NEURAL INFORMATION
PROCESSING SYSTEMS

1st and 3nd Solutions to FaceBook AI Image Similarity Challenge

Speaker: Wenhao Wang

VisionForce (Wenhao Wang, Yifan Sun, Weipu Zhang and Yi Yang)

Baidu Research

NEURAL INFORMATION
PROCESSING SYSTEMS

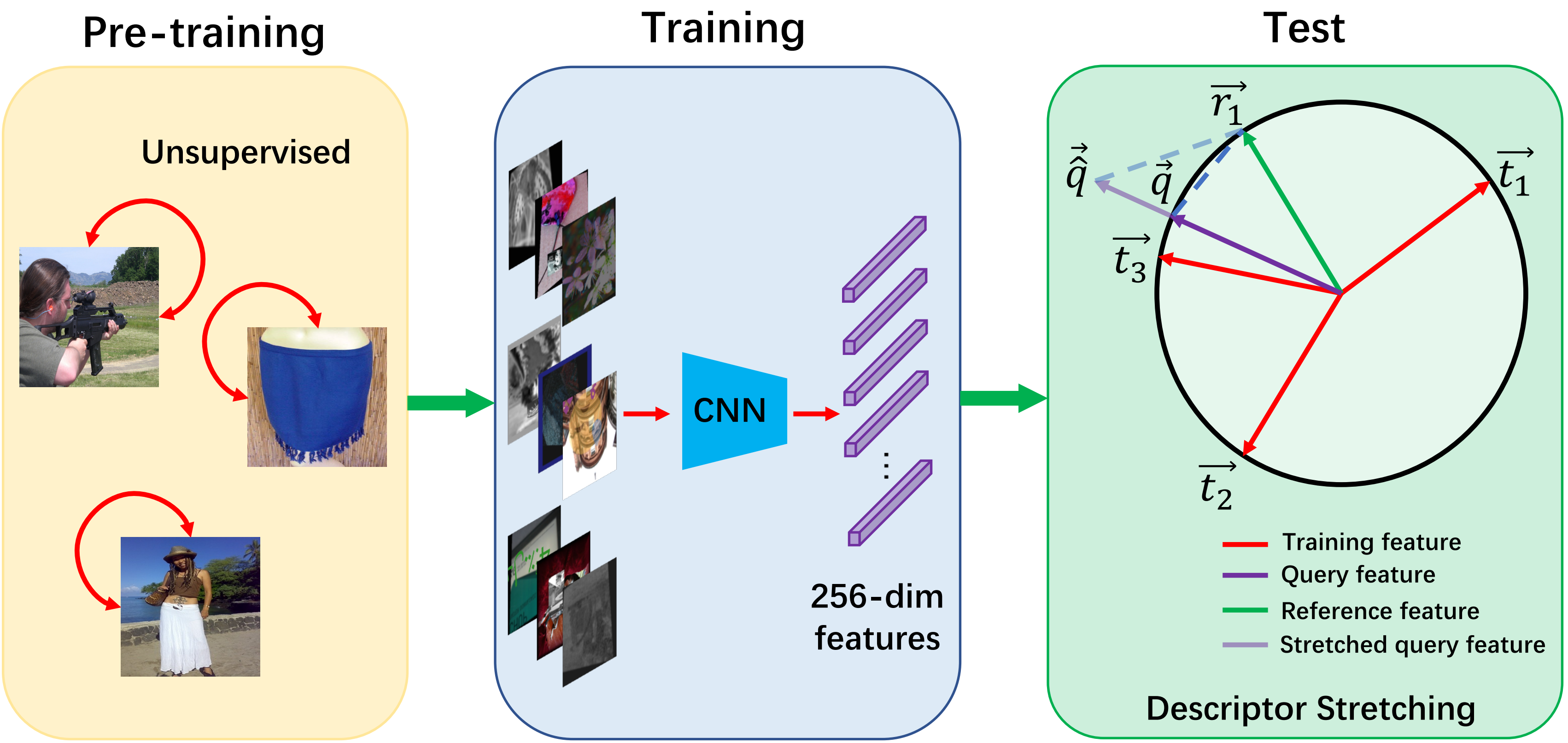
Bag of Tricks and A Strong Baseline For Image Copy Detection

3nd Solution to Descriptor Track

Authors: Wenhao Wang, Weipu Zhang, Yifan Sun, Yi Yang

Baidu Research

Pipeline



Pre-training

Unsupervised

Training

CNN

256-dim features

Test

🟥 Training feature

🟪 Query feature

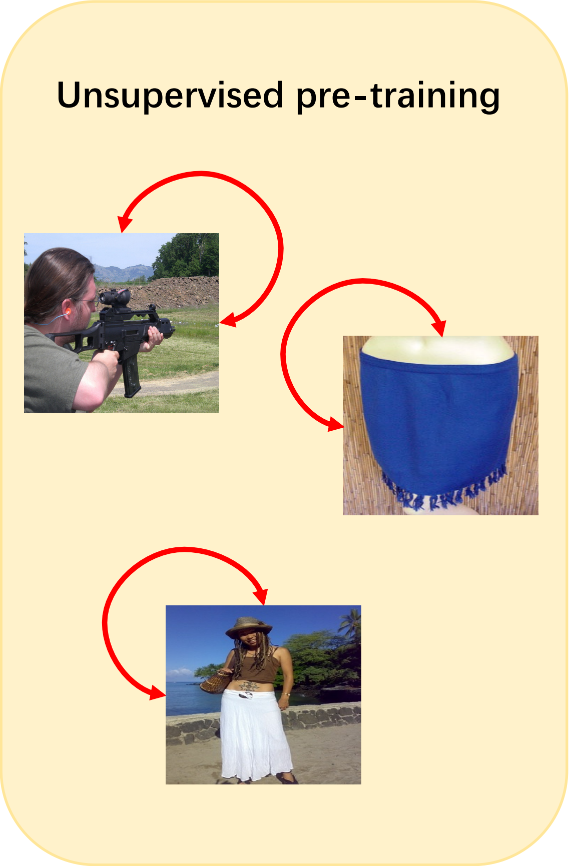
🟩 Reference feature

🟪 Stretched query feature

Descriptor Stretching

Pre-training

Unsupervised pre-training on ImageNet using Barlow Twins [1].



Unsupervised pre-training

Why?

The granularity of a category is the same in ISC2021 and self-supervised learning.

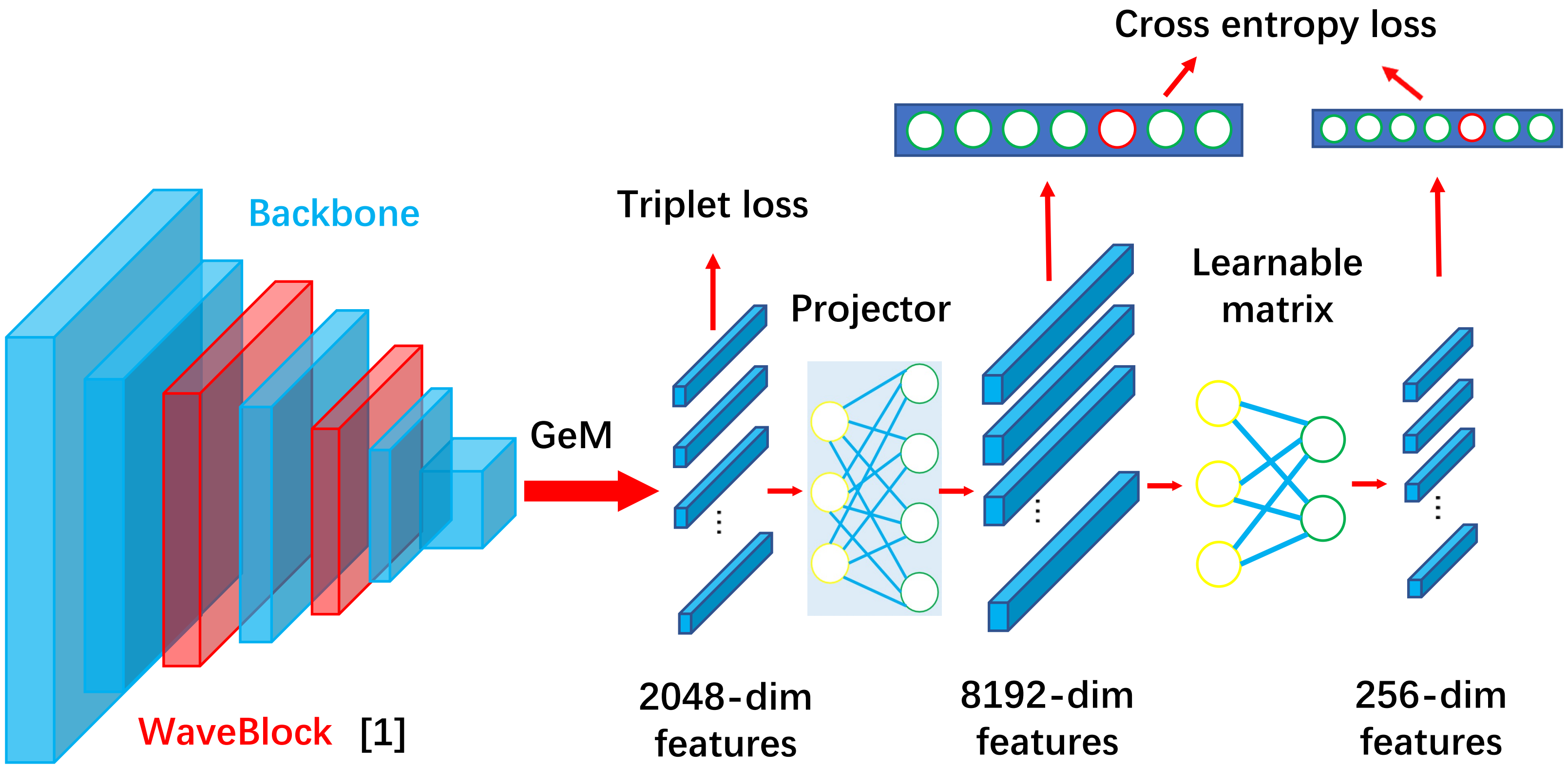
Choice?

