Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

For foundation-model registration in correlative microscopy, cross-modal appearance matters more than field of view

Frank Cai

cai485@purdue.edu

Purdue University System

Article

Keywords: correlative microscopy, image registration, dense feature matching, foundation models, domain fine-tuning, catastrophic forgetting

Posted Date: July 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10311464/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Correlative materials microscopy pairs images of one specimen across modalities (SEM, EBSD, TEM, optical) that share little appearance and can differ in field of view (FOV) by orders of magnitude, defeating classical registration. We tested whether a scale-aware pyramidal wrapper around pretrained dense matchers (RoMa, ELoFTR, MatchAnything) could lift cross-modal registration on AmalgaMatch (187 pairs, 19 subsets). A naive tiling pyramid degrades matchers badly: because they never abstain, every tile floods robust estimation with thousands of confident matches (median error 76→1794 px). A redesigned verified coarse-to-fine wrapper recovers a small but significant gain (SR@10 0.10→0.12, p = 0.017) without losing any pair, yet FOV ≤ 5% success stays zero. The largest lever was the backbone: cross-modal-trained MatchAnything-RoMa gave the only significant gain over zero-shot (SR@10 + 0.032, p = 0.018), among high-FOV pairs. A controlled FOV ladder isolating scale (appearance fixed) shows the wrapper triples success at 10% FOV (0.07→0.23, p = 0.0014), so the scale mechanism is sound. Decoder-only fine-tuning cut in-distribution TEM error ~ 5×, but across eight seeds regressed SR@20 (0.393→0.26) by forgetting untrained modalities, which L2-SP did not fix. Appearance, not scale, is the dominant uncontrolled factor on AmalgaMatch, reached by elimination rather than a direct appearance sweep.

Introduction

Correlative microscopy spatially associates measurements of a single specimen acquired by different instruments: scanning electron microscopy (SEM), electron backscatter diffraction (EBSD), transmission electron microscopy (TEM), optical light microscopy (LOM), and derived modalities such as image-quality and inverse-pole-figure maps. The aim is to read out structure, chemistry, crystallography, and mechanics at the same physical location [1]. The scientific payoff is direct: linking a dislocation configuration in TEM to a grain orientation in EBSD, or slip partitioning in SEM-DIC to the underlying microstructure, requires that the two images be brought into a common coordinate frame to sub-grain precision. Registration is that step, and it is the bottleneck. Unlike natural-image or remote-sensing registration, correlative microscopy couples two compounding difficulties. The modalities share little mutual information, because a secondary-electron contrast image and a crystallographic orientation map are generated by entirely different physics. And the field of view (FOV) can differ by more than an order of magnitude, with the narrow-FOV image covering as little as 2% of the wide-FOV frame. Classical intensity- and feature-based methods assume shared appearance and comparable scale, so they fail outright on this regime [2].

Learned dense feature matchers have transformed correspondence estimation for natural images. Detector-free transformers such as LoFTR [3] and its efficient successor [4] produce semi-dense matches in texture-poor scenes; RoMa [5] couples frozen DINOv2 features [6] with a specialised fine-feature branch and a match-classification decoder to yield pixel-dense warps with calibrated certainty; and MatchAnything [7] retrains such backbones on large-scale synthetic cross-modal signals, reporting strong generalisation across more than eight unseen cross-modality tasks with a single weight set. One

property of these detector-free dense matchers becomes central below. Sparse learned matchers built on keypoint detection and attention-based assignment [15, 16] can decline to match; dense matchers instead emit a confident correspondence at essentially every pixel and never abstain. All of these models are trained almost entirely on macroscale imagery, which raises the question of whether they transfer to materials microstructures, and of whether their failures at severe FOV mismatch can be repaired without retraining.

A natural hypothesis is that the FOV gap is the dominant problem and that a scale-aware wrapper can close it: crop the wide-FOV source into a pyramid of target-resolution tiles, match each tile against the narrow-FOV target, pool the correspondences, and fit a single global transform by robust estimation. The appeal is that it requires no training and treats the matcher as a black box. The project hypotheses were **(H1)** that such pyramidal patching yields ≥ 35% relative gain at FOV ≤ 5% over the best zero-shot baseline; **(H2)** that RoMa's dense flow tolerates partial overlap better than the ELoFTR family at low FOV; and **(H3)** that an affine model is sufficient for AmalgaMatch, with full homography offering no significant benefit.

Here we test these hypotheses end-to-end on all 187 AmalgaMatch pairs, with controls, ablations, and a controlled decoupling experiment, and the result is as much diagnostic as methodological. The naive pyramid does not merely fail to help. It destroys the matchers it wraps, for a structural reason we trace to the fact that dense matchers never abstain. A redesigned verified coarse-to-fine wrapper recovers a small, real gain, but it cannot reach the severe-FOV regime. The single largest off-the-shelf lever turns out to be the backbone rather than the wrapper. A controlled FOV ladder then separates scale from appearance on real pairs and shows that the wrapper mechanism is in fact sound for scale; the real benchmark simply never presents scale as the isolated failure mode, because the low-FOV pairs are also the most appearance-divergent. Finally we attack appearance directly with materials-domain fine-tuning, observe forgetting of untrained modality combinations, and use a three-seed study to separate what a one-line weight anchor can and cannot fix. The contributions are four. First, a mechanistic account of why pyramidal wrappers break non-abstaining dense matchers, together with a verified coarse-to-fine wrapper that does not (Fig. 1a). Second, a controlled FOV-ladder protocol that decouples scale from appearance and validates the wrapper's scale mechanism (Fig. 1b). Third, evidence that cross-modal appearance, not FOV, is the dominant constraint on this benchmark. Fourth, a three-seed robustness analysis of decoder-only fine-tuning that diagnoses breadth of modality coverage, not the choice of optimiser or weight anchor, as the binding constraint on forgetting, together with the fine-tuning recipe itself (a large, seed-robust in-distribution gain with a modest, seed-noisy held-out regression).

Figure 1. Method overview. (**a**) The verified coarse-to-fine wrapper (pyramid v2): a direct match yields an incumbent transform; candidate tile-search and zoom stages are invoked only on weak support and accepted only when a mutual-information-on-overlap verifier beats the incumbent, making the wrapper monotone with respect to the direct fit. (**b**) The FOV-ladder protocol crops real base-matchable pairs to shrinking field-of-view ratios with appearance, modality, and pixel size fixed, isolating scale as the only variable.

