Cross-Detector Descriptor Fusion

Scale Control and Spatial Alignment for Local Feature Matching

Frank Sossi | School of STEM, University of Washington Bothell

Committee: Prof. Clark Olson (Chair), Prof. Min Chen, Prof. Dong Si

2026

Thesis in One Sentence

Selecting high-quality keypoints through detector consensus and scale filtering, then fusing complementary descriptors with proper magnitude matching, yields large improvements in local feature matching performance.

Outline

Framing

Process

Results

Lessons Learned

The Problem

- Local feature matching is fundamental to SLAM, structure from motion, image retrieval, and visual place recognition [6]
- Two stages: **detection** (find salient locations) and **description** (encode local appearance)
- Different descriptor families have **complementary strengths** — but combining them is not straightforward

Opportunity

Can we systematically combine descriptors and improve keypoint selection to achieve better matching than any single method alone?

State of the Art When This Work Began

What existed:

- SIFT [6]: gradient histograms, 128-D
- SURF [4]: Haar wavelets, 64-D
- HardNet [7]: CNN, triplet-loss trained
- SOSNet [10]: second-order similarity CNN
- HoNC [8]: color normal histograms

What was missing:

- Systematic study of **cross-family fusion**
- Understanding of **when fusion helps vs. hurts**
- Role of **keypoint quality** (scale, detector agreement)
- Color-capable **patch benchmark** for fair comparison

Key Concepts

Detection → Description → Matching → Evaluation

Keypoint
A salient image location (corner, blob) likely to be re-detected under viewpoint or lighting change

Descriptor
A fixed-length vector encoding the appearance around a keypoint — used to find correspondences via nearest-neighbor search

mAP
Mean Average Precision — our primary metric. Higher = better matching. Evaluated per Bojanic et al. [3]

Research Goals & Measures of Success

RQ1 Does detector consensus provide a keypoint quality signal?
Success: intersection keypoints outperform full sets at matched count

RQ2 Can color descriptors improve fusion?
Success: HoNC+CNN $>$ CNN alone on color patch benchmark

RQ3 What compatibility patterns govern descriptor fusion?
Success: identify when fusion helps vs. hurts, and why

RQ4 How does keypoint scale impact performance?
Success: quantify the relationship between scale and mAP

Evaluated on HPatches benchmark [1]: 116 sequences, 59 viewpoint + 57 illumination, ground-truth homographies.

Benefits & Beneficiaries

Research community:

- Systematic fusion compatibility rules
- Evidence that keypoint quality $\geq$ descriptor choice
- Color HPatches benchmark (new resource)
- Discriminator–Matcher framework for predicting fusion outcomes

Practitioners:

- Concrete recipes: which descriptors to fuse, and how
- Scale filtering as a free performance boost
- Open-source DescriptorWorkbench framework
- Applicable to SLAM, SfM, visual localization

Key Decision: Two Evaluation Pipelines

Full-Image Pipeline

Detection $\rightarrow$ Description $\rightarrow$ Matching

Tests detector **and** descriptor jointly. Used for scale control and intersection experiments.

Patch Benchmark Pipeline

Pre-extracted patches $\rightarrow$ Description

Holds keypoint quality **constant**. Isolates descriptor fusion effects.

Why two pipelines? Fusion experiments on full images confound keypoint quality with descriptor quality. The patch pipeline lets us study fusion in isolation [1].

DescriptorWorkbench Architecture

C++ framework with:

- 10 descriptor types (SIFT, RGBSIFT, HoNC, DSP-SIFT, HardNet, SOSNet, ...)
- YAML-driven experiment configuration
- SQLite database for 100+ experiments
- Three metrics from Bojanic et al. [3]: matching, verification, retrieval

Tech Stack

- OpenCV 4.13 + LibTorch 2.10
- CUDA 13.1 (RTX 4090)
- Google Test for unit tests
- Python for analysis

