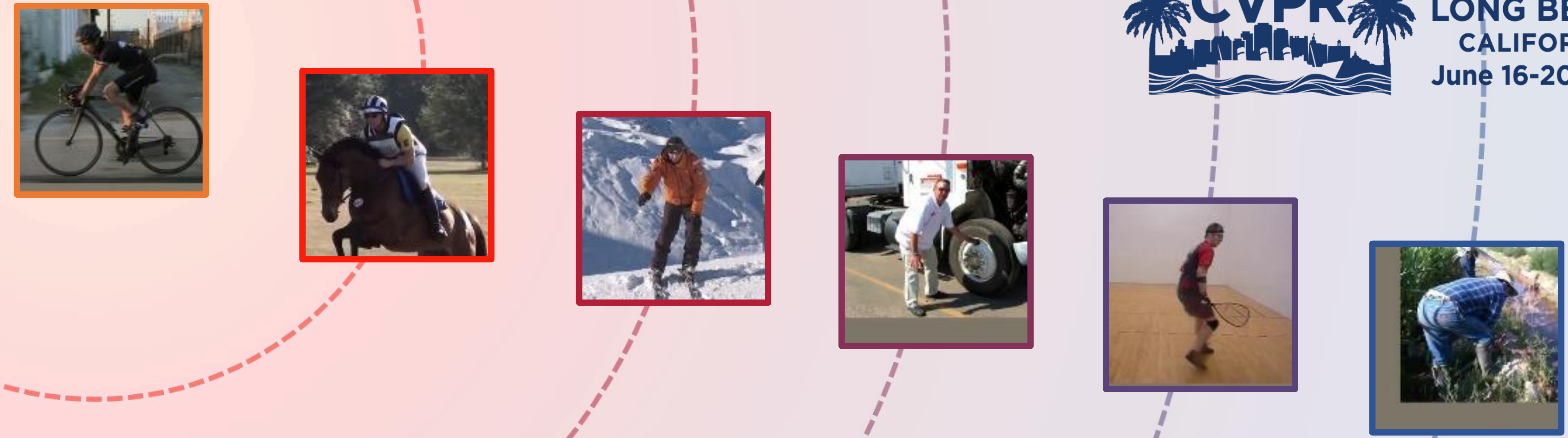
CVPR LONG BEACH CALIFORNIA June 16-20, 2019



Deep Metric Learning Beyond Binary Supervision

Sungyeon Kim Minkyo Seo Ivan Laptev Minsu Cho Suha Kwak

{tjddus9597, mkseo, mscho, suha.kwak}@postech.ac.kr, ivan.laptev@inria.fr

POSTECH Inria

How much similar/dissimilar?

D ( , ) $<$ D ( , )

Metric: Function that quantifies a distance

Metric Learning: Learning a metric from a set of data

Deep Metric Learning

$D(f(\mathbf{x}_1), f(\mathbf{x}_2))$ $D(f(\mathbf{x}_1), f(\mathbf{x}_3))$

$\mathbf{x}_2$

$\mathbf{x}_1$

$\mathbf{x}_3$

Pairwise relation

$D(f_1, f_2) \downarrow, D(f_1, f_3) \uparrow$

Triplet relation

$D(f_1, f_2) < D(f_1, f_3)$

…

Deep Metric Learning

Learning a deep neural net f that satisfies the relations

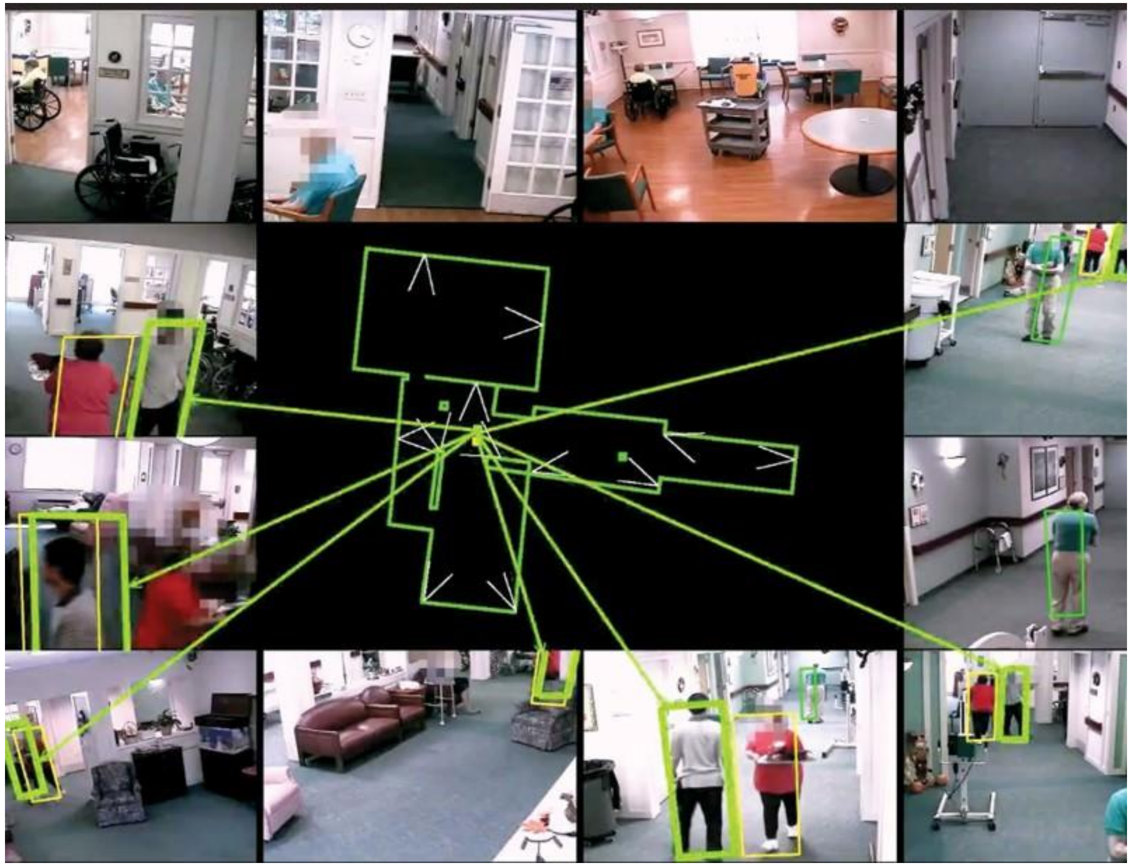
Applications

Content-based image retrieval

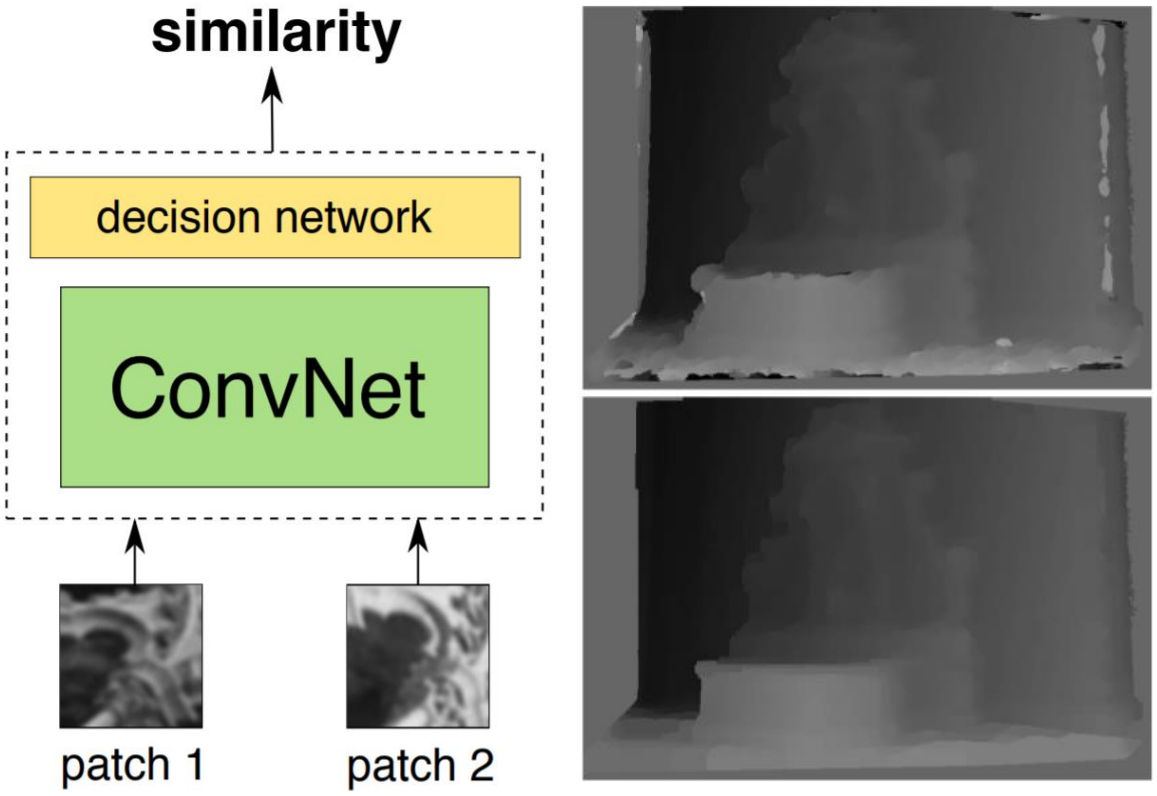
Face verification/identification[1]

[1] FaceNet: A unified embedding for face recognition and clustering, CVPR 2015

Applications



Person re-identification[2]



