Supplementary Material – Learning Affine Correspondences by Integrating Geometric Constraints

Pengju Sun[1, 2] Banglei Guan[1, 2(✉)] Zhenbao Yu[1, 2] Yang Shang[1, 2] Qifeng Yu[1, 2] Daniel Barath[3, 4]

[1]College of Aerospace Science and Engineering, National University of Defense Technology, China.
[2] Hunan Provincial Key Laboratory of Image Measurement and Vision Navigation, China.
[3]ETH Zurich, Switzerland. [4] HUN-REN SZTAKI, Hungary.

1. Overview

In this supplementary material, we provide the details of the loss function and additional experiments. In Sec. 2, the geometric constraints based on Sampson Distance are derived. In Sec. 3, we demonstrate the performance of our method on datasets with large viewpoint changes. In Sec. 4, we demonstrate that the affine correspondences obtained by our method lead to improved relative pose accuracy compared with methods obtaining point correspondences on the indoor dataset. In Sec. 5, we show the failed cases.

2. Details of Sampson Distance for Geometric Constraints

Sampson Distance was originally introduced for conic fitting. The method finds the refined parameters that reduce the overall fitting errors iteratively [18]. Recently, Sampson Distance has also been used to model the measurement residuals of the correspondences between two views in computer vision [6]. It can be regarded as a first-order approximation of geometric error and offers an efficient and effective alternative to traditional error metrics. Characterized by its reduced computational complexity, it provides an estimate of error that is comparable in accuracy to the geometrical error [18]. In previous work, Zhou et al. [19] proposed that how much a match prediction fulfills the epipolar geometry can be precisely measured by the Sampson distance. In this paper, A novel affine transformation loss, represented by the Affine Sampson Distance, is introduced to further enhance the conformity of affine correspondences with the scene geometry. Given an AC satisfying $G_E(\widehat{X}) = 0$, where $G_E(X)$ is the geometric constraint approximated by a Taylor expansion:

$$G_E\left(X+\delta_X\right) \approx G_E(X) + \frac{\partial G_E}{\partial X}\delta_X, \tag{1}$$

δ_X quantifies the measurement residual. Letting

$$J = \frac{\partial G_E}{\partial X}, \tag{2}$$

$$\epsilon = G_E(X) - G_E(\widehat{X}), \tag{3}$$

namely,

$$J\delta_X = -\epsilon, \tag{4}$$

the goal is to find δ_X that minimizes $\|\delta_X\|$ subject to Eq. 1. The problem can be solved by Lagrange Multipliers and the Sampson Distance is defined as the squared norm of δ_X.

$$\left\|\delta_X\right\|^2 = \epsilon^T\left(JJ^T\right)^{-1}\epsilon. \tag{5}$$

For the epipolar constraints,

$$G_E(X) = p_2^T F p_1. \tag{6}$$

We take partial derivatives of x_1, y_1, x_2, y_2. Let $Z_0 = G(X)$. The remaining terms are $Z_1 = \frac{\partial Z_0}{\partial x_1}$, $Z_2 = \frac{\partial Z_0}{\partial y_1}$, $Z_3 = \frac{\partial Z_0}{\partial x_2}$, $Z_4 = \frac{\partial Z_0}{\partial y_2}$, we can obtain Eq. 7.

