JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Adaptive Spot-Guided Transformer for Consistent Local Feature Matching

Jiahuan Yu*, Jiahao Chang*, Jianfeng He, Tianzhu Zhang†, Feng Wu

University of Science and Technology of China

* Equal contribution † Corresponding author

Project Homepage:
https://astr2023.github.io

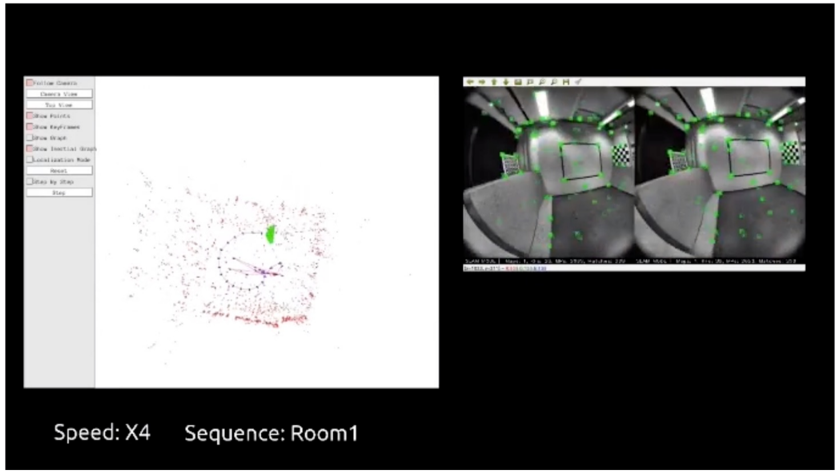
Contents

- **Introduction**
- Motivation
- Novelty
- Evaluation
- Visualization
- Conclusion