Open-source on GitHub

Key Decision: Building a Color Patch Benchmark

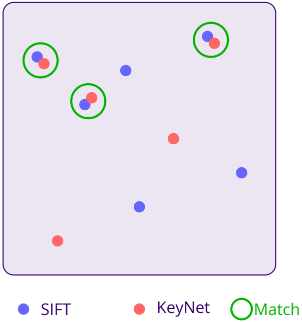
- Original HPatches provides only **grayscale** 65×65 patches [1]
- Color descriptors (HoNC, RGBSIFT) **cannot be evaluated** on grayscale data
- We re-extracted color patches from original images using stored keypoint locations + ground-truth homographies
- Validation: SIFT baseline 22.9% mAP (vs. 25.47% original grayscale)

Impact

Without this, we could not answer RQ2 (color descriptor fusion) at all.

Spatial Intersection Algorithm

1. Detect keypoints with SIFT detector
2. Detect keypoints with KeyNet detector [2]
3. Find **spatial matches** within tolerance (*r* pixels)
4. Keep only agreed-upon locations
5. Describe with any descriptor



Intuition: If two very different detectors agree a location is interesting, it is likely a *high-quality* feature.

Scale Characteristics of Detectors

SIFT Detector [6]

- Avg scale: **4.45 px**
- Many small-scale keypoints
- $\sim$2.5M total keypoints

KeyNet Detector [2]

- Avg scale: **49.83 px**
- Larger, more distinctive regions
- $\sim$2.8M total keypoints

Key insight: Larger keypoint scale $\Rightarrow$ more informative patches $\Rightarrow$ better descriptors.

A 4 px keypoint samples $\sim$16$\times$16 pixels; a 10 px keypoint samples $\sim$40$\times$40 pixels.

Same-Family Fusion Does Not Help

Hypothesis: SIFT + RGBSIFT fusion should add color information to grayscale SIFT.

Result (patch benchmark):

Configuration	mAP
RGBSIFT alone	24.6%
SIFT alone	22.9%
SIFT + RGBSIFT concat	10.4%

Why it failed

SIFT-family descriptors capture **correlated** gradient histogram information — even when one uses color channels. Fusion doubles dimensionality, adding noise rather than signal.

Cross-Family Fusion Requires Magnitude Matching

Problem: SIFT + HardNet fusion performs **worse** than either descriptor alone.

Root cause:

- SIFT values: 0–512
- HardNet values: -0.3 to $+0.3$
- In L2 distance, SIFT **dominates**

	Raw	**After L2**
SIFT	[0, 512]	[0, 0.3]
HardNet	[$-$0.3, 0.3]	[$-$0.3, 0.3]
HoNC	[0, 1]	[0, 0.3]

Solution: Pre-Fusion L2 Normalization

Solution: L2-normalize each descriptor component *before* fusion:

$$d_{\text{fused}} = \text{fuse}\left(\frac{d_A}{\|d_A\|_2}, \frac{d_B}{\|d_B\|_2}\right)$$

Without normalization

SIFT + HardNet: **failed**

SIFT magnitudes dominate distance

With normalization

SIFT + HardNet: **46.0% mAP**

Equal contribution from each component

This led us to implement `normalize_before_fusion` as a configurable option in the framework.

Keypoint Quality Matters More Than Descriptor Choice

- Descriptor research focuses on better **encoding algorithms**
- Yet **keypoint selection** has an equal or greater effect on performance

What changed	mAP gain
Better descriptor (SIFT $\rightarrow$ HardNet)	+20%
Better keypoints (scale filter on SIFT)	+18%
Better keypoints (intersection on HardNet)	+18%

Keypoint quality improvements are comparable to switching descriptor families entirely [6, 7].

