Human Understanding:
ReID, Pose, Attributes & Activity

Chi Zhang
Megvii (Face++)
zhangchi@megvii.com
Nov 2017

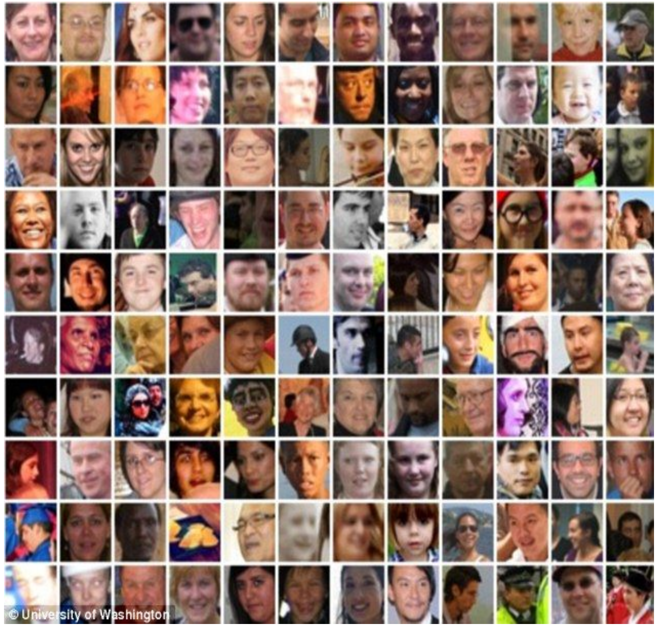
Face++ 旷视

Outline

- Person Re-Identification
  - Metric Learning
  - Mutual Learning
  - Feature Alignments
  - Re-Ranking
- Enhance ReID
  - Pose Estimation
  - Attributes
  - Tracklets

ReID: From Face to Person

- Face Recognition
  - Applications
    - 1:1 Verification
    - 1:N Identification
    - N:N Clustering
  - Limits
    - Size : 32*32
    - Horizontal : -30 ~ 30
    - Vertical : -20 ~ 20
    - Little Occlusion



ReID: From Face to Person

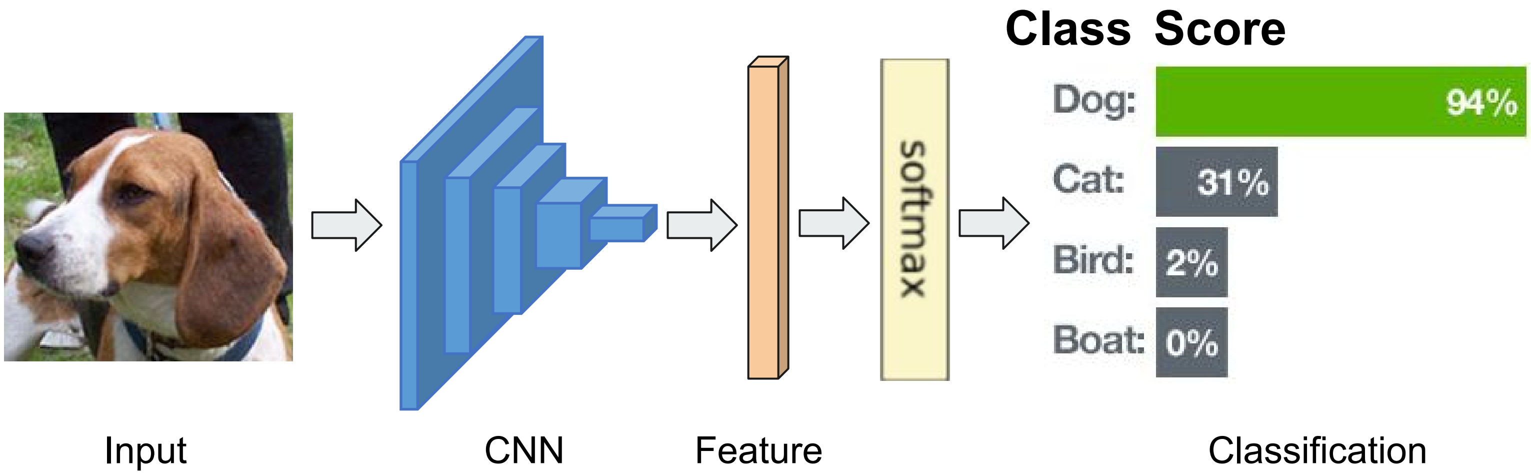
- Person Re-Identification
  - Applications
    - Tracking across cameras
    - Searching in image/video gallery
  - Challenges
    - Inaccurate detection
    - Misalignment
    - Illumination difference
    - Occlusion

Deep Metric Learning

- From Classification to Metric Learning
- Losses in Metric Learning
  - Pairwise Loss
  - Triplet Loss
    - Improved Triplet Loss
  - Quadruplet Loss
- Hard Sample Mining
  - Batched Hard Sample Mining in Triplet
  - Soft Hard Sample Mining
  - Margin Sample Mining

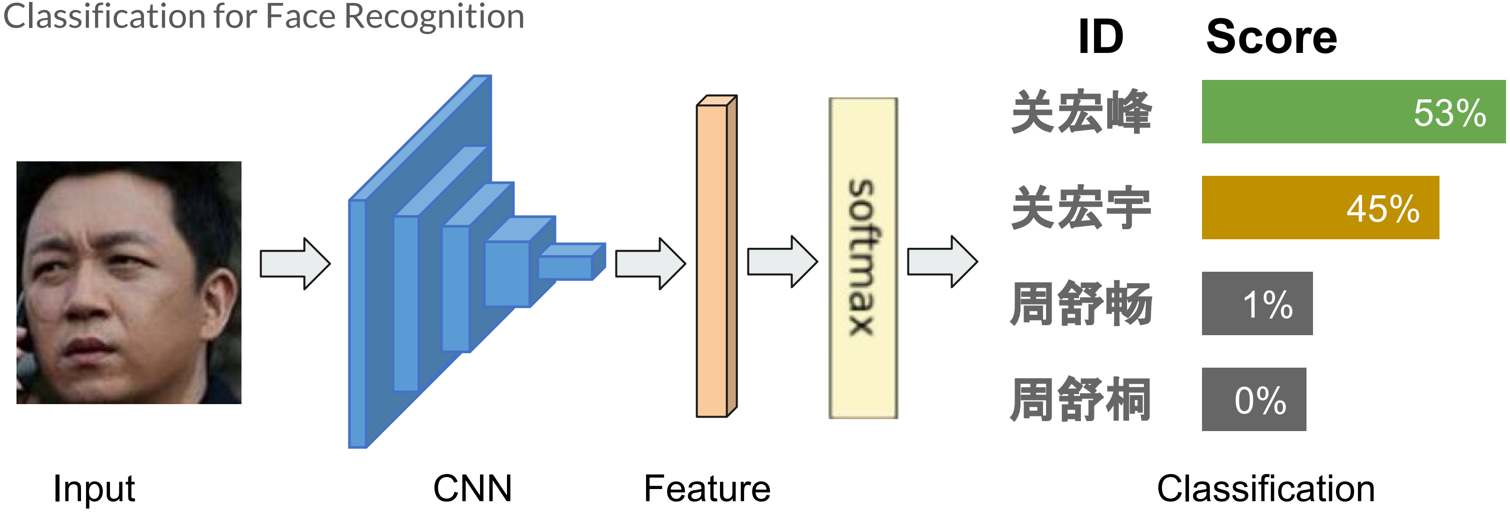
From Classification to Metric Learning

- General Classification in Deep Learning



From Classification to Metric Learning

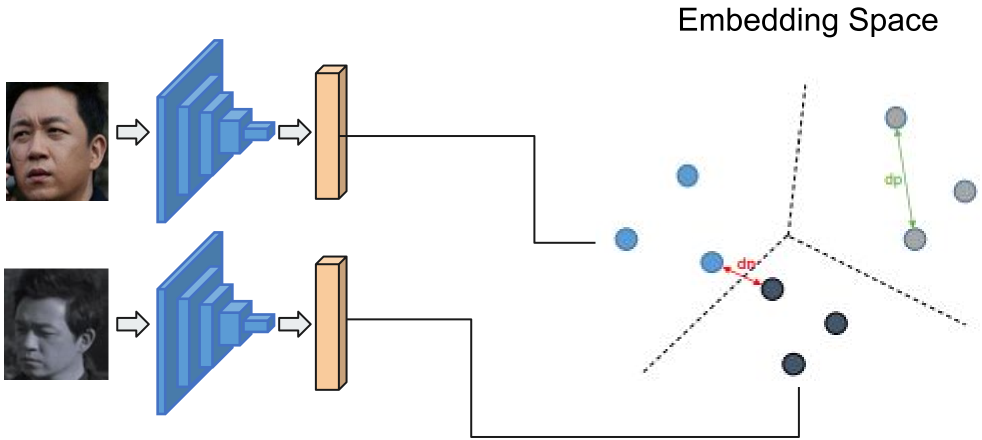
- Classification for Face Recognition



Face++ 旷视

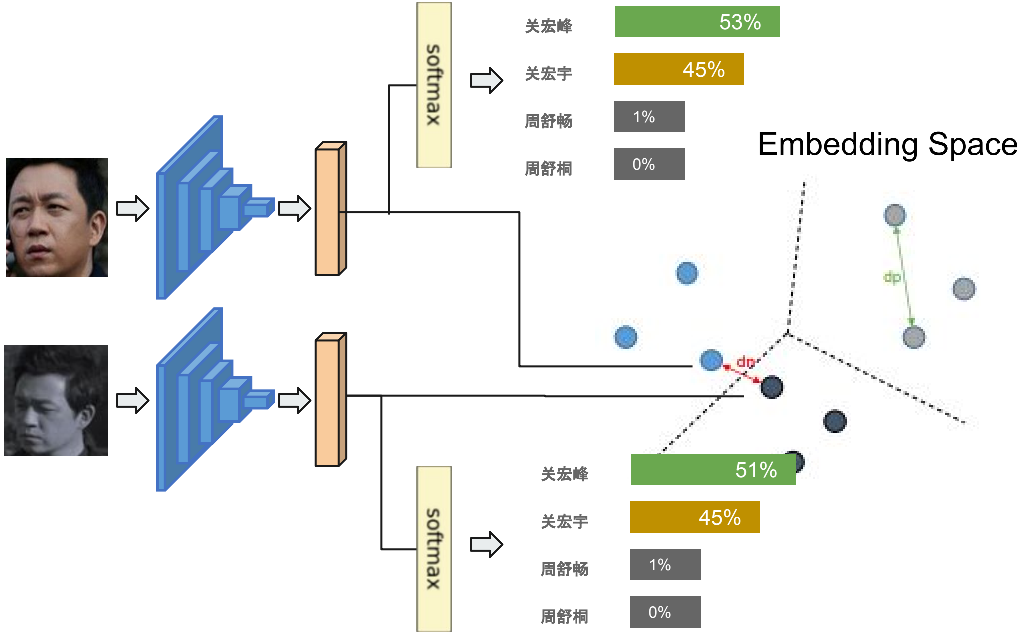
From Classification to Metric Learning

- Feature Learned in Classification
- Similarity is decided by the distance of features
  - Usually L2 distance



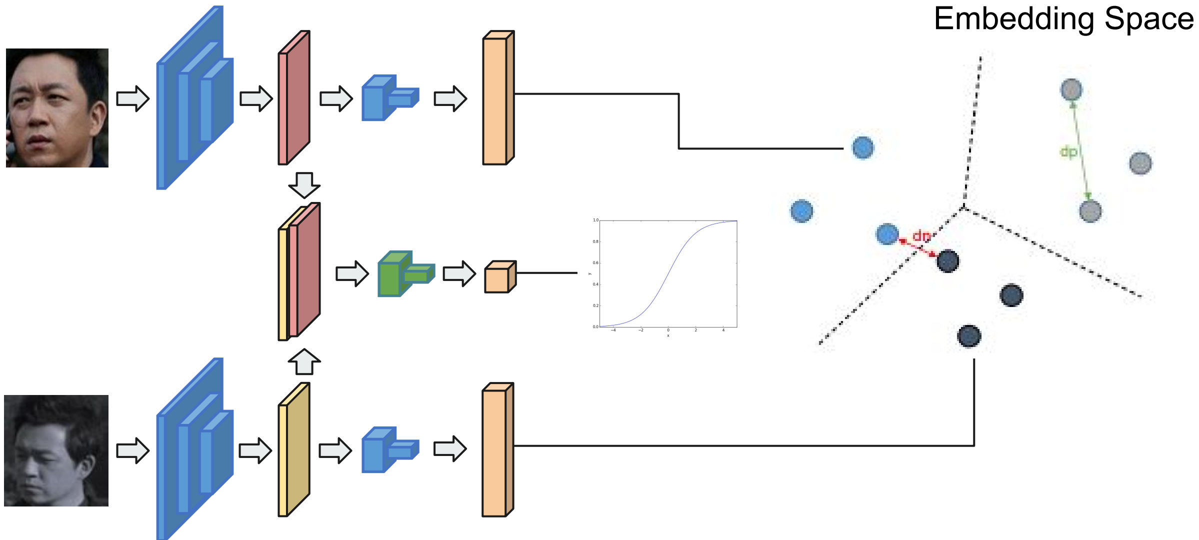
From Classification to Metric Learning

- Feature Learned with Metric Learning Loss
  - Pre-trained in Classification
  - Trained together with classification