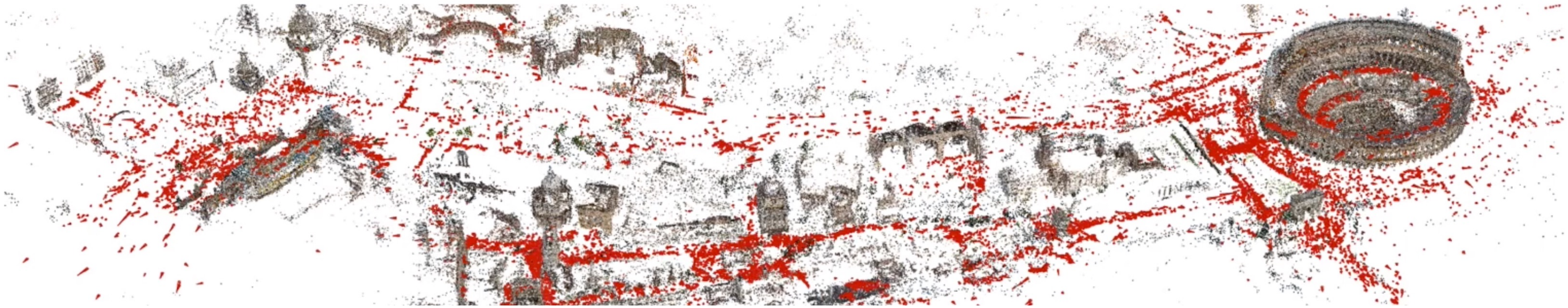
Introduction

- Local feature matching serves as a fundamental task in many 3D vision tasks

Visual Localization



SLAM



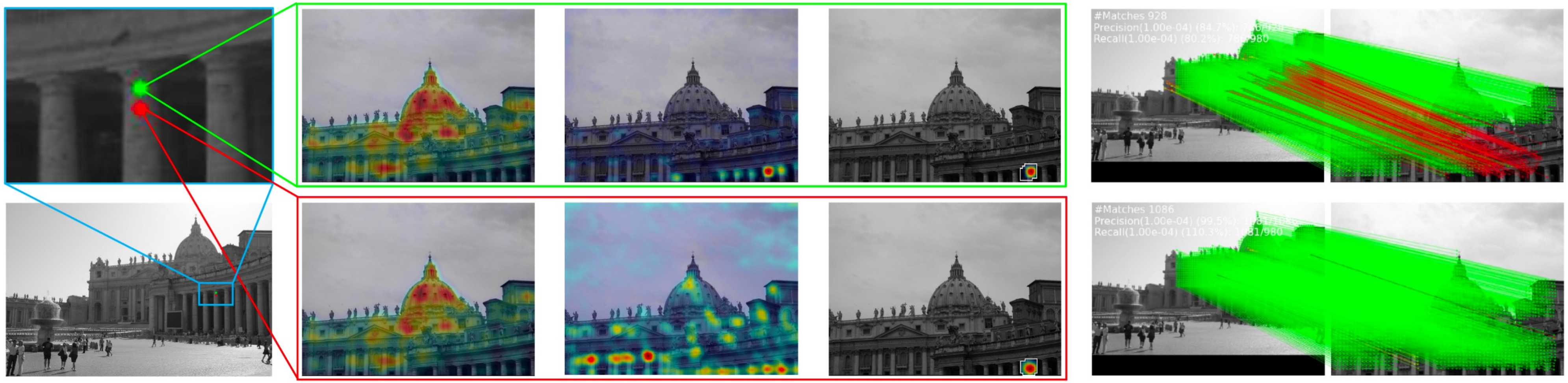
Structure from Motion (SfM)

Contents

- Introduction
- **Motivation**
- Novelty
- Evaluation
- Visualization
- Conclusion

Motivation

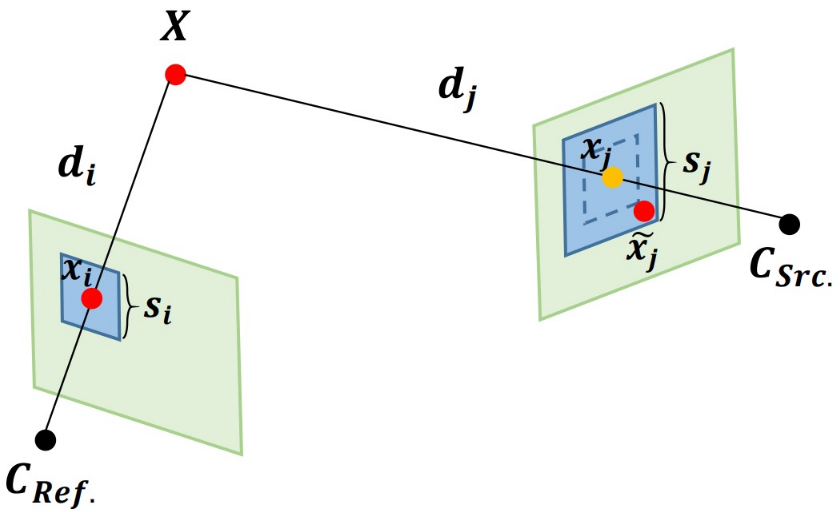
- **Local consistency** is ignored in Transformer-based methods:
  - For two **similar adjacent** pixels in reference image (red and green), the corresponding attention maps with source image:
    - are quite different (see Vanilla Attention)
    - include too many irrelevant areas (see Linear Attention)
  - Leading to **inconsistent** matching results between similar adjacent pixels



(a)Reference (b)Linear Attention (c)Vanilla Attention (d)Ours (e)Matching Result

Motivation

- **Scale variation** is not properly handled in existing coarse-to-fine methods:
  - Existing coarse-to-fine manner: refine coarse matching result in **fixed-size** fine stage windows
  - When scale variation is large, correct matching pixel may be **out of** fine stage window
  - Coarse matching: $(x_i,\ x_j)$, correct matching: $(x_i, \tilde{x}_j)$, fixed window size: s_i

