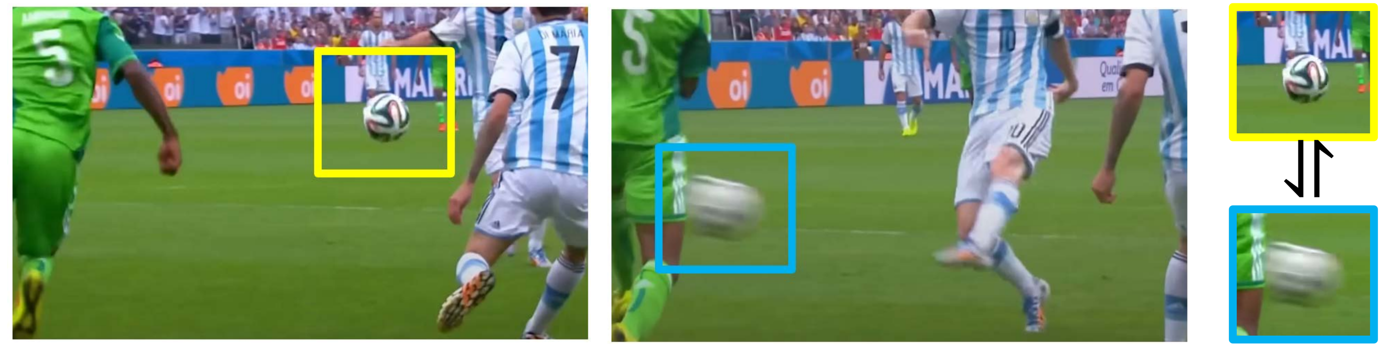
Retrieval Robust to Object Motion Blur

Rong Zou[1], Marc Pollefeys[1,2], Denys Rozumnyi[1,3]
[1]ETH Zürich, [2]Microsoft, [3]Czech Technical University in Prague

Code, Data, Model

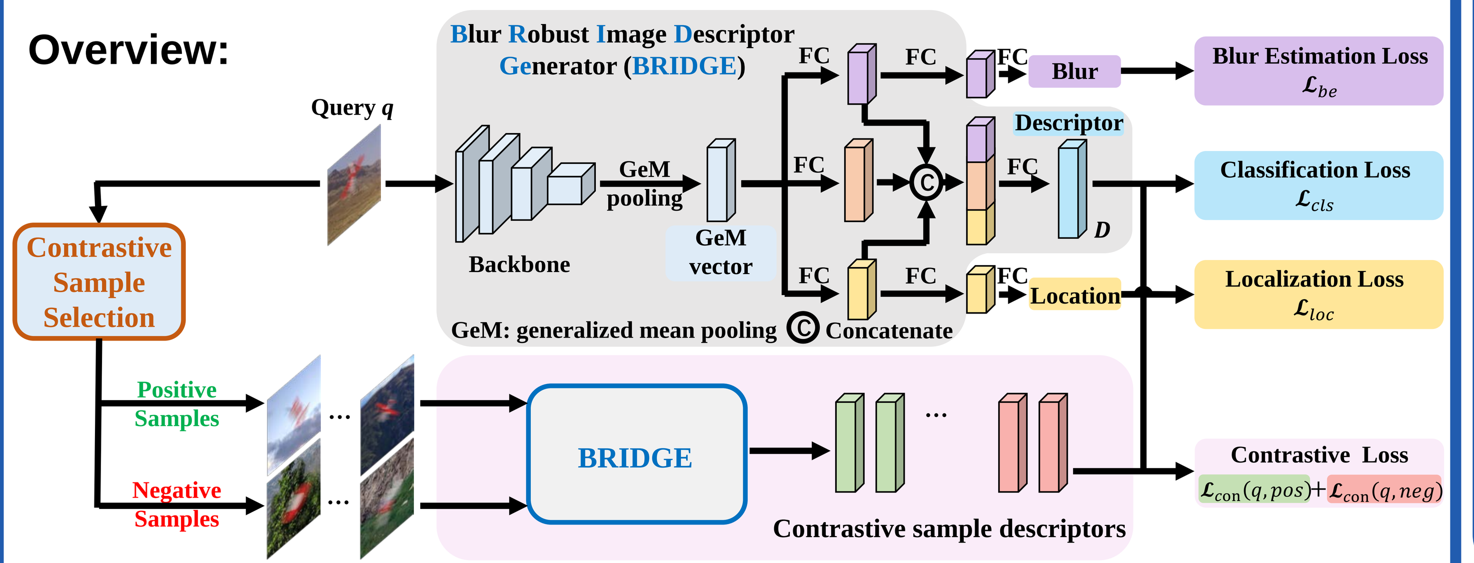
Introduction

Dynamic objects often appear blurred in images. Robust object retrieval in the presence of motion blur is an unstudied area and has practical significance in applications such as security surveillance & sports analysis.



