西安電子科技大學
XIDIAN UNIVERSITY

SGAD: Semantic and Geometric-aware Descriptor for Local Feature Matching

Xiangzeng Liu[1*], Chi Wang[1*†], Guanglu Shi[1], Xiaodong Zhang[1], Qiguang Miao[1†], Miao Fan[2]

[1]Xidian University [2]Navinfo Europe B.V

Outline

1. Introduction
2. Method
3. Experiments
4. Discussion
5. Conclusion

Introduction

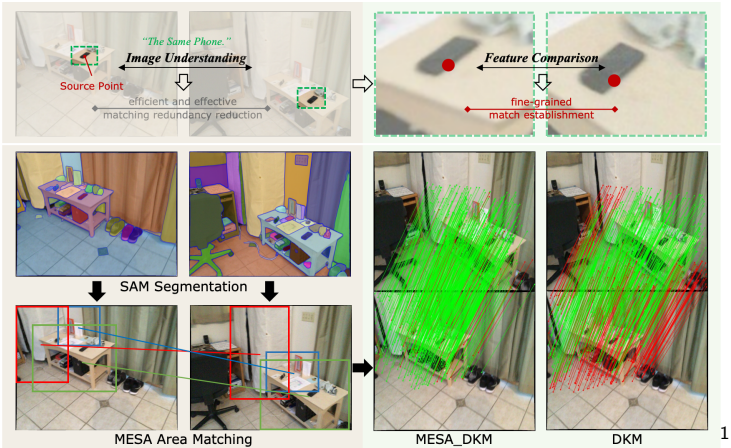
Why Local Feature Matching Matters

- Foundational for SfM, SLAM, visual localization, image retrieval, and many multi-view geometry tasks.
- Accurate pixel correspondences remain challenging under scale, viewpoint, illumination, and repetition changes.
- Achieving both high accuracy and high efficiency is essential for real-time, large-scale deployment.

Pixel-wise Matching Challenges

- Dense comparisons across entire images waste computation on irrelevant pixels.
- Highly sensitive to illumination changes, wide baselines, and repetitive structures, leading to unstable correspondences.
- Scaling to high-resolution or multi-image settings quickly becomes impractical.

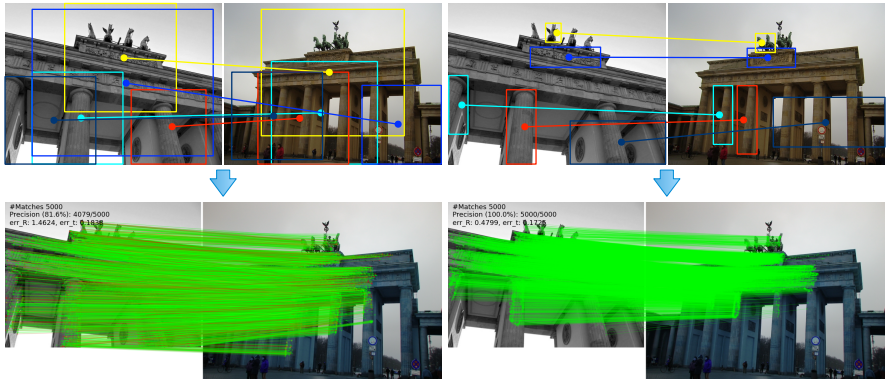
Opportunities for Area-to-Point Matching



- Segmenting overlapping regions first narrows the search space for point matchers.
- Region-level semantics deliver context that complements low-level descriptors.
- Motivates learning compact, reusable area representations.

[1]Illustration adapted from MESA [9].

Limitations of MESA



(a) MESA (b) SGAD

- **Inefficiency:** pixel-level activity maps and graph matching dominate runtime.
- **Inconsistency:** merged or mismatched regions reduce subsequent point-matching accuracy.
- Motivates a descriptor-driven, single-pass alternative.

Problem Statement

Core Question

Can we **learn discriminative, geometry-aware area descriptors** and accomplish **global-consistent matching in a single forward pass**?

- Avoid the efficiency bottlenecks of pixel-wise comparisons and graph optimization.
- Provide reliable region priors that strengthen downstream point matchers.

