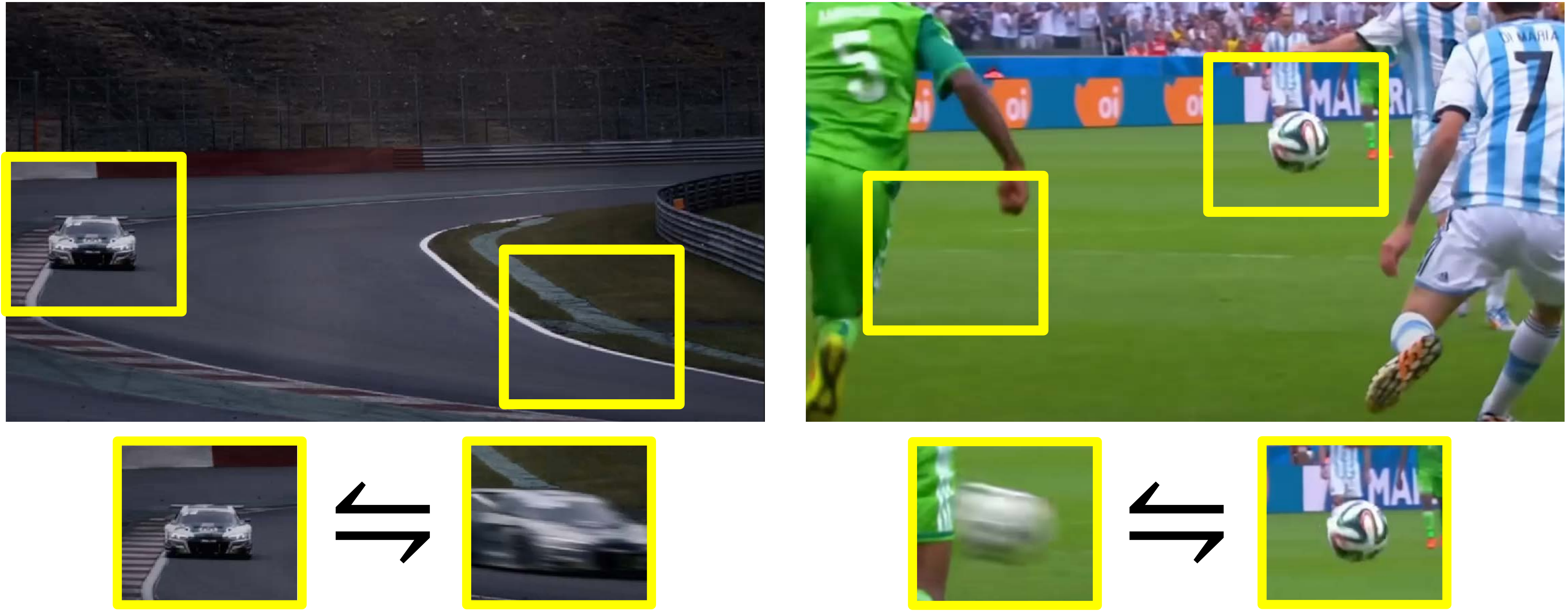
# Retrieval Robust to Object Motion Blur

