Feature-compatible Progressive Learning for Video Copy Detection

Wenhao Wang, Yifan Sun, Yi Yang

ReLER, University of Technology Sydney

Baidu Inc.

Zhejiang University

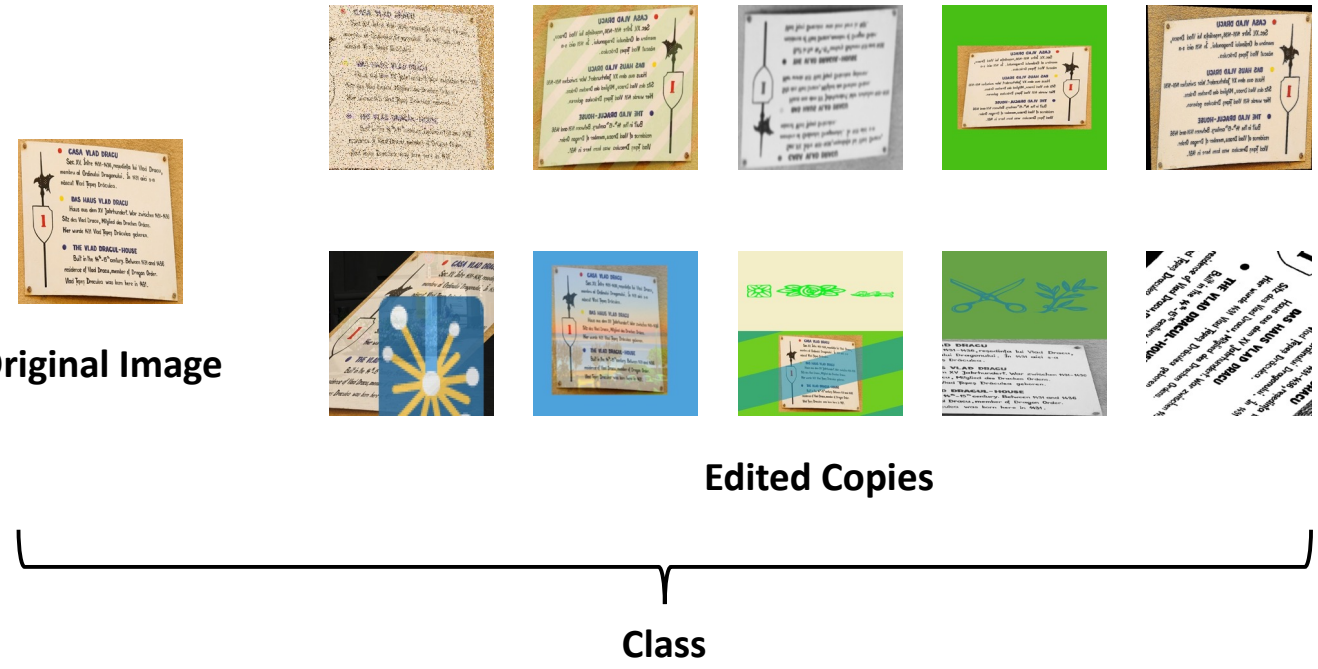
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Background

- Our second place solutions to the Meta AI Video Similarity Challenge (VSC22), CVPR 2023
- Matching track: identifies specific clips
- Descriptor track: 512-dimensional vector representations
- Inspirations from: ISC21-winning solutions (FOSSL [1], and CNNCL [2])
- Built on: our previous work (BoT, D^2LV, ASL)