Overview and Contributions

SGAD in a Nutshell

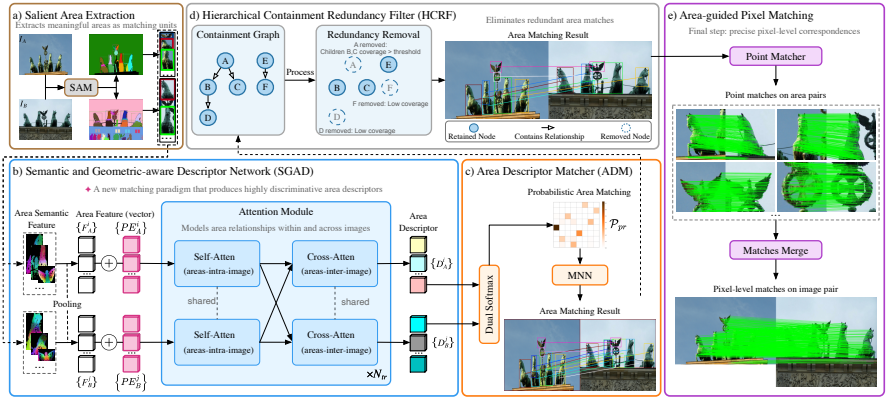
Area descriptors first: combine semantic cues and geometry to produce highly discriminative descriptors that a lightweight matcher can align efficiently.

Key Contributions

- Propose SGAD, an area descriptor network with alternating attentions and geometric positional encoding, enabling **single-pass** region matching.
- Introduce a **dual-task supervision** (classification + ranking) to jointly learn absolute and relative similarity.
- Design a **Hierarchical Containment Redundancy Filter (HCRF)** to prune overlapping areas; deliver consistent gains across matchers while being up to 60$\times$ faster than MESA.

Method

SGAD Overview



- SAM segments salient regions; DINOv2 provides semantic descriptors.
- Geometric positional encoding injects cross-region relations; alternating self/cross attention fuses intra- and inter-image context.
- Dual-softmax + MNN yield matches; HCRF removes redundant areas before local point matching.

Semantic and Geometric Area Representation

Semantic Features

- SAM generates instance masks; we take bounding boxes as region proposals.
- A frozen DINOv2-L backbone extracts token features, average-pooled into area vectors F^i.

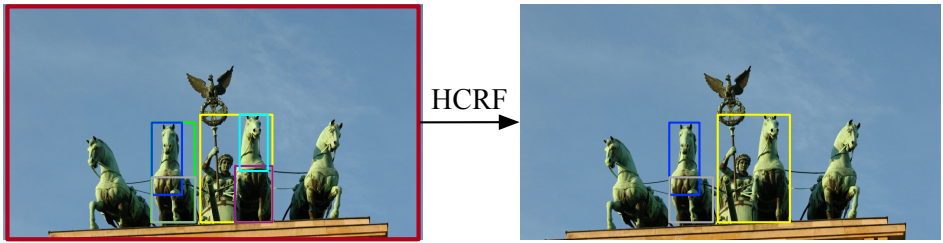
Geometric Positional Encoding

- Compute pairwise distances and angles between region centers to capture global relations.
- Feed statistics through an MLP and add to semantic features: $\hat{F}^i = F^i + PE^i$.
- Mitigates confusion between visually similar but spatially distinct regions.

Alternating Attention and Descriptor Matching

- **Alternating attention:** each layer performs intra-image self-attention followed by cross-image attention; $N_{tr} = 4$ layers output descriptors D_A^i, D_B^j.
- **Matching probabilities:** temperature-scaled dot products feed dual-softmax to obtain confidence matrix $\mathcal{P}_{pr}$.
- **Mutual nearest neighbor:** thresholding plus MNN enforce one-to-one matches and discard outliers.

Hierarchical Containment Redundancy Filter



- Build a containment graph with overlap and coverage thresholds to capture nested regions.
- Depth-first traversal decides whether to keep parents or children, removing redundant overlaps.
- Cuts downstream computations while preserving informative regions for point matching.

Dual-Task Supervision

Label Construction

- Use camera poses and depth to project regions across views; IoU provides ground-truth scores $\mathcal{P}_{gt}$.
- Threshold IoU to derive binary labels $\mathcal{P}_{gt}^{cls}$ for the classification branch.

Loss Functions

- **Classification:** focal loss on $\mathcal{P}_{pr}$ vs. $\mathcal{P}_{gt}^{cls}$ balances easy/hard and pos/neg pairs.
- **Ranking:** ListMLE ensures predicted scores respect ground-truth ordering for each source region.
- Joint objective $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rank}$ captures both absolute and relative similarity cues.

Experiments

Experimental Setup

- **Datasets:** ScanNet1500 (indoor), MegaDepth1500 (outdoor), HPatches (homography).
- **Training:** separate indoor/outdoor models following LoFTR splits; AdamW, lr 1×10^{-4}, batch size 64, trained 2 days on one RTX A6000.
- **Backbones:** DINOv2-L frozen for semantics; SAM regions generated offline; $N_{tr} = 4$ transformer layers.
- **Metrics:** Area AUC, Homography AUC@3/5/10, Pose AUC@$5^\circ/10^\circ/20^\circ$, runtime.

