Structure-Regularized Attention for Deformable Object Representation
SUPPLEMENTARY MATERIAL

A Module Implementation

The Structure-Regularized Attention (StRAttention) block is comprised of two operations, *local attention* and *mode attention*. The schema of the two operations, local attention, and mode attention, is shown in Figure 1. The input of the local attention operation is produced by a 1×1 convolution and the output of the module is processed by another 1×1 convolution when instantiating it as the drop-in replacement of a bottleneck residual block. As shown in Figure 1 in the main paper, input feature maps are dealt with local attention and then fed into mode attention. The outputs of the two operations are added through the use of skip connection.

The strategy of generating modal vectors is of importance for aggregating the information within modes. In this work the modes are expected to represent spatial structural factorization (e.g., parts or body landmarks), which is modelled in an unsupervised manner. 1×1 group convolutions equipped with softmax function generate the normalized spatial mask $\mathbf{M}^g \in [0, 1]^{H \times W}$ for each one of G modes. The modal vector is then computed by weighted summation over the output of local attention $\mathbf{S}_g$. Each mode is consequentially represented by the most representative nodes. We impose a diversity regularization on the training loss [19] to encourage modes to detect disjoint positions, forming a soft constraint for modeling diversified latent factors. The term is formulated as $\mathcal{L}_d = G - \sum_{ij} \max_{g=1,\dots,G} M_{ij}^g$, which is non-negative and the minimum (i.e., zero value) can be only achieved if disjoint positions are activated with the value 1.

B Experimental Setup

B.1 Person ReID

Database: Market1501 [22] contains $32,668$ images from $1,501$ identities whose samples are captured under 6 camera viewpoints. $12,936$ images from 751 identities are used for training and the left images (including $3,368$ query images and $19,732$ gallery images of 750 persons) are used for testing.

Configuration: We conduct all the experiments in the single-query setting without a re-ranking algorithm. The results are reported on the cropped images based on detection boxes. We report performance based on two measures: Cumulative matching characteristics (CMC) rank-1 accuracy and mean average precision (mAP). Post-processing (*e.g.*, re-ranking and multi-query fusion) is not applied for all the experiments.

ResNet-50 is used as the backbone architecture, where the last spatial downsampling operation is removed following conventional settings [18, 11, 5] and a dimensionality-reduction layer is used after average pooling layer, leading to a 512-D feature vector, analogous to [18, 9].

The model weights are initialized by the parameters of models trained on the ImageNet dataset. StRAttention variant is constructed by replacing the three residual blocks at the last stage by ours

NeurIPS 2020 Workshop on Object Representations for Learning and Reasoning, Vancouver, Canada.

