contamination but declines more sharply as contamination increases. CandidateShift (Mix) leads all three metrics for $r \geq 0.6$; the SIFT-only variant leads its training group for $r \geq 0.7$. The advantage is concentrated in outlier-heavy inputs rather than being uniform across the range.

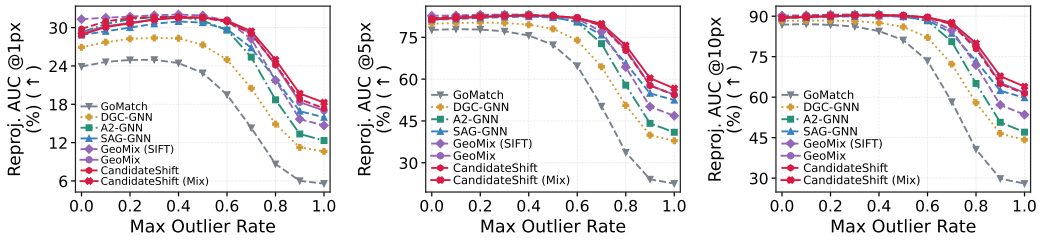

Figure 3: **Localization under controlled outlier sampling.** All models use SIFT queries; the horizontal axis bounds the sampled outlier rate, not the realized fraction per image.

4.3 Contribution of the two training terms

Table 2 separates supervised perturbations from the added pair objective. The perturbation-only reference includes FiLM and group DRO: it raises top-1 AUC@1px over GeoMix (SIFT) by 1.71 percentage points, and the pair objective adds 0.99. Holding these components fixed isolates the additional losses, not the individual effects of perturbations, FiLM, or group weighting. Identity supervision gives the larger gain, particularly for upper-quartile pose errors.

The joint objective attains the lowest reported translation quantiles and the highest AUC@1px at both retrieval depths. It also produces fewer invalid top-1 poses than identity-only training (35 versus 44). Identity-only training gives higher top-1 AUC@5/10px, lower upper-quartile rotation error, and higher matching precision and recall; both variants tie at top-10 AUC@10px. The joint objective also leads at top-10 AUC@5px. Thus correspondence statistics and pose-estimation outcomes favor different configurations.

Table 2: MegaDepth loss ablations. Supervised perturbations include FiLM and group DRO; checkmarks indicate the active training components.

	Variant	Supervised perturbations	$\mathcal{L}_{\text{avail}}$	$\mathcal{L}_{\text{id}}$	Reproj. AUC (%) @1/5/10px (↑)	Rotation (°) Quantile@25/50/75% (↓)	Translation (m) Quantile@25/50/75% (↓)
Top-1	f_θ = GeoMix (SIFT) (Zhang et al., 2026b)				14.99 / 47.35 / 54.11	0.10 / 0.39 / 20.43	0.0087 / 0.0388 / 2.1545
	Perturbations only	✓			16.70 / 52.58 / 59.52	**0.08** / 0.26 / 12.58	0.0076 / 0.0255 / 1.3260
	+ Availability	✓	✓		17.43 / 54.31 / 61.20	**0.08** / **0.23** / 10.32	0.0073 / 0.0221 / 1.0727
	+ Identity	✓		✓	17.64 / **54.90** / **61.86**	**0.08** / **0.23** / **8.51**	0.0072 / 0.0223 / 0.8886
	CandidateShift	✓	✓	✓	**17.69** / 54.78 / 61.68	**0.08** / **0.23** / 8.68	**0.0071** / **0.0220** / **0.8866**
Top-10	f_θ = GeoMix (SIFT) (Zhang et al., 2026b)				18.44 / 57.73 / 65.68	0.06 / 0.16 / 2.37	0.0047 / 0.0148 / 0.2318
	Perturbations only	✓			20.61 / 62.66 / 70.51	**0.05** / 0.12 / 0.88	0.0042 / 0.0113 / 0.0829
	+ Availability	✓	✓		21.22 / 64.17 / 72.05	**0.05** / **0.11** / 0.65	0.0040 / 0.0107 / 0.0639
	+ Identity	✓		✓	21.49 / 64.80 / **72.75**	**0.05** / **0.11** / 0.57	**0.0039** / 0.0103 / 0.0553
	CandidateShift	✓	✓	✓	**21.59** / **64.90** / **72.75**	**0.05** / **0.11** / **0.56**	**0.0039** / **0.0101** / **0.0543**

4.4 Behavior under candidate deletion

Figure 4 compares four loss variants on 512 fixed MegaDepth SIFT validation queries at nested retention fractions $0.1, 0.2, 0.3, 0.4, \ldots, 0.9$, with identical subsets and masks. At retention 0.1, nine queries fail because too few candidates remain for neighborhood convolution, and 499 successful queries contribute to identity metrics; other fractions complete all 512 queries. Both metrics use keypoints with surviving ground-truth matches, averaging first within each image and then across images with eligible keypoints. Identity accuracy tests the most probable real candidate without an availability gate. Conditional KL compares the restricted, renormalized full-set prediction with the recomputed reduced-set distribution, excluding unmatched entries and using the full-to-reduced direction. CandidateShift improves both measures over supervised perturbations at every fraction. Identity-only training has higher accuracy at four fractions. These metrics assess identity selection conditional on a surviving match; availability supervision instead covers all keypoints, including those with no remaining valid correspondence.

