Reliability-Oriented One-Shot Anatomical Landmark Detection under Extreme Annotation Scarcity

Yinbing Tian[1], Xuefei Li[5], Ziyang Wang[6], and Li Guo[1,2,3,4⋆]

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Engineering Research Center of Blockchain and Network Convergence Technology, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] National Engineering Research Center for Mobile Internet Security Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4] Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[5] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[6] School of Computer Science and Digital Technologies, Aston University, Birmingham B4 7ET, United Kingdom

Abstract. Extreme annotation scarcity shifts anatomical landmark detection from fully supervised prediction toward template-based landmark transfer. In this setting, average accuracy is not the only concern: mean radial error can obscure the failure mode that matters most for downstream use, where a small number of severe landmark swaps may invalidate measurements even when the average error remains competitive. This paper studies one-shot hand X-ray landmark detection from a reliability-oriented perspective under a strict protocol that uses one labeled template and no unlabeled target images. Rather than proposing a new full detector, we analyze Topology-Aware Foundation-Model Matching (TAFM), an inference-time safeguard built on top of template-only matching. TAFM estimates an image-specific anatomical prior from coarse correspondences, uses the prior to constrain candidate scoring, and gates local refinement against topology-inconsistent predictions. We define reliability metrics that complement mean radial error, including landmark-level severe-failure counts, image-level severe-case rates, maximum per-image error, and review-trigger statistics. On the 300-image Hand X-ray test split, TAFM achieves state-of-the-art performance among the compared one-shot methods, reducing MRE from 1.655 mm to 1.340 mm compared with the reproduced FM-OSD baseline and decreasing image-level > 20 mm severe failures from 13.33% to 1.67%.

Keywords: One-Shot Landmark Detection · Medical Image Reliability · Severe Failure Analysis · Topology-Guided Matching · Hand X-ray

⋆ Corresponding author: guoli@bupt.edu.cn

1 Introduction

Anatomical landmarks are compact geometric anchors for medical image interpretation. They support diagnosis, growth assessment, registration initialization, surgical planning, and longitudinal measurement [4,3,13,10]. Modern supervised landmark detectors can achieve strong accuracy when abundant annotations are available [10,16,18], but expert landmark annotation is slow, anatomically demanding, and difficult to scale across sites. This annotation barrier motivates one-shot landmark detection, where landmarks are transferred from a template image to a query image [15,11,14,17,9].

Foundation-model descriptors have substantially improved this one-shot setting. In particular, FM-OSD uses frozen foundation-model features with global-to-local matching and reports markedly better hand X-ray localization than earlier template-only approaches [2,1,9]. However, the remaining errors are not merely small localization noise. Because the template and a test image may differ in pose, scale, bone maturity, projection, and repeated finger structures, correspondence-based detectors can still assign a landmark to a visually similar but anatomically wrong location. In hard cases, most landmarks can be reasonable while a small number of severe swaps produce visibly unsafe predictions.

These outliers matter both clinically and statistically. Standard landmark metrics include mean radial error (MRE) and successful detection rate (SDR) [13,10]. However, MRE can be strongly affected by a few very large errors while still hiding which images contain unsafe landmarks. In hand X-rays, a single cross-finger or wrist-cluster swap can invalidate downstream measurements even when the average error appears competitive. Thus, one-shot landmark detection should reduce not only average MRE but also the tail errors that dominate reliability risk and distort aggregate performance.

Topology provides a natural way to turn this reliability concern into a concrete constraint. Anatomical landmarks are not independent points: classical active shape and appearance models used shape priors to keep configurations plausible [4,3], and modern heatmap-regression methods also incorporate spatial configuration or global structure so that local evidence is interpreted in anatomical context [10,7]. This principle is especially relevant to hand X-rays, where multiple phalangeal joints have similar local appearance but remain constrained by finger order, relative spacing, and wrist-to-finger geometry. Once a topology prior identifies a prediction as visually plausible but anatomically inconsistent, the remaining question is how the system should respond. Reliability-oriented prediction and selective prediction argue that models should expose high-risk outputs rather than silently committing to every prediction [8,5,6], and medical image quality-control mechanisms similarly aim to identify cases requiring additional inspection [12]. For one-shot landmark transfer, these two lines of work motivate a conservative strategy: keep the strong foundation-model matching pipeline, but add an anatomical check that suppresses large correspondence errors and flags topology-inconsistent local refinements.