From Classification to Metric Learning

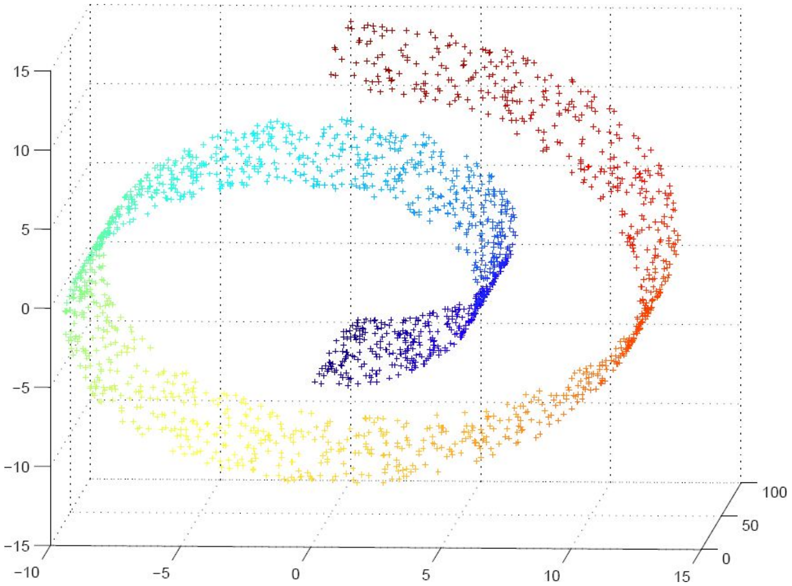
- Fusing intermediate features
  - Not practical in applications: needs to store (large) feature maps



Face++ 旷视

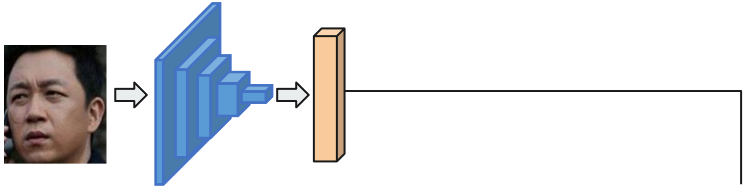
Metric Learning

- Goal
  - Learn a function that measures how similar two objects are.
  - Compared to classification which works in a closed-word, metric learning deals with an open-world.
- Applications
  - Ranking
  - Identification
  - Verification
  - Clustering

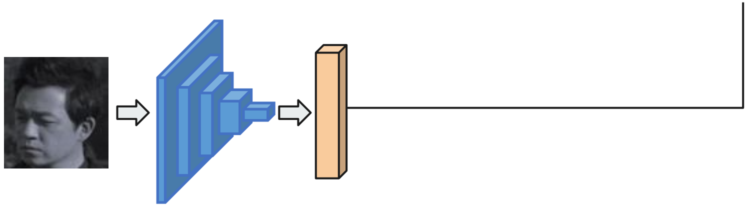


Metric Learning: Contrastive Loss

- δ is Kronecker Delta
- α is the margin for different identities



$$L_{pairwise} = \delta(I_A, I_B) \cdot ||f_A - f_B||_2 + (1 - \delta(I_A, I_B))(\alpha - ||f_A - f_B||_2)_+$$

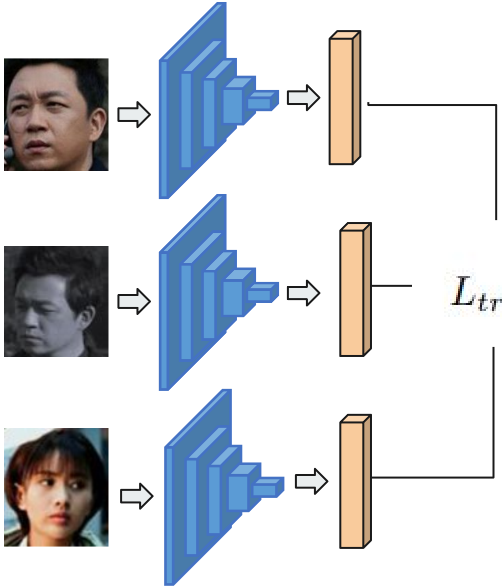


Metric Learning: Contrastive Loss

- The distance of images with the same identity should be smaller
- The distance of images with different identities should be larger

R. R. Varior et al., Gated siamese convolutional neural network architecture for human re-identification. ECCV. 2016

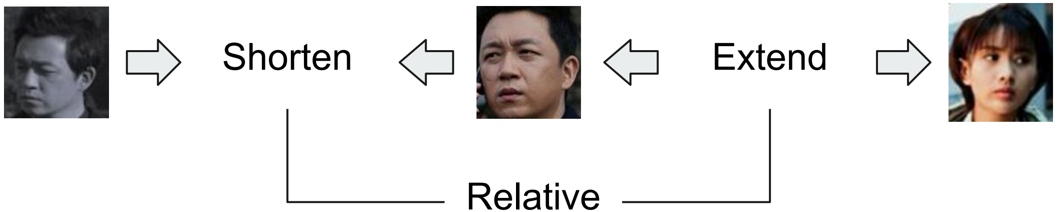
Metric Learning: Triplet Loss



$$L_{trp} = \frac{1}{N}\sum^{N}\left(\|f_A - f_{A'}\|_2 - \|f_A - f_B\|_2 + \alpha\right)_+$$

Metric Learning: Triplet Loss

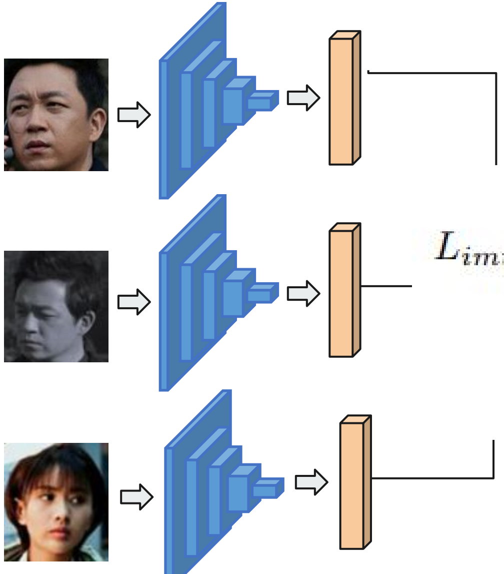
- A batch of triplets (A, A’, B) are trained in each iteration
  - A and A’ share the same identity
  - B has a different identity
- The distance of A and A’ should be smaller than that of A and B



H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing, 2017

Metric Learning: Improved Triplet Loss

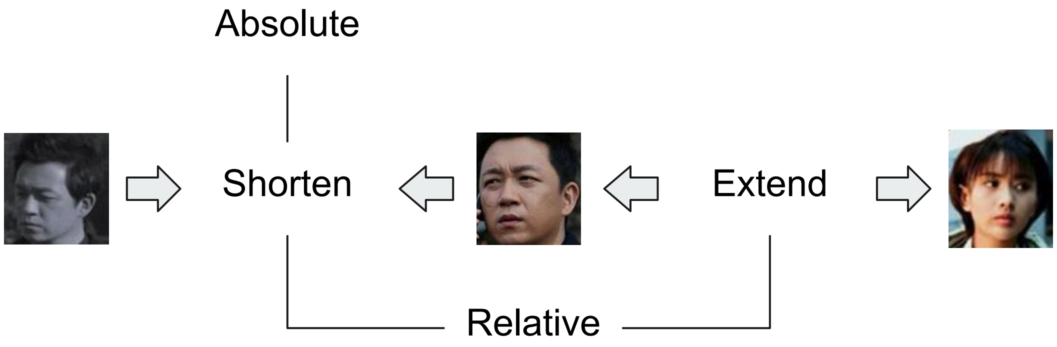
- β-term penalizes distance between features of A and A'



$$L_{imtrp} = \frac{1}{N}\sum^{N}(\|f_A - f_{A'}\|_2 - \|f_A - f_B\|_2 + \alpha)_+ + \frac{1}{N}\sum^{N}(\|f_A - f_{A'}\|_2 - \beta)_+$$

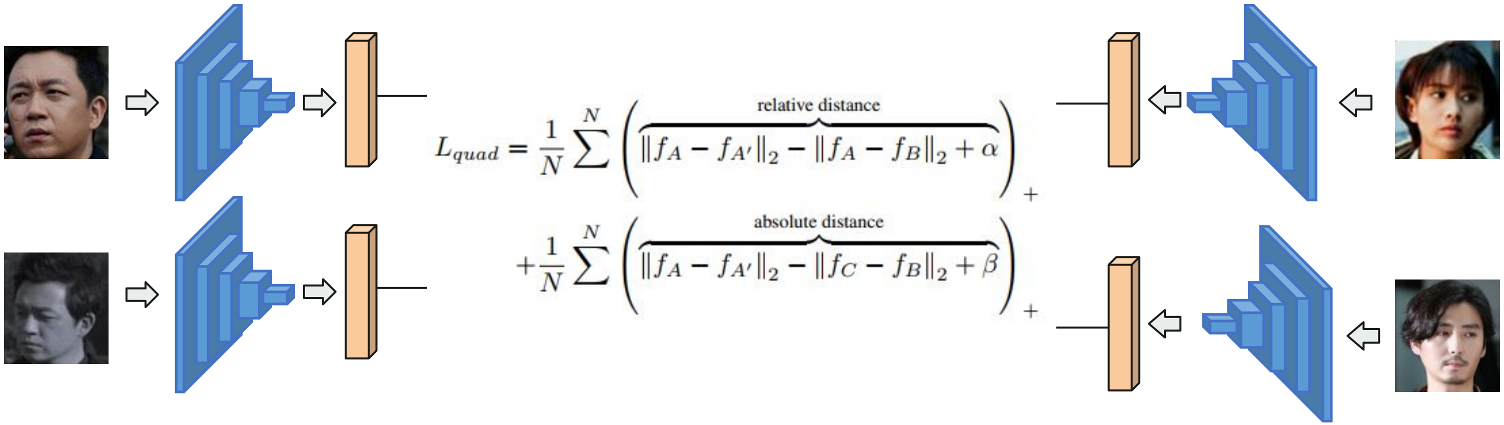
Metric Learning: Improved Triplet Loss

- Triplet Loss with Contrastive Loss
- Only consider image pairs with the same identity



D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. CVPR2016

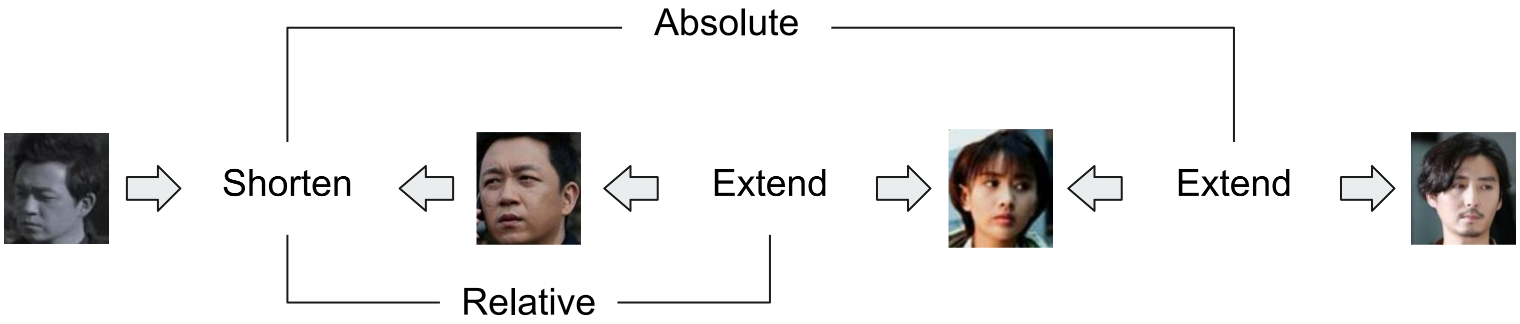
Metric Learning: Quadruplet Loss



$$L_{quad} = \frac{1}{N}\sum^{N}\Big(\overbrace{\|f_A - f_{A'}\|_2 - \|f_A - f_B\|_2 + \alpha}^{\text{relative distance}}\Big)_+ + \frac{1}{N}\sum^{N}\Big(\overbrace{\|f_A - f_{A'}\|_2 - \|f_C - f_B\|_2 + \beta}^{\text{absolute distance}}\Big)_+$$

Metric Learning: Quadruplet Loss

- Triplet Loss & Pairwise Loss
- Distance between any identical images should be smaller than that between different images



W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. arXiv preprint arXiv:1704.01719, 2017.