Moco, BYOL, SwAV, ***Barlow Twins***, SimSiam, …

[1] Jure Zbontar, et al. Barlow twins: Self-supervised learning via redundancy reduction. In ICML, 2021.

Training

Training methods

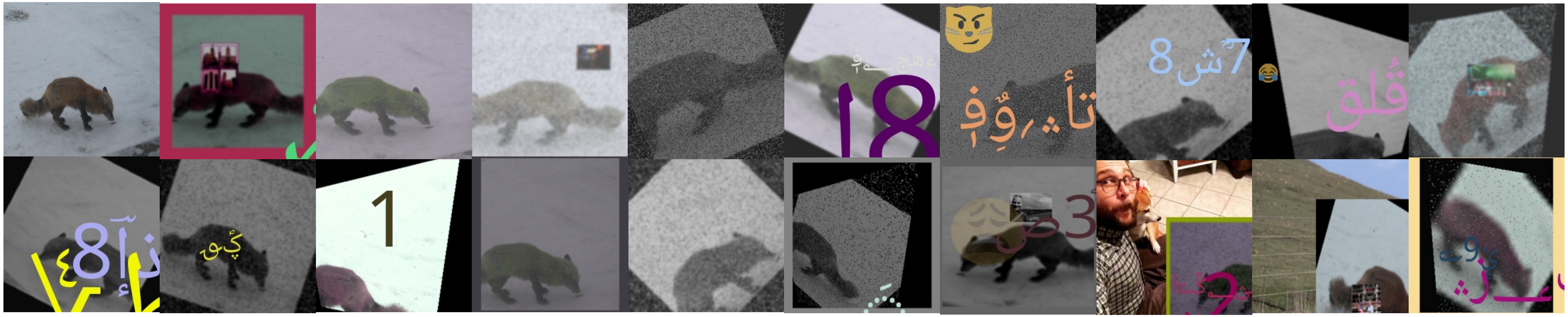


[1] Wenhao Wang, et al. Attentive WaveBlock: Complementarity-enhanced Mutual Networks for Unsupervised Domain Adaptation in Person Re-identification and Beyond. In Preprint, 2020.

Training

One set of designed augmentations

Basic augmentation



Test

Descriptor Stretching *VS* Score Normalization

Descriptor Stretching

1. Purpose: To make the similarity values comparable across different queries;
2. Subject: ***Features***.

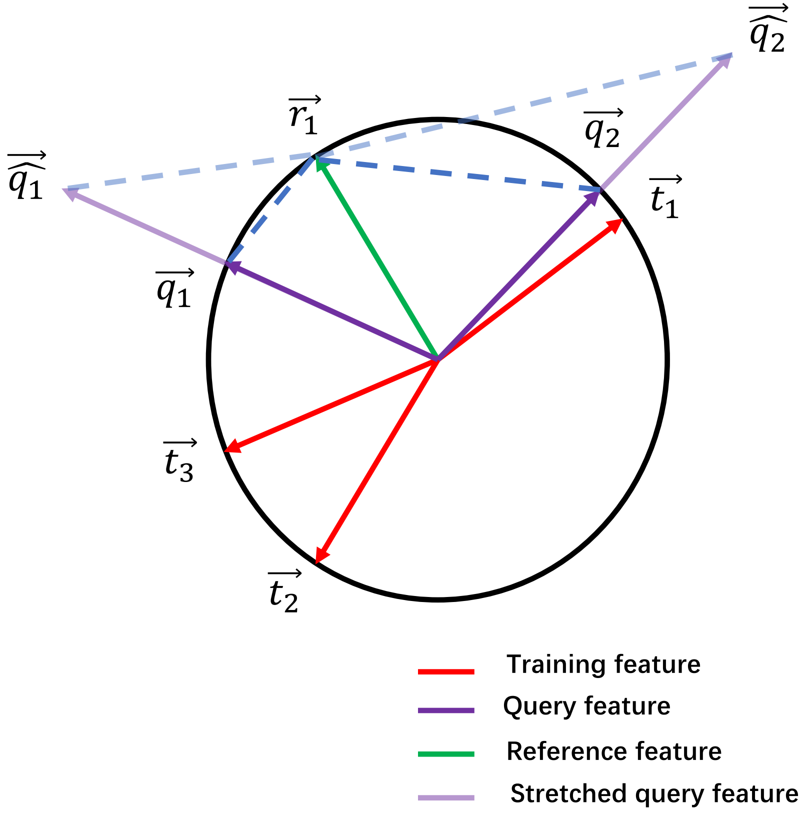
Score Normalization

1. Purpose: To make the similarity values comparable across different queries;
2. Subject: ***Scores***.

Therefore, in this track, we use ***Descriptor Stretching*** to replace Score Normalization.

Test

Descriptor Stretching



Given the feature of a query image $\overrightarrow{q_1}$, and a reference image $\overrightarrow{r_1}$, the original score $s_1$is defined as

$$s_1 = |\overrightarrow{q_1} - \overrightarrow{r_1}|.$$

Similarly, we have:

$$s_2 = |\overrightarrow{q_2} - \overrightarrow{r_1}|.$$

If $s_1 > s_2$, $\overrightarrow{q_2}$ is more similar to $\overrightarrow{r_1}$ than $\overrightarrow{q_1}$, and vice versa.

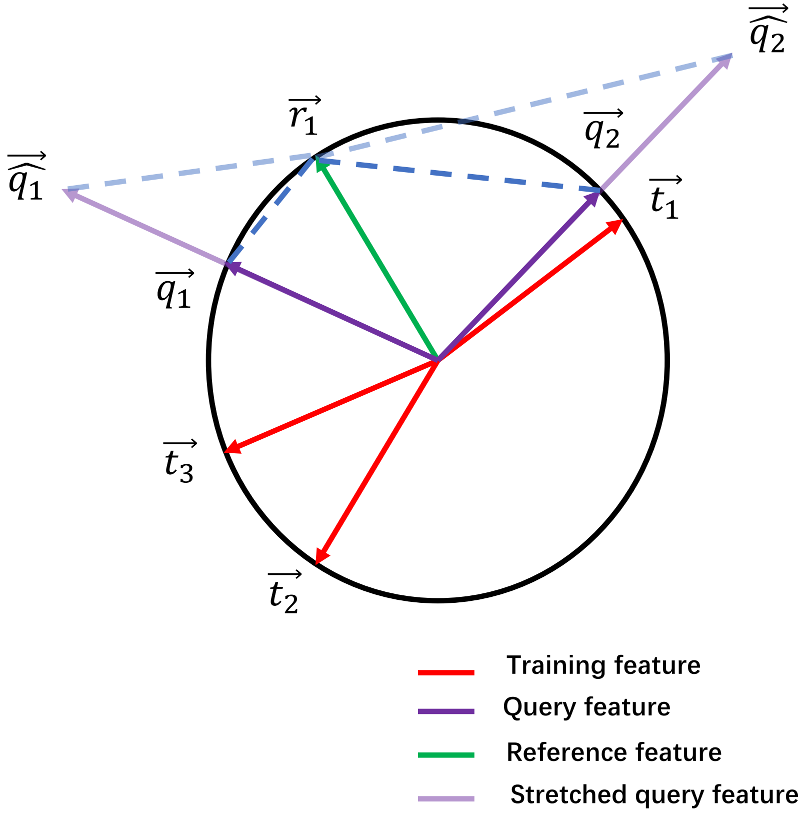
The definition of descriptor stretching is

$$\overrightarrow{\widehat{q_1}} = \alpha \cdot s_{n_1} \cdot \overrightarrow{q_1},$$

Test

$$\overrightarrow{\widehat{q_1}} = \alpha \cdot s_{n_1} \cdot \overrightarrow{q_1},$$

Descriptor Stretching



where: α is a hyper-parameter, and $s_{n_1}$is the mean of top n inner product scores between $\overrightarrow{q_1}$ and the features of images from the training set. Then the stretched score $\widehat{s_1}$ is defined as:

$$\widehat{s_1}=\left|\overrightarrow{\widehat{q_1}} - \overrightarrow{r_1}\right|.$$

Similarly, we have:

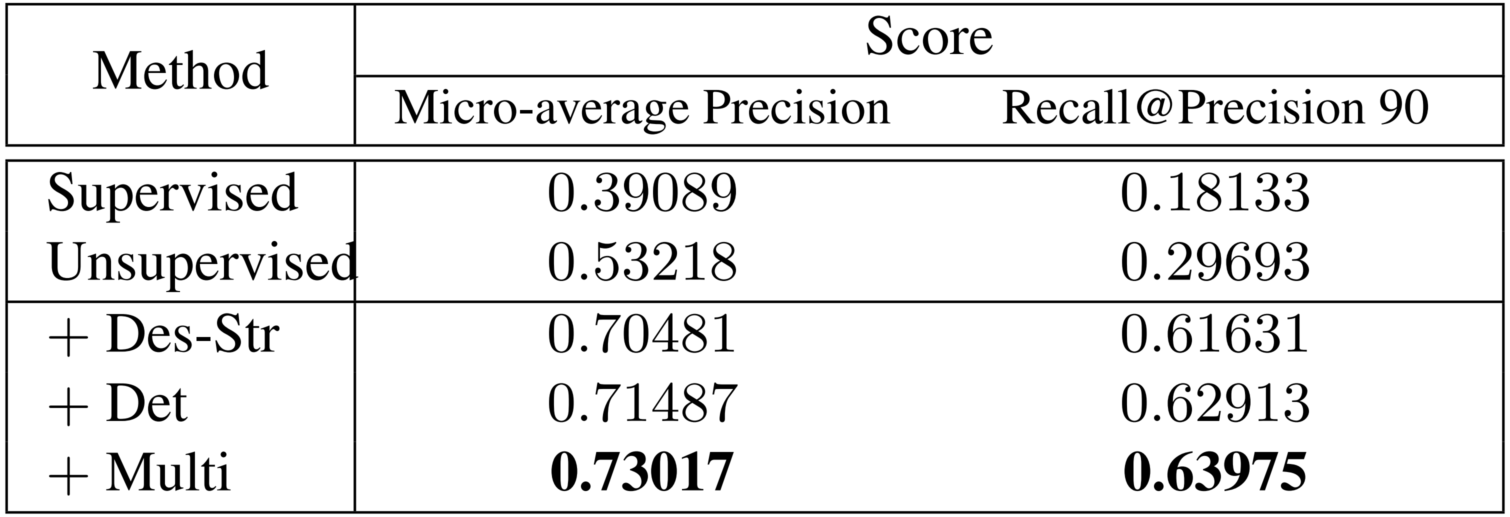
$$\overrightarrow{\widehat{q_2}} = \alpha \cdot s_{n_2} \cdot \overrightarrow{q_2},$$

$$\widehat{s_2}=\left|\overrightarrow{\widehat{q_2}} - \overrightarrow{r_1}\right|.$$

After stretching, we use the stretched feature of a query image as its final descriptor.

Experiments

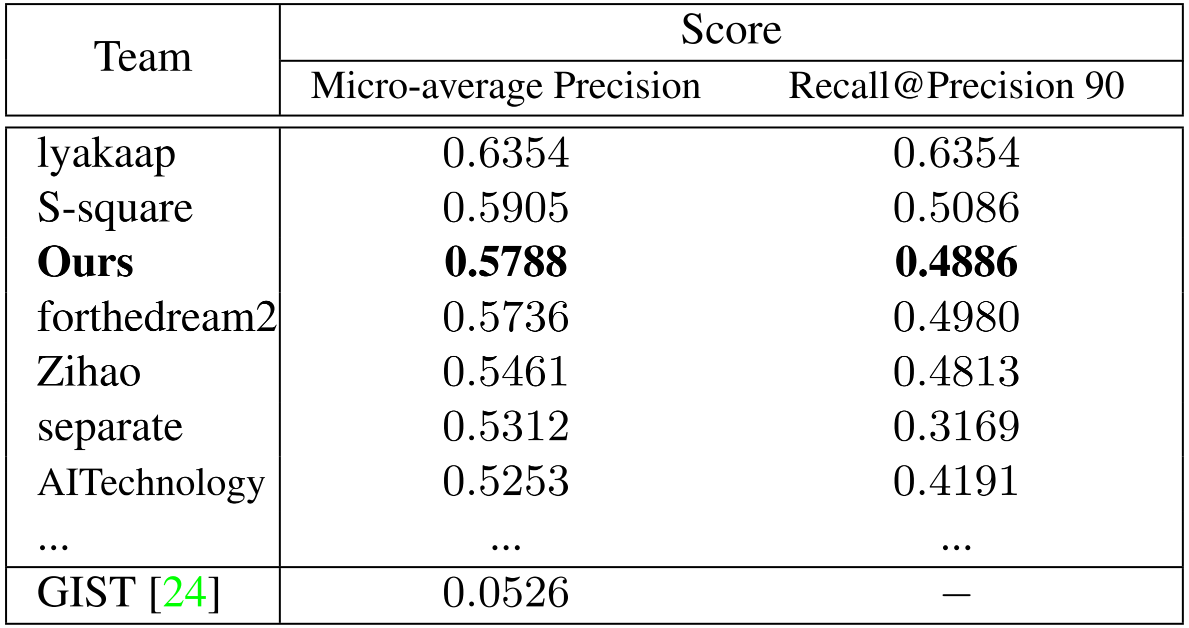
Ablation Studies



Method	Score	
	Micro-average Precision	Recall@Precision 90
Supervised	0.39089	0.18133
Unsupervised	0.53218	0.29693
+ Des-Str	0.70481	0.61631
+ Det	0.71487	0.62913
+ Multi	**0.73017**	**0.63975**

Experiments

Comparison with State-of-the-Arts



Team	Score	
	Micro-average Precision	Recall@Precision 90
lyakaap	0.6354	0.6354
S-square	0.5905	0.5086
Ours	**0.5788**	**0.4886**
forthedream2	0.5736	0.4980
Zihao	0.5461	0.4813
separate	0.5312	0.3169
AITechnology	0.5253	0.4191
...	...	...
GIST [24]	0.0526	–

NEURAL INFORMATION PROCESSING SYSTEMS

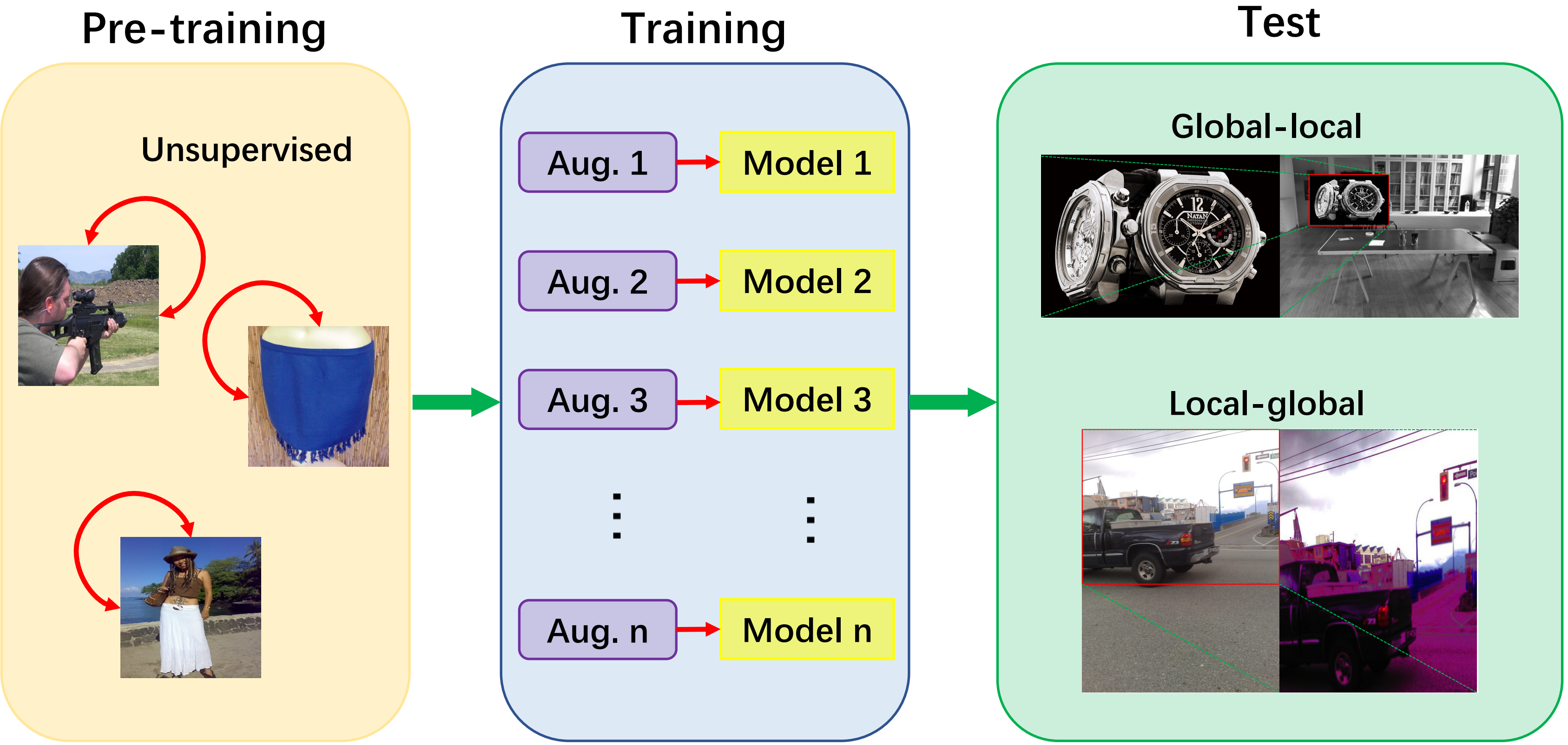
D^{2}LV: A Data-Driven and Local-Verification Approach for Image Copy Detection

1st Solution to Matching Track

Authors: Wenhao Wang, Yifan Sun, Weipu Zhang, Yi Yang

Baidu Research

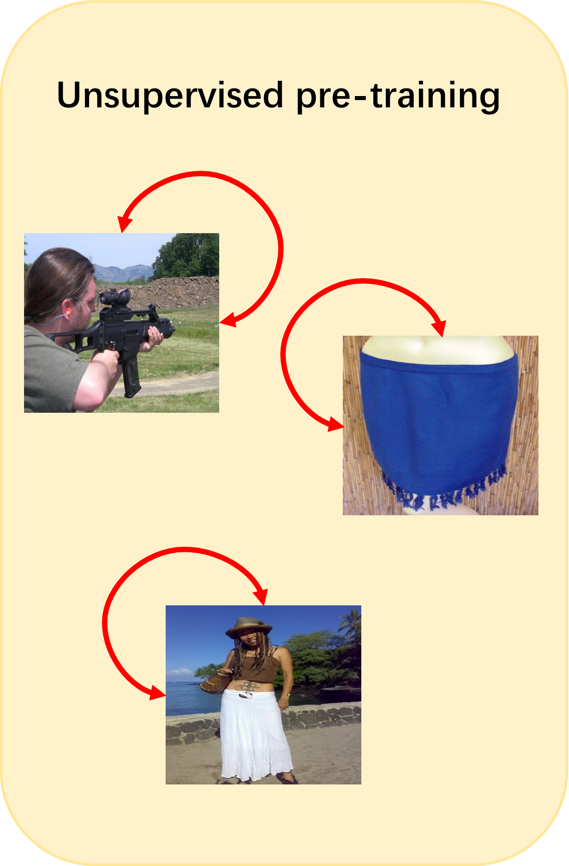
Pipeline



Pre-training
Unsupervised
Training
Aug. 1
Model 1
Aug. 2
Model 2
Aug. 3
Model 3
⋮
⋮
Aug. n
Model n
Test
Global-local
Local-global

Pre-training

Unsupervised pre-training on ImageNet using BYOL [1] and Barlow Twins [2].