European Conference on Computer Vision (ECCV) 2024

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

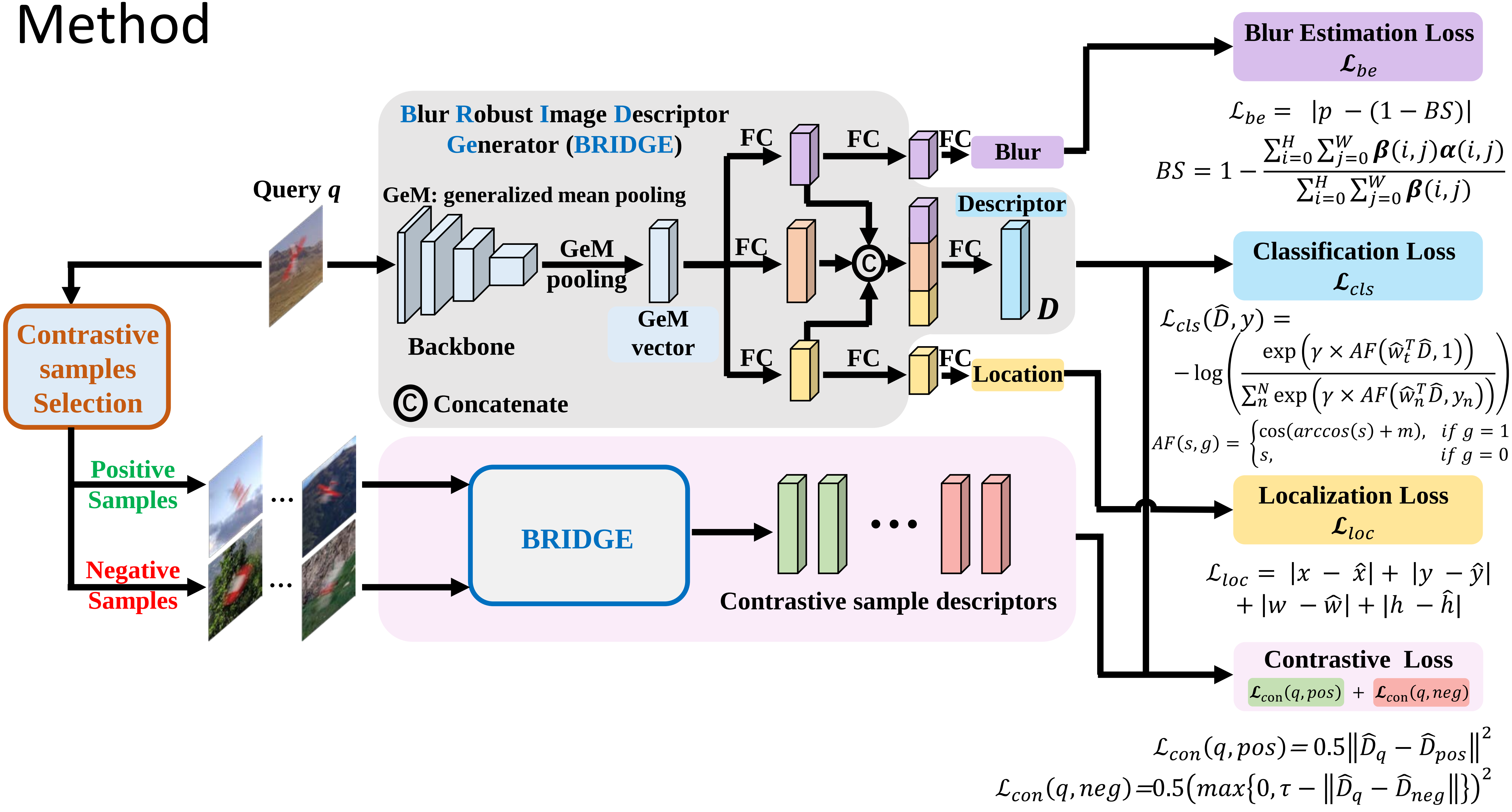
# Motivation

- Dynamic objects often appear blurred in images



- **Robust object retrieval in the presence of motion blur** has practical significance
- **Goal:** create **blur-robust** image representations for **bidirectional** matching of motion-blurred objects and their deblurred counterparts