Results

Benchmark, protocol, and baselines

We evaluate on AmalgaMatch [2]: 187 image pairs spanning six correlative-microscopy matching tasks (orientation mapping, serial sectioning, multiscale, slip partitioning, fracture surfaces, dislocation characterisation) across 19 material subsets, each pair carrying ground-truth (GT) point correspondences. Per pair we match, robustly fit a transform with MAGSAC++ [8] (5.5 px reprojection threshold; affine vs. homography selected by a BIC-style criterion on inlier residuals), refine on inliers with a thin-plate spline [9], and report the mean Euclidean distance (ED) between GT correspondences projected into source coordinates. Aggregates are success rates SR@{5,10,20} px and median ED over pairs; significance is assessed by paired bootstrap (B = 10,000, 95 % CIs). One protocol floor matters: the GT itself fits a global affine only to ~10.3 px median residual, so px-threshold metrics are floored well above sub-pixel, and sub-5 px claims on AmalgaMatch measure the GT more than the method. FOV strata use the GT-implied area ratio; only four pairs sit below ratio 0.05 under any definition.

Table 1 reports the headline comparison. Classical baselines reproduce the benchmark's own finding: SIFT [10] succeeds on roughly 3 of 187 pairs (Control A), and mutual-information intensity refinement [11] (Control B) moves median ED but flips no pair to success. Among zero-shot foundation matchers, RoMa is the strongest single backbone (median ED 76.3 px, SR@10 0.10, SR@20 0.23), and it dominates the MatchAnything-ELoFTR configuration everywhere (SR@10 0.10–0.13 vs. 0.01–0.02), including every low-FOV stratum, which supports H2. Figure 2 shows the full success-rate bars and Figure 3 the per-task-group breakdown.

Table 1. Headline registration accuracy on all 187 AmalgaMatch pairs.

Configuration	med ED (px)	SR@5	SR@10	SR@20
SIFT (Control A)	908	0.01	0.02	0.02
SIFT + MI (Control B)	824	—	~0.02	—
LoFTR	270	0.03	0.06	0.10
MatchAnything-ELoFTR	510	0.01	0.01	0.02
RoMa (zero-shot)	76.3	0.05	0.10	0.23
RoMa + pyramid v1	1794	0.00	0.01	0.02
RoMa + pyramid v2	74.0	0.05	0.12	0.25
MatchAnything-RoMa	84.0	0.04	0.13	0.24
MatchAnything-RoMa + pyramid v2	**72.7**	0.05	**0.13**	0.24

Figure 2

Figure 2. Success rates across all configurations on AmalgaMatch. Pyramid v1 collapses RoMa; pyramid v2 recovers a small significant gain without losing pairs; the MatchAnything-RoMa backbone provides the only significant headline gain over the zero-shot bar.

Figure 3

Figure 3. Per-task-group success rates. Performance is concentrated in orientation-mapping and serial-sectioning groups; dislocation characterisation and low-mutual-information pairs remain hard for every configuration, consistent with appearance (not scale) being the dominant constraint.

The naive pyramid catastrophically degrades dense matchers

The straightforward implementation of the scale hypothesis (pyramid v1) tiles the source, matches each tile, pools all correspondences, and fits one transform. It does not help; it destroys the backbone. RoMa's median ED rose from 76 to 1794 px, and 106 of 187 pairs became hard failures (Fig. 2). The mechanism is structural, and it is the central negative finding about the wrapper class. Dense matchers never abstain: every tile, whether or not it overlaps the target, returns on the order of 10^4 confident matches. Pooling across a pyramid therefore floods the robust estimator with mostly-spurious correspondences, collapses the inlier fraction from 0.114 (direct) to 0.005, and leaves MAGSAC++ no consensus to find. This is a property of the non-abstaining dense-matcher class rather than a tuning artefact, and it predicts that any pool-then-fit pyramid will fail on these backbones regardless of tile size or overlap.

A verified coarse-to-fine wrapper recovers a small, real gain

We redesigned the wrapper so that it never pools blindly (pyramid v2; Fig. 1a). It matches the full pair directly first. It then invokes candidate stages, namely tile search over a scaled source grid followed by zoom refinement on the best region, but only when direct support is weak, and it accepts a candidate transform only if a mutual-information-on-overlap verifier scores it above the incumbent. This makes the wrapper monotone by construction: it can replace a bad direct fit, but tile noise cannot drag it below one. On RoMa, pyramid v2 lifts SR@10 from 0.10 to 0.12 (+0.021, 95 % CI [+0.005, +0.043], $p = 0.017$) with zero pairs lost, and it produces the only non-zero severe-stratum results in the entire study (0.00 → 0.03). Two ablations confirm the design is at its useful limit. Iterated zoom, which chains refinements, is not significantly better than a single zoom and is borderline worse, because each stage multiplies verifier error. Certainty gating at 0.5 discards low-certainty matches before fitting, and it is significantly worse than plain v2 (SR@20 −0.037, CI [−0.070, −0.011]), because it starves the estimator on exactly the hard pairs where everything is low-certainty. Plain v2 is therefore the final wrapper.

On the stronger MatchAnything-RoMa backbone, however, the wrapper's headline contribution vanishes (SR@10 flat) and the no-regression property breaks (two pairs gained, two lost, all straddling the 7–13 px threshold band). The mutual-information verifier is good enough to rescue a weak backbone's coarse

failures, but it cannot discriminate between transforms that already agree to within ~10 px. This verifier ceiling is the wrapper's intrinsic limit, and it reappears below.

The largest off-the-shelf lever is the backbone

MatchAnything's released RoMa checkpoint is key-for-key compatible with the roma_outdoor architecture (603/603 tensors, all retrained), so it drops into the identical wrapper and pipeline with no code changes. Swapping it in for stock RoMa is the only intervention in the study that clears the zero-shot bar with a significant headline gain: SR@10 0.10 → 0.13 (+0.032, 95 % CI [+0.005, +0.064], p = 0.018). The gain sits entirely in the ≥ 0.5 FOV stratum (0.13 → 0.18): cross-modal training repairs appearance on pairs that were nearly matchable, with an all-or-nothing profile (slightly worse SR@5, higher median ED when it misses). Severe-FOV pairs, which are small, low-texture, and modality-divergent, remain at zero for every configuration tested. A change of weights rather than of geometry produced the project's only headline gain, and that is the first direct sign that appearance, not scale, is the operative constraint. The finding is consistent with the AmalgaMatch authors' own evaluation, which independently reports that MatchAnything-RoMa outperforms MatchAnything-ELoFTR across most subsets [1]; our contribution is to quantify where that gain lives (high FOV) and to show what it cannot reach (severe FOV).

