TUE-AM-276

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

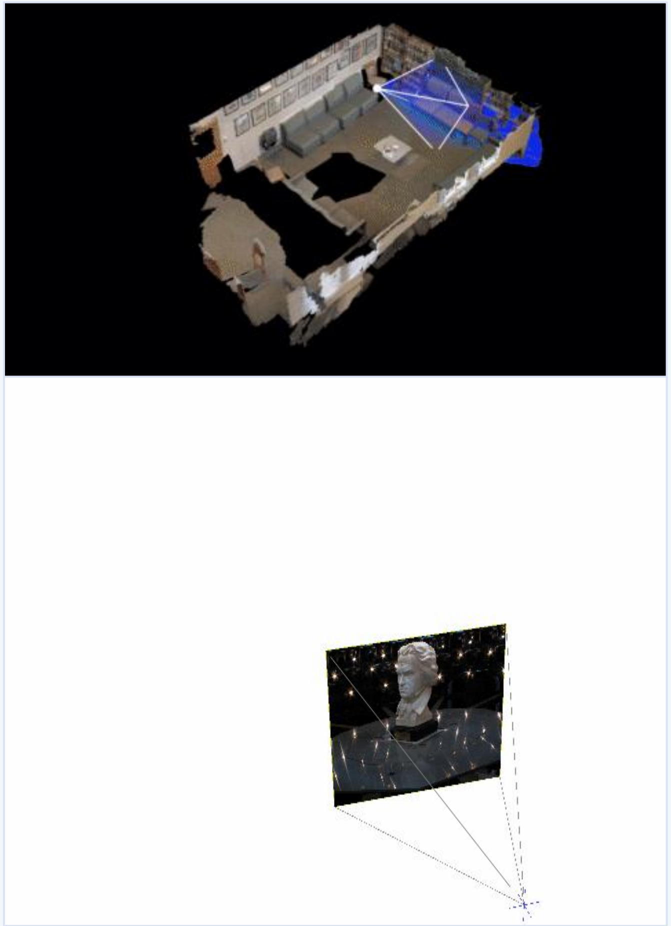
D²Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers

Jianfeng He[1*] Yuan Gao[1*] Tianzhu Zhang[1,2] Zhe Zhang[2] Feng Wu[1]

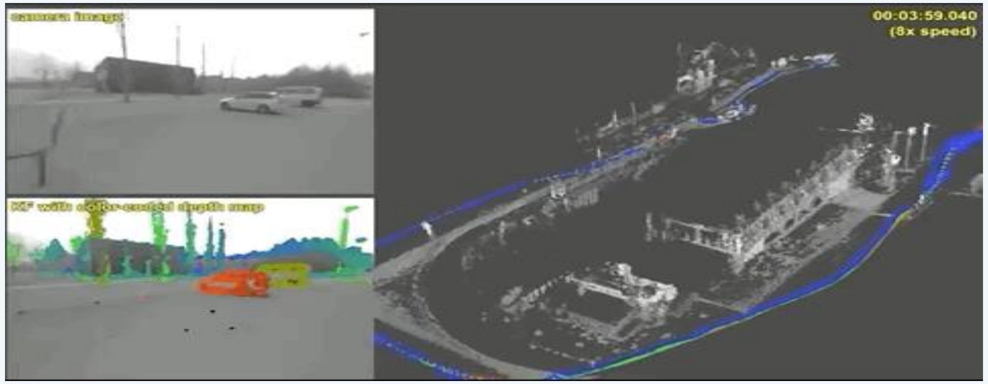
[1]University of Science and Technology of China
[2]Deep Space Exploration Laboratory