Unsupervised pre-training

Why?

The granularity of a category is the same in ISC2021 and self-supervised learning.

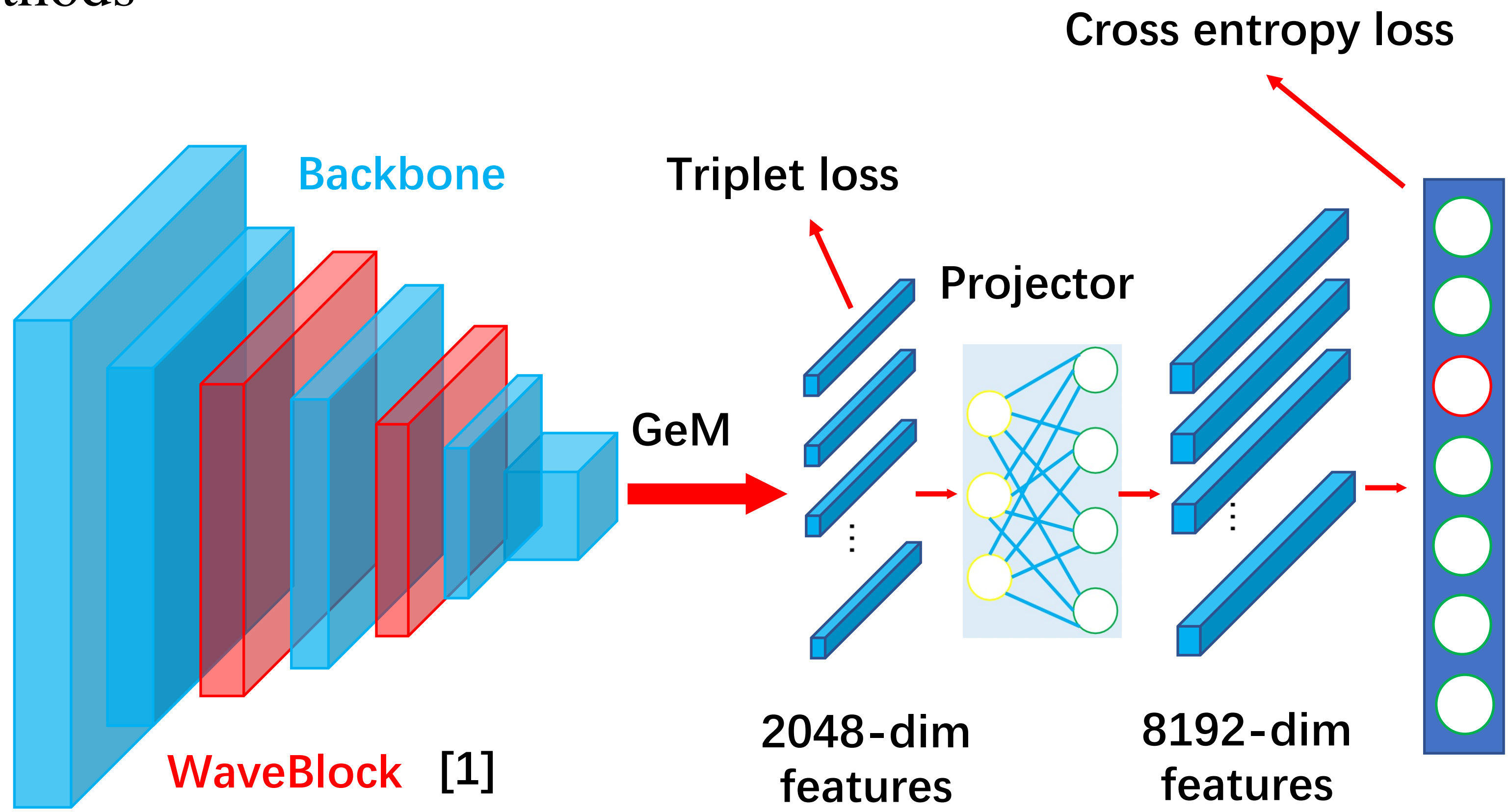
Choice?

Moco, ***BYOL***, SwAV, ***Barlow Twins***, SimSiam, …

[1] Grill Jean-Bastien, et al. Bootstrap your own latent: a new approach to self-supervised learning. NIPS 2020,
[2] Jure Zbontar, et al. Barlow twins: Self-supervised learning via redundancy reduction. In ICML, 2021.

Training

Training methods



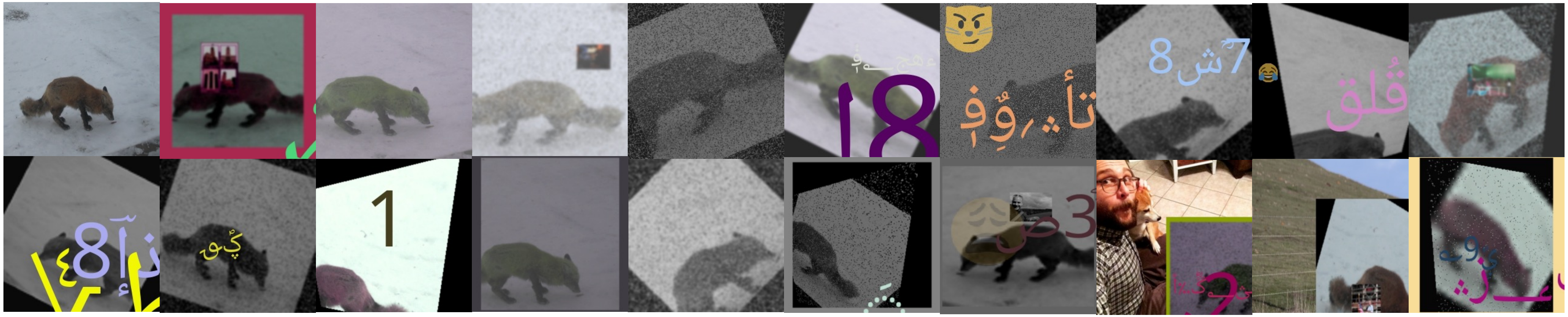
[1] Wenhao Wang, et al. Attentive WaveBlock: Complementarity-enhanced Mutual Networks for Unsupervised Domain Adaptation in Person Re-identification and Beyond. In Preprint, 2020.

Training

11 sets of designed augmentations generate 11 datasets:

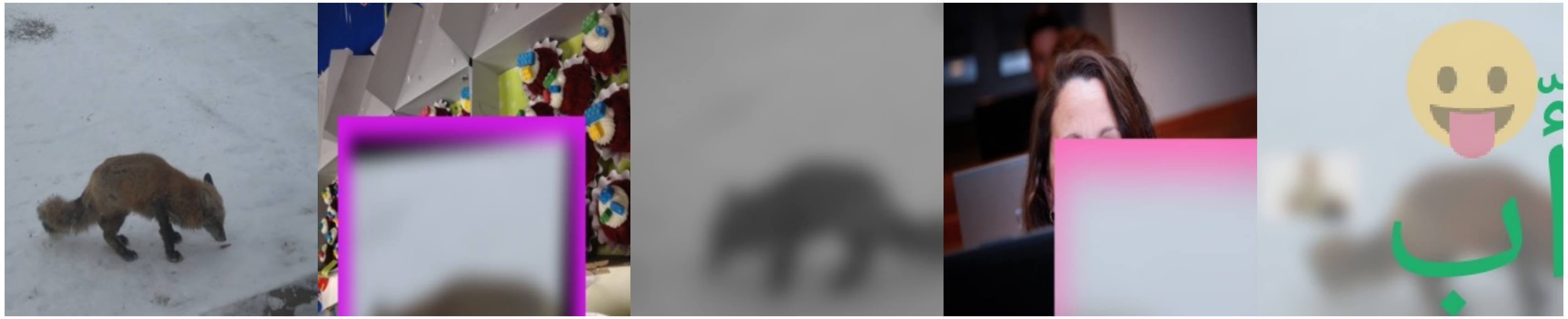
Training on each dataset *separately*.

