Person Identification:

Face Recognition & Person Re-Identification

Chi Zhang
Megvii (Face++)
zhangchi@megvii.com
Jun 2018

Face++ 旷视

Outline

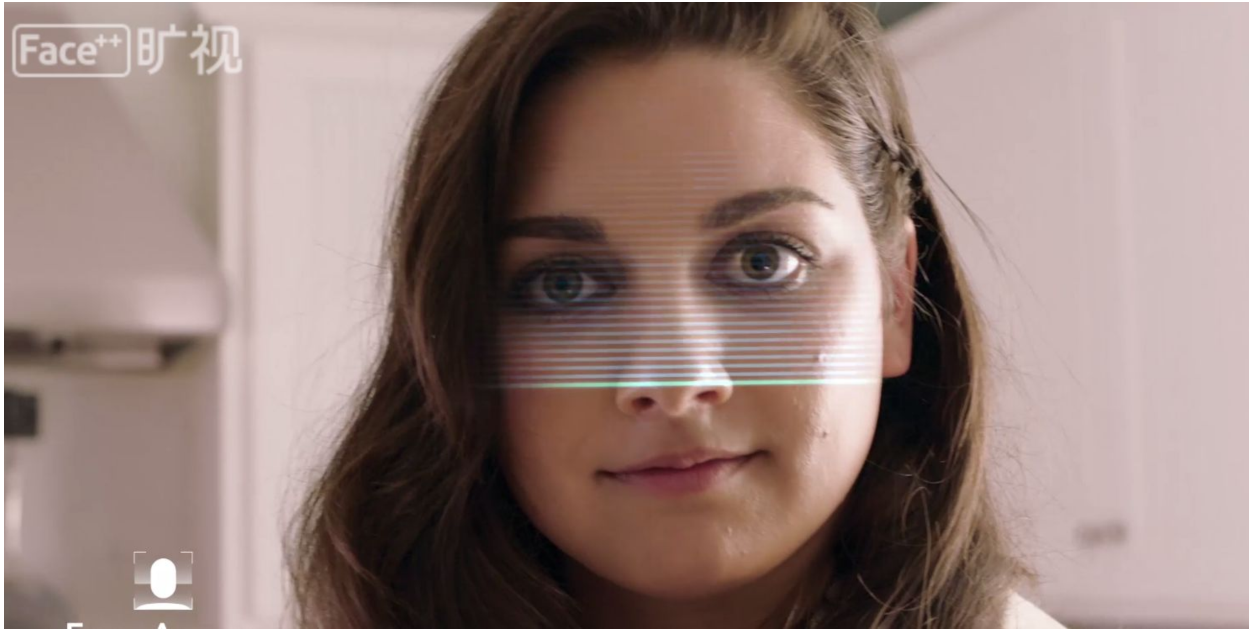
- **Face Recognition**
  - Applications
  - Classification
  - Metric Learning
  - Hard Sample Mining
- Person Re-Identification
  - Applications
  - Feature Alignments
  - ReID with Pose Estimation

Face Recognition: Applications

Face++ 旷视

Face Recognition: Applications

- Mobile Phone



Face Recognition: Applications

- City Brain

Face++ 旷视

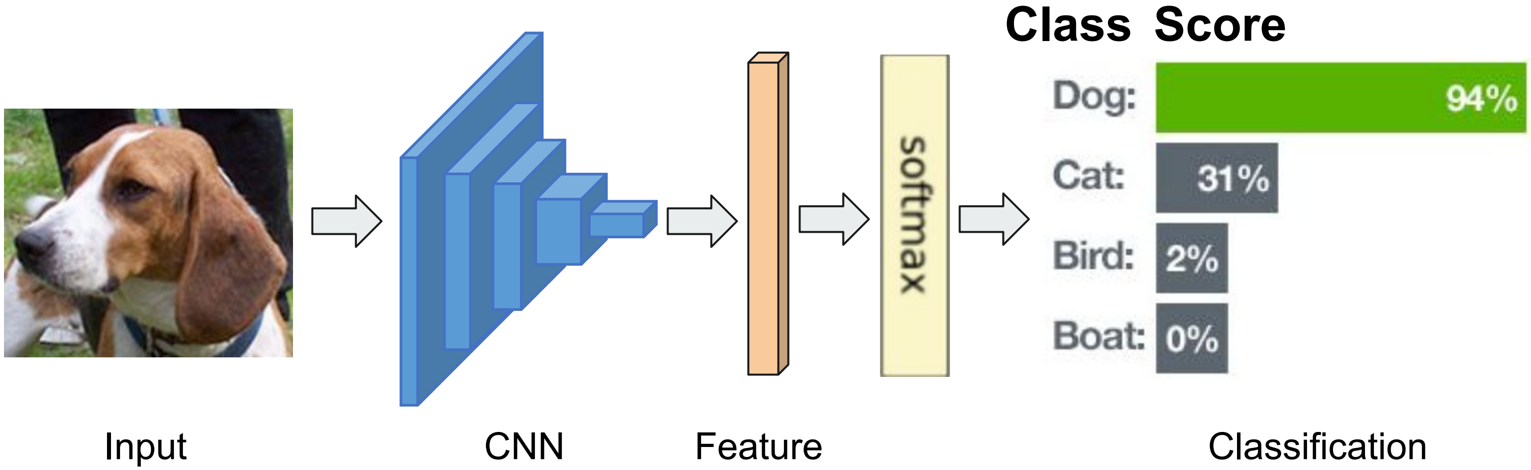
Face Recognition: Applications

- New Retail

Face++ 旷视

Face Recognition: Classification

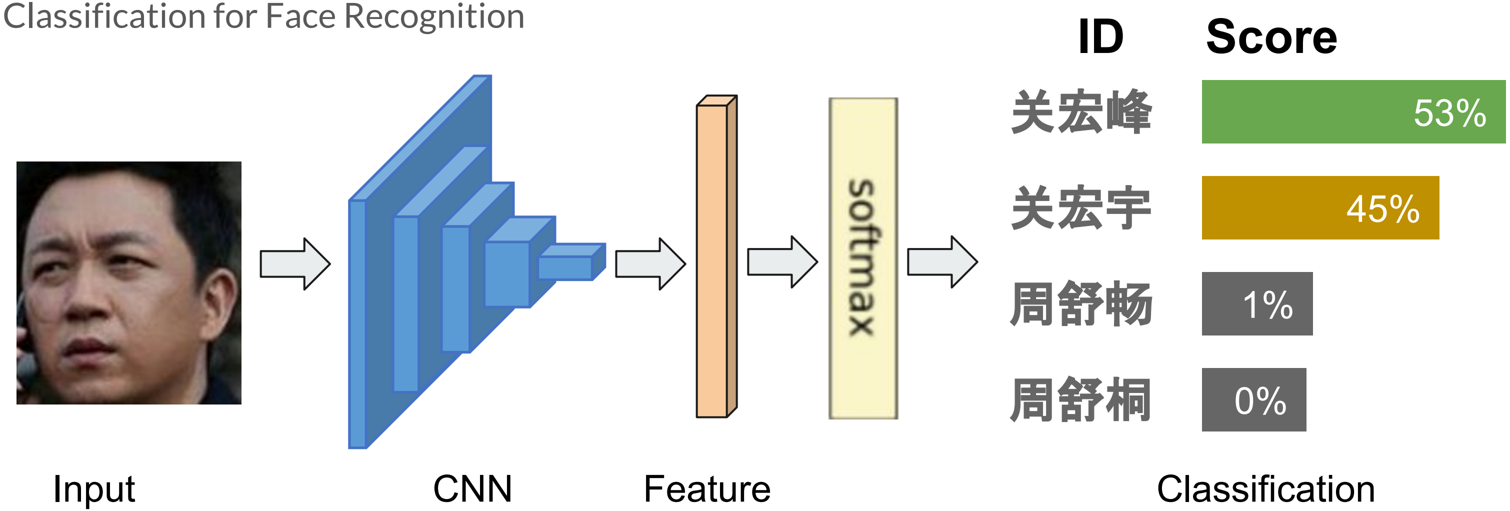
- General Classification in Deep Learning