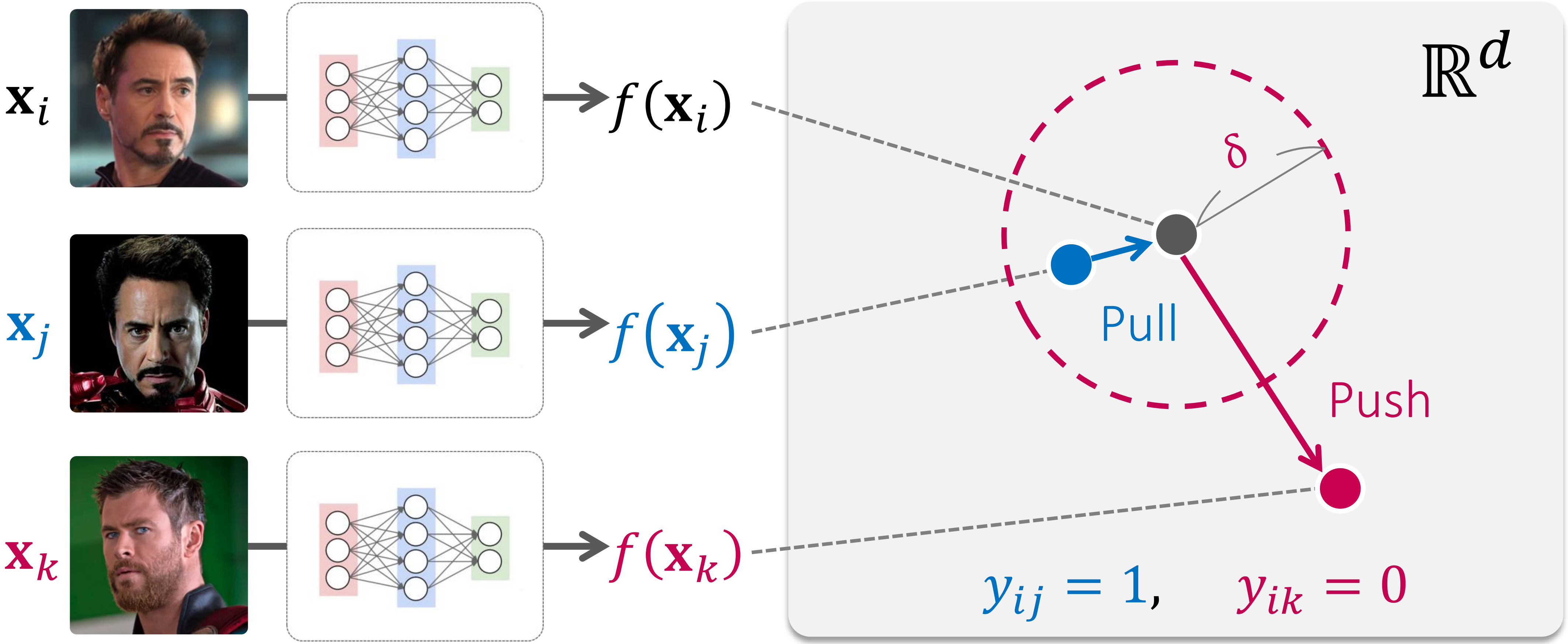
Patch matching/stereo imaging[3]

[2] Beyond triplet loss: a deep quadruplet network for person re-identification, CVPR 2017
[3] Learning to compare image patches via convolutional neural networks, CVPR 2015

Existing Approaches

- Contrastive loss for Siamese networks[4]

$$\ell_{\mathrm{ctr}}(i,j) = y_{ij} D\left(f_i, f_j\right)^2 + \left(1 - y_{ij}\right)\left[\delta - D(f_i, f_j)\right]_+^2$$

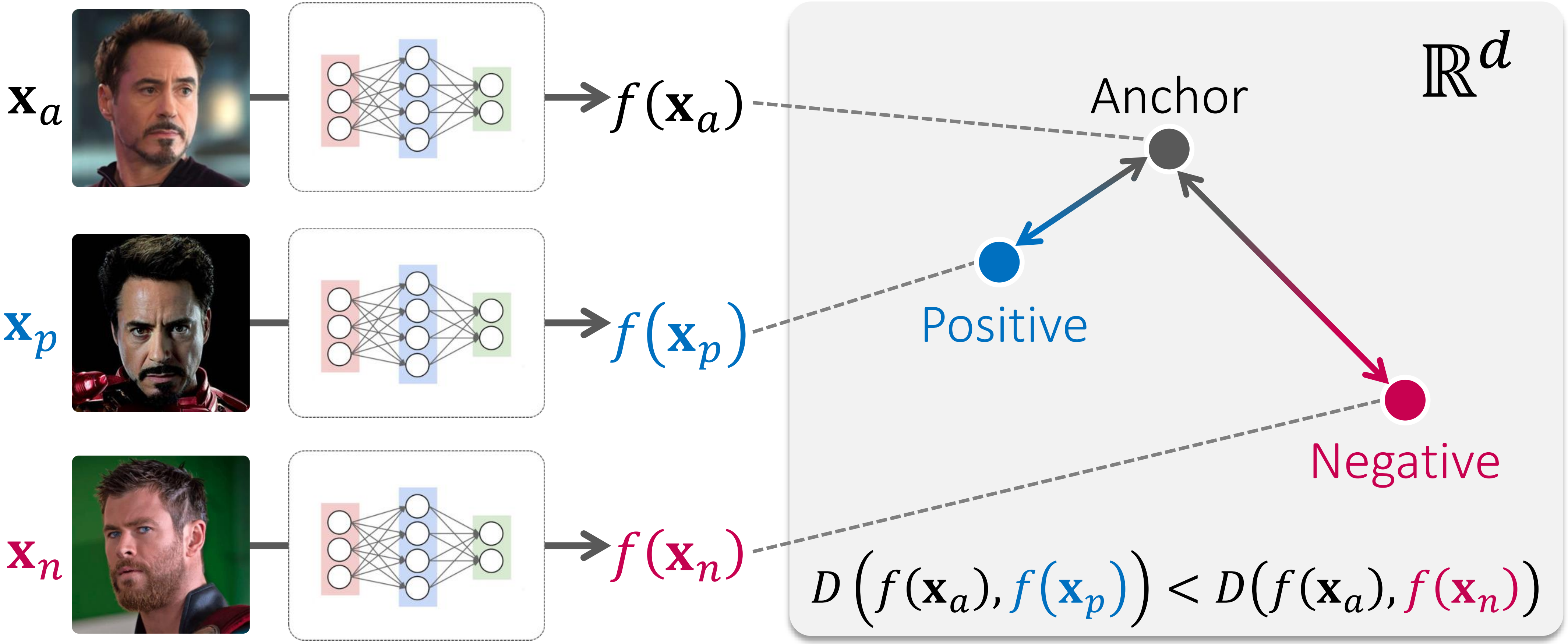


[4] Learning a similarity metric discriminatively with application to face verification, CVPR 2005

Existing Approaches

- Triplet rank loss for triplet networks[1]

$$\ell_{\text{tri}}(a,p,n) = \left[D\left(f_a, f_p\right) - D\left(f_a, f_n\right) + \delta\right]_+$$



[1] FaceNet: A unified embedding for face recognition and clustering, CVPR 2015

Existing Approaches

- A common issue
  - Existing (deep) metric learning approaches rely on binary relations between images: "*same*" or "*not*".

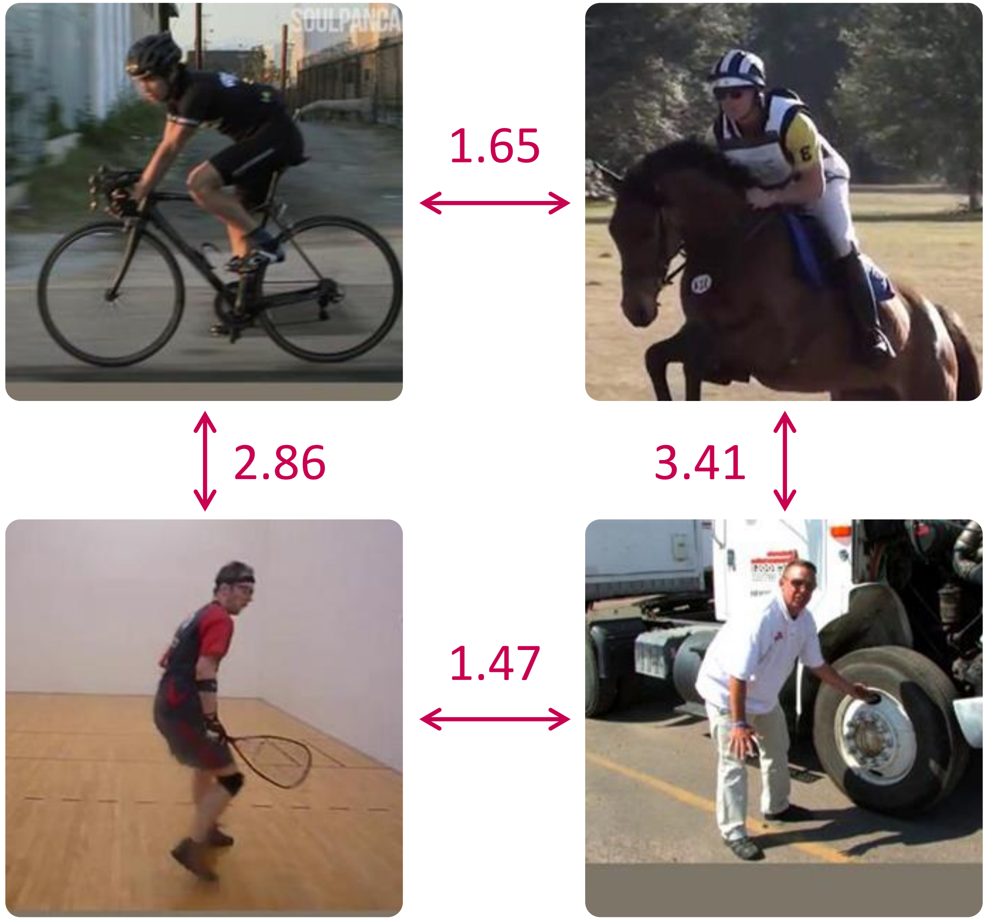
Face verification

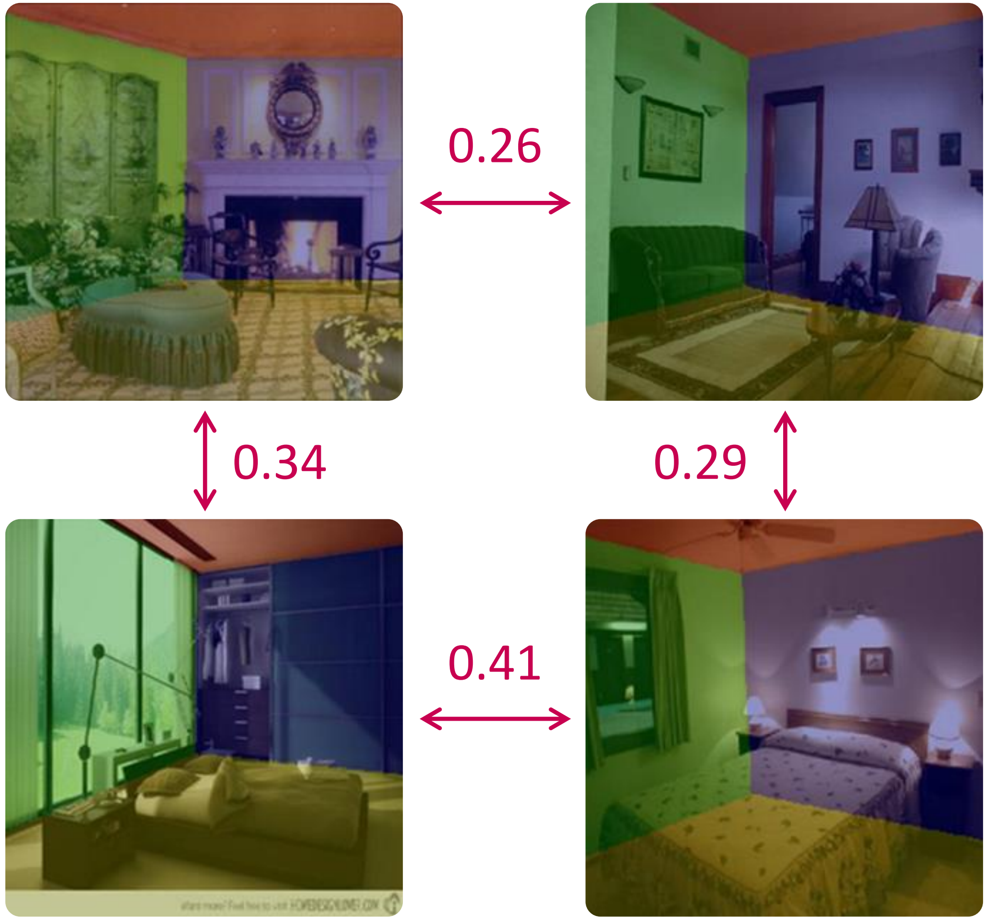
Content-based image retrieval

Person re-identification

Existing Approaches

- A common issue
  - However, relations between real world images are *not binary* but often represented as *continuous similarities*.