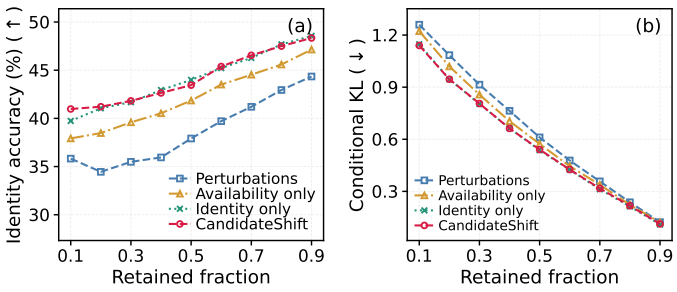

Figure 4: Deletion diagnostics.

4.5 Generalization to unseen scenes

We evaluate without fine-tuning on four Cambridge Landmarks (Kendall et al., 2015) scenes (1,158 queries) and all seven scenes of 7Scenes (Shotton et al., 2013) (17,000 queries), with top-10 retrieval. Table 3 uses SIFT query keypoints on Cambridge and SuperPoint on 7Scenes. Pose medians use valid estimates. SAG-GNN additionally uses locally reproduced semantic attributes; CandidateShift uses coordinates and RGB. Appendix D reports the complementary detector settings.

CandidateShift lowers translation medians in every scene relative to GeoMix (SIFT). Under Mix training, it lowers both Cambridge/SIFT averages relative to GeoMix. SAG-GNN has lower Cambridge averages than SIFT-trained CandidateShift. On 7Scenes/SuperPoint, CandidateShift (Mix)

improves several scenes but has higher averages than GeoMix, with the largest increase on Stairs. The loss ablations show a detector-dependent ordering: joint supervision gives the lowest 7Scenes/SIFT averages, while supervised perturbations give the lowest SuperPoint averages.

Table 3: Zero-shot localization with top-10 retrieval. Cells: median translation/rotation (cm/°, ↓); Avg.: mean of unrounded scene medians.

Method	Cambridge / SIFT (cm/°) (↓)					7Scenes / SuperPoint (cm/°) (↓)							
	King's	Hospital	Shop	St Mary's	Avg.	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
GoMatch (Zhou et al., 2022)	107/2.85	419/11.38	544/14.51	1347/27.44	604/14.05	4/1.58	12/3.78	5/3.47	7/1.81	28/5.82	15/3.14	57/13.45	18/4.72
DGC-GNN (Wang et al., 2024)	80/2.21	250/6.55	308/20.69	804/20.22	360/12.42	**3**/1.47	6/2.12	4/2.63	6/1.62	11/2.47	9/2.25	43/10.34	12/3.27
A2-GNN (Zhang et al., 2025)	108/3.02	377/10.16	491/12.98	1256/29.25	558/13.85	**3**/1.40	6/1.89	3/2.39	**5**/1.45	9/2.31	8/2.05	62/13.74	14/3.60
SAG-GNN (Zhang et al., 2026a)	51/1.29	**100**/3.73	**51/3.46**	**312/10.35**	**128/4.71**	**3**/1.43	**5**/1.70	3/2.41	6/1.51	**8**/2.06	8/2.07	52/12.15	12/3.33
GeoMix (SIFT) (Zhang et al., 2026b)	73/1.92	298/8.18	197/10.58	944/24.49	378/11.29	**3/1.35**	**5**/1.74	3/2.24	**5**/1.43	**8**/2.06	**7**/1.98	51/10.97	12/3.11
CandidateShift	**48/1.09**	118/**3.69**	61/4.04	412/11.47	160/5.07	**3**/1.37	**5/1.67**	**2/2.16**	**5/1.41**	**8/2.04**	**7/1.91**	**39/7.91**	**10/2.64**
GeoMix (Zhang et al., 2026b)	48/**1.21**	150/4.92	63/4.42	546/14.09	202/6.16	**3/1.34**	**4/1.58**	**2**/2.19	**5**/1.42	**8**/1.96	**7/1.91**	**36/7.44**	**9/2.55**
CandidateShift (Mix)	**46**/1.31	**130/3.15**	**32/2.44**	**467/12.50**	**169/4.85**	**3**/1.35	**4**/1.60	**2/2.17**	**5/1.41**	**8/1.90**	**7/1.91**	43/8.67	10/2.72

4.6 Transfer to a second matching architecture

We instantiate f_θ with A2-GNN (Zhang et al., 2025), replacing the GeoMix representation with angular and annular graph aggregation. The same availability labels and conditional identity targets supervise A2-GNN assignments without changing the inference architecture. The augmented model improves all six AUC values in Table 4. Against a separately trained A2-GNN control, AUC@10px also increases by 3.96 and 6.39 percentage points for top-1 and top-10. The gains across both tested geometric representations support transfer of the objective beyond GeoMix. This comparison addresses architectural transfer on MegaDepth; the cross-scene and cross-detector tests use the GeoMix-based model.

