ByteDance

UCF Center for Research in Computer Vision

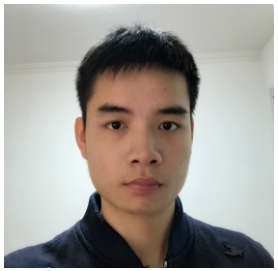
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

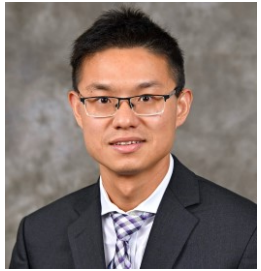
THU-AM-275 Highlight

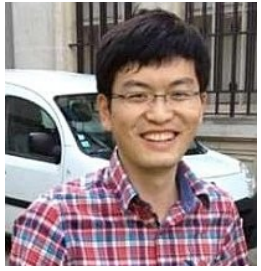
$R^2Former$: Unified Retrieval and Reranking Transformer for Place Recognition

Sijie Zhu[1,2], Linjie Yang[1], Chen Chen[2], Mubarak Shah[2], Xiaohui Shen[1], Heng Wang[1]

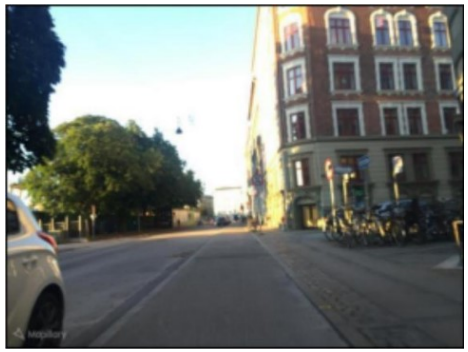
[1]ByteDance [2]Center for Research in Computer Vision, University of Central Florida





Visual Place Recognition (VPR)



