LoFTR: Detector-Free Local Feature Matching with Transformers

CVPR 2021   Zhejiang University CG&CAD

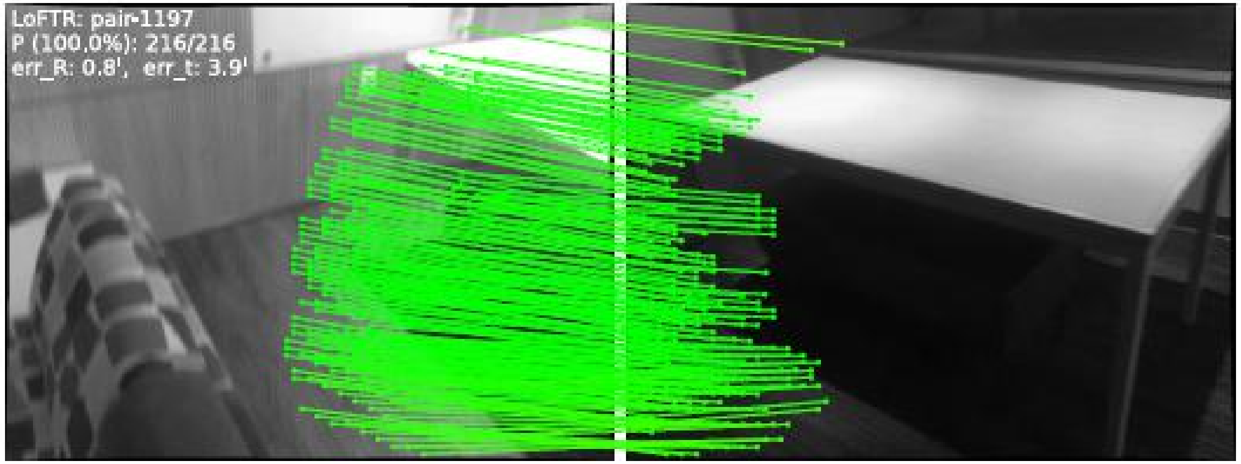
Introduction

Problem setting: Point matching between images

Input: A pair of image

Output: Sparse points correspondence between images

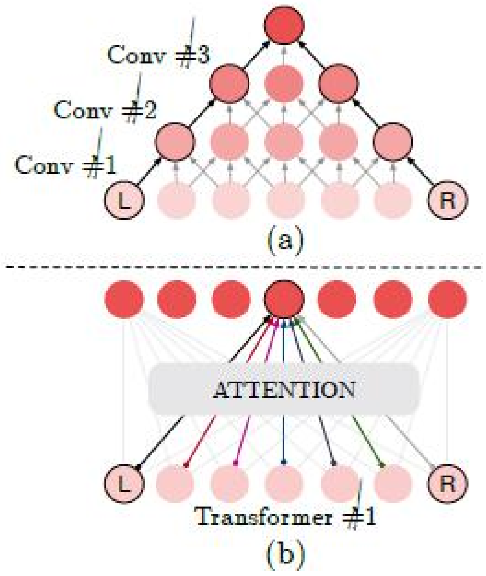
Application: SfM, SLAM, etc.



