Structure-Regularized Attention Networks

Anonymous ECCV submission

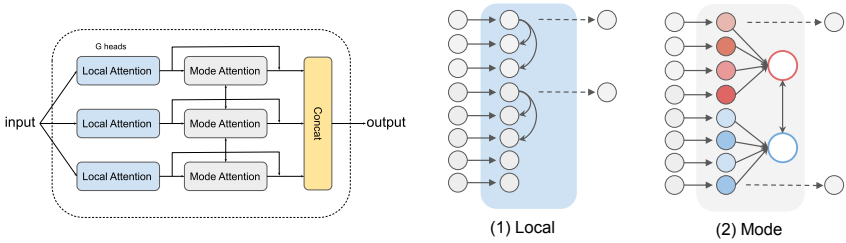
Paper ID 3180

Abstract. Capturing contextual dependencies has proven useful to improve the representational power of deep neural networks. Recent approaches that focus on modeling global context, such as self attention and non-local operation, achieve the objective by enabling unconstrained pair-wise interactions between elements. In this work, we consider the visual tasks which may require or highly benefit from the introduction of structural priors. Inspired from feature subspace algorithms, we propose a new attention mechanism that formulates feature interactions in a structural manner. We develop it as a network module and evaluate the proposed method on person re-identification and face recognition tasks. Comprehensive studies and empirical comparisons with the state-of-the-art attention networks demonstrate its effectiveness on both of performance and model complexity. We further investigate the effect of the mechanism on network representations, showing that discriminative patterns corresponding to different parts can be effectively captured and enhanced by the method without the need of extra supervision.

Keywords: Attention Module, Structural Factorization, Representation Learning

1 Introduction

Attention is capable of learning to focus on the most informative or relevant components of input and has proven to be an effective approach for boosting the performance of neural networks in a wide range of tasks [14, 19, 50, 52]. Self-attention [50] is an instantiation of attention which weights the context elements by leveraging pairwise dependencies between the representations of query and every contextual elements. The ability of exploiting the entire context with variable length has made it successfully integrated into the encoder-decoder framework for sequence transduction. [52] interprets it from another perspective, i.e., as non-local means [2], and adapts it to convolutional neural networks for visual applications. However, it is computationally expensive where the complexity is quadratic to input length (e.g., spatial dimensions for image and spatial-temporal dimensions for video sequence). The approach captures long-range association within global context by allowing each node (e.g., a pixel on feature maps) to attend over every other positions, forming a complete and unconstrained graph which may be intractable for extracting informative patterns in practice. Relative position embedding has proven useful in alleviating the issue by taking positional



(a) Overview of StRAttention module (b) The concept of the two attentions.

Fig. 1: **Left:** Overview of the StRAttention module and **Right:** The concept of local attention and mode attention. Local attention is designed to integrate the information from spatially adjacent regions. Mode attention, as the complementary, captures the long-range context in a structure manner (the nodes with red and blue colors refers to two modes here).

information into account [1, 43] but the structural information specific to tasks is not effectively exploited.

In this work, we consider the visual tasks which may require or highly benefit from the modeling of structural dependencies, e.g., part-based modeling for human bodies [10, 62]. Our goal is to explore the attention mechanism which is efficient to capture such structural dependencies. To this end, we introduce a simple but efficient attention module which we term the "Structure-Regularized Attention" (StRAttention). It is formalized as the composition of two-level operations, namely local and mode attention, to model feature correlations between nodes. The fundamental of the two operations is illuminated in Figure 1. The local attention, functioned as spatial expansion on local regions, captures informative patterns by virtue of pairwise relationships between neighbor nodes. The higher-level contextual information can be accessed through the mode attention that enhances the diversity of features. The mechanism enables each node to attend to (theoretically) global context in a structural manner. The basic structure of the StrAttention module is depicted in Figure 2.

Inspired from subspace algorithms [37], the design is built upon the hypothesis of data factorizability, i.e., projecting the data into multiple feature subspaces, defined as modes, which are expected to be more compact and typically represent certain component (e.g., associate with semantic concepts) of the data. A set of parameterized transformations project nodes into multiple modes that perform to capture diversified and discriminative representations. We deploy the method into person re-identification and face recognition tasks in which the module tends to capture diversified patterns referring to different parts in an end-to-end trainable and unsupervised way.