Table 4: A2-GNN transfer. Reproj. AUC (%) @1/5/10px (↑).

Method	Top-1	Top-10
A2-GNN (Zhang et al., 2025)	12.43/41.46/47.61	16.59/52.86/60.82
+ CandidateShift	**13.64/45.40/51.94**	**19.19/59.46/67.08**

4.7 Inference cost and remaining limitations

Table 1 reports CUDA-event matching times from two rounds on 128 fixed queries, with batch size one, 16 warm-ups, and reversed model order. PnP, input construction, retrieval, and semantic extraction are excluded. CandidateShift and GeoMix have similar inference costs. Paired training increases update time, while the deployed matcher remains unchanged; Section C.2 quantifies this training cost. Results reflect one training run per configuration. The ablations do not separate availability supervision from unmatched-loss reweighting. Stopped-gradient identity targets can retain errors in the full-set prediction.

5 Conclusion

CandidateShift trains descriptor-free 2D-3D matchers to account for changes in candidate availability. Geometric labels determine whether a valid correspondence survives, while restricted full-set predictions provide conditional identity targets for the remaining matchable keypoints. The two terms distinguish match existence from landmark identity. Localization and controlled-deletion experiments show improved pose accuracy, landmark selection, and cross-set prediction consistency relative to supervised perturbations. Cross-scene gains vary with the detector; improvements on a second matching architecture require no additional inference modules. Ablations identify conditional identity supervision as the stronger contributor and reveal metric-dependent trade-offs when availability supervision is added. The results support treating the candidate set as part of the prediction problem, rather than assuming that every input perturbation preserves the correct matching outcome.

REPRODUCIBILITY STATEMENT

Sections 3, 4.1, and 4.4 specify the mathematical objective, perturbation groups, numerical safeguards, training constants, diagnostic deletion masks, and evaluation protocols. Complete implementation hyperparameters, candidate merging details, and computational profiling environments are documented in Appendices A and C. All evaluations rely on public benchmark datasets. The complete source code, configuration files, and pretrained checkpoints will be made publicly available upon publication.

ETHICS STATEMENT

Visual localization is a foundational technology for autonomous navigation, robotics, and augmented reality. While descriptor-free 2D-3D matching eliminates high-dimensional appearance descriptors that could potentially be inverted into raw images, geometric coordinates and estimated camera poses still inevitably disclose spatial location. This study investigates geometric matching accuracy and candidate robustness rather than formal privacy guarantees. All experiments are conducted strictly on established public research benchmarks (MegaDepth, Cambridge Landmarks, and 7Scenes) and involve no personal or biometric data collection.

AI USE STATEMENT

During the preparation of this manuscript, ChatGPT was used solely for language polishing, grammatical corrections, and stylistic refinement to enhance text clarity, as well as minor LaTeX typesetting adjustments. The authors thoroughly reviewed and edited all generated text. The authors take full responsibility for the paper's originality, scientific claims, methodology, experimental implementation, and references.

REFERENCES

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Arandjelovic_NetVLAD_CNN_Architecture_CVPR_2016_paper.html.

Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15859–15869, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Bai_PointDSC_Robust_Point_Cloud_Registration_Using_Deep_Spatial_Consistency_CVPR_2021_paper.html.

Dylan Campbell, Liu Liu, and Stephen Gould. Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization. In *European Conference on Computer Vision (ECCV)*, pp. 244–261, 2020. URL https://arxiv.org/abs/2007.14628.

Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7653–7662, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Camposeco_Hybrid_Scene_Compression_for_Visual_Localization_CVPR_2019_paper.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018. URL https://arxiv.org/abs/1712.07629.

Mihai Dusmanu, Ondrej Miksik, Johannes L. Schönberger, and Marc Pollefeys. Cross-descriptor visual localization and mapping. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2012.01377.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015. URL https://arxiv.org/abs/1505.07427.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2041–2050, 2018. URL https://www.cs.cornell.edu/projects/megadepth/.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17627–17638, 2023. URL https://arxiv.org/abs/2306.13643.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://arxiv.org/abs/1709.07871.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.08731.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12716–12725, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Sarlin_From_Coarse_to_Fine_Robust_Hierarchical_Localization_at_Large_Scale_CVPR_2019_paper.html.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020. URL https://arxiv.org/abs/1911.11763.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2937, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Shotton_Scene_Coordinate_Regression_2013_CVPR_paper.html.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Sun_LoFTR_Detector-Free_Local_Feature_Matching_With_Transformers_CVPR_2021_paper.html.