Face++ 旷视

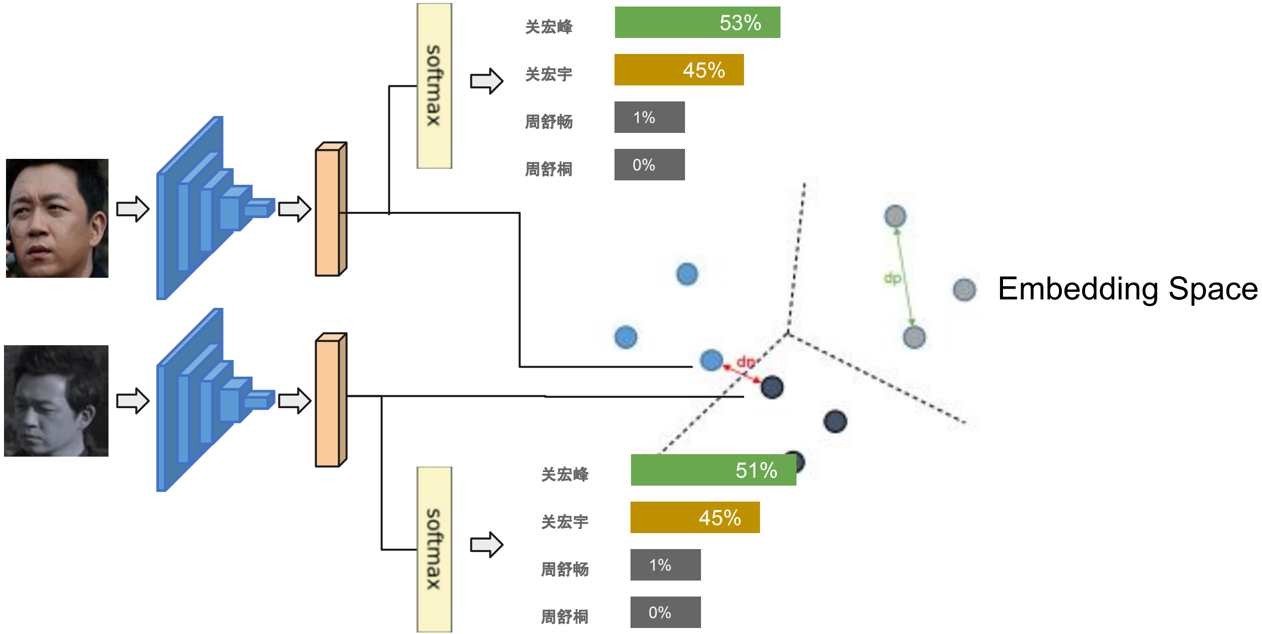
Face Recognition: Classification

- Classification for Face Recognition



Face++ 旷视

Face Recognition: Classification



Face++ 旷视

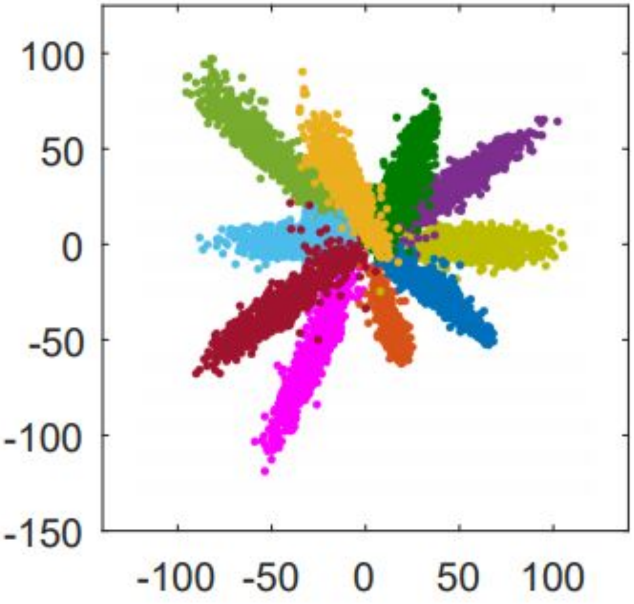
Face Recognition: Classification

- Softmax

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}}$$

$$f_j = \boldsymbol{W}_j^T \boldsymbol{x}_i + b_j$$

$$-\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\cos(\theta_{y_i})}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

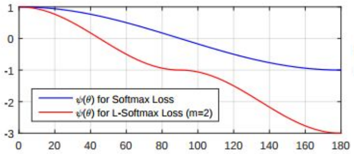


Face++ 旷视

Face Recognition: Classification

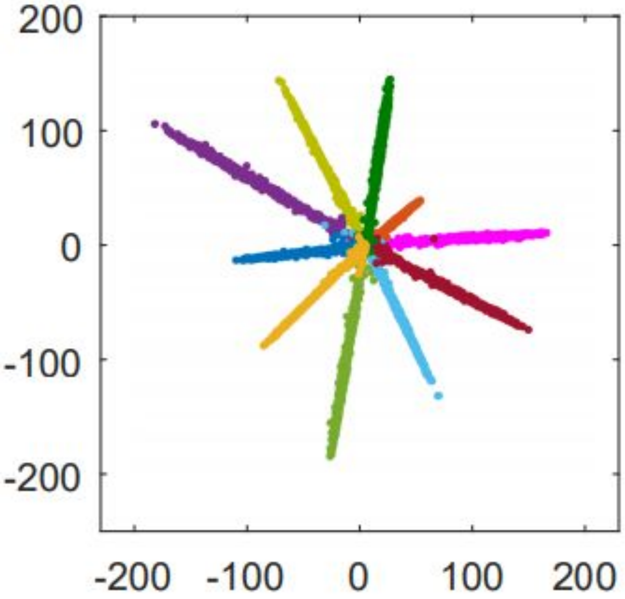
- L-Softmax

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}}$$



$$\|\boldsymbol{W}_1\|\|\boldsymbol{x}\|\cos(\theta_1) \geq \|\boldsymbol{W}_1\|\|\boldsymbol{x}\|\cos(m\theta_1) > \|\boldsymbol{W}_2\|\|\boldsymbol{x}\|\cos(\theta_2).$$

$$-\log\left(\frac{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})}}{e^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|\psi(\theta_{y_i})} + \sum_{j\neq y_i} e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$



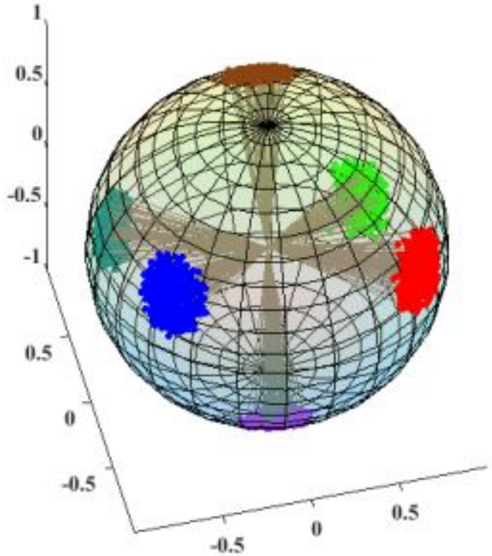
Face Recognition: Classification

- A-Softmax (SphereFace)

$$L_s = \frac{1}{N}\sum_{i=1}^{N} - \log p_i = \frac{1}{N}\sum_{i=1}^{N} - \log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}}$$

Normalize weights

$$- \log \left(\frac{e^{\|\boldsymbol{x}_i\| \psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\| \psi(\theta_{y_i,i})} + \sum_{j \neq y_i} e^{\|\boldsymbol{x}_i\| \cos(\theta_{j,i})}} \right)$$



Face++ 旷视

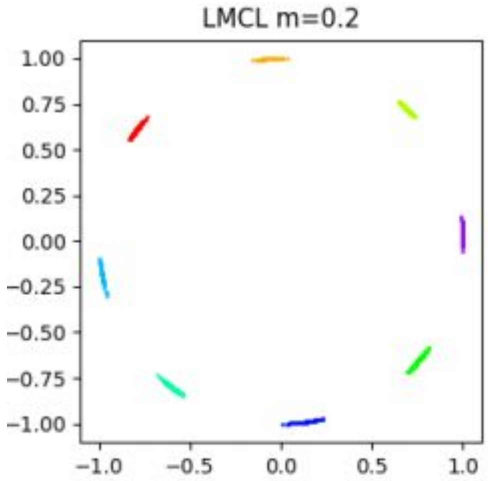
Face Recognition: Classification

- Large Margin Cosine Loss (CosFace)

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}}$$

Normalize weights
Normalize features
Replace angular margin by cosine margin

$$-\log \frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i} e^{s\cos(\theta_{j,i})}}$$



Face++ 旷视

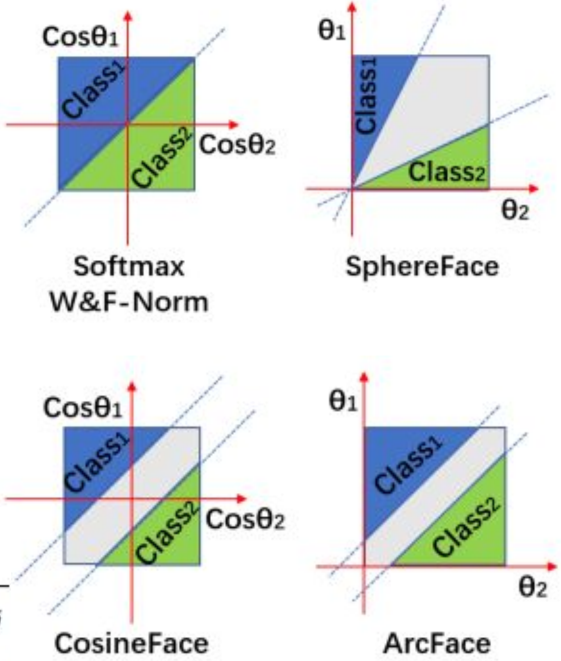
Face Recognition: Classification

- ArcFace

$$L_s = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_{y_i}}}{\sum_{j=1}^{C} e^{f_j}}$$