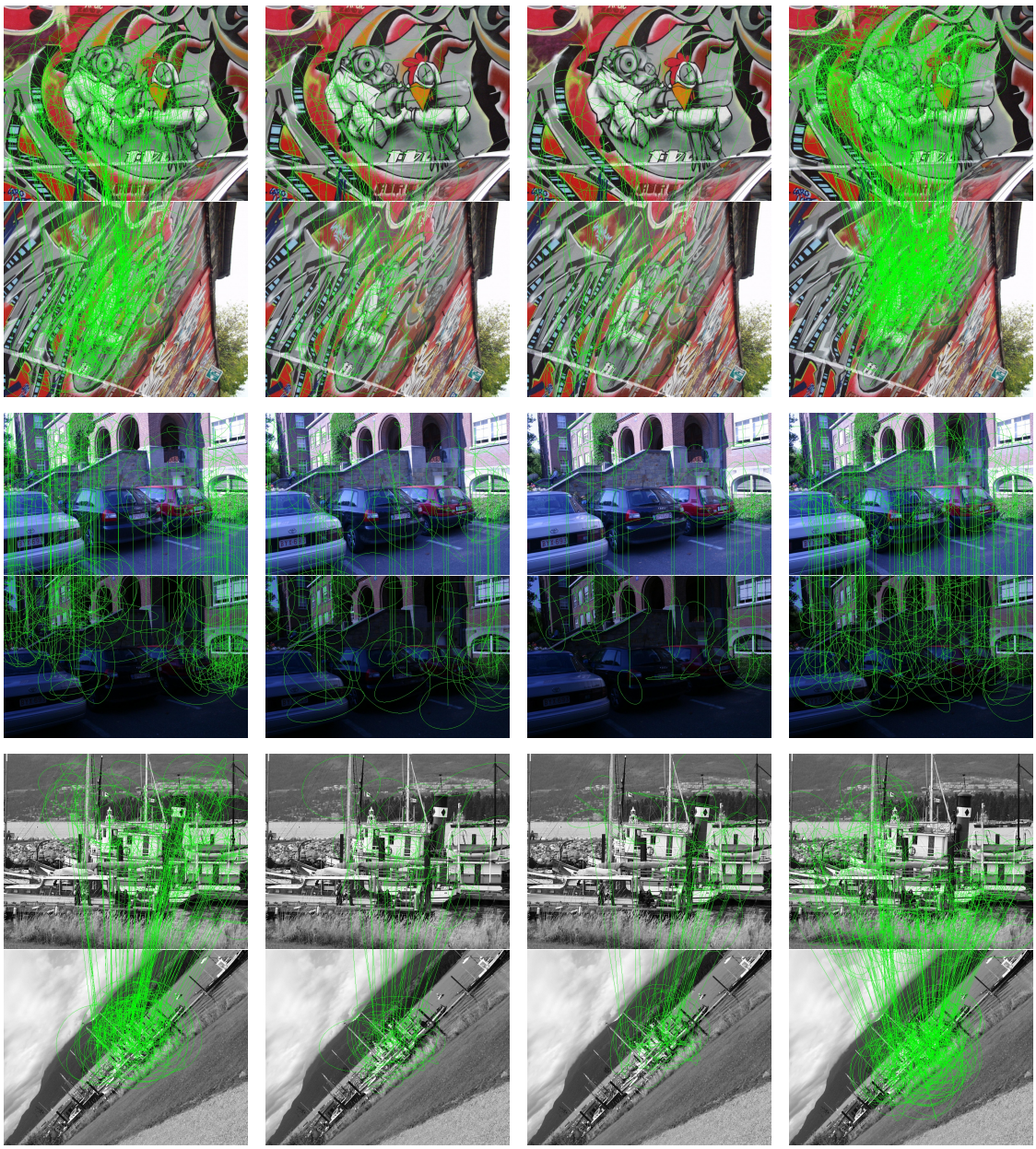
$$SD_P(E_{PC}) = \frac{Z_0^2}{Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2}, \tag{7}$$

where

$$\begin{cases} Z_0 = x_1(f_{31} + f_{11}x_2 + f_{21}y_2) + y_1(f_{32} + \\ f_{12}x_2 + f_{22}y_2) + f_{13}x_2 + f_{23}y_2 + f_{33}, \\ Z_1 = f_{31} + f_{11}x_2 + f_{21}y_2, \\ Z_2 = f_{32} + f_{12}x_2 + f_{22}y_2, \\ Z_3 = f_{13} + f_{11}x_1 + f_{12}y_1, \\ Z_4 = f_{23} + f_{22}y_1 + f_{21}x_1, \end{cases} \tag{8}$$

the f_{ij}, $(i, j \in \{1, 2, 3\})$ is an element in the fundamental matrix. The affine transformation constraints is as follows:

$$SD_A\left(E_{AC}\right)_{(1:2)} = SD_A\left(A^{-T}\left(F^T p_2\right)_{(1:2)} + \left(F p_1\right)_{(1:2)}\right). \tag{9}$$



(a) ASIFT (b) VLFeat (c) AffNet (d) Ours

Figure 1. The image matching results in the Extreme View Dataset [9]. Our method could finds the highest number of correct matches.