# Method

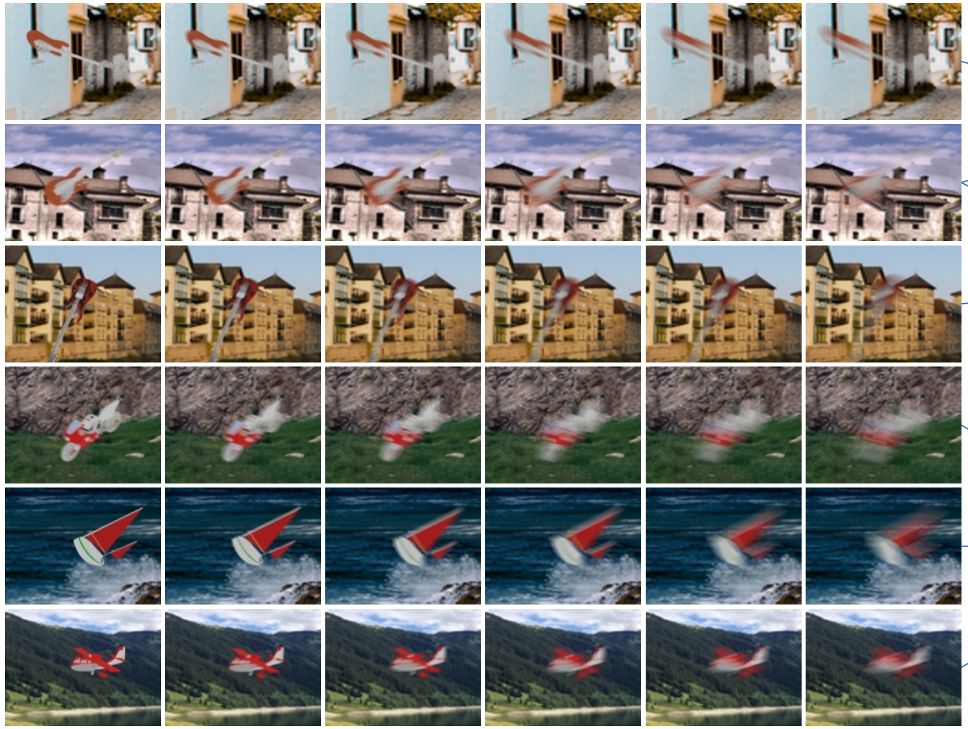


Blur Robust Image Descriptor Generator (BRIDGE)
GeM: generalized mean pooling
Query $q$
Backbone
GeM pooling
GeM vector
FC
FC
FC
FC
FC
FC
FC
FC
Blur
Descriptor
$D$
Location
Concatenate
Contrastive samples Selection
Positive Samples
Negative Samples
BRIDGE
Contrastive sample descriptors
Blur Estimation Loss $\mathcal{L}_{be}$
$\mathcal{L}_{be} = |p - (1 - BS)|$
$BS = 1 - \frac{\sum_{i=0}^{H}\sum_{j=0}^{W}\boldsymbol{\beta}(i,j)\boldsymbol{\alpha}(i,j)}{\sum_{i=0}^{H}\sum_{j=0}^{W}\boldsymbol{\beta}(i,j)}$
Classification Loss $\mathcal{L}_{cls}$
$\mathcal{L}_{cls}(\widehat{D}, y) = -\log\left(\frac{\exp\left(\gamma \times AF\left(\widehat{w}_t^T\widehat{D}, 1\right)\right)}{\sum_n^N \exp\left(\gamma \times AF\left(\widehat{w}_n^T\widehat{D}, y_n\right)\right)}\right)$
$AF(s,g) = \begin{cases} \cos(\arccos(s) + m), & if\ g = 1 \\ s, & if\ g = 0 \end{cases}$
Localization Loss $\mathcal{L}_{loc}$
$\mathcal{L}_{loc} = |x - \hat{x}| + |y - \hat{y}| + |w - \widehat{w}| + |h - \hat{h}|$
Contrastive Loss
$\mathcal{L}_{con}(q,pos) + \mathcal{L}_{con}(q,neg)$
$\mathcal{L}_{con}(q,pos) = 0.5\left\|\widehat{D}_q - \widehat{D}_{pos}\right\|^2$
$\mathcal{L}_{con}(q,neg) = 0.5\left(max\{0, \tau - \left\|\widehat{D}_q - \widehat{D}_{neg}\right\|\}\right)^2$

# Datasets – Synthetic

- No existing dataset for this novel retrieval task
- We developed a simulator to generate motion-blurred data under controlled conditions
- Simulating 1,138 objects from 39 categories moving along random trajectories
- Capturing images with different camera exposure time in the simulator
- Each image is assigned a Blur Level ($BL$) according to its Blur Severity ($BS$): $BL = \lceil 10 \cdot BS \rceil$
- Examples:

$BL\ 1$ $BL\ 2$ $BL\ 3$ $BL\ 4$ $BL\ 5$ $BL\ 6$



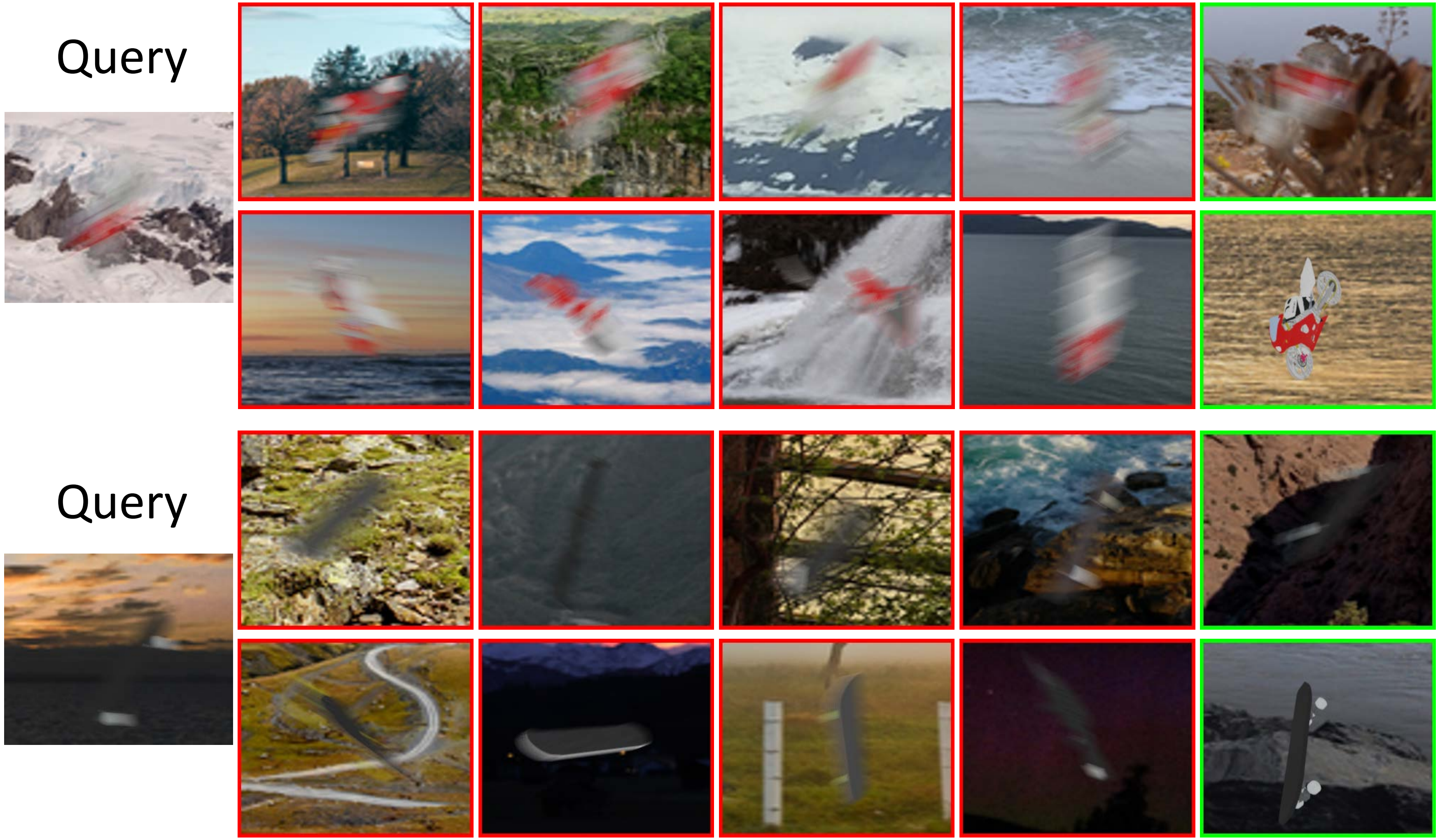
Same object, different trajectories

Same category, different objects (intra-class similarity)