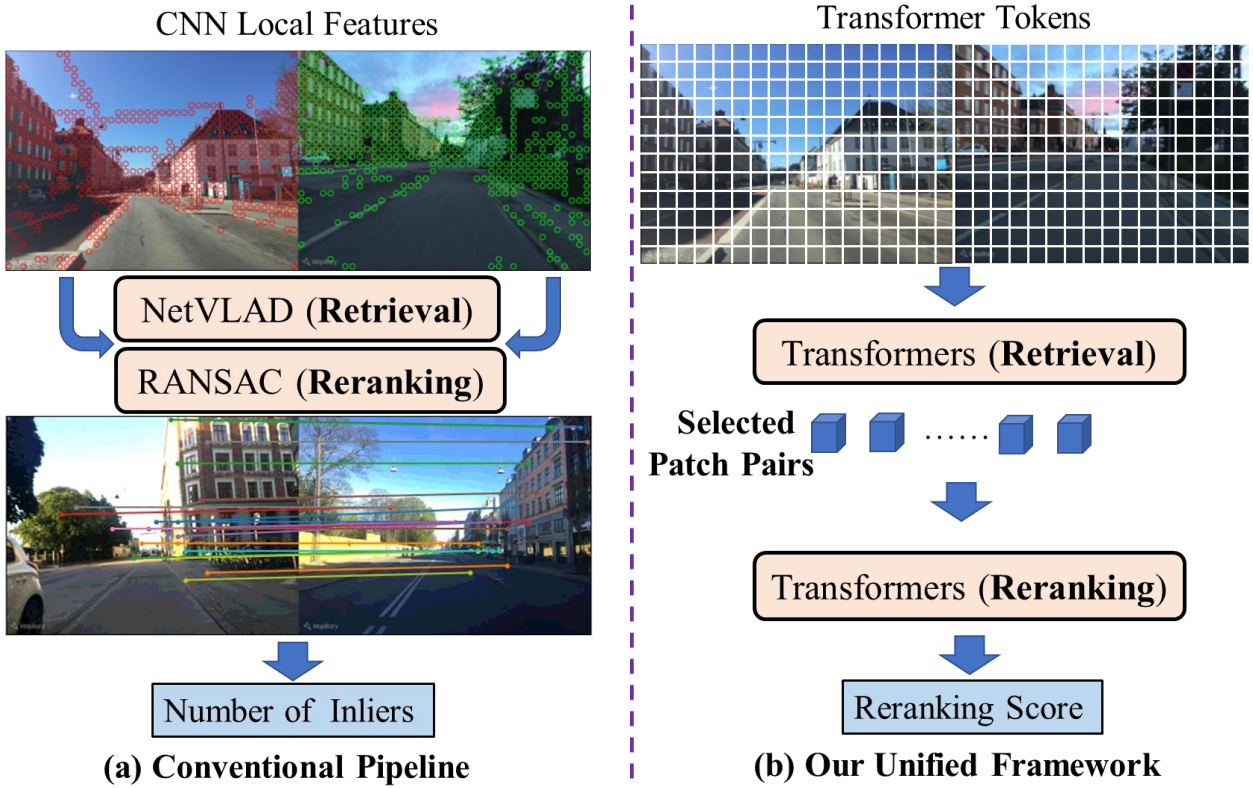
Query Image from
Unknown Location

Search

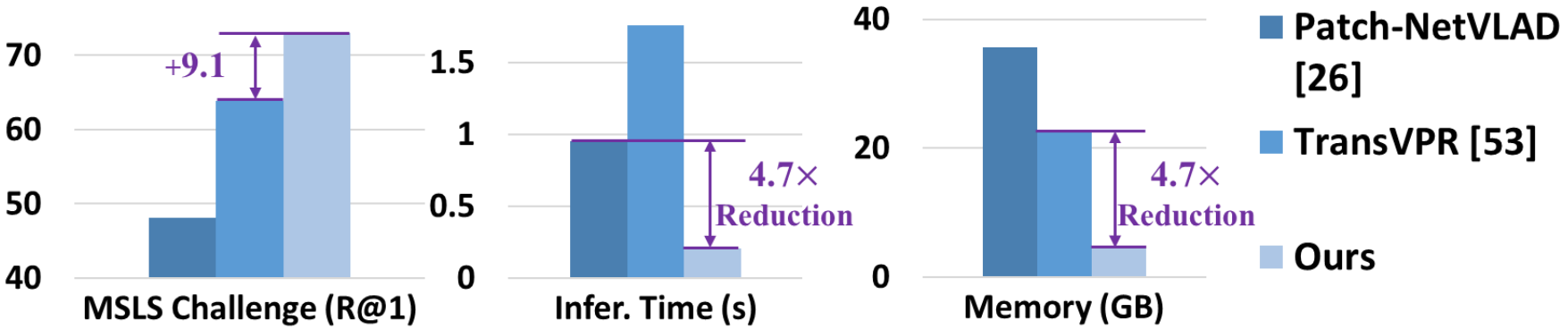
Reference Images from
Known Locations

Retrieved Image

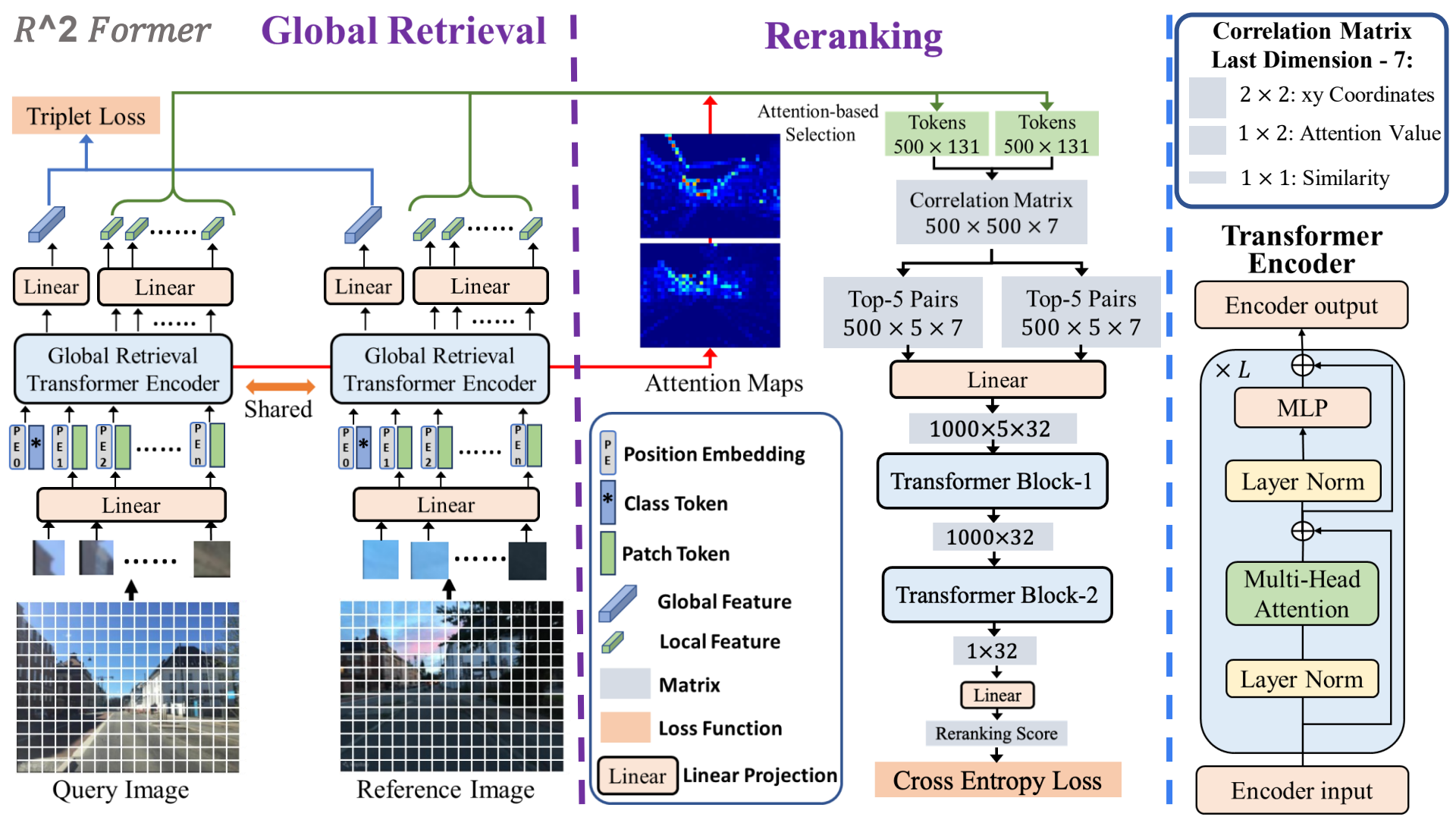
$R^2Former$ vs Conventional Pipeline



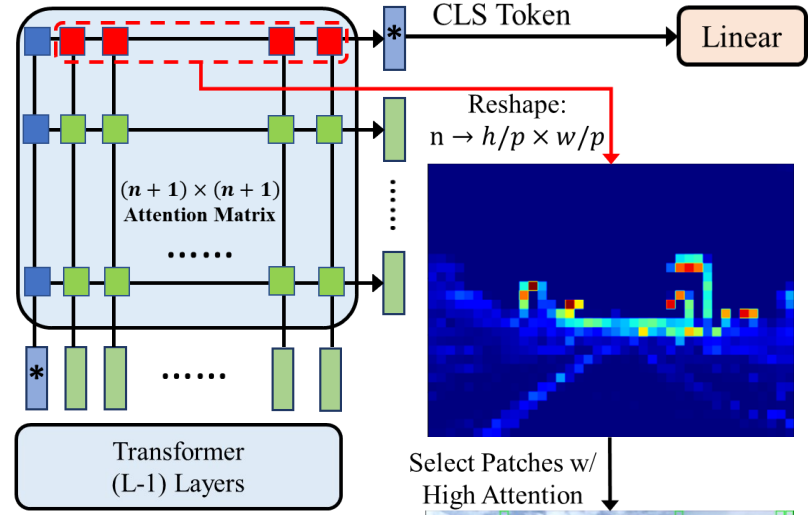
Effective and Efficient

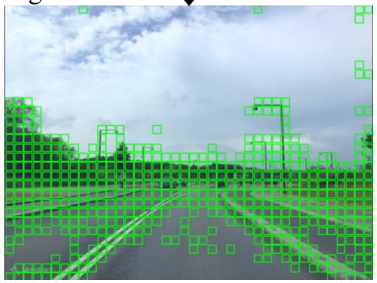


(c) Comparison with State-of-the-art Methods


R^2 Former
Global Retrieval
Reranking
Triplet Loss
Linear
Global Retrieval Transformer Encoder
Shared
Query Image
Reference Image
Attention-based Selection
Attention Maps
Tokens 500 × 131
Correlation Matrix 500 × 500 × 7
Top-5 Pairs 500 × 5 × 7
1000×5×32
Transformer Block-1
1000×32
Transformer Block-2
1×32
Reranking Score
Cross Entropy Loss
Position Embedding
Class Token
Patch Token
Global Feature
Local Feature
Matrix
Loss Function