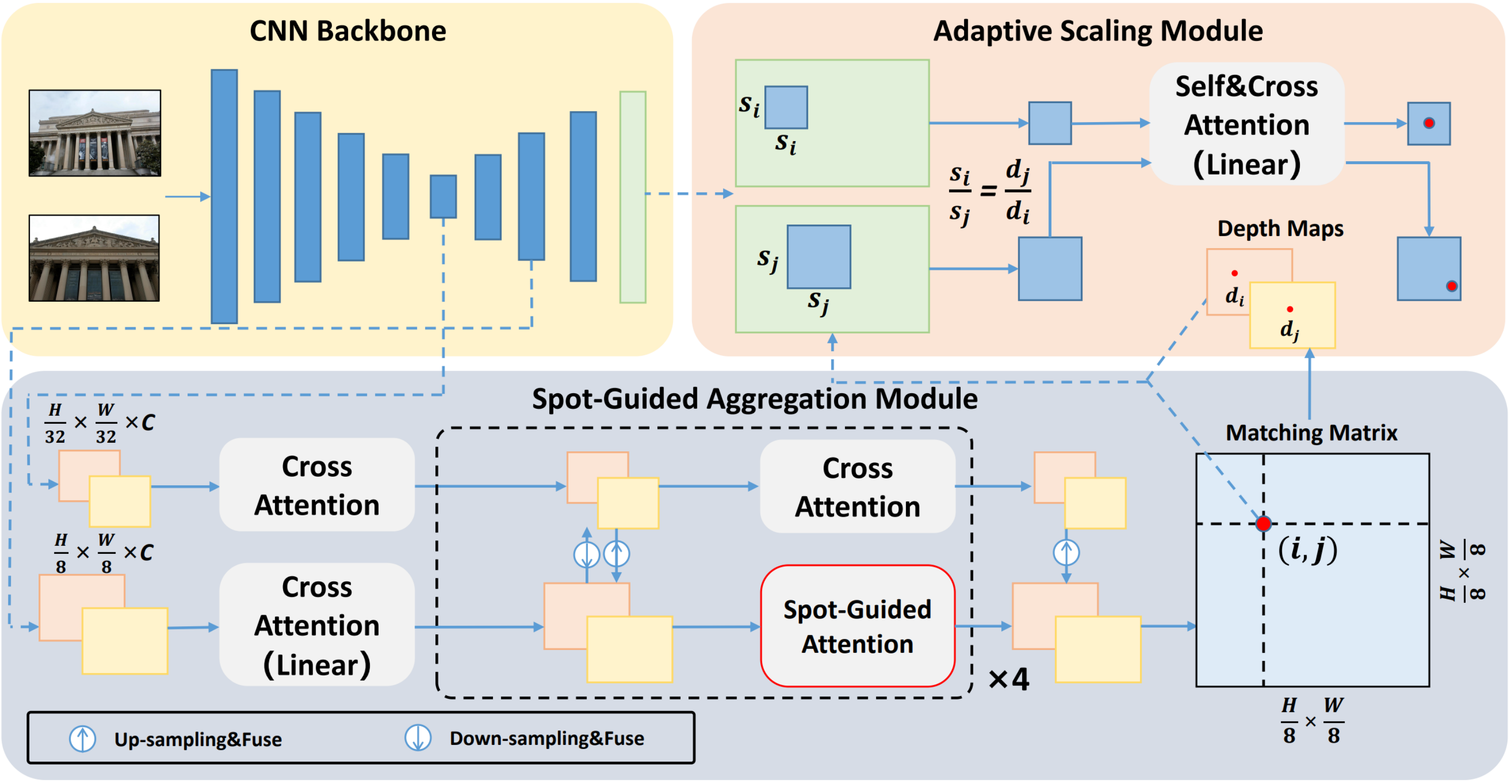


Contents

- Introduction
- Motivation
- **Novelty**
- Evaluation
- Visualization
- Conclusion