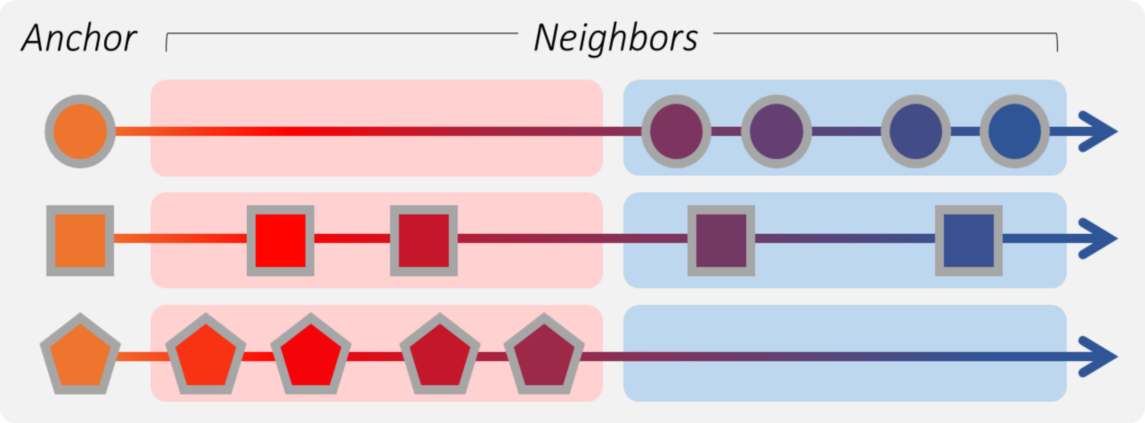




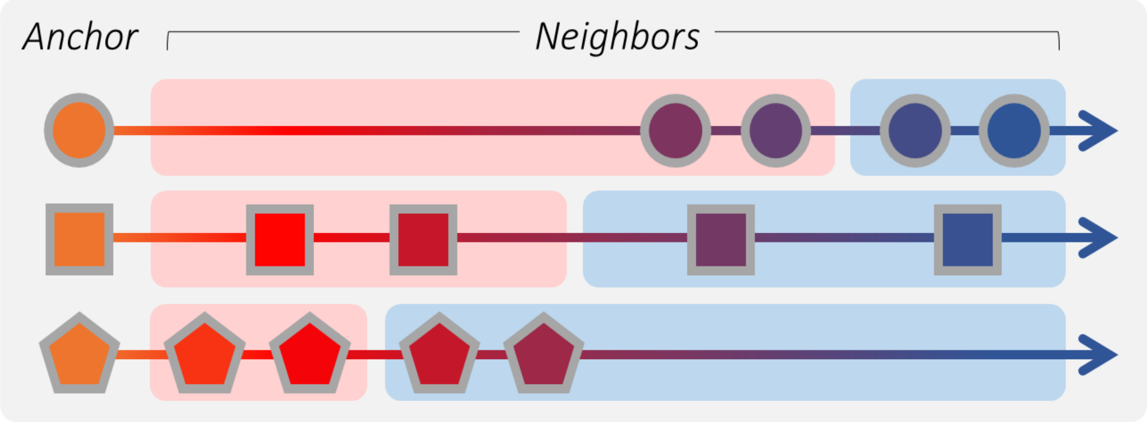
Existing Approaches

- Conventional approaches to handle the issue
  - Existing metric learning loss + *similarity quantization*

Binary thresholding[5]
Populations of positive and negative examples would be significantly imbalanced.



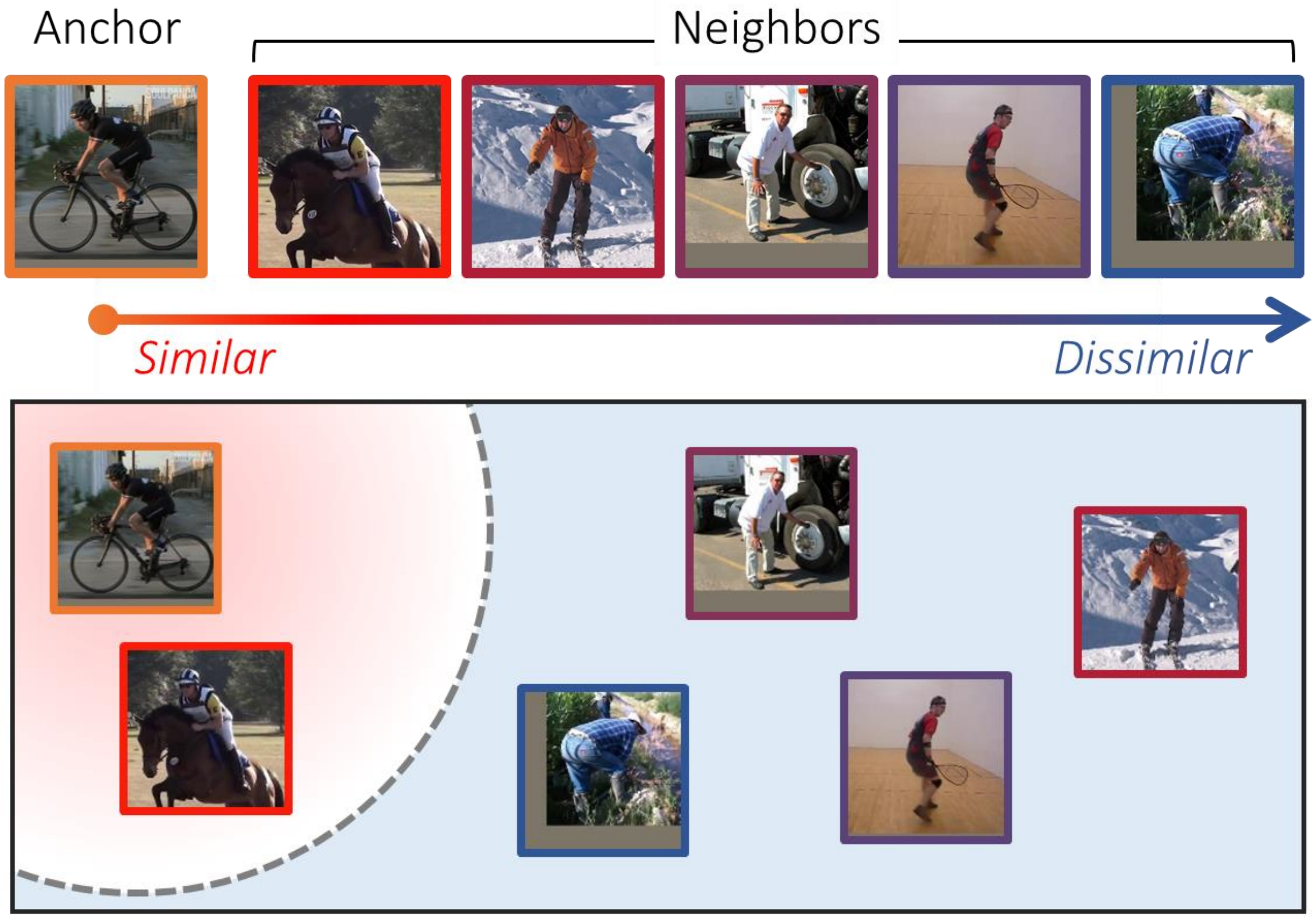
Nearest neighbor search[6]
Positive neighbors of a rare example would be dissimilar and negative neighbors of a common example would be too similar.



[5] Pose embeddings: A deep architecture for learning to match human poses, arXiv 2015
[6] Thin-slicing for pose: Learning to understand pose without explicit pose estimation, CVPR 2016