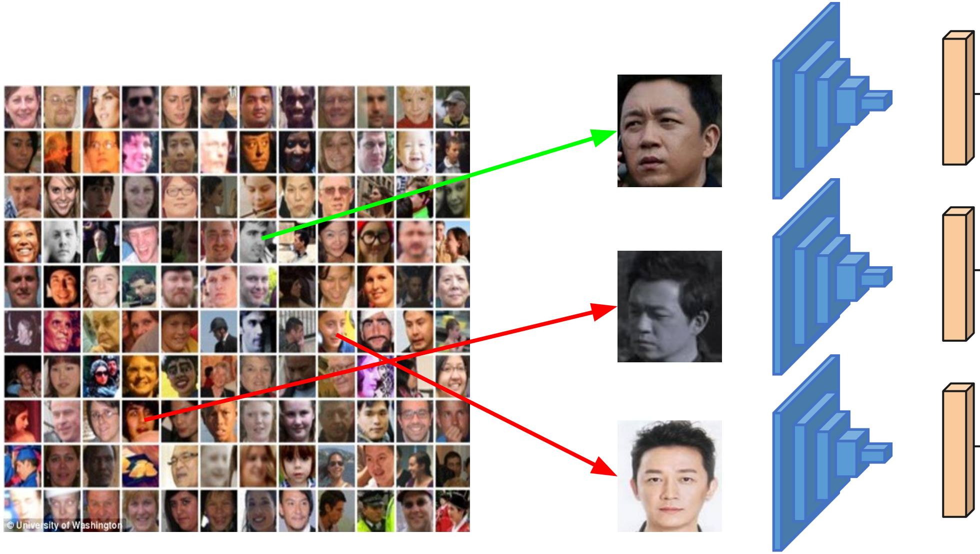
Hard Sample Mining

- The possible number of triplets grows cubically
- Trivial triplets quickly become uninformative
- The fraction of trivial triplets are large

Trivial:

Non-Trivial:

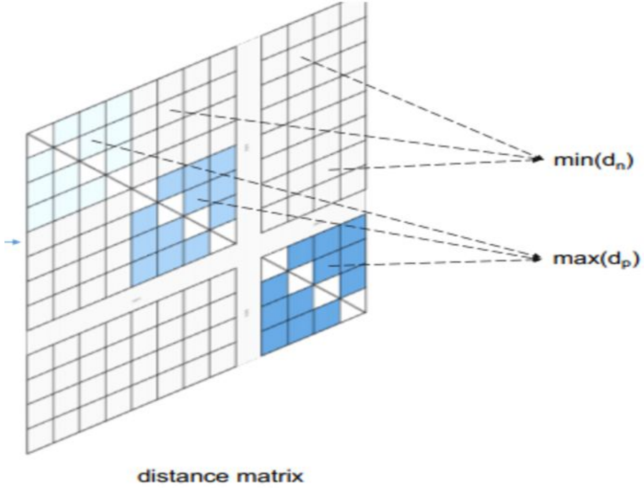
Hard Sample Mining: Triplet Hard Loss



$$L_{trihard} = \frac{1}{N} \sum_{A \in batch} \left(\overbrace{\max_{A'}(\| f_A - f_{A'} \|_2)}^{\text{hard positive pair}} - \overbrace{\min_{B}(\| f_A - f_B \|_2)}^{\text{hard negative pair}} + \alpha \right)_+$$

Hard Sample Mining: Triplet Hard Loss

- Each batch contains K identities, each identities contains L images
- Compute the distance between each images in the batch
- Distance matrix
  - Diagonal Blocks are distance between images with the same identity
  - Others are distance between images with different identities

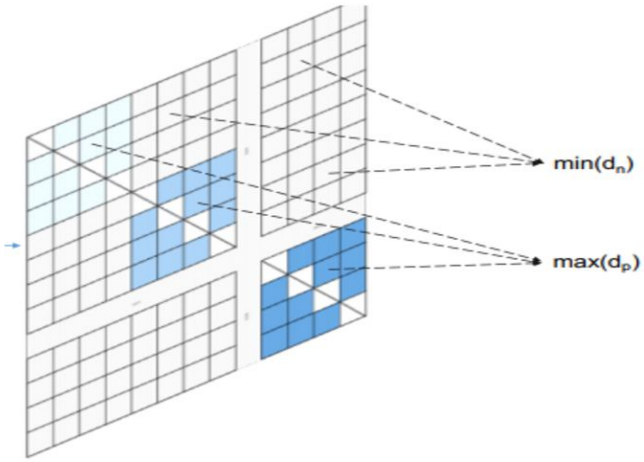


A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017

Face++ 旷视

Hard Sample Mining: Triplet Hard Loss

- Generate a triplet from **each line** in the matrix
  - Each image in the batch
- The largest distance in the diagonal block
  - The most unsimilar image with the same identity
- The smallest distance in other places
  - The most similar image with a different identity

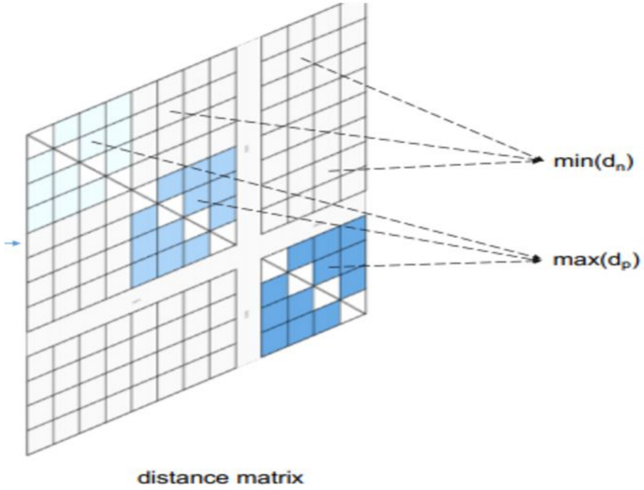


distance matrix

Face++ 旷视

Hard Sample Mining: Soft Triplet Hard Loss

- Generate a triplet from **each line** in the matrix
  - Each image in the batch
- The weighted average distance in the diagonal block
  - Softmax(d_ij)
- The weighted average distance in the diagonal block
  - Softmax(-d_ik)



Hard Sample Mining: Lifted Structured Embedding

- Lifted Structured Embedding
  - Only mining hard negative samples
  - The negative samples are weighted by the margin constraint violation

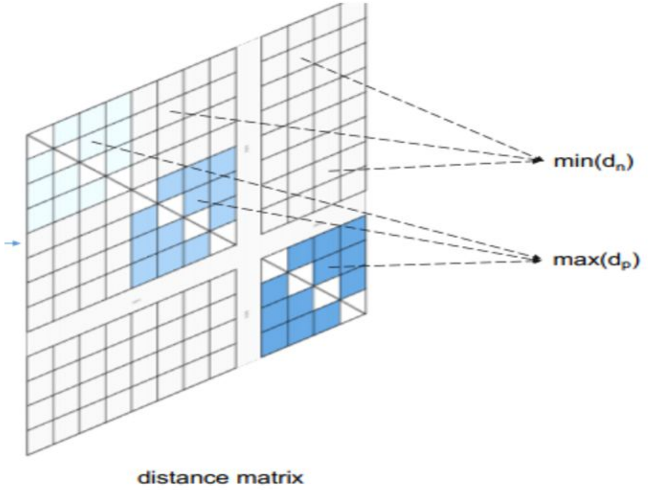
$$\ell(X, \mathbf{y}) = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[\log \left(\sum_{(i,k)\in\mathcal{N}} \exp\{\alpha - D_{i,k}\} + \sum_{(j,l)\in\mathcal{N}} \exp\{\alpha - D_{j,l}\} \right) + D_{i,j} \right]_+^2$$

H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In CVPR2016.

Face++ 旷视

Hard Sample Mining

- Margin Sample Mining
    - Generate only one triplet from **each batch**
    - The largest distance in the diagonal block
        - The most unsimilar image pair with the same identity in the batch
    - The smallest distance in other places
        - The most similar image pair with different identities in the batch



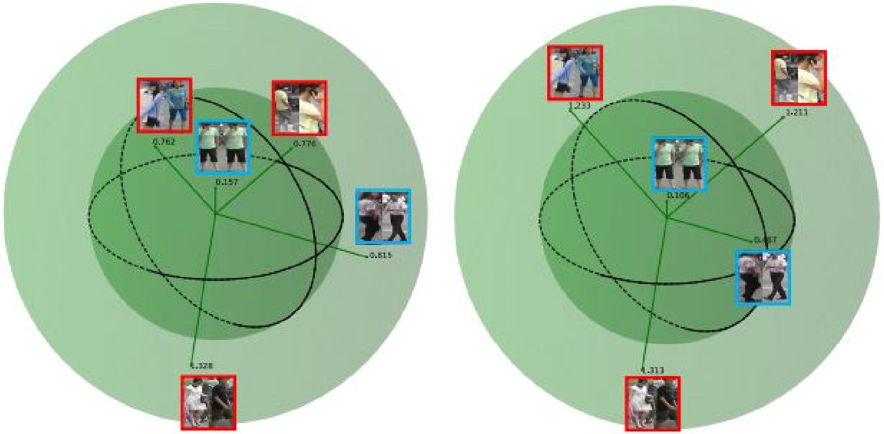
Face++ 旷视

[Q. Xiao, H. Luo, C. Zhang, Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification, arXiv: 1710.00478]

Hard Sample Mining

- Margin Sample Mining

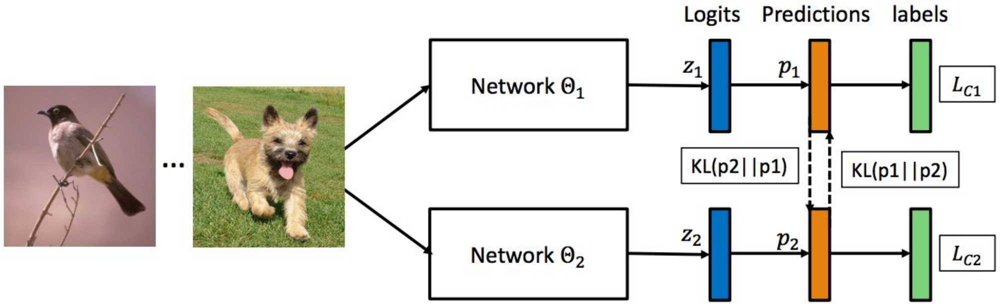
$$L_{eml} = \Big(\overbrace{\max_{A,A'}(\|f_A - f_{A'}\|_2)}^{\text{hardest positive pair}} - \overbrace{\min_{C,B}(\|f_C - f_B\|_2)}^{\text{hardest negative pair}} + \alpha \Big)_+$$



(a) TriHard (b) MSML

Mutual Learning

- Knowledge Distill
  - A smaller, faster student model learn from a powerful teacher model
- Mutual Learning
  - A set of student models learn from each other



Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. arXiv preprint arXiv:1706.00384, 2017

Mutual Learning

- Mutual Learning in Classification

$$D_{KL}(\boldsymbol{p}_2 \| \boldsymbol{p}_1) = \sum_{i=1}^{N} \sum_{m=1}^{M} p_2^m(\boldsymbol{x}_i) \log \frac{p_2^m(\boldsymbol{x}_i)}{p_1^m(\boldsymbol{x}_i)}$$

- Mutual Learning in Ranking

$$P(\pi|\mathbf{X}) = \prod_{i=1}^{n} \frac{\exp[S(\mathbf{x}_{\pi(i)})]}{\sum_{k=i}^{n} \exp[S(\mathbf{x}_{\pi(i)})]}$$

[Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. arXiv preprint arXiv:1707.01220, 2017.]

Face++ 旷视

Mutual Learning in Metric Learning

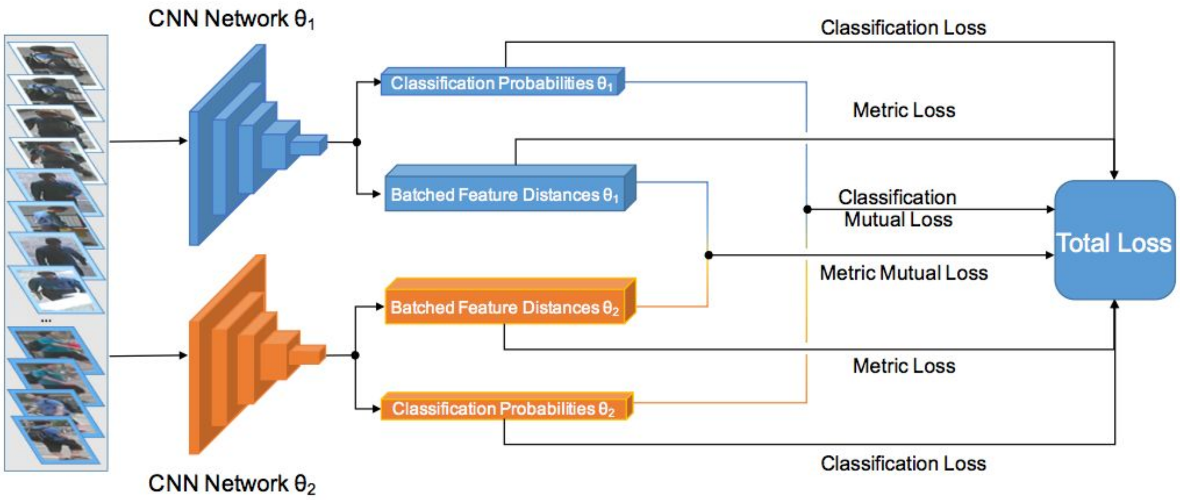
- Batched Distance Matrix
- Metric Mutual Learning

$$L_M = \frac{1}{N^2} \sum_i^N \sum_j^N \Big([ZG(M_{ij}^{\theta_1}) - M_{ij}^{\theta_2}]^2 + [M_{ij}^{\theta_1} - ZG(M_{ij}^{\theta_2})]^2 \Big)$$

X. Zhang et al, AlignedReID: Surpassing Human-Level Performance in Person Re-Identification, arXiv: 1711.08184

Face++ 旷视

Mutual Learning in Metric Learning



Face++ 旷视

Difficulties in Person Re-Identification

- Different Directions

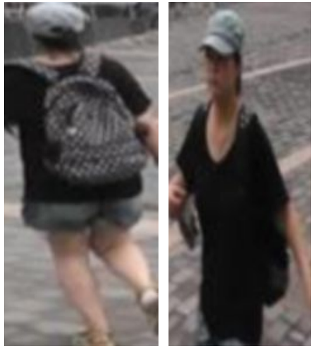
- Non-rigid Body Deformation

- Different Illumination

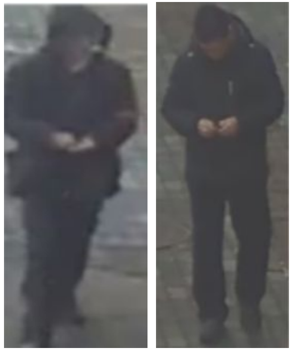
Difficulties in Person Re-Identification