In summary, the main contribution of the work is that we propose a new attention mechanism which allows the capture of long-range dependencies effectively through a structural manner. As another contribution, we present a formulation of local attention which is simple and efficient. Detailed descriptions

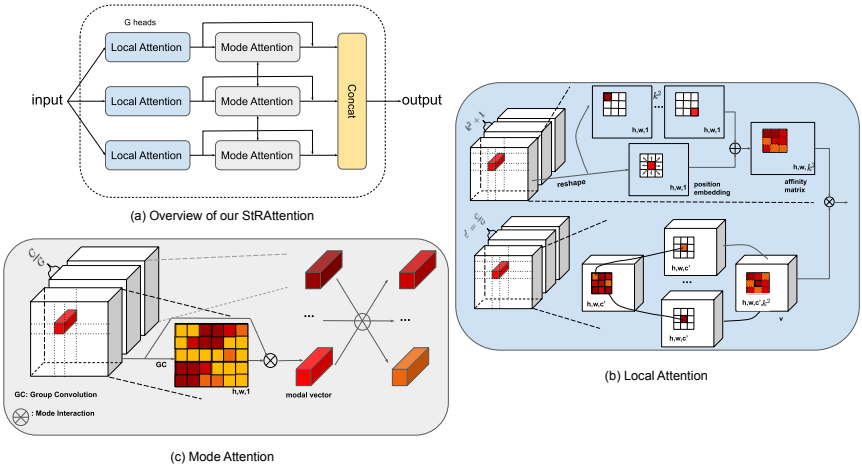


Fig. 2: The illustration of the Structure-Regularized Attention Block. (a) Basic structure of the block with G heads (modes). (b) Local attention operation. (c) Mode attention operation. More details can be found in Sect. 4.

for the method as well as discussions are present in the Section 3 and Section 4. We then conduct a series of experiments to validate the approach and compare it with the state-of-the-art attention methods across multiple datasets in Section 5 so that a comprehensive understanding can be provided for the method. In order to further understand the behavior of the module, we investigate its effect on network representations, showing that the mechanism can effectively capture and enhance the distinct patterns describing different parts automatically without the need of extra supervision.

2 Related Work

Deep architectures. Convolutional neural networks [23] have proven to be powerful models for representation learning. Convolutions, as the fundamental modules in networks, function as linear mappings for combining local spatial connectivities along channels. Many researches are proposed to improve network capability by tuning the form of modules. 1×1 convolutions are used to increase the nonlinearity by learning combinations between channels [29]. Residual blocks [12] have shown the advantages of learning very deep networks by virtue of skip connections. ResNext extends the axis of module design that increases the model capacity through the use of group convolutions [54]. Multi branch convolutions comprised of compositional operations in [46] generalize the concept of group convolutions to a flexible manner. Depthwise convolutions [7] have shown effectiveness for learning efficient networks. Capsule reformulates the unit of activation as a group of neurons whose activity vectors and association with lower-level capsules are jointly inferred in an iterative procedure [39, 13].

Attention mechanisms. Attention mechanisms are effective to boost discriminative ability of neural networks by enabling model learning to mostly focus on the important components of data. Spatial attention [20, 55] encourages the use of the information from informative regions of inputs. The work [14] re-weights the channels of convolutional features by encoding the information from the full extent of inputs. Self attention mechanisms have shown the effectiveness of modeling long-range relationships for sequence learning [50]. The work [52] interprets it from the perspective of non-local mean and the proposed non-local modules capture the pair-wise relationship between positions to achieve global context through a fully connected graph. Some works are proposed to address the issue of computational cost and memory overhead when performing on the complete graph. [17] stacks criss-cross operations to achieve global access. [27] formalizes it as an iterative estimation of bases and assignment. Both methods show the benefits on semantic segmentation. Self attention operations are adapted into local regions as the substitute of convolutions in [38].

Graph-based modeling. Graphical models have proven to be a popular tool for modeling relationships. Conditional random fields (CRF) [24] is effective for boosting robustness and performance in segmentation tasks [4]. Graph convolutions have been successfully used for capturing long range relations between the nodes defined by parts [56], objects [28] or distinct regions [6] in a wide range of vision tasks.

Part models. The structure of objects are typically described by the composition of discriminative parts. Deformable Parts Model [10] is a very successful instantiation for modeling parts in an unsupervised manner. Some methods learn part oriented representations and have shown the effectiveness for dealing with deformable objects (such as face [57], human body [26, 31]) where extra manual annotations or rules induced from prior knowledge are typically introduced for supervision. [48] proposes to capture distinct parts (landmarks) in an unsupervised way.