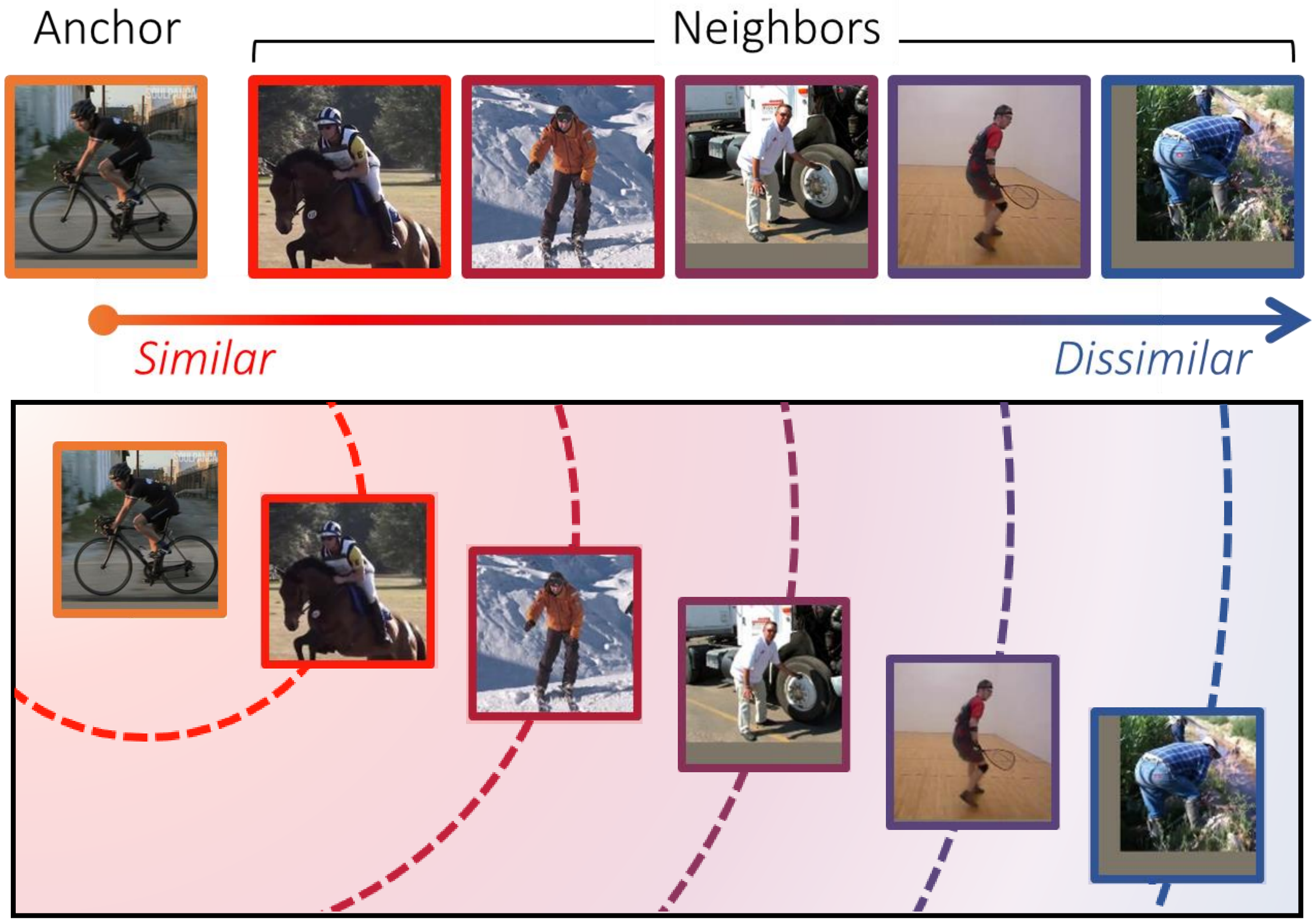
Existing Approaches

- Conventional approaches to handle the issue
  - Degree of similarity is ignored in the learned embedding space.



Our Approach

- Our goal
  - Learning a metric space that reflects the degree of similarity directly



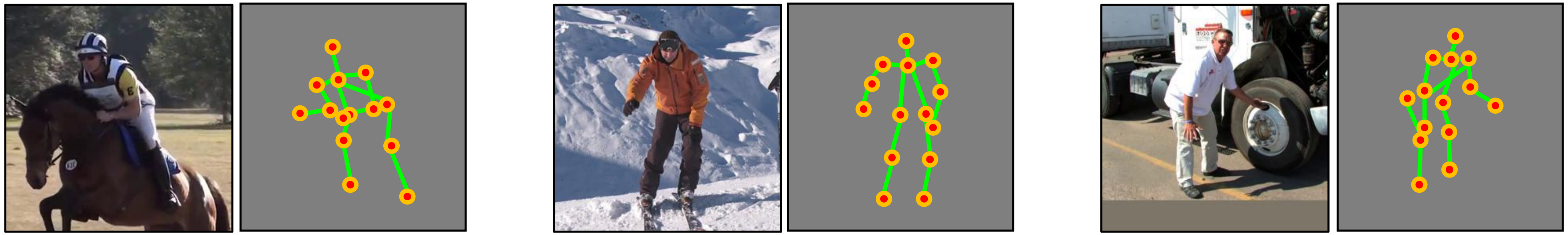
Our Approach

- Our goal
  - Learning a metric space that reflects the degree of similarity directly

- Contributions
  - A new triplet loss: *Log-ratio loss*
  - A new triplet sampling technique: *Dense triplet sampling*
  - Various applications
    - Human pose retrieval
    - Room layout retrieval
    - Caption-aware image retrieval
    - Representation learning for image captioning

Log-ratio Loss

- Definition



$$\mathbf{x}_a \quad \boldsymbol{y}_a \qquad \mathbf{x}_i \quad \boldsymbol{y}_i \qquad \mathbf{x}_j \quad \boldsymbol{y}_j$$

$$\ell_{\mathrm{lr}}(a,i,j) = \left\{\log \frac{D(f_a, f_i)}{D(f_a, f_j)} - \log \frac{D_{\boldsymbol{y}}(\boldsymbol{y}_a, \boldsymbol{y}_i)}{D_{\boldsymbol{y}}(\boldsymbol{y}_a, \boldsymbol{y}_j)}\right\}^2$$

where $f_i \coloneqq f(\mathbf{x}_i)$ is the embedding vector of image i, and $D(\cdot)$ denotes the squared Euclidean distance.

The distance between two images in the learned metric space will be proportional to their distance in the label space.

Log-ratio Loss

- Analysis on its gradients

$$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_a} = -\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_i} - \frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_j}$$

$$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_i} = \frac{(f_i - f_a)}{D(f_a, f_i)} \cdot \ell'_{\mathrm{lr}}(a,i,j)$$

$$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_j} = \frac{(f_a - f_j)}{D(f_a, f_j)} \cdot \ell'_{\mathrm{lr}}(a,i,j)$$

Direction between
the anchor and neighbors

Discrepancy between
the label distance ratio and
the embedding distance ratio

$$4\left\{\log \frac{D(f_a, f_i)}{D(f_a, f_j)} - \log \frac{D_{\mathbf{y}}(\mathbf{y}_a, \mathbf{y}_i)}{D_{\mathbf{y}}(\mathbf{y}_a, \mathbf{y}_j)}\right\}$$

Log-ratio Loss

- Comparison to the triplet rank loss

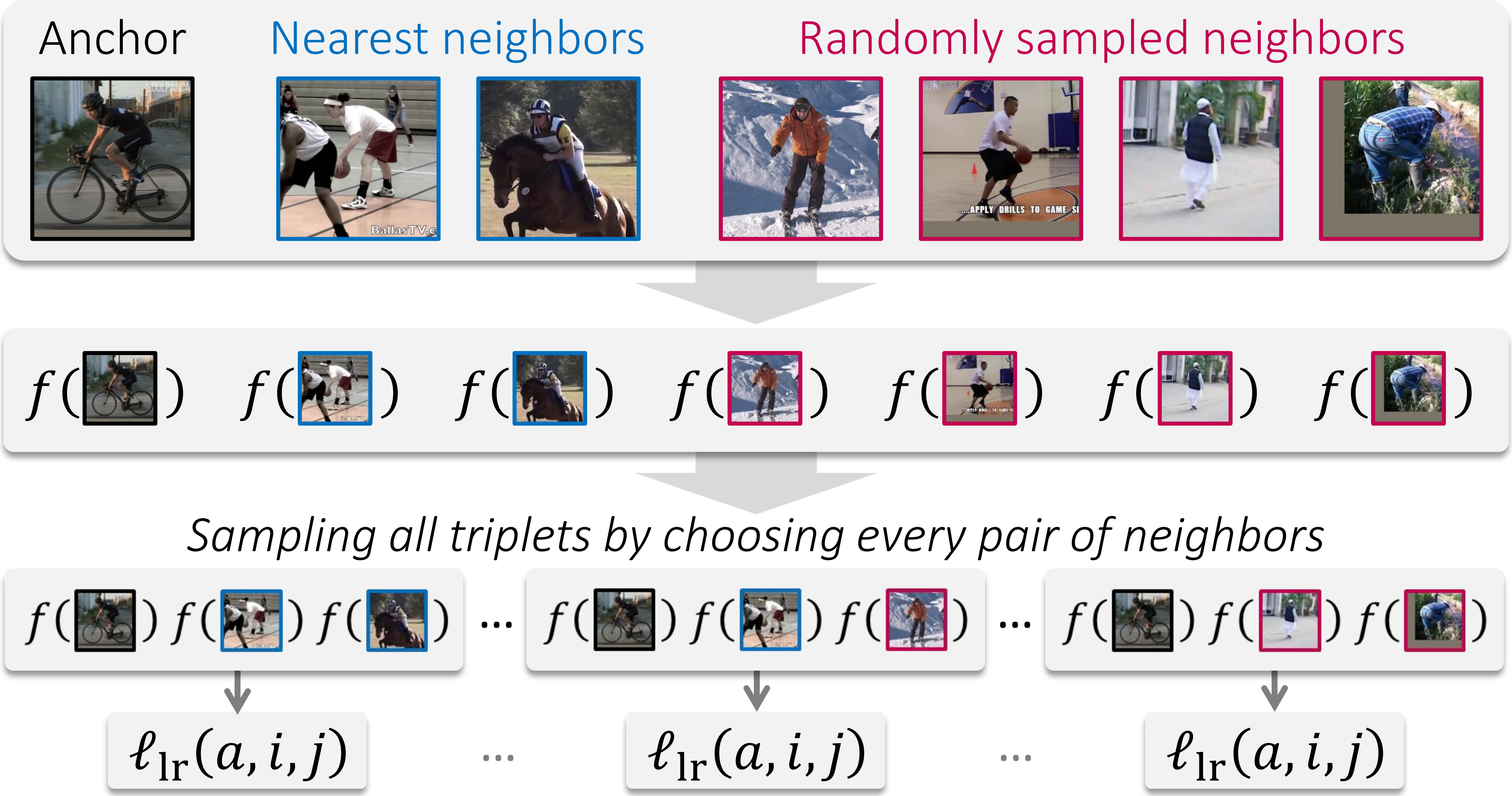
Log-ratio loss	*Triplet rank loss*
$\ell_{\mathrm{lr}}(a,i,j) = \left\{\log \frac{D(f_a,f_i)}{D(f_a,f_j)} - \log \frac{D(y_a,y_i)}{D(y_a,y_j)}\right\}^2$	$\ell_{\mathrm{tri}}(a,i,j) = \left[D(f_a,f_i) - D(f_a,f_j) + \delta\right]_+$
$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_a} = -\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_i} - \frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_j}$	$\frac{\partial \ell_{\mathrm{tri}}(a,i,j)}{\partial f_a} = -\frac{\partial \ell_{\mathrm{tri}}(a,i,j)}{\partial f_i} - \frac{\partial \ell_{\mathrm{tri}}(a,i,j)}{\partial f_j}$
$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_i} = \frac{(f_i - f_a)}{D(f_a,f_i)} \cdot \ell'_{\mathrm{lr}}(a,i,j)$	$\frac{\partial \ell_{\mathrm{tri}}(a,i,j)}{\partial f_i} = 2(f_i - f_a) \cdot \mathbb{I}(\ell_{\mathrm{tri}}(a,i,j) > 0)$
$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_j} = \frac{(f_a - f_j)}{D(f_a,f_j)} \cdot \ell'_{\mathrm{lr}}(a,i,j)$	$\frac{\partial \ell_{\mathrm{tri}}(a,i,j)}{\partial f_j} = 2(f_a - f_j) \cdot \mathbb{I}(\ell_{\mathrm{tri}}(a,i,j) > 0)$
Although the rank constraint holds, the gradients' magnitudes could be significant if $\ell'_{\mathrm{lr}}(a,i,j)$ is large.	The gradients are zero if the triplet satisfies the rank constraint due to the indicator $\mathbb{I}(\ell_{\mathrm{tri}}(a,i,j) > 0)$.