- Occlusion

- Incomplete



- Similar Appearance



ReID Evaluation Criteria

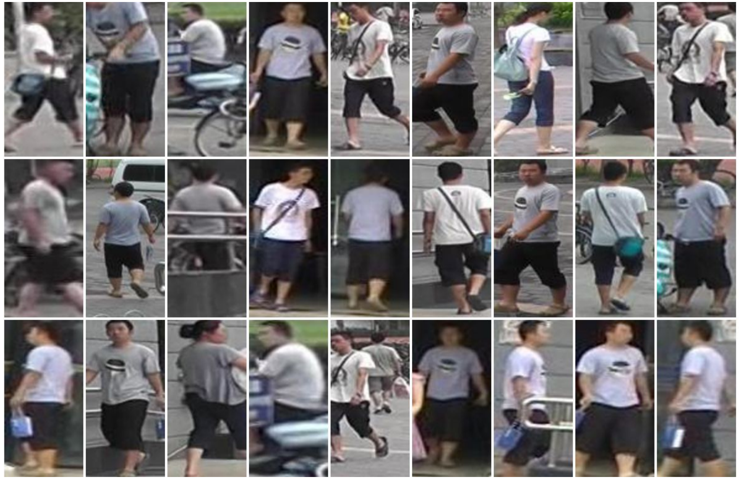
- CMC (cumulative match characteristic)
    - Rank-1, Rank-5, Rank-10
- mAP
    - Precision : fraction of ground truths in the results
    - AP: average of precision in top-k results, where the k-th is a ground truth

$$AvgP = \frac{1}{N_{Rel}} \sum_{d_i \in Rel} \frac{i}{Rank(d_i)}$$

    - mAP: average of AP for all queries

Re-Identification Datasets

- Marke1501
  - 1501 persons, 32643 bounding boxes
  - 6 cameras in Tsinghua



Methods	mAP	r=1
Temporal [23]	22.3	47.9
Learning [47]	35.7	61.0
Gated [32]	39.6	65.9
Person [5]	45.5	71.8
Re-ranking [57]	63.6	77.1
Pose [52]	56.0	79.3
Scalable [1]	68.8	82.2
Improving [16]	64.7	84.3
In [13]	69.1	84.9
In (RK)[13]	**81.1**	86.7
Spindle[50]	-	76.9
Deep[49]*	68.8	87.7
DarkRank[4]*	74.3	89.8
GLAD[37]*	73.9	**89.9**
HydraPlus-Net[20]*	-	76.9
AlignedReID	82.3	92.6
AlignedReID (RK)	**91.2**	**94.0**

Face++ 旷视

Re-Identification Datasets

- CUHK03
    - 1360 persons, 13164 bounding boxes
    - 2 cameras in CUHK

Methods	r=1	r=5	r=10
Person [15]	44.6	-	-
Learning [47]	62.6	90.0	94.8
Gated [32]	61.8	-	-
A [34]	57.3	80.1	88.3
Re-ranking [57]	64.0	-	-
In [13]	75.5	95.2	99.2
Joint [42]	77.5	-	-
Deep [10]*	84.1	-	-
Looking [2]*	72.4	95.2	95.8
Unlabeled [56]	84.6	97.6	98.9
A [55]*	83.4	97.1	98.7
Spindle[50]	88.5	97.8	98.6
DarkRank[4]*	89.7	**98.4**	**99.2**
GLAD[37]*	85.0	97.9	99.1
HydraPlus-Net[20]*	**91.8**	**98.4**	99.1
AlignedReID	91.9	98.7	99.4
AlignedReID (RK)	**96.1**	**99.5**	**99.6**

Face++ 旷视

Re-Identification Datasets

- DukeMTMC-reid
  - 702 persons, 16522 bounding boxes
  - 8 cameras in Duke

Method	Rank-1	mAP
BOW+kissme [38]	25.13	12.17
LOMO+XQDA [18]	30.75	17.04
IDE [39]	65.22	44.99
GAN [40]	67.68	47.13
OIM [29]	68.1	47.4
TriNet [10]	72.44	53.50
ACRN [20]	72.58	51.96
SVDNet [24]	76.7	56.8
SVDNet+Ours	**79.31**	**62.44**
SVDNet+Ours+re [41]	**84.02**	**78.28**

Re-Identification Dataset

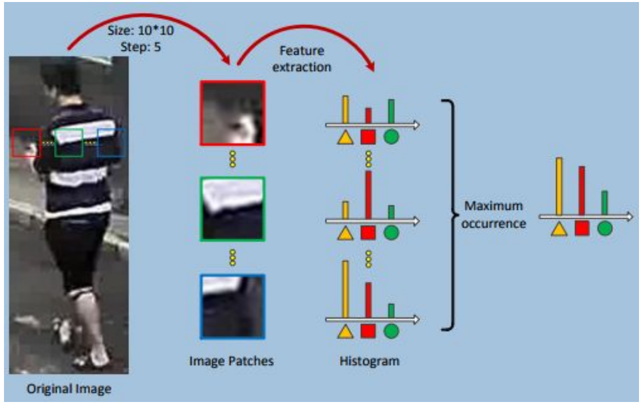
- MARS
  - 1261 persons, 20478 tracklets
  - 6 cameras in Tsinghua

Methods	mAP	r=1
Re-ranking [57]	68.5	73.9
Learning [48]*	-	55.5
Multi [31]*	-	68.2
MARS [30]	49.3	68.3
In [13]	67.7	79.8
In (RK)[13]	**77.4**	**81.2**
Quality [21]*	51.7	73.7
See [58]	50.7	70.6
AlignedReID	79.1	86.8
AlignedReID (RK)	**85.6**	**87.5**

Face++ 旷视

Traditional Methods

- Colors
  - HSV after Retinex Algorithm
- Texture
  - Scale Invariant Local Ternary Pattern (SILTP)
- Image Representation
  - Local Maximal Occurrence Feature (LOMO)
- Methods
  - KISSME
  - Linear Discriminant Analysis (LDA)
  - Cross-view Quadratic Discriminant Analysis (XQDA)

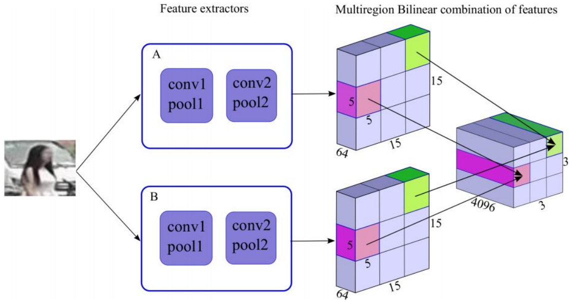


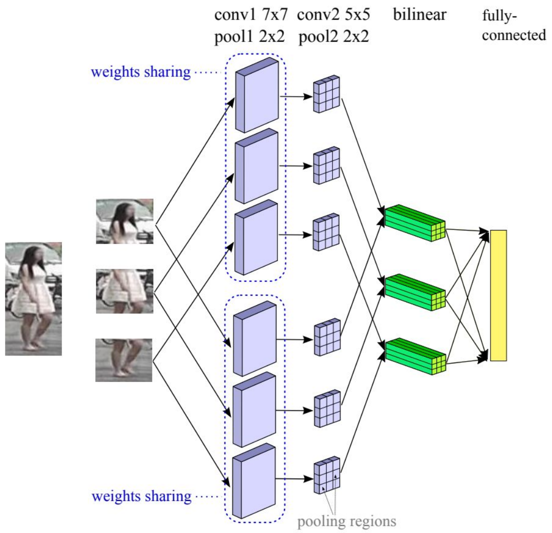
Feature Alignments in Deep Re-Identification

- Motivations
  - Person is highly structured
  - Local similarity plays a key role to decide the identity
- Methods
  - Local Features from local regions
  - Local Feature Alignment
    - Fusion by LSTM
    - Alignment in PL-Net
    - Alignment in AlignedReID

Local Features from Local Regions

- Extract features in multiple regions
    - Without Alignment

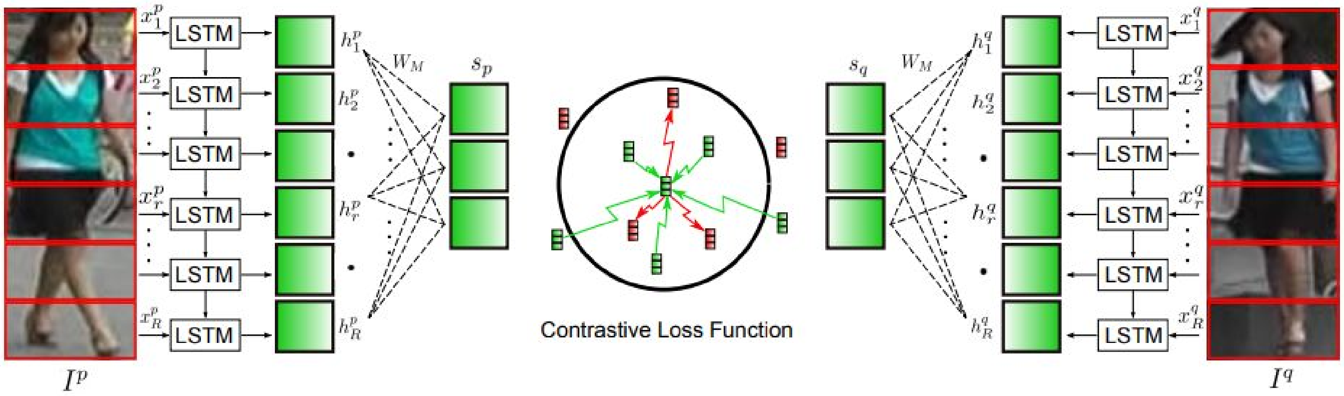




[Multi-region Bilinear Convolutional Neural Networks for Person Re-Identification]()

Local Feature Alignment

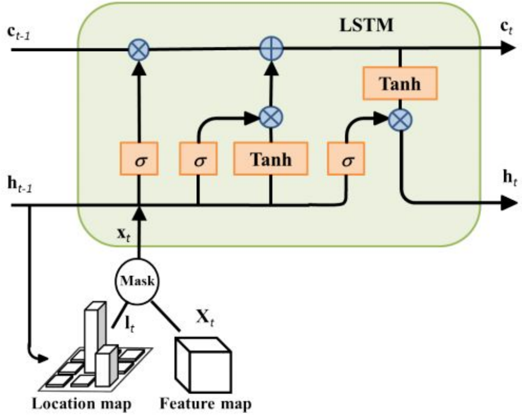
- Fusion by LSTM (Long Short-Term Memory) RNN

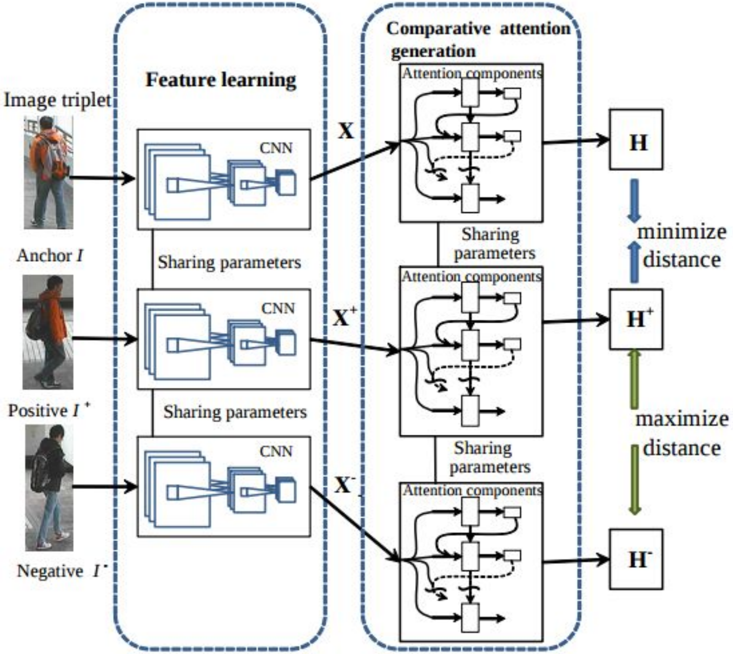


[R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human reidentification. In European Conference on Computer Vision, pages 135–153. Springer, 2016]

Fusion Local Feature by LSTM

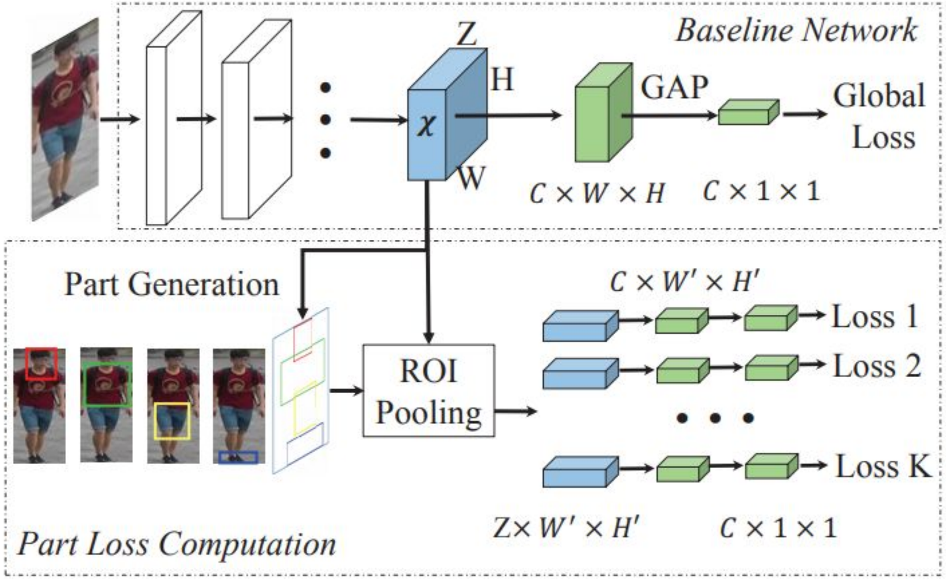
- LSTM with mask
  - Hard Local Mask
  - Soft Attention Mask