Method

Overview:



Blur Estimation: $\mathcal{L}_{be} = |p - (1 - BS)|$, $BS = 1 - \frac{\sum_{i=0}^{H}\sum_{j=0}^{W}\boldsymbol{\beta}(i,j)\boldsymbol{\alpha}(i,j)}{\sum_{i=0}^{H}\sum_{j=0}^{W}\boldsymbol{\beta}(i,j)}$

Blur severity (*BS*) reflects a signal-to-noise ratio (SNR) as severer motion blur implies greater ratio of background (noise) to foreground (signal). $\mathcal{L}_{be}$ implicitly forces the model to identify blur and estimate SNR, helping it focus on and extract signal information in the blurred region while ignoring background noises.

Localization: $\mathcal{L}_{loc} = |x - \hat{x}| + |y - \hat{y}| + |w - \hat{w}| + |h - \hat{h}|$

$\mathcal{L}_{loc}$ requires the network to distinguish motion-blurred objects from backgrounds under various blur conditions, forcing it to recognize blur, identify objects at different blur levels, and further, focus more on the foreground objects.

Classification:

$$\mathcal{L}_{cls}(\widehat{D}, y) = -\log\left(\frac{\exp\left(\gamma \times AF\left(\widehat{w}_t^T\widehat{D}, 1\right)\right)}{\sum_n^N \exp\left(\gamma \times AF\left(\widehat{w}_n^T\widehat{D}, y_n\right)\right)}\right), AF(s,g) = \begin{cases}\cos(arccos(s)+m), & g = 1\\ s, & g = 0\end{cases}$$

$\mathcal{L}_{cls}$ requires the model to group sharp or blurred images of the same object together while separating different objects, which may appear similar, especially when motion-blurred, implicitly forcing the model to focus on blur-invariant features.

Contrastive: $\mathcal{L}_{con}(q,pos) = 0.5\left\|\widehat{D}_q - \widehat{D}_{pos}\right\|^2$

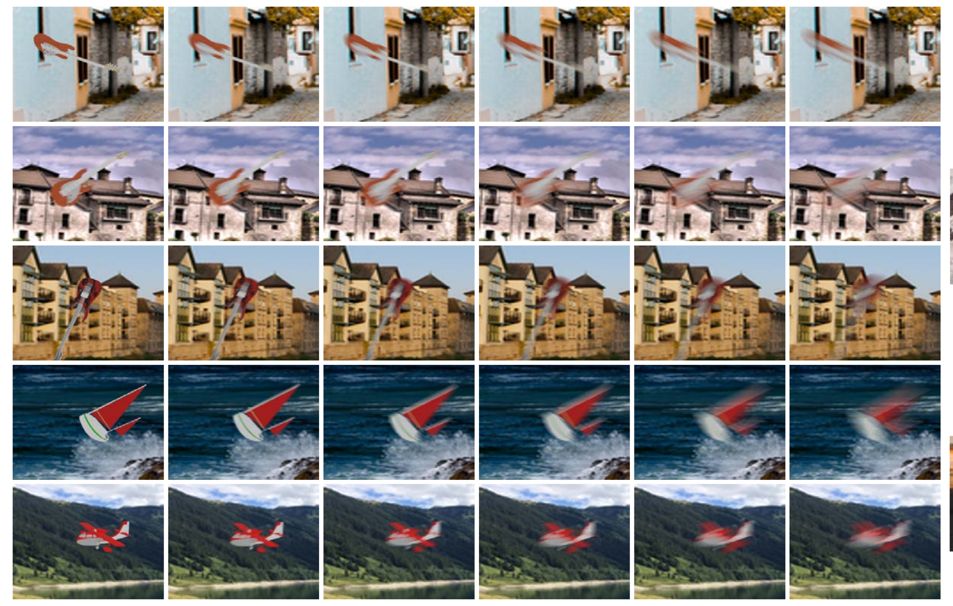
$\mathcal{L}_{con}(q,neg) = 0.5\left(max\{0, \tau - \left\|\widehat{D}_q - \widehat{D}_{neg}\right\|\}\right)^2$

$\mathcal{L}_{con}$ further enhances this focus by explicitly forcing descriptors of the same object, whether sharp or motion-blurred, to be close together, and those of different objects to be far apart.