Log-ratio Loss

- Compared to the triplet rank loss, our loss
  - Captures continuous similarities between images better,
    (the triplet rank loss focuses only on partial ranks of similarities.)
  - Does not require any hyperparameter,
    (for the triplet rank loss the margin should be tuned carefully.)
  - Does not demand L_2 normalization of the embedding vectors,
    (such a normalization is essential for the triplet rank loss.)
  - Performs much better with a low embedding dimension.

Dense Triplet Sampling

- Main idea: Using all triplets within a minibatch



Dense Triplet Sampling

- Why not using existing sampling techniques[1,7]
  - They rely on binary relations between images.
  - They are designed to be combined with conventional triplet losses.
  - The notion of hardness is not clear in our setting.

- Our sampling strategy is well matched with the log-ratio loss.
  - The log-ratio loss enables every triplet to well contribute to training.

$$\frac{\partial \ell_{\mathrm{lr}}(a,i,j)}{\partial f_i} = \frac{(f_i - f_a)}{D(f_a, f_i)} \cdot 4 \left\{ \log \frac{D(f_a, f_i)}{D(f_a, f_j)} - \log \frac{D_{\boldsymbol{y}}(\boldsymbol{y}_a, \boldsymbol{y}_i)}{D_{\boldsymbol{y}}(\boldsymbol{y}_a, \boldsymbol{y}_j)} \right\}$$

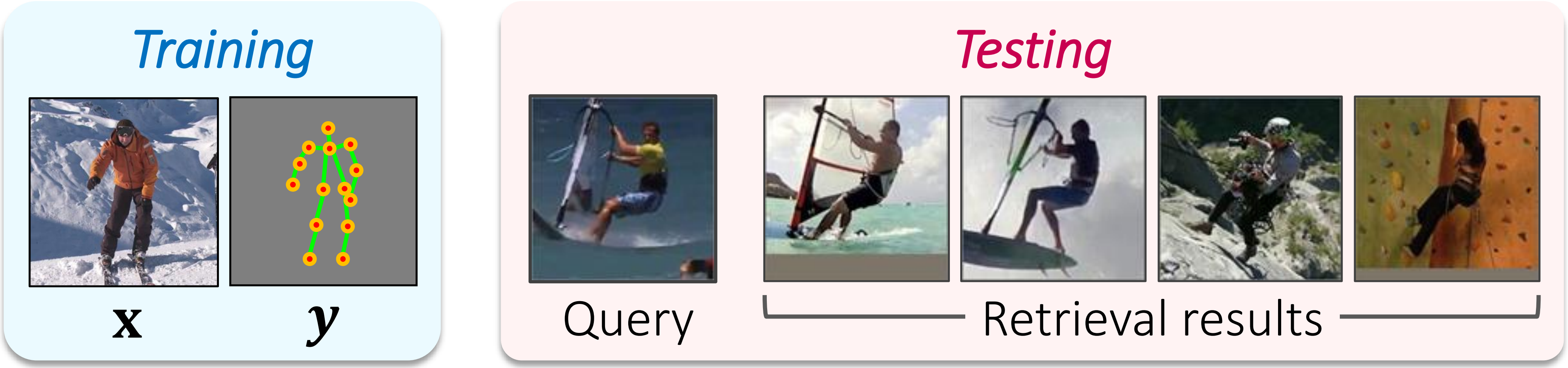
Non-trivial even if the triplet complies the rank constraint

  - *Exploiting all triplets improves embedding performance*.

[1] FaceNet: A unified embedding for face recognition and clustering, CVPR 2015
[7] Sampling matters in deep embedding learning, ICCV 2017

Experiments – Three Retrieval Tasks

- Human pose retrieval



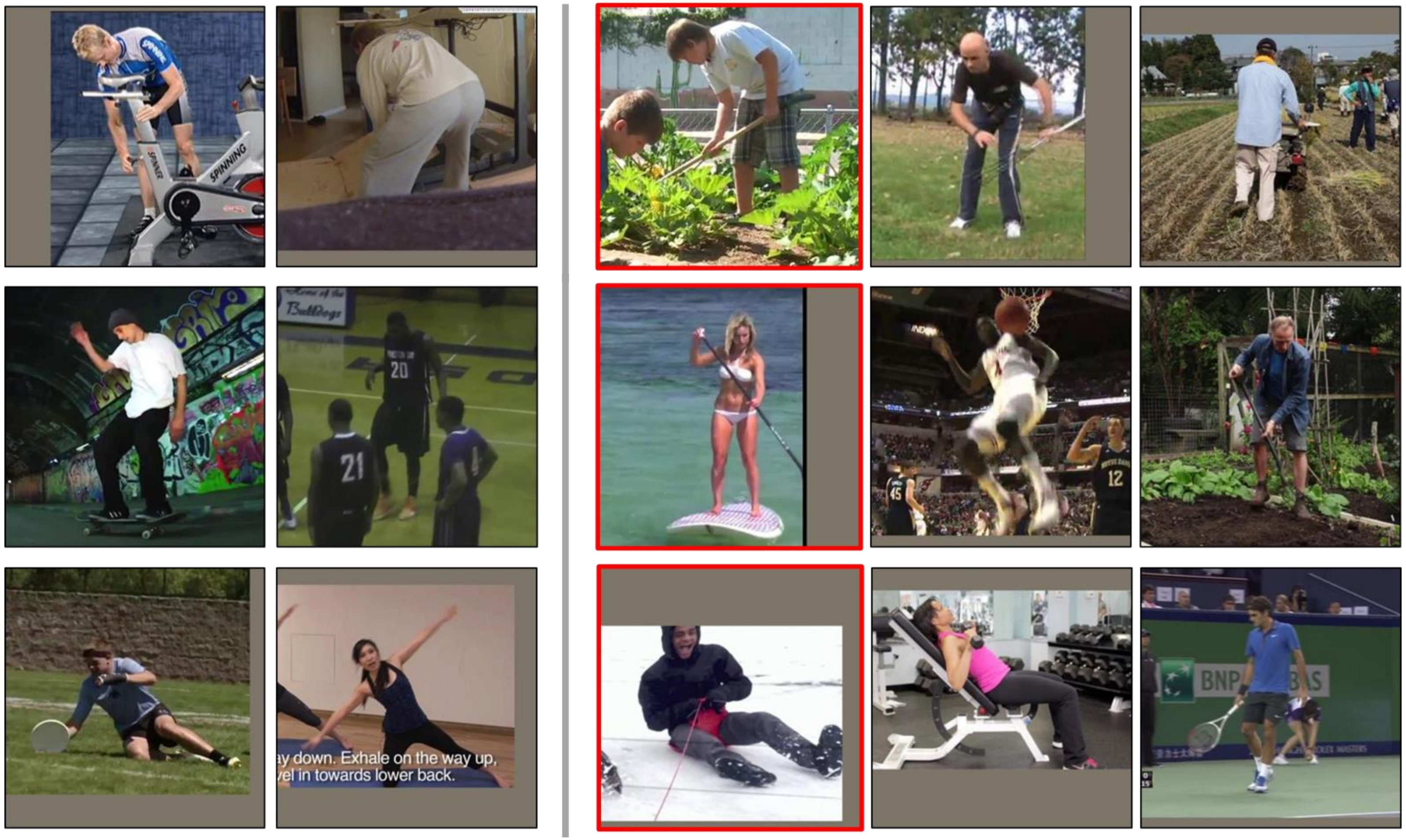
  - Conducted on the *MPII human pose dataset*
  - Similarity between images: *inverse pose distances*
  - Application: *pose-aware representation for action recognition*
  - Label distance between images:

$$D_{\boldsymbol{y}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2,$$

Experiments – Three Retrieval Tasks

- Human pose retrieval

Query Oracle Ours Thin-slicing ResNet34



<u>ResNet34</u>: ImageNet pre-trained network

Typically focuses on objects or background other than human poses.

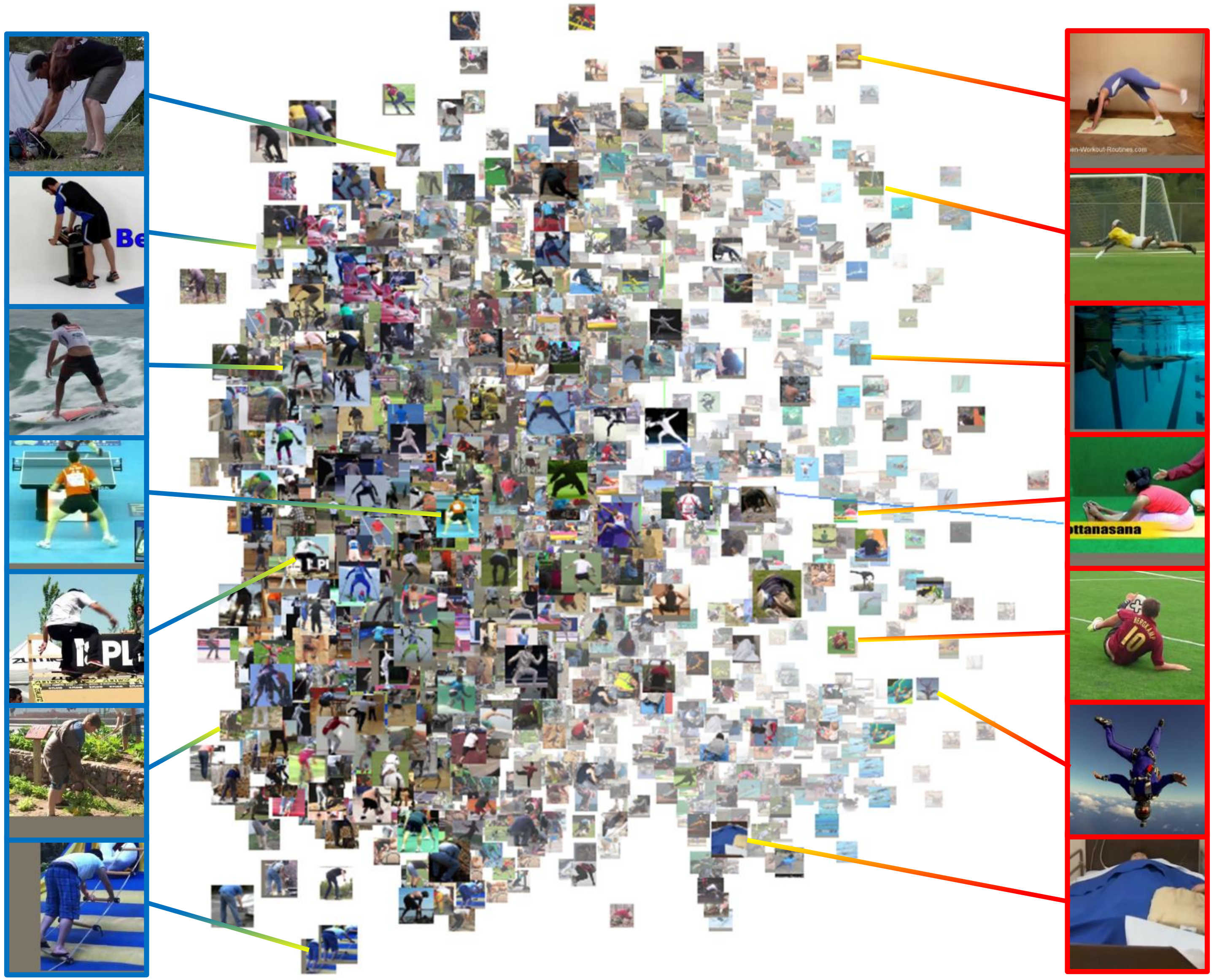
<u>Thin-slicing</u>[6]: A previous work on pose embedding

Often fails to address rare human poses.