When $G(X)$ is the constraint on the first row in Eq. 9 in the paper. We take partial derivatives of $x_1 \dots a_{22}$. Let $M_0 = G(X)$. The remaining terms are $M_1 = \frac{\partial M_0}{\partial a_{11}}$, $M_2 = \frac{\partial M_0}{\partial y_1}$, $M_3 = \frac{\partial M_0}{\partial x_2}$, $M_4 = \frac{\partial M_0}{\partial a_{21}}$, $M_5 = \frac{\partial M_0}{\partial x_1}$, $M_6 = \frac{\partial M_0}{\partial y_2}$. The first one can be formulated as follows:

$$SD_A(E_{AC})_{(1)} = \frac{M_0^2}{M_1^2 + M_2^2 + M_3^2 + M_4^2 + M_5^2 + M_6^2}, \tag{10}$$

where

$$\begin{cases} M_0 = x_1(a_{11}f_{11} + a_{21}f_{21}) + y_1(a_{11}f_{12} + a_{21}f_{22}) \\ +a_{11}f_{13} + a_{21}f_{23} + f_{11}x_2 + f_{21}y_2 + f_{31}, \\ M_1 = f_{13} + f_{11}x_1 + f_{12}y_1, \\ M_2 = a_{11}f_{12} + a_{21}f_{22}, \\ M_3 = f_{11}, \\ M_4 = f_{23} + f_{21}x_1 + f_{22}y_1, \\ M_5 = a_{11}f_{11} + a_{21}f_{21}, \\ M_6 = f_{21}, \end{cases} \tag{11}$$

(a) VLFeat (b) AffNet (c) ASIFT (d) Ours

Figure 2. Failure modes. Other methods also fail.

Similarly, the second one is formulated as

$$SD_A(E_{AC})_{(2)} = \frac{N_0^2}{N_1^2 + N_2^2 + N_3^2 + N_4^2 + N_5^2 + N_6^2}, \tag{12}$$

where

$$\begin{cases} N_0 = x_1(a_{12}f_{11} + a_{22}f_{21}) + y_1(a_{12}f_{12} + a_{22}f_{22}) \\ +a_{12}f_{13} + a_{22}f_{23} + f_{12x_2} + f_{22}y_2 + f_{32}, \\ N_1 = f_{13} + f_{11}x_1 + f_{12}y_1, \\ N_2 = a_{12}f_{11} + a_{22}f_{21}, \\ N_3 = f_{12}, \\ N_4 = f_{23} + f_{21}x_1 + f_{22}y_1, \\ N_5 = a_{12}f_{12} + a_{22}f_{22}, \\ N_6 = f_{22}, \end{cases} \tag{13}$$

3. Image Matching on EVD

Affine features are beneficial for matching images with large viewpoint changes because they utilize further geometric information compared to their point-based counterparts. We now show additional results on the Extreme View dataset [10], whose average viewpoint change is substantially larger than that of the HPatches dataset [1]. The dataset with the ground truth is available on the web-page[1].

Evaluation protocol. We compare the proposed method with the view-synthesis-based Affine-SIFT (ASIFT) [13], the VLFeat library [17], and the learning-based AffNet [12]. Following the protocol [12], we report the number of successfully matched image pairs and the average number of correct inliers per matched pair.

Results. The average inlier numbers and the number of successfully matched image pairs are shown in Table 1. Example results are shown in the Fig. 1. Only the correct matches are displayed. Our method has a significant advantage in terms of matching quantity at the same pixel error threshold.

[1]http://cmp.felk.cvut.cz/wbs/index.html

Benefiting from the use of dense matching, and through the estimation of affine features, our method obtains more accurate matches in the case of large viewpoint change. This experiment demonstrates that our method is more robust than other affine-based ones to large viewpoint changes. This signifies that the affine correspondences we extract are of better quality. This can be attributed to our pipeline design for affine correspondence extraction, leveraging a combination of geometric constraints.

Table 1. The comparison of affine extractors on a wide baseline stereo dataset EVD [9] following the protocol in [11]. The number of successfully matched image pairs (N) and the average number of correct inliers (inl.) are presented. The best result is in bold.

