CrossScore: Towards Multi-View Image Evaluation and Scoring (Supplementary)

Zirui Wang Wenjing Bian Victor Adrian Prisacariu
University of Oxford

A Additional Practical Details

At test time, evaluating images at resolution 518×690 with batch size 16 takes 20GB GPU memory with an average time consumption of 69ms/img on an NVIDIA RTX 4090 GPU. Our model consists of 26M parameters, of which 23M are from the pre-trained DINOv2-small backbone Φ_{enc}, and 3M are from our cross attention module Φ_{cross} and our score regression head Φ_{dec}.

B Additional Quantitative Details for Table 1

We offer detailed results for Tab. 1 in Tables 5 to 7. For no-reference baselines, assessments are conducted using Matlab with their default settings and feature models.

Table 5: *Correlation between various metrics and SSIM on the Map-Free Relocalisation (MFR) dataset.*

	MFR					
Scene	FR		NR			CR
	SSIM ↑	PSNR ↑	BRISQUE ↓	NIQE ↓	PIQE ↓	Ours ↑
s00004	0.40	15.88	19.40	3.00	34.10	0.46
s00010	0.64	19.16	20.82	3.21	28.39	0.72
s00034	0.66	21.91	25.23	2.66	29.74	0.72
s00082	0.44	16.35	23.47	2.68	34.54	0.48
s00103	0.59	16.43	30.50	3.16	47.39	0.64
s00135	0.64	20.12	20.90	2.88	42.21	0.75
s00175	0.51	17.32	24.83	3.24	31.79	0.56
s00238	0.61	16.74	27.15	2.74	34.17	0.61
s00244	0.50	18.03	25.57	3.24	42.06	0.66
s00284	0.61	19.46	25.79	2.78	30.60	0.58
s00311	0.55	17.82	24.08	3.21	26.75	0.65
s00345	0.56	18.82	19.71	2.83	41.05	0.72
s00426	0.74	22.10	24.41	2.66	32.16	0.82
s00441	0.58	20.36	26.03	2.51	39.36	0.71
Correlation	1.00	0.78	0.23	-0.30	-0.11	0.83

Table 6: *Correlation between various metrics and SSIM on the Mip360 dataset.*

Mip360						
Scene	FR		NR			CR
	SSIM ↑	PSNR ↑	BRISQUE ↓	NIQE ↓	PIQE ↓	Ours ↑
bicycle	0.85	26.66	22.30	2.67	33.41	0.82
bonsai	0.95	32.48	27.08	3.46	52.90	0.89
counter	0.92	29.82	25.61	2.78	50.57	0.87
flowers	0.72	25.31	26.53	2.57	31.46	0.64
garden	0.92	31.19	13.18	2.37	31.38	0.87
kitchen	0.95	32.48	30.73	3.03	43.82	0.85
room	0.94	33.42	33.43	2.93	53.95	0.91
stump	0.83	30.43	22.52	2.97	21.35	0.81
treehill	0.74	25.25	23.58	2.30	31.22	0.73
Correlation	1.00	0.91	0.19	0.61	0.69	0.95

Table 7: *Correlation between various metrics and SSIM on the RealEstate10K (RE10K) dataset.*

RealEstate10K						
Scene	FR		NR			CR
	SSIM ↑	PSNR ↑	BRISQUE ↓	NIQE ↓	PIQE ↓	Ours ↑
00407b3f1bad1493	0.90	26.06	44.33	3.70	69.66	0.91
004ed278c2b168f1	0.73	20.13	53.48	4.39	54.82	0.77
0065a058603dfca4	0.88	22.98	49.01	4.18	75.67	0.90
00703cbf7531ef11	0.56	17.74	30.67	2.57	42.89	0.67
00761c6dcec91853	0.95	31.34	44.70	3.83	62.81	0.93
007ac6cef80a692c	0.90	22.76	33.68	3.25	70.71	0.91
0081cfd790d7ad74	0.02	10.45	NaN	NaN	NaN	0.23
009664cb1b8d351a	0.74	17.00	52.77	4.39	82.10	0.74
00a50bfbce75d465	0.86	23.86	38.82	3.22	65.02	0.88
00a9f110ad222aa4	0.81	22.34	32.61	2.25	49.61	0.82
00b52b21e0d54a42	0.89	22.64	43.64	3.70	72.21	0.90
00b9a7963f9bd9c6	0.37	14.60	34.06	3.30	67.52	0.38
00c8250efd605554	0.15	8.73	32.17	2.98	60.74	0.19
Correlation	1.00	0.92	0.46	0.32	0.27	0.99