[6] Thin-slicing for pose: Learning to understand pose without explicit pose estimation, CVPR 2016

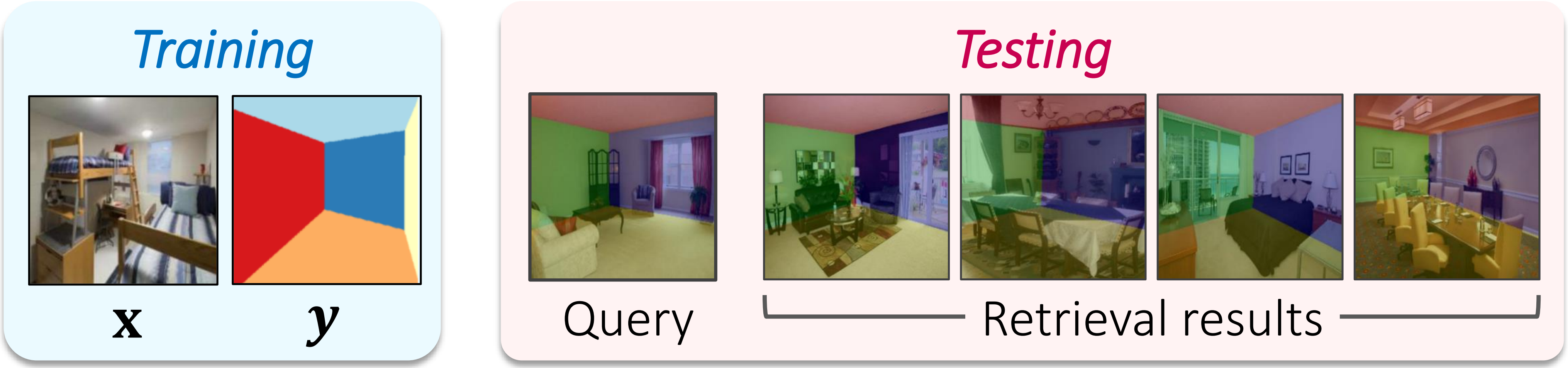
Experiments – Three Retrieval Tasks

- Human pose retrieval



Experiments – Three Retrieval Tasks

- Room layout retrieval



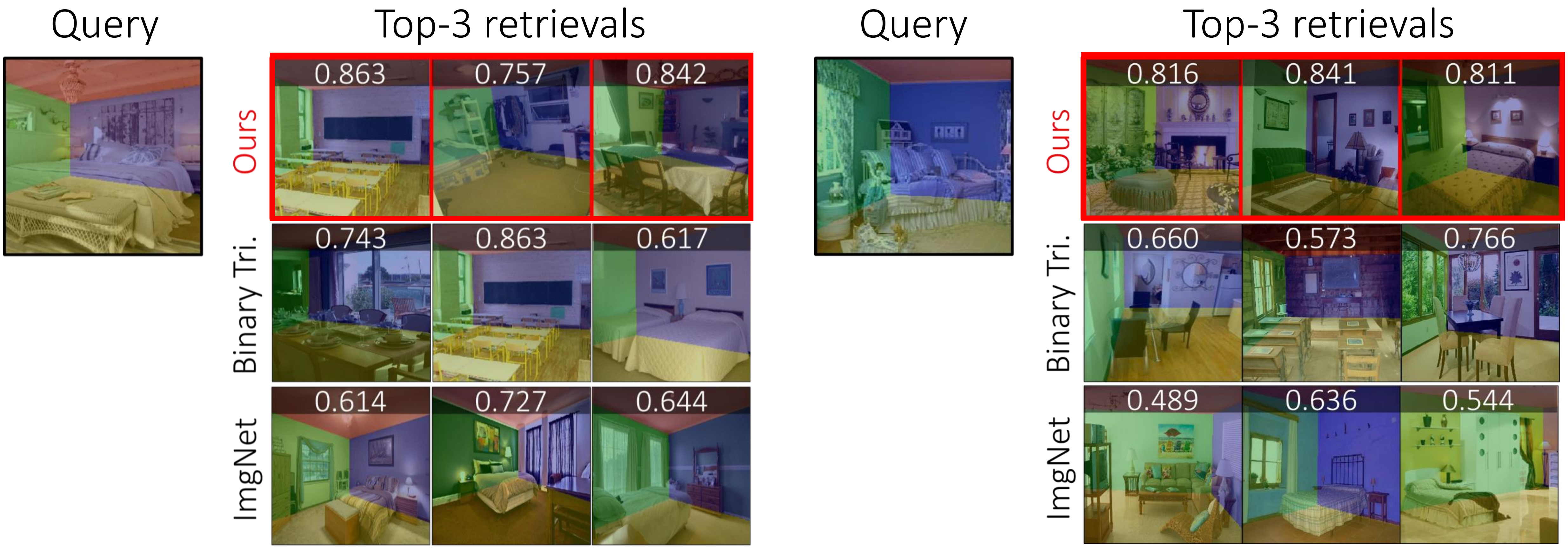
  - Conducted on the *LSUN room layout dataset*
  - Label distance between images:

$$D_{\boldsymbol{y}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = 1 - \mathrm{mIoU}(\boldsymbol{y}_i, \boldsymbol{y}_j),$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ denote groundtruth room segmentations

Experiments – Three Retrieval Tasks

- Room layout retrieval



Binary Tri.: Triplet rank loss + Binary thresholding

ImgNet: ImageNet pre-trained ResNet101

Experiments – Three Retrieval Tasks

- Caption-aware image retrieval

Training

A girl holding a frisbee

$\mathbf{x}$ $\boldsymbol{y}$

Testing

Query

Retrieval results

- Conducted on the *MS-COCO 2014 caption dataset*
- Label distance between images:

$$D_{\boldsymbol{y}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sum_{c_i \in \boldsymbol{y}_i} \min_{c_j \in \boldsymbol{y}_j} W(c_i, c_j) + \sum_{c_j \in \boldsymbol{y}_j} \min_{c_i \in \boldsymbol{y}_i} W(c_i, c_j)\,,$$

where $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are sets of 5 captions and $W(\cdot)$ is the WMD[8] between two captions

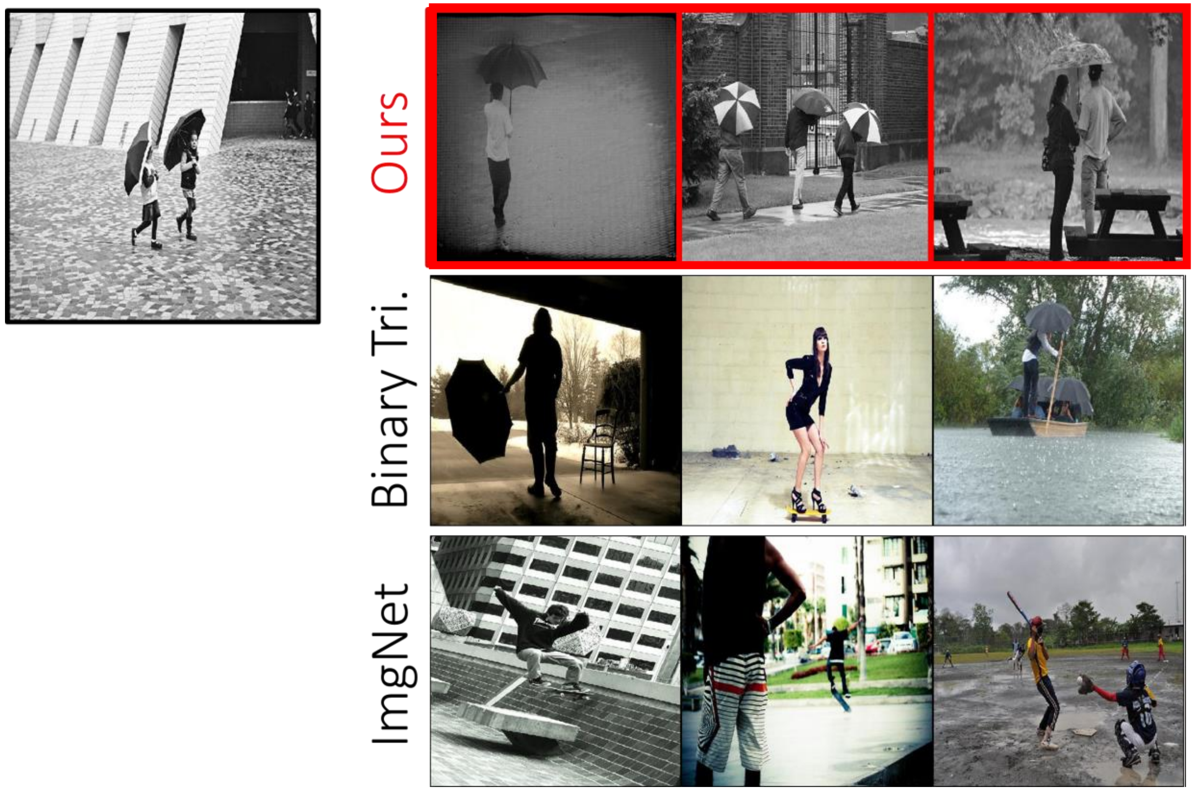
[8] From word embeddings to document distances, ICML 2015