Back to Angular Space

$$\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}}$$



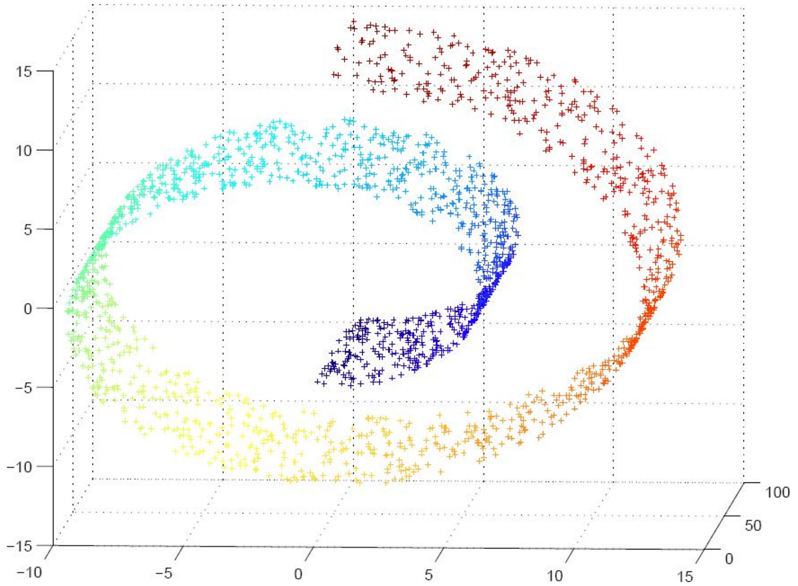
Face++ 旷视

Face Recognition: Classification

- Paradox
  - Classification can only discriminate the “seen” objects
- To recognize “unseen” objects
  - The similarity of the features learned in classification
  - Similar Classification Probability to Closer Feature Distance
- Beyond Softmax
  - Large Margin Cosine Loss is effective and easy to train

From Classification to Metric Learning

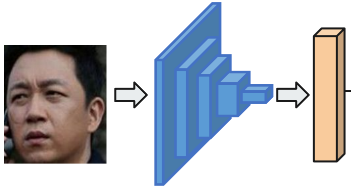
- Directly train model from Loss of feature distances
  - Learn a function that measures how similar two objects are
  - Compared to classification which works in a closed-word, metric learning deals with an open-world.
  - Metric Learning can be done together with Classification



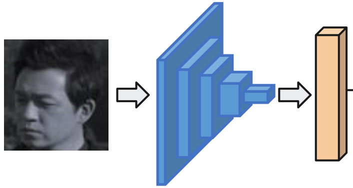
Face++ 旷视

Metric Learning: Contrastive Loss

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



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

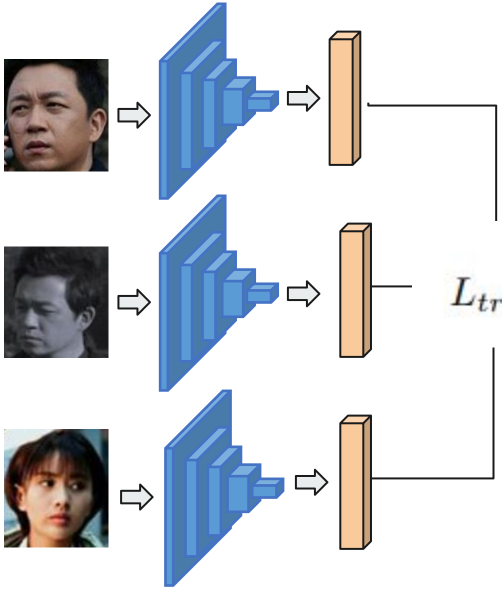


Metric Learning: Contrastive Loss

- The distance of images with the same identity (positive pairs) should be smaller
- The distance of images with different identities (negative pairs) should be larger
- α is used to ignore the “naive” negative pairs

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

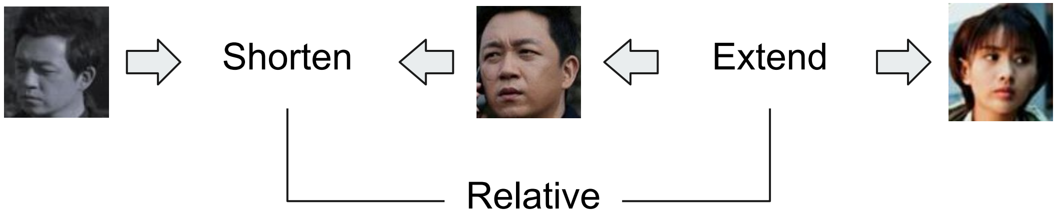
Metric Learning: Triplet Loss



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

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
- α is the margin between negative and positive pairs.
- Without α, all distance converge to zero.



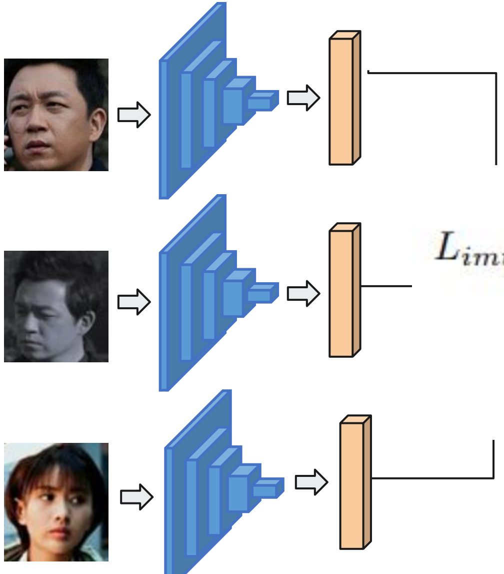
H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing, 2017

Contrastive Loss vs. Triplet Loss

- Contrastive Loss:
  - Margin between all positive pairs and negative pairs
  - Positive & negative pairs are also constrained
  - Positive pairs are always trained
  - Negative pairs are trained until it is greater than the margin
- Triplet Loss
  - Margin between positive paris and negative pairs **given the query**
  - Stop training positive(negative) pairs that are smaller(larger) than all negative(positive) pairs with a margin
  - Pay more attention to samples that disobey the order
  - Suffers from lack of generality
- Complementary to Triplet Loss
  - Improved Triplet Loss
  - Quadruplet Loss

Metric Learning: Improved Triplet Loss

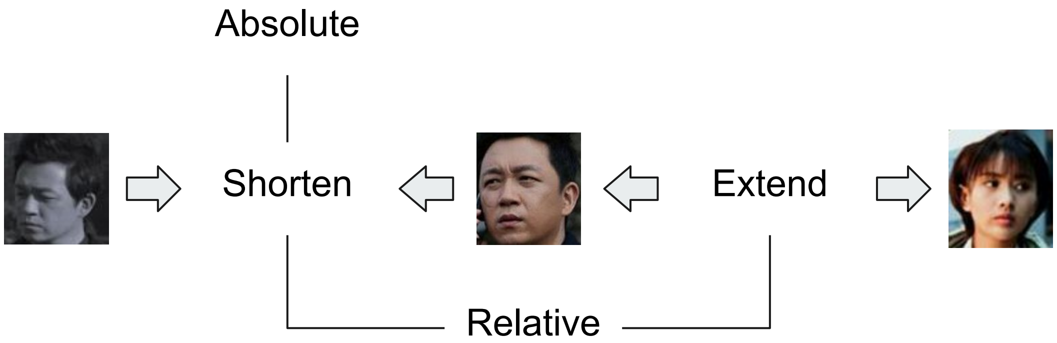
- β-term penalizes distance between features of A and A'