1. Basic augmentation

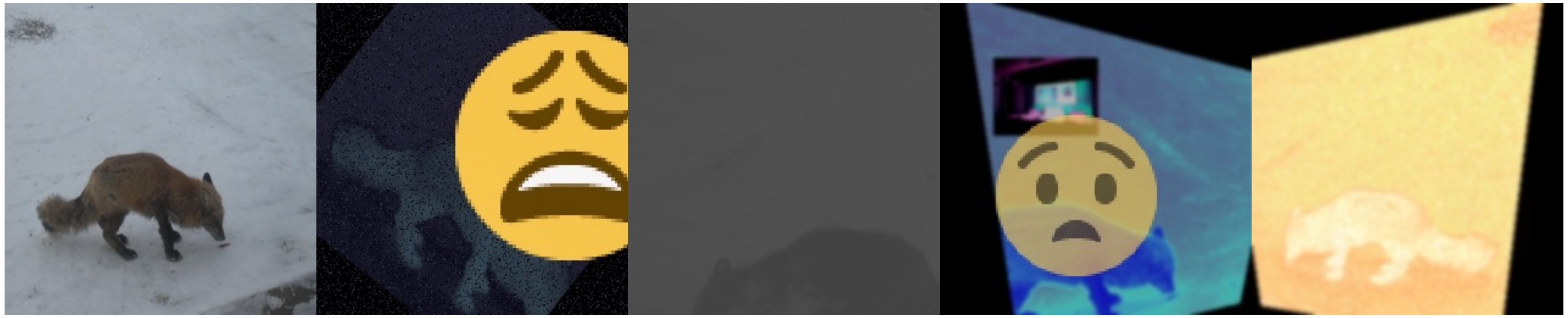


Training

2. Basic + Super-blur augmentation

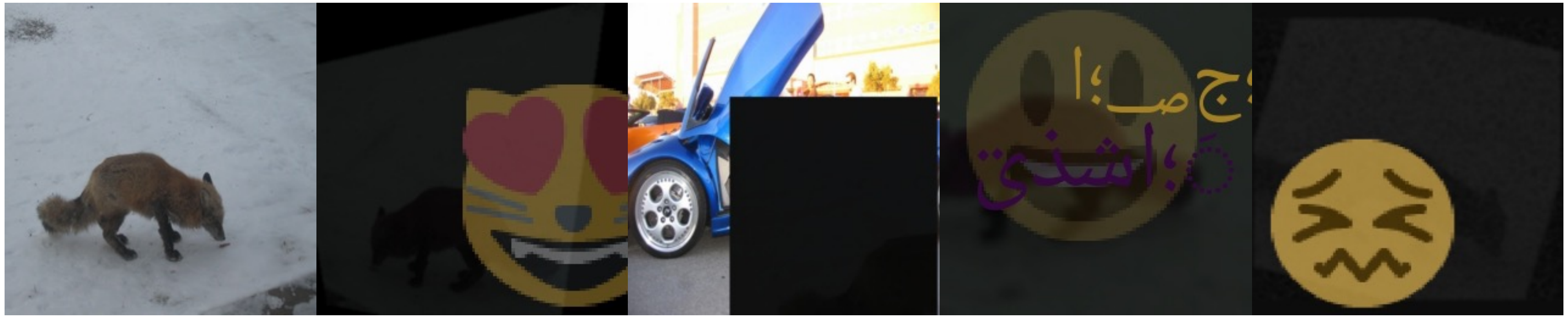


3. Basic + Super-color augmentation

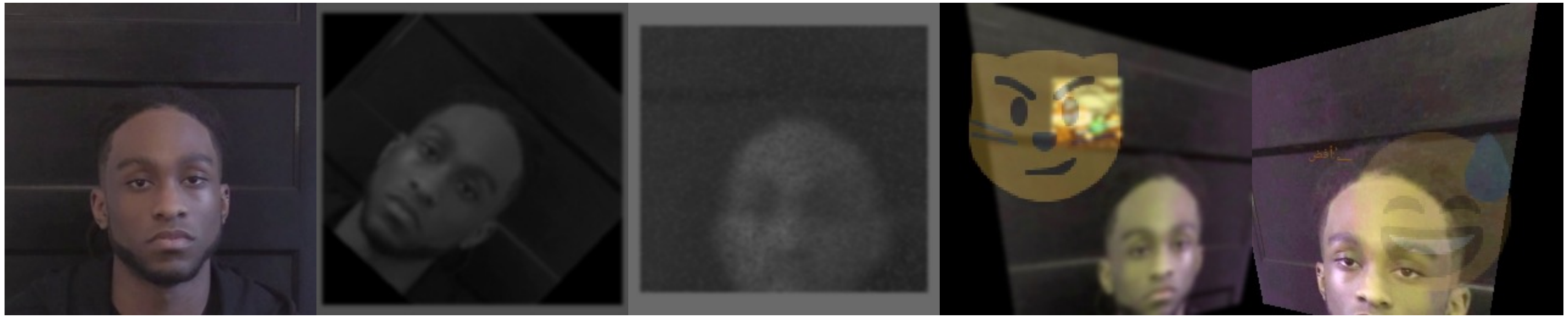


Training

4. Basic + Super-dark augmentation

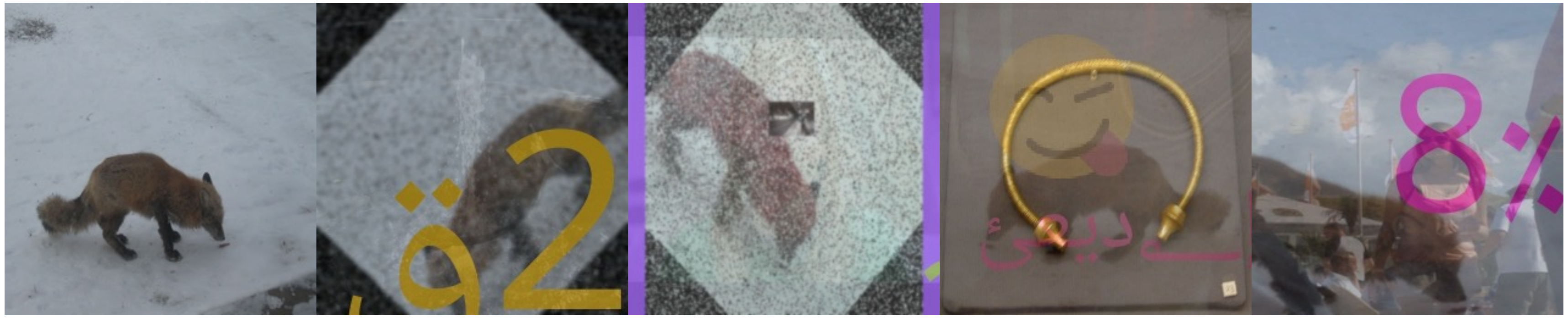


5. Basic + Super-face augmentation

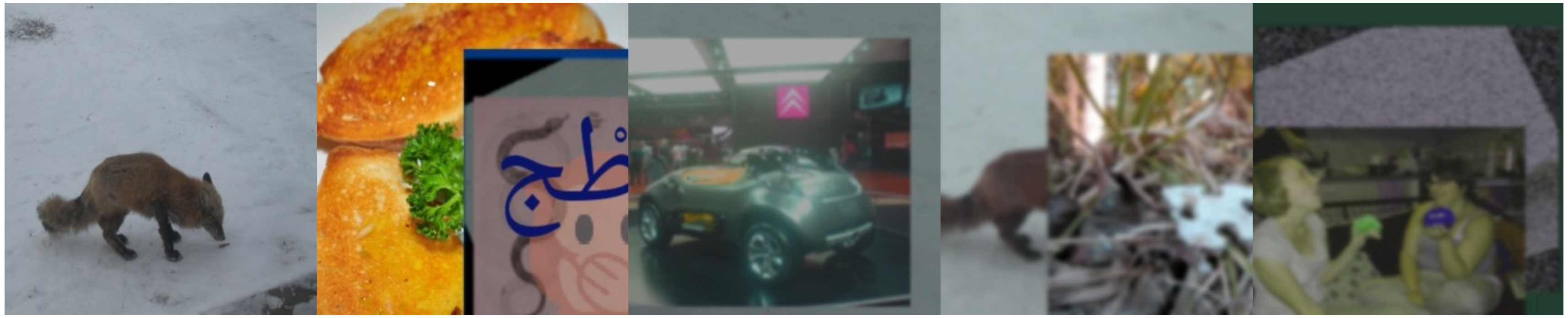


Training

6. Basic + Super-opaque augmentation



7. Basic + Super-occlude augmentation



Training

Grayscale augmentation

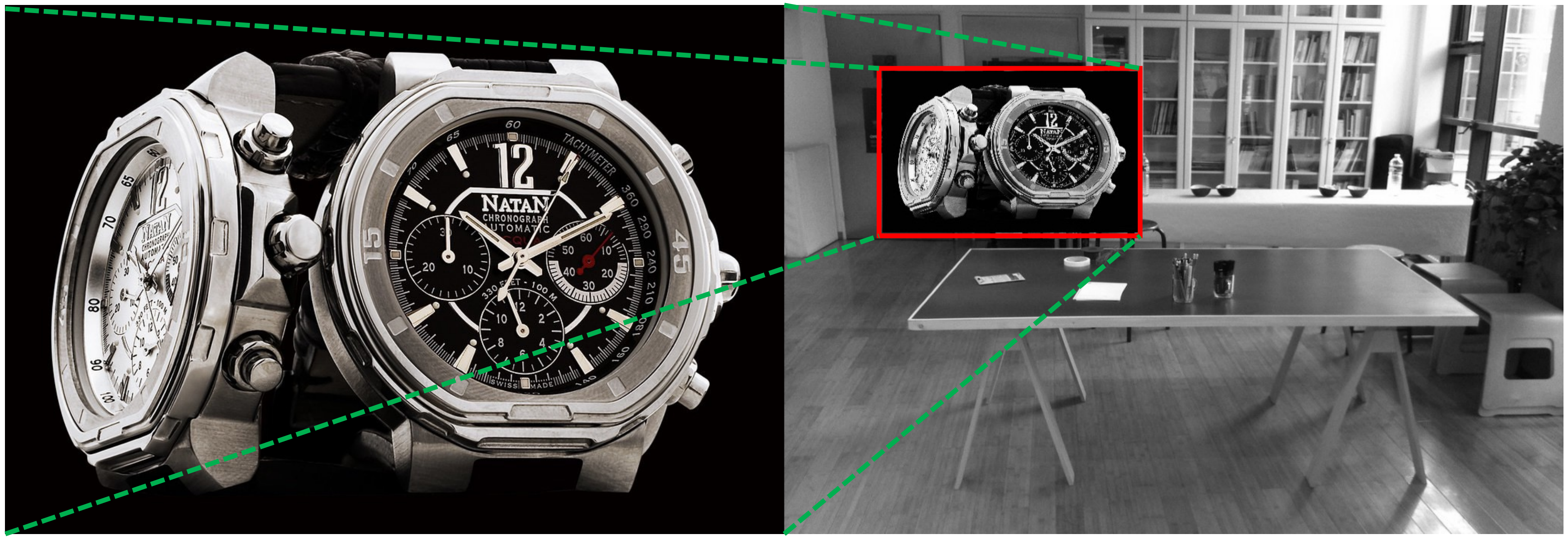
The augmentation changes all the color images into ***grayscale style***.