Experiments – Three Retrieval Tasks

- Caption-aware image retrieval

Query | Top-3 retrievals

Query | Top-3 retrievals

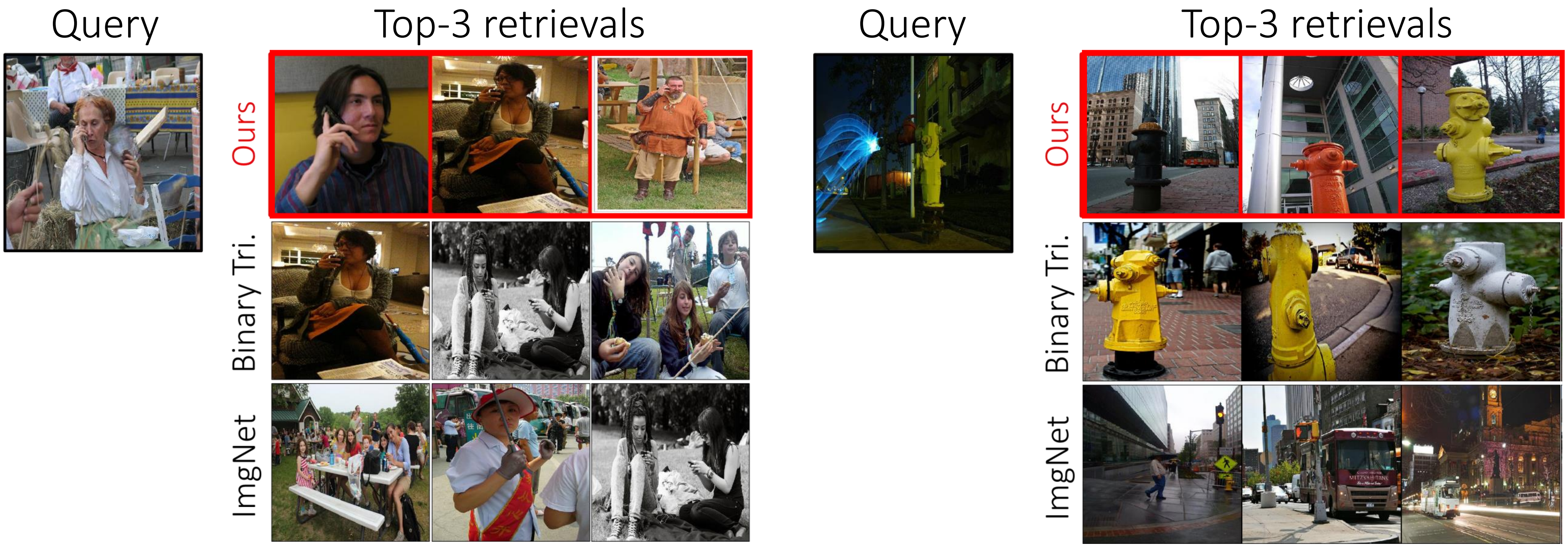


Binary Tri.: Triplet rank loss + Binary thresholding

ImgNet: ImageNet pre-trained ResNet101

Experiments – Three Retrieval Tasks

- Caption-aware image retrieval

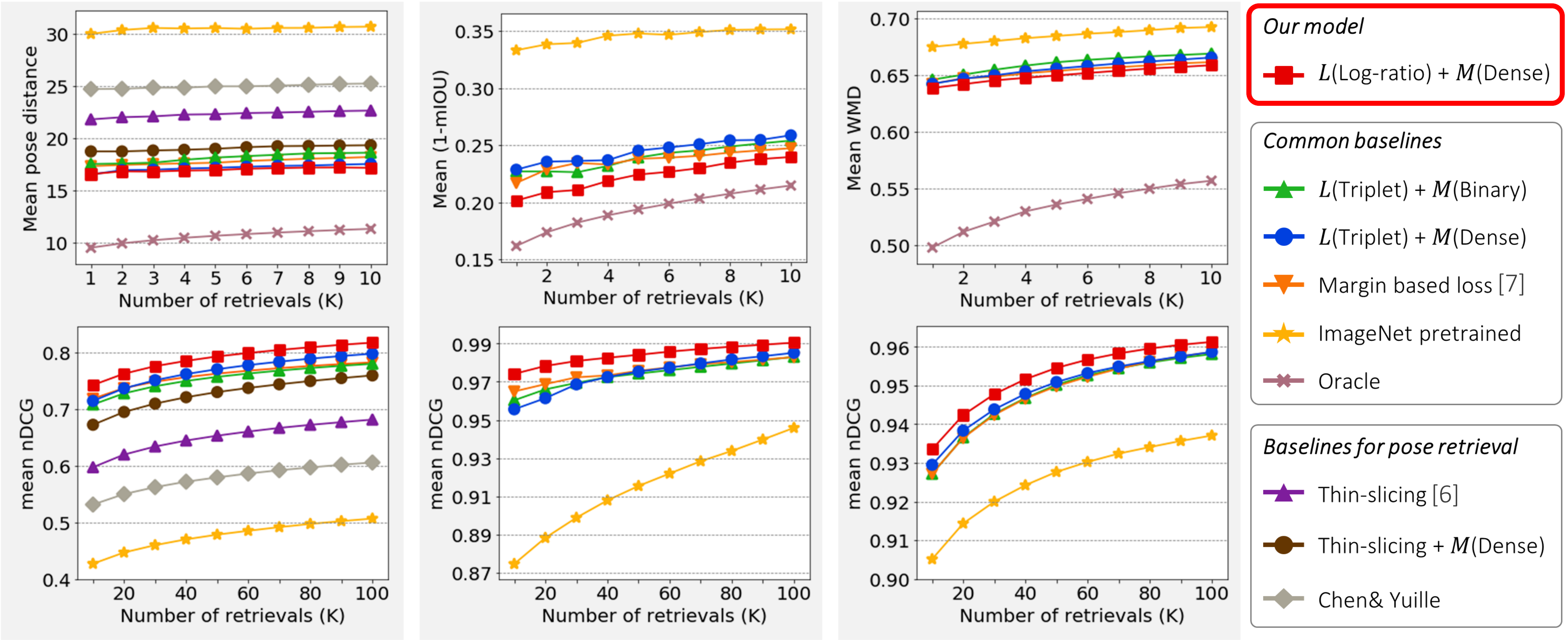


Binary Tri.: Triplet rank loss + Binary thresholding

ImgNet: ImageNet pre-trained ResNet101

Experiments – Three Retrieval Tasks

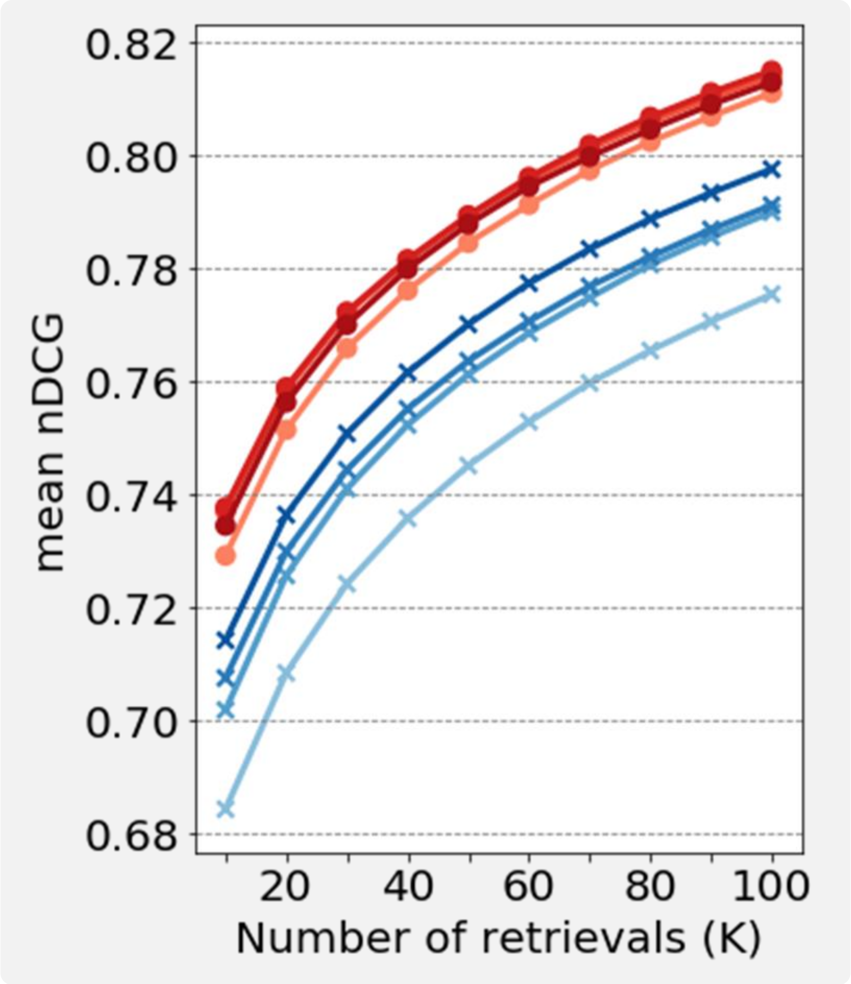
- Quantitative performance analysis

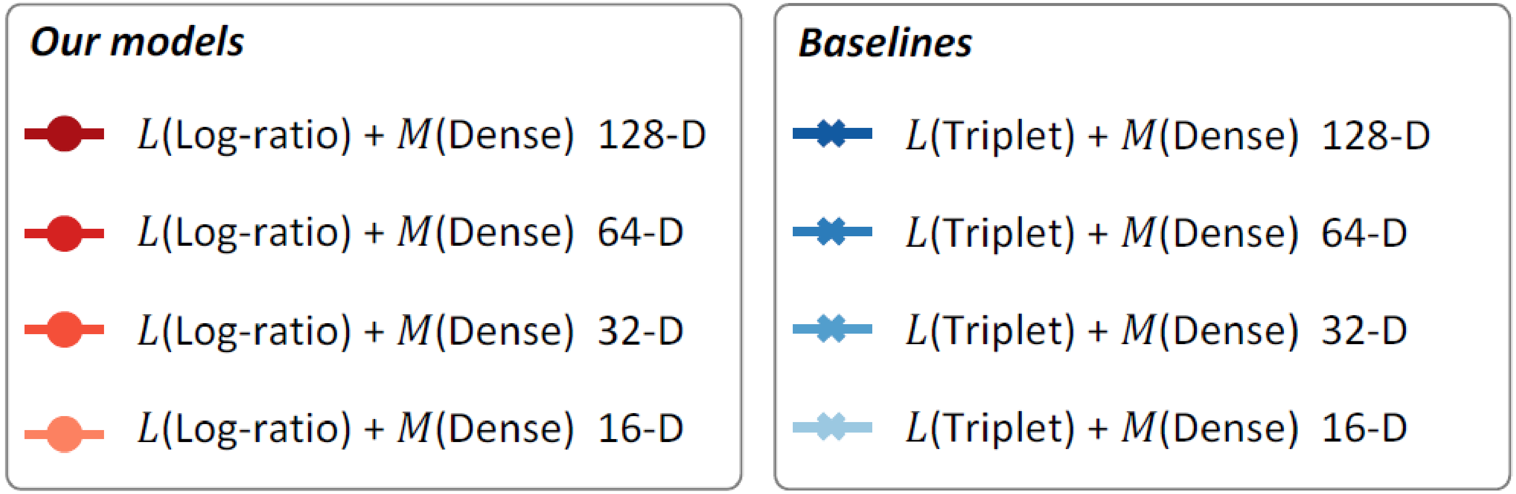


Human pose retrieval | Room layout retrieval | Caption-aware image retrieval

Experiments – Three Retrieval Tasks

- Embedding dimension vs. retrieval performance



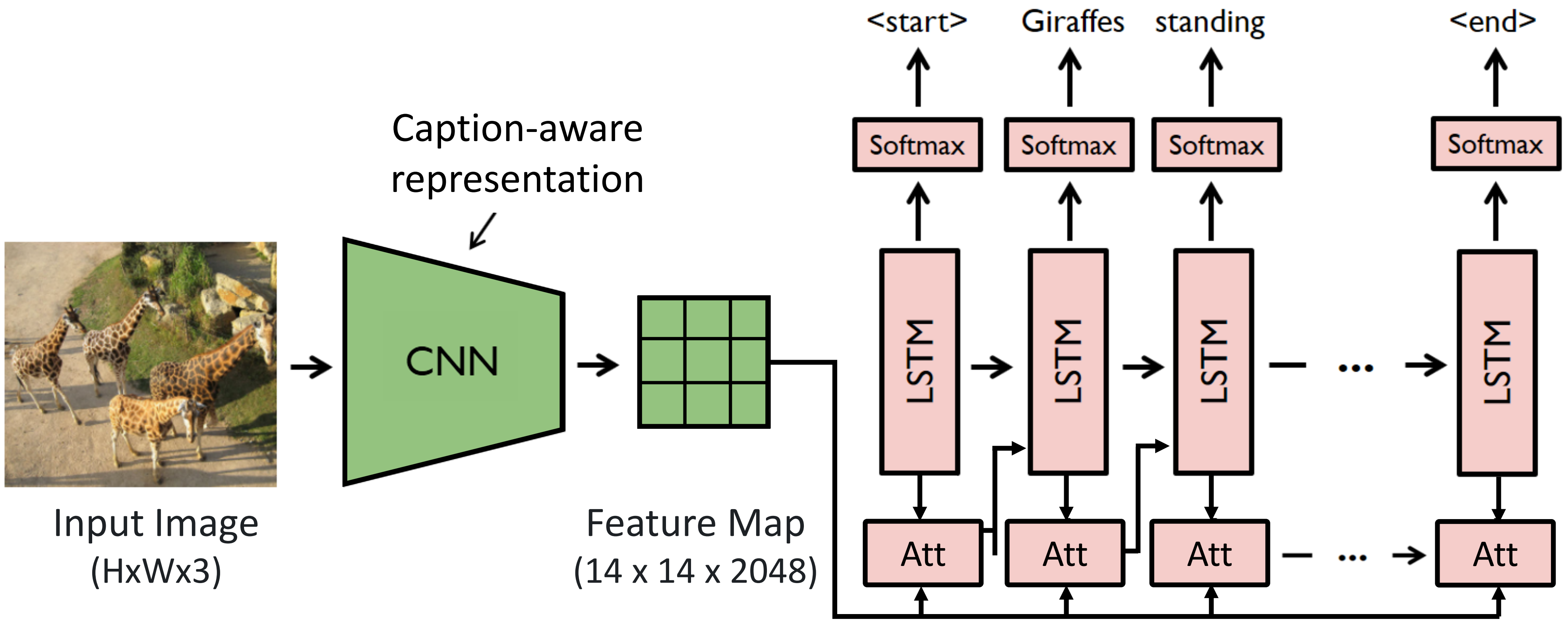


<u>L(Log-ratio) + M(Dense)</u>: Log-ratio loss + Dense triplet sampling

<u>L(Triplet) + M(Dense)</u>: Triplet rank loss + Dense triplet sampling

Experiments – Representation Learning

- Representation learning for image captioning



Our approach

Using the caption embedding network trained with caption similarities as an initial visual representation for image captioning

Experiments – Representation Learning

- Quantitative results

115.9 in CIDEr
Caption-aware feature + RL

+2.5%

113.1 in CIDEr
ImageNet pretrained feature + RL

34.65 in BLEU-4
Caption-aware feature + RL

+3.5%

33.48 in BLEU-4
ImageNet pretrained feature + RL

[9] Self-critical sequence training for image captioning, CVPR 2017
[10] Bottom-up and top-down attention for image captioning and visual question answering, CVPR 2018

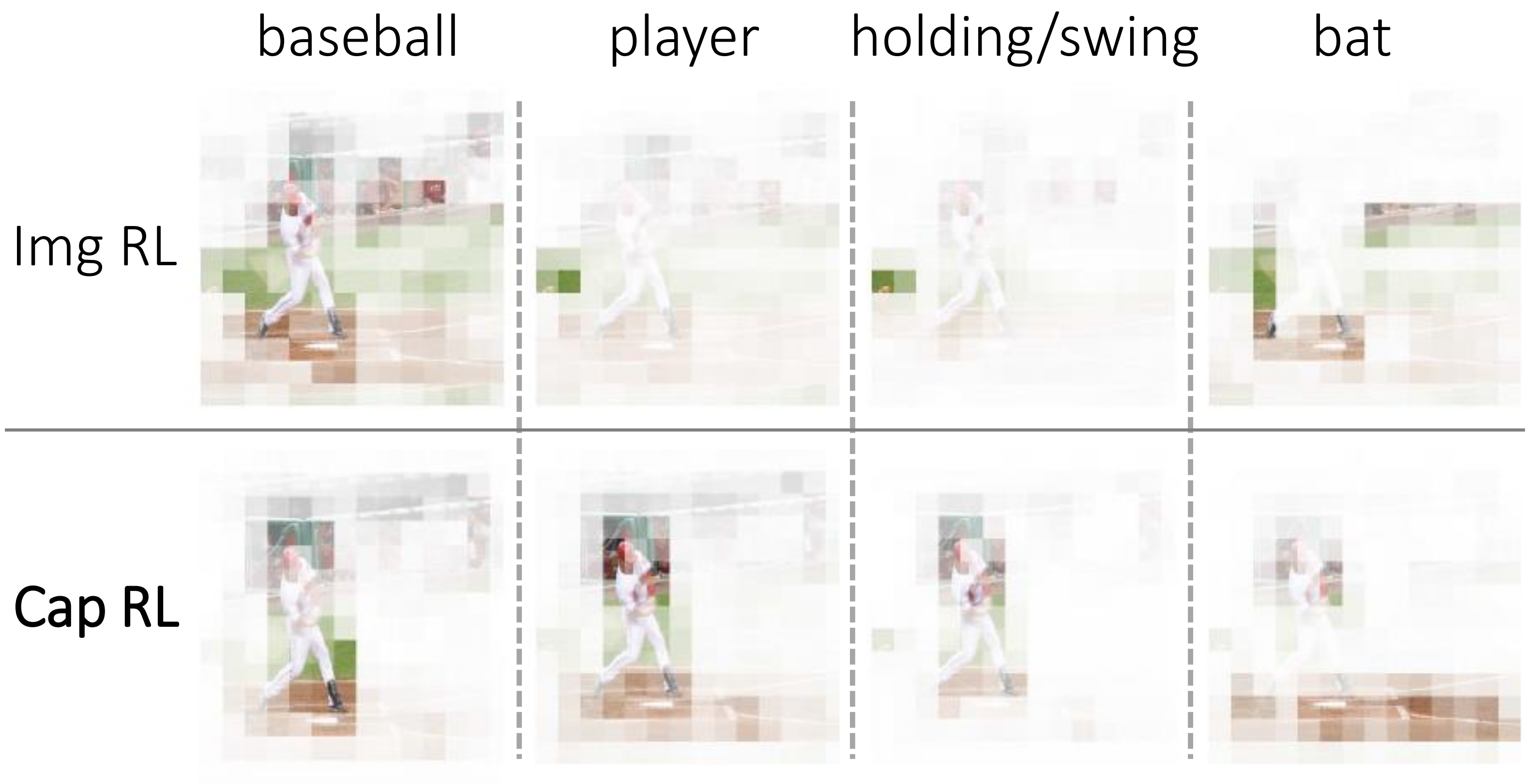
Experiments – Representation Learning

- Qualitative results obtained by the top-down attention model

GT1	There are some zebras standing in a grassy field
GT2	A field with tall grass, bushes and trees, that has zebra standing in the field
Img XE	A group of zebras grazing in a field