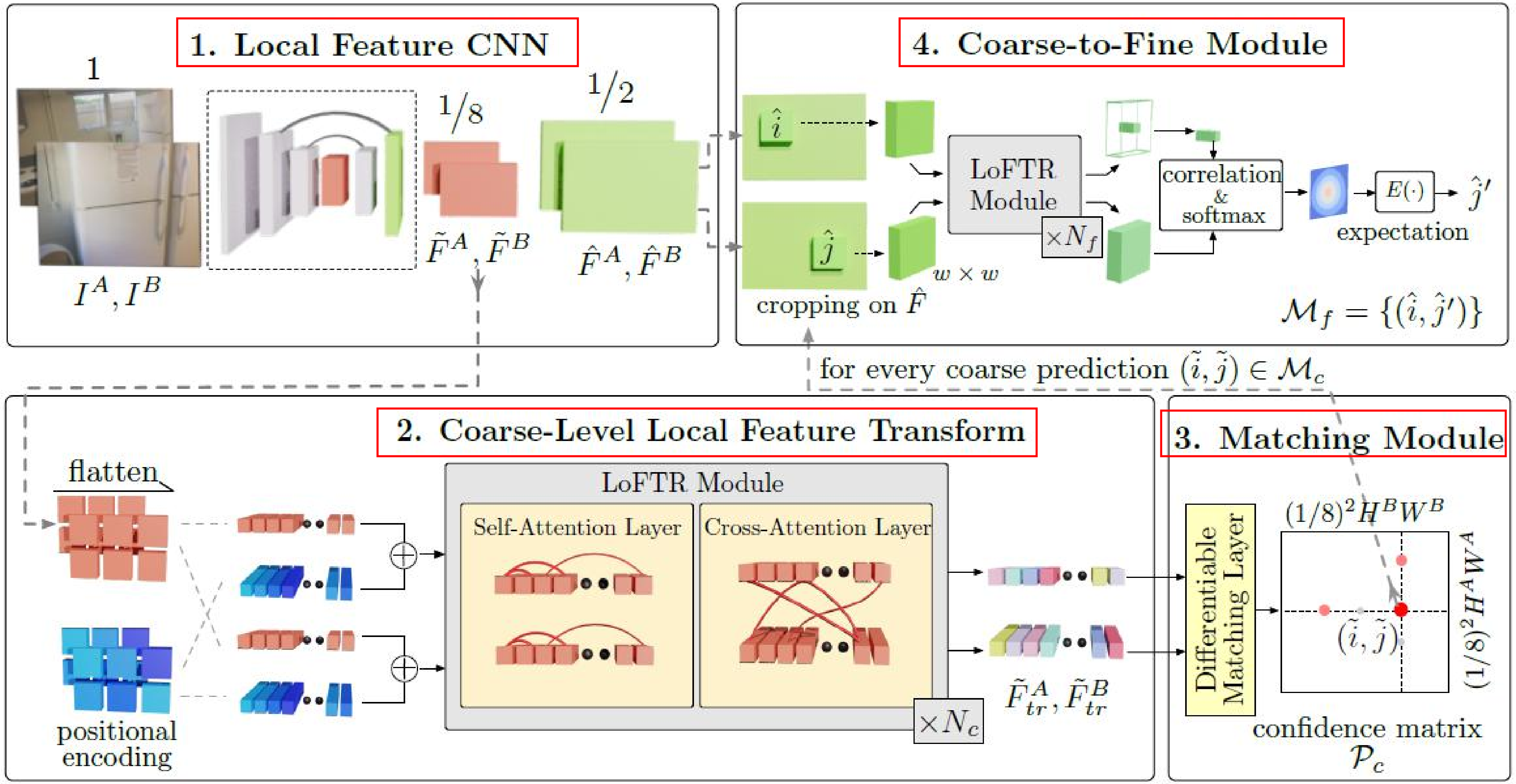
Motivation

1. Hand-crafted point descriptor fail to extract **enough interest points** due to various factors: e.g. poor texture, repetitive patterns, illumination variation
   → **pixel-wise dense matches** by CNNs and selected by **high confidence scores**

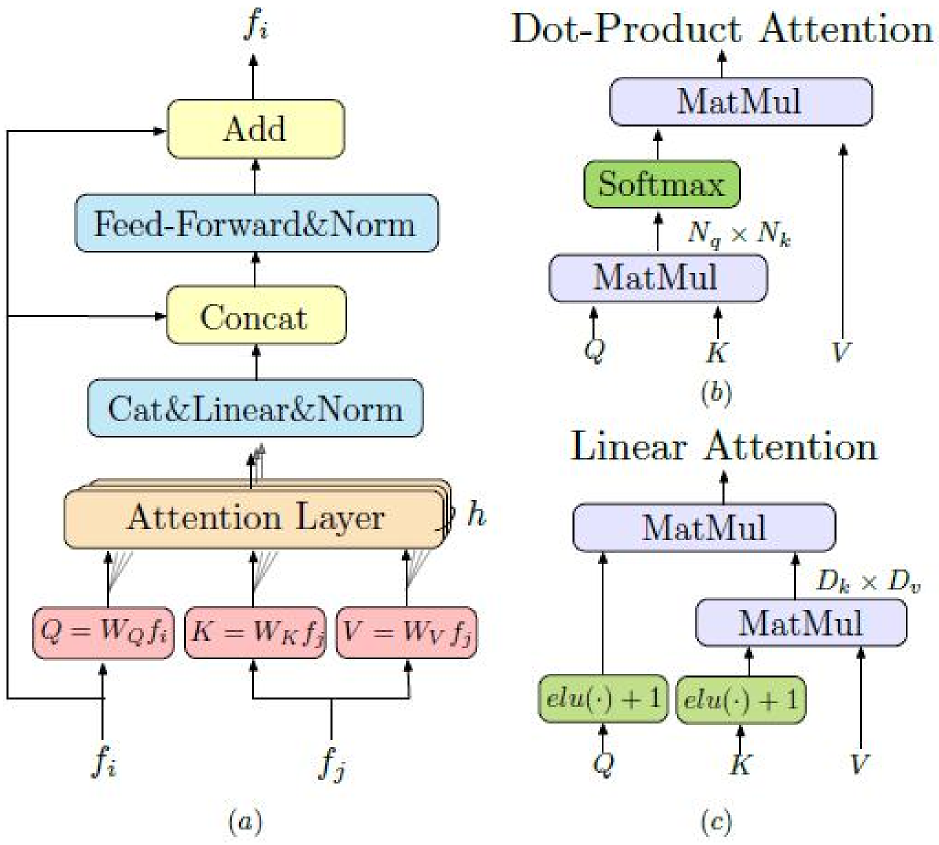
2. CNNs suffer from **limited receptive field**, which may fail in large indistinctive regions
   → human can find correspondences with a **larger global receptive field** -> transformer