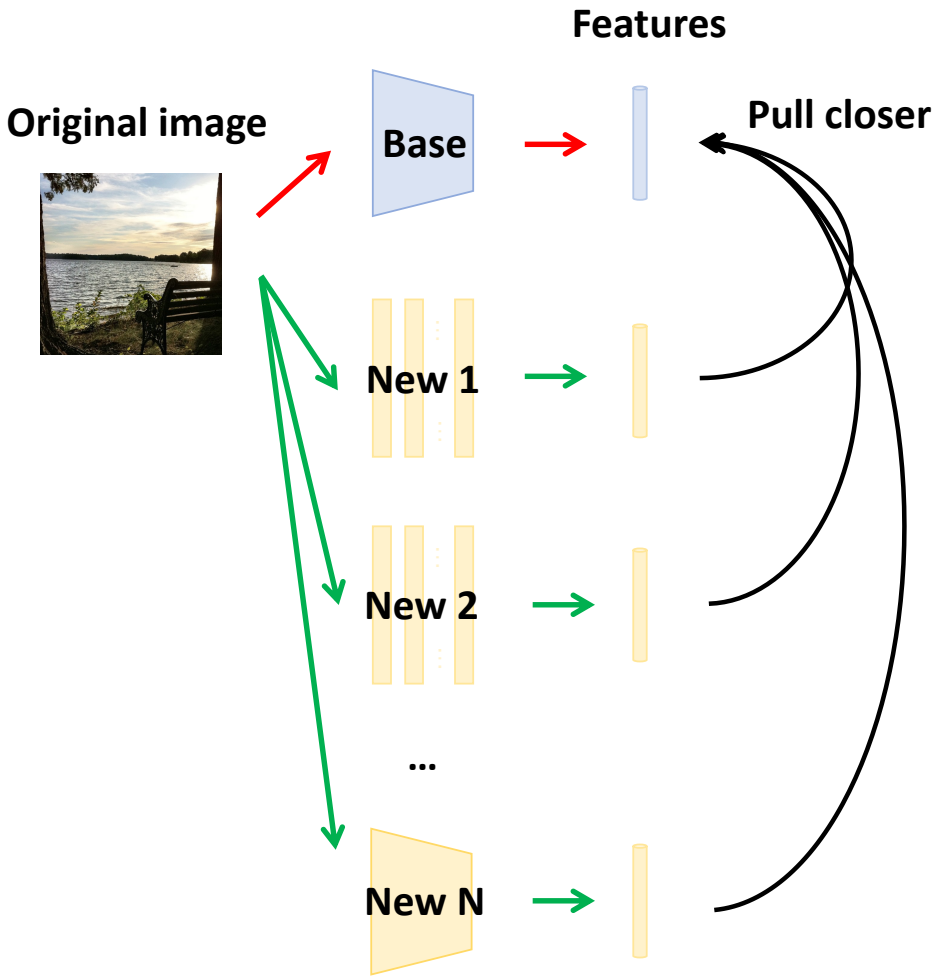
[1] Fossl: Feature compatible self-superviseded learning for large-scale image similarity detection.
[2] Contrastive learning with large memory bank and negative embedding subtraction for accurate copy detection.

Base Training



- Generate edited copies
  - Use training image from ISC21
  - Design 20+ transformations
- Perform deep metric learning
  - CosFace Loss
  - Backbone: CotNet

Feature-compatible Learning



- L_2 distance

$$\mathcal{L}_{\text{com}} = \sum_{i=0}^{N} \left\| \frac{f(x_{o_i})}{\|f(x_{o_i})\|_2} - \frac{g(x_{o_i})}{\|g(x_{o_i})\|_2} \right\|_2$$

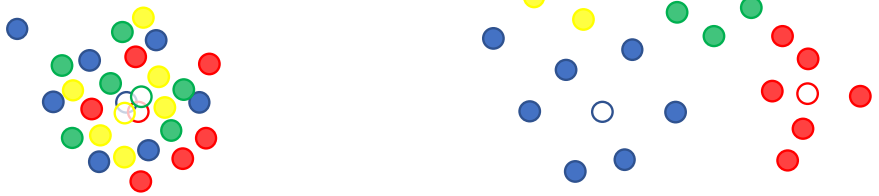
- Final loss

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{mtr}} + \lambda_r \cdot \mathcal{L}_{\text{com}}$$

- Backbones

ResNet-50, ResNext-50, SKNet-50, ViT, Swin Transformer, and T2T-ViT

Feature-compatible Learning



With Compatible-Learning

Without Compatible-Learning

- Base feature of edit copies
- New 1 feature of edit copies
- New 2 feature of edit copies
- New N feature of edit copies
- Base feature of an original image
- New 1 feature of an original image
- New 2 feature of an original image
- New N feature of an original image