Local Feature Alignment

- Alignment in PL-Net (Part Loss Network)
  - Unsupervised “detect” human body parts
  - Extract local features by ROI Pooling
  - Concatenate global feature and local features

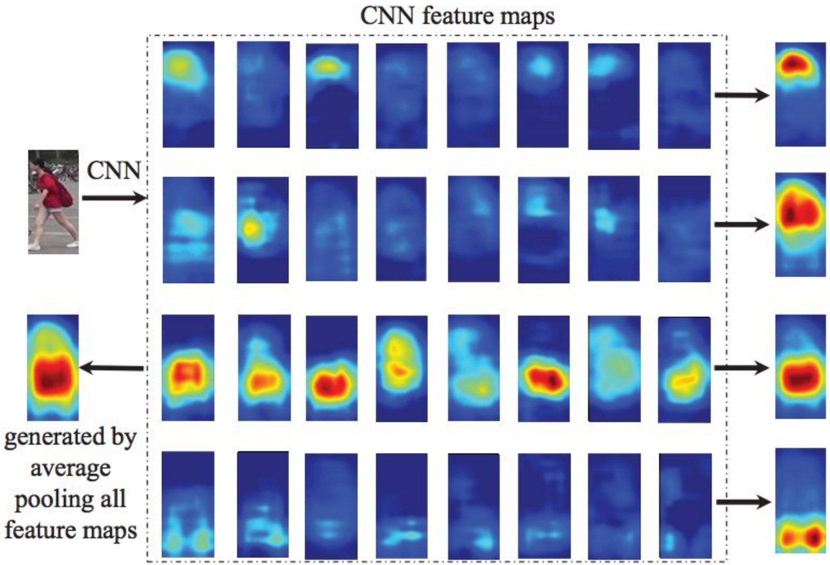


Alignment in PL-Net

- Unsupervised Part Detection
    - Compute maximum activation position on each feature map

$$(h_z, w_z) = \arg\max_{h,w} \mathcal{X}_z(h, w)$$

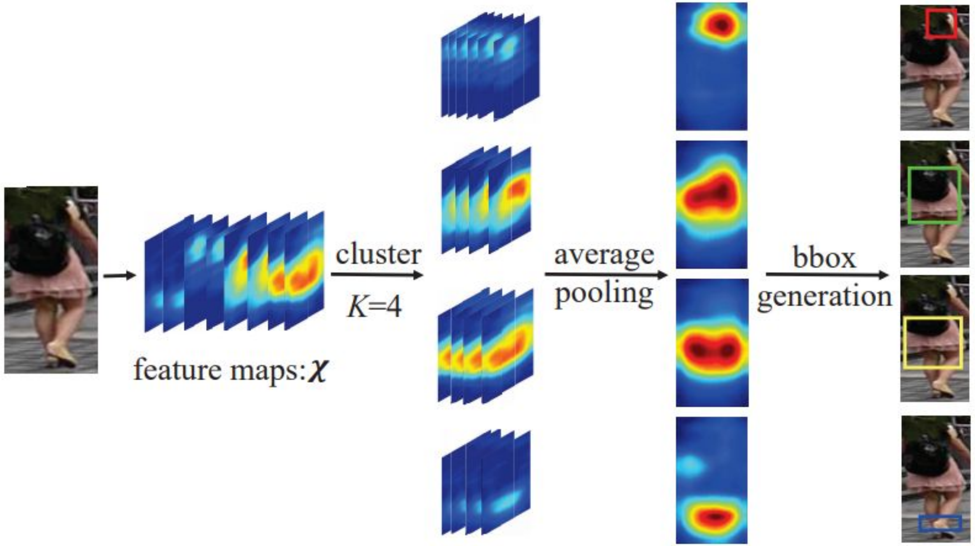
    - Clustering feature maps with similar maximum responses



H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. arXiv preprint arXiv:1707.00798, 2017.

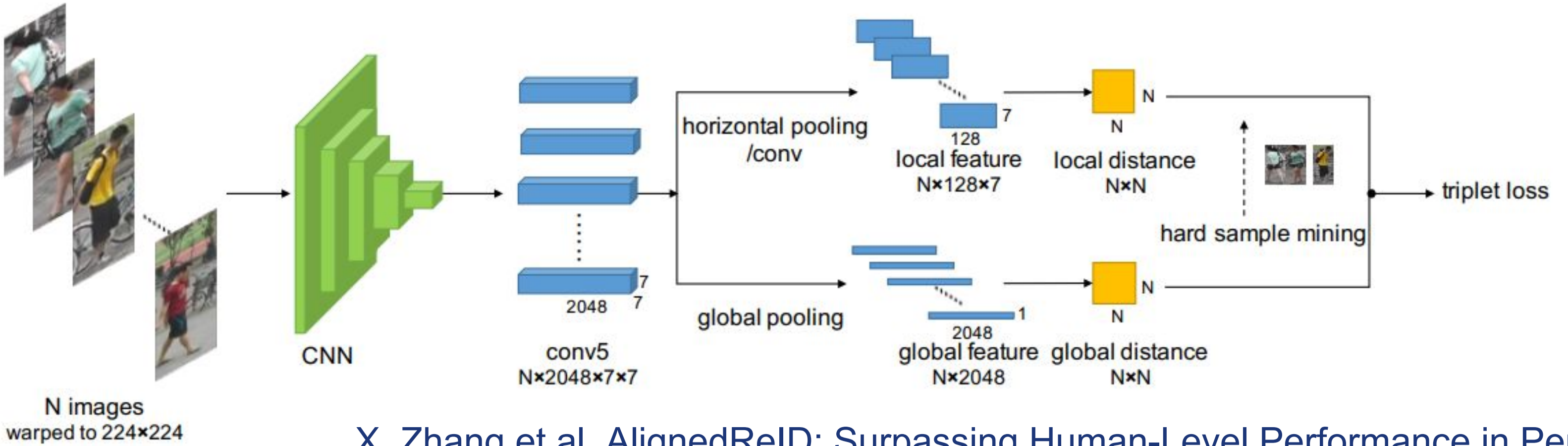
Alignment in PL-Net

- Unsupervised Part Detection



Local Feature Alignment

- AlignedReID
    - The first ReID model surpassing human-level performance



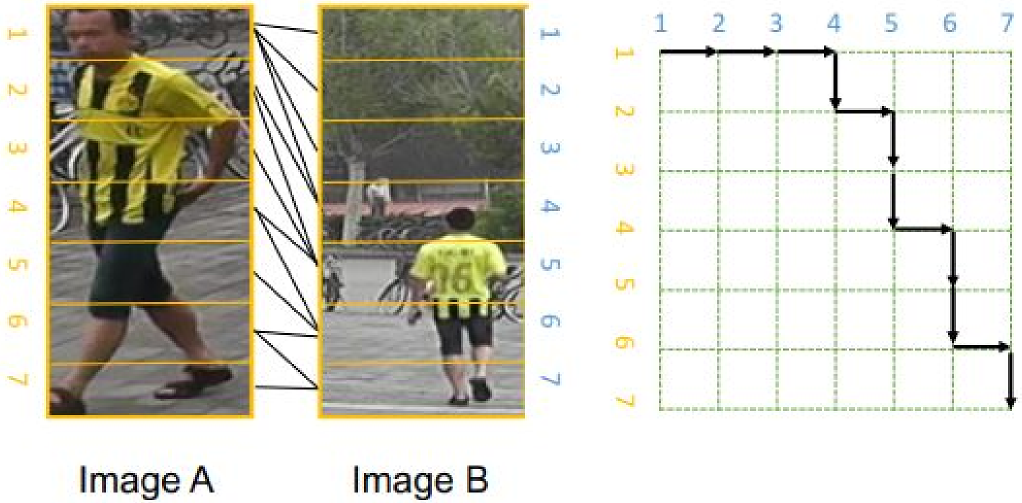
[X. Zhang et al, AlignedReID: Surpassing Human-Level Performance in Person Re-Identification, arXiv: 1711.08184]()

AlignedReID

- Distance matrix of local features

$$d_{i,j} = \frac{e^{||f_i - g_j||_2} - 1}{e^{||f_i - g_j||_2} + 1}$$

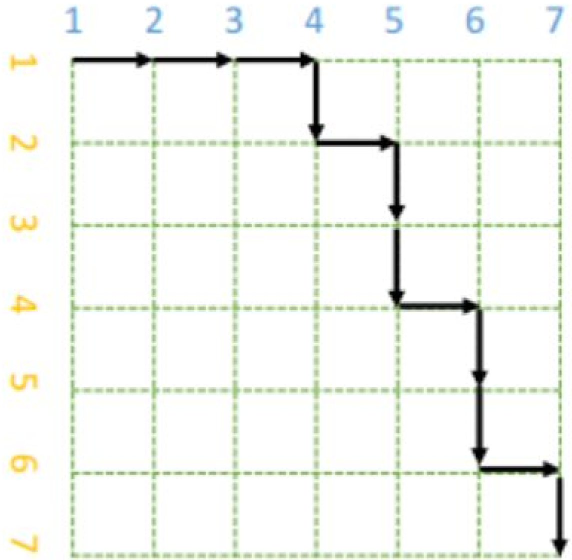
- The alignment is the one with minimum total distance



AlignedReID

- Find the shortest path by dynamic programming

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ S_{i,j-1} + d_{i,j} & i = 1, j \neq 1 \\ min(S_{i-1,j}, S_{i,j-1}) + d_{i,j} & i \neq 1, j \neq 1 \end{cases}$$



AlignedReID

- Robust to inaccurate detection, occlusion
- Discriminative to similar appearance

(a) (b) (c) (d)

Re-Ranking

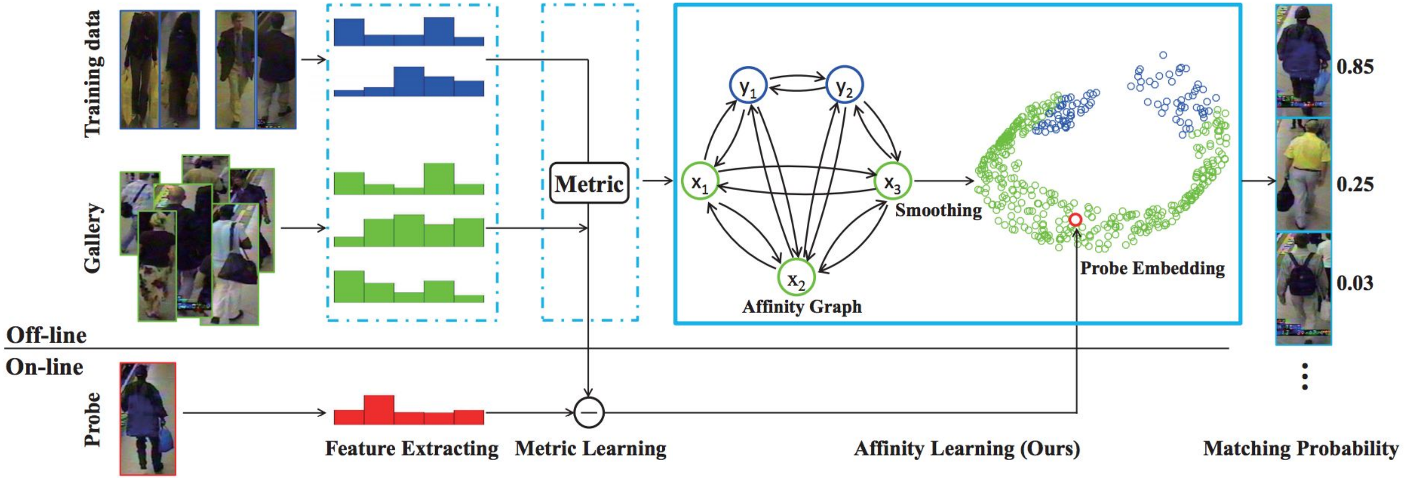
- After obtaining an initial ranking list, the re-ranking step with the relevant images will receive higher ranks
  - Re-rank on Supervised Smoothed Manifold
  - Re-rank by K-reciprocal Encoding

S. Bai, X. Bai, and Q. Tian. Scalable person reidentification on supervised smoothed manifold. arXiv preprint arXiv:1703.08359, 2017

Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. arXiv preprint arXiv:1701.08398, 2017