Background

Image matching has broad 3D vision applications



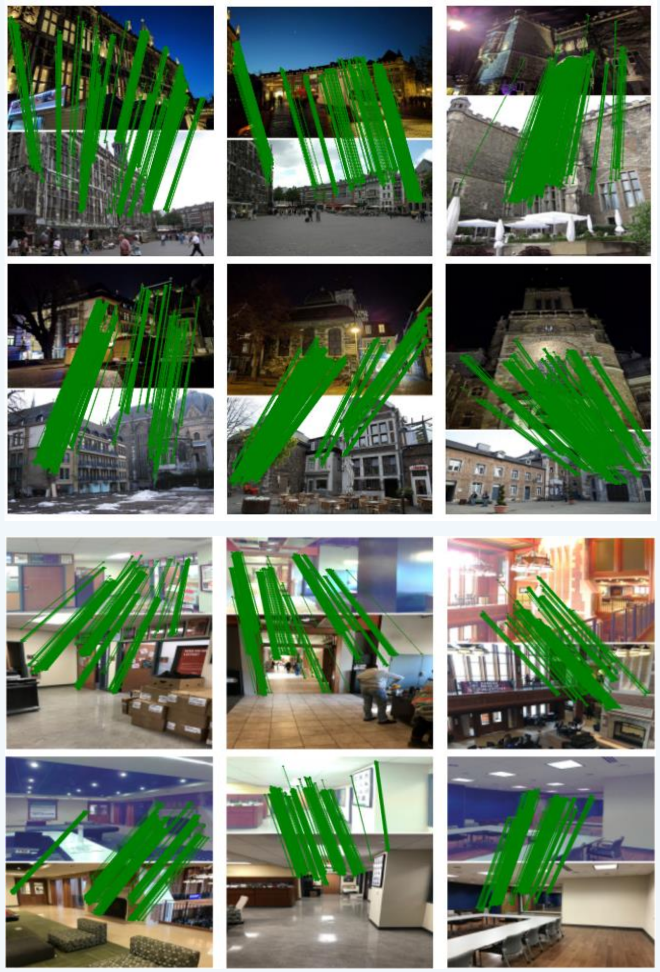
3D reconstruction



Simultaneous localization and mapping (SLAM)