Cap XE	Two zebras are standing in a grassy field
Img RL	A group of zebras are grazing in a field
Cap RL	A couple of zebras and a zebra standing in a field

GT1	A baseball batter swinging a bat over home plate
GT2	A baseball player swings a bat at a game
Img XE	A baseball player holding a bat on a field
Cap XE	A baseball player swinging a bat on top of a field
Img RL	A baseball player holding a bat on a field
Cap RL	A baseball player swinging a bat at a ball

- Visualization of attentions drawn by the Att2all2 model



Img RL	A baseball player holding a bat on a field
Cap RL	A baseball player swinging a bat at a ball

Conclusion

- Summary
  - A new framework for metric learning with continuous labels
  - Various applications including visual representation learning
  - Performance boost over existing approaches

- Future directions
  - A better distance metric for continuous and structured labels
  - A hard triplet mining technique for continuous metric learning
  - More applications of semantic nearest neighbor search
  - A new benchmark for continuous metric learning

References

[1] FaceNet: A unified embedding for face recognition and clustering, CVPR 2015

[2] Beyond triplet loss: a deep quadruplet network for person re-identification, CVPR 2017

[3] Learning to compare image patches via convolutional neural networks, CVPR 2015

[4] Learning a similarity metric discriminatively with application to face verification, CVPR 2005

[5] Pose embeddings: A deep architecture for learning to match human poses, arXiv 2015

[6] Thin-slicing for pose: Learning to understand pose without explicit pose estimation, CVPR 2016

[7] Sampling matters in deep embedding learning, ICCV 2017

[8] From word embeddings to document distances, ICML 2015

[9] Self-critical sequence training for image captioning, CVPR 2017

[10] Bottom-up and top-down attention for image captioning and visual question answering, CVPR 2018