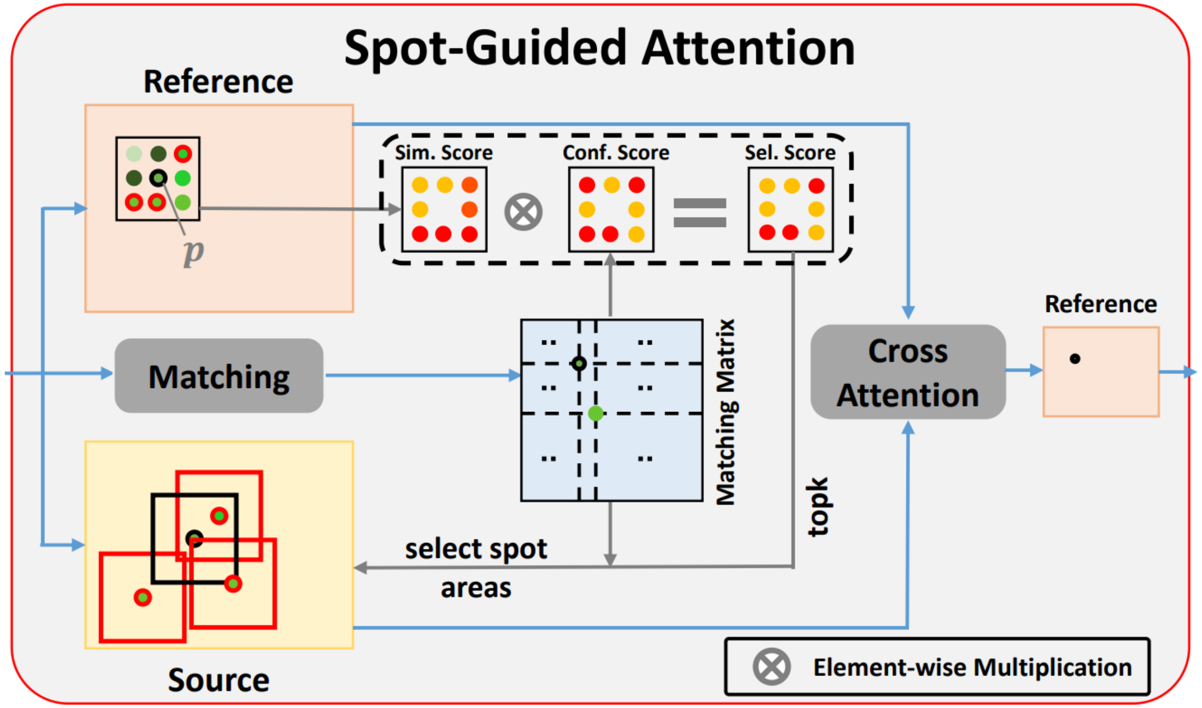
Novelty

- A unified coarse-to-fine architecture named Adaptive Spot-Guided Transformer (**ASTR**) taking local consistency and scale variation into consideration
  - **Spot-Guided Attention**: maintain local consistency
  - **Adaptive Scaling**: handle large scale variation



Novelty -- Spot-Guided Attention

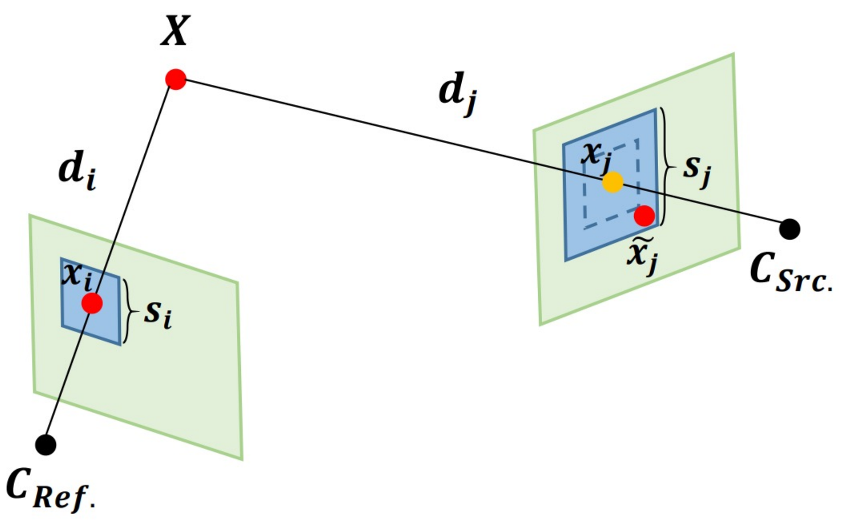
- For each pixel P in reference image:
  - $N(P)$: adjacent area of P
  - Similarity score: similarity between P and $N(P)$
  - Confidence score: matching confidence of $N(P)$
  - Selection score = Similarity score × Confidence score
  - Spot area: adjacent area of correspondence pixel of $\{P\} \cup \mathrm{topk}(N(P))$
- Do attention between P and spot area
- **Adjacent** and similar pixel share similar spot area
- Filter irrelevant area