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

# Datasets – Synthetic

- Distractors: 1,560 objects from the same categories to increase retrieval difficulty



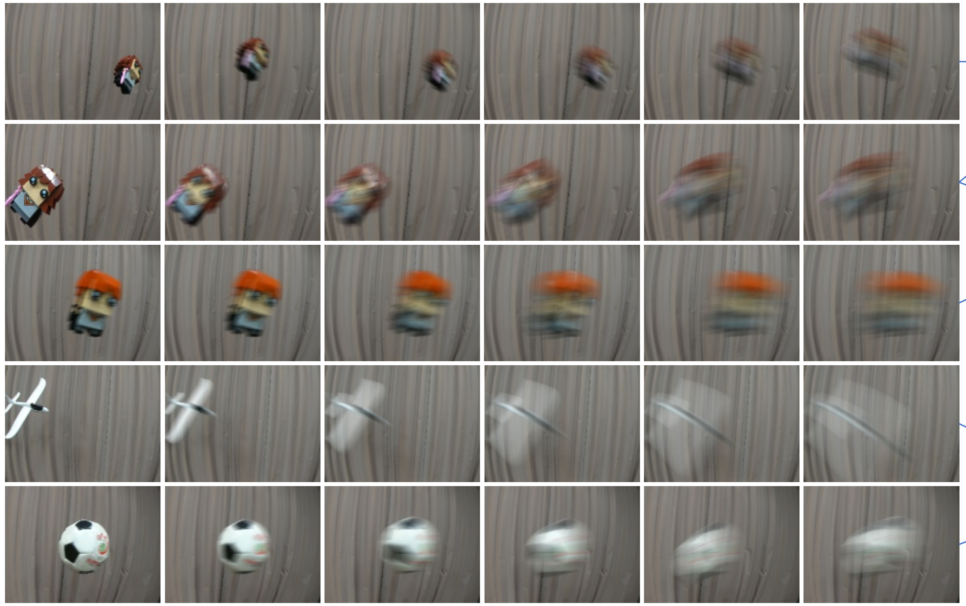

Red
difficult distractors

Green
positives in database
(top: motion-blurred
bottom: sharp)

# Datasets – Real

- We recorded high-frame-rate (240fps) videos of objects moving along random trajectories
- 35 carefully selected objects, ensuring a balanced difficulty in terms of both intra- and inter-class similarity; None of them are in synthetic data
- Averaging different numbers of consecutive frames to obtain images with various amounts of motion blur
- Each real image is manually assigned a Blur Level based on the perceived blur ($BL^r$, r denotes real data)
- Examples:

$BL^r$ 1 $BL^r$ 2 $BL^r$ 3 $BL^r$ 4 $BL^r$ 5 $BL^r$ 6



Same object, different trajectories

Same category, different objects (intra-class similarity)

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

# Datasets – Statistics

- Statistics of synthetic evaluation data for different $BLs$

| Dataset | # 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 |
| 1M Distractors | 1,091,939 | 214,364 | 177,869 | 222,542 | 235,662 | 149,828 | 91,674 |

- Statistics of real evaluation data for different $BLs$:

| Dataset | # 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 |

# Results — Quantitative on Synthetic (+1M distractors)

- All methods are retrained on the same synthetic data
- Metric: mean average precision (mAP) of top 100

| Method | mAP (all queries) | mAP (subset of queries for each $BL$) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| DELG [Cao, ECCV 2020] | 68.19 | 73.64 | 75.40 | 73.34 | 68.05 | 58.28 | 42.46 |
| DOLG [Yang, ICCV 2021] | 69.97 | 75.75 | 77.47 | 75.01 | 70.10 | 60.01 | 42.49 |
| Token [Wu, AAAI 2022] | 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** |

- The database contains images of all blur levels ($BL$ 1 to 6)