Visual localization

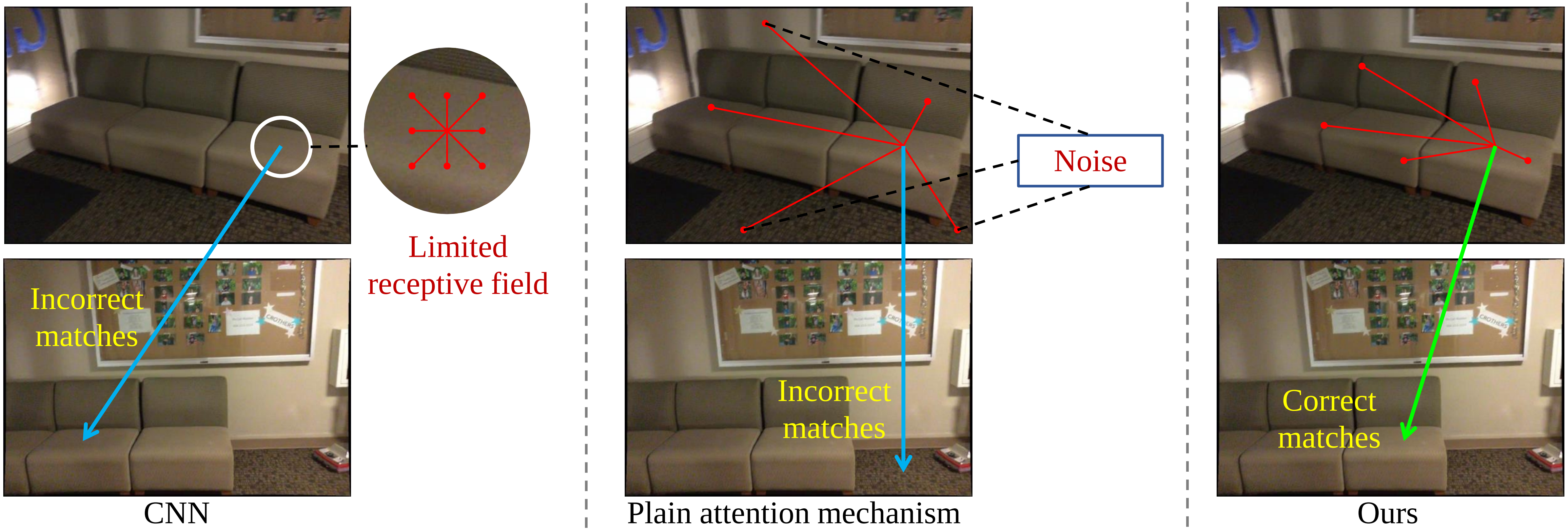
Matching Examples



Challenges

◻ Limited feature discriminability

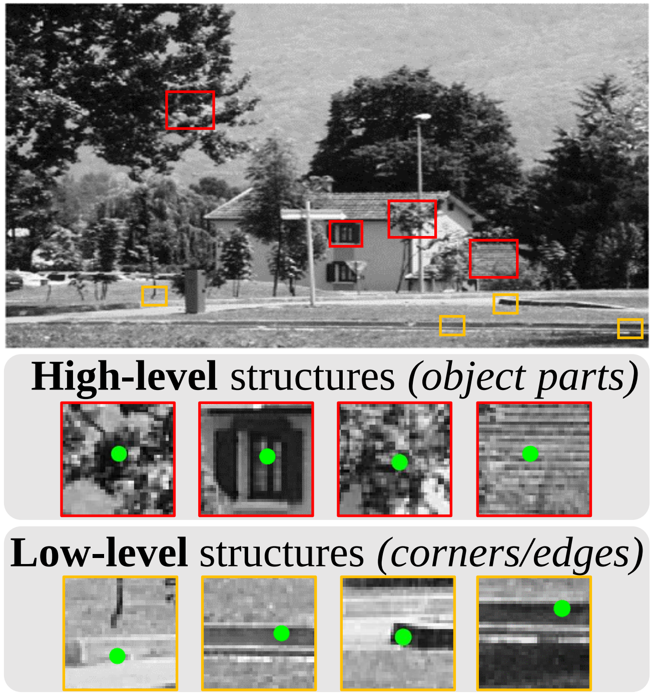
- The receptive field of features extracted by CNN is limited.
- The plain attention mechanism may aggregate irrelevant noise.
- The above ways to extract features would lack discriminative ability in texture-less regions.