Area Matching Accuracy

Dataset	AUC@0.2	AUC@0.3	AUC@0.4	AUC@0.5
ScanNet1500	95.46	96.18	96.78	97.28
MegaDepth1500	97.39	97.81	98.02	98.18

Table: SGAD maintains $\geq$95% area-matching AUC on both indoor and outdoor benchmarks.

- Unlike MESA/DMESA, SGAD outputs a full confidence matrix, simplifying evaluation and downstream integration.
- High-quality area correspondences provide strong priors for fine-grained point matchers.

Pose Estimation Improvements (ScanNet)

Pose estimation AUC	ScanNet1500 benchmark								
	1296×968			880×640			640×480		
	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°
TopicFM [4] AAAI'23	19.14	36.55	52.68	19.23	36.48	52.49	19.45	36.57	52.75
TopicFM+ [5] TIP'24	20.26	37.83	54.06	19.73	37.19	53.64	20.00	37.79	53.98
SP[1]+SG[6] CVPR'19	22.62	42.89	61.44	23.37	43.68	62.76	21.73	41.64	60.41
SGAD+SPSG	25.74$_{+13.79\%}$	45.95$_{+7.13\%}$	63.77$_{+3.79\%}$	25.17$_{+7.70\%}$	45.46$_{+4.08\%}$	63.48$_{+1.15\%}$	24.31$_{+11.87\%}$	44.64$_{+7.20\%}$	62.92$_{+4.15\%}$
LoFTR [7] CVPR'21	8.91	18.31	29.56	20.01	39.50	57.70	28.44	50.43	68.80
SGAD+LoFTR	29.69$_{+233.22\%}$	51.50$_{+181.27\%}$	69.58$_{+135.39\%}$	28.79$_{+43.88\%}$	50.68$_{+28.30\%}$	68.92$_{+19.45\%}$	28.69$_{+0.88\%}$	51.18$_{+1.49\%}$	69.40$_{+0.87\%}$
DKM [2] CVPR'23	24.16	44.03	61.34	29.30	51.02	68.51	30.17	51.80	69.52
MESA [9]+DKM CVPR'24	30.07$_{+24.46\%}$	51.57$_{+17.12\%}$	68.98$_{+12.46\%}$	30.76$_{+4.98\%}$	52.59$_{+3.08\%}$	69.74$_{+1.80\%}$	30.85$_{+2.25\%}$	52.57$_{+1.49\%}$	69.97$_{+0.65\%}$
DMESA [8]+DKM ArXiv'24	28.89$_{+19.58\%}$	49.34$_{+12.06\%}$	66.28$_{+8.05\%}$	30.61$_{+4.47\%}$	51.95$_{+1.82\%}$	69.14$_{+0.92\%}$	30.90$_{+2.42\%}$	52.31$_{+0.98\%}$	69.86$_{+0.49\%}$
SGAD+DKM	31.63$_{+30.92\%}$	52.98$_{+20.33\%}$	69.93$_{+14.00\%}$	31.85$_{+8.70\%}$	53.11$_{+4.10\%}$	70.12$_{+2.35\%}$	31.65$_{+4.91\%}$	53.12$_{+2.55\%}$	70.69$_{+1.68\%}$
SGAD†+DKM	31.51$_{+30.42\%}$	52.87$_{+20.08\%}$	69.87$_{+13.91\%}$	31.76$_{+8.40\%}$	53.02$_{+3.92\%}$	70.05$_{+2.25\%}$	31.49$_{+4.38\%}$	52.92$_{+2.16\%}$	70.55$_{+1.48\%}$
ROMA [3] CVPR'24	31.51	53.44	71.10	31.99	53.90	71.39	31.80	53.92	71.29
SGAD+ROMA	**33.84**$_{+7.39\%}$	**55.37**$_{+3.61\%}$	**72.23**$_{+1.59\%}$	**33.61**$_{+5.06\%}$	**55.19**$_{+2.39\%}$	**72.11**$_{+1.01\%}$	**33.49**$_{+5.31\%}$	**55.14**$_{+2.26\%}$	**72.16**$_{+1.22\%}$

Table: Relative pose estimation on ScanNet1500. †: model trained on MegaDepth.