This paper builds on the FM-OSD global-to-local structure and asks whether topology-aware inference can suppress severe outliers without introducing a new

heavy detector or changing the original branch training logic. The resulting method, Topology-Aware Foundation-Model Matching (TAFM), estimates an image-specific anatomical prior from coarse correspondences, uses the prior to re-score global candidates, and accepts local refinement only when it remains consistent with the estimated layout. When local evidence conflicts with the topology prior, the method falls back to the topology-aware coarse point and records a review trigger. This conservative design aims to improve TAFM on the cases that most influence MRE and clinical reliability: rare but large landmark mistakes.

The contributions are summarized as follows:

1. We introduce reliability metrics for one-shot landmark detection, including severe landmark failures, image-level failure rates, maximum per-image error, hard-case behavior, and review-trigger statistics.
2. We propose TAFM, which constrains foundation-model matching and local refinement using an estimated anatomical prior, with the aim of reducing severe swaps and the MRE impact of outliers.
3. We provide a hand-only reliability analysis showing that TAFM reaches the strongest performance among the compared one-shot methods while substantially reducing severe-failure behavior.

2 Problem Setting

2.1 Template-Only Landmark Detection

Let I_t be a single labeled template hand X-ray with landmark set $P_t = \{p_i^t\}_{i=1}^N$, where $N = 37$ for the Hand benchmark [10]. Given a query image I_q, the task is to estimate the corresponding landmark set $\hat{P}_q = \{\hat{p}_i^q\}_{i=1}^N$. The strict template-only protocol permits the method to use the template image and its annotation, but it does not permit additional unlabeled target images for representation learning, pseudo-labeling, registration training, or test-time adaptation.

This protocol represents a realistic low-resource setting: a clinician or operator may provide one annotated reference case, while a model must assign landmark identities in new cases without collecting a target cohort in advance. It also stresses a different failure mode from standard supervised landmark detection. The main risk is not only small average error, but whether an anatomically implausible correspondence is produced for a small subset of landmarks.

2.2 Severe Swaps as Reliability Failures

In template-only landmark detection, the main reliability risk is an identity-preserving failure rather than ordinary local imprecision. A visually plausible match may land on the wrong phalangeal joint or wrist structure, producing a severe swap that is anatomically inconsistent with the rest of the hand. Such errors are rare relative to all landmark predictions, but a single swapped landmark can dominate image-level measurements and make the case unsafe for downstream use.

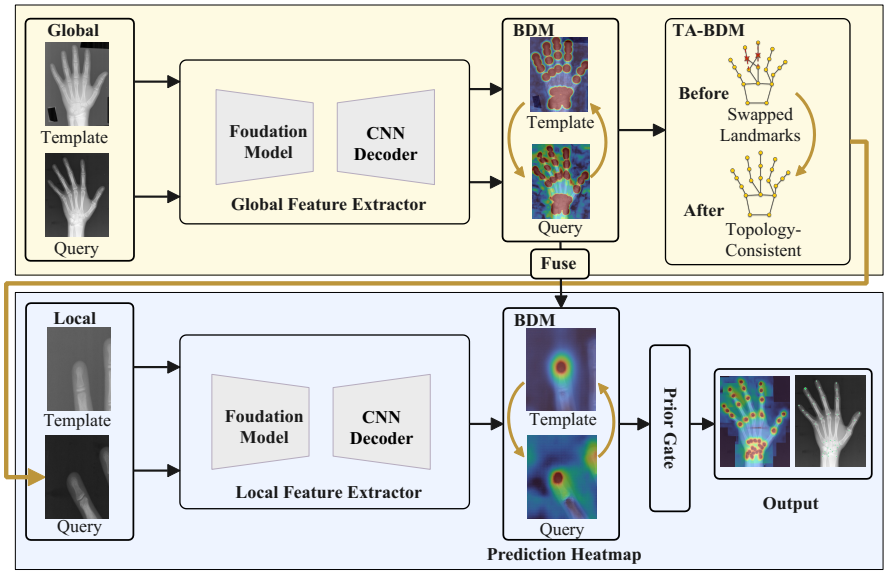

Fig. 1. Overview of the proposed TopoKAN framework. A frozen DINO encoder extracts template and query features, Conv-KAN decoders densify global and local representations.

A reliability-oriented detector should therefore reduce both average localization error and the risk of anatomically implausible outliers. This distinction is important because a method can improve typical landmarks while still leaving a small number of unsafe correspondences unresolved.

3 Topology-Aware Foundation-Model Matching

The proposed method is Topology-Aware Foundation-Model Matching (TAFM). It performs inference-time global-to-local template matching using the FM-OSD structure as the underlying detector [9]. The design principle is conservative: TAFM targets severe outliers without adding many trainable parameters or modifying the global and local branch training logic. Instead, it estimates an anatomical prior from the model's own coarse predictions and uses this prior to correct or flag unreliable matches at inference time.