Re-Ranking

- Supervised Smoothed Manifold



Supervised Smoothed Manifold

- Learning smooth similarity matrix Q from initial similarity matrix W
- The data manifold is modeled as a weighted affinity graph

$$P(i \to j) = P_{ij} = \frac{W_{ij}}{\sum_{j'=1}^{N} W_{ij'}}$$

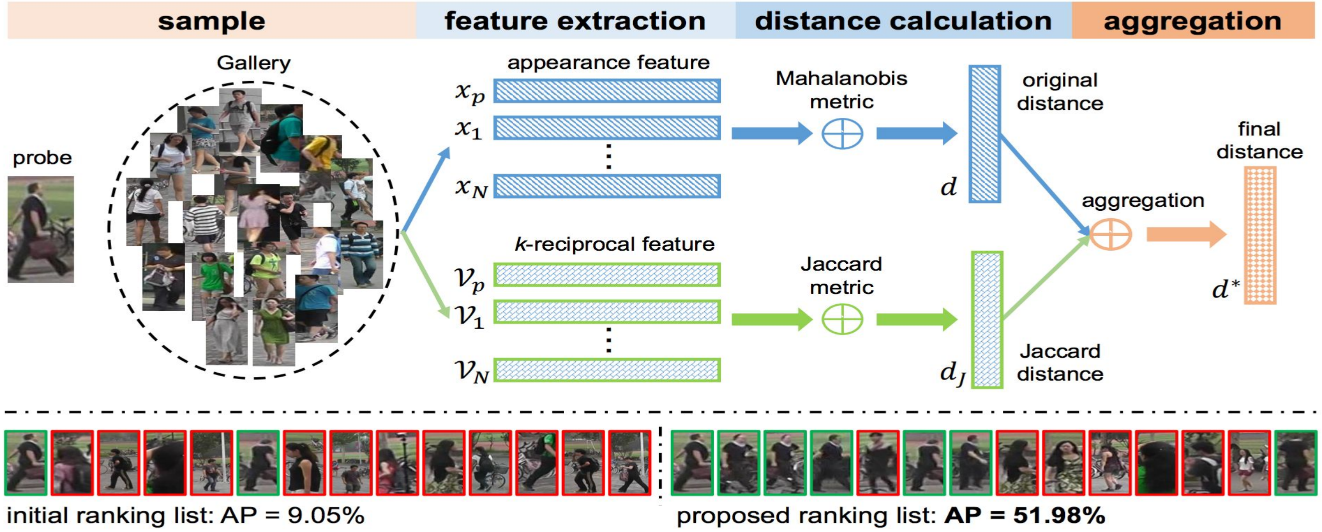
- A random walk is on the graph with edge weights

$$Q_{ki}^{(t+1)} = \alpha \sum_{l,j}^{N} \mathcal{P}(ki \to lj) Q_{lj}^{(t)} + (1-\alpha) L_{ki}$$

where $\mathcal{P}(ki \to lj) = P(k \to l) P(i \to j) = P_{kl} P_{ij}$

Re-Ranking

- K-reciprocal Encoding



Face++ 旷视

K-reciprocal Encoding

- K-nearest neighbours

$$N(p,k) = \{g_1^0, g_2^0, ..., g_k^0\}, |N(p,k)| = k$$

- K-reciprocal nearest neighbours

$$\mathcal{R}(p,k) = \{g_i \mid (g_i \in N(p,k)) \land (p \in N(g_i,k))\}$$

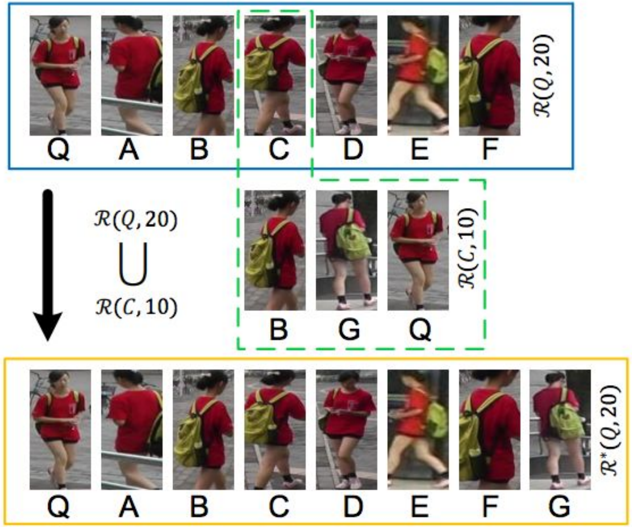
K-reciprocal Encoding

- Extend K-reciprocal nearest neighbours

$$\mathcal{R}^*(p,k) \leftarrow \mathcal{R}(p,k) \cup \mathcal{R}(q, \tfrac{1}{2}k)$$

$$s.t. \left|\mathcal{R}(p,k) \cap \mathcal{R}(q, \tfrac{1}{2}k)\right| \geqslant \tfrac{2}{3}\left|\mathcal{R}(q, \tfrac{1}{2}k)\right|,$$

$$\forall q \in \mathcal{R}(p,k)$$



Face++ 旷视

K-reciprocal Encoding

- Recalculate similarity between images
  - Jaccard distance of their k-reciprocal sets

$$d_J(p, g_i) = 1 - \frac{|\mathcal{R}^*(p,k) \cap \mathcal{R}^*(g_i,k)|}{|\mathcal{R}^*(p,k) \cup \mathcal{R}^*(g_i,k)|}$$

  - Revised Jaccard distance

$$d_J(p, g_i) = 1 - \frac{\sum_{j=1}^{N} \min(\mathcal{V}_{p,g_j}, \mathcal{V}_{g_i,g_j})}{\sum_{j=1}^{N} \max(\mathcal{V}_{p,g_j}, \mathcal{V}_{g_i,g_j})}$$

where $$\mathcal{V}_{p,g_i} = \begin{cases} \mathrm{e}^{-d(p,g_i)} & \text{if } g_i \in \mathcal{R}^*(p,k) \\ 0 & \text{otherwise.} \end{cases}$$

  - New distance

$$d^*(p, g_i) = (1 - \lambda) d_J(p, g_i) + \lambda d(p, g_i)$$

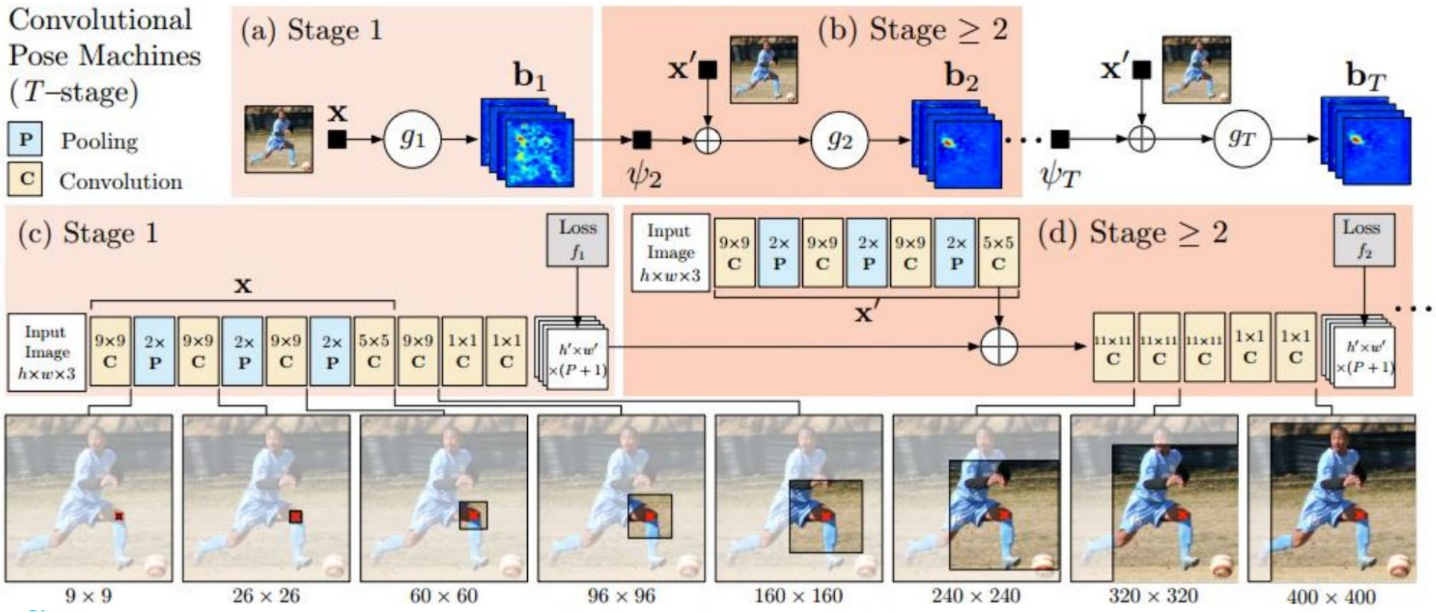
ReID with Pose Estimation

- Pose Estimation
- Global-Local Alignment Descriptor (GLAD)
    - Vertical alignment by pose estimation
- SpindleNet
    - Fusing local features from regions proposed by pose estimation

Pose Estimation

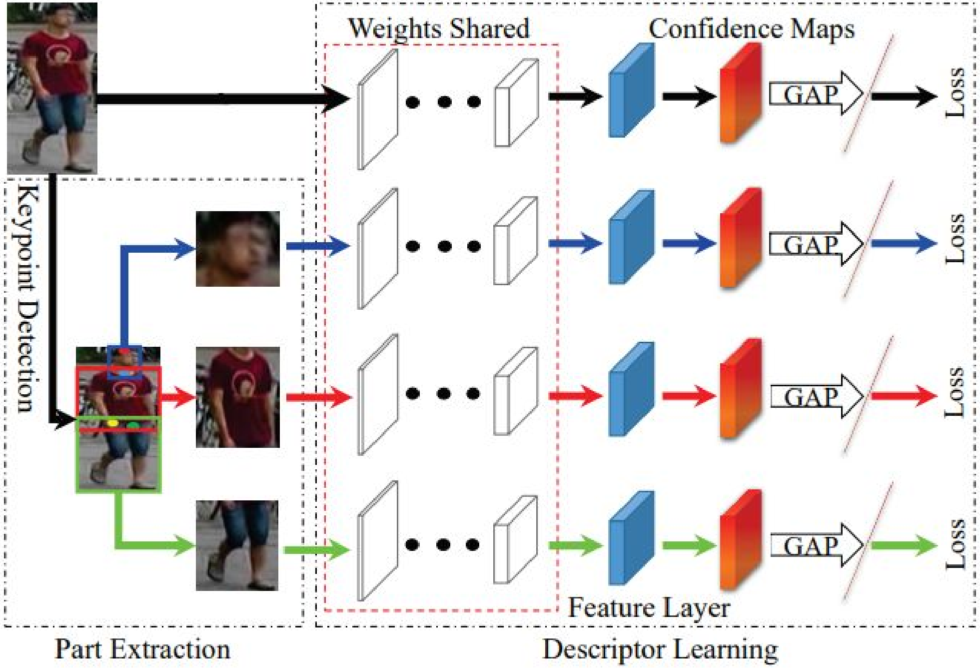
- Single Person Skeleton
  - Convolutional Pose Machines (CPM)
- Multi-Person Skeleton
  - DeeperCut

Convolutional Pose Machine (CPM)



Global-Local Alignment Descriptor (GLAD)

- Pose Estimation
  - Deeper Cut
- Part Extraction
  - Head, Upper Body, Lower Body
- Descriptor Learning
  - Concate global & local features



Global-Local Alignment Descriptor (GLAD)

- Estimate four key points of body
  - upper-head, neck, right-hip, left-hip
- Head

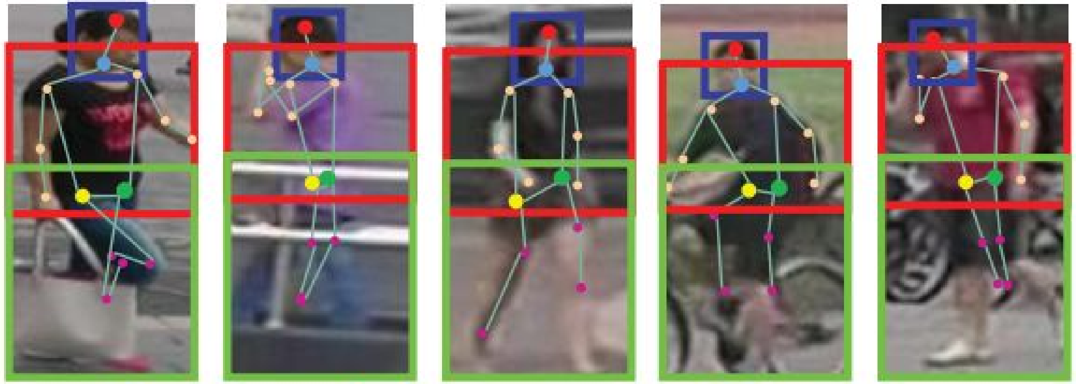
$$B^h = [(x_c - w/2, y_1 - \alpha), (x_c + w/2, y_2 + \alpha)],$$
$$w = y_2 - y_1 + 2 \cdot \alpha,$$
$$x_c = (x_1 + x_2)/2,$$

- Upper & Lower Body

$$B^{ub} = [(0, y_2 - 2 \cdot \alpha), (W - 1, y_c + 2 \cdot \alpha)],$$
$$B^{lb} = [(0, y_c - 2 \cdot \alpha), (W - 1, H - 1)],$$
$$y_c = (y_3 + y_4)/2,$$

Global-Local Alignment Descriptor (GLAD)

- Part Extraction



Head region   Upper-body region   Lower-body region

Face++ 旷视

Global-Local Alignment Descriptor (GLAD)