- Ensemble

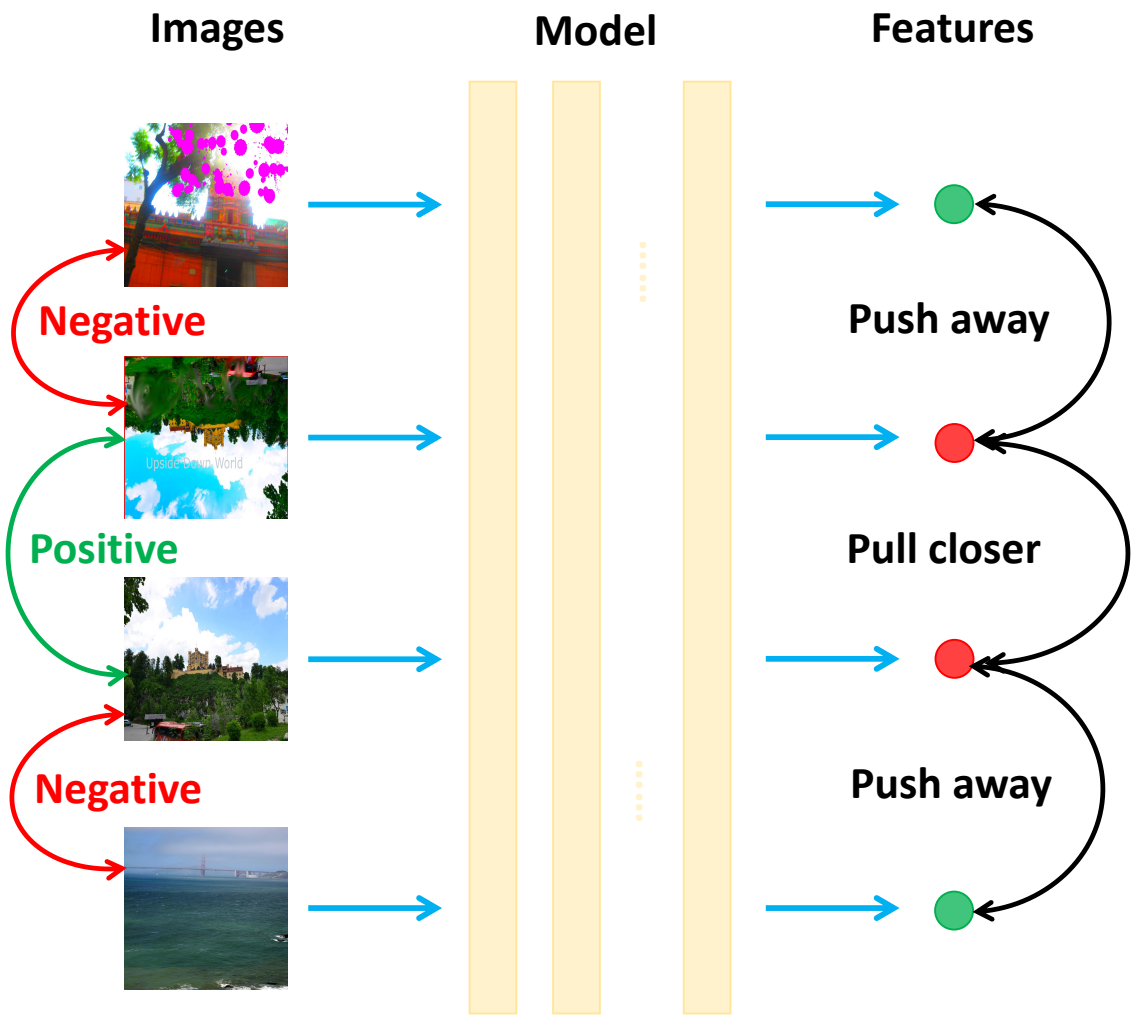
- query:

$$\frac{1}{N}\sum_{i=1}^{N} g_i(q)$$

- reference:

$$\frac{1}{N}\sum_{i=1}^{N} g_i(r)$$

Fine-tuning on the GT pairs

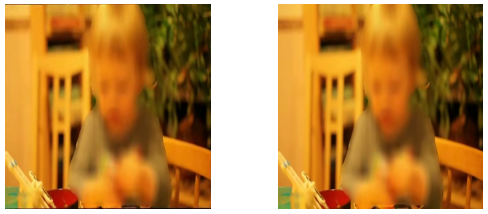


$$\mathcal{L}_{\text{pos}} = \sum_{i=0}^{M} \left\| \frac{g_t\left(x_{p_i}^1\right)}{\left\|g_t\left(x_{p_i}^1\right)\right\|_2} - \frac{g_t\left(x_{p_i}^2\right)}{\left\|g_t\left(x_{p_i}^2\right)\right\|_2} \right\|_2$$