# Results — Qualitative on Synthetic (+1M)

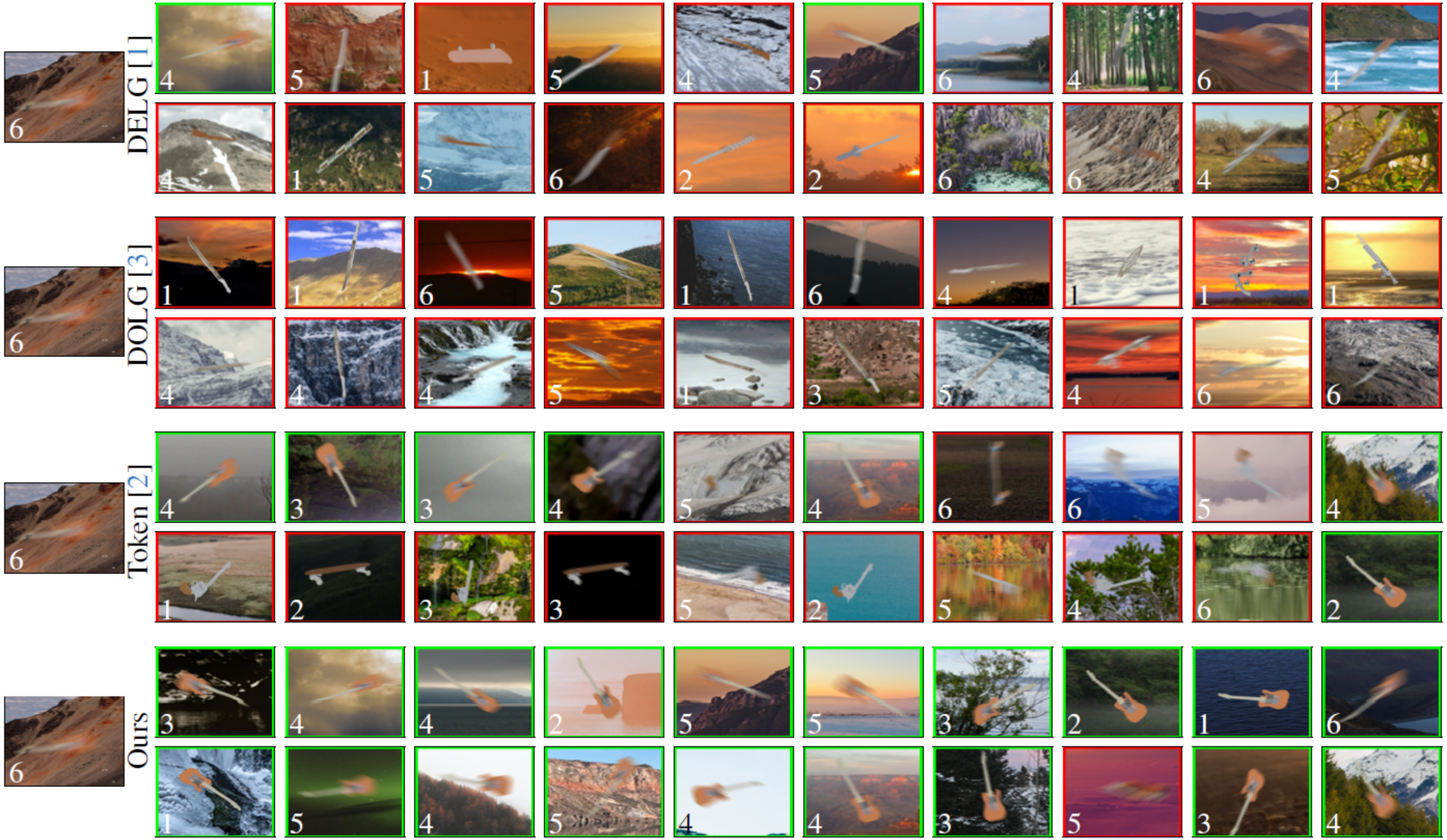
- Illustration of retrieval difficulty in terms of intra-class similarity

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

# Results — Qualitative on Synthetic (+1M)

- Illustration of retrieval difficulty in terms of inter-class similarity

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



# Results — Quantitative on Real

- Metric: mean average precision (mAP) of all retrieved images

| Method | mAP (all queries) | mAP (subset of queries for each $BL^r$ ) | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| DELG [Cao, ECCV 2020] | 54.82 | 49.13 | 63.43 | 57.25 | 55.01 | 53.77 | 42.92 |
| DOLG [Yang, ICCV 2021] | 54.64 | 43.93 | 60.59 | 58.36 | 59.06 | 58.58 | 45.78 |
| Token [Wu, AAAI 2022] | 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** |