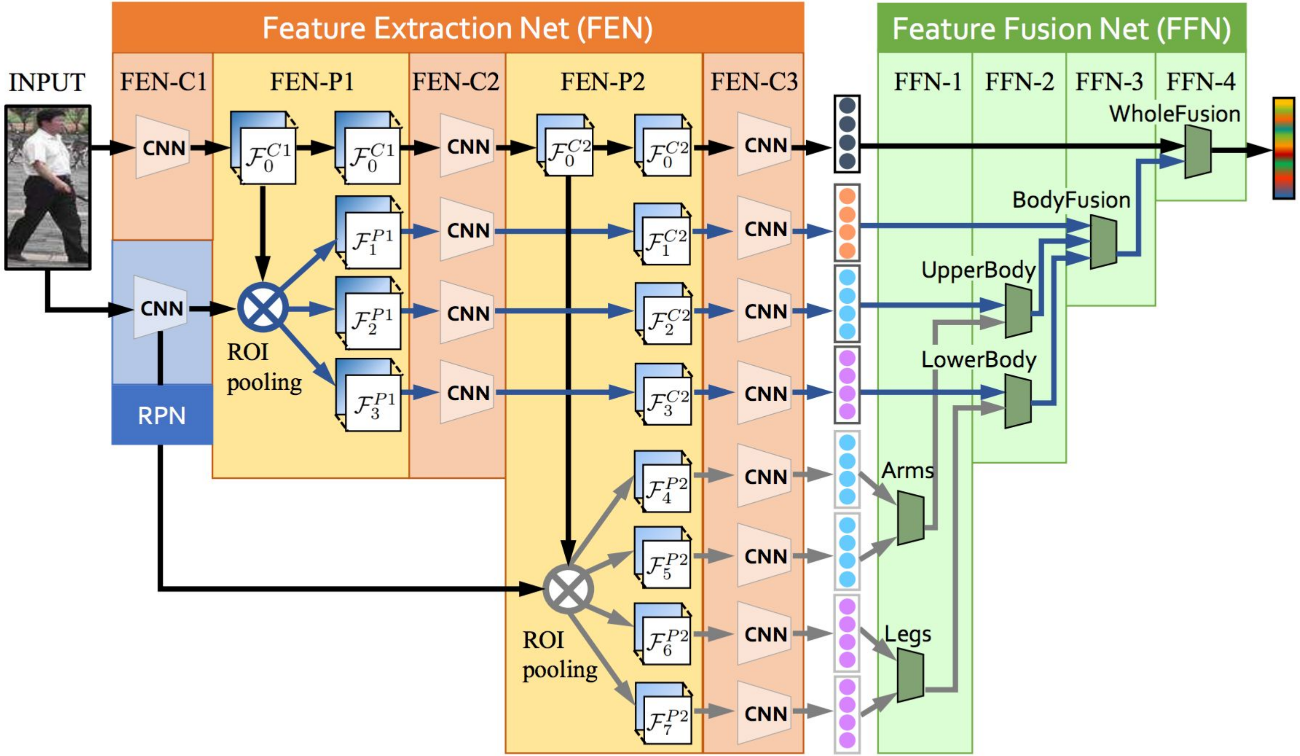
- Descriptor Learning
  - Replace FC with Global Pooling
  - Only Classification in Training
    - Global Loss for the whole body
    - Local Losses for body regions
  - Concate features in the inference stage

[L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. arXiv preprint arXiv:1709.04329, 2017]

SpindleNet

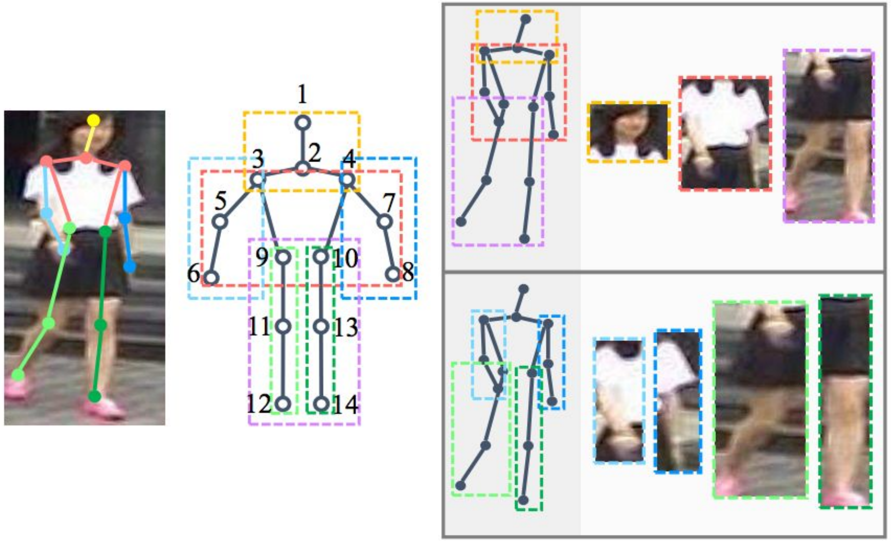
- Region Proposal Network (RPN)
  - Propose seven body regions
- Feature Extraction Network (FEN)
  - Extract semantic features from body regions
- Feature Fusion Network (FFN)
  - Merge local features with competitive scheme

H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. CVPR, 2017.


Feature Extraction Net (FEN)
Feature Fusion Net (FFN)
INPUT
FEN-C1
FEN-P1
FEN-C2
FEN-P2
FEN-C3
FFN-1
FFN-2
FFN-3
FFN-4
CNN
RPN
ROI pooling
WholeFusion
BodyFusion
UpperBody
LowerBody
Arms
Legs

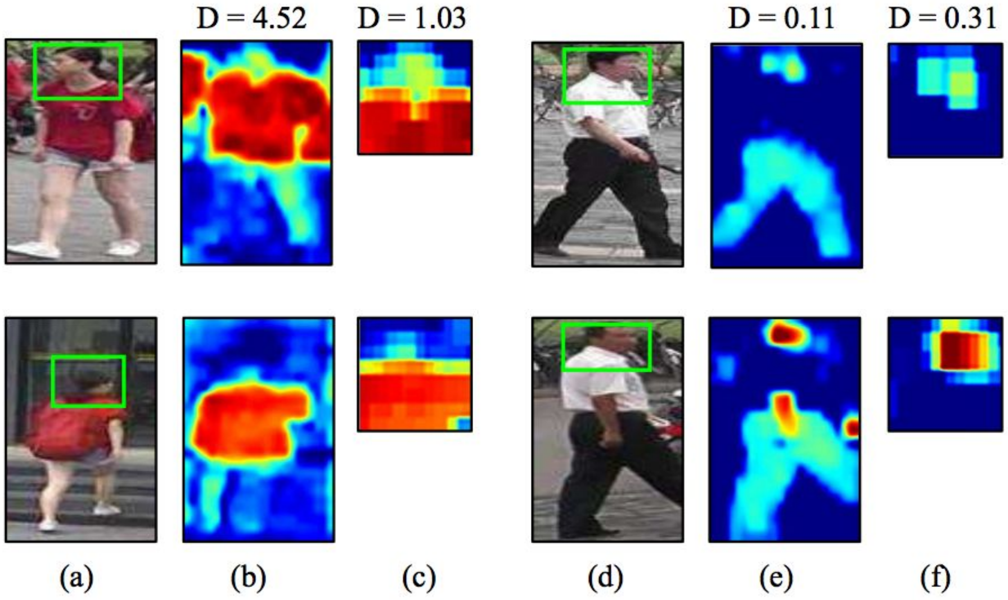
Region Proposal Network (RPN)

- CPM for body keypoints
- Minimum bounding box to contain corresponding keypoints

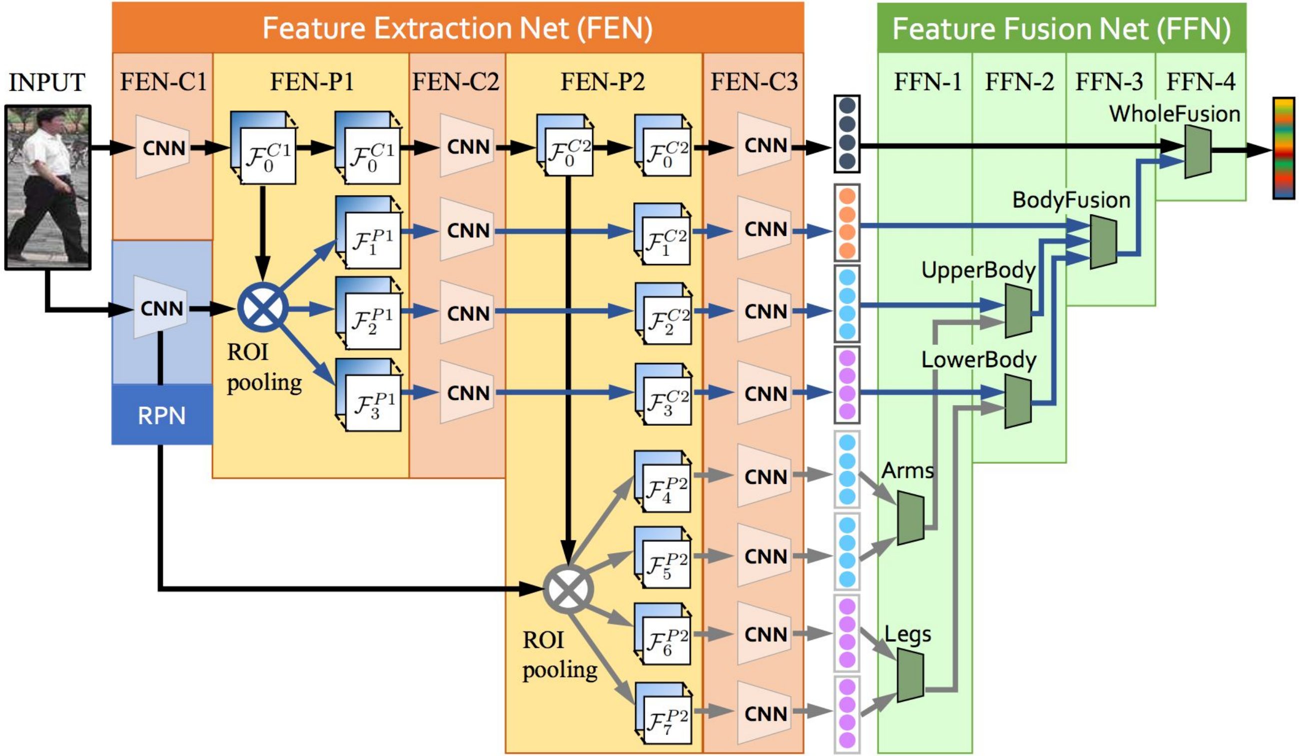


Feature Extraction Network (FEN)

- Sub-region features cropped at different stages
  - the three macro features are pooled out after the first convolution stage (FEN-C1)
  - the four micro features are pooled out after the second convolution stage (FEN-C2)



Face++ 旷视

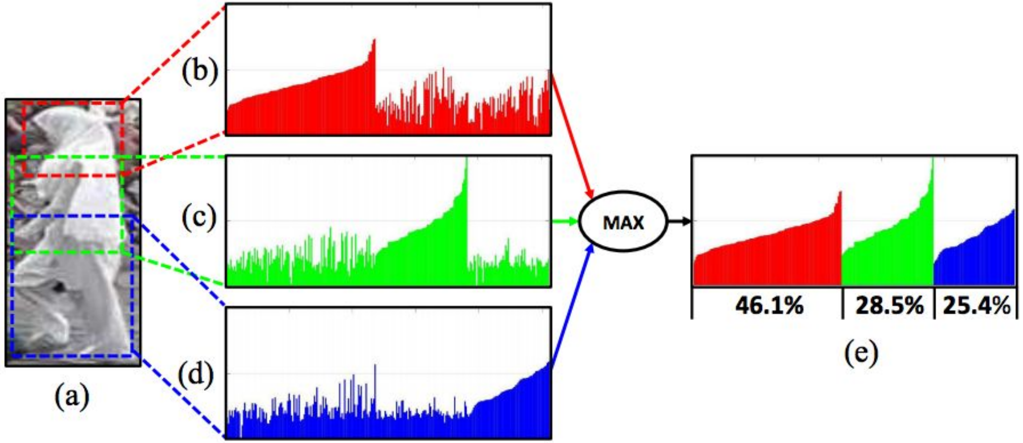


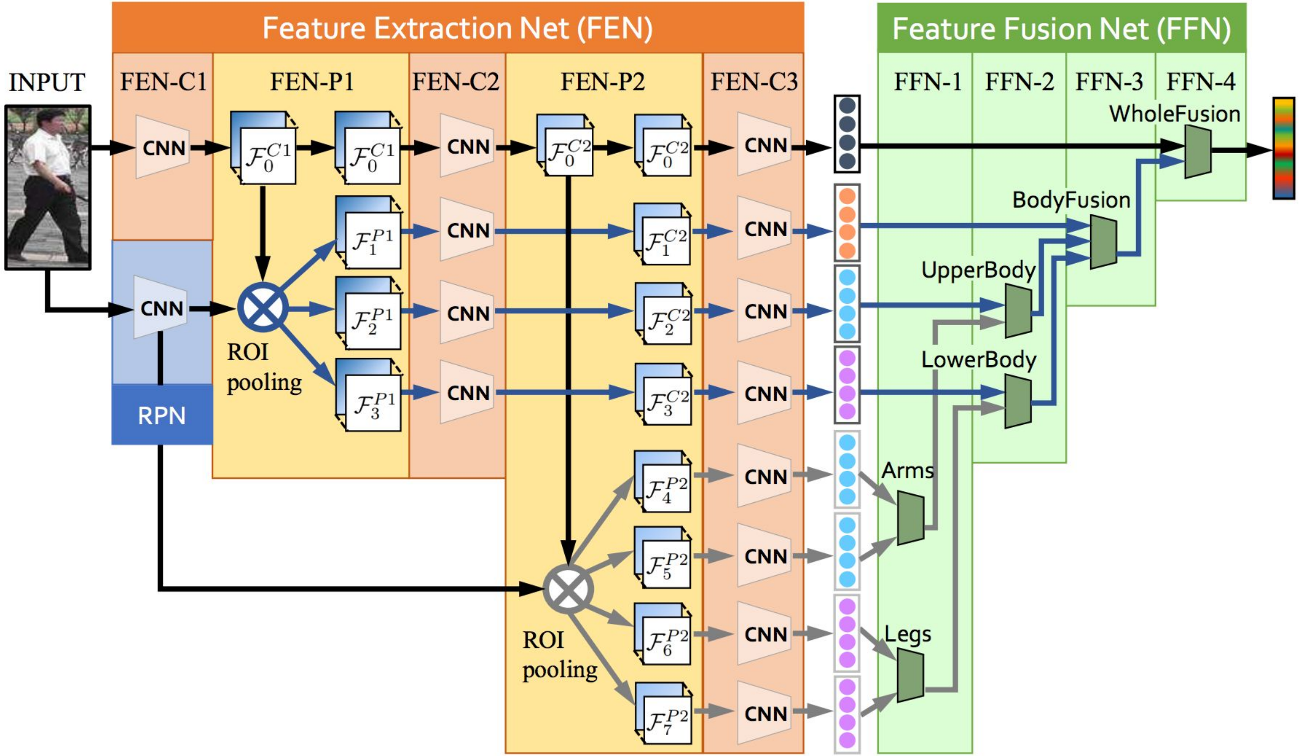
Feature Extraction Net (FEN)
Feature Fusion Net (FFN)
INPUT
FEN-C1
FEN-P1
FEN-C2
FEN-P2
FEN-C3
FFN-1
FFN-2
FFN-3
FFN-4
CNN
RPN
ROI pooling
$\mathcal{F}_0^{C1}$
$\mathcal{F}_1^{P1}$
$\mathcal{F}_2^{P1}$
$\mathcal{F}_3^{P1}$
$\mathcal{F}_0^{C2}$
$\mathcal{F}_1^{C2}$
$\mathcal{F}_2^{C2}$
$\mathcal{F}_3^{C2}$
$\mathcal{F}_4^{P2}$
$\mathcal{F}_5^{P2}$
$\mathcal{F}_6^{P2}$
$\mathcal{F}_7^{P2}$
WholeFusion
BodyFusion
UpperBody
LowerBody
Arms
Legs