Pose Estimation Improvements (MegaDepth)

Pose estimation AUC	MegaDepth1500 benchmark								
	1200×1200			832×832			640×640		
	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°
SP[1]+SG[6] CVPR'19	56.83	71.90	83.03	53.32	68.75	80.66	47.28	63.57	76.50
TopicFM [4] AAAI'23	52.68	69.44	81.42	49.36	67.28	80.01	46.53	64.15	77.73
TopicFM+ [5] TIP'24	56.52	71.93	82.87	55.03	70.18	81.49	49.53	65.31	77.49
LoFTR [7] CVPR'21	62.37	76.34	85.96	60.64	74.82	84.83	56.42	71.80	82.65
SGAD+LoFTR	65.98$_{+5.79\%}$	78.77$_{+3.18\%}$	87.13$_{+1.36\%}$	65.16$_{+7.45\%}$	77.94$_{+4.17\%}$	86.62$_{+2.11\%}$	63.94$_{+13.33\%}$	76.90$_{+7.10\%}$	85.78$_{+3.79\%}$
DKM [2] CVPR'23	61.11	74.63	84.02	62.42	75.88	85.11	63.26	76.13	84.97
MESA [9]+DKM CVPR'24	62.31$_{+1.96\%}$	76.11$_{+1.98\%}$	85.56$_{+1.83\%}$	62.68$_{+0.42\%}$	75.96$_{+0.11\%}$	85.35$_{+0.28\%}$	63.02$_{-0.38\%}$	76.31$_{+0.24\%}$	85.60$_{+0.74\%}$
DMESA [8]+DKM ArXiv'24	63.52$_{+3.94\%}$	76.29$_{+2.22\%}$	85.31$_{+1.54\%}$	64.02$_{+2.56\%}$	76.69$_{+1.07\%}$	85.54$_{+0.51\%}$	65.24$_{+3.13\%}$	77.98$_{+2.43\%}$	86.55$_{+1.86\%}$
SGAD+DKM	66.40$_{+8.66\%}$	78.38$_{+5.02\%}$	86.51$_{+2.96\%}$	66.49$_{+6.52\%}$	78.80$_{+3.85\%}$	87.23$_{+2.49\%}$	66.75$_{+5.52\%}$	78.80$_{+3.51\%}$	87.13$_{+2.54\%}$
SGAD†+DKM	66.04$_{+8.08\%}$	78.07$_{+4.61\%}$	86.50$_{+2.95\%}$	65.97$_{+5.69\%}$	78.35$_{+3.26\%}$	86.88$_{+2.08\%}$	66.73$_{+5.49\%}$	78.39$_{+2.97\%}$	86.67$_{+2.00\%}$
ROMA [3] CVPR'24	65.68	78.15	86.68	65.91	78.41	86.95	65.29	78.01	86.68
SGAD+ROMA	**67.85**$_{+3.30\%}$	**79.87**$_{+2.20\%}$	**88.02**$_{+1.55\%}$	**68.34**$_{+3.69\%}$	**80.27**$_{+2.37\%}$	**88.34**$_{+1.57\%}$	**67.94**$_{+4.06\%}$	**80.09**$_{+2.67\%}$	**88.40**$_{+1.98\%}$

Table: Relative pose estimation on MegaDepth1500. †: model trained on ScanNet.

Runtime Comparison

Single Forward-Pass Advantage

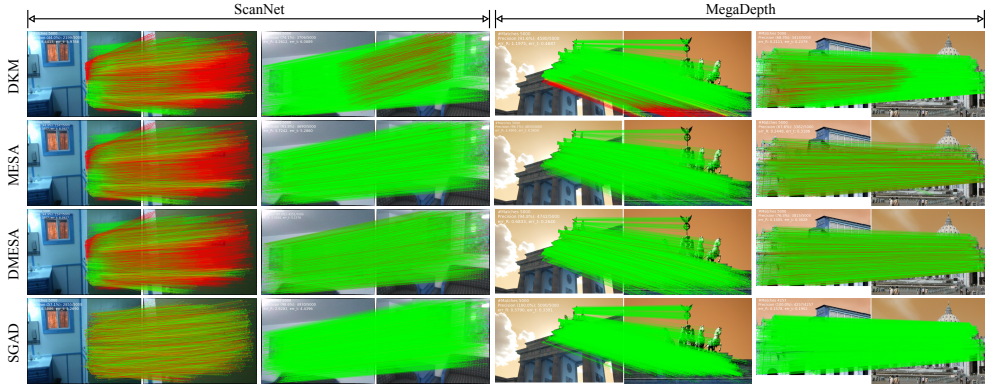
Matching relies on descriptor dot products and MNN filtering, which are fully parallelizable and eliminate graph-optimization bottlenecks.