3.1 Global and Local Matching Branches

We follow the FM-OSD global-to-local template matching structure [9]. Let ϕ_G denote the frozen global feature extractor followed by the official decoder. The template and query feature maps are

$$F_t^G = \phi_G(I_t), \qquad F_q^G = \phi_G(I_q). \tag{1}$$

For landmark i, the annotated template point p_i^t is mapped to a feature-grid coordinate u_i^t, and its descriptor is matched against all query-grid locations u. We use p for image-space points on the resized canvas and u for feature-grid coordinates; the two are related by the fixed scaling induced by the decoder stride, while millimetre conversion is applied only during metric computation. The visual matching score is

$$s_i(u) = \cos\left(F_q^G(u), F_t^G(u_i^t)\right). \tag{2}$$

A top-k candidate set $\mathcal{C}_i = \mathrm{TopK}_u(s_i(u))$ is filtered by bidirectional matching: each query candidate is mapped back to the template feature map, and the candidate whose reverse match is most consistent with u_i^t is selected. After converting the selected query-grid location to image coordinates, the global branch outputs the coarse landmark p_i^G.

The local branch refines a coarse point by cropping around it and applying the official local decoder. In the original global-to-local pipeline, the crop is centered on the global prediction p_i^G and the local decoder outputs a displacement correction. This refinement is useful when p_i^G is already on the correct anatomical structure, but it can also sharpen an incorrect correspondence if the crop is centered on a visually similar wrong joint. This motivates treating local refinement as a conditional update rather than an unconditional replacement.

Both branches keep the official FM-OSD training objective [9], which follows the common heatmap-regression practice of supervising landmark responses with Gaussian target maps [10]. For a training image feature map F and landmark grid coordinate u_i, the decoder is optimized so that the cosine-similarity response from the landmark descriptor to the dense feature map approximates a Gaussian heatmap,

$$H_{i,\sigma_h}(u) = \exp\left(-\frac{\|u - u_i\|_2^2}{2\sigma_h^2}\right), \qquad S_i(u) = \cos\left(F(u), F(u_i)\right). \tag{3}$$

The branch loss is the heatmap mean-squared error

$$\mathcal{L}_{\mathrm{hm}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{u}\left(S_i(u) - H_{i,\sigma_h}(u)\right)^2, \tag{4}$$

where σ_h is the heatmap bandwidth; a broader heatmap is used for the global branch and a narrower heatmap for local crops. TAFM does not introduce an additional training loss; it uses the trained responses in Eq. (2) and modifies only inference-time candidate selection and fallback.

TAFM is placed around these two branches. The global and local feature extractors, decoders, and training procedure remain unchanged. TAFM first uses the original coarse set $P_G = \{p_i^G\}_{i=1}^N$ to estimate a prior location $\bar{p}_i$ for each landmark, then produces a topology-aware coarse point p_i^P by re-scoring the global candidates, and finally obtains a topology-aware local proposal

$$p_i^{PL} = p_i^P + \Delta_i^L(p_i^P). \tag{5}$$

The final prediction $\hat{p}_i$ is selected by the topology gate in Sec. 3.3. TAFM adds only inference-time scoring and gating, keeps the parameter overhead negligible, and targets rare but severe outlier landmarks.

3.2 Coarse Correspondence and Anatomical Prior

For each landmark, the template feature at the annotated position is matched against query-image features to obtain a coarse candidate. A bidirectional check improves correspondence consistency by requiring the query candidate to map back near the original template landmark. However, bidirectional matching still treats landmarks independently and may confuse visually similar structures.

To introduce structure, we estimate an image-specific anatomical prior from the coarse landmark set. Let $P_G = \{p_i^G\}_{i=1}^N$ be the coarse predictions. We fit an affine transformation $T(p) = Ap + t$ from the template landmarks to P_G using residual trimming, producing an expected query location $\bar{p}_i = T(p_i^t)$ for each landmark. The corresponding feature-grid location is $\bar{u}_i = \Pi(\bar{p}_i)$, where Π is the canvas-to-grid scaling map. This prior does not assume a fixed hand shape; it estimates a coarse global layout for the current query image and downweights candidates that are inconsistent with that layout:

$$\tilde{s}_i(u) = s_i(u) - \lambda_{\text{pos}} \frac{\|u - \bar{u}_i\|_2^2}{\sigma_p^2}, \tag{6}$$

where $s_i(u)$ is the visual similarity score, $\bar{u}_i$ is the prior location on the feature grid, and σ_p is a fixed spatial bandwidth that controls how strongly the prior penalizes distant candidates.