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

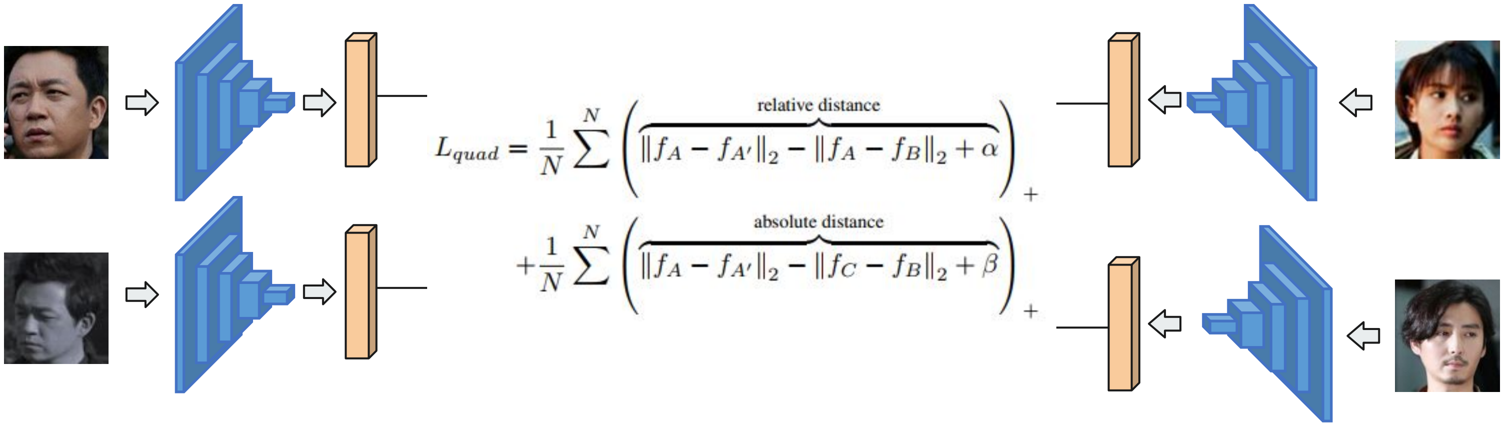
Metric Learning: Improved Triplet Loss

- 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

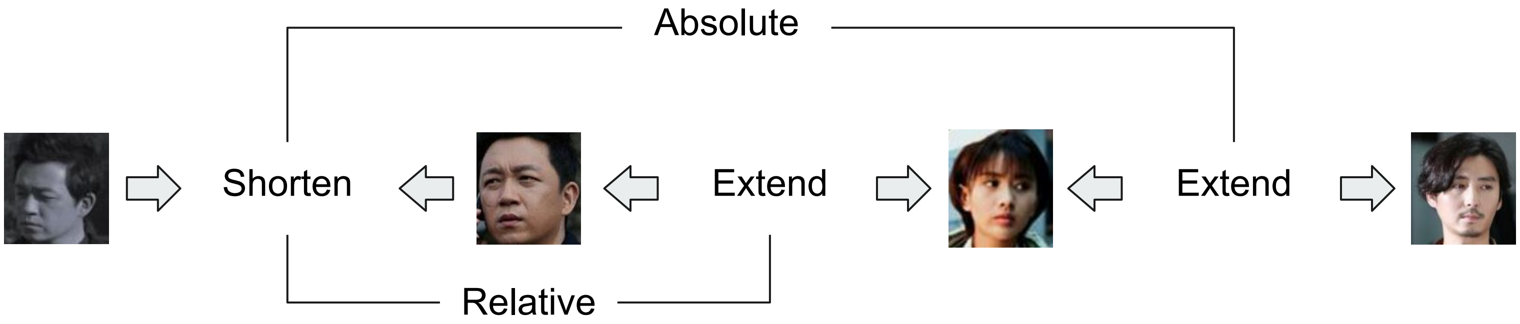
Metric Learning: Quadruplet Loss



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

Metric Learning: Quadruplet Loss

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



Face++ 旷视

[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.]

Improved Triplet Loss & Quadruplet Loss

- Common
  - Introduce loss to “strengthen” triplet loss
  - Samples are still trained when triplet constraint is satisfied
- Difference
  - Improved Triplet Loss
    - An absolute margin is given for positive pairs
  - Quadruplet Loss
    - A relative margin between all positive pairs and negative pairs
- What if?

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

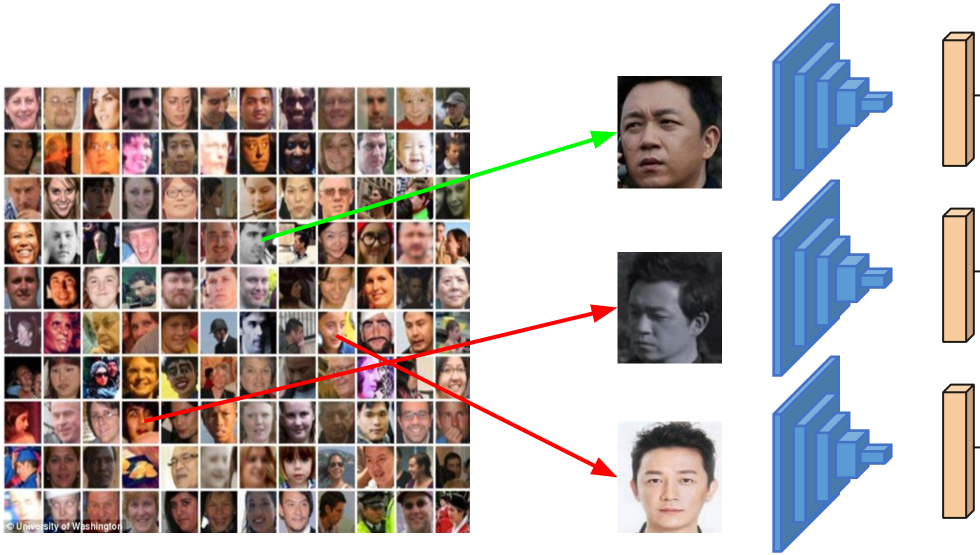
Hard Sample Mining

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

Trivial:

Non-Trivial:

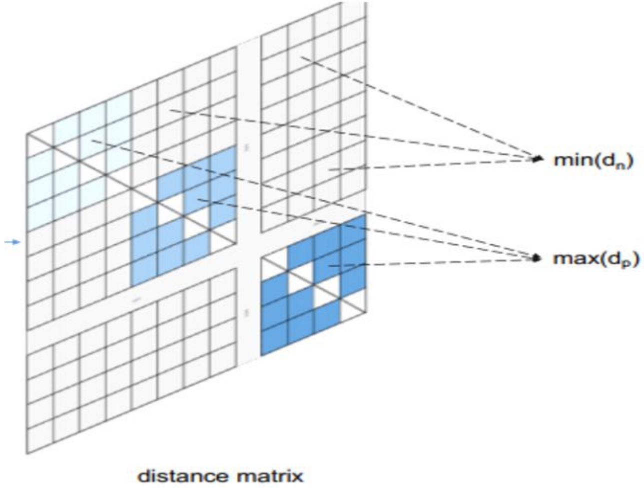
Hard Sample Mining: Triplet Hard Loss



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

Hard Sample Mining: Triplet Hard Loss

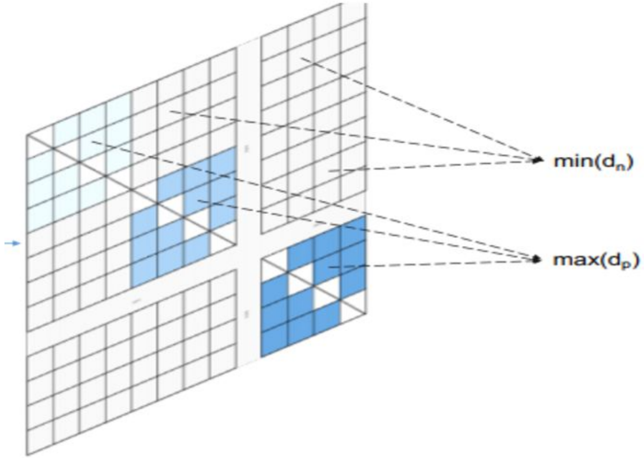
- Each batch contains K identities, each identities contains L images
- Compute the distance between each images in the batch
- Distance matrix
  - Diagonal Blocks are distance between images with the same identity
  - Others are distance between images with different identities



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

Hard Sample Mining: Triplet Hard Loss

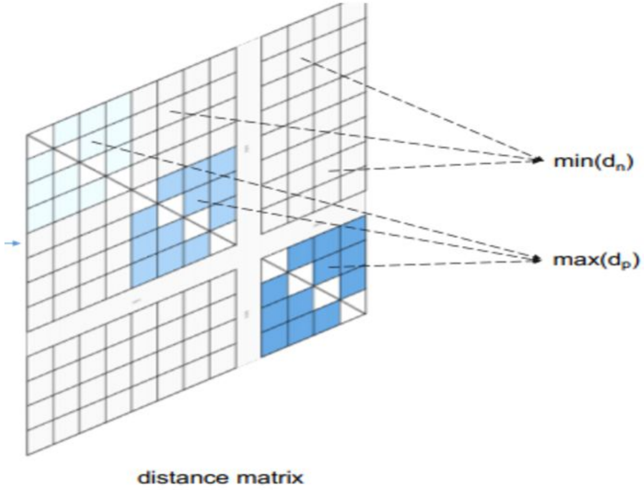
- Generate a triplet from **each line** in the matrix
  - Each image in the batch
- The largest distance in the diagonal block
  - The most unsimilar image with the same identity
- The smallest distance in other places
  - The most similar image with a different identity



distance matrix

Hard Sample Mining: Soft Triplet Hard Loss

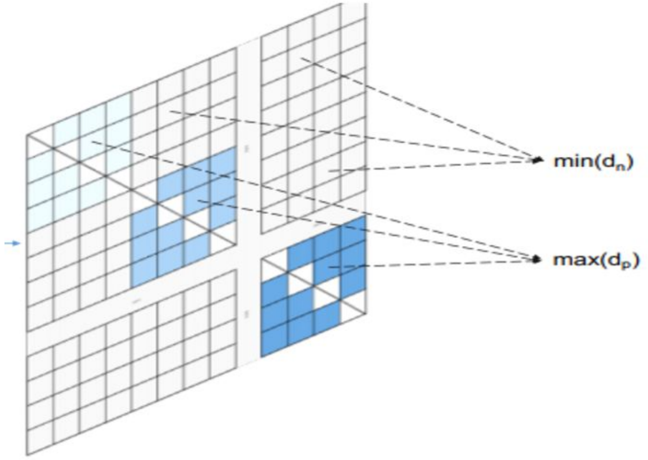
- Generate a triplet from **each line** in the matrix
    - Each image in the batch
- The weighted average distance in the diagonal block
    - Softmax(d_ij)
- The weighted average distance in the diagonal block
    - Softmax(-d_ik)
- The harder samples with larger weights