A controlled FOV ladder decouples scale from appearance

The real dataset cannot answer the question that H1 was about, namely the failure FOV per backbone, because the low-FOV pairs are simultaneously the most appearance-divergent and only four pairs sit below area ratio 0.05. We therefore constructed a controlled FOV ladder (Fig. 1b). For every base-matchable pair (direct mean ED < 20 px; 36–40 pairs per backbone) we cropped the target to absolute area ratios {0.5, 0.25, 0.1, 0.05, 0.02}, holding appearance, modality gap, and pixel sizes fixed while sweeping FOV. All GT is kept for evaluation, so out-of-crop points test the global transform's extrapolation. With scale thus isolated (Fig. 4):

- **Direct matching collapses between FOV 0.25 and 0.1** for both RoMa and MatchAnything-RoMa: SR@10 holds near base levels through 0.5, bends at 0.25, and collapses at 0.1, with a hard floor at 0.02 for every configuration.
- **With scale isolated, the wrapper does what it was designed to do.** At FOV 0.1, MatchAnything-RoMa with pyramid v2 holds SR@10 0.23 against 0.07 direct, a gain of +0.150 (95 % CI [+0.050, +0.275], p = 0.0014) and a more-than-threefold relative improvement in the regime H1 targeted.

This resolves the project's central tension. The wrapper mechanism is sound for scale, and the usable-FOV envelope extends from about 0.25 (direct) to about 0.1 (wrapped, strong backbone). H1 as stated remains rejected on real pairs, but only because the real distribution never isolates scale as the failure

mode: the severe-FOV pairs fail on appearance first, and no scale wrapper can touch that. Figure 5 shows the per-backbone FOV breakdown on the native (uncropped) pairs for comparison.

Figure 4

Figure 4. Controlled FOV ladder with appearance fixed. Cropping base-matchable pairs to shrinking FOV ratios shows direct matching collapsing between 0.25 and 0.1, while the verified wrapper restores SR@10 at the 0.1 rung (0.07 → 0.23, p = 0.0014): the wrapper's scale mechanism is sound once scale is isolated.

Figure 5

Figure 5. Success rate versus field-of-view stratum on native pairs. On the real distribution, low-FOV strata are simultaneously the most appearance-divergent, confounding scale and appearance; severe-FOV success is at the floor for every configuration.

Materials-domain fine-tuning: the appearance lever and its cost

The sections above leave appearance as the dominant constraint and name materials-domain fine-tuning as the credible path, so we tested it directly. A scene-level 131/28/28 train/val/test split prevents leakage, because within-scene pairs share images and a pair-level split would leak. All headline numbers come from the 28 test pairs only, with validation driving model selection and the test set touched once. Fine-tuning is decoder-only: the VGG+DINOv2 encoder is frozen in evaluation mode so that its normalisation statistics cannot drift, and ~100M decoder parameters are trained on densified sparse GT (sparse correspondences splined into a dense A→B warp, supervised inside the GT-support region) with FOV-crop and photometric augmentation. Selection by median validation ED chose an early checkpoint (validation median 17.6 vs. 21.7 px zero-shot).

The aggregate test result (Table 2) hides two opposite effects that a per-subclass cut separates. In-distribution, fine-tuning works hard: the 16 TEM dislocation test pairs (50 TEM pairs in training) improved median ED from 320.8 to 61.8 px, a 5.2× reduction and the single largest metric movement we saw, although the improvement runs mostly from ~320 to ~60 px and so stays above the 20 px bar, where SR barely registers it. Off-distribution, the model forgets: the C103 SEM-SE↔LOM-height subclass has zero training pairs and exists only in val/test, so it probes catastrophic forgetting directly. This subclass holds two scenes (two test pairs each). Across the three seeds examined at the per-scene level (seeds 1–3) the second scene is unrecoverable in every run, plain or anchored (residual 370–1400 px against a ~12 px zero-shot baseline), which marks a genuine modality-coverage gap rather than an optimisation failure. The first scene is more fragile than seed-stable: the plain fine-tune retains it across all three seeds (≤31 px), but it is exactly this scene that the originally reported single run (seed 0) broke (80/1220 px), so the headline "catastrophic forgetting" of the recoverable scene is itself a seed-dependent event. What is seed-robust is the unrecoverable second scene (seeds 1–3) and the resulting modest SR@20 regression relative to zero-shot (0.393 → 0.26 ± 0.02 over eight seeds); with only two C103 pairs in validation, a

median selection metric cannot see this damage in any run. A FOV-ladder analysis on the fine-tuned backbone (same 63-pair testbed) confirms that the direction is right but partly redundant with the wrapper. Fine-tuning helps the moderate-FOV regime, holding SR@10 0.37–0.39 at rung 0.25 against 0.28–0.30 for the zero-shot model, and the pyramid still adds a significant but roughly halved scale lift at the severe 0.1 rung (+0.078, CI [+0.020, +0.157], p = 0.0154, against +0.150 on stock MatchAnything-RoMa). Fine-tuning and the wrapper address overlapping scale and appearance failures, so they do not fully stack.

Table 2. Decoder-only fine-tuning on the held-out 28-pair test split: mean ± SD over eight training seeds (TPS-refined ED; zero-shot deterministic). Both deployment protocols — direct match and the verified pyramid-v2 wrapper — are evaluated on the same per-seed checkpoints. SR@5 ≈ 0 throughout (the ~10.3 px GT affine-residual floor) and is omitted.

Method (8-seed mean ± SD)	Protocol	SR@10	SR@20	med ED (px)
MatchAnything-RoMa (zero-shot)	direct	0.214	0.393	83.5
	pyramid v2	0.214	0.393	69.1
MA-RoMa plain fine-tune (λ = 0)	direct	0.236 ± 0.019	0.264 ± 0.019	49.6 ± 3.7
	pyramid v2	0.228 ± 0.027	0.264 ± 0.019	49.5 ± 3.1
MA-RoMa L2-SP (λ = 0.01)	direct	0.250 ± 0.000	0.268 ± 0.019	50.8 ± 8.1
	pyramid v2	0.245 ± 0.023	0.272 ± 0.019	50.2 ± 6.6

Across eight seeds the plain and L2-SP fine-tunes are statistically indistinguishable on every metric (overlapping ± SD), so the anchor is a zero-cost floor rather than a reliable improvement at this data budget. Both regress SR@20 modestly relative to zero-shot (0.393 → 0.26 / 0.27), within the resolution of a 28-pair test, while sharply improving in-distribution TEM median ED (~5× from ~321 px, seed-robust). Because each direct/pyramid pair shares a checkpoint, the verified wrapper does not stack on the fine-tune (SR@20 within ±0.01, median ED within a few px). The regression tracks forgetting of the untrained C103 SEM-SE↔LOM-height combination, of which one scene is unrecoverable at every seed and anchor. An earlier, non-retained draw of seeds 1–3 sat lower at direct SR@10 0.095 ± 0.021: run-to-run scatter can exceed seed-to-seed scatter at 28 pairs, a caution against over-reading any single-run SR number here.