- The database contains images of all blur levels ($BL^r$ 1 to 6)
- All methods are trained on synthetic data and evaluated on real data without finetuning

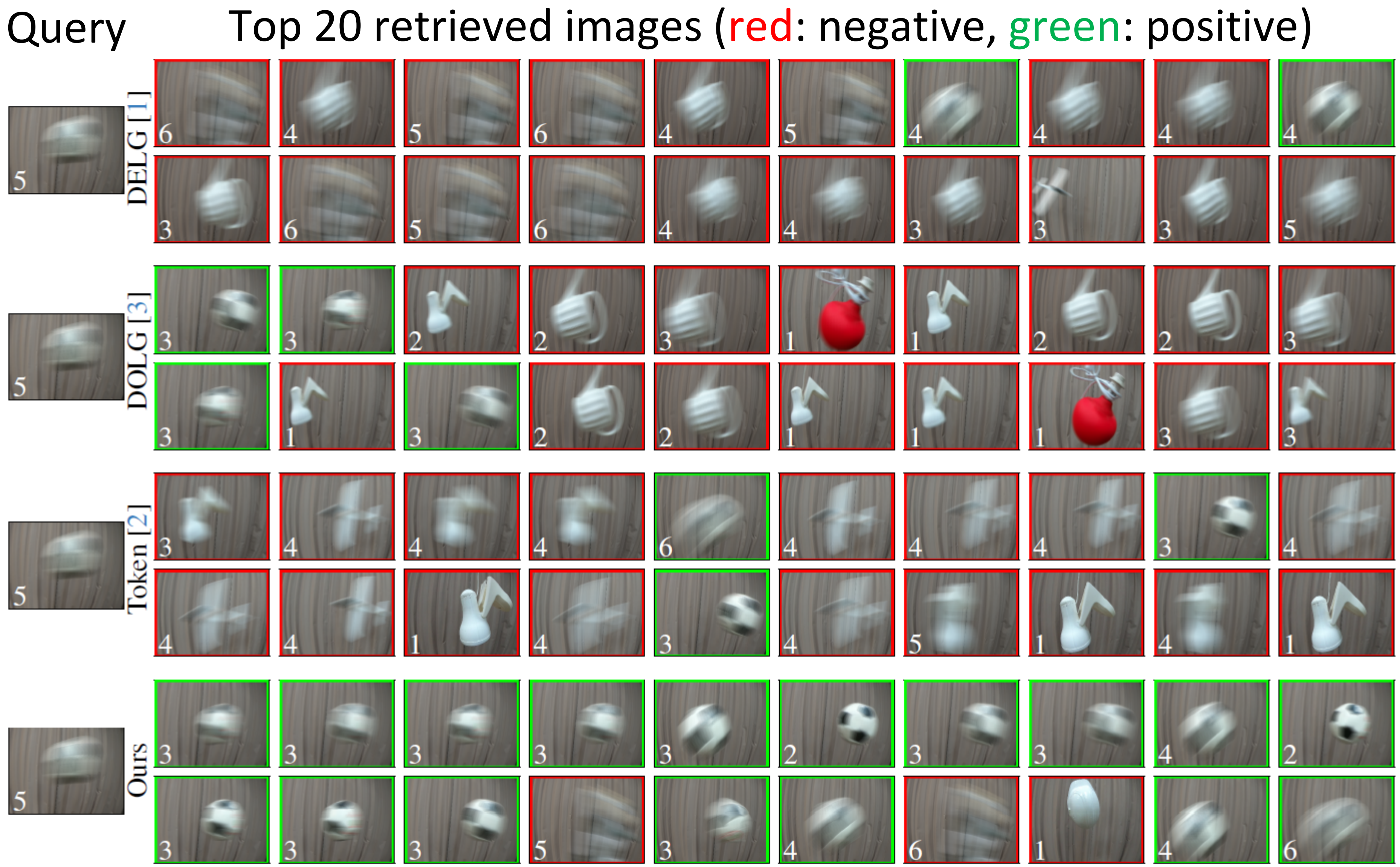
# Results — Qualitative on Real

- Illustration of retrieval difficulty in terms of intra-class similarity

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

# Results — Qualitative on Real

- Illustration of retrieval difficulty in terms of inter-class similarity



# Results — Ablation Study on Synthetic

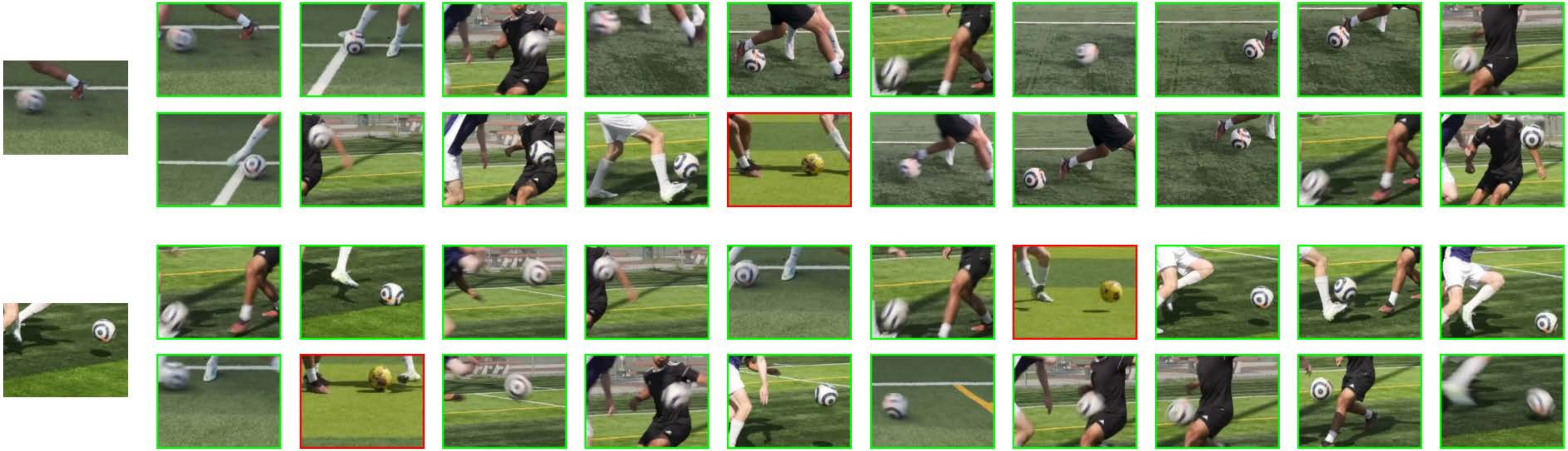
- Ablation study on loss components

| $\mathcal{L}_{con}$ | $\mathcal{L}_{cls}$ | $\mathcal{L}_{be}$ | $\mathcal{L}_{loc}$ | 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** |

# Application to real-world video data

- We extracted 190 images of the same ball from a YouTube soccer video as query & database
- Adding 4,600 hard distractors: 4,431 sports ball images from MSCOCO [Lin, ECCV 2014];
  169 images of a different ball extracted from the same video

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



- Illustration of our method's effectiveness in handling various blur conditions and complex and diverse backgrounds in the real world

https://www.youtube.com/watch?v=U8WCRz0Yh4Q

# Conclusion

- We introduce **a novel retrieval task** involving motion blur; this task holds practical significance with applications in real-world dynamic scenarios.
- We present **the first method** specifically designed to tackle this task, which is trained with specialized loss functions tailored to improve model's understanding of motion blur.
- We introduce **a new benchmark** featuring synthetic and real-world datasets specifically constructed for this task. The datasets are large-scale, meticulously processed, and directly applicable for future research in blur retrieval.
- We conducted extensive experiments, showing that our method achieves **higher mAP** and exhibits **superior robustness to motion blur** compared to SOTA standard retrieval methods.

Thank you!