Face++ 旷视

Hard Sample Mining: Margin 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

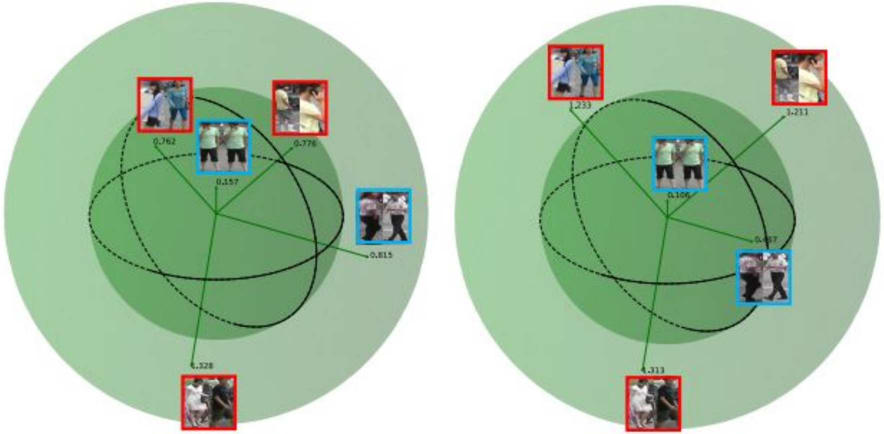


[Q. Xiao, H. Luo, C. Zhang, Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification, arXiv: 1710.00478]

Hard Sample Mining: Margin Sample Mining

- Margin Sample Mining

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



(a) TriHard (b) MSML

Face Recognition : Conclusion

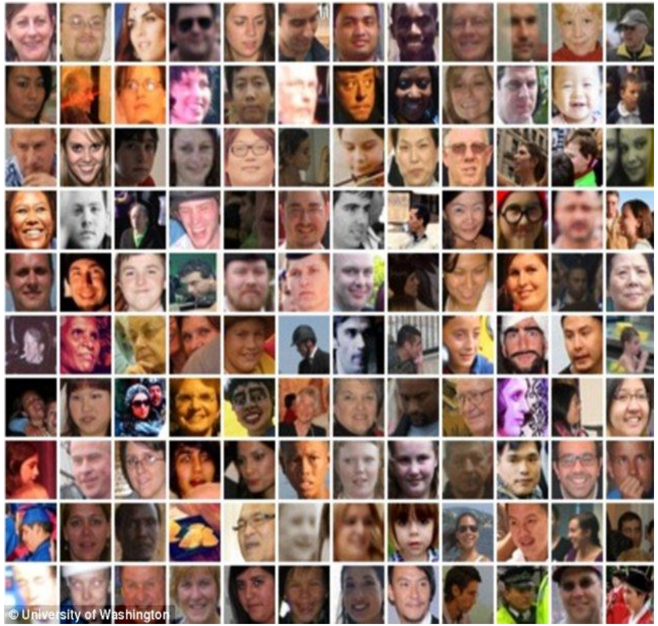
- Embedding images to feature space
  - Similar instances should be closer in the space
- Classification vs. Metric Learning
  - Triplet Loss (and its improvements) performs better than contrastive loss
  - Advanced classification, such as Large Margin Cosine Loss, comparable to Triplet Loss
  - Combining classification and metric learning always performs better
- Hard Sample Mining
  - Critical to achieve high accuracy

Outlines

- Face Recognition
  - Applications
  - Classification
  - Metric Learning
  - Hard Sample Mining
- **Person Re-Identification**
  - Applications
  - Feature Alignments
  - ReID with Skeleton
  - ReID with Attributes

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



Face++ 旷视

From Face to Person

- Person Re-Identification
  - Applications
    - Tracking in a single camera
    - Tracking across multiple cameras
    - Searching a person in a set of videos
    - Clustering persons in a set of photos
  - Challenges
    - Inaccurate detection
    - Misalignment
    - Illumination difference
    - Occlusion