3 Context Modeling by Structural Factorization

We will use "pixel" and "node" interchangeably in the following descriptions. Formally, let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote an input, e.g., feature maps of an image, with spatial dimensions $H \times W$ and channels C, and $\mathbf{x}_i$ denote the feature on the pixel x_i, where $i \in \mathcal{N}_G \equiv \{1, \cdots, H\} \times \{1, \cdots, W\}$. The context feature is captured by virtue of the transformation $f : \mathbb{R}^C \times \mathbb{R}^C \to [0, 1]$ [50, 52],

$$\mathbf{y}_i = \sum_{j\in\mathcal{N}_G} f(\mathbf{x}_i, \mathbf{x}_j)u(\mathbf{x}_j), \tag{1}$$

where u represents the unitary transformation on single node and f captures the pair-wise correlation between nodes within the global context. f forms a complete graph that each node can attend to every other node and global context is consequently able to be directly accessed at each position. It brings about the

challenge of quadratic computational complexity and memory overhead with respect to the size of $\mathcal{N}_G$ [17, 27]. More importantly, the structural prior the data has is not fully exploited.

The use of hierarchical structure is believed to play a critical role for capturing the statistics in images independent of learnable parameters [49]. In reality, most data (e.g. deformable objects) can be assumed to live on low dimensional manifolds. Our goal is to encourage the nodes to interact in a structure-regularized manner so that information can be efficiently delivered by taking advantage of the natural characteristics of data.

To this regard, we formalize the problem as a form of structural factorization. We want to learn a set of transformations to project the data onto multiple diversified subspaces, $\Phi := \{\Phi_g\}_{g=1}^G$ where $\Phi_g : \mathcal{X} \to \mathcal{S}_g$ corresponds to the projection onto the g-th subspace which we call "mode" here. The corresponding output for node x_i is $\mathbf{s}_i^g$. Each mode is expected to represent a certain factor the data consists of and the contextual information with respect to each node can be effectively modeled by virtue of these modes. Suppose we have the values of the models $\mathbf{Z} = \{\mathbf{z}_g\}_{g=1}^G$ (e.g., centroid vector), $r_{ig} \in [0, 1]$ indicates the degree of the node x_i correlated to the g-th mode. The context for the node x_i can be formulated as a combination of the information derived from each mode $\mathbf{y}_i := \bigcup \mathbf{y}_i^g$, and

$$\mathbf{y}_i^g = r_{ig} \cdot \mathbf{z}_g, \quad r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g), \tag{2}$$

where the information between modes can be further correlated through a function $\rho : \mathcal{Z} \to \mathcal{Z}$ that captures the relation between modal values and propagates such higher-level context to each node.

4 Structure-Regularized Attention

We instantiate the method of section 3 concretely with a two-level attention mechanism which is comprised of local attention and mode attention. Local attention projects inputs onto a set of feature subspaces and simultaneously capture local correlation within neighborhoods. Mode attention is responsible for modeling the relation between nodes and modes as well as higher-level relation between modes.

4.1 Local Attention

It has been shown that self-attention performing on local neighborhood is a comparative alternative of convolution [38], which is formulated as

$$\begin{aligned} \mathbf{s}_i &= \sum_{j \in \mathcal{N}_K(i)} a_{ij} u(\mathbf{x}_j), \\ a_{ij} &= \sigma_m \left(q(\mathbf{x}_i)^T k(\mathbf{x}_j) + q(\mathbf{x}_i)^T \mathbf{r}_{j-i} \right). \end{aligned} \tag{3}$$

σ_m denotes the softmax function, q, k and u denote the unitary transformations (e.g., linear mappings). The affinity matrix $A_i = \{a_{ij}\}_{j \in \mathcal{N}_K(i)} \in [0, 1]^{K \times K}$ is

produced by the sum of two terms. The first captures the relation between the node x_i and its spatially-close neighbors $x_j \in \mathcal{N}_k(i)$, i.e., a $k \times k$ local region. The second term is used to supplement the lack of position information by introducing learnable relative position embedding $\mathbf{r}_{j-i}$, where $j - i$ denotes the relative distance of x_i with respect to x_j. In other word, the affinity matrix A_i is expected to generate a proper *data-dependent* local softmask with limited size $K \times K$ at each node x_i for local context aggregation.

Here we propose a simplified variant inspired by the dynamic convolution in [53] which is designed for sequence-to-sequence modeling. Formally, the affinity matrix A_i is predicted by the transformation on x_i,

$$a_{ij} = \sigma_m \left(\omega(\mathbf{x}_i)_j + \nu(\mathbf{x}_j) \right), \tag{4}$$

where the transformation $\omega : R^C \to R^{K \cdot K}$ and $\nu : R^C \to R$ are associated with a set of learnable parameters and the output is used as the logits of softmax-normalization. The relative importance of neighbors with respect to the target node x_i is predicted by ω which can also encode spatial layout. The contribution from the neighbor itself is derived from ν. Compared to the formulation (3), it also expresses the neighborhood spatial relationship in pairs while though an implicit but very efficient manner.