Some examples

Test

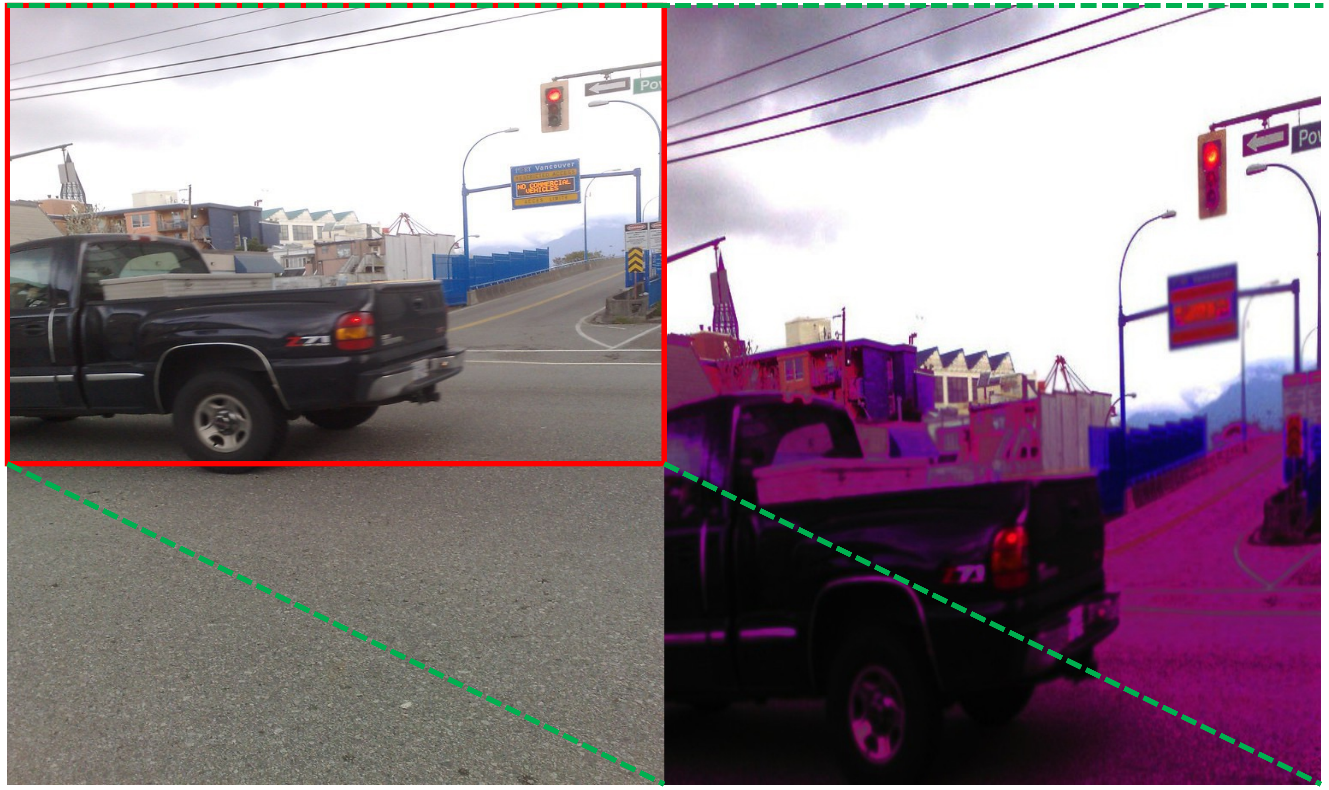
Two corner cases:

(1) Some query images are generated by overlaying a reference image on top of a distractor image.

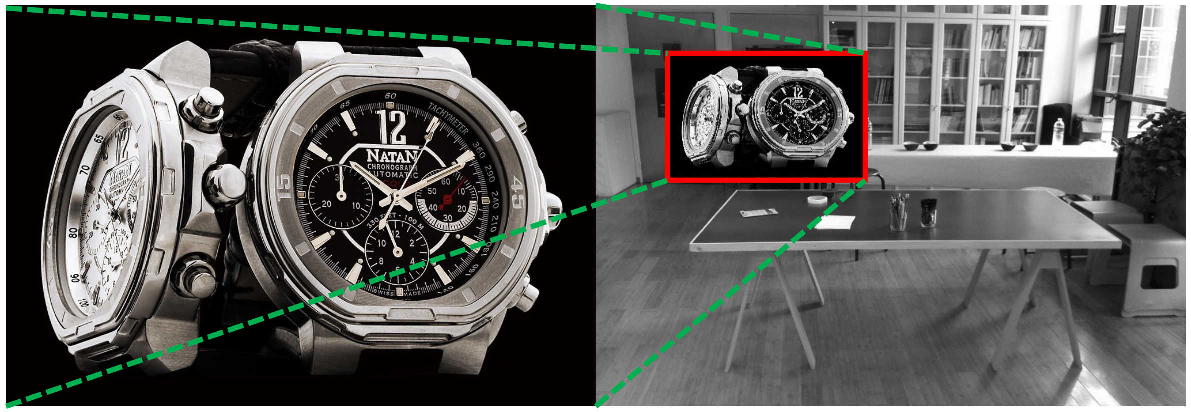


Test

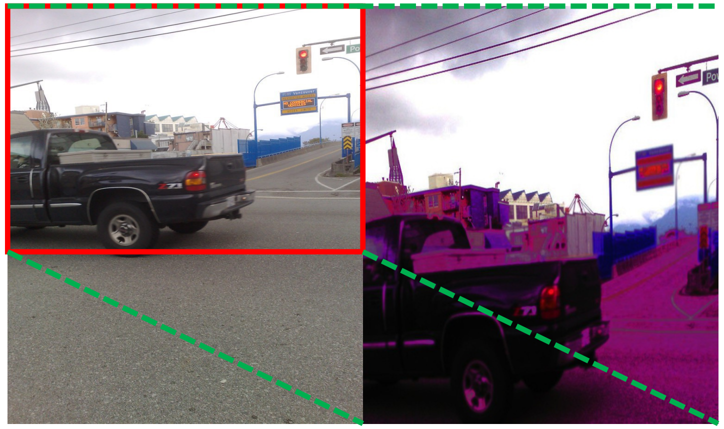
(2) Some queries are cropped from the reference images and thus only contain parts of the reference images.