Method: Pipeline



Method: Local Feature Transformer(LoFTR) Module



(a) Transformer encoder layer; (b) Vanilla dot-product attention with O(N*N) complexity;
(c) Linear attention layer with O(N) complexity;

Method: Establishing Coarse-level Matches

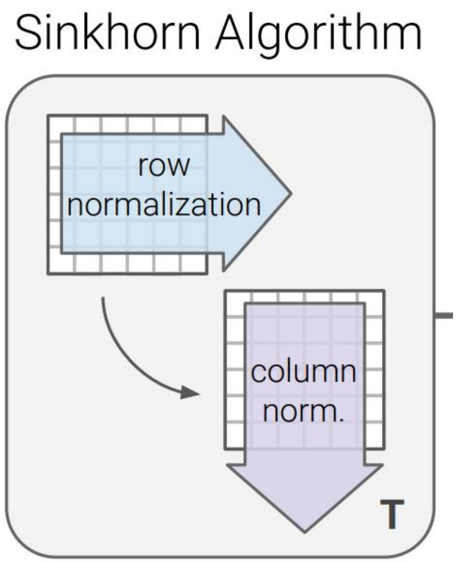
Differentiable matching layers

Dual-softmax

$$\hat{c}_{ijkl} = r^A_{ijkl} r^B_{ijkl} c_{ijkl},$$

$$r^A_{ijkl} = \frac{c_{ijkl}}{\max_{ab} c_{abkl}}, \quad \text{and} \quad r^B_{ijkl} = \frac{c_{ijkl}}{\max_{cd} c_{ijcd}}.$$

Sinkhorn

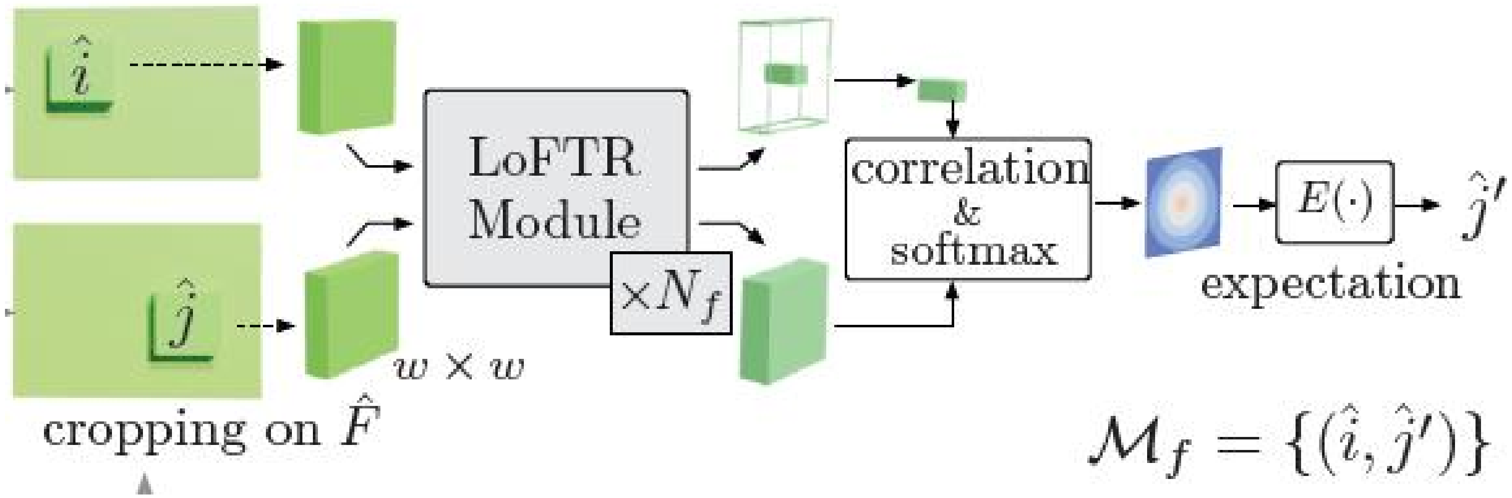


Match Selection

$$\mathcal{M}_c = \{(\tilde{i},\tilde{j}) \mid \forall (\tilde{i},\tilde{j}) \in \mathrm{MNN}(\mathcal{P}_c),\ \mathcal{P}_c(\tilde{i},\tilde{j}) \geq \theta_c\}$$

Simply saying,
iteratively normalize the matrix row and col,
finally will converge.