We introduce the multi-head paradigm [50, 38] to project inputs onto multiple modes. If G projections are expected to learn, the function on g-th mode is comprised of the transformation ω_g and ν_g and outputs $\mathbf{S}_g = \{\mathbf{s}_i^g\}$ for $\mathbf{X}$.

Discussion. We refer the local attention mechanism as capturing correlations in local regions, which functions as convolutions. Compared to linear summation over neighbor nodes in convolutions, local attention expresses the patterns in a second-order manner though both of them build upon local connectivities. When performing convolutions, identical weights are shared over different positions. In contrast, local attention generates data-dependent weights at each position that consequently introduces dynamics into networks. This motivation shares some similarity with dynamic filters [21] while the work generates the weights conditioned on extra inputs which are encoded in a separate encoder-decoder network trunk.

4.2 Mode Attention

Local attention expresses the patterns within local region which is insufficient to capture the correlations from larger (or theoretically global) receptive fields. The goal of mode attention is to supplement the ability in an efficient manner by virtue of structural factorizability intrinsically exists on data. For example, a deformable object can be effectively described by a combination of representations towards different parts [10]. In this regard, we expect each mode responsible for one distinct component of representation distribution. The intrinsic properties (e.g., statistics or centroids) of each mode are represented by the modal value. We use the parameterized transformation $\xi_g : \mathbf{S}_g \mapsto \mathbf{z}_g$ to generate a feature vector for the g-th mode. The attention coefficient r_{ig} in (2) is then measured

by using the similarity between the modal value and the corresponding feature vector for the node x_i,

$$r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g) = \sigma(\langle \mathbf{s}_i^g, \mathbf{z}_g \rangle). \tag{5}$$

σ denotes the gating function which can be defined with either softmax or sigmoid. The first assumes the coefficients satisfy with multinomial distribution that will encourage selecting one of the G modes and the second measures the correlation over different modes independently. Both forms can achieve the goal that the nodes correlated to the same mode will share some context to enhance their desired representations.

Furthermore, higher-level correlation can be conveniently modeled through $\rho : \mathcal{Z} \to \mathcal{Z}$,

$$\mathbf{z}'_g = \sum_{j=1}^{G} \sigma_m(\langle \mathbf{z}_g, \mathbf{z}_j \rangle) \cdot \mathbf{z}_j. \tag{6}$$

The updated $\mathbf{Z}$ through across-mode interactions can substitute the one in (5) to generate the contextual feature which is complementary to the output of local attention.

Discussion. The design of correlating nodes to multiple modes is related to soft-clustering and mixture models [33] which learn clusters by updating central vectors and node assignments iteratively through Expectation-Maximization algorithm [8]. Such iterative process is substituted by forward and backward propagations in the framework where the associated parameters are learned by gradient descent. During inference the modal vectors and the attention coefficients are computed once, which is more efficient and suitable for neural network paradigms.

4.3 Module Instantiation

We instantiate the module as the replacement of a bottleneck residual block [11] which is comprised of a 1×1 convolution for dimensionality reduction, a 3×3 convolution for spatial expansion and a 1×1 convolution for dimensionality increase, where the 3×3 convolutions are replaced by the operations of local and mode attention.

Local Attention. The formula in (4) is implemented with two branches. One branch is used to directly predict the $K \times K$ logits conditioned on each position. The other one simply generates a feature map with spatial dimension $H \times W$ and unfolds it to $K \times K$ feature maps. We use two-layer 1×1 group convolutions equipped with nonlinearity to implement the transformation. The schema of the local attention block is shown in the Supplemental Material.

Mode Attention. 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). The prior can be simply introduced through extra supervision (e.g., part or landmark detection, pose estimation) while it may restrict the method. We

instead model it 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 feature vector representing the modal value is then computed by weighted summation over $\mathbf{S}_g$ (in section 4.2). The schema of mode attention is shown in the Supplemental Material. Each mode is represented by the most representative nodes. We impose a diversity regularization to the training loss [48] to encourage different modes to detect disjoint positions, forming a soft constraint for modeling parts. The term is formulated as $\mathcal{L}_d = G - \sum_{ij} \max_{g=1,\ldots,G} M_{ij}^g$. The regularization term is non-negative and the minimum (i.e., zero value) can be only achieved if disjoint positions are activated with the value 1. The contextual features induced by mode attention is added on the outputs of local attention to boost representation discriminability.