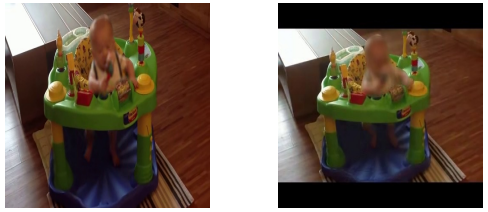
$$\mathcal{L}_{neg} = \frac{1}{2} \sum_{i=0}^{M} \left(\left\| \frac{g_t\left(x_{p_i}^1\right)}{\left\|g_t\left(x_{p_i}^1\right)\right\|_2} - \frac{g_t\left(x_{n_i}^1\right)}{\left\|g_t\left(x_{n_i}^1\right)\right\|_2} \right\|_2 + \left\| \frac{g_t\left(x_{p_i}^2\right)}{\left\|g_t\left(x_{p_i}^2\right)\right\|_2} - \frac{g_t\left(x_{n_i}^2\right)}{\left\|g_t\left(x_{n_i}^2\right)\right\|_2} \right\|_2 \right)$$

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{mtr}} + \lambda_r \cdot \mathcal{L}_{\text{com}} + \lambda_{\text{pn}} \cdot (\mathcal{L}_{\text{pos}} - \mathcal{L}_{\text{neg}})$$

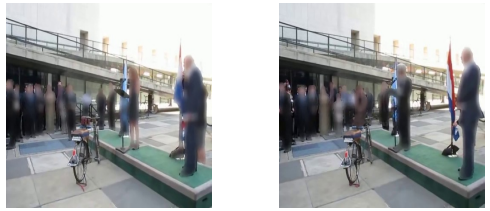
Visualization



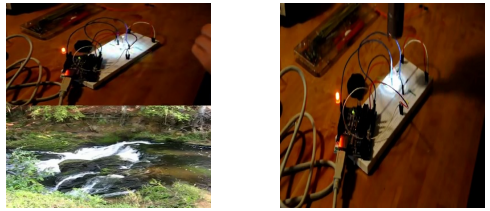
Similarity = 0.55



Similarity = 0.45



Similarity = 0.35

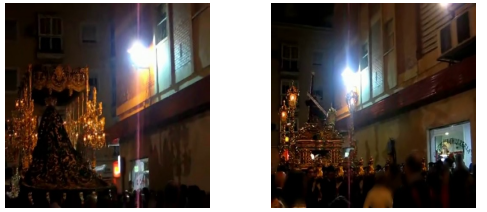


Similarity = 0.25

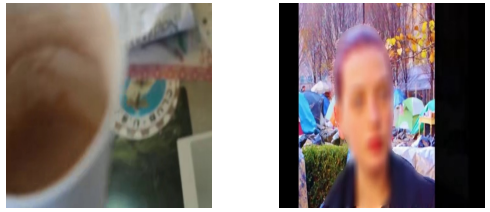
Similarity = 0.15

Similarity = 0.05

Similarity = -0.05



Similarity = -0.15



Similarity = -0.25

Comparison

Matching track

Team	$\mu AP(\%) \uparrow$
do something more	91.53
CompetitionSecond (Ours)	**77.11**
cvl-matching	70.36
People-AI	50.72
Baseline	44.11
...	...

Descriptor track

Team	$\mu AP(\%) \uparrow$
do something	87.17
FriendshipFirst (Ours)	**85.14**
cvl-descriptor	83.62
Zihao	77.29
People-AI	68.84
Baseline	60.47
...	...

Thanks for your listening