Linear Projection
Correlation Matrix Last Dimension - 7:
2 × 2: xy Coordinates
1 × 2: Attention Value
1 × 1: Similarity
Transformer Encoder
Encoder output
× L
MLP
Layer Norm
Multi-Head Attention
Encoder input

Attention Map



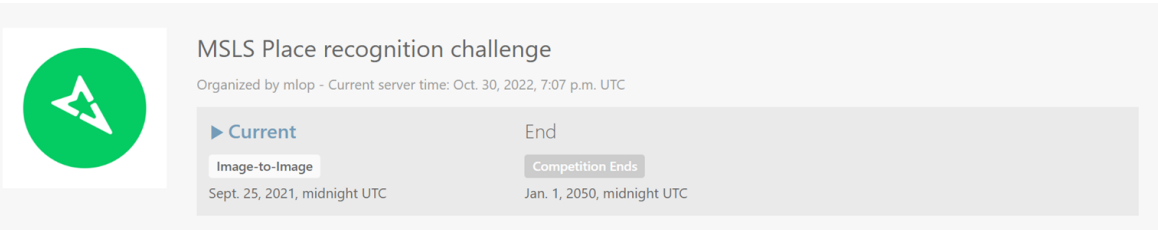


Performance on Major Datasets

	MSLS Val [55]			MSLS Challenge [55]			Pitts30k [50]			Tokyo 24/7 [49]		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [3]	60.8	74.3	79.5	35.1	47.4	51.7	81.9	91.2	93.7	64.8	78.4	81.6
SFRS [23]	69.2	80.3	83.1	41.5	52.0	56.3	89.4	94.7	95.9	85.4	91.1	**93.3**
SP-SuperGlue [15, 44]	78.1	81.9	84.3	50.6	56.9	58.3	87.2	94.8	**96.4**	88.2	90.2	90.2
Patch-NetVLAD [26]	79.5	86.2	87.7	48.1	57.6	60.5	88.7	94.5	95.9	86.0	88.6	90.5
TransVPR [53]	86.8	91.2	92.4	63.9	74.0	77.5	89.0	94.9	96.2	79.0	82.2	85.1