A one-line weight anchor, tested across seeds

The forgetting is the only thing standing between this fine-tune and a net-positive intervention, so we tested the lightest-touch mitigation, L2-SP weight anchoring [12]. The identical decoder-only recipe gains a single penalty $(\lambda/2)\|\theta_{dec} - \theta_{dec}^{0}\|^2$ that pulls the decoder back toward its zero-shot initialisation, with no replay corpus, no architecture change, and no extra parameters. We swept $\lambda \in \{0, 0.01, 0.1, 1.0\}$,

where λ = 0 reproduces the plain fine-tune, selecting the step per run by minimum validation median ED and the λ by held-out *test*-split retention, because validation cannot see the C103 damage.

In a single seed-0 run the anchor looked like a clear win. At λ = 0.01 the recoverable C103 scene, the scene that run broke, dropped from 80/1220 px back to 16/24 px, close to zero-shot's ~12 px, while the other C103 scene stayed broken at every λ; heavier anchoring (λ = 0.1, 1.0) bought no further retention and cost both in-distribution gain and training stability, as the certainty head decalibrated into transient all-fail validation epochs whose onset moved earlier as λ grew. On that single run, L2-SP λ = 0.01 improved median ED over the plain fine-tune by 15.6 px (CI [−49.8, −2.4], $p = 0.0046$) and shrank the SR@20 deficit against zero-shot to −0.071 (CI [−0.214, +0.071], $p = 0.91$). Neither edge generalises: the eight-seed analysis below, under both the direct and the pyramid deployment protocols (Table 2), leaves L2-SP and the plain fine-tune indistinguishable.

Eight training seeds temper that reading. Re-running the plain (λ = 0) and anchored (λ = 0.01) fine-tunes across eight seeds with the identical recipe (Table 2; per-seed values for the five new seeds in Table 3) leaves the two configurations statistically indistinguishable on every test metric: direct-match SR@10 0.236 ± 0.019 (plain) versus 0.250 ± 0.000 (L2-SP), SR@20 0.264 ± 0.019 versus 0.268 ± 0.019, median ED 49.6 ± 3.7 versus 50.8 ± 8.1 px. The single-run median-ED edge does not survive as a cross-seed effect, and on the recoverable C103 scene the anchor is if anything less stable than the plain fine-tune, retaining it in one of three seeds against three of three for λ = 0. The one effect that is robust across all six runs is the opposite of an anchor success: the second C103 scene is unrecoverable at every seed and every λ, which marks a modality-coverage gap rather than an optimisation choice. We therefore report L2-SP as a deliberately zero-cost floor that does not worsen the fine-tune rather than as a reliable forgetting fix, and we expect stronger continual-learning methods such as experience replay, elastic weight consolidation [14], or LoRA-style low-rank updates [13] to be needed for a seed-robust improvement; quantifying that is left to future work. With only 28 test pairs the SR@20 comparisons are low-power, so the regression should be read as "modest and within seed noise" rather than as proven equivalence, and a formal equivalence test on a larger held-out set is planned. Evaluating both deployment protocols on the same per-seed checkpoints (Table 2) shows the verified wrapper neither helps nor hurts the fine-tuned backbone on the real test split, with pyramid and direct SR@20 coinciding at 0.264 (plain) and within 0.004 for L2-SP (0.268 direct, 0.272 pyramid); this is consistent with the FOV-ladder result that scale and appearance failures overlap in the real distribution, so here the wrapper and the fine-tune do not stack. Lifting SR@20 *above* zero-shot additionally requires closing the unrecoverable-modality gap, which needs training coverage of that modality combination and not a better optimiser.

Table 3. Per-seed direct-match results for the five new training seeds (seeds 4–8; 28-pair test split, TPS-refined ED) underlying the eight-seed aggregate in Table 2; seeds 1–3 of the same retrained draw contribute to that aggregate but predate this run.

Config	Seed	SR@10	SR@20	med ED (px)
plain fine-tune (λ = 0)	4	0.250	0.286	48.3
plain fine-tune (λ = 0)	5	0.214	0.250	47.5
plain fine-tune (λ = 0)	6	0.250	0.250	49.3
plain fine-tune (λ = 0)	7	0.214	0.286	47.2
plain fine-tune (λ = 0)	8	0.214	0.250	47.1
L2-SP (λ = 0.01)	4	0.250	0.250	70.2
L2-SP (λ = 0.01)	5	0.250	0.286	50.2
L2-SP (λ = 0.01)	6	0.250	0.250	43.4
L2-SP (λ = 0.01)	7	0.250	0.250	50.6
L2-SP (λ = 0.01)	8	0.250	0.286	47.8

Across the five new seeds SR@10 spans 0.214–0.250 and SR@20 0.250–0.286 for both configurations, overlapping rather than separating. Median ED differs in scatter rather than central tendency (plain 47.1–49.3 vs. L2-SP 43.4–70.2 px, the L2-SP range widened by a single high-residual seed), so no per-seed separation supports calling L2-SP an improvement.

Hypothesis verdicts

H1 (pyramid ≥ 35 % gain at FOV ≤ 5 %) is **rejected** on real pairs, because severe-FOV success is zero for every configuration and those pairs fail on appearance rather than scale; the wrapper mechanism is nonetheless **validated** independently by the controlled FOV ladder. **H2** (RoMa beats the ELoFTR family at low FOV) is **supported**. **H3** (affine sufficient vs. homography) is **mostly supported**: the BIC selector picks affine on 69 % of well-registered pairs and homography on the remaining 31 %, so the correct protocol is automatic selection rather than a hard affine restriction.

Discussion

The conclusion we draw, for this benchmark, is that cross-modal appearance rather than scale mismatch is the dominant constraint on foundation-model registration. Three lines of evidence point to it. First, the only off-the-shelf intervention that produced a significant headline gain was a change of *weights* (cross-modal-trained MatchAnything-RoMa), and that gain lived entirely among high-FOV, nearly-matchable pairs. Second, a scale-only wrapper that triples success when scale is isolated experimentally moves real severe-FOV pairs almost not at all, because in the real distribution low FOV and high appearance divergence are confounded. Third, attacking appearance directly with domain fine-tuning produced the single largest error reduction we observed (~5× on in-distribution TEM; 5.2× in the seed-0 run, and seed-

robust in direction across the three seeds). The pyramid solves a problem these pairs mostly do not have, and it cannot touch the one they do.