3.3 Topology-Gated Local Refinement

Local refinement can improve fine localization, but it can also amplify an incorrect coarse match if the crop is centered on the wrong anatomical structure. The safeguard therefore treats local refinement as conditional. A refined local prediction p_i^{PL} is accepted only when it remains close to the prior location:

$$\hat{p}_i = \begin{cases} p_i^{PL}, & \|p_i^{PL} - \bar{p}_i\|_2 \le \gamma, \\ p_i^{P}, & \|p_i^{PL} - \bar{p}_i\|_2 > \gamma, \end{cases} \tag{7}$$

where p_i^P is the topology-aware coarse prediction and γ is the gate threshold in image-space pixels. This rule is intentionally conservative: when local evidence conflicts with the estimated anatomical layout, the pipeline falls back to the topology-aware coarse result. The same disagreement can also be interpreted as a review trigger.

Algorithm 1 summarizes the inference procedure for TAFM. The algorithm first obtains the original independent coarse matches P_G, fits a trimmed affine prior from P_t to P_G, and then re-scores each landmark candidate set using the prior-aware score. Local refinement is applied only after the topology-aware

Algorithm 1 Topology-aware bidirectional matching with reliability gate

Require: Template landmarks P_t; global feature maps F_t^G, F_q^G; local decoder Δ^L; top-k; prior weight λ_{pos}; bandwidth σ_p; gate threshold γ
Ensure: Final predictions $\hat{P}_q$ and review flags R
Initialize $P_G \leftarrow \emptyset$
for each landmark $i = 1, \ldots, N$ **do**
Compute $s_i(u) = \cos(F_q^G(u), F_t^G(u_i^t))$ as in Eq. (2)
Select $\mathcal{C}_i \leftarrow \text{TopK}_u(s_i(u))$
Reverse-check each $c \in \mathcal{C}_i$ against F_t^G and keep the candidate returning closest to u_i^t
Add the image-coordinate coarse prediction p_i^G to P_G
end for
Fit a trimmed affine prior $T(p) = Ap + t$ from P_t to P_G
for each landmark $i = 1, \ldots, N$ **do**
Compute prior location $\bar{p}_i = T(p_i^t)$ and grid prior $\bar{u}_i$
Re-score candidates with $\tilde{s}_i(u) = s_i(u) - \lambda_{\text{pos}} \|u - \bar{u}_i\|_2^2 / \sigma_p^2$
Apply top-k reverse checking with $\tilde{s}_i$ to obtain topology-aware coarse point p_i^P
Run local refinement in a crop centered at p_i^P to obtain p_i^{PL} as in Eq. (5)
if $\|p_i^{PL} - \bar{p}_i\|_2 \leq \gamma$ **then**
$\hat{p}_i \leftarrow p_i^{PL}$, $R_i \leftarrow 0$
else
$\hat{p}_i \leftarrow p_i^P$, $R_i \leftarrow 1$ ▷ fallback and review trigger
end if
end for
return $\hat{P}_q = \{\hat{p}_i\}_{i=1}^N$ and $R = \{R_i\}_{i=1}^N$

coarse point p_i^P is selected. The gate in Eq. (7) then determines whether the refined point p_i^{PL} is accepted or replaced by p_i^P, while the binary flag R_i records whether the landmark should be treated as topology-inconsistent for reliability analysis.

4 Reliability Metrics

We report conventional localization metrics only as context. The primary goal is to characterize severe failures and review behavior.

Average localization. We report MRE and successful detection rate (SDR) at 2, 4, and 10 mm. These metrics allow comparison with prior Hand X-ray landmark work but are not sufficient for reliability analysis.

Landmark-level severe failures. We count landmarks with error larger than 10 mm and 20 mm over the full Hand test set. This directly measures the frequency of large anatomical mistakes.

Image-level severe failures. We report the percentage of test images containing at least one landmark error larger than 10 mm or 20 mm. This estimates how often a case would require attention before downstream measurements are trusted.

Maximum per-image error. For each test image, we compute the maximum landmark error. We summarize the distribution using mean, median, and 95th percentile. This highlights tail behavior that is hidden by MRE.