Novelty -- Adaptive Scaling

- Coarse matching $(x_i,\ x_j)$ is obtained in coarse stage
- Correct matching $(x_i, \tilde{x}_j)$, if window size s_i is fixed, $\tilde{x}_j$ may be **out of window**
- Use coarse matching $(x_i,\ x_j)$ and RANSAC algorithm to calculate **relative depth** d_i/d_j, and scale the windows:

$$\frac{s_i}{s_j} = \frac{d_j}{d_i}$$



Contents

- Introduction
- Motivation
- Novelty
- **Evaluation**
- Visualization
- Conclusion

Evaluation

- ## Homography Estimation (HPatches)

Category	Method	Homography est. AUC			matches
		@3px	@5px	@10px	
Detector-based	D2Net [15]+NN	23.2	35.9	53.6	0.2K
	R2D2 [42]+NN	50.6	63.9	76.8	0.5K
	DISK [55]+NN	52.3	64.9	78.9	1.1K
	SP [14]+SuperGlue [47]	53.9	68.3	81.7	0.6K
	Patch2Pix [64]	46.4	59.2	73.1	1.0k
Detector-free	Sparse-NCNet [43]	48.9	54.2	67.1	1.0K
	COTR [24]	41.9	57.7	74.0	1.0K
	DRC-Net [27]	50.6	56.2	68.3	1.0K
	LoFTR [50]	65.9	75.6	84.6	1.0K
	PDC-Net+ [54]	66.7	76.8	85.8	1.0k
	ASTR(ours)	**71.7**	**80.3**	**88.0**	1.0K

- ## Visual Localization (InLoc & Aachen)

InLoc

Method	DUC1	DUC2
	$(0.25m, 10^\circ)$ / $(0.5m, 10^\circ)$ / $(1m, 10^\circ)$	
Patch2Pix [64](w.SP [47]+CAPS [58])	42.4 / 62.6 / 76.3	43.5 / 61.1 / 71.0
LoFTR [50]	47.5 / 72.2 / 84.8	54.2 / 74.8 / **85.5**
MatchFormer [57]	46.5 / 73.2 / 85.9	**55.7** / 71.8 / 81.7
ASpanFormer [9]	51.5 / **73.7** / 86.4	55.0 / 74.0 / 81.7
ASTR(ours)	**53.0 / 73.7 / 87.4**	52.7 / **76.3** / 84.0

Aachen

Method	Day	Night
	$(0.25m, 2^\circ)$ / $(0.5m, 5^\circ)$ / $(1m, 10^\circ)$	
Localization with matching pairs provided in dataset		
R2D2 [42]+NN	-	71.2 / 86.9 / 98.9
ASLFeat [36]+NN	-	72.3 / 86.4 / 97.9
SP [14]+SuperGlue [47]	-	73.3 / 88.0 / 98.4
SP [14]+SGMNet [8]	-	72.3 / 85.3 / 97.9
Localization with matching pairs generated by HLoc		
LoFTR [50]	88.7 / 95.6 / 99.0	78.5 / 90.6 / 99.0
ASpanFormer [9]	89.4 / 95.6 / 99.0	77.5 / 91.6 / 99.0
AdaMatcher [22]	89.2 / **95.9** / **99.2**	**79.1** / **92.1** / **99.5**
ASTR(ours)	**89.9** / 95.6 / **99.2**	76.4 / **92.1** / **99.5**

- ## Relative Pose Estimation (MegaDepth & ScanNet)

MegaDepth

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	SP [14]+SuperGlue [47]	42.2	59.0	73.6
	SP [14]+SGMNet [8]	40.5	59.0	73.6
Detector-free	DRC-Net [27]	27.0	42.9	58.3
	PDC-Net+(H) [54]	43.1	61.9	76.1
	LoFTR [50]	52.8	69.2	81.2
	MatchFormer [57]	53.3	69.7	81.8
	QuadTree [52]	54.6	70.5	82.2
	ASpanFormer [9]	55.3	71.5	83.1
	ASTR(ours)	**58.4**	**73.1**	**83.8**

ScanNet (* train on MegaDepth)