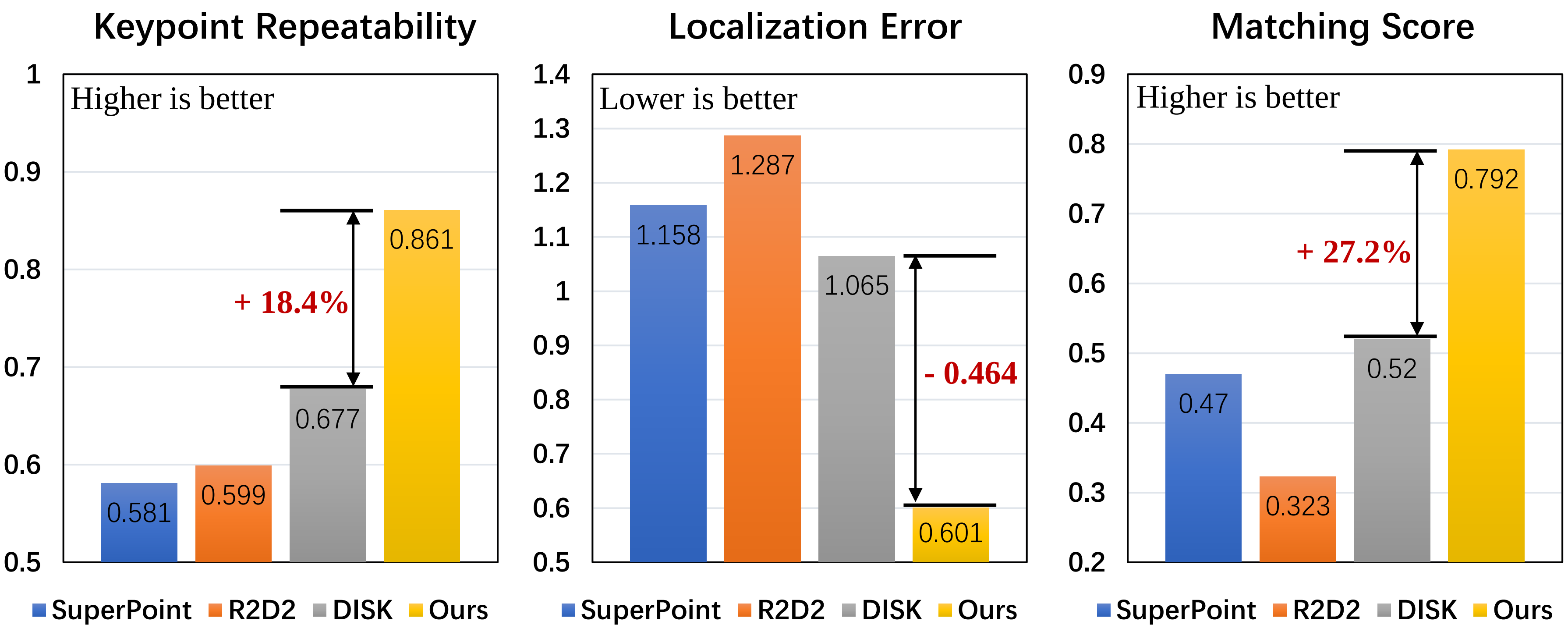
Challenges

◻ Detecting keypoints of different structures

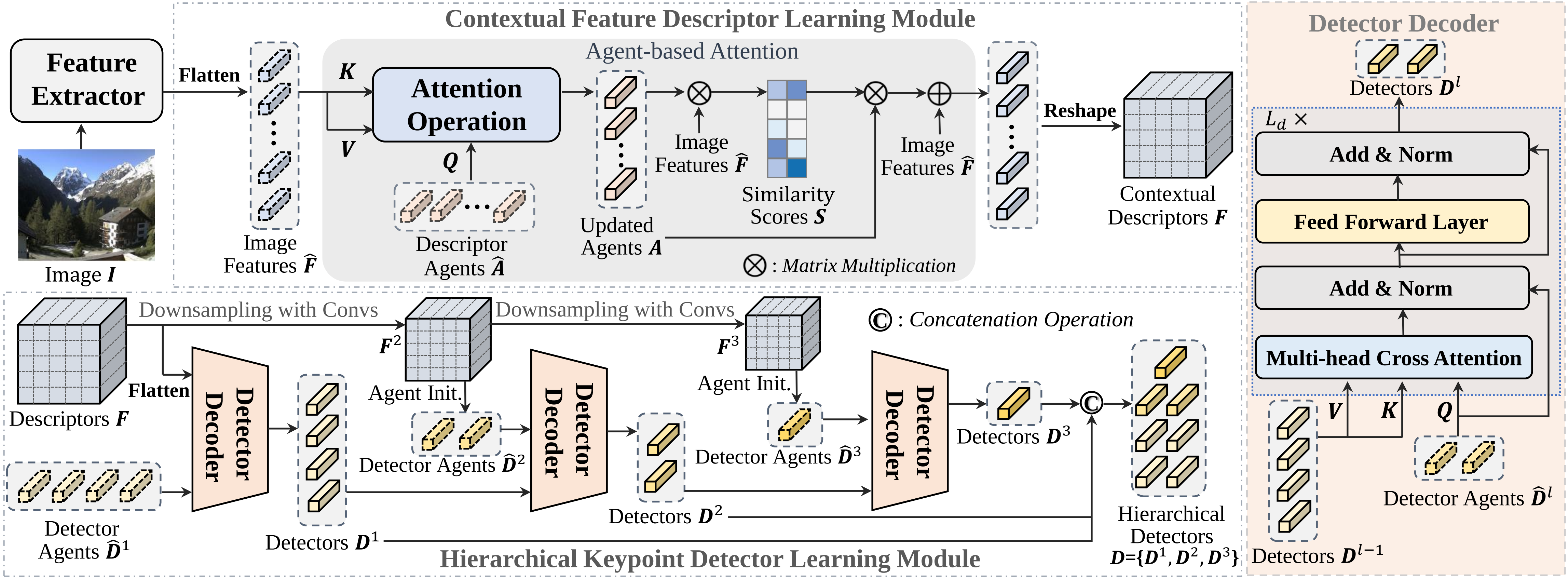
- There are diverse levels of structures in an image, from simple corner points to complex object parts.
- Existing keypoint detectors are usually good at identifying keypoints with a specific level of structure.
- The ability to detect keypoints with diverse levels of structures is needed.



Diverse levels of structures



Our Approach



We propose a novel image matching model **by jointly learning detectors and descriptors** via Agent-based Transformers，including a Contextual Feature Descriptor Learning Module and a Hierarchical Keypoint Detector Learning Module.

Contextual Feature Descriptor Learning (CFDL)