Review-trigger statistics. For topology-gated refinement, we report the number of accepted and rejected local refinements, the image-level rate of review-triggered cases, and the recall of severe failures among triggered cases. For landmark i, the review trigger is defined as

$$R_i = \mathbb{1}\left[\|p_i^{PL} - \bar{p}_i\|_2 > \gamma\right], \tag{8}$$

and an image is review-triggered if any landmark in that image has $R_i = 1$. A prevented severe failure at threshold $\tau \in \{10\,\text{mm}, 20\,\text{mm}\}$ is counted when the local prediction would be severe, the gate rejects it, and the final fallback is no longer severe:

$$\mathbb{1}\left[e_i(p_i^{PL}) > \tau\right]\mathbb{1}\left[R_i = 1\right]\mathbb{1}\left[e_i(\hat{p}_i) \le \tau\right]. \tag{9}$$

These metrics evaluate whether the safeguard can act as a practical warning signal rather than only an accuracy-improving post-processing step.

Hard-case distribution. To evaluate tail behavior without selecting only one method's failures, we form a symmetric hard-case subset. We take the union of the top-20 test images ranked by per-image MRE for FM-OSD and the top-20 images ranked by per-image MRE for the full safeguard. If the union contains fewer than 40 unique images, the remaining images are selected by the average per-image MRE of the two methods. This subset is used only for reliability analysis and visualization; the main quantitative claims remain based on the full test split.

5 Experiments

5.1 Dataset and Protocol

We use the Hand X-ray benchmark [10], which contains 909 radiographs with 37 annotated landmarks. Image 10 is used as the template, and evaluation is performed on the standard 300-image test split, giving 11,100 landmark predictions per method. Predictions are made on the resized 840×640 canvas and converted to millimeters using the dataset scale metadata.

5.2 Compared Variants

All reproduced FM-OSD and TAFM variants use the official FM-OSD decoder architecture and pretrained weights, so the comparison isolates the reliability effect of topology-aware matching and fallback from changes in model architecture. We compare four variants:

- **FM-OSD Global**: reproduced official FM-OSD global branch with standard bidirectional matching.
- **FM-OSD**: reproduced official FM-OSD global branch plus local crop refinement.
- **TAFM Global**: TAFM coarse matching using a robust affine prior estimated from standard coarse predictions; no local refinement is used in the final output.
- **TAFM**: TAFM Global plus official FM-OSD local refinement and topology gating. If the local prediction deviates from the fitted topology prior by more than γ, the method falls back to the topology-aware coarse point.

We also include the FM-OSD paper's reported Hand results as reference rows where applicable; outlier counts are reported only for reproduced predictions because they require per-landmark outputs.

5.3 Implementation Details

We use the official FM-OSD global and local decoders for all reproduced and TAFM variants, so the comparison isolates the effect of topology-aware inference rather than architectural or training changes. The feature extractor is DINO ViT-S/8 with key features from layer 8 [2]. Matching uses stride-4 features, an 840×640 input canvas, a load size of 224, and top-$k = 5$ bidirectional matching. TAFM uses the same top-k setting with position weight $\lambda_{\text{pos}} = 0.07$, spatial bandwidth $\sigma_p = 38$ pixels, robust affine trimming ratio 0.85, and topology-gate threshold $\gamma = 80$ pixels. All quantitative results are recomputed from per-landmark predictions on the full test split.

6 Results

6.1 Full-Test Reliability

Table 1 reports conventional accuracy and severe-failure metrics on the full Hand test split. The reproduced FM-OSD baseline confirms the value of local refinement for average localization accuracy, but it also shows a key limitation: refining a local crop does not by itself remove severe correspondence errors inherited from the global stage. This is the regime in which topology is most useful. TAFM Global removes many anatomically implausible coarse matches even without local refinement, showing that the fitted prior directly improves the reliability of template matching. The full TAFM variant then combines this safer coarse

Table 1. Reliability analysis on the Hand X-ray dataset. "Report" denotes values reported by FM-OSD [9]; "Reproduce" denotes our official FM-OSD reproduction. Outlier metrics are computed from per-landmark predictions and are therefore reported only for reproduced methods.

Method	Source	Branch	MRE↓ (mm)	SDR@2↑ (%)	SDR@4↑ (%)	SDR@10↑ (%)	>10↓ count	>20↓ count	Img. >10↓ (%)	Img. >20↓ (%)
FM-OSD Global	Report	Global	1.86	76.75	96.60	99.13	–	–	–	–
FM-OSD	Report	Global+Local	1.41	**86.66**	96.66	99.11	–	–	–	–
FM-OSD Global	Reproduce	Global	1.992	72.74	94.89	99.23	86	79	15.00	13.33
FM-OSD	Reproduce	Global+Local	1.655	83.76	96.23	99.24	84	79	14.67	13.33
TAFM Global	Ours	Global	1.690	72.94	95.23	99.72	31	25	3.67	2.33
TAFM	Ours	Global+Local	**1.340**	84.21	**96.71**	**99.74**	**29**	**23**	**3.67**	**1.67**

stage with local refinement and conservative fallback, yielding the best overall performance among the compared one-shot methods.