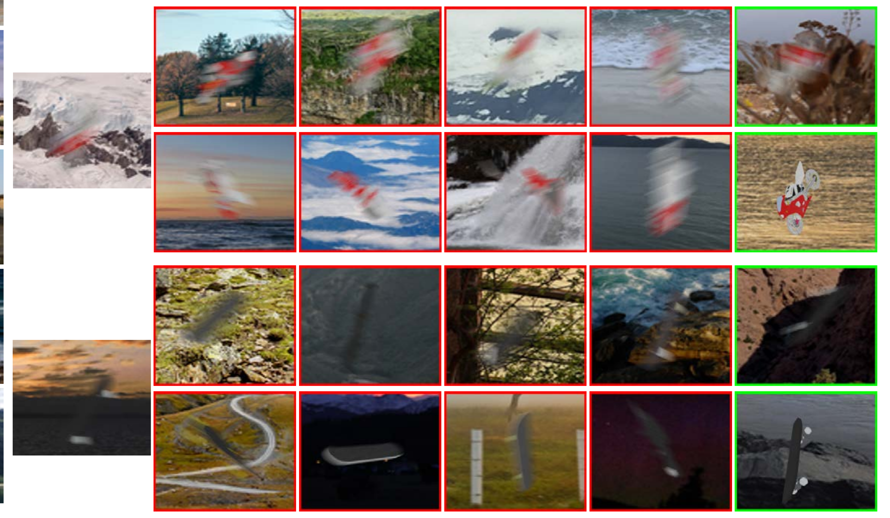
Synthetic Dataset

- Captured moving objects with different camera exposure times to obtain images with various amounts of motion blur
- 1,138 objects from 39 categories moving along random trajectories
- Each image is assigned a Blur Level (BL) according to its Blur Severity (BS): $BL = \lceil 10 \cdot BS \rceil$

BL 1 BL 2 BL 3 BL 4 BL 5 BL 6



- Distractors: 1,560 objects from the same categories to increase retrieval difficulty in terms of intra-class similarity

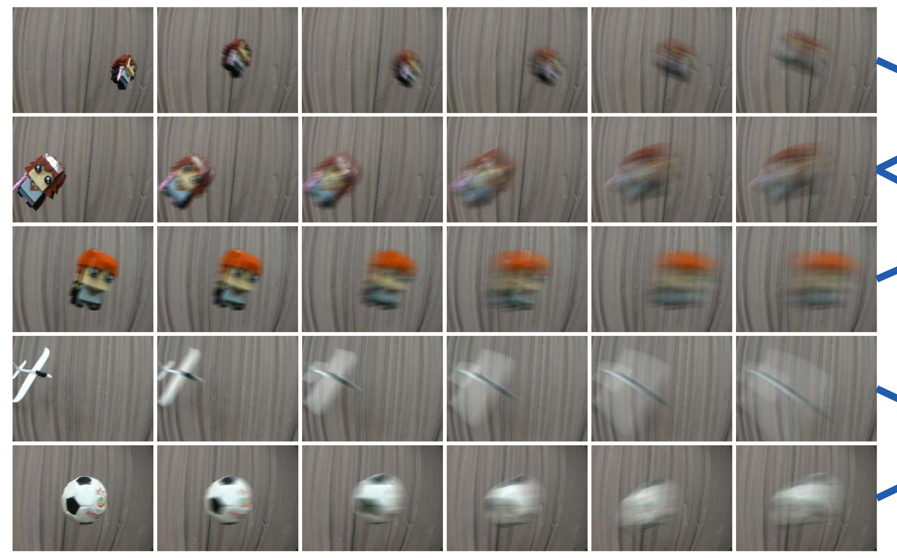


Syn. Data	# Total Images	# images each BL					
		1	2	3	4	5	6
Query	20,995	4,288	3,932	4,078	4,089	2,930	1,678
Database	91,621	18,871	17,508	17,888	18,029	12,546	6,779
Distractors	1,091,939	214,364	177,869	222,542	235,662	149,828	91,674