◻ Agent-based attention mechanism

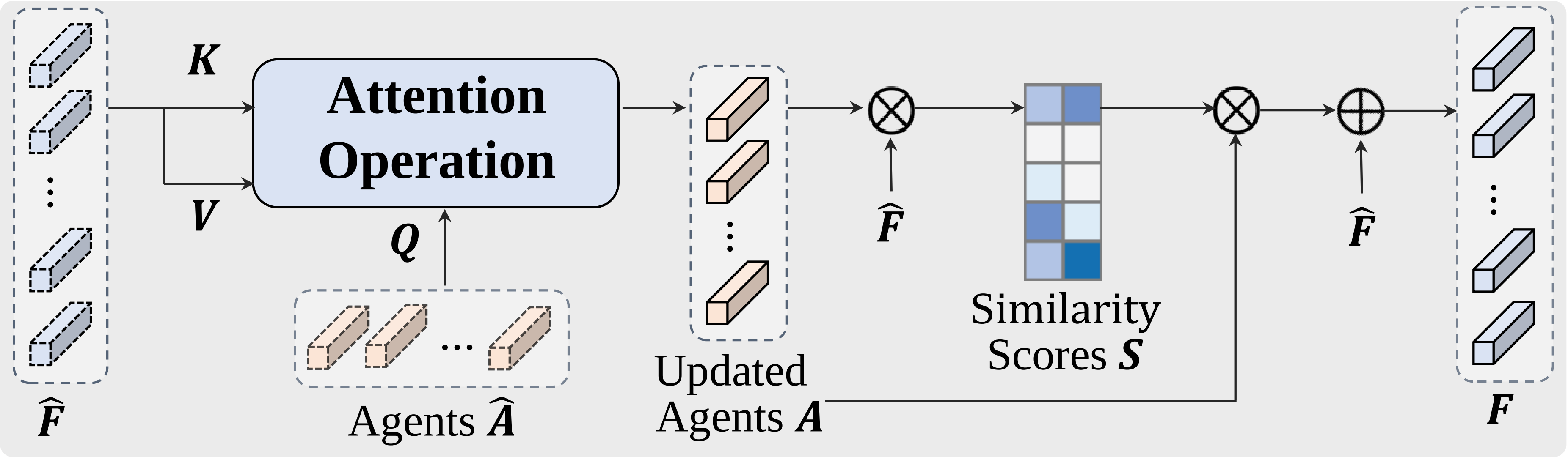
➢ Descriptor agents $\boldsymbol{A}$ are learned by interacting with image features $\widehat{\boldsymbol{F}}$ via the attention operation:

$$\boldsymbol{Q} = \boldsymbol{W}^{\mathcal{Q}}\widehat{\mathbf{A}}, \mathbf{K} = \boldsymbol{W}^{\mathcal{K}}\widehat{\boldsymbol{F}}, \boldsymbol{V} = \boldsymbol{W}^{\mathcal{V}}\widehat{\boldsymbol{F}}$$

$$\boldsymbol{A} = \mathbf{V} \cdot \mathrm{Softmax}(\boldsymbol{K}^{\boldsymbol{T}}\mathbf{Q})$$

➢ Contextual feature descriptors $\boldsymbol{F}$ are obtained by fusing $\boldsymbol{A}$ and $\widehat{\boldsymbol{F}}$:

$$\boldsymbol{F} = \widehat{\boldsymbol{F}} + \boldsymbol{AS}, \text{where } \boldsymbol{S} = \boldsymbol{A}^{\boldsymbol{T}}\widehat{\boldsymbol{F}}$$



Hierarchical Keypoint Detector Learning (HKDL)

◻ Agent Initialization

- Contextual Features $\boldsymbol{F}$ are down-sampled at each levels to obtain $\boldsymbol{F}^l$
- Multiple convolution layers are applied on $\boldsymbol{F}^l$ to produce masks
- Detector agents $\widehat{\boldsymbol{D}}^l$ are initialized via the mask pooling operation on $\boldsymbol{F}^l$