Test



Global-local matching strategy

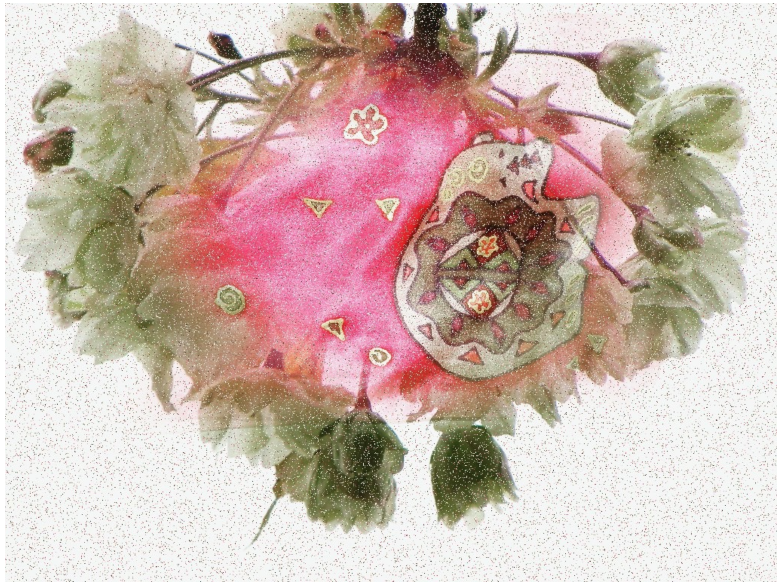


Local-global matching strategy

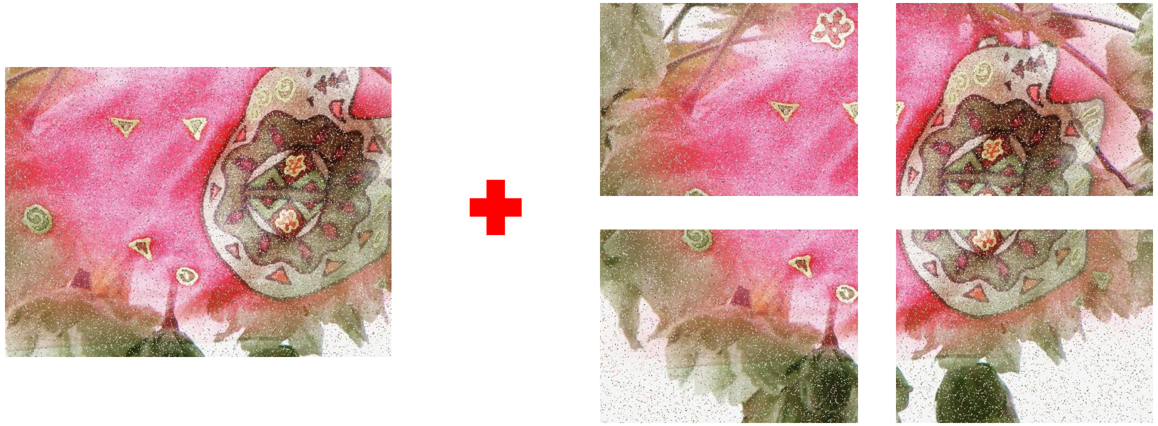
Test

Generate local features of query images

Crop centers



Original image

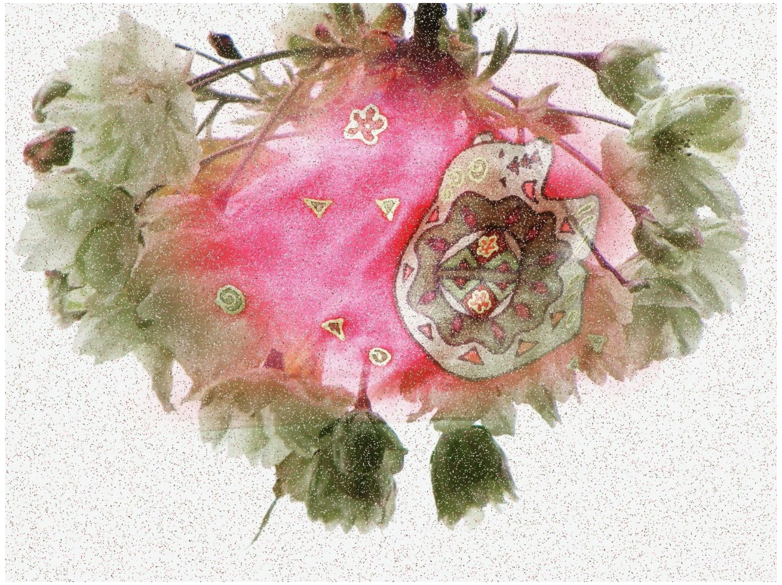


Cropped centers

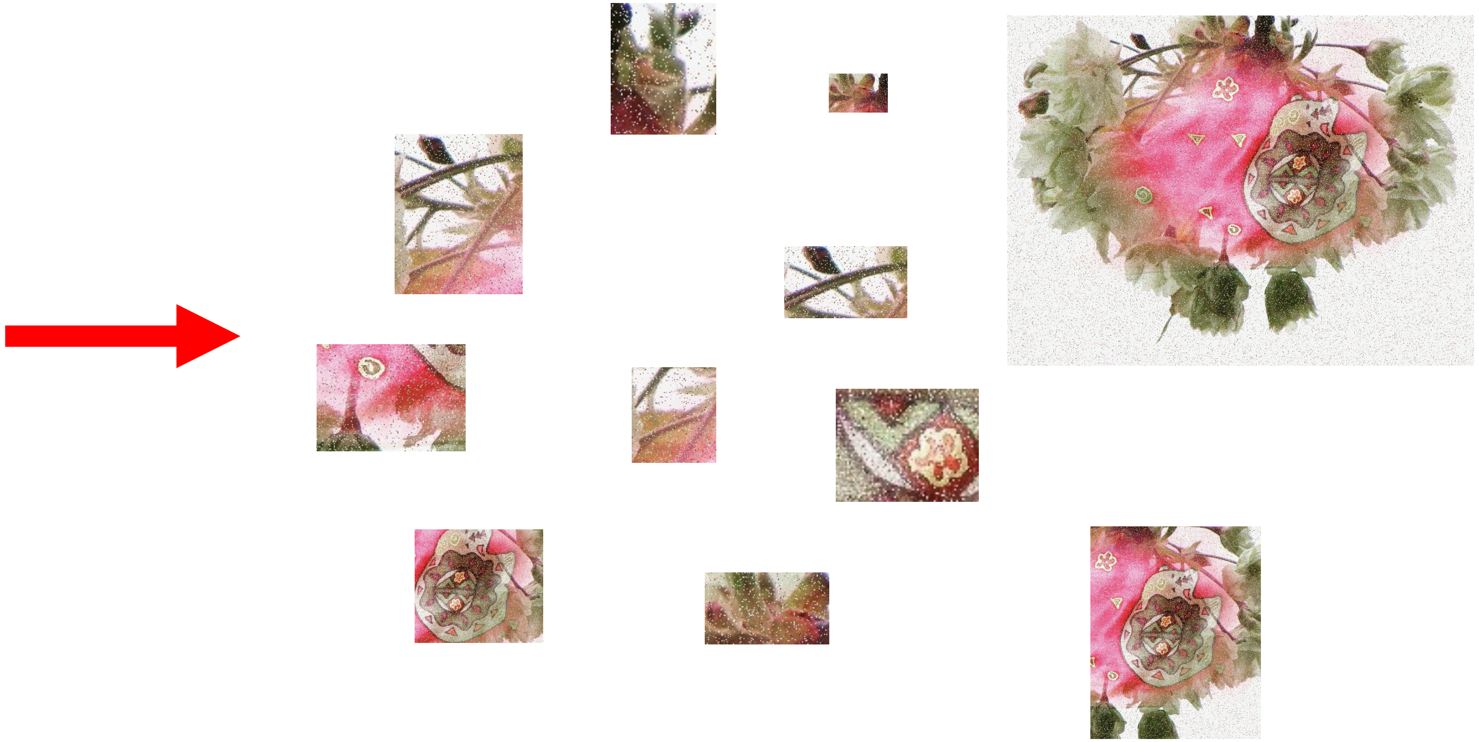
Test

Generate local features of query images

Selective search



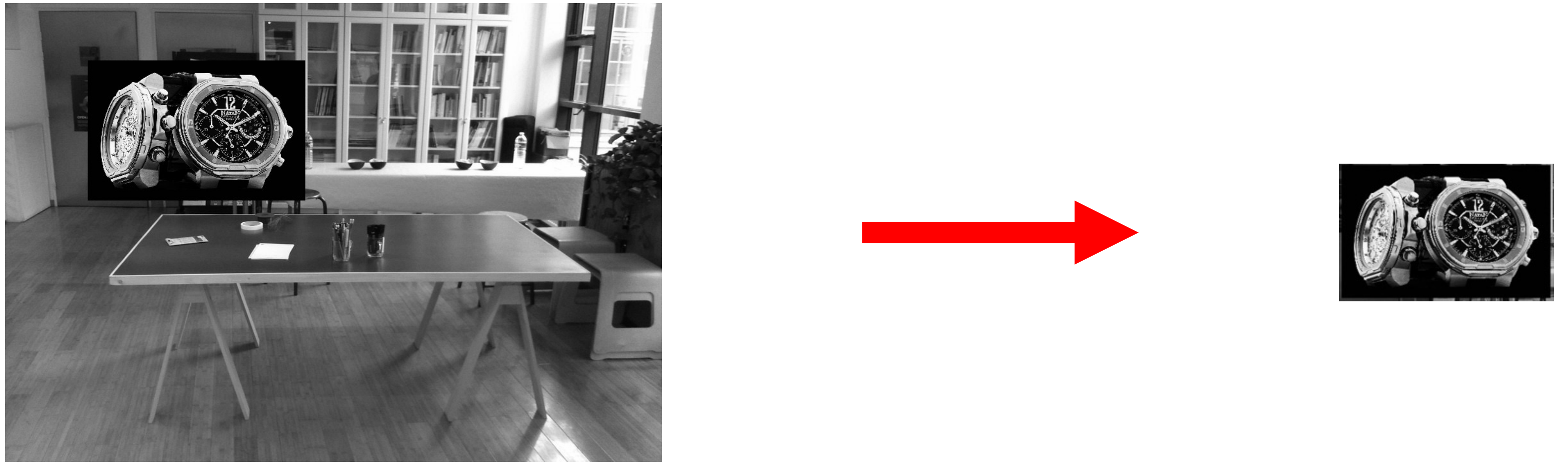
Original image



Test

Generate local features of query images

Detection

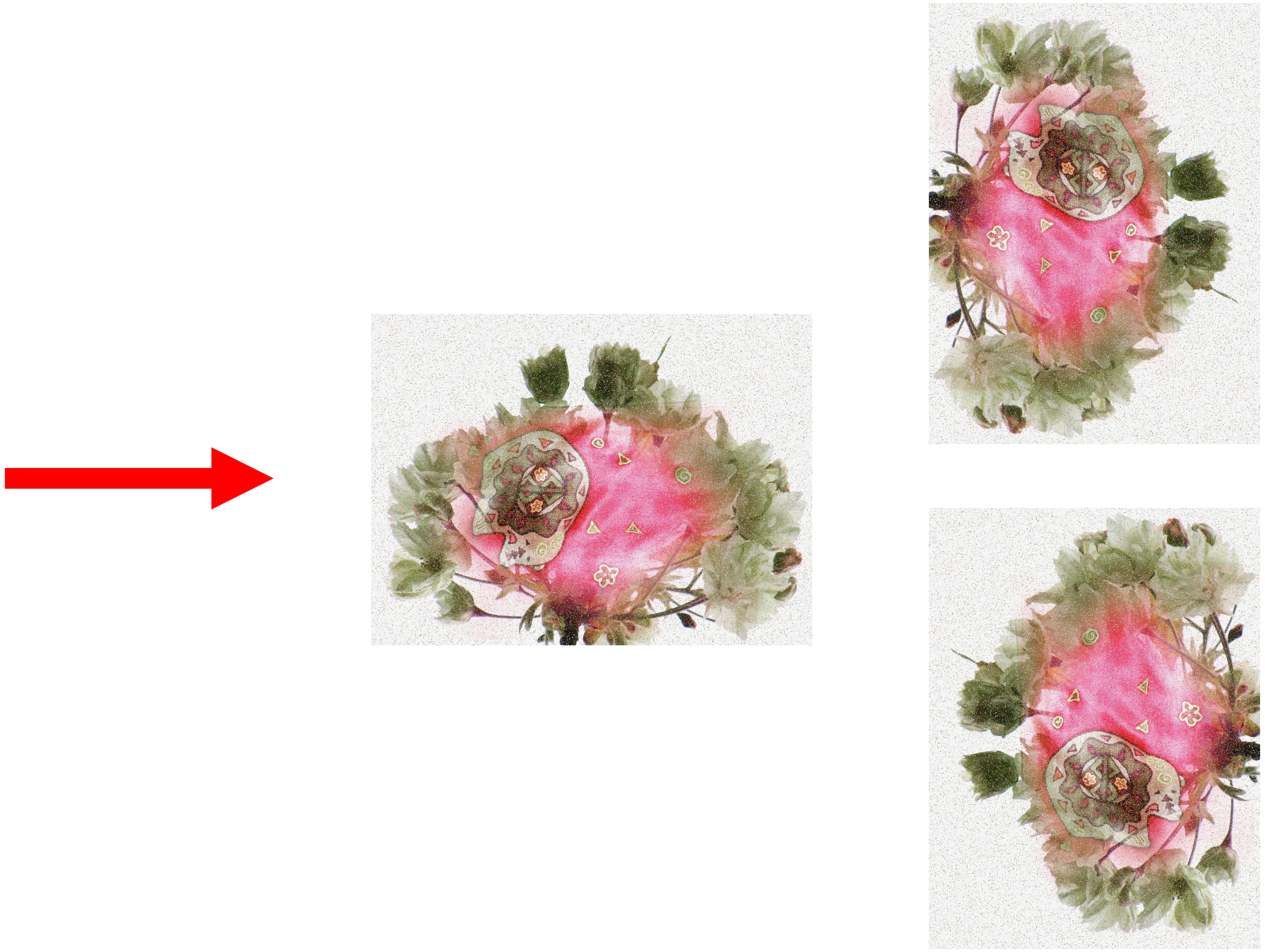


Original image

Test

Rotating

Original image

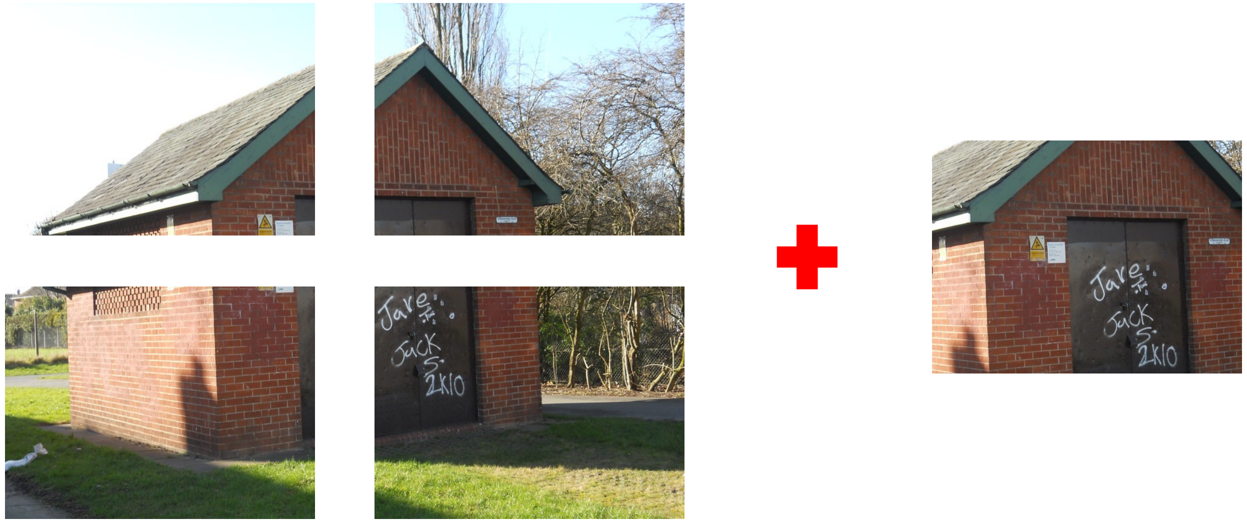


Test

Generate local features of reference images

1) Dividing into 5 large parts

Original image



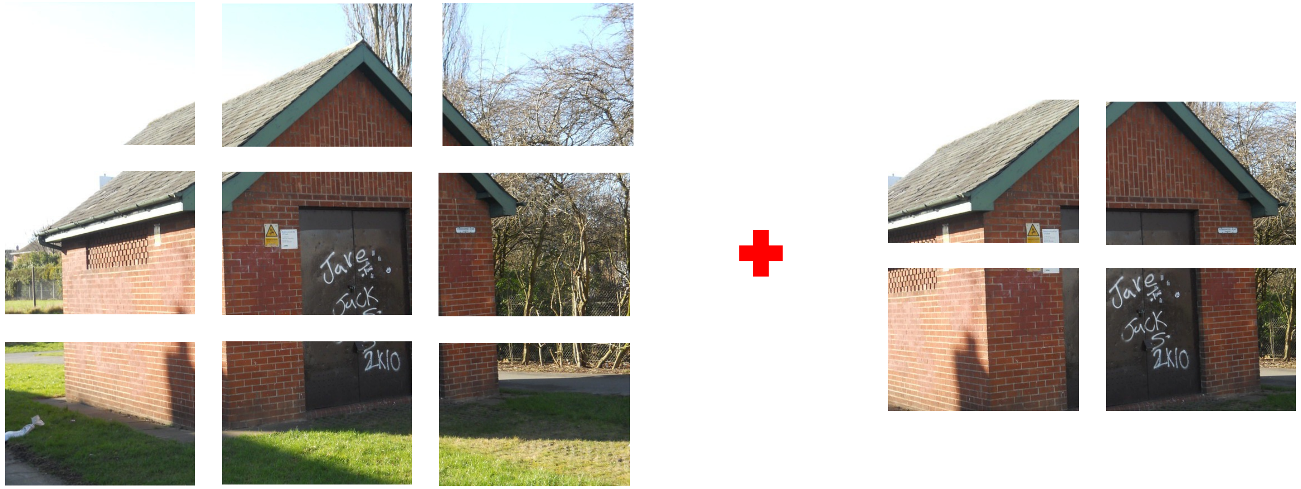
Divided images

Test

Generate local features of reference images