5 Experiments

To validate the effectiveness of the proposed Structure-Regularized Attention, we conduct the experiments on two kinds of widely studied structural data in the community: human body and human face. In the following experiments, we will focus on two tasks: person re-identification and face recognition across multiple datasets.

5.1 Person re-identification

Datasets. We evaluate our method on three widely used person ReID datasets including Market1501 [63], CUHK03 [25], DukeMTMC-reID [67]. Market1501 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. CUHK03 dataset contains $13,164$ images from $1,467$ identities. We report the results following the training/testing protocol proposed in [68]. DukeMTMC-ReID dataset contains $16,522$ training images from 702 identities, $2,228$ query images of the left 702 identities and $17,661$ gallery images.

We conduct all the experiments in single-query setting without re-ranking algorithm [68] due to the consideration of efficiency. For Market1501 and CUHK03, the results are reported on the cropped images based on detection boxes. We evaluate the performance based on two measures: cumulative matching characteristics (CMC) scores and mean average precision (mAP). Post-processing (*e.g.*, re-ranking and multi-query fusion) is not applied for all the experiments.

Implementation Details. We use ResNet-50 [11] as the backbone architecture and replace the three building blocks at the last stage with our modules. The output of the module is then fed into global average pooling and a 2048-D feature vector is produced. We use the euclidean distance between query images and

Table 1: Performance (%) comparisons with the state of the art methods. The listed methods are categorized into 3 groups: **group 1** contains methods using additional data; **group 2** represents part based methods and methods in **group 3** adopt attention mechanism. The three groups are divided by horizontal lines. The last row is our results.

		Market1501		CUHK03		DukeMTMC-reID	
	backbone	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SPReID [22]	Inception-V3	76.6	90.8	–	–	63.3	80.5
PGFA [34]	ResNet-50	76.8	91.2	–	–	65.5	82.6
PSE_ECN [41]	ResNet-50	80.5	90.4	–	–	75.7	84.5
DPFL [5]	Inception-V3	73.1	88.9	37.0	40.7	60.6	79.2
VCFL[30]	GoogLeNet	74.5	89.3	55.6	61.4	–	–
AlignedReID [60]	ResNet-50	79.3	91.8	–	–	–	–
AOS [16]	ResNet-50	70.4	86.5	43.3	47.1	62.1	79.2
HA-CNN [26]	–	75.7	91.2	38.6	41.7	63.8	80.5
DLPA[62]	GoogLeNet	75.7	91.2	38.6	41.7	63.8	80.5
DuATM[44]	DenseNet-121	76.6	91.4	–	–	64.6	81.8
MltB + L_id [58]	ResNet-50	79.0	91.6	57.6	58.5	65.8	80.7
0urs	ResNet-50	**82.6**	**93.3**	**58.5**	**58.9**	**67.5**	**83.6**

gallery images for testing. Classification loss (*i.e.*, cross entropy loss) based on the labels of identities are used in the training stage. By following a conventional setting [64], the networks of the baseline and our attention counterparts are initialized with the parameters pre-trained on ImageNet. More training details and hyperparameter setting can be found at Supplementary Material.

Comparison with the SOTAs. We compare our method with the state-of-the-art (SOTA) methods on the above three datasets. We classify current SOTA deep models into three categories, *i.e.*, part (mainly refers to stripes and grids) based models, attention based models and the models benefited from additional supervision or datasets. The proposed method can be categorized to an attention based network without extra supervision.

The performance comparison on the three datasets is shown in Table 1, demonstrating that our method can achieve competitive performance on all the three datasets. It is to note that compared to the methods training only with the data and identity labels provided by the dataset, better performance is likely to be achieved by introducing extra datasets or supervision including image synthesis [31, 47], models from other tasks (*e.g.*, human pose estimation [31, 61] or recognition of human attributes[30]). Instead, the structure prior is embedded by our method in an end-to-end trainable and unsupervised manner.

model	mAP	Rank1	Flops
ResNet50 [11]	74.3	88.8	4.05G
SASA [38]	77.5	90.2	3.19G
SENet [14]	77.9	91.5	4.49G
Non-local [52]	80.2	91.9	7.28G
ResNet50_StRA	**82.6**	**93.3**	**3.17G**
mnetv2 [40]	71.7	88.7	**0.37G**
mnetv2_StRA	74.2	89.3	**0.37G**
mnetv2(1.4) [40]	72.5	89.0	0.68G
mnetv2(1.4)_StRA	**74.6**	**89.9**	0.72G

Table 2: Comparison on Market1501. *mnetv2* is the abbreviation for mobilenetv2 [40]. Models ending with *StRA* stands for integrating StRAttention.