Method: Coarse-to-Fine Module



By computing expectation over the probability distribution, we get the final position with sub-pixel accuracy.

Method: Supervision

Coarse-level Supervision

$$\mathcal{L}_c = -\frac{1}{|\mathcal{M}_c^{gt}|} \sum_{(\tilde{i},\tilde{j}) \in \mathcal{M}_c^{gt}} \log \mathcal{P}_c\left(\tilde{i},\tilde{j}\right)$$

Fine-level Supervision

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i},\hat{j}') \in \mathcal{M}_f} \boxed{\frac{1}{\sigma^2(\hat{i})}} \left\|\hat{j}' - \hat{j}'_{gt}\right\|_2$$

Focusing on low uncertainty points

Experiments: Homography Estimation

Category	Method	Homography est. AUC			#matches
		@3px	@5px	@10px	
Detector-based	D2Net [11]+NN	23.2	35.9	53.6	0.2K
	R2D2 [32]+NN	50.6	63.9	76.8	0.5K
	DISK [47]+NN	52.3	64.9	78.9	1.1K
	SP [9]+SuperGlue [37]	53.9	68.3	81.7	0.6K
Detector-free	Sparse-NCNet [33]	48.9	54.2	67.1	1.0K
	DRC-Net [19]	50.6	56.2	68.3	1.0K
	LoFTR-DS	**65.9**	**75.6**	**84.6**	1.0K

Corner error

Experiments: Pose Estimation

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	ORB [35]+GMS [2]	5.21	13.65	25.36
	D2-Net [11]+NN	5.25	14.53	27.96
	ContextDesc [27]+Ratio Test [26]	6.64	15.01	25.75
	SP [9]+NN	9.43	21.53	36.40
	SP [9]+PointCN [52]	11.40	25.47	41.41
	SP [9]+OANet [53]	11.76	26.90	43.85
	SP [9]+SuperGlue [37]	16.16	33.81	51.84
Detector-free	DRC-Net † [19]	7.69	17.93	30.49
	LoFTR-OT†	16.88	33.62	50.62
	LoFTR-OT	21.51	40.39	**57.96**
	LoFTR-DS	**22.06**	**40.8**	57.62

ScanNet (Indoor)

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	SP [9]+SuperGlue [37]	42.18	61.16	75.96
Detector-free	DRC-Net [19]	27.01	42.96	58.31
	LoFTR-OT	50.31	67.14	79.93
	LoFTR-DS	**52.8**	**69.19**	**81.18**

MegaDepth (Outdoor)

Experiments: Visual Localization

Method	Day	Night
	(0.25m,2°) / (0.5m,5°) / (1.0m,10°)	
Local Feature Evaluation on Night-time Queries		
R2D2 [32]+NN	-	71.2 / 86.9 / 98.9
LISRD [31]+SP [9]+AdaLam [4]	-	**73.3** / 86.9 / 97.9
ISRF [29]+NN	-	69.1 / 87.4 / 98.4
SP [9]+SuperGlue [37]	-	**73.3** / 88.0 / 98.4
LoFTR-DS	-	72.8 / **88.5** / **99.0**
Full Visual Localization with HLoc		
SP [9]+SuperGlue [37]	**89.8** / **96.1** / **99.4**	77.0 / **90.6** / **100.0**
LoFTR-OT	88.7 / 95.6 / 99.0	**78.5** / **90.6** / 99.0

Aachen Day-Night (outdoor)

Method	DUC1	DUC2
	(0.25m,10°) / (0.5m,10°) / (1.0m,10°)	
ISRF [29]	39.4 / 58.1 / 70.2	41.2 / 61.1 / 69.5
KAPTURE [14]+R2D2 [32]	41.4 / 60.1 / 73.7	47.3 / 67.2 / 73.3
HLoc [36]+SP [9]+SuperGlue [37]	**49.0** / 68.7 / 80.8	53.4 / **77.1** / 82.4
HLoc [36]+**LoFTR-OT**	47.5 / **72.2** / **84.8**	**54.2** / 74.8 / **85.5**

InLoc benchmark

Experiments

Using DETR-style [3] Transformer architecture which has positional encoding at each layer, leads to a noticeably declined result.

Method	Pose estimation AUC		
	@5°	@10°	@20°
1) replace LoFTR with convolution	14.98	32.04	49.92
2) 1/16 coarse-resolution + 1/4 fine-resolution	16.75	34.82	54.0
3) positional encoding per layer	18.02	35.64	52.77
4) larger model with $N_c = 8$, $N_f = 2$	**20.87**	40.23	57.56
Full ($N_c = 4$, $N_f = 1$)	20.06	**40.8**	**57.62**