	VLFeat [17]	AffNet [12]	ASIFT [13]	Ours
N.	2	4	2	**11**
inl.	56	34	64	**137**

4. Relative Pose Estimation on ScanNet-1500

The ScanNet [5] is a large-scale indoor dataset that is used to target the task of indoor pose estimation. This dataset is challenging since it contains image pairs with wide baselines and extensive texture-less regions. We follow the evaluation in SuperGlue [14].

Evaluation protocol. Following [14] and [15], we report the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$. To compare with existing methods on the same baseline, we utilize RANSAC as implemented in the OpenCV library to solve for the essential matrix from predicted matches as previous methods do [15]. To demonstrate that the estimated affine frames are beneficial for pose estimation, we also run the affine correspondence-based Graph-Cut RANSAC [2, 3], designed specifically to leverage affine shapes together with the point locations.

Result. As shown in Table 2, the proposed method with

Table 2. Relative pose Accuracy Under the recall Curve (AUC; higher is better) thresholded at 5°, 10°, and 20° on the ScanNet-1500 [5]. All methods run RANSAC-based essential matrix estimation, except for the last row, where we run the affine-based GC-RANSAC [2, 3], benefiting from the affine correspondences that we obtain.The best results are in bold.

AUC@ $\rightarrow$	$5^\circ \uparrow$	$10^\circ \uparrow$	$20^\circ \uparrow$
LoFTR [15] $_{CVPR^{21}}$	22.1	40.8	57.6
ASpanFormer [4] $_{ECCV^{22}}$	25.6	46.0	63.3
PDC-Net+ [16] $_{TPAMI^{23}}$	20.3	39.4	57.1
DKM [7] $_{CVPR^{23}}$	29.4	50.7	68.3
RoMA[8] $_{CVPR^{24}}$	31.8	53.4	70.9
Ours(RANSAC)	30.7	51.7	69.0
Ours(aff. GC-RANSAC)	**33.1**	**55.9**	**73.4**

RANSAC achieves good results. When leveraging the estimated affine shapes with GC-RANSAC, the proposed method achieves the best performance.

5. Failed Cases

Fig. 2 shows the failure cases caused by significant viewpoint changes and large scale variations. However, all other tested baselines fail in these cases.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2017. 3

[2] Daniel Barath and Jiri Matas. Graph-Cut RANSAC: Local optimization on spatially coherent structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4961–4974, 2022. 3, 4

[3] Daniel Barath, Michal Polic, Wolfgang Förstner, Torsten Sattler, Tomas Pajdla, and Zuzana Kukelova. Making affine correspondences work in camera geometry computation. In *European Conference on Computer Vision*, pages 723–740, 2020. 3, 4

[4] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David N. R. McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36, 2022. 4

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 3, 4

[6] Yuchao Dai, Hongdong Li, and Laurent Kneip. Rolling shutter camera relative pose: Generalized epipolar geometry. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4132–4140, 2016. 1

[7] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 4

[8] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 4

[9] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. Mods: Fast and robust method for two-view matching. *Computer Vision and Image Understanding*, 141:81–93, 2015. 2, 3

[10] Dmytro Mishkin, Jiri Matas, and Michal Perdoch. MODS: Fast and robust method for two-view matching. *Computer vision and image understanding*, 141:81–93, 2015. 3

[11] Dmytro Mishkin, Jiri Matas, Michal Perdoch, and Karel Lenc. Wxbs: Wide baseline stereo generalizations. *ArXiv*, abs/1504.06603, 2015. 3

[12] Dmytro Mishkin, Filip Radenović, and Jiři Matas. Repeatability is not enough: Learning affine regions via discriminability. In *European Conference on Computer Vision*, pages 287–304, 2018. 3

[13] Jean-Michel Morel and Guoshen Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, page 438–469, 2009. 3

[14] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4937–4946, 2020. 3

[15] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 3, 4

[16] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced probabilistic dense correspondence network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10247–10266, 2023. 4

[17] Andrea Vedaldi and Brian Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. *ACM international conference on Multimedia*, page 1469–1472, 2010. 3

[18] He Zhang and Cang Ye. Sampson distance: A new approach to improving visual-inertial odometry's accuracy. In *IEEE International Conference on Intelligent Robots and Systems*, pages 9184–9189, 2021. 1

[19] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2Pix: Epipolar-guided pixel-level correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4667–4676, 2021. 1