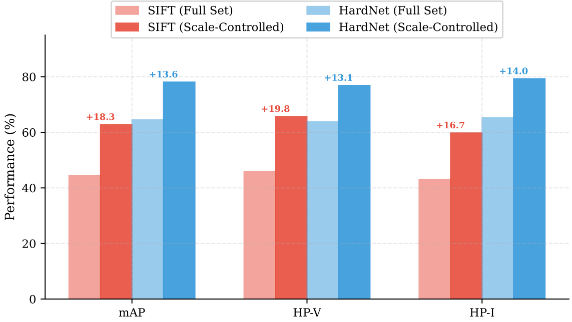
Baseline Performance

Descriptor	Detector	mAP	HP-V	HP-I
SIFT	SIFT	44.5%	45.9%	43.1%
RootSIFT	SIFT	46.7%	46.2%	47.2%
HardNet	KeyNet	64.5%	63.8%	65.3%
SOSNet	KeyNet	64.3%	63.4%	65.2%

- Learned descriptors outperform SIFT by $\sim$20% mAP
- SIFT slightly favors viewpoint; CNN slightly favors illumination
- These are our baselines — all improvements measured from here

Evaluated on HPatches [1] with metrics from Bojanic et al. [3]

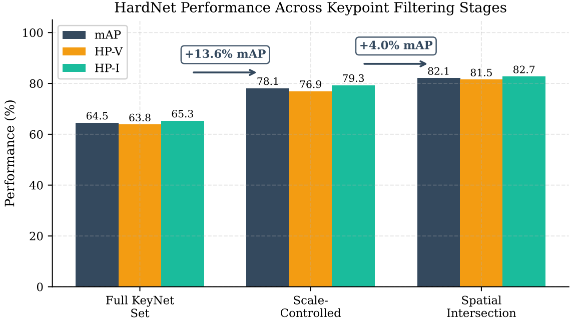
Result: Scale Control Impact



- SIFT: 44.5% $\rightarrow$ **62.8%** mAP
  +18.3% absolute
- HardNet: 64.5% $\rightarrow$ **78.1%** mAP
  +13.6% absolute
- Filter: keep top 25% by scale
- **Quality over quantity**

Answers **RQ4**: scale has a large, consistent impact across descriptor families [6].

Result: Detector Intersection Progression



HardNet mAP across stages:

1. Full KeyNet set: 64.5%
2. Scale-controlled: 78.1%
3. Spatial intersection: **82.1%**

Detector consensus provides quality signal **beyond** scale alone.

Answers **RQ1**: Yes, intersection keypoints are more distinctive keypoints [2].

Validating the Intersection Mechanism (1/2)

The +18% mAP gain required verification. We ruled out four alternatives:

✗ Not better repeatability

Intersection: 28.0% vs Scale: 29.4% ($p < 0.0001$). Less repeatable, yet better matches.

✗ Not more distinctive descriptors

Correct-match NN ratios identical ($\sim$0.44).
Descriptors equally good; locations differ.

✗ Not higher response values

Both sets: avg response $\sim$0.035.
Keypoint strength is identical.

✗ Not spatial crowding

Quadrant distributions identical:
TL 13%, TR 18%, BL 24%, BR 45%.
Filtering is uniform across image.

Validating the Intersection Mechanism (2/2)

Confirmed: Detector consensus removes “confusing” keypoints at repetitive textures.

✓ Higher precision

At threshold 0.8: 71.2% vs 66.2%

✓ Fewer false positives

0.40 vs 0.51 FP per TP (22% reduction)

✓ All descriptors benefit equally

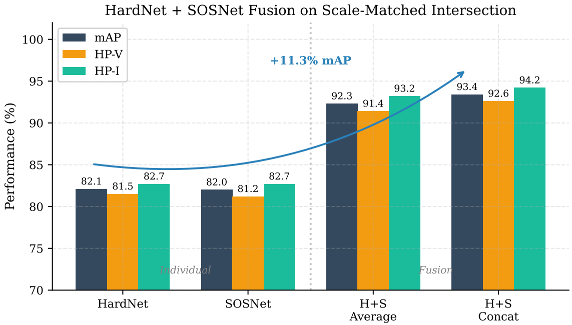
Proves it’s **location quality**.

Multi-descriptor gains over baseline:

DSPSIFT	+28%
RGBSIFT	+31%
HoNC	+32%
HardNet	+17%
SOSNet	+17%

Mechanism: Locations where both SIFT and KeyNet detect keypoints have unique local structure.