Face++ 旷视

Feature Fusion Network (FFN)

- Feature vectors of different body subregions are merged in different stages in tree-structured
- Feature competition with element-wise max operation




Feature Extraction Net (FEN)
Feature Fusion Net (FFN)
INPUT
FEN-C1
FEN-P1
FEN-C2
FEN-P2
FEN-C3
FFN-1
FFN-2
FFN-3
FFN-4
CNN
RPN
ROI pooling
WholeFusion
BodyFusion
UpperBody
LowerBody
Arms
Legs

ReID with Person Attributes

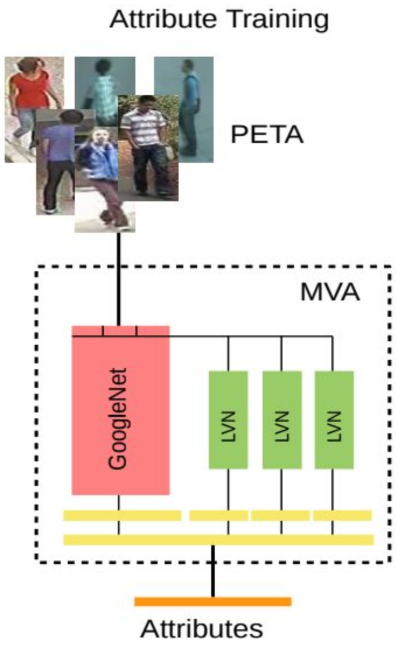
- Attribute Complementary ReID Network
    - Train an attribute classifier on separate data
    - Train reid model with the attribute loss
    - Inference with the attribute & reid representation

CUHK3	Market-1501	DukeMTMC-reID
Male Long Hair Jacket Backpack	Backpack Skirt Male Long Hair	Jacket Casual Upper Trousers Male
Sandals V-neck	Hat V-Neck	Sandals Messenger Bag

ReID with Person Attributes

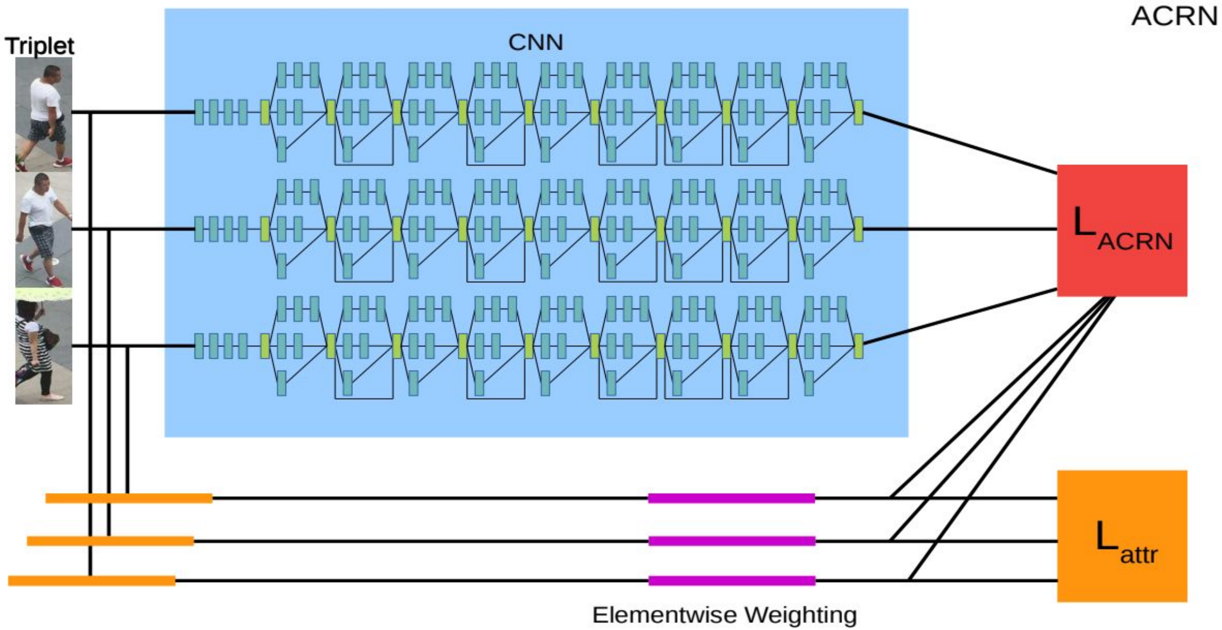
- Train Attribute Classifier
  - Base Network for global feature
  - Local View Networks for region features
  - a single multi-class cross-entropy loss, instead of one loss for each attribute
  - Weighting attributes in loss

$$L_{wce} = \sum_{i=1}^{L} \frac{1}{2w_i} * p_i * log(q_i) + \frac{1}{2(1-w_i)}(1-p_i) * log(1-q_i)$$



Face++ 旷视

ReID with Person Attributes

- 

ReID with Person Attributes

- Attribute Complementary ReID Network
  - Incorporate attribute loss into the triplet loss
  - Weighting different attributes
  - In inference stage, the ReID representation is used together with the weighted attribute representation

$$L_{ACRN} = \frac{1}{N} \sum_{i=1}^{N} d_i^{f^p} - d_i^{f^n} + m + \gamma(d_i^{att^p} - d_i^{att^n})$$

$$d_i^{att^p} = \|att_i^a - att_i^p\|_2^2$$

$$d_i^{att^n} = \|att_i^a - att_i^n\|_2^2$$

Face++ 旷视

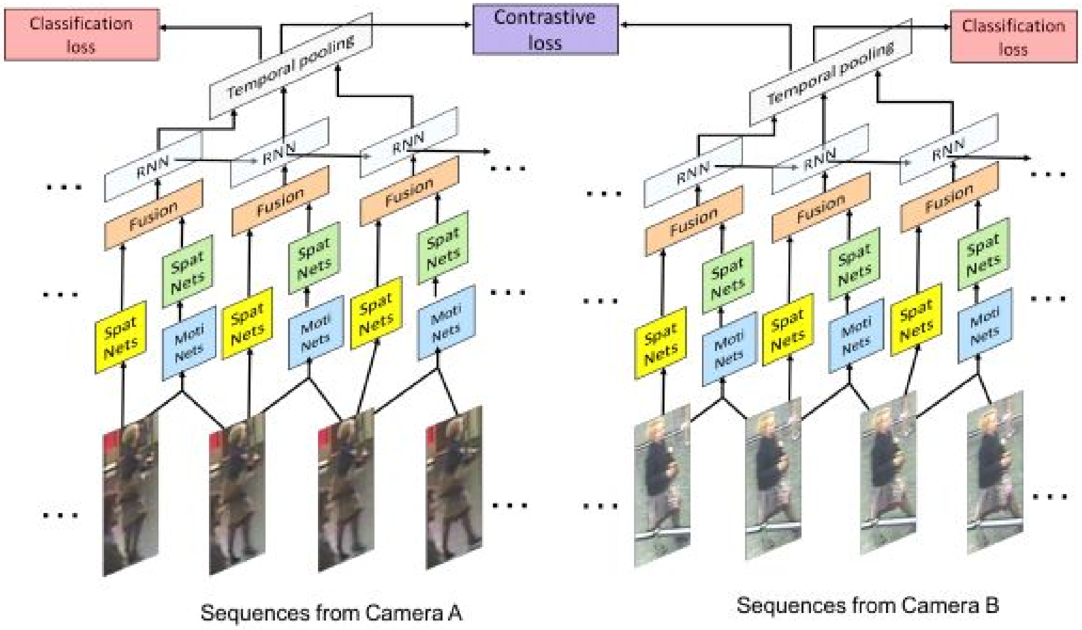
A. Schumann, R. Stiefelhagen, Person Re-Identification by Deep Learning Attribute-Complementary Information, CVPR2017

ReID for Tracklets

- Accumulative Motion Context (AMOC) network
  - Motion context from trained optical flow
  - Two-stream spatial-temporal architecture
  - RNN to summarize long-term information

Accumulative Motion Context (AMOC) Network

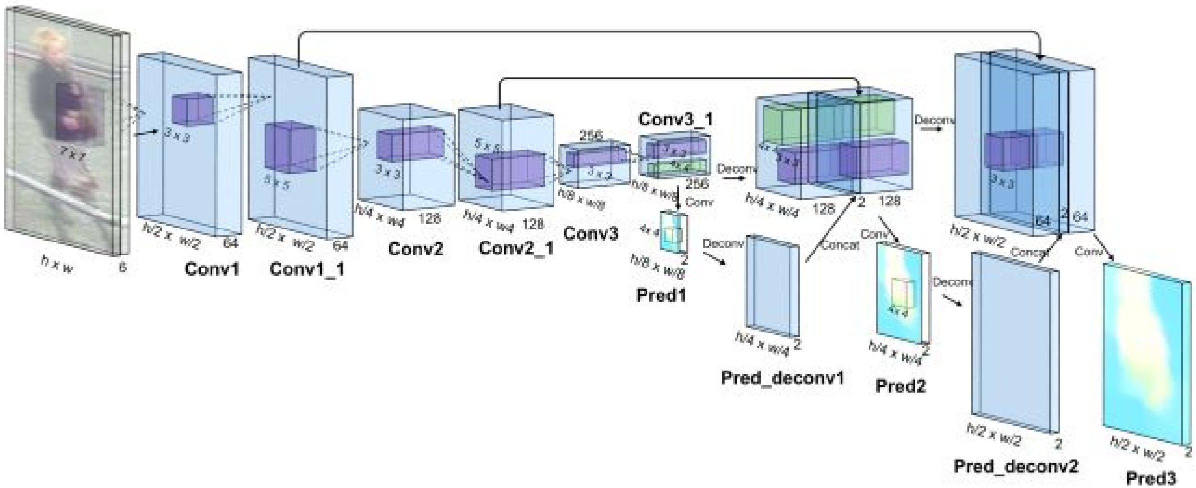
- AMOC for video-based ReID
  - Motion Net
  - Spatial Net
  - RNN
  - Temporal Pooling



Face++ 旷视

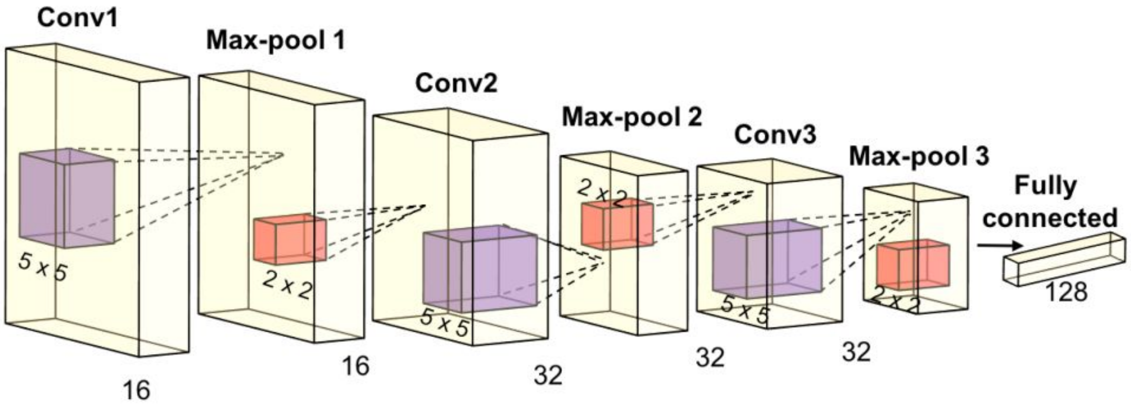
Accumulative Motion Context (AMOC) Network

- Motion Net



Accumulative Motion Context (AMOC) Network

- Spatial Network
- Spatial Fusion
- Temporal RNN



[H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, Video-based Person Re-identification with Accumulative Motion Context, arXiv:1701.00193, 2017]

Face++ 旷视