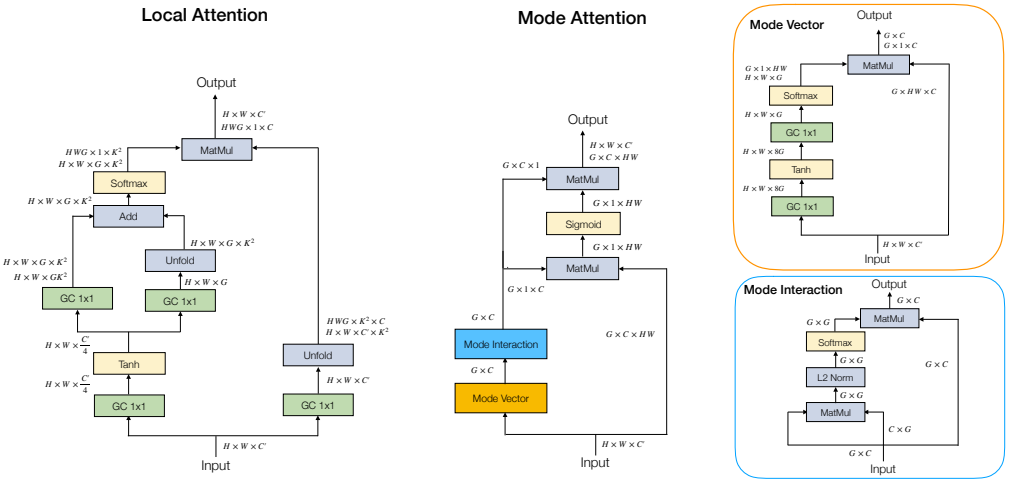

Figure 1: The schema of an StRAttention module. **Left:** *Local Attention* operation (Eq. 2 in the main paper). The two branches performing with "Add" operation correspond to the transformation ω and ν, respectively. The normalized local softmasks (i.e., affinity matrix) multiply with the output of transformation u and produce the operation output. **Right:** *Mode Attention* operation, which is comprised of mode vector unit and mode interaction unit. G modal vectors are generated by respective normalized spatial masks for each mode. The production of modal vectors and node representations passes through the sigmoid gating function which produces the attention coefficient in Eq. 3. GC denotes group convolutions with the group number set to G. C' denotes channel dimension, and $C' = C \cdot G$. H and W denote spatial dimensions. K is the local neighborhood size. Batch Normalization [7] is used after group convolutions by default. The implementation may also require reshaping or permuting operations, which are not explicitly illustrated in this figure.

where the weights are initialized randomly. Only classification loss (*i.e.*, cross-entropy loss based on identities) is used when training. The normalized (to unit ℓ_2 norm) feature vectors of query images and gallery images are compared by using the Euclidean distance metric for testing.

When training on Market1501, the parameters of the models from stage 1 to 4 are frozen at the first 8 epochs that could facilitate convergence. Images are resized to 256×128 and simply augmented by random flipping, cropping, and erasing. We use Adam [8] as the optimizer, where the initial learning rate is set to $3\mathrm{e}{-4}$, and decayed (multiplied) by 0.2 every 20 epochs. Batch size is set to 32 and weight decay is $5\mathrm{e}{-4}$. We train models for 100 epochs with two NVIDIA Tesla P40 GPUs, based on the Pytorch framework. The weight of the divergence loss $\mathcal{L}_d$ added to the objective is set to 1.0.

B.2 Face Recognition

Database: We use a collection of multiple public training datasets [21, 10, 2, 13] as the medium-size training set, and use VGGFace2 [1] to show the effectiveness of the proposed method on a larger scale dataset.

For evaluation, we apply the following verification datasets which are typically used for evaluating face models. LFW [6] contains $13,233$ face images from $5,749$ subjects collected from the website. We report the network performance following the standard *unrestricted with labeled outside data* protocol as in [6]. CFP-FP [16] dataset aims to evaluate the models when pose variation is high and extreme pose exists. AgeDB-30 [12] contains face images with high age variance. CPLFW [23] and CALFW [24] contain the same identities as LFW while focusing on the evaluation with large pose and age variation, respectively, requiring good generalization of the features extracted from networks.

Configuration: ResNet-50 [4] is adopted as the backbone network, where conventional global average pooling is replaced by an BN [7]-Dropout [17]-FC-BN module following [3], and finally produces a 512-D feature vector. The attention modules are used at the last stage of the architecture.

All the models are trained from scratch. Classification loss (*i.e.*, cross-entropy loss) is used as the objective for training. When evaluating on the test set, the feature vectors extracted from the original images and flipped ones are concatenated and then normalized for comparison. The verification accuracy is conducted with the best threshold on the Euclidean distance metric (in the range of [0,4]) following [20, 3].

All the models (including baselines and ours) are trained from scratch. Standard SGD with momentum is used for optimization. Batch size is set to 512 (on 8 GPUs, *i.e.*, 64 per GPU) and weight decay is $5\mathrm{e}{-4}$. For training on the medium-scale dataset, models are trained for 20 epochs, and the learning rate is initially set to 0.1 and multiplied by 0.1 at the 8-th and 12-th epochs. Models are trained for 50 epochs on VGGFace2 dataset [1], and the learning rate is initially set to 0.1 and multiplied by 0.1 at the 20-th,30-th,38-th,44-th and 48-th epochs.