From Face to Person

- Different Directions

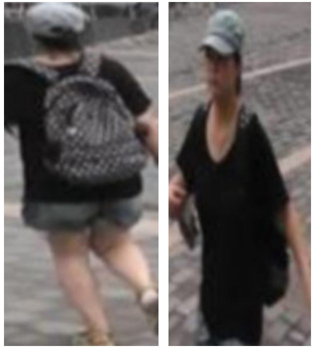
- Non-rigid Body Deformation

- Different Illumination

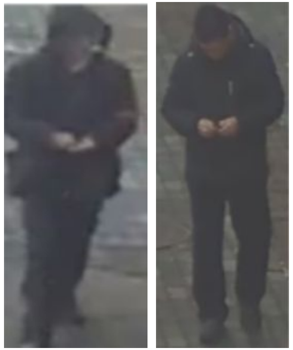
From Face to Person

- Occlusion

- Incomplete



- Similar Appearance



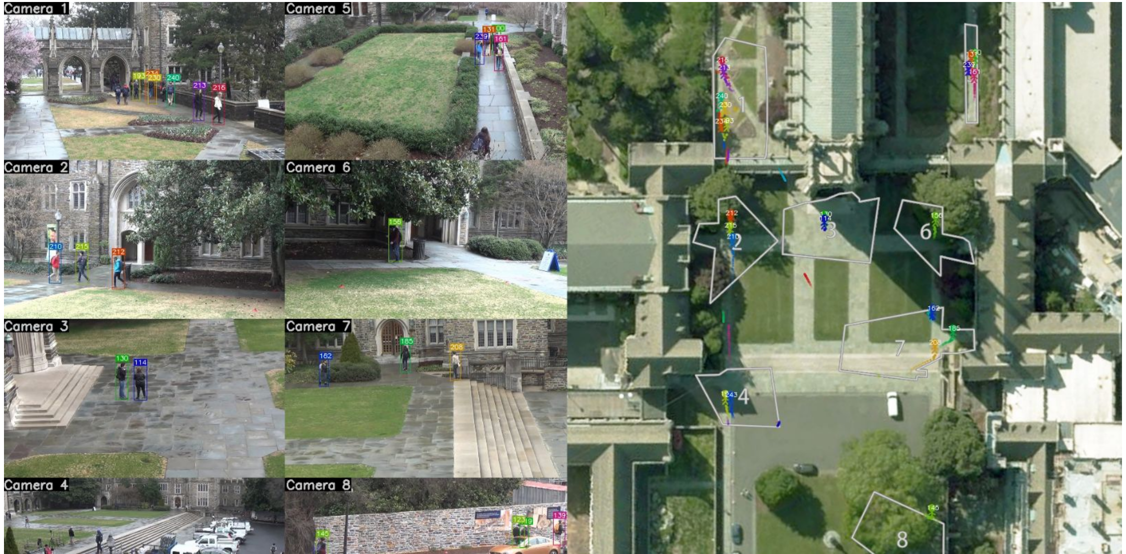
Re-IDentification: Applications

- Single Camera Tracking

Face++ 旷视

Re-IDentification: Applications

- Multiple Camera Tracking

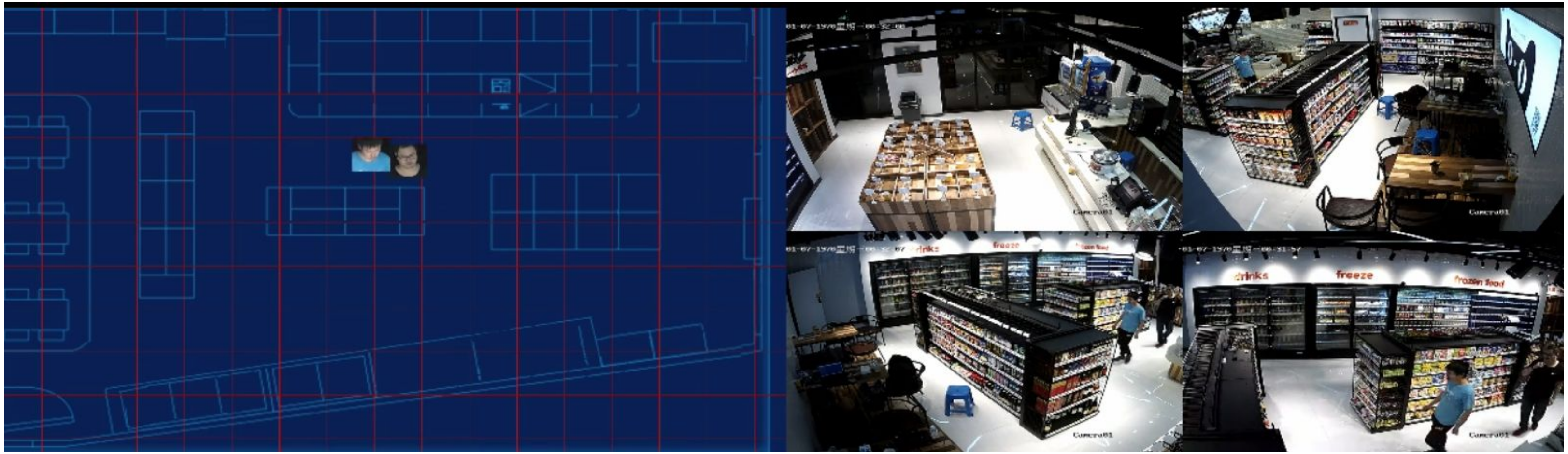


Re-IDentification: Applications

- Searching a person

Re-IDentification: Applications

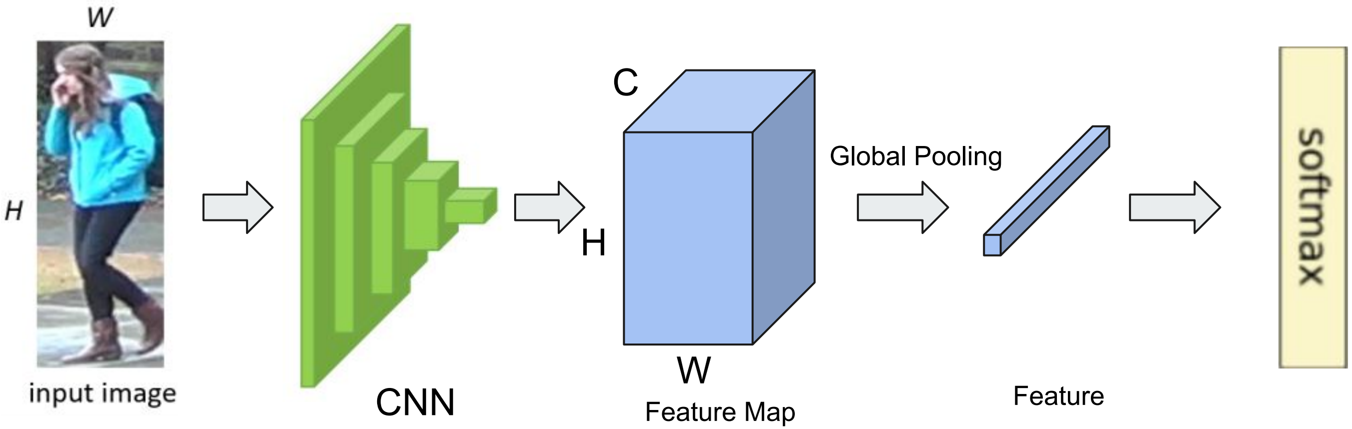
- Locating a person



Face++ 旷视

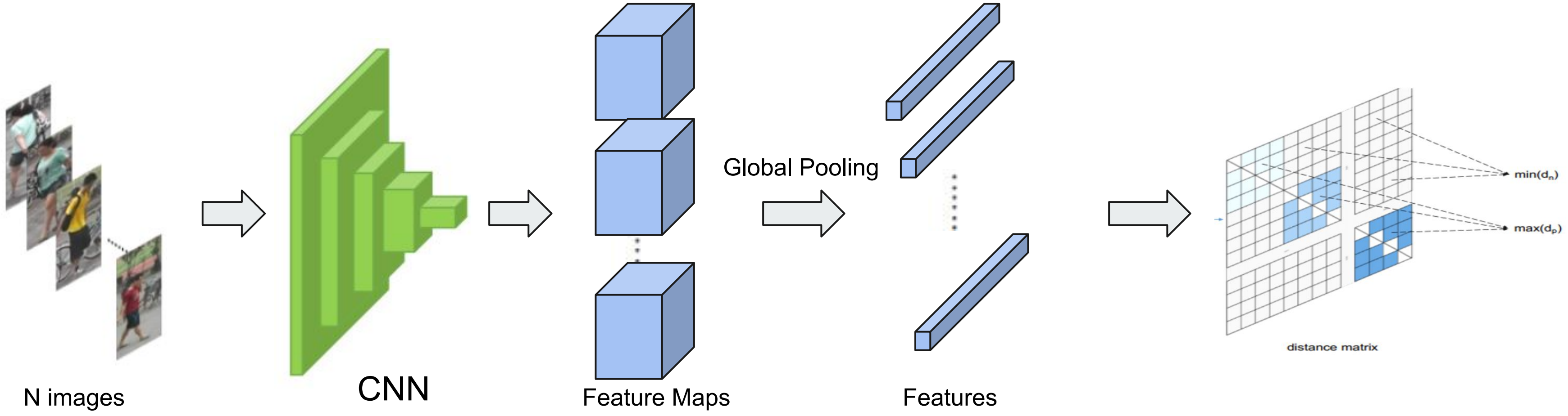
Re-Identification: Baseline

- Train ReID Model as Classification



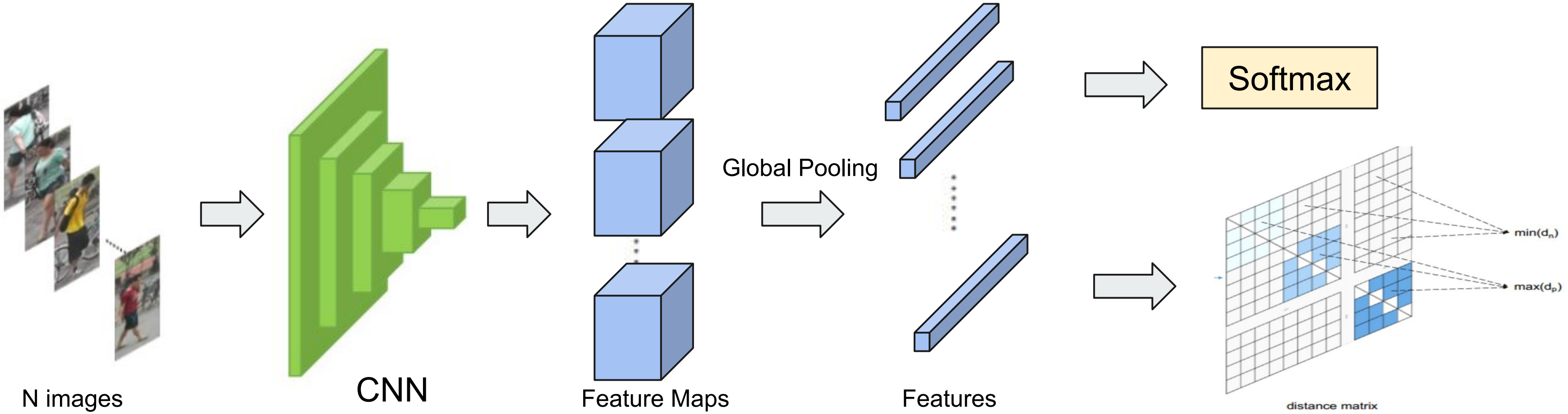
Re-Identification: Baseline

- Train ReID Model by Triplet Loss with Hard Mining



Re-Identification: Baseline

- Combing Triplet Loss and Classification



Re-Identification: Baseline

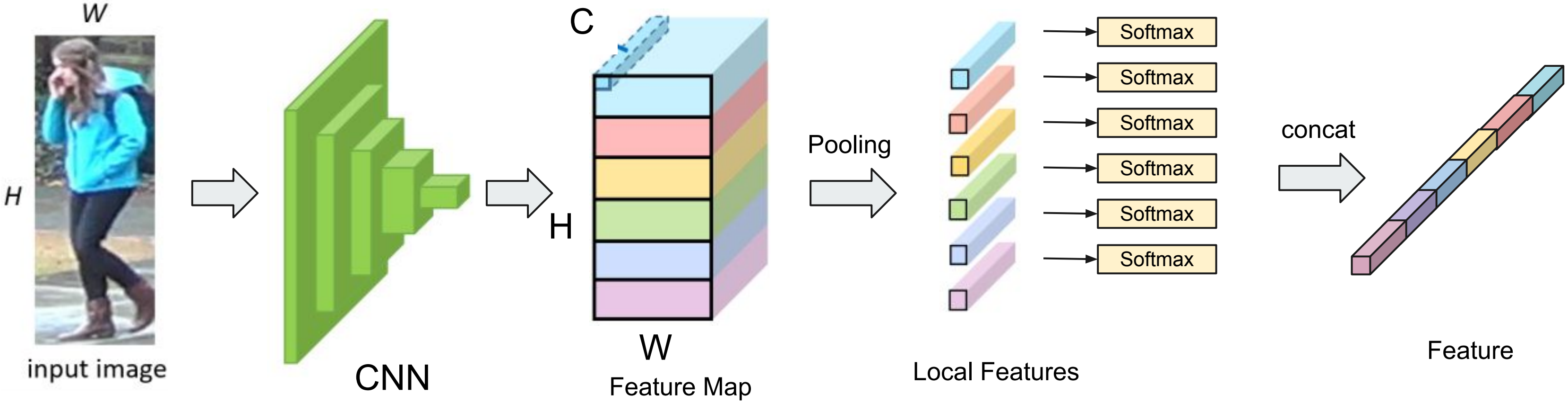
- Bottleneck is important in Classification
- Hard mining is important in Triplet Loss
- Triplet Loss usually achieves higher accuracy than classification in the same dataset
- However, Classification is more robust among different datasets
- After all, Classification with triplet loss always achieves better performance

Re-Identification: Baseline

- Disadvantage
  - Only global information is obtained
  - Local similarity plays a key role to decide the identity
- Motivations
  - Person is highly structured
  - In different views, the order of horizontal division keeps the same.