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	D2-Net [15]+NN	5.3	14.5	28.0
	SP [14]+OANet [61]	11.8	26.9	43.9
	SP [14]+SuperGlue [47]	16.2	33.8	51.8
Detector-free	DRC-Net [27]*	7.7	17.9	30.5
	MatchFormer [57]*	15.8	32.0	48.0
	LoFTR-OT [50]*	16.9	33.6	50.6
	Quadtree [52]*	19.0	37.3	53.5
	ASTR(ours)*	**19.4**	**37.6**	**54.4**

Evaluation

- Ablation study on MegaDepth

Proposed Module

Index	Multi-Level	Spot-Guided ($l=5, k=4$)	Scaling	Pose estimation AUC @5°	@10°	@20°
1				45.6	62.2	75.3
2	✓			46.7	63.1	76.3
3	✓	✓		47.7	64.5	77.4
4	✓	✓	✓	**48.3**	**65.0**	**77.7**

Different Adjacent Area Size l and top-k

$k(l=5)$	Pose estimation AUC @5°	@10°	@20°
1	46.0	62.7	76.2
2	47.5	64.0	77.1
3	47.3	63.8	76.7
4	**47.7**	**64.5**	**77.4**
5	47.1	63.7	77.0
6	46.9	63.6	76.6

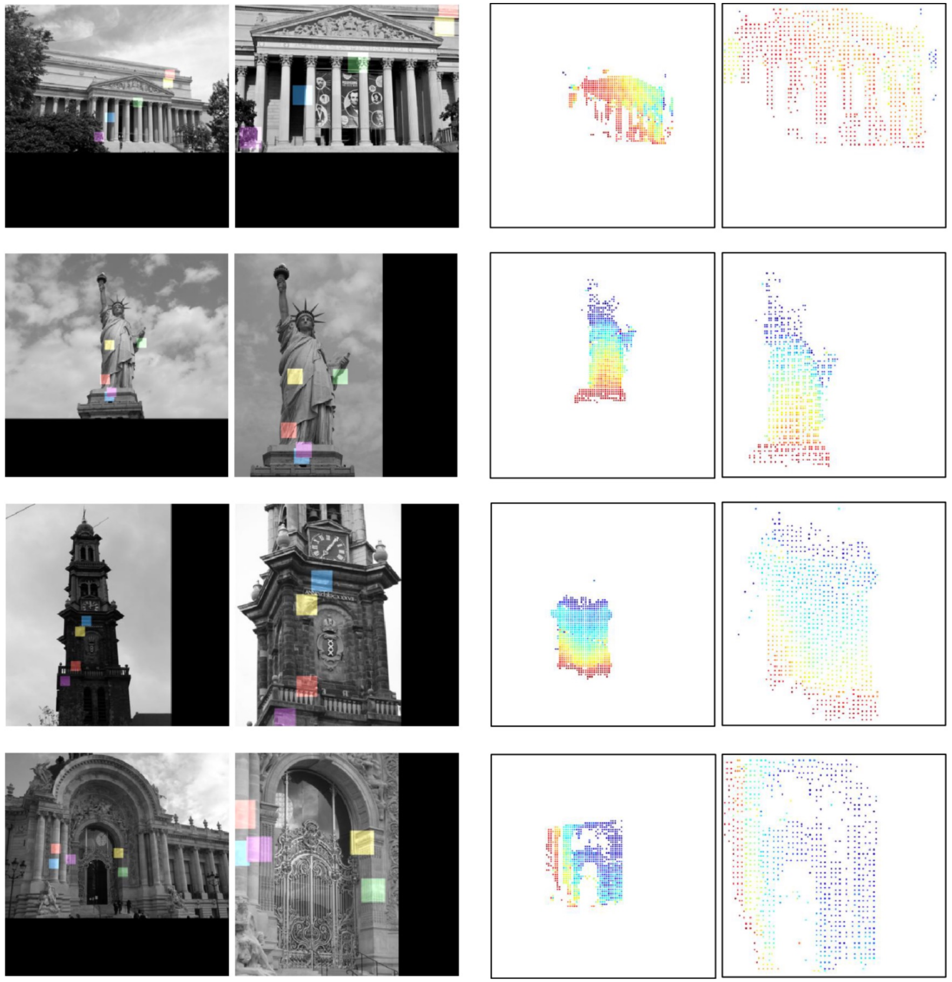
$l(k=4)$	Pose estimation AUC @5°	@10°	@20°
3	46.7	63.2	76.1
5	**47.7**	**64.5**	**77.4**
7	47.2	63.4	76.8
9	43.0	60.5	74.8