We are deliberately careful about the asymmetry in this evidence. Scale is decoupled by a direct, controlled experiment, the FOV ladder. Appearance is implicated by elimination rather than by an equivalent “scale fixed, appearance swept” manipulation, and the FOV ladder itself runs only on base-matchable pairs, which are already selected to have tractable appearance. The appearance conclusion is therefore weaker than the scale conclusion, and we present it as the dominant *uncontrolled* factor on AmalgaMatch rather than as a demonstrated law of cross-modal microscopy. A controlled appearance sweep, for example holding geometry fixed while degrading mutual information, is the natural complement to the FOV ladder and is left to future work.

This sharpens, rather than contradicts, the picture from the AmalgaMatch authors’ own foundation-model evaluation [1]. They establish that a macroscale-trained cross-modal matcher generalises surprisingly well to microscopy and that severe FOV, low mutual information, and dislocation imaging remain hard. We add the causal decomposition: of those three, FOV is addressable by a training-free wrapper once it is the isolated variable, whereas the low-mutual-information and dislocation cases are appearance failures that only domain data repairs.

The practical implication for correlative-microscopy pipelines is a clear ordering. (i) Choose a cross-modal-trained backbone: MatchAnything-RoMa is the strongest off-the-shelf option and loads into a stock roma_outdoor pipeline unchanged. (ii) Fit transforms with automatic affine-vs-homography selection rather than a hard affine restriction (affine wins 69 % of pairs, homography the rest). (iii) Add a decoder-only domain fine-tune when in-domain data exists, accepting that it helps in-distribution modalities and risks forgetting untrained ones; in our setup this is ~100M decoder parameters, AdamW for 1500 steps, about 90 min on a single RTX 5090, and a 445 MB checkpoint, optionally with an L2-SP anchor at $\lambda = 0.01$ as a zero-cost regulariser (it does not, on its own, prevent loss of untrained modality combinations; see below). (iv) Apply the verified wrapper only for genuinely multiscale tasks, where it adds a real scale lift without regressing the rest. A practical caveat on coverage: the fine-tune’s benefit is confined to modality combinations represented in training (here TEM dislocation imaging and several EBSD subclasses), and combinations absent from training (the C103 SEM-SE↔LOM-height case) should be run with the zero-shot backbone, or the fine-tune re-trained with that combination included, until coverage is broadened.

Several limitations bound these claims. The mutual-information verifier cannot rank transforms within ~10 px of one another, which both caps the wrapper’s benefit on strong backbones and prevents the iterated-zoom machinery from paying off; a learned or GT-free residual verifier is the obvious next lever. The ~10.3 px GT affine residual floors all px-threshold metrics, so AmalgaMatch cannot certify sub-5 px registration and our SR@5 numbers should be read with that floor in mind. The FOV ladder cleanly isolates scale, but it crops real targets rather than acquiring genuinely low-FOV images, so it does not reproduce the texture, detector-noise, and resolution differences of native low-FOV acquisition; its

wrapper gains should be read as a controlled lower bound on the mechanism, not as a literal forecast for natively low-FOV pairs. Both the direct-match and the pyramid-protocol fine-tuning numbers are now reported as mean ± SD over eight training seeds, evaluated on the same per-seed checkpoints (Table 2). An earlier three-seed draw whose checkpoints were not retained sat well below this, at direct SR@10 0.095 ± 0.021; decoder fine-tuning is not bit-reproducible at fixed seed on our stack (non-deterministic CUDA kernels and multi-worker data loading), and that an independent draw moves the SR@10 mean from 0.095 to ~0.24 shows run-to-run scatter exceeding seed-to-seed scatter at 28 test pairs — itself a caution against over-reading any single-run SR number on this benchmark. The reliable comparison is the within-checkpoint one (direct versus pyramid in Table 2), which shows the recommended deployment configuration, the verified wrapper on the fine-tune, neither gains nor loses over direct matching across seeds. The 28-pair test split is small, so SR comparisons are low-power and the seed-to-seed scatter is correspondingly large, which is why we read the fine-tune as a large in-distribution gain accompanied by a modest, seed-noisy SR regression rather than a clean net improvement, and why we do not claim the L2-SP anchor as a reliable fix. A formal equivalence test on a larger held-out set and at least one stronger continual-learning baseline (LoRA-style low-rank adaptation) remain planned and not yet run. Finally, the fine-tuning study rests on 131 training pairs over 12 subclasses; the unrecoverable C103 scene, broken at every seed and anchor strength, shows that breadth of modality coverage rather than optimisation is the remaining bottleneck. The concrete next experiment is to add training pairs in that modality combination and re-evaluate on the same held-out split and FOV ladder used here. We expect a forgetting-robust fine-tune over broad modality coverage, with the verified wrapper reserved for multiscale tasks, to be the deployable configuration for cross-modal microscopy registration.

Methods

Dataset and ground truth. AmalgaMatch [2] provides 187 correlative-microscopy image pairs across six matching tasks and 19 material subsets, each with GT point correspondences and per-image physical pixel size where available. We use all pairs; no pair is excluded from the zero-shot/wrapper evaluation. FOV strata are defined by the GT-implied source/target area ratio. Because within-scene pairs can share source images, all train/val/test partitioning for the fine-tuning experiments is performed at the scene level to prevent image leakage.

Registration pipeline and metrics. For each pair the matcher returns correspondences, which are fed to MAGSAC++ [8] (a marginalising robust estimator in the RANSAC [17] lineage that replaces a hard inlier threshold with marginalisation over noise scale) with a 5.5 px reprojection threshold. Both a 6-DoF affine and an 8-DoF homography are fit and the model is selected by a BIC-style penalty on inlier residuals; the selected transform is refined on its inliers with a thin-plate spline [9]. Accuracy is the mean Euclidean distance between GT target points mapped into source coordinates and their GT source positions. Reported aggregates are median ED and success rates SR@{5,10,20} px over pairs. Classical controls are SIFT [10] with the same MAGSAC++ fitting (Control A) and a mutual-information intensity registration [11] (Control B).

Statistics. All method-vs-method comparisons use a paired bootstrap over per-pair errors (B = 10,000 resamples) with 95 % confidence intervals on the difference of the relevant aggregate; reported p-values are two-sided bootstrap probabilities.

Backbones and weights. We evaluate SIFT, LoFTR [3], the MatchAnything-ELoFTR [4,7] configuration, RoMa [5], and MatchAnything-RoMa [7]. RoMa couples frozen DINOv2 features [6] with a fine-feature ConvNet and a match-classification decoder. The released MatchAnything-RoMa checkpoint is key-for-key compatible with the roma_outdoor architecture (603/603 tensors), so it is loaded into the identical pipeline without modification.