Ours	**89.7**	**95.0**	**96.2**	**73.0**	**85.9**	**88.8**	**91.1**	**95.2**	96.3	**88.6**	**91.4**	91.7

UCF

Top-1 Result on MSLS Challenge



MSLS Place recognition challenge

Organized by mlop - Current server time: Oct. 30, 2022, 7:07 p.m. UTC

▶ Current

Image-to-Image

Sept. 25, 2021, midnight UTC

End

Competition Ends

Jan. 1, 2050, midnight UTC

#	User	Entries	Date of Last Entry	recall@5 ▲
1	SijieZhu	1	03/14/23	0.88 (1)
2	changxinyuan.cxy	11	07/27/22	0.82 (2)
3	izquierdo	9	05/18/23	0.80 (3)
4	gberton	2	04/21/22	0.80 (4)
5	sobremesa	10	03/01/22	0.77 (5)
6	Jincheng2	3	10/16/22	0.77 (6)
7	MAX-OTW3	9	10/16/22	0.76 (7)
8	lijinchengECN	3	10/10/22	0.74 (8)
9	Cheng	15	10/22/22	0.74 (9)
10	qilongwu	5	04/04/23	0.74 (10)
11	jiang_163	4	05/23/23	0.73 (11)
12	lib2000	4	02/20/22	0.71 (12)
13	LSL10	3	04/07/23	0.69 (13)
14	Jincheng_LI	5	10/10/22	0.67 (14)
15	haiyang_hit	11	05/06/23	0.51 (15)