Real Dataset

- Recorded high-frame-rate videos (240fps), averaging different numbers of consecutive frames to obtain images with various amounts of motion blur
- 35 carefully selected objects, ensuring a balanced difficulty in terms of both intra- and inter-class similarity; None of them are in synthetic data
- 139 videos of objects moving along random trajectories

BL^r1 BL^r2 BL^r3 BL^r4 BL^r5 BL^r6



- Each real image is manually assigned a Blur Level (BL^r) based on the perceived blur

Different trajectories of the same object

Different objects from the same category (showing intra-class similarity)

Different categories of objects with similar textures (showing inter-class similarity)

Real Data	# Total Images	# images each BL^r					
		1	2	3	4	5	6
Query	2,753	612	620	561	396	315	249
Database	10,340	1,923	1,803	2,080	1,745	1,375	1,414

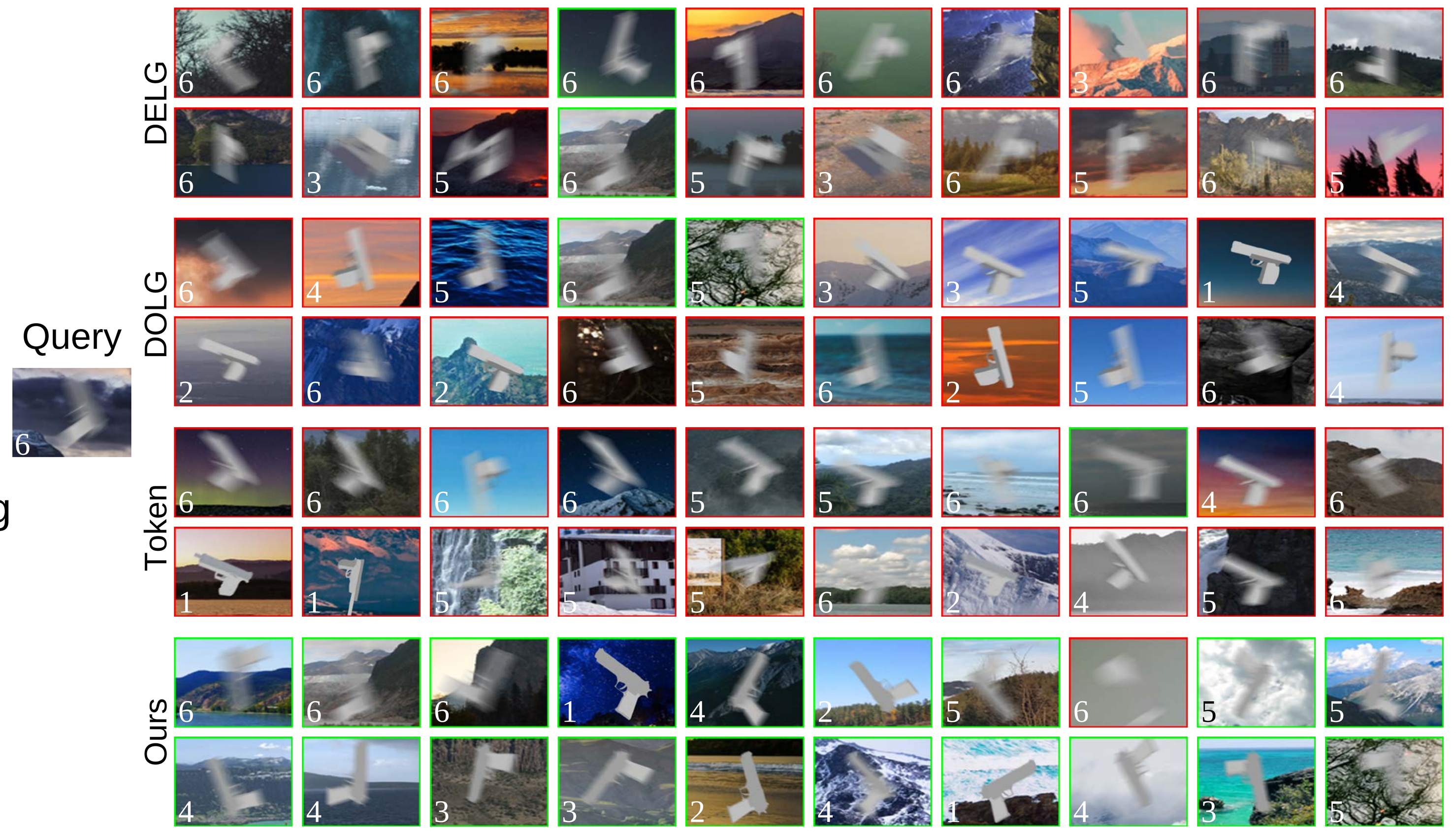
Experiments

Retrieval Results on Synthetic Data with 1M Distractors (mAP@100):

- All methods are retrained on the same synthetic data

Method	mAP (all queries)	mAP (subset of queries for each BL)					
		1	2	3	4	5	6
DELG	68.19	73.64	75.40	73.34	68.05	58.28	42.46
DOLG	69.97	75.75	77.47	75.01	70.10	60.01	42.49
Token	70.65	75.32	77.66	75.51	70.24	61.19	48.05
Ours-sharp	32.64	71.93	43.88	27.18	15.41	7.94	4.27
Ours	**84.09**	**88.74**	**89.56**	**87.68**	**84.41**	**76.89**	**62.42**

Top 20 retrieved images (red: negative, green: positive)



Retrieval Results on Real Data (mAP@all):

- All models are trained on synthetic and tested on real without finetuning

Method	mAP (all queries)	mAP (subset of queries for each BL^r)					
		1	2	3	4	5	6
DELG	54.82	49.13	63.43	57.25	55.01	53.77	42.92
DOLG	54.64	43.93	60.59	58.36	59.06	58.58	45.78
Token	43.33	38.71	47.08	50.79	46.44	42.71	24.43
Ours-sharp	40.24	49.55	45.02	41.33	33.23	29.40	27.91
Ours	**62.88**	**57.50**	**70.38**	**66.77**	**63.18**	**64.48**	**46.14**

Ablation Results on Synthetic Data (mAP@all):