Result: CNN + CNN Fusion

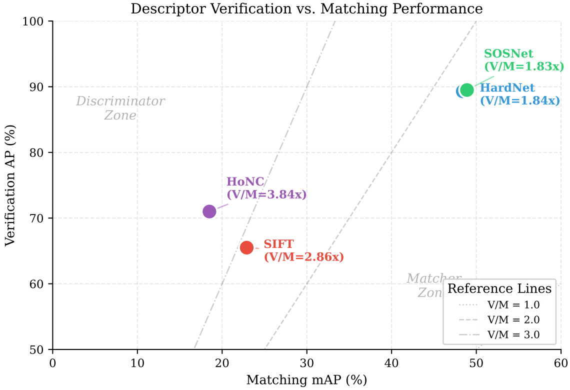


On intersection keypoints:

- HardNet alone: 82.1%
- SOSNet alone: 82.0%
- Concatenation: **93.4%**
  +11.3% absolute
- Averaging: 92.3%

HardNet [7] and SOSNet [10] learn complementary representations despite similar training.

The Discriminator--Matcher Framework

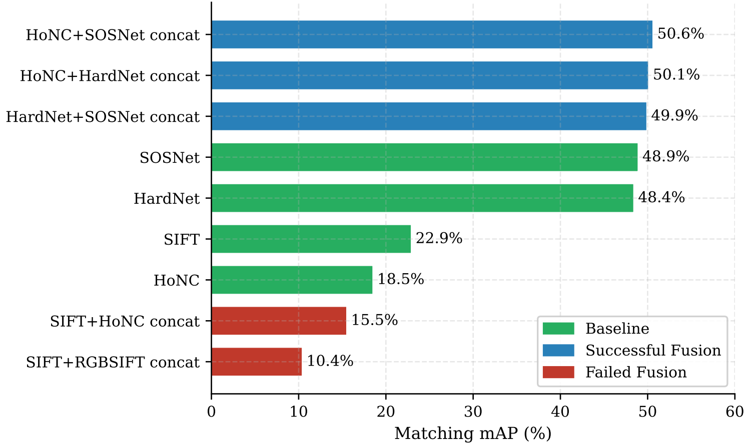


V/M Ratio = Verification / Matching:

- HoNC [8]: 3.84× **(discriminator)**
  Good at rejecting false matches
- HardNet: 1.84× **(matcher)**
  Trained for correspondence

Prediction: pairing a discriminator with a matcher yields the best fusion.

Result: Patch Benchmark Fusion



- **Best:** HoNC + SOSNet concat = **50.6%** mAP (color + learned)

Answering the Research Questions

RQ1 **Yes** — detector intersection improves HardNet by +18% mAP. Consensus keypoints are more repeatable.

RQ2 **Yes** — HoNC + SOSNet (50.6%) outperforms SOSNet alone (48.9%). Color adds complementary discrimination.

RQ3 **Complementarity determines success.** Discriminator + Matcher works; similar + similar fails. Cross-family requires magnitude matching.

RQ4 **Scale is a dominant factor.** +18% for SIFT, +14% for HardNet with top-25% filtering. Comparable to switching descriptor families entirely.

What I Learned: Technical Insights

1. **Keypoint quality deserves more attention**
   The 39% gain from scale control and 25% from intersection exceed most algorithmic advances, yet these strategies are rarely discussed in the literature
2. **Failure is informative**
   The magnitude mismatch discovery came from a “failed” fusion experiment — investigating *why* something fails is as valuable as demonstrating success
3. **Two pipelines prevent false conclusions**
   Full-image experiments confound detector effects with descriptor effects — the patch benchmark was essential for clean fusion analysis

What I Learned: Engineering & Process

1. **Experiment infrastructure pays off**
   SQLite tracking + YAML configs + automated metrics enabled running 100+ experiments systematically
2. **Reproducibility requires tooling**
   Building DescriptorWorkbench took significant effort, but every result in the thesis can be reproduced from a single YAML file
3. **Data analysis reveals what code cannot**
   The V/M ratio framework emerged from plotting verification vs. matching — a pattern invisible in raw numbers

Future Work

Short-term:

- Learned fusion weights (attention-based, per-dimension)
- Tolerance sensitivity analysis for intersection radius
- Additional descriptors (DISK, ALIKE, SuperPoint)

Long-term:

- End-to-end learned detect + describe + fuse pipeline
- Validation on MegaDepth [5], Oxford5k [9]
- Real-time deployment (mobile SLAM)

Limitations

- **Single dataset:** All results on HPatches [1] — may not generalize to extreme viewpoint ($>$60°) or different domains
- **Computational overhead:** Concatenation doubles descriptor dimensionality (128-D $\rightarrow$ 256-D) — though keypoint filtering reduces total cost by $28\times$
- **Detector dependency:** Best results use KeyNet [2] for CNN descriptors — findings may not transfer to other detectors
- **Fixed fusion:** We use equal weighting ($\alpha = 0.5$) — learned weights could improve results further

Summary of Contributions

1. **Detector intersection** as quality filter: HardNet 82.1% mAP (+25% relative)
2. **Color HPatches benchmark**: enables color descriptor evaluation
3. **Complementary fusion**: HoNC + SOSNet = 50.6% mAP on patches
4. **Magnitude matching**: L2 normalization enables cross-family fusion
5. **Scale control**: +39% SIFT, +21% CNN with top-25% filtering
6. **DescriptorWorkbench**: open-source framework, 100+ experiments

Keypoint selection strategy can matter as much as descriptor algorithm choice.

Thank You

Questions?

Frank Sossi · University of Washington Bothell

https://github.com/F-Sossi/DescriptorWorkbench

References I

[1] Vassileios Balntas et al. “HPatches: A benchmark and evaluation of handcrafted and learned local descriptors”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. PDF in repo: research/Balntas_HPatches_A_Benchmark_CVPR_2017_paper.pdf. 2017, pp. 5173–5182.

References II

[2] Axel Barroso-Laguna et al. "Key. net: Keypoint detection by handcrafted and learned CNN filters". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. PDF in repo: re-
search/Key.Net_Keypoint_Detection_by_Handcrafted_and_Learned_CN
2019, pp. 5836–5844.

[3] Kristijan Bartol et al. "On the comparison of classic and deep keypoint detector and descriptor methods". In: *arXiv preprint arXiv:2007.10000* (2020). PDF in repo: research/On the Comparison of Classic and Deep\\nKeypoint Detector and Descriptor Methods.pdf.

References III

[4] Herbert Bay et al. "SURF: Speeded up robust features". In: *European conference on computer vision*. PDF in repo: research/ComputerVisionECCV2006_hasSurfPaper.pdf (SURF paper starts on p. 404 of proceedings). Springer. 2006, pp. 404–417.

[5] Zhengqi Li and Noah Snavely. "MegaDepth: Learning single-view depth prediction from internet photos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2041–2050.

References IV

[6] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60.2 (2004). PDF in repo: research/Lowe_2004_SIFT_IJCV.pdf, pp. 91–110.

[7] Anastasiia Mishchuk et al. “Working hard to know your neighbor’s margins: Local descriptor learning loss”. In: *Advances in Neural Information Processing Systems*. Vol. 30. PDF in repo: research/Mishchuk_2017_HardNet.pdf. 2017.

References V

[8] Clark F Olson and Siqi Zhang. "Keypoint recognition with histograms of normalized colors". In: *2016 13th Conference on Computer and Robot Vision (CRV)*. PDF in repo: research/Keypoint Recognition with Histograms of Normalized Colors.pdf. IEEE. 2016, pp. 311–318.

[9] James Philbin et al. "Object retrieval with large vocabularies and fast spatial matching". In: *2007 IEEE conference on computer vision and pattern recognition*. PDF in repo: research/Philbin_Object_Retrieval_CVPR_2007.pdf. IEEE. 2007, pp. 1–8.

References VI

[10] Yurun Tian et al. "SOSNet: Second order similarity regularization for local descriptor learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. PDF in repo: research/Tian_2019_SOSNet_CVPR.pdf. 2019, pp. 11016–11025.