UCF

Computational Efficiency

	Feature Dim ↓		Latency per Query (ms) ↓			Memory Footprint (GB) ↓	
	Global	Local	Extraction	Retrieval	Reranking	MSLS Val	1M Images
ResNet101 + NetVLAD [3, 6]	65536	N/A	9.60	2.33	N/A	4.79	244.14
Patch-NetVLAD-s [26]	512	936×512	9.29	0.08	952.85	37.60	1917.29
Patch-NetVLAD-p [26]	4096	2826×4096	9.36	0.19	8377.17	908.30	46315.85
TransVPR [53]	**256**	1200×256	**6.20**	**0.07**	1757.70	22.72	1158.53
Ours	**256**	**$500 \times (128+3)$**	8.81	**0.07**	**202.37**	**4.79**	**244.01**

UCF

Comparison with Other Reranking Methods

	R@1	R@5	R@10
No Reranking	79.3	90.8	92.6
RANSAC [19]	84.9	93.0	94.5
RRT [48]	81.2	91.9	93.1
CVNet [32]	73.4	86.8	91.4
Ours	**89.7**	**95.0**	**96.2**

Transformer Token vs CNN Local Feature

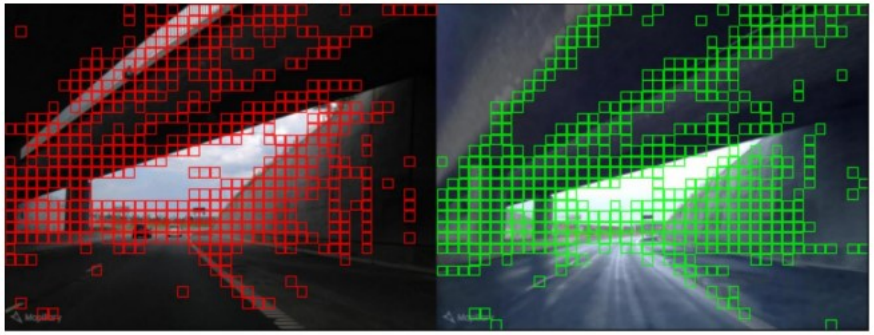
	Architecture	R@1	R@5	R@10
Ours w/o Reranking	ViT-Small	79.3	90.8	92.6
	ResNet50 + GeM	79.6	90.9	92.6
	ViT-Base	84.9	92.7	94.5
Ours w/ RANSAC	ViT-Small	84.9	93.0	94.5
	ResNet50 + GeM	84.3	91.4	93.0
	ViT-Base	87.0	93.0	94.6
Ours	ViT-Small	89.7	95.0	96.2
	ResNet50 + GeM	88.4	93.6	95.3
	ViT-Base	**90.0**	**95.1**	**96.9**

UCF

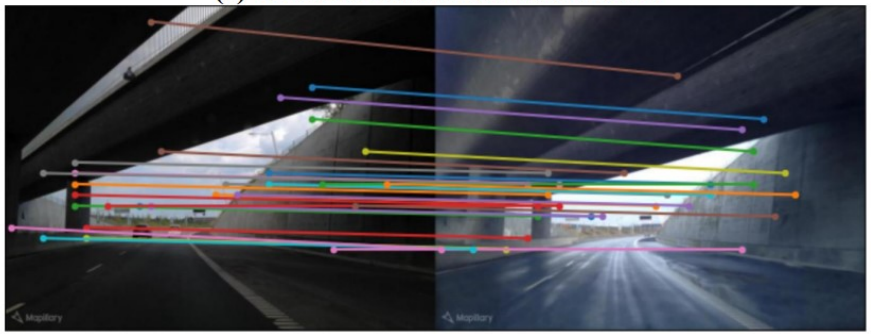
Interpretability

(a) Image Pair

(b) Selected Tokens



(c) RANSAC Matched Local Pairs



(d) Ours Top-20 Highlighted Local Pairs

Summary

- A unified retrieval and reranking framework for place recognition employing only transformers, which demonstrates that vision transformer tokens are comparable and sometimes better than CNN local features in terms of reranking or local matching.

- A novel transformer-based reranking module that learns to attend to the correlation of informative local feature pairs. It can be combined with either CNN or transformer backbones with better performance and efficiency than other reranking methods, e.g. RANSAC.

- Code: https://github.com/Jeff-Zilence/R2Former

Thank You!

UCF