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. URL https://arxiv.org/abs/1703.01780.

Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/a42a596fc71e17828440030074d15e74-Abstract.html.

Shuzhe Wang, Juho Kannala, and Daniel Barath. DGC-GNN: Leveraging geometry and color cues for visual descriptor-free 2D-3D matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20881–20891, 2024. doi: 10.1109/CVPR52733.2024.01973. URL https://arxiv.org/abs/2306.12547.

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2666–2674, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Yi_Learning_to_Find_CVPR_2018_paper.html.

Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5845–5854, 2019. URL https://arxiv.org/abs/1908.04964.

Shihua Zhang, Tianhao Xu, Zizhuo Li, Qing Ma, and Jiayi Ma. SAG-GNN: Semantic-aware guided GNN for descriptor-free 2D-3D matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31410–31420, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_SAG-GNN_Semantic-Aware_Guided_GNN_for_Descriptor-Free_2D-3D_Matching_CVPR_2026_paper.html.

Yejun Zhang, Shuzhe Wang, and Juho Kannala. A2-GNN: Angle-annular GNN for visual descriptor-free camera relocalization. In *International Conference on 3D Vision (3DV)*, pp. 357–368, 2025. doi: 10.1109/3DV66043.2025.00038. URL https://arxiv.org/abs/2502.20036.

Yejun Zhang, Xinjue Wang, Zihan Wang, Esa Rahtu, and Juho Kannala. GeoMix: Descriptor-free visual localization via global context and multi-detector training. In *European Conference on Computer Vision (ECCV)*, 2026b. URL https://arxiv.org/abs/2607.02486. arXiv:2607.02486.

Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6464–6473, 2021. URL https://arxiv.org/abs/2101.00591.

Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *European Conference on Computer Vision (ECCV)*, pp. 407–425, 2022. doi: 10.1007/978-3-031-20080-9_24. URL https://arxiv.org/abs/2203.12979.

A Implementation details

The FiLM lookup table has five rows of twenty parameters. Four rows correspond to the perturbation groups; the zero-indexed row is reserved. Each active row contains separate scale and bias vectors for two-dimensional query bearings, two-dimensional landmark bearings, query RGB, and landmark RGB. Zero initialization gives the identity transformation. The reference branch and all evaluation calls bypass this module.

Color corruption first sets one channel to zero and then quantizes all channels. With sixteen levels on $[-1, 1]$, zero is not itself a quantization level; in the supplied implementation the reset channel becomes $1/15$. This operation changes the input cue while keeping the geometric correspondence labels fixed.

A retrieval list without an alternative view advances the sampler to the next example. Minimum-size constraints can leave density or outlier subsets unchanged. These pairs retain both supervised losses but are excluded from $\mathcal{B}_{\mathrm{del}}$ in Eq. (12).

Evaluation records and matching order. In Table 1, historical top-1 evaluations contain 12,388 records for GoMatch, DGC-GNN, A2-GNN, and SAG-GNN, and 12,399 for GeoMix and CandidateShift. SAG-GNN merges candidates before matching; the other rows merge correspondences after matching. Component ablations use aligned records: top-1 has 12,399 records for 12,388 unique queries because invalid inputs are replaced, and top-10 has 12,399 unique queries. The controlled-outlier experiment instead uses candidate merging before matching for every model, with 12,399 records in the AUC denominator.

B Probability interpretation of candidate deletion

Equation (3) collects all removed outcomes into the unmatched state. Its entries sum to one because the retained landmarks, removed landmarks, and original unmatched state partition the full outcome space. If $K \subset J$ is retained later, $T_{J\to K}$ maps distributions on $J \cup \{\perp\}$ to $K \cup \{\perp\}$ by the same rule. Each removed landmark contributes its probability once, so $T_{J\to K} T_J \mathbf{p}_i = T_K \mathbf{p}_i$.

For two distributions $\mathbf{v}, \mathbf{w}$ over $J \cup \{\perp\}$, define $\rho_v = \sum_{j\in J} v(j)$ and $\rho_w = \sum_{j\in J} w(j)$. When both lie strictly between zero and one, write $\bar{v}(j) = v(j)/\rho_v$ and $\bar{w}(j) = w(j)/\rho_w$ on J. With the usual zero-probability conventions, splitting ordinary KL into real and unmatched outcomes gives

$$D_{\mathrm{KL}}(\mathbf{v}\|\mathbf{w}) = \sum_{j\in J} \rho_v \bar{v}(j) \log \frac{\rho_v \bar{v}(j)}{\rho_w \bar{w}(j)} + (1-\rho_v) \log \frac{1-\rho_v}{1-\rho_w} \tag{13}$$

$$= D_{\mathrm{KL}}(\mathrm{Bern}(\rho_v)\|\,\mathrm{Bern}(\rho_w)) + \rho_v D_{\mathrm{KL}}(\bar{\mathbf{v}}\|\bar{\mathbf{w}}), \tag{14}$$

where $\mathrm{Bern}(\rho)$ is the Bernoulli distribution with success probability ρ. Taking $\mathbf{v} = T_J \mathbf{p}_i$ and $\mathbf{w} = \mathbf{p}'_i$ illustrates how a direct distributional loss would weight identity by the retained mass predicted on the full set. The training objective instead uses class-balanced geometric availability labels and equal identity weight on I_+; it is not this joint KL. Adding removed probability to $\perp$ therefore has no effect on the conditional identity target in Eq. (6).