The relative pattern is more informative than the absolute MRE alone. FM-OSD improves the center of the error distribution, whereas TAFM mainly changes the tail: large landmark swaps become much rarer and severe cases concentrate in fewer images. This distinction supports the reliability-oriented design of TAFM. The topology prior does not merely shift all predictions slightly closer to the ground truth; it selectively suppresses anatomically implausible correspondences that dominate failure risk.

Fig. 2 provides a qualitative view of this tail behavior on representative hard cases. FM-OSD Global and FM-OSD can make isolated but severe finger-level or wrist-level swaps, visible as long yellow displacement lines among otherwise plausible landmarks. TAFM Global reduces many of these coarse mismatches through the fitted anatomical prior, and TAFM preserves local refinement when it agrees with the prior while falling back to the topology-aware coarse point when the local prediction is unreliable.

Fig. 3 summarizes the per-image MRE distribution on the full 300-image test split. FM-OSD Global and FM-OSD retain a heavier high-error tail, whereas TAFM Global and TAFM shift the distribution downward. This pattern is consistent with the table-level observation that the topology prior acts mainly on difficult images rather than uniformly displacing all predictions.

6.2 Topology Gate as a Review Signal

Table 2 reports topology-gate behavior for TAFM at $\gamma = 80$ pixels. The gate is sparse: it rejects 12 of 11,100 local refinements and triggers 11 of 300 images. Despite this low trigger rate, the triggered set captures 50% of images whose pre-gate local prediction contains a > 20 mm severe failure. The precision is low because the gate is designed as a conservative warning signal rather than a final diagnosis of error; in practice, triggered cases should be interpreted as candidates for review. This tradeoff is expected because the gate triggers at the landmark level but is evaluated at the image level: a single topology-inconsistent refinement flags the whole case even when the remaining landmarks are accurate. The operating point should therefore be chosen according to the cost asymmetry between missing a severe failure and sending a safe case for review.

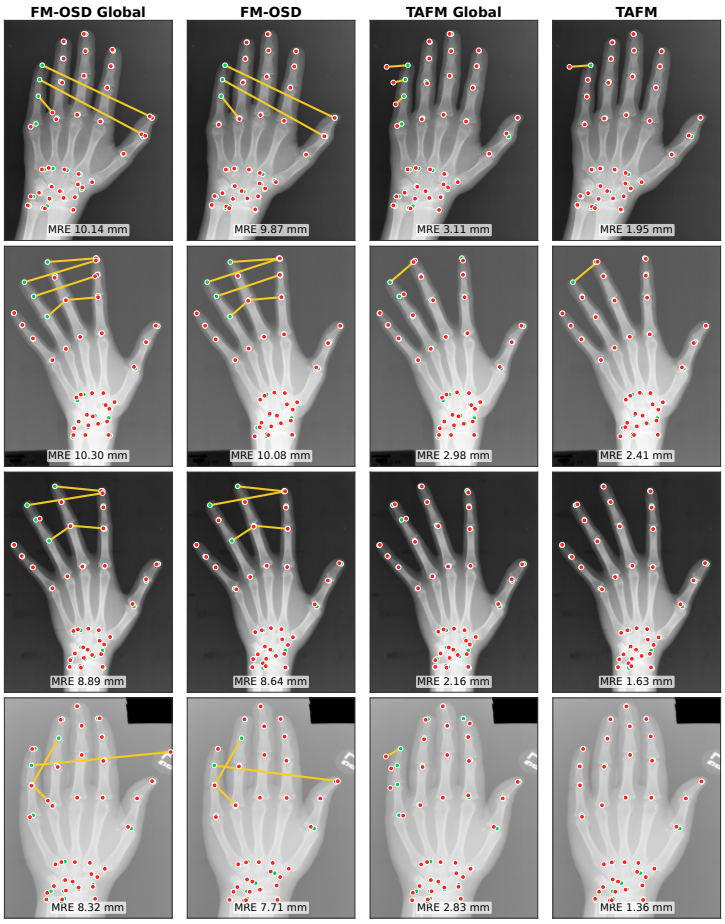

Fig. 2. Qualitative reliability examples on hard Hand X-ray test cases. Green and red points denote ground truth and predictions, respectively, while yellow lines connect corresponding landmarks. Compared with FM-OSD Global and FM-OSD, TAFM reduces severe landmark swaps by combining topology-aware coarse matching with local refinement and fallback.