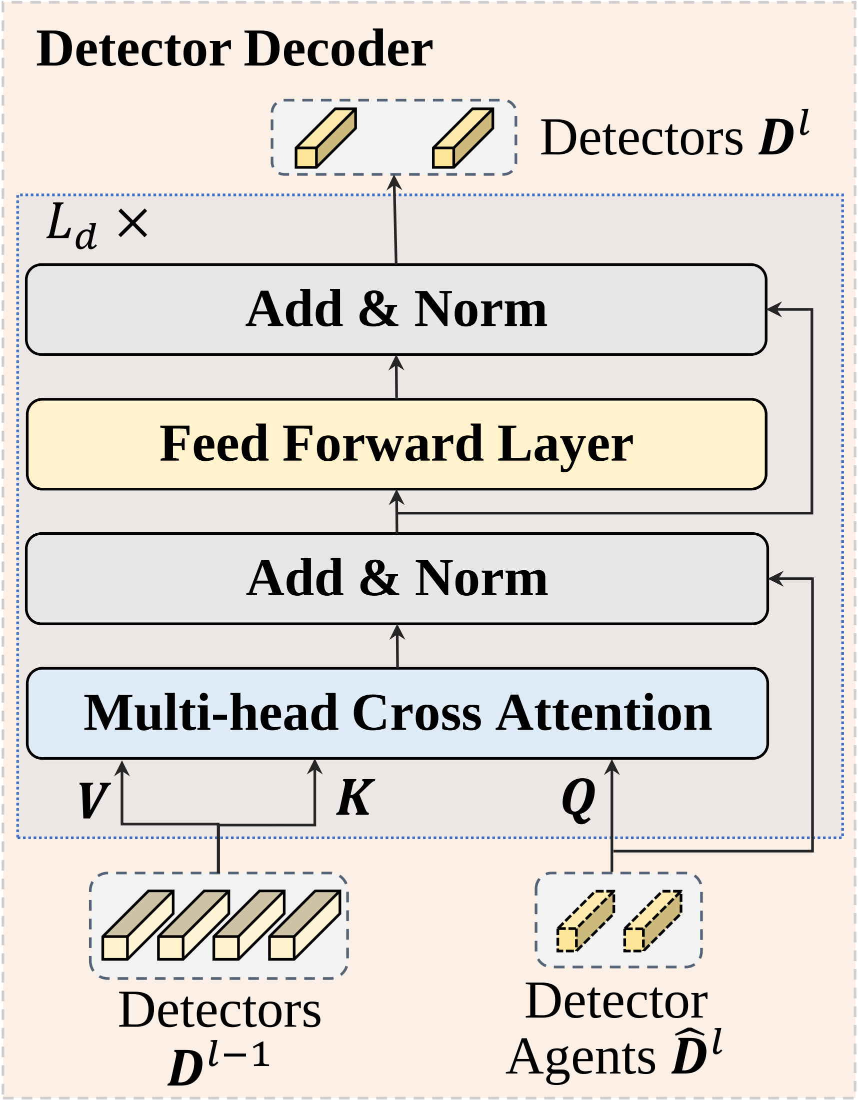
◻ Detector Decoder

- We aggregate low-level keypoint detectors to form high-level keypoint detectors in a hierarchical way:

$$\boldsymbol{D}^l = \text{Detector\_decoder}\left(\widehat{\boldsymbol{D}}^l, \boldsymbol{D}^{l-1}\right)$$

- Hierarchical keypoint detectors $\boldsymbol{D}$ are obtained by concatenating keypoint detectors $\boldsymbol{D}^l$ at different levels:

$$\boldsymbol{D} = concat(\{\boldsymbol{D}^l\}_{l=1}^{3}$$



Final Keypoint Detetcion and Matching

□ Keypoint Detection

- Given contextual descriptors $\boldsymbol{F}$ and hierarchical keypoint detectors $\boldsymbol{D}$, multiple score maps $\boldsymbol{S}_N$ are generated by the dot production: $\boldsymbol{S}_N = \boldsymbol{D}^T\boldsymbol{F}$
- The final keypoint detection score map $\boldsymbol{S}_c \in \mathbb{R}^{1\times h\times w}$ is obtained by averaging $\boldsymbol{S}_N$ on the first channel.
- Keypoints can be obtained by applying the local maxima filtering and the threshold constraint on the score map $\boldsymbol{S}_c$.

□ Keypoint Matching

- Given detected keypoints, their corresponding descriptors are acquired from the contextual feature maps $\boldsymbol{F}$.
- According to the keypoint feature distances, matches are established by the Nearest Neighbor (NN) matcher.

Experimental Results

❖ Quantitative Results

Results on the Hpatches dataset

Methods	AUC@3px	AUC@5px	AUC@10px
Sparse-NCNet [32]	48.9	54.2	67.1
DRC-Net [16]	50.6	56.2	68.3
LoFTR [42]	65 .9	75.6	84.6
D2-Net [12] + NN	23.2	35.9	53.6
R2D2 [31] + NN	50.6	63.9	76.8
DISK [47] + NN	52.3	64.9	78.9
SuperPoint [9] + SuperGlue [34]	53.9	68.3	81.7
D^{2}Former + NN (ours)	**71.6**	**81.3**	**89.7**

+ 5.7%
+ 5.7%
+ 5.1%

Results on the ScanNet dataset

Methods	AUC@5°	AUC@10°	AUC@20°
DRC-Net [16]	7.69	17.93	30.49
LoFTR [42]	22.06	40.80	57.62
ASpanFormer [7]	25.60	46.00	63.30
D2-Net [12] + NN	5.25	14.53	27.96
R2D2 [31] + NN	7.43	17.45	28.64
SuperPoint [9] + NN	9.43	21.53	36.40
SuperPoint [9] + PointCN [52]	11.40	25.47	41.41
SuperPoint [9] + OANet [53]	11.76	26.90	43.85
SuperPoint [9] + SuperGlue [34]	16.16	33.81	51.84
D^{2}Former + NN (ours)	**31.03**	**51.69**	**69.17**

+ 5.43%
+ 5.69%
+ 5.87%

Results on the YFCC100M dataset

Methods	AUC@5°	AUC@10°	AUC@20°
LoFTR [42]	40.28	61.17	77.80
SIFT [18] + SuperGlue [34]	30.49	51.29	69.72
R2D2 [31] + NN	33.85	52.44	68.53
SuperPoint [9] + NN	16.94	30.39	45.72
SuperPoint [9] + OANet [53]	26.82	45.04	62.17
SuperPoint [9] + SuperGlue [34]	38.72	59.13	75.81
D^{2}Former + NN (ours)	**56.78**	**73.71**	**85.37**