C Additional Qualitative Results

We invite readers to check out a video with additional qualitative results on our project page: https://crossscore.active.vision.

D Discussion: Relationships with Visual Place Recognition (VPR) Systems

One straightforward way to evaluate an image with a full-reference metric such as SSIM without aligned ground truth is to utilise a nearby frame, for example, a temporal neighbour, or a visually similar frame obtained from image retrieval or visual place recognition (VPR) systems. Fig. 8 and Tab. 8 demonstrate that SSIM scores computed using misaligned images (nearest frames) are significantly different from GT SSIM scores, whereas our multi-view-based scores are similar to GT SSIM scores.

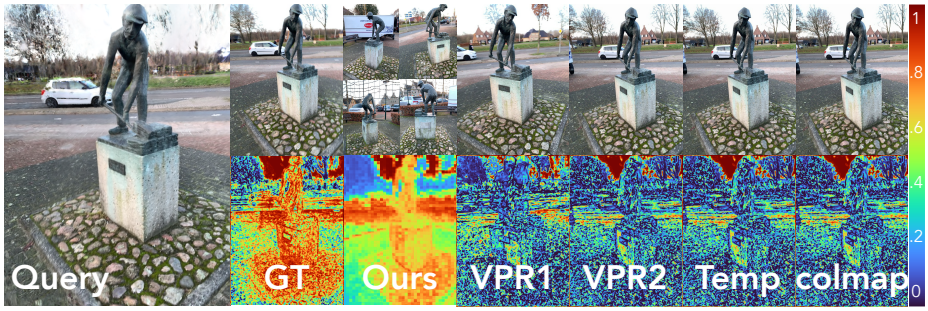

Fig. 8: Our method vs. computing SSIM between a query image and a nearest frame*. The nearest frame is selected through various strategies:* ***VPR1****: SALAD [19],* ***VPR2****: CricaVPR [27],* ***Temp****: temporal nearest frame, and* ***colmap****: vocabulary-tree-based image retrieval module in COLMAP [45]. We show that the SSIM scores computed using misaligned images (nearest frames) are significantly different from GT SSIM scores, whereas our multi-view-based scores are similar to GT SSIM scores.*

Table 8: Correlation between GT SSIM, our score, and SSIM scores computed using various nearest frames. *Each nearest frame is selected through the following strategies:* ***VPR1****: SALAD [19],* ***VPR2****: CricaVPR [27],* ***Temp****: temporal nearest frame, and* ***COLMAP****: vocabulary-tree-based image retrieval module in COLMAP [45]. We show that the SSIM scores computed using misaligned images (nearest frames) are significantly different from GT SSIM scores, whereas our multi-view-based scores are similar to GT SSIM scores.*

	GT	Ours	VPR1 [19]	VPR2 [27]	Temp	COLMAP [45]
Corr	1.0	**0.83**	0.37	0.38	0.37	0.39

E Social Impact

Our cross-reference image quality assessment method has limited negative social impact. It enhances image evaluations for applications like novel view synthesis without using human data, thus avoiding privacy issues. Our method does not facilitate harmful activities and focuses on technical improvements. With low misuse potential and significant benefits for fields like computer graphics and virtual reality, this advancement positively impacts technological and creative industries without significant ethical concerns.