Table 2. Topology-gate behavior for TAFM. Severe images are defined on the pre-gate local prediction p^{PL}; triggered images contain at least one rejected local refinement.

Setting	Accepted landmarks	Rejected landmarks	Triggered images	Triggered images (%)	Severe Img.>10	Severe Img.>20	Recall >10	Recall >20	Precision >20
$\gamma = 80$ px	11088	12	11	3.67	10	4	0.20	0.50	0.18

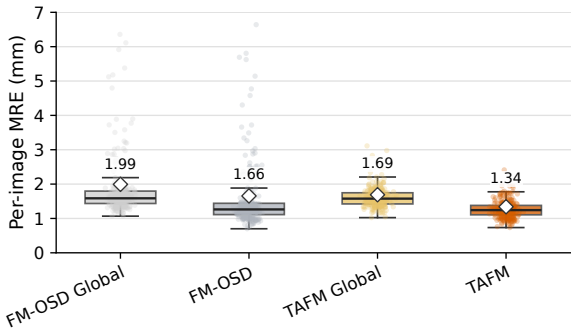

Fig. 3. Per-image MRE distribution on the 300-image Hand X-ray test set. Boxes show the interquartile range, horizontal lines denote medians, diamonds denote means, and jittered points show individual test images. TAFM compresses the high-error tail while maintaining the central accuracy of the baseline.

6.3 Hard-Case Analysis

The hard-case subset is designed to evaluate tail behavior without favoring either method. We take the union of the 20 test images with the largest per-image MRE under FM-OSD and the 20 test images with the largest per-image MRE under TAFM; if the union contains fewer than 40 unique images, the remaining cases are filled by the average per-image MRE of the two methods. Table 3 reports reliability on this 40-image subset. The column N is the number of hard cases, Mean MRE and Median MRE summarize the per-image average landmark error, and Img.$>$10/Img.$>$20 give the percentage of hard cases containing at least one landmark error above 10 mm or 20 mm.

The symmetric construction is important because the subset is the union of failures from both FM-OSD and TAFM, rather than a cherry-picked set for either method. The comparison shows that the hard subset is dominated by severe swaps in the FM-OSD variants. Local refinement reduces average error only modestly on this subset, and the image-level severe-failure rates remain unchanged, indicating that local crops do not correct many inherited correspondence mistakes. In contrast, TAFM Global substantially lowers both mean and median hard-case error, showing that the topology prior directly improves coarse matching. The full TAFM variant further benefits from local refinement after the coarse prediction has been made more reliable, reducing mean MRE from 4.226 mm to 2.222 mm and reducing hard cases with $>$ 20 mm failures from 67.5% to 12.5%. This supports the main reliability claim: topology-aware inference is most valuable in the difficult images where template-only matching produces anatomically implausible outliers.

7 Discussion

The results support the reliability-oriented framing of template-only landmark detection. FM-OSD improves average error over FM-OSD Global, but its severe-

Table 3. Hard-case reliability on the Hand X-ray dataset. The subset contains 40 images selected symmetrically from FM-OSD and TAFM failures.

Method	Branch	N	Mean MRE↓ (mm)	Median MRE↓ (mm)	Img.>10↓ (%)	Img.>20↓ (%)
FM-OSD Global	Global	40	4.509	3.542	70.0	67.5
FM-OSD	Global+Local	40	4.226	3.148	70.0	67.5
TAFM Global	Global	40	2.619	1.932	22.5	17.5
TAFM	Global+Local	40	**2.222**	**1.674**	**17.5**	**12.5**

failure counts remain close to the global-only baseline. This suggests that local refinement alone does not eliminate anatomically implausible correspondences. By contrast, TAFM Global substantially reduces severe outliers even without local refinement, showing that a fitted anatomical prior can correct many coarse swaps.

TAFM combines the complementary strengths of topology-aware matching and local refinement. The method obtains state-of-the-art performance among the compared one-shot Hand X-ray methods while also reducing the most severe image-level failures. This is important because downstream measurements can be invalidated by a single badly swapped landmark. The improvement in image-level $> 20\,\mathrm{mm}$ failures, from 13.33% to 1.67%, is therefore more informative for reliability than the MRE gain alone.

The gate analysis should be interpreted as a review signal rather than a complete failure detector. With $\gamma = 80$ pixels, the gate rejects only 12 landmarks and triggers 11 images, so it is intentionally sparse. Its recall for pre-gate $> 20\,\mathrm{mm}$ severe images is 0.50, indicating that topology disagreement can identify a meaningful subset of high-risk cases. However, the current trigger is not sufficient as a standalone quality-control system; future work should calibrate the gate threshold and combine topology residuals with uncertainty or image-level confidence scores.