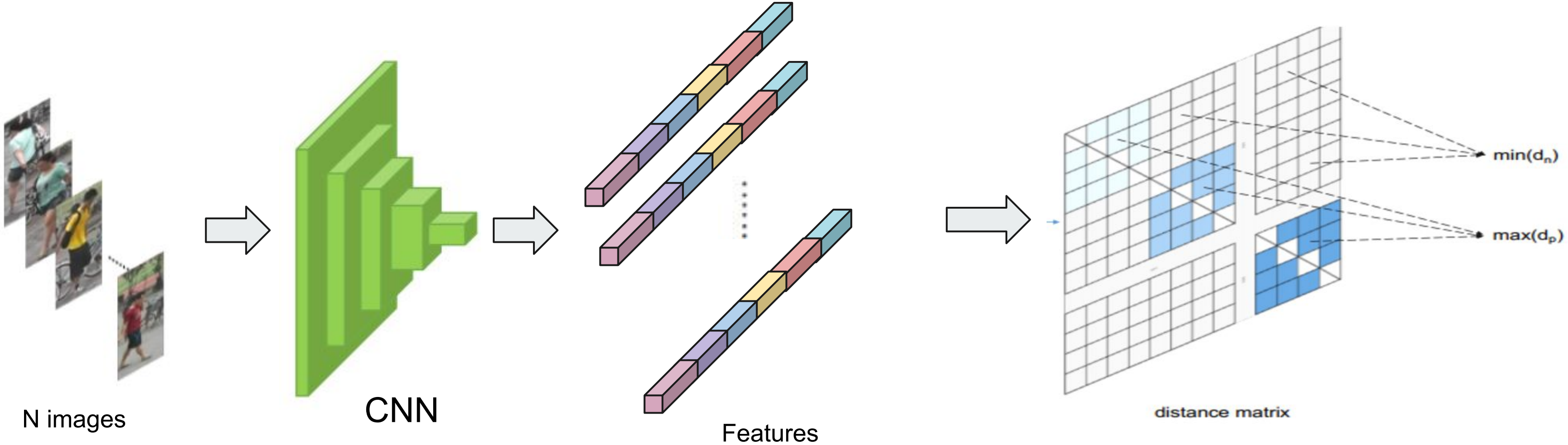
Re-Identification: Part-based Model

- Divide Feature Map to obtain local features
- Concat local features to obtain final feature



Re-Identification: Part-based Model

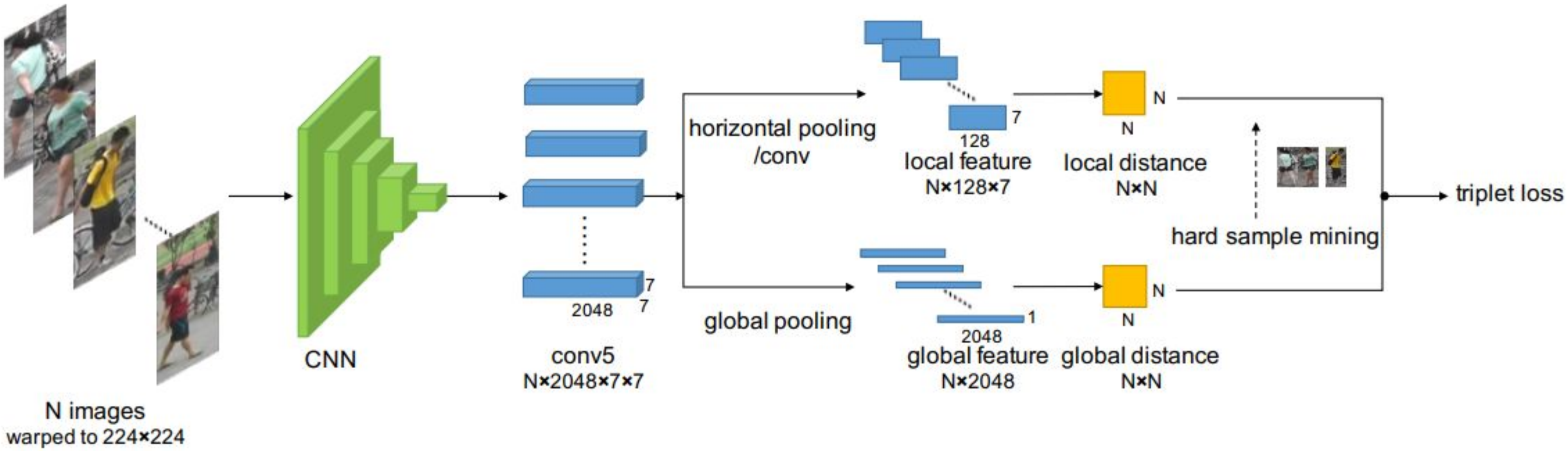
- Triplet Loss for global features



Re-Identification: Part-based Model

- Classification for the local features
    - Triplet Loss is not suitable here
- Triplet Loss with hard mining for the global feature is helpful
- Disadvantage
    - Alignment is rigid
    - Suffer from misalignment and incompletion
- Motivation
    - Automatic alignment

Re-Identification: AlignedReID



Face++ 旷视

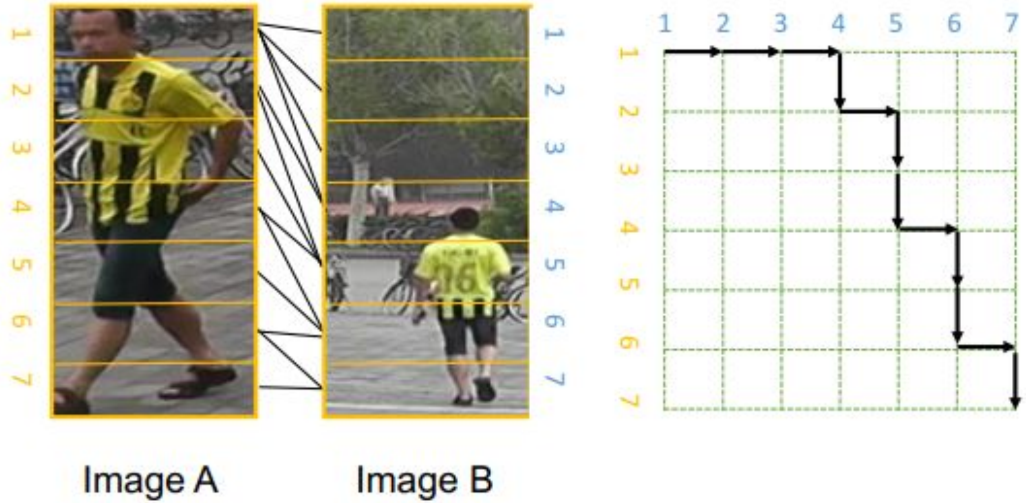
[X. Zhang et al, AlignedReID: Surpassing Human-Level Performance in Person Re-Identification, arXiv: 1711.08184]

Re-Identification: AlignedReID

- Distance matrix of local features

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

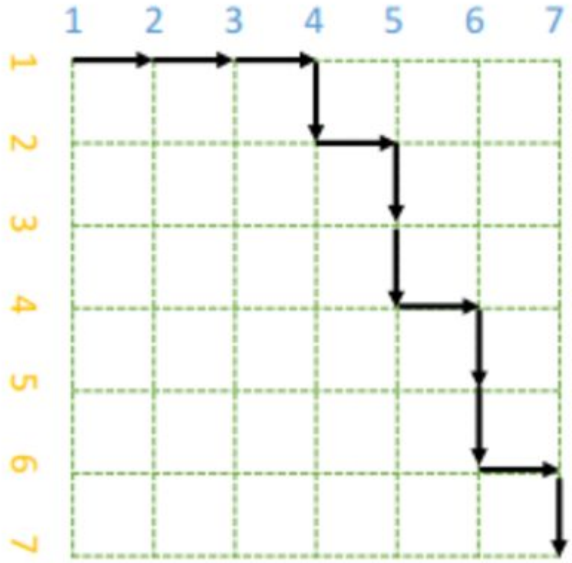
- The alignment is the one with minimum total distance