1) Dividing into 5 large parts 2) Dividing into 13 small parts

Original image



Divided images

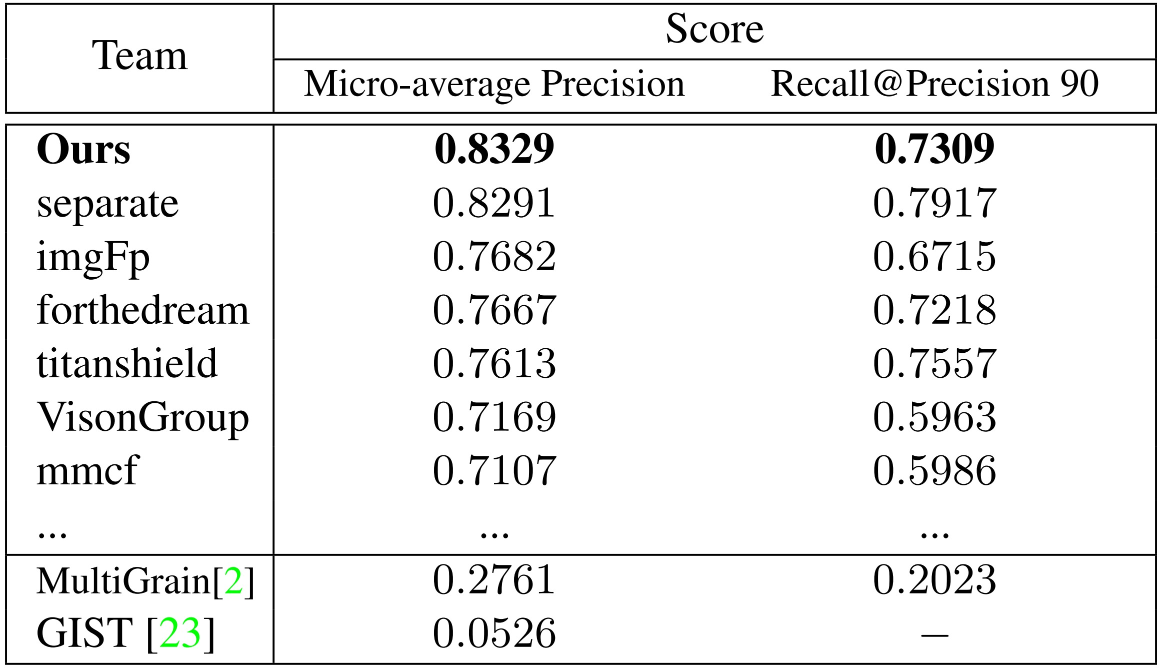
Experiments

Ablation Studies

Method	Score	
	Micro-average Precision	Recall@Precision 90
Supervised	0.68726	0.54678
Unsupervised	0.70813	0.62773
Global-local	0.82726	0.74755
Both	0.83720	0.75155
Adv-Aug	0.88640	0.80124
Multi+Tricks	**0.90035**	**0.81887**

Experiments

Comparison with State-of-the-Arts



Team	Score	
	Micro-average Precision	Recall@Precision 90
Ours	**0.8329**	**0.7309**
separate	0.8291	0.7917
imgFp	0.7682	0.6715
forthedream	0.7667	0.7218
titanshield	0.7613	0.7557
VisonGroup	0.7169	0.5963
mmcf	0.7107	0.5986
...	...	...
MultiGrain[2]	0.2761	0.2023
GIST [23]	0.0526	–

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS FOR LISTENING

Speaker: Wenhao Wang

Baidu Research