Three limitations qualify the conclusions. First, all experiments use the Hand X-ray benchmark and one fixed template, so template-selection sensitivity and transfer to other anatomical sites remain to be tested. Second, TAFM relies on a trimmed affine prior, which is appropriate for many global hand pose variations but may underfit anatomy with stronger articulated deformation; piecewise or nonrigid priors are natural extensions. Third, the topology gate uses a fixed threshold $\gamma = 80$ pixels. Per-landmark calibration or combination with descriptor-level confidence may improve the precision of the review signal without sacrificing recall for severe failures.

8 Conclusion

We presented a reliability-oriented formulation of one-shot anatomical landmark detection under extreme annotation scarcity. TAFM treats the most important failure mode as an inference-time reliability problem: an image-specific anatomical prior re-scores global candidates, gates local refinement, and exposes topology-inconsistent predictions for review. On the Hand X-ray benchmark,

TAFM reduces MRE from 1.655 mm to 1.340 mm, image-level > 20 mm severe failures from 13.33% to 1.67%, and hard-case mean MRE from 4.226 mm to 2.222 mm. These results show that lightweight topology-aware priors can improve the tail reliability of foundation-model template matching without changing the detector architecture or retraining the underlying branches.

Acknowledgements This work was supported by the Key Research and Development Program of Hainan Province (No. ZDYF2025(LALH)002) and the Beijing Natural Science Foundation (No. L232039).

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep ViT Features as Dense Visual Descriptors. arXiv preprint arXiv:2112.05814 (2021)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6), 681–685 (2001). https://doi.org/10.1109/34.927467
4. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. Computer Vision and Image Understanding **61**(1), 38–59 (1995). https://doi.org/10.1006/cviu.1995.1004
5. Geifman, Y., El-Yaniv, R.: Selective Classification for Deep Neural Networks. In: Advances in Neural Information Processing Systems 30 (NeurIPS) (2017)
6. Geifman, Y., El-Yaniv, R.: SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2151–2159. PMLR (2019), https://proceedings.mlr.press/v97/geifman19a.html
7. Jiang, Y., Li, Y., Wang, X., Tao, Y., Lin, J., Lin, H.: CephalFormer: Incorporating Global Structure Constraint into Visual Features for General Cephalometric Landmark Detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 227–237. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_22
8. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Advances in Neural Information Processing Systems 30 (NeurIPS) (2017)
9. Miao, J., Chen, C., Zhang, K., Chuai, J., Li, Q., Heng, P.A.: FM-OSD: Foundation Model-Enabled One-Shot Detection of Anatomical Landmarks. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 297–307. Springer (2024)
10. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating Spatial Configuration into Heatmap Regression Based CNNs for Landmark Localization. Medical Image Analysis **54**, 207–219 (2019)
11. Quan, Q., Yao, Q., Li, J., Zhou, S.K.: Which Images to Label for Few-Shot Medical Landmark Detection? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20606–20616 (2022)

12. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Alzheimer's Disease Neuroimaging Initiative: Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-Wise Quality Control. NeuroImage **195**, 11–22 (2019). https://doi.org/10.1016/j.neuroimage.2019.03.042
13. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., Siao, M.J., Lai, T.M., Ibragimov, B., Vrtovec, T., Ronneberger, O., et al.: A Benchmark for Comparison of Dental Radiography Analysis Algorithms. Medical Image Analysis **31**, 63–76 (2016)
14. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images. IEEE Transactions on Medical Imaging **41**(10), 2658–2669 (2022)
15. Yao, Q., Quan, Q., Xiao, L., Zhou, S.K.: One-Shot Medical Landmark Detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 177–188. Springer (2021)
16. Zhou, G.Q., Miao, J., Yang, X., Li, R., Huo, E.Z., Shi, W., Huang, Y., Qian, J., Chen, C., Ni, D.: Learn Fine-Grained Adaptive Loss for Multiple Anatomical Landmark Detection in Medical Images. IEEE Journal of Biomedical and Health Informatics **25**(10), 3854–3864 (2021)
17. Zhu, H., Quan, Q., Yao, Q., Liu, Z., Zhou, S.K.: UOD: Universal One-Shot Detection of Anatomical Landmarks. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 24–34. Springer (2023)
18. Zhu, H., Yao, Q., Xiao, L., Zhou, S.K.: You Only Learn Once: Universal Anatomical Landmark Detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 85–95. Springer (2021)