Re-Identification: AlignedReID

- Find the shortest path by dynamic programming

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



Re-Identification: AlignedReID

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

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

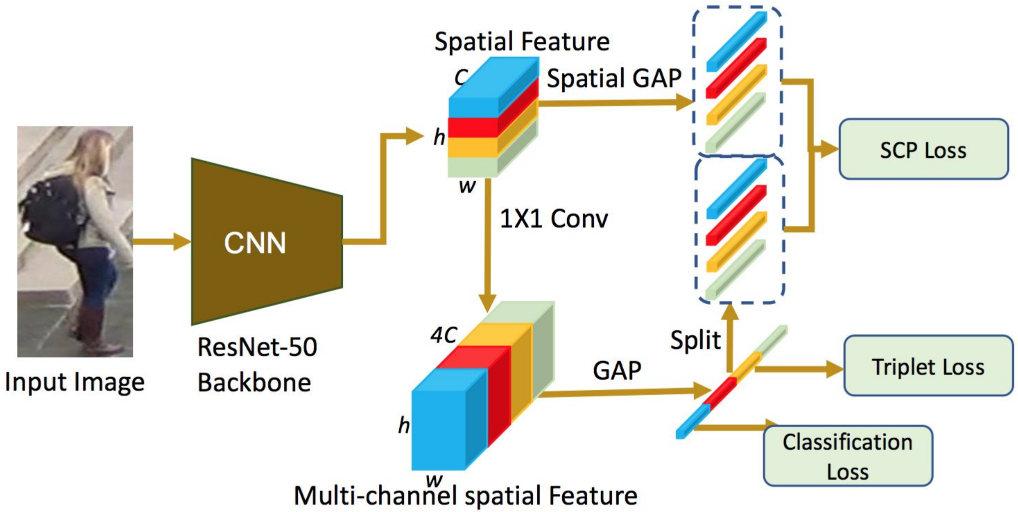
Face++ 旷视

Re-Identification: AlignedReID

- Mismatched parts have little contribution during back-propagation
- Local features help to learn a better global feature
- Disadvantage
    - Local features are obtained from small receptive field
    - Channels in the global feature has no relationship with spatial locality
- Motivation
    - Build spatial-channel relationship
    - Benefit for partial person re-identification

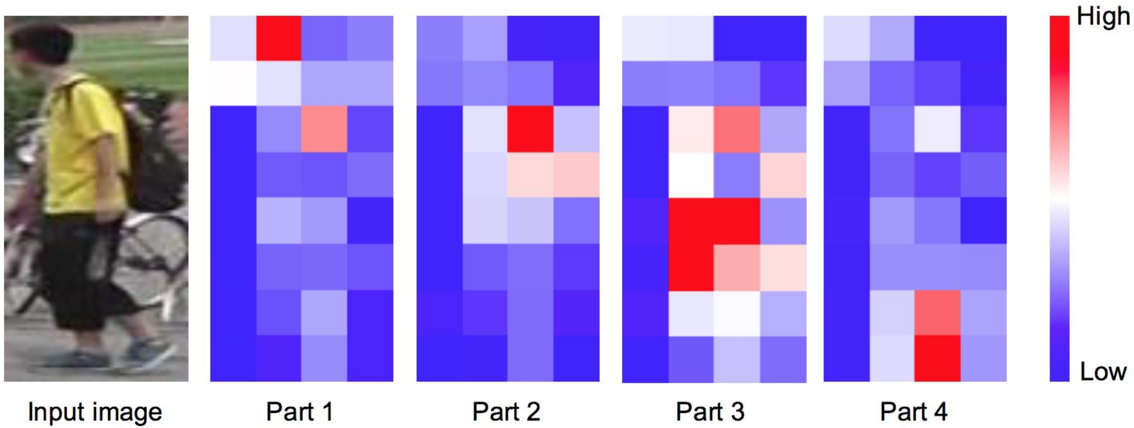
Re-Identification: Spatial-Channel Parallelism

- Local features obtained from local spatial part.
- Global feature obtained from the whole feature map.
- Each part of the global feature is related to a local feature.
- The relationship is implemented by adding their L2 distance in the loss function



Re-Identification: Spatial-Channel Parallelism

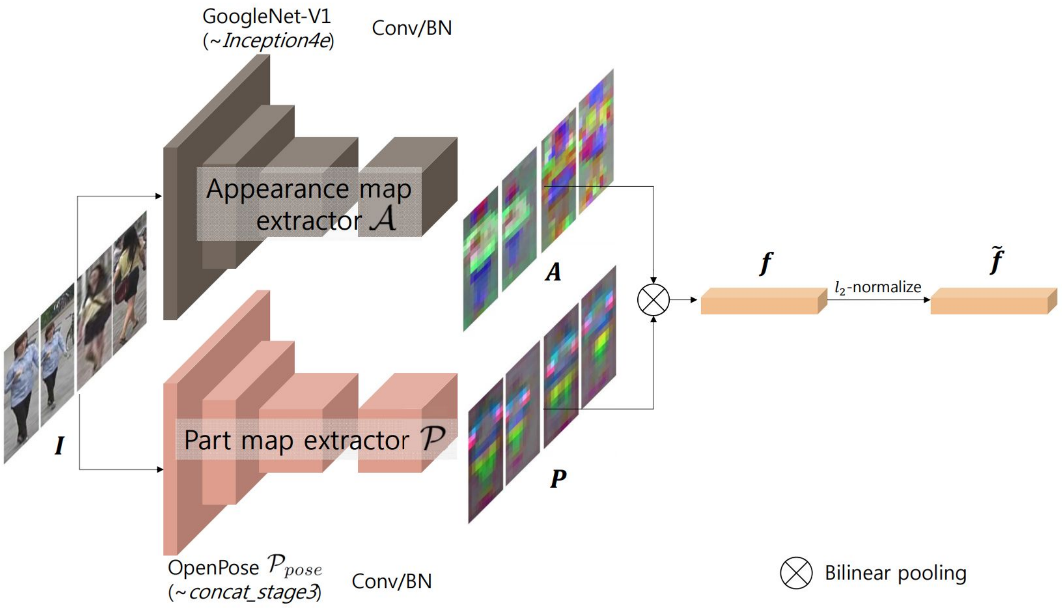
- The learned global feature shows the relationship of their channels to the corresponding spatial parts.
- Disadvantage
  - Only horizontal mapping
- Motivation
  - Apply Pose Estimation



Face++ 旷视

Re-Identification : ReID with Skeleton

- One branch is extracted reid feature map
- The other branch is extracted pose estimation
- The feature is obtained by the bilinear pooling of these two branches
- Pose estimation branch is pre-trained, then finetune in the reid training process



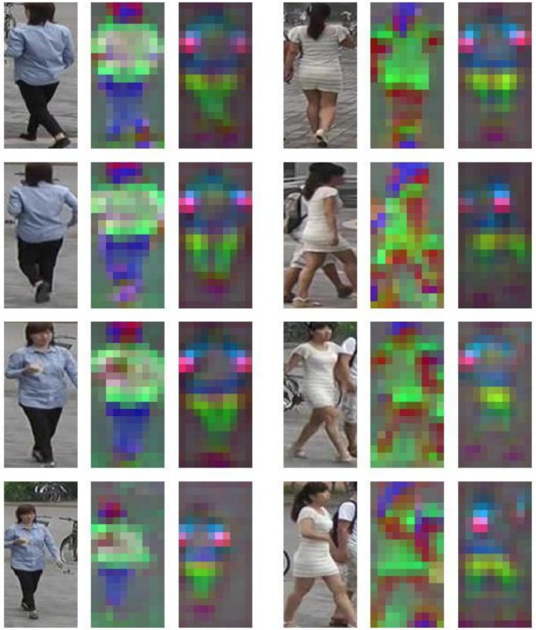
Re-Identification：ReID with Skeleton

- The reid feature maps show the similarity between color or texture, regardless of parts
- The pose estimation maps show the similarity between body parts, regardless of appearance similarity

(a) Appearance features (b) Part features

Re-Identification : ReID with Skeleton

- Similar color shows the similarity in appearance or locality
- Robust to body deformation and inaccurate detection
- Disadvantage
  - Extra training data is needed
  - Bilinear pooling is consuming
  - Accuracy is not high enough
- Motivations
  - Better pose estimation
  - Skeleton keypoints are not necessary
  - Body Segmentation may be better



Face++ 旷视

Summary

- Re-Identification can be considered as a kind of metric learning
  - Better trained together with classification
  - Triplet Loss, or its improvements, usually works well
  - Hard sample mining is critical
- End-to-end learning with structure prior is more powerful than a “blind” end-to-end learning
  - Local Feature with alignment can significantly improve the accuracy
  - The alignment can be helped by pose estimation
    - However pose estimation is not always dependable
  - The alignment can be learned automatically
- Relationship with Human Attributes
  - ReID provides more discriminative details than human attributes