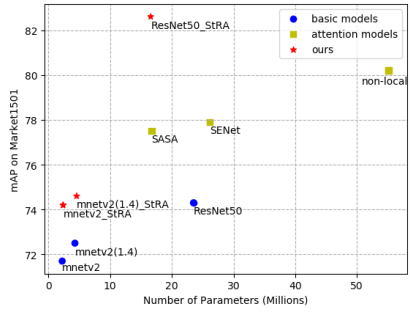


Fig. 3: **Model size** and **mAP** comparison on Market1501. Compared to the counterparts, Our StRAttention module can boost mAP while keeping the model small and efficient.

Comparison with modern attention networks. We compare our method with three modern attention networks. Non-local networks [52] and SENets [14] are two widely-used attention models which integrate the modeling of global information. We follow the network configuration in the original papers. SASA [38] is an instantiation of local attention (in Eq.3) which has shown good performance on general object recognition. We implement it by inserting the attention blocks at the last stage of ResNet-50, following the implementation in the section 5.1. Same experimental setup is used for all the experiments.

We compare the methods in terms of performance and model complexity. The results in Table 2 show that our method outperforms the baseline and other attention networks by a large margin on mAP and rank-1 metrics with the highest efficiency (i.e., Flops). We also compare the model size versus mAP of different methods in Fig.3. Benefiting from the use of structural factorization, the proposed attention module can achieve much higher mAP with comparable parameter size over SASA and much less parameters over SENets and Non-local networks. Our method achieves the best trade-off between performance and model complexity compared to the other attention methods.

Evaluation on efficient architecture. To verify the robustness of our method, we insert the modules into the mobilenetv2 [40] which is a representative architecture for mobile setting. We use the configure of backbone with the width multiplier setting as 1 and 1.4 for the experiments. The counterparts of integrating the proposed attention module is implemented by inserting the modules at the last stage.

Model complexity is the critical factor for evaluating mobile-setting architectures. The results in Table 2 demonstrates that our attention module is capable of boosting the performance while keeping efficiency. Figure3 verify the superiority

Table 3: Ablations on Market1501 dataset for person re-identification

(a) **Module components**: Effect of each component. *Local attention* indicates using the unit defined in Eq. 4 only. *Mode Attention/wo* indicates the variant of using mode attention without mode interaction.

	mAP	rank1
resnet50 baseline	74.3	88.8
Local Attention	77.8	90.3
Mode Attention/wo	80.3	91.5
StR Attention	**82.6**	**93.3**

(b) **Gating function**: softmax and sigmoid function at *mode attention* module, which controls the propagation of contextual information.

gating function	mAP	rank1
sigmoid	**82.6**	**93.3**
softmax	80.1	91.1

(c) **Kernel size**: different kernel size in the *local attention* module

kernel size	mAP	rank1
3×3	**82.6**	93.3
5×5	82.0	**93.4**
7×7	80.4	91.9

(d) **Head number**: the number of modes (subspace projections) expected to learn.

head number	mAP	rank1
4	79.4	91.1
8	**82.6**	**93.3**
16	80.2	92.1
32	79.6	91.5

of our model which is able to achieve higher performance in a computationally efficient and light-weight manner. It further validate the motivation of the module design that taking advantage of structural dependency into the representation learning could enable features to interact in an effective and efficient manner, facilitating the delivery of contextual information in networks.

Ablation study on module components. The StRAttention block is comprised of two components, i.e., local attention which integrates the information over spatially-adjacent regions, and mode attention which models the longe-range contextual relationships in a structural manner. To evaluate the effectiveness of each component, all the experiments are made with the same setting except the factor to be studied. The results are reported in table 3(a). We can observe that using local attention (in Eq. 4) can simply yield obvious performance improvement over the baseline, by 3.5% on mAP and 1.7% on Rank-1 score. Incorporating mode attention without mode interaction can further boost the performance. Using the default configuration of adding the interaction between modes is able to push the result further, demonstrating the necessary of all the components in the module. We can conclude that each module plays an important role, and combining the benefits from all of them can effectively learn discriminative feature representations.

Ablation study on module configurations. In this section, we conduct extensive ablation studies to fully investigate the effect of module configurations for performance which we hope to provide a comprehensive understanding for the module. All the following studies are conducted on the Market1501 dataset.

Gating function: we compare the different gating functions for distributing the contextual information in the mode attention unit. The results show that the sigmod function works better than the softmax function. We conjecture that the use of sigmod relaxes the strong assumption held by the softmax function, *i.e.*, each node should belongs to one mode, as redundant information (such as some background) always exist and may not be useful for describing discriminative patterns. By the consideration of performance, we apply the sigmoid function by default.