+ 16.50%
+ 12.54%
+ 7.57%

Results on the MegaDepth dataset

Methods	AUC@5°	AUC@10°	AUC@20°
DRC-Net [16]	27.01	42.96	58.31
LoFTR [42]	52.80	69.19	81.18
ASpanFormer [7]	55.30	71.50	83.10
R2D2 [31] + NN	37.14	55.09	69.65
SuperPoint [9] + SuperGlue [34]	42.18	61.16	75.96
D^{2}Former + NN (ours)	**66.27**	**78.44**	**86.81**

+ 10.97%
+ 6.94%
+ 3.71%

Effectiveness of each component on the ScanNet

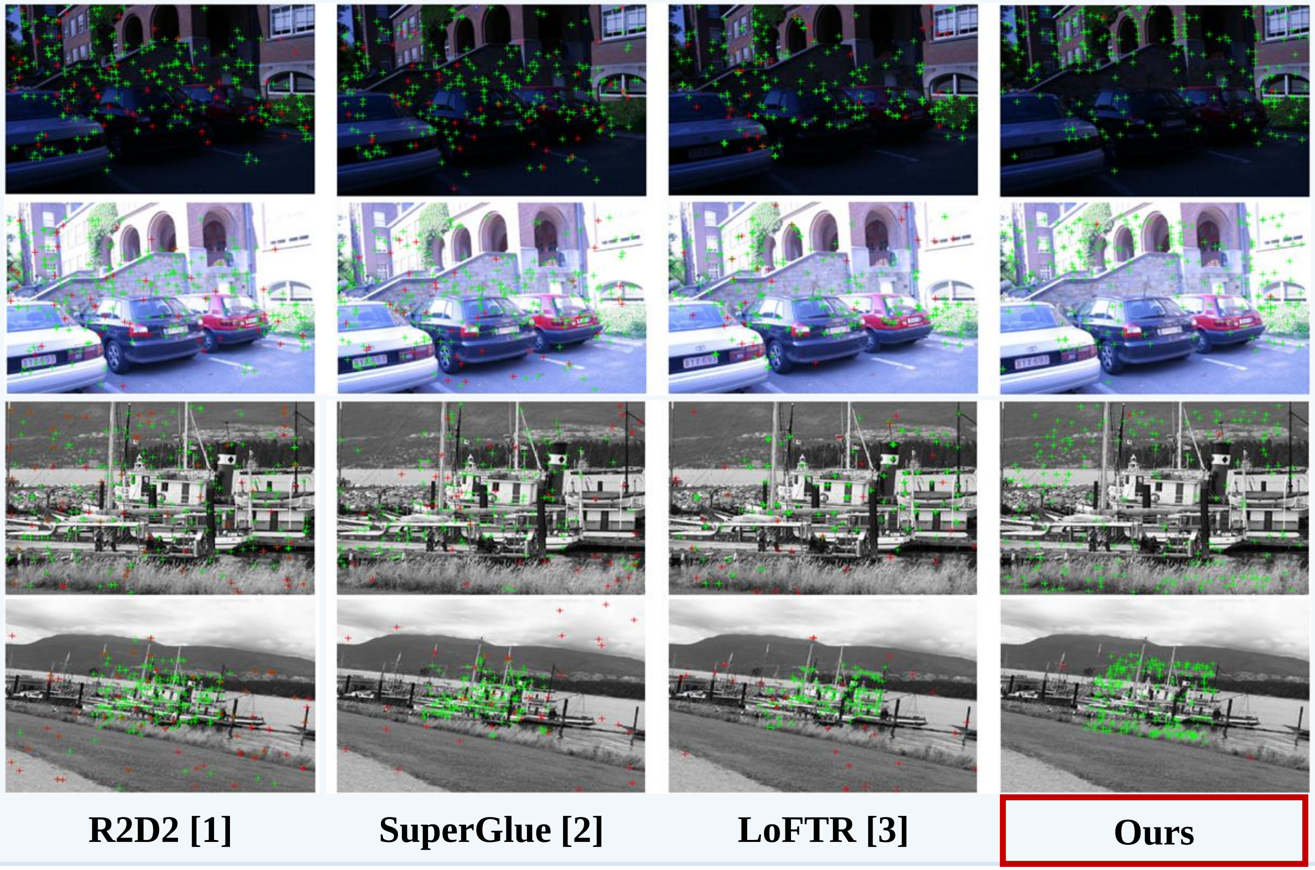
Models	HKDL	CFDL	AUC@5°	AUC@10°	AUC@20°
[A]	✗	✗	7.43	17.45	28.64
[B]	✗	✓	18.68	36.49	55.17
[C]	✓	✗	27.64	48.34	67.05
[D]	✓	✓	**31.03**	**51.69**	**69.17**