C Additional computational measurements

C.1 Measured latency and backbone cost

Table 5 reports deployed parameters, matching-plus-PnP latency, and raw-matcher GFLOPs. Timing uses two rounds on the same 128 queries, reversing model order in the second round, with batch size one and 16 warm-up iterations. The pipeline uses synchronized wall-clock timing and excludes retrieval, input construction, and external semantic extraction.

Operation counts cover the matching backbone before the outlier classifier, using FP32, batch size one, and 1024 query points and landmarks in one retrieved pair. A real input is repeated or truncated to this size for profiling only. Fused multiply–add counts as two operations, scalar arithmetic and exponentials/logarithms as one, softmax as approximately $5N$, and reductions as N for N

elements. Uncovered neighborhood, clustering, and indexing operations are excluded; this is neither a full-pipeline nor a top-10-query count.

CandidateShift and GeoMix have the same deployed parameter count and fixed-shape backbone cost, with similar measured latency. Operation counts alone do not determine runtime: SAG-GNN is faster in the measured pipeline. The number of predicted correspondences also affects classifier and PnP cost.

Table 5: Inference cost: deployed parameters (M), matching-plus-PnP latency (ms/query), and raw-matcher GFLOPs.

Method	Params (M) (↓)	Top-1			Top-10			GFLOPs (↓)
		Mean (↓)	Median (↓)	P95 (↓)	Mean (↓)	Median (↓)	P95 (↓)	
GoMatch (Zhou et al., 2022)	1.32	60.94	61.85	68.31	304.52	334.57	464.24	13.31
DGC-GNN (Wang et al., 2024)	5.65	136.05	137.47	149.03	852.56	908.52	1262.32	51.15
A2-GNN (Zhang et al., 2025)	2.67	78.91	81.88	87.59	440.10	492.83	647.58	21.88
SAG-GNN (Zhang et al., 2026a)	4.18	91.85	95.11	102.11	548.25	618.95	789.43	35.34
GeoMix (SIFT) (Zhang et al., 2026b)	3.47	98.50	102.14	109.88	576.93	644.87	845.58	22.76
CandidateShift	3.47	95.30	100.89	108.48	569.33	637.88	822.73	22.76
GeoMix (Zhang et al., 2026b)	3.47	94.24	99.04	107.23	579.64	651.01	842.60	22.76
CandidateShift (Mix)	3.47	95.12	100.24	108.41	574.77	645.50	834.45	22.76

C.2 Training-update cost

We measure 100 updates per configuration across three identical GPUs (34/33/33 updates), with ten warm-ups on each device. Initial matcher parameters and ordered batches are shared, using SIFT, batch size 16, FP32, and Adam. GeoMix receives only reference inputs; paired variants share reference and perturbed inputs. Synchronized timing covers gradient clearing, forward and loss computation, backpropagation, and Adam updates, excluding loading, host-to-device transfer, validation, logging, and saving.

Table 6 reports mean update time and sample standard deviation. Paired training is slower than reference-only training; the three deletion-aware variants have similar means. Perturbation-only timing is more variable. Single-loss implementations compute both auxiliary terms before selecting the active term, so these measurements compare training paths rather than isolated loss costs.

Table 6: Training-update time (ms/update), measured over 100 updates per configuration.

Variant	Supervised perturbations	$\mathcal{L}_{\text{avail}}$	$\mathcal{L}_{\text{id}}$	Mean (ms) (↓)	Standard deviation (ms)
f_θ = GeoMix (SIFT) (Zhang et al., 2026b)				1668.51	59.00
Perturbations only	✓			3505.30	165.48
+ Availability	✓	✓		3414.02	54.70
+ Identity	✓		✓	3393.30	50.11
CandidateShift	✓	✓	✓	3411.19	61.88

D Complementary detector settings

Table 7 complements the main comparison with Cambridge/SuperPoint and 7Scenes/SIFT. Both use top-10 retrieval, the same pose estimator, and no target-scene fine-tuning.

SIFT-trained CandidateShift lowers both errors relative to GeoMix (SIFT) in every scene. On Cambridge/SuperPoint, Mix-trained GeoMix retains lower averages; CandidateShift improves translation only in Shop. On 7Scenes/SIFT, CandidateShift (Mix) lowers both averages relative to GeoMix, with the largest scene-level reduction on Stairs. SIFT-trained CandidateShift has the lowest 7Scenes/SIFT averages, showing that detector diversity does not improve transfer uniformly.