Kernel size: The kernel size of local attention module may play an important role as it determines the spatial extent the local attention unit can cover (in Table 3(c)). Among 3×3, 5×5 and 7×7, kernels with the size of 3×3 and 5×5 achieve similar performance. By the consideration of efficiency, we adopt 3×3 kernel size by default.

Head numbers: Intuitively, we expect each attention head to represent a certain component of the data, *e.g.*, head, shoes, arms with sleeves in person re-identification task (details and visualization can be found at Section 6). Experimental results are listed in Table 3(d). The number of heads must not be too large for a valid mode representation, and also not too small for effective feature subspace projection. We follow the best setting, *i.e.*, total 8 heads by default in the experiments.

5.2 Face recognition

To investigate the generalization of the method, we conduct the experiments on face recognition. The discriminative power of feature representations is of importance for the task. Challenges of face recognition may come from a variety of factors, such as pose and expression variations and occlusions. We show that our attention mechanism can boost the performance on multiple validation datasets. We further demonstrate the gain achieved by the module is complementary to the advance of the loss functions whose advantages have been verified on the task of face recognition [51, 9].

Datasets. We use multiple public training datasets [59, 32, 3, 36] for training models.

For evaluation, we apply the following verification datasets which are typically used for evaluating face models. LFW [15] contains 13,233 face images from 5,749 subjects collected from the website. We report the network performance following the standard *unrestricted with labelled outside data* protocol as in [15]. CFP-FP [42] dataset aims to evaluate the models when pose variation is high and extreme pose exists. Agedb-30 [35] contains face images with high age variance.

Table 4: Face recognition results

loss function	network architecture	LFW	CPLFW	CALFW	CFP-FP	AgeDB-30
Softmax	ResNet50	99.1	94.4	82.0	89.1	93.1
	ResNet50-StRA	99.1	95.0	82.4	89.5	93.3
cos loss [51]	ResNet50	99.0	94.4	82.1	90.1	93.7
	ResNet50-StRA	**99.2**	**95.8**	**83.6**	90.5	94.1
arc loss [9]	ResNet50	99.1	94.9	82.0	90.6	93.8
	ResNet50-StRA	99.1	95.7	83.1	**91.1**	**94.4**

CPLFW [65] and CALFW [66] contains the same identities as LFW, focusing on the evaluation with large pose and age variation respectively, requiring good generalization of the features extracted from networks.

Implementation Details. ResNet-50 [11] is adopted as backbone network. The attention modules are used at the last stage of the architecture. BN [18]-Dropout [45]-FC-BN structure used in [9] stacks on the top to obtain a 512-D feature vector. Classification loss (*i.e.*, cross entropy loss) or large margin losses (cosine loss [51] and additive angular margin loss [9]) is used as the objective for training. When evaluating on test set, the above feature vectors of original images and flipped images are concatenated and normalized for comparison. Face verification is conducted by thresholding the euclidean distance. Threshold is selected in the range of $[0, 4]$ which maximizes the accuracy, following the previous work [9]. Training details and hyperparameters are shown at Supplementary Material.

Experimental Results. We compare the baseline model and the counterpart equipped with Structure-Regularized Attention modules with three loss functions. *softmax* is the standard setting that training models with softmax loss. *cos loss* and *arc loss* denotes that the models are trained with large margin cosine loss (CosFace) [51] and additive angular margin loss (ArcFace) [9], respectively.

The results are shown in Table4. We can observe that when training with the softmax function, the proposed network module can benefit the performance on all the test set. The advantage coming from the architecture design can be accumulated into the advance of improved loss functions, showing that the benefit of StRAttention is not restricted into specific losses or tasks.

6 Interpretation and Discussion

In order to further understand the behavior of the module, we provide the comparison by visualizing the activations of feature maps for the four heads in figure 4, which can give a much clearer understanding for the mechanism. We visualize the activations in three settings corresponding to the three volumes in the