+ 23.60%
+ 34.24%
+ 40.53%

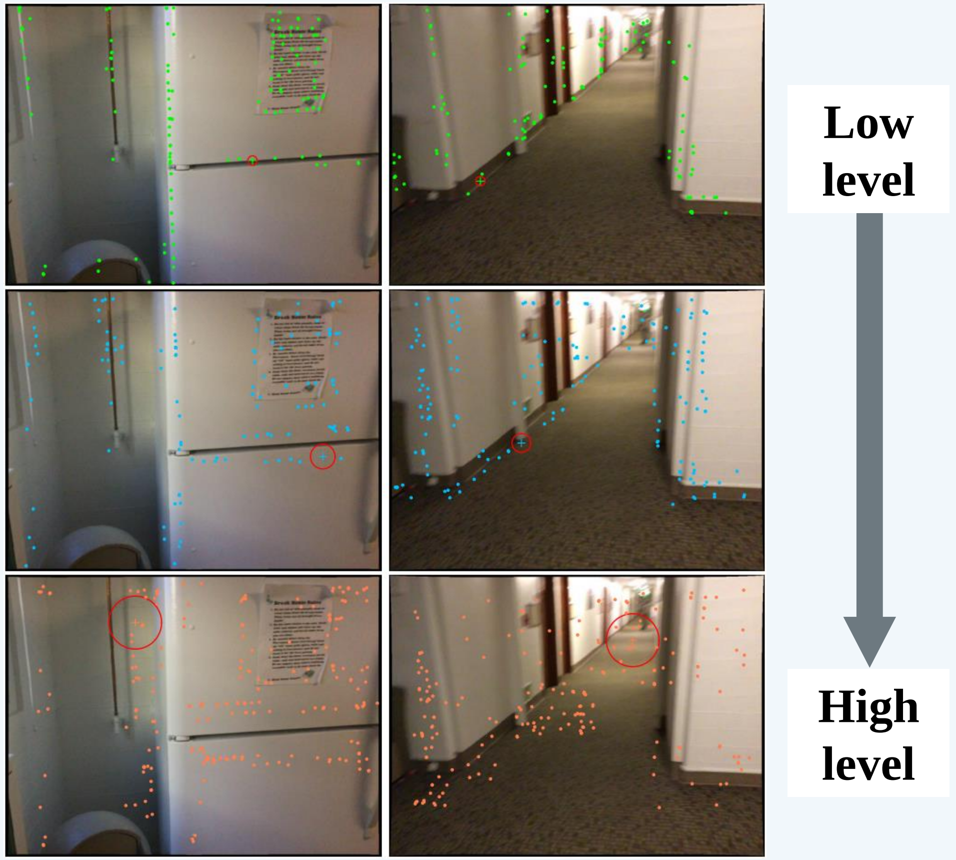
Experimental Results

❖ Qualitative Results

Qualitative comparisons with previous state-of-the-art methods



Keypoint detection results for different levels



[1] Revaud J, De Souza C, Humenberger M, et al. R2d2: Reliable and repeatable detector and descriptor[J]. Advances in neural information processing systems, 2019, 32.
[2] Sarlin P E, DeTone D, Malisiewicz T, et al. Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2020: 4938-4947.
[3] Sun J, Shen Z, Wang Y, et al. LoFTR: Detector-free local feature matching with transformers. Proceedings of the IEEE conference on computer vision and pattern recognition. 2021: 8922-8931.

Conclusion

- We propose a novel image matching model by **Jointly Learning** Hierarchical Detectors and Contextual Descriptors via **Agent-based Transformers.**

- D^2Former can extract **discriminative features** and realize **robust keypoint detection** under some extremely challenging scenarios.

- **D^2Former** outperforms previous state-of-the-art methods by **a large margin** on four challenging benchmarks.

TUE-AM-276

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

D2Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers

Thanks for watching!

Jianfeng He[1*] Yuan Gao[1*] Tianzhu Zhang[1,2] Zhe Zhang[2] Feng Wu[1]

[1]University of Science and Technology of China
[2]Deep Space Exploration Laboratory

DEEP SPACE EXPLORATION LAB
深空探测实验室