- Results are grouped based on the loss applied directly to the descriptor

$\mathcal{L}_{con}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{be}$	$\mathcal{L}_{loc}$	mAP (all queries)	mAP (subset of queries for each BL)					
					1	2	3	4	5	6
√				78.13	80.51	81.70	81.16	79.20	73.93	61.24
√		√		81.66	83.49	85.01	84.43	82.69	77.64	67.08
√			√	85.94	87.54	88.25	87.83	86.52	83.08	75.58
√		√	√	**87.48**	**88.69**	**89.89**	**89.40**	**88.24**	**84.97**	**76.74**
	√			78.73	81.53	82.97	82.83	79.86	72.93	59.00
	√	√		83.67	85.19	87.56	87.40	85.10	79.22	65.93
	√		√	88.74	89.89	90.91	90.88	89.75	86.07	77.67
	√	√	√	**91.23**	**92.02**	**93.16**	**93.09**	**91.97**	**89.00**	**82.27**
√	√			85.06	87.42	88.29	87.66	85.85	81.20	69.96
√	√	√		87.17	89.03	90.03	89.55	88.07	83.91	73.48
√	√		√	90.39	91.85	92.45	92.14	91.20	88.20	79.36
√	√	√	√	**91.78**	**93.05**	**93.48**	**93.14**	**92.25**	**90.20**	**82.86**

Top 20 retrieved images (red: negative, green: positive)

Application to Real-world Video Data:

- Extracted 190 images of the same ball from a YouTube soccer video as query & database (https://www.youtube.com/watch?v=U8WCRz0Yh4Q)
- 4,600 hard distractors (4,431 sports ball images from MSCOCO, and 169 images of a different ball extracted from the same video)

Top 20 retrieved images (red: negative, green: positive)

Conclusion

- We introduce a novel and practical retrieval task involving object motion blur and propose the first method designed to create blur-robust image representations for bidirectional matching of motion-blurred objects and their deblurred counterparts.
- We present a new benchmark featuring synthetic and real-world data specifically constructed for this task, which is carefully processed and directly applicable for future research in blur-robust retrieval.
- Our method outperforms state-of-the-art standard retrieval methods and demonstrates superior robustness to motion blur.