For both training and evaluation sets, images are pre-processed by following standard strategies [20], *i.e.*, detecting face area and then aligning it to canonical views by performing similarity transformation based on five detected landmarks. Models are trained with center crops (the size is 112×112) of images whose shorter edges are resized to 112 on the medium-scale dataset. Models trained on VGGFace2 are based on inputs first resized to 224×192. Each pixel is subtracted 127.5 and divided by 128 for normalization. Only random horizontal flipping is used as data augmentation during training. The weight of the divergence loss $\mathcal{L}_d$ added to the objective is set to 0.1.

B.3 Facial Expression Recognition

Database: The Facial Expression Recognition 2013 (FER2013) database contains $35,887$ images. The dataset contains $28,709$ training images, $3,589$ validation (public test) images, and another $3,589$ (private) test images. Faces are labeled as any of the seven expressions: "Anger", "Disgust", "Fear", "Happiness", "Sadness", "Surprise" and "Neutral".

Configuration: Images are resized to 44×44 and random horizontal flip is adopted for training. All the models (including baselines and ours) are trained from scratch by using the classification loss. The baseline architecture is a variant of bottleneck residual network, where the kernel size and the stride of the first convolutional layer is set to 3 and 1 and followed by BN and ReLU units, max-pooling layer is omitted, and each of the following four stages is comprised of 3 blocks. We implement the StRA variant by applying the modules at the last stage of the baseline network. The last 1×1 convolution, which typically fuses feature maps across channels, is omitted in order to facilitate the understanding of behavior among modes. The weight of the divergence loss is set to 1. Batch size is set to 128 (on single GPU). We use SGD with momentum 0.9 for optimization. Weight decay is set to $5\mathrm{e}{-4}$. The learning rate is initially set to 0.01 and decayed by 0.9 every 5 epochs after 80 epochs, following [14]. Models are trained for 190 epochs in total.

C Ablation Study on Module Configuration

We conduct ablation studies on Market1501 as Table 1 in the main paper. The StRAttention block consists of two components, i.e., Local Attention which integrates information over spatially-adjacent regions, and Mode Attention which models the long-range contextual relationships in a node-to-mode manner where mode interaction is exploited to allow mode interactions.

Table 1: Ablation studies on module components.

Model	Component			mAP	Rank1
	Local Attn.	Mode Attn. w/o Interact.	Mode Attn.		
ResNet50				77.1	90.6
StRAttention	✓			79.9	92.3
	✓	✓		83.3	93.4
	✓	✓	✓	**84.1**	**93.8**

Table 2: Ablation studies on Mode Attention.

Method	mAP	Rank1	Params	FLOPs
Conv	77.1	90.6	24.6M	4.05G
SASA [15]	79.5	92.3	17.8M	3.19G
Conv + Mode	82.1	93.2	24.6M	4.06G
Group Conv + Mode	82.8	93.3	18.4M	3.26G
SASA + Mode	83.0	93.6	17.8M	3.19G
Local + Mode (ours)	84.1	93.8	17.6M	3.17G

We first assess each component and the results are shown in Table 1. Using our Local Attention can simply yield obvious performance improvement over the baseline ResNet-50, by 2.8% on mAP and 1.7% on Rank-1 score. Incorporating Mode Attention without mode interaction can further boost performance. Using the default configuration is able to push the performance further, demonstrating the necessity of all the components in the module. We can conclude that each component plays an important role, and accumulated benefits can be achieved by combining them.

To validate the effectiveness of the proposed mode attention, we conduct ablation studies that replace Local Attention with convolutions, group convolutions and SASA [15]. Results in Table 2 show that replacing the local attention module by classical convolutions or self-attention layers can still obtain superior performance over the counterparts without mode attention, demonstrating the generalizability of the proposed structure regularized paradigm on representation learning.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[2] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pages 768–783. Springer, 2014.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[5] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019.

[6] Gary Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. *Tech. rep.*, 10 2008.

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[11] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019.

[12] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

[13] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.

[14] Zhenyue Qin and Jie Wu. Visual saliency maps can apply to facial expression recognition. *arXiv preprint arXiv:1811.04544*, 2018.

[15] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019.

[16] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[18] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018.

[19] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5916–5925, 2017.

[20] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[21] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[22] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[23] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.

[24] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.