Pyramid wrappers. Pyramid v1 tiles the source into a sliding-window pyramid (50 % overlap) whose levels match the target's nominal physical resolution, matches each tile against the target, back-projects and pools all correspondences, then fits one global transform. Pyramid v2 (Fig. 1a) is verified coarse-to-fine: it computes a direct match and incumbent transform first, then, only on weak direct support, evaluates candidate stages (tile search over the scaled source grid, then zoom refinement on the best-scoring region) and accepts a candidate only if a mutual-information-on-overlap verifier improves on the incumbent, making the wrapper monotone with respect to the direct fit. Ablations cover iterated vs. single zoom and certainty gating at 0.5.

FOV-ladder construction. For each backbone we take its base-matchable pairs (direct mean ED < 20 px) and crop the target to absolute source/target area ratios {0.5, 0.25, 0.1, 0.05, 0.02}, holding appearance, modality, and pixel size fixed. Full GT is retained for evaluation so that out-of-crop correspondences test extrapolation of the fitted global transform. Reported metrics are transform-based mean ED with the same fitting and bootstrap protocol. The fine-tuned-backbone ladder is restricted to a fixed 63-pair testbed so that direct rows do not expand the eligible set.

Domain fine-tuning and L2-SP. Fine-tuning is decoder-only: the VGG+DINOv2 encoder is held in evaluation mode (frozen, with normalisation statistics fixed) while ~100M decoder parameters are trained. Supervision densifies sparse GT correspondences into an A→B warp via thin-plate splines and supervises inside the GT-support region under a sparse-GT-robust loss, with FOV-crop and photometric augmentation. Optimisation uses AdamW at 2×10^{-5} for 1500 steps; the checkpoint is selected by minimum validation median ED. The weight-anchored variant adds an L2-SP penalty [12] $(\lambda/2)\|\theta_dec - \theta_dec^{0}\|^2$ anchoring the decoder to its zero-shot initialisation; $\lambda \in \{0, 0.01, 0.1, 1.0\}$ is swept, with the step selected per run by validation median ED and λ by held-out test-split retention. The plain ($\lambda = 0$) and anchored ($\lambda = 0.01$) recipes are each run at eight seeds for the variance analysis. We considered replay-based and low-rank [13] alternatives and elastic weight consolidation [14] as heavier mitigations; L2-SP was chosen as the zero-cost baseline.

Declarations

Data availability

The AmalgaMatch dataset is publicly available from the Fraunhofer Fordatis repository (doi:10.24406/fordatis/436) [2]. The analysis code, evaluation harness, and result-regeneration scripts are openly available at https://github.com/fronkt/correlative-microscopy-alignment and archived at Zenodo (doi:10.5281/zenodo.20819649); all tables and figures regenerate from the released result files via the scripts referenced therein. Fine-tuned model checkpoints (445 MB) are available from the author on reasonable request.

Code availability

All code is released under the repository above, with a single-command evaluation entry point and a synthetic-pair acceptance test.

Acknowledgements

The author thanks the AmalgaMatch authors for releasing the dataset and the developers of RoMa, MatchAnything, Efficient LoFTR, and MAGSAC++ for releasing their code and weights.

Author contributions

F.C. conceived the study, implemented the methods, conducted the experiments, analysed the results, and wrote the manuscript.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Durmaz, A. R., Lamb, J. D., Echlin, M. P. & Pollock, T. M. Foundation models for multimodal image data fusion in materials science. *Frontiers in Materials***13** (2026). doi:10.3389/fmats.2026.1815017.
2. Durmaz, A. R., Lamb, J. D., Echlin, M. P. & Pollock, T. M. AmalgaMatch: a benchmark dataset for cross-modal image matching in correlative materials microscopy. Fordatis, Fraunhofer Research Data Repository (2026). doi:10.24406/fordatis/436.

3. Sun, J., Shen, Z., Wang, Y., Bao, H. & Zhou, X. LoFTR: Detector-free local feature matching with transformers. In *Proc. IEEE/CVF CVPR*, 8922–8931 (2021).
4. Wang, Y., He, X., Peng, S., Tan, D. & Zhou, X. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proc. IEEE/CVF CVPR* (2024).
5. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M. & Felsberg, M. RoMa: Robust dense feature matching. In *Proc. IEEE/CVF CVPR*, 19790–19800 (2024).
6. Oquab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* (2024).
7. He, X. *et al.* MatchAnything: Universal cross-modality image matching with large-scale pre-training. *arXiv:2501.07556* (2025).
8. Barath, D., Noskova, J., Ivashechkin, M. & Matas, J. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proc. IEEE/CVF CVPR*, 1304–1312 (2020).
9. Bookstein, F. L. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.***11**, 567–585 (1989).
10. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.***60**, 91–110 (2004).
11. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. & Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging***16**, 187–198 (1997).
12. Li, X., Grandvalet, Y. & Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *Proc. 35th ICML*, PMLR **80**, 2825–2834 (2018).
13. Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. In *ICLR* (2022).
14. Kirkpatrick, J. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA***114**, 3521–3526 (2017).
15. Sarlin, P.-E., DeTone, D., Malisiewicz, T. & Rabinovich, A. SuperGlue: Learning feature matching with graph neural networks. In *Proc. IEEE/CVF CVPR*, 4938–4947 (2020).
16. Lindenberger, P., Sarlin, P.-E. & Pollefeys, M. LightGlue: Local feature matching at light speed. In *Proc. IEEE/CVF ICCV*, 17627–17638 (2023).
17. Fischler, M. A. & Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM***24**, 381–395 (1981). *(Reference numbering in this Markdown render follows insertion order; the LaTeX/BibTeX build renumbers strictly by order of first appearance.)*