Contents

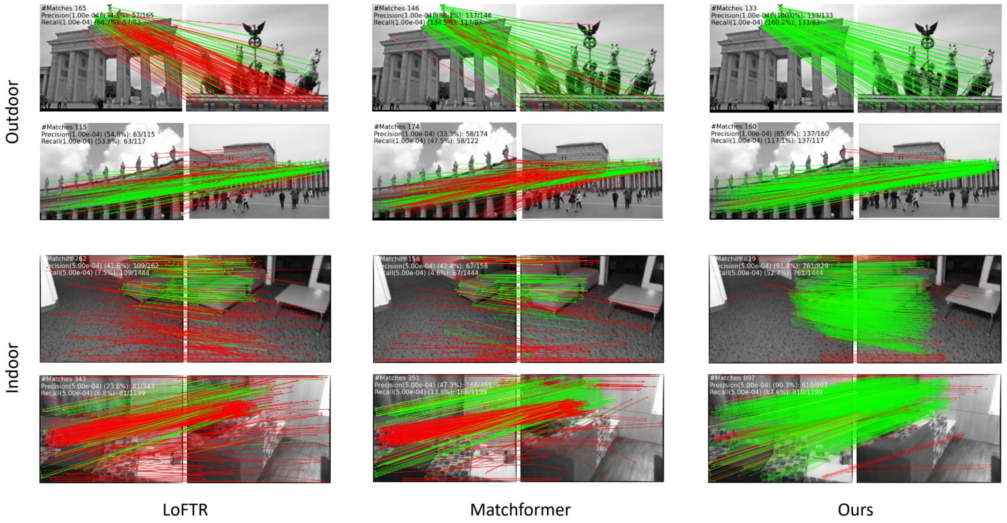
- Introduction
- Motivation
- Novelty
- Evaluation
- **Visualization**
- Conclusion

Visualization

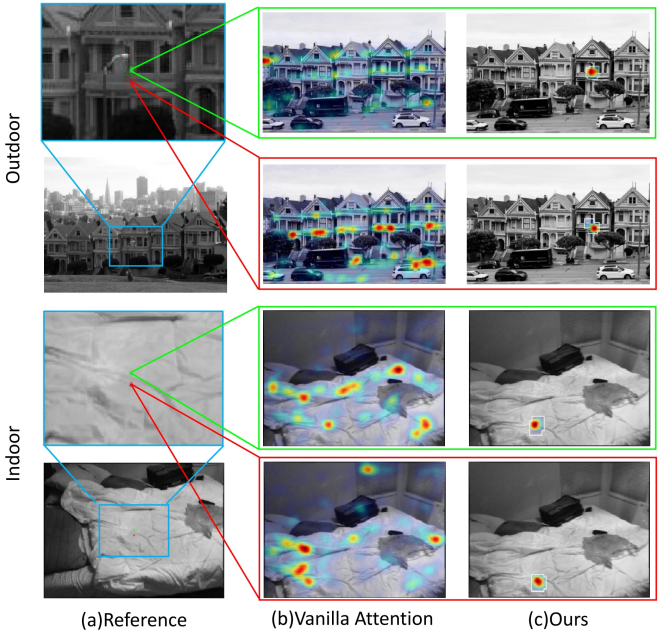
Fine Stage Window Scaling and Depth Estimation



Qualitative Comparison



Attention Heatmap



Contents

- Introduction
- Motivation
- Novelty
- Evaluation
- Visualization
- **Conclusion**

Conclusion

- A novel Adaptive Spot-Guided Transformer (**ASTR**) for local feature matching
- Two novel module:
  - **Spot-Guided Attention**: maintain local consistency, filter irrelevant attention areas
  - **Adaptive Scaling**: scale fine stage window to handle large scale variation
- SOTA performance in extensive experimental

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks!

Adaptive Spot-Guided Transformer for Consistent Local Feature Matching

Jiahuan Yu*, Jiahao Chang*, Jianfeng He, Tianzhu Zhang†, Feng Wu

University of Science and Technology of China

* Equal contribution † Corresponding author

Project Homepage:
https://astr2023.github.io