Table 7: Complementary-detector localization with top-10 retrieval. Cells: median translation/rotation (cm/°, ↓); Avg.: mean of unrounded scene medians.

Method	Cambridge / SuperPoint (cm/°) (↓)					7Scenes / SIFT (cm/°) (↓)							
	King's	Hospital	Shop	St Mary's	Avg.	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
GoMatch (Zhou et al., 2022)	27/0.67	288/8.84	48/3.83	283/8.96	162/5.58	4/1.61	11/3.51	8/4.94	9/2.19	13/2.89	12/2.64	76/19.66	19/5.35
DGC-GNN (Wang et al., 2024)	17/0.40	85/2.70	12/1.44	63/2.27	44/1.70	**3**/1.40	5/1.78	4/2.98	7/1.65	8/1.96	8/2.07	87/20.25	17/4.58
A2-GNN (Zhang et al., 2025)	16/0.40	48/1.62	14/1.79	64/2.01	35/1.46	**3**/1.37	5/1.76	4/2.60	6/1.56	**7**/1.91	8/2.02	74/19.27	15/4.36
SAG-GNN (Zhang et al., 2026a)	**14**/0.38	**33/0.98**	**7/0.82**	**31**/1.11	**21/0.82**	**3**/1.38	5/1.81	**3**/2.56	6/1.62	**7**/1.96	8/2.06	85/19.17	17/4.37
GeoMix (SIFT) (Zhang et al., 2026b)	18/0.45	64/1.98	14/1.56	49/1.75	36/1.44	**3**/1.35	5/1.76	**3**/2.37	6/1.50	**7**/1.86	**7**/1.96	60/15.09	13/3.70
CandidateShift	15/**0.34**	35/1.04	10/0.84	35/**1.10**	24/0.83	**3/1.33**	**4/1.55**	**3/2.29**	**5/1.46**	**7/1.77**	**7/1.87**	**39/9.44**	**10/2.82**
GeoMix (Zhang et al., 2026b)	**14/0.36**	**32/0.77**	8/**0.67**	**20/0.78**	**18/0.65**	**3**/1.33	**4/1.47**	**3/2.32**	**5/1.46**	**7/1.75**	**7**/1.89	48/11.99	11/3.17
CandidateShift (Mix)	15/**0.36**	34/0.94	**7**/0.83	21/0.79	19/0.73	**3/1.32**	**4**/1.51	**3**/2.36	**5/1.46**	**7**/1.77	**7/1.88**	**45/10.70**	**10/3.00**

D.1 Cross-scene component ablations

Tables 8 and 9 compare four SIFT-trained loss configurations with top-10 retrieval and no fine-tuning. Medians use valid poses; scene averages are computed before rounding. On Cambridge, identity-only training gives the lowest average translation and rotation errors with either query detector. The joint objective lowers SIFT translation error at St Mary's, but its SuperPoint average translation error exceeds that of supervised perturbations.

Table 8: Cambridge loss ablations with SIFT training and top-10 retrieval. Median translation/rotation (cm/°, ↓); Avg.: four-scene mean.

Query detector	Variant	Supervised perturbations	$\mathcal{L}_{\text{avail}}$	$\mathcal{L}_{\text{id}}$	King's	Hospital	Shop	St Mary's	Avg.
SIFT	Perturbations only	✓			52/1.34	166/4.52	117/6.44	571/13.95	226/6.56
	+ Availability	✓	✓		52/1.29	168/5.22	81/5.01	569/11.02	217/5.64
	+ Identity	✓		✓	**46/1.09**	**91/3.65**	**53/2.62**	430/**9.89**	**155/4.31**
	CandidateShift	✓	✓	✓	48/**1.09**	118/3.69	61/4.04	**412**/11.47	160/5.07
SuperPoint	Perturbations only	✓			**15**/0.37	**30/0.94**	12/1.02	**29**/1.02	**21**/0.84
	+ Availability	✓	✓		**15**/0.36	35/1.25	**10**/1.04	30/1.26	23/0.98
	+ Identity	✓		✓	**15**/0.38	32/1.03	**10/0.84**	30/**1.01**	**21/0.82**
	CandidateShift	✓	✓	✓	**15/0.34**	35/1.04	**10/0.84**	35/1.10	24/0.83

7Scenes. Table 9 covers all seven scenes with both query detectors. With SIFT, the joint objective has the lowest average translation and rotation errors; its largest reduction relative to supervised perturbations occurs on Stairs. Identity-only training remains stronger on Fire and Pumpkin. With SuperPoint, supervised perturbations have the lowest averages, followed by identity-only training. The joint objective improves some scene-level errors but does not improve the cross-detector average.

Table 9: 7Scenes loss ablations with SIFT training and top-10 retrieval. Median translation/rotation (cm/°, ↓); Avg.: mean of all seven scene medians.