	LoFTR	DKM	MESA+LoFTR	DMESA+LoFTR	SGAD+LoFTR
MegaDepth	0.38	1.51	60.23	1.84	**0.82**
ScanNet	0.28	0.72	33.44	1.38	**0.67**

Table: Average inference time (seconds) on an RTX A6000. SAM preprocessing excluded.

- SGAD+LoFTR is $\sim$73$\times$ faster than MESA+LoFTR and $\sim$2.2$\times$ faster than DMESA+LoFTR.
- Semi-dense SGAD+LoFTR even surpasses dense DKM in both speed and accuracy.

Qualitative Comparisons



- SGAD preserves more correct matches under large viewpoint changes and low-texture conditions.
- Rotation and translation errors are consistently lower than MESA and DMESA.

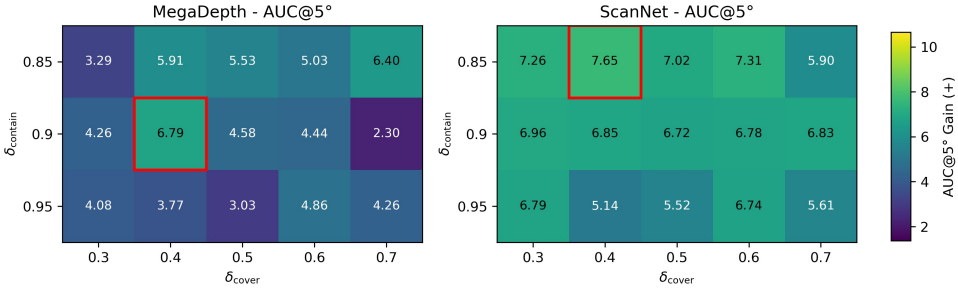
Ablation Study

	DINOv2	Attention	PE	$\mathcal{L}_{cls}$	$\mathcal{L}_{rank}$	$\mathcal{L}_{Triplet}$	AUC@0.2	AUC@0.3
1)	✓	✗	✗	✓	✓	✗	79.74	82.62
2)	✓	✓	✗	✓	✓	✗	92.52	94.23
3)	✓	✓	✓	✓	✓	✗	**95.46**	**96.18**
4)	✓	✓	✓	✓	✗	✗	94.82	95.61
5)	✓	✓	✓	✗	✗	✓	91.81	93.55

Table: Ablation on ScanNet1500 validating each component of SGAD.

- Geometric encoding and alternating attention drive descriptor discriminability.
- Ranking supervision boosts relative ordering beyond pure classification.

HCRF Sensitivity



- Coverage in [0.4, 0.5] and contain threshold in [0.85, 0.9] work robustly across datasets.
- Gains of $+6.79$ and $+7.65$ AUC@5° over LoFTR on MegaDepth and ScanNet, respectively.

Discussion

Comparison to MESA

- **Matching strategy:** SGAD replaces pixel-level activity maps and graph optimization with descriptor matching, drastically cutting complexity.
- **Descriptor strength:** SGAD aggregates global context, positional cues, and cross-image relationships, producing stable area correspondences without graph optimization.
- **Scalability:** training and inference are GPU-friendly and combine seamlessly with existing point matchers.

Takeaway

Descriptor-centric region matching offers a lightweight, scalable alternative to previous A2PM pipelines.

Limitations and Future Work

- Average pooling into compact vectors may still struggle with extreme geometric distortions—richer shape modeling could help.
- HCRF reduces redundancy, yet partial overlaps remain; exploring end-to-end region selection is promising.
- Current design cascades with point matchers; joint training or adaptive scheduling could further improve performance.

Conclusion

Conclusion

- SGAD delivers single-pass, parallel area matching that balances accuracy and efficiency.
- Dual-task supervision and HCRF strengthen match quality and downstream pose estimation.
- Verified gains across LoFTR, DKM, ROMA, establishing new state of the art on multiple benchmarks.

Resources

Project page: https://mr-chiwang.github.io/SGAD/

References I

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.

[2] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023.

[3] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024.

[4] Khang Truong Giang, Soohwan Song, and Sungho Jo. TopicFM: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2447–2455, 2023.

[5] Khang Truong Giang, Soohwan Song, and Sungho Jo. TopicFM+: Boosting accuracy and efficiency of topic-assisted feature matching. *IEEE Transactions on Image Processing*, 2024.

[6] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.

References II

[7] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[8] Yesheng Zhang and Xu Zhao. DMESA: Densely matching everything by segmenting anything. *arXiv preprint arXiv:2408.00279*, 2024.

[9] Yesheng Zhang and Xu Zhao. MESA: Matching everything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20217–20226, 2024.

Thank You

Questions and discussion are welcome.