Figures

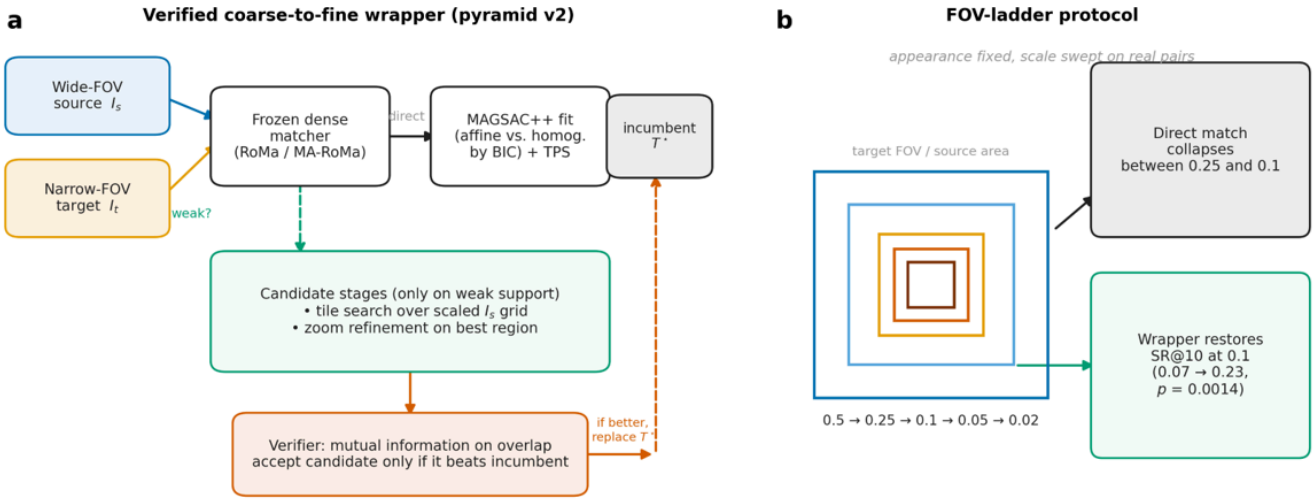


Figure 1

Method overview. (**a**) The verified coarse-to-fine wrapper (pyramid v2): a direct match yields an incumbent transform; candidate tile-search and zoom stages are invoked only on weak support and accepted only when a mutual-information-on-overlap verifier beats the incumbent, making the wrapper monotone with respect to the direct fit. (**b**) The FOV-ladder protocol crops real base-matchable pairs to shrinking field-of-view ratios with appearance, modality, and pixel size fixed, isolating scale as the only variable.

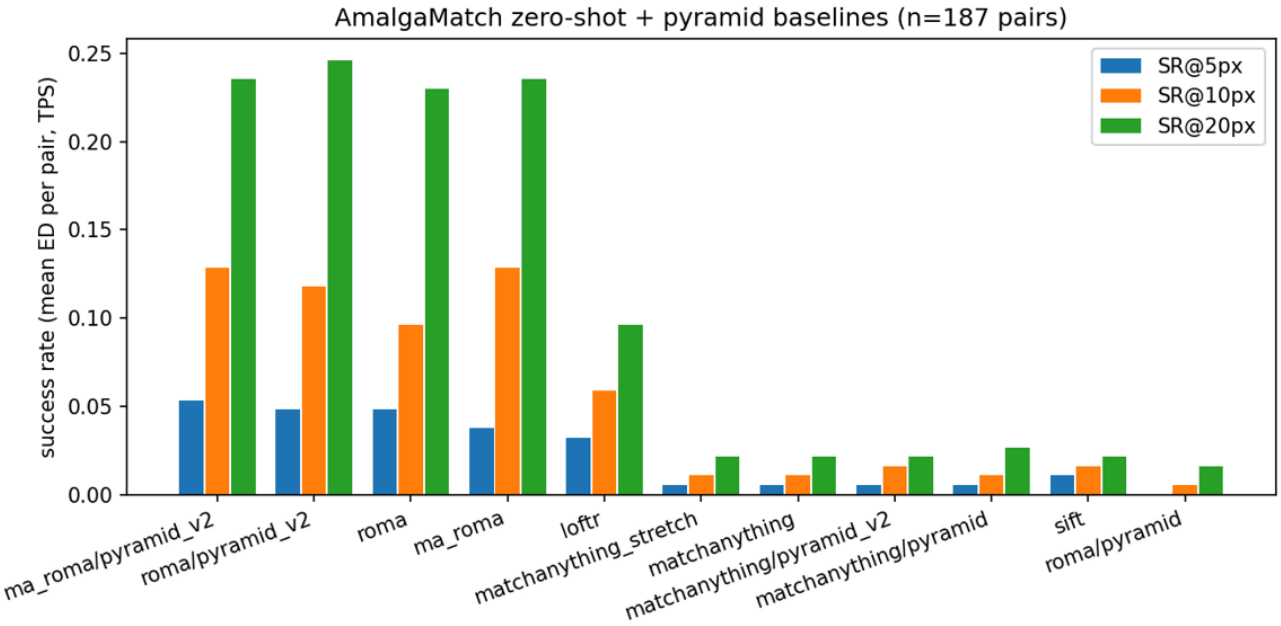


Figure 2

Success rates across all configurations on AmalgaMatch. Pyramid v1 collapses RoMa; pyramid v2 recovers a small significant gain without losing pairs; the MatchAnything-RoMa backbone provides the

only significant headline gain over the zero-shot bar.

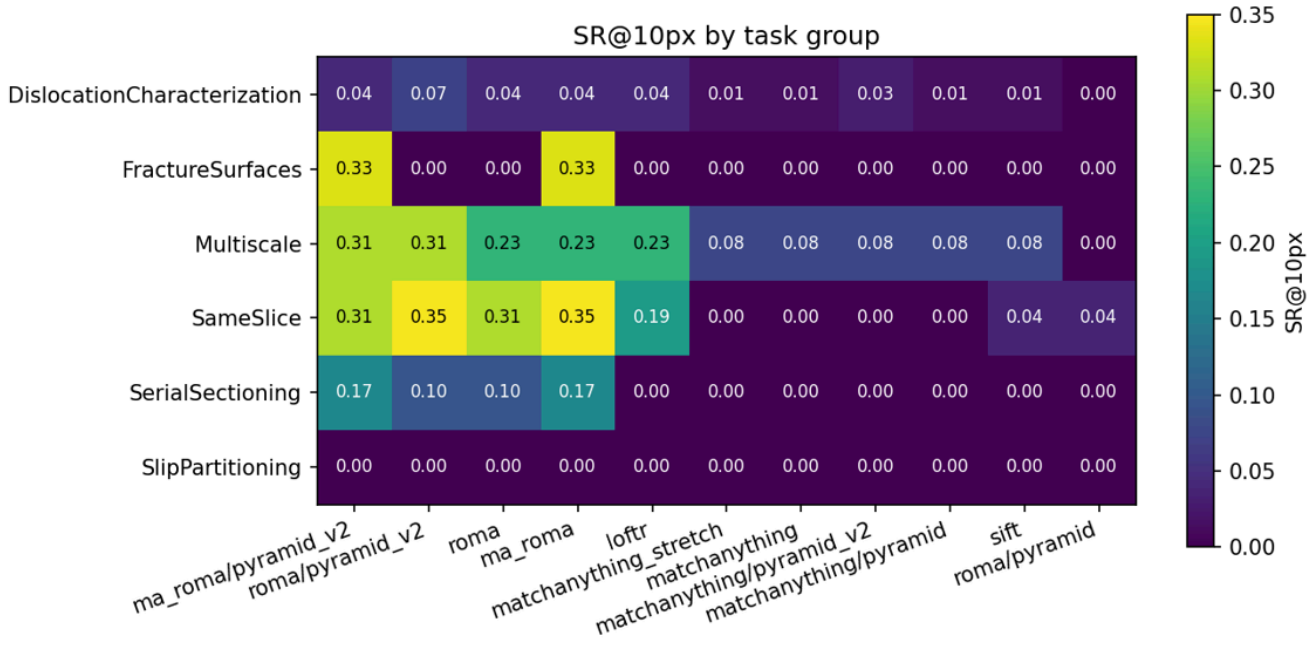


Figure 3

Per-task-group success rates. Performance is concentrated in orientation-mapping and serial-sectioning groups; dislocation characterisation and low-mutual-information pairs remain hard for every configuration, consistent with appearance (not scale) being the dominant constraint.