Query detector	Variant	Supervised perturbations	$\mathcal{L}_{\text{avail}}$	$\mathcal{L}_{\text{id}}$	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
SIFT	Perturbations only	✓			**3/1.32**	**4**/1.62	**3**/2.35	6/1.49	**7**/1.79	**7**/1.92	49/12.16	11/3.24
	+ Availability	✓	✓		**3**/1.36	**4**/1.58	**3**/2.34	6/1.47	**7**/1.78	**7**/1.92	45/11.46	11/3.13
	+ Identity	✓		✓	**3**/1.34	**4/1.49**	**3**/2.30	**5/1.46**	**7/1.75**	**7**/1.89	45/10.37	**10**/2.94
	CandidateShift	✓	✓	✓	**3**/1.33	**4**/1.55	**3/2.29**	**5/1.46**	**7**/1.77	**7/1.87**	**39/9.44**	**10/2.82**
SuperPoint	Perturbations only	✓			**3**/1.36	**5/1.63**	**2**/2.19	**5/1.41**	**8**/1.99	**7/1.91**	**37/6.99**	**10/2.50**
	+ Availability	✓	✓		**3**/1.36	**5**/1.75	**2/2.16**	**5**/1.43	**8**/1.98	**7**/1.93	42/8.08	**10**/2.67
	+ Identity	✓		✓	**3/1.35**	**5**/1.64	**2**/2.19	**5/1.41**	**8/1.96**	**7**/1.92	38/7.38	**10**/2.55
	CandidateShift	✓	✓	✓	**3**/1.37	**5**/1.67	**2/2.16**	**5/1.41**	**8**/2.04	**7/1.91**	39/7.91	**10**/2.64

E ADDITIONAL MATCHING DIAGNOSTICS

E.1 MATCHING QUALITY AND POSE FAILURES

Table 10 reports query-macro precision, recall, and F1 on aligned records. Correct denotes the mean number of correct correspondences. A pose failure has a negative or nonfinite recorded rotation or translation error; a finite inaccurate pose is not a failure under this definition. Identity-only training leads the correspondence metrics, while joint supervision yields fewer invalid top-1 poses. Both have zero top-10 failures in these records.

Table 10: Query-macro matching quality (%) and pose failures.

	Variant	Supervised perturbations	$\mathcal{L}_{\text{avail}}$	$\mathcal{L}_{\text{id}}$	Precision (%) (↑)	Recall (%) (↑)	F1 (%) (↑)	Correct (↑)	Failures (↓)
Top-1	Perturbations only	✓			26.62	31.79	28.45	47.76	45
	+ Availability	✓	✓		27.90	34.27	30.21	50.71	39
	+ Identity	✓		✓	**28.37**	**34.63**	**30.64**	**51.03**	44
	CandidateShift	✓	✓	✓	28.16	34.52	30.47	50.88	**35**
Top-10	Perturbations only	✓			17.77	46.40	25.05	146.64	1
	+ Availability	✓	✓		18.46	48.84	26.13	153.03	1
	+ Identity	✓		✓	**18.87**	**49.16**	**26.61**	**153.66**	**0**
	CandidateShift	✓	✓	✓	18.75	49.10	26.48	153.43	**0**

E.2 CANDIDATE-SET SIZE

Figure 5 reports ΔAUC, the reprojection AUC of CandidateShift minus that of supervised perturbations, in percentage points. Gains are positive at all thresholds in every top-1 stratum and four of the five top-10 strata. Only AUC@5/10px decreases in the smallest top-10 stratum, which contains 14 records. Candidate count therefore does not yield a monotonic ordering of the observed gains.

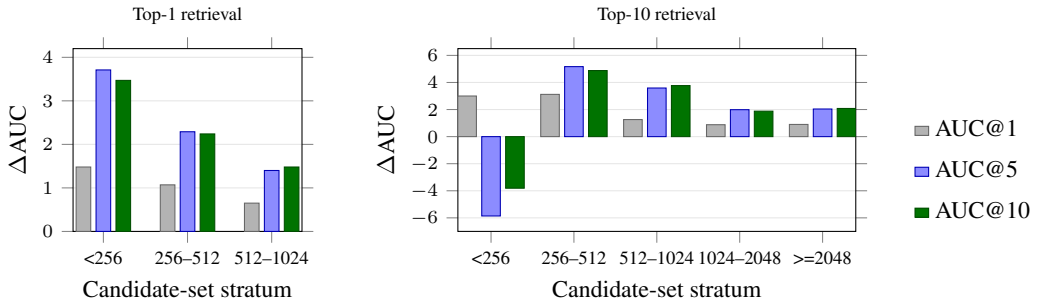

Figure 5: Candidate-set-size analysis. ΔAUC (percentage points) is measured relative to supervised perturbations; positive bars favor CandidateShift.

F QUALITATIVE 2D-3D MATCHING

Figures 6 and 7 show SIFT-only and Mix-trained comparisons, respectively. TP counts correct correspondences; P and F1 are percentages. CandidateShift improves all three measures in the SIFT examples. With Mix training, TP and F1 improve in every pair, while precision improves in nine. The first tower pair illustrates the trade-off: TP rises from 127 to 151, but precision falls from 87.6% to 86.8%. Comparisons are within query pairs because the two figures use different queries.

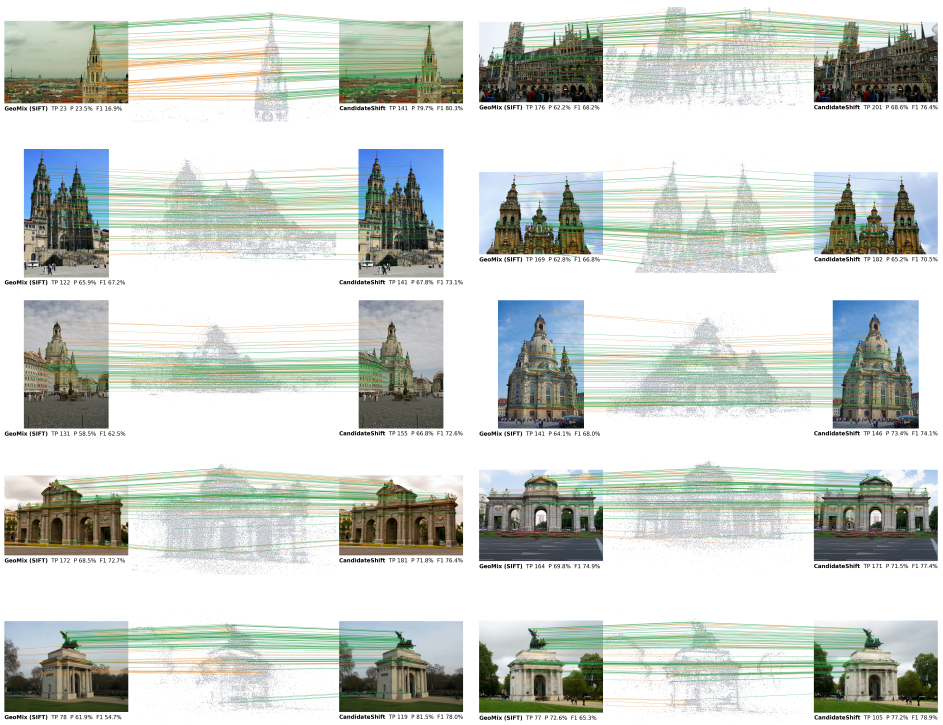

Figure 6: **SIFT-only matching.** GeoMix (SIFT), left; CandidateShift, right.

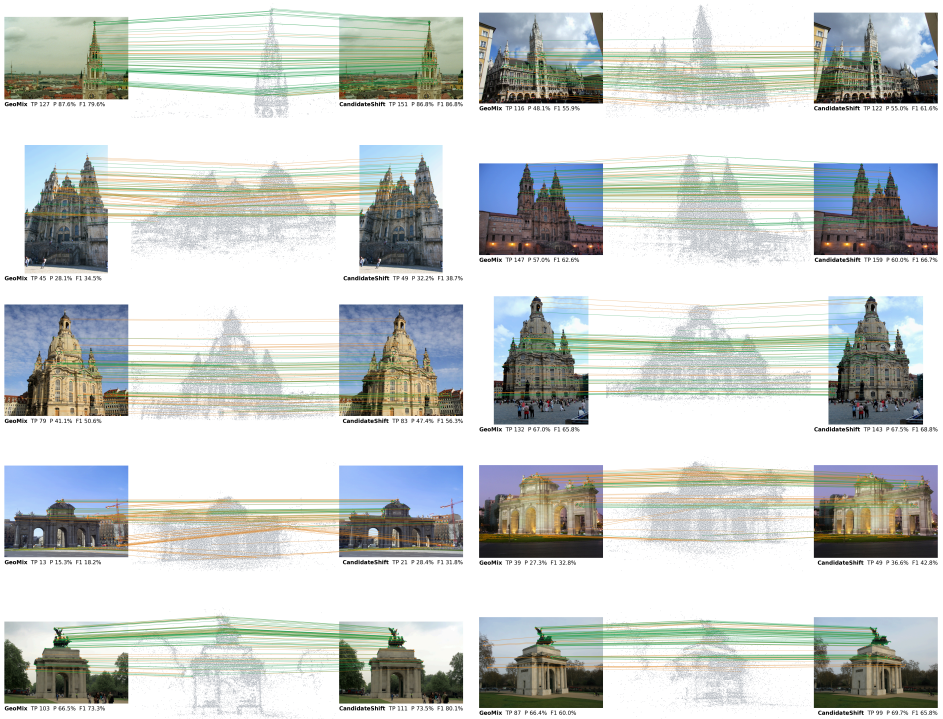

Figure 7: **Mix-trained matching.** GeoMix, left; CandidateShift (Mix), right.