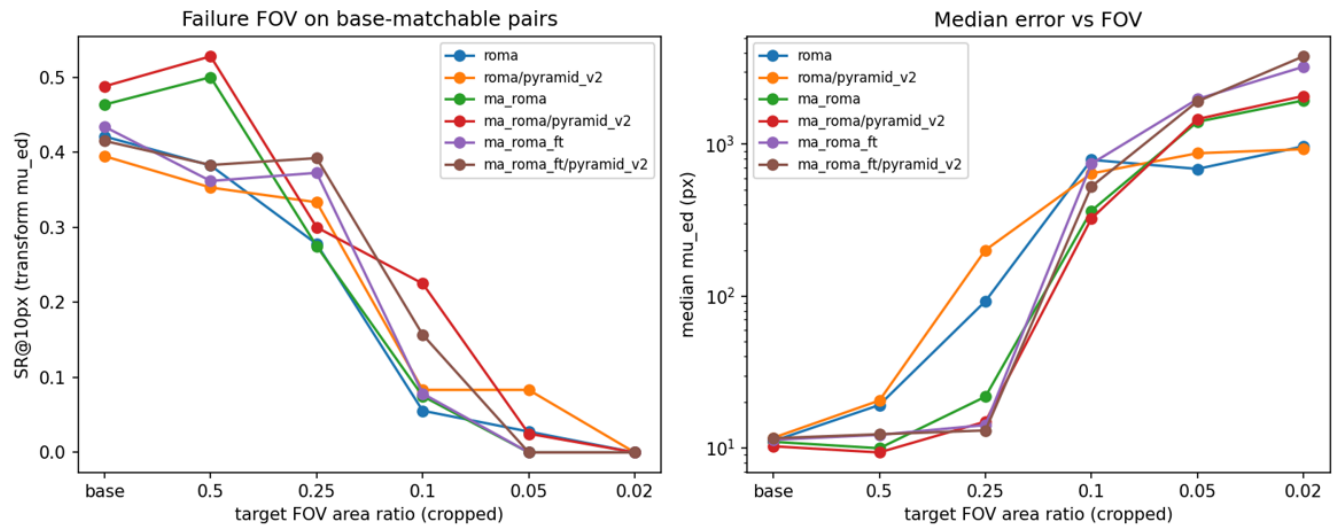


Figure 4

Controlled FOV ladder with appearance fixed. Cropping base-matchable pairs to shrinking FOV ratios shows direct matching collapsing between 0.25 and 0.1, while the verified wrapper restores SR@10 at the 0.1 rung (0.07 → 0.23, p = 0.0014): the wrapper's scale mechanism is sound once scale is isolated.

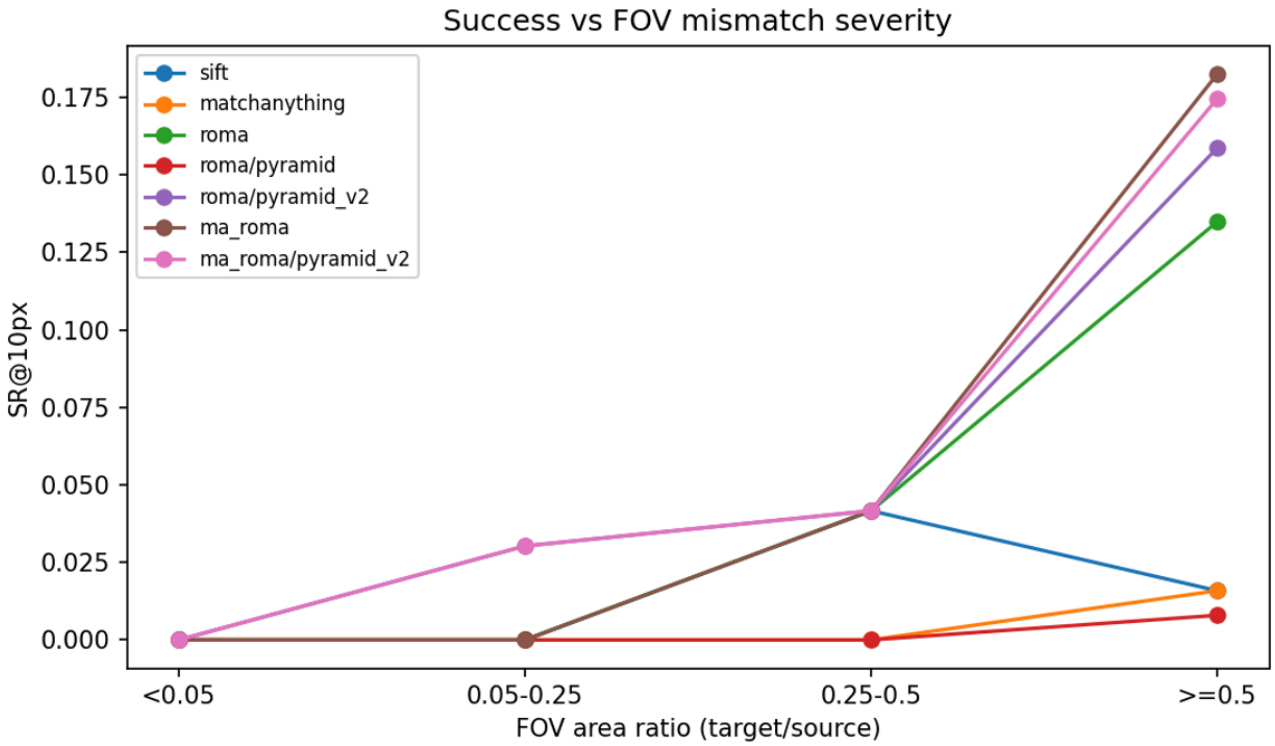


Figure 5

Success rate versus field-of-view stratum on native pairs. On the real distribution, low-FOV strata are simultaneously the most appearance-divergent, confounding scale and appearance; severe-FOV success is at the floor for every configuration.