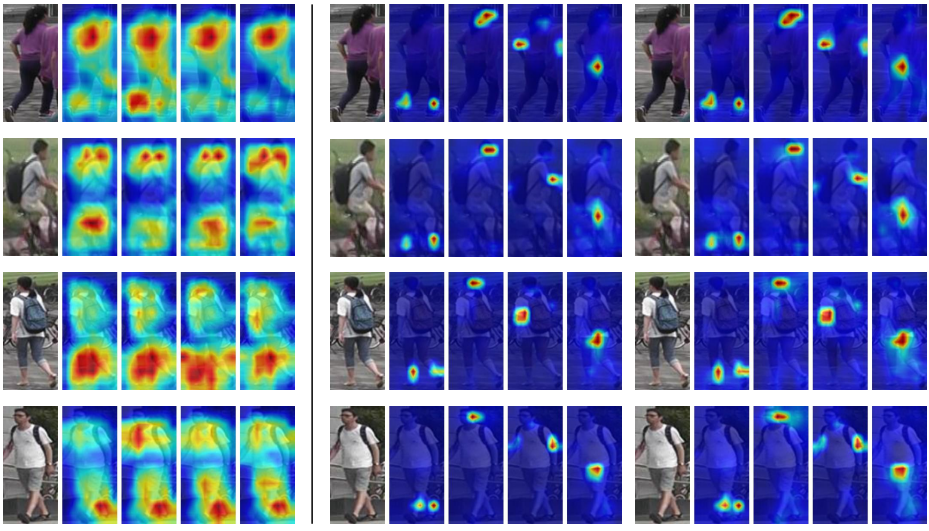


Fig. 4: Visualization for the activation of local attention only variant, attention coefficients of the *mode operation* and the output feature maps of the module.

figure 4, *i.e.*, the output of local attention-only module running only with the local attention unit, the attention coefficients derived from the mode operation and the final output of the module. The difference between the heatmaps generated by local attention only variant is marginal. In contrast, incorporating the regularization of the mode attention unit can diversify the feature learning on different groups. Although the multi-head transformations are assumed to detect distinct patterns, the diversity between heads are still difficult to achieve in practice. The structural regularization by integrating the function of local and mode attentions can benefit the learning of discriminative patterns.

7 Conclusion

In this work we introduce a novel attention module which can effectively capture the longe-range dependency through the use of structural factorization on data. The comprised components, i.e., local attention and mode attention, are complementary for capturing the informative patterns and the combination is capable of improving the discriminative power of models. Extensive experiments on person reid and face recognition across multiple datasets have shown the effectiveness of the proposed method. The structure prior is assumed to be spatial factorization in the work, which would be interesting to generalize it to disentangling factors (e.g., describing the factor of age and emotions for face perception) and may be helpful for representing learning in generative models.

References

1. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: ICCV (2019)
2. Buades, A., Coll, B., Morel, J.M.: Non-local means denoising. Image Processing On Line **1**, 208–212 (2011)
3. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: European conference on computer vision. pp. 768–783. Springer (2014)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: Proceedings of the IEEE international conference on computer vision. pp. 2590–2600 (2017)
6. Chen, Y., Rohrbach, M., Yan, Z., Shuicheng, Y., Feng, J., Kalantidis, Y.: Graph-based global reasoning networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 433–442 (2019)
7. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22 (1977)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
10. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hinton, G.E., Sabour, S., Frosst, N.: Matrix capsules with EM routing. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=HJWLfGWRb
14. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
15. Huang, G., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. Tech. rep. (10 2008)
16. Huang, H., Li, D., Zhang, Z., Chen, X., Huang, K.: Adversarially occluded samples for person re-identification. In: CVPR (2018)
17. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
19. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience (2001)

20. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NeurIPS (2015)
21. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: NeurIPS (2016)
22. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: CVPR. pp. 1062–1071 (2018)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
24. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282289. ICML 01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
25. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)
26. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
27. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9167–9176 (2019)
28. Li, Y., Gupta, A.: Beyond grids: Learning graph representations for visual recognition. In: Advances in Neural Information Processing Systems. pp. 9225–9235 (2018)
29. Lin, M., Chen, Q., Yan, S.: Network in network. In: ICLR (2014)
30. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6639–6648 (2019)
31. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4099–4108 (2018)
32. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
33. McLachlan, G.J., Basford, K.E.: Mixture models: Inference and applications to clustering, vol. 84. M. Dekker New York (1988)
34. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV (2019)
35. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 51–59 (2017)
36. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
37. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. Acm Sigkdd Explorations Newsletter (2004)
38. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: NeurIPS (2019)
39. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in neural information processing systems. pp. 3856–3866 (2017)
40. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)

41. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR (2018)
42. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
43. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
44. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR (2018)
45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
46. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
47. Tang, Z., Naphade, M., Birchfield, S., Tremblay, J., Hodge, W., Kumar, R., Wang, S., Yang, X.: Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
48. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5916–5925 (2017)
49. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: CVPR (2018)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
51. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
52. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
53. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. arXiv preprint arXiv:1901.10430 (2019)
54. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
55. Xie, W., Shen, L., Zisserman, A.: Comparator networks. In: ECCV (2018)
56. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
57. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: A deep learning approach. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). pp. 3676–3684 (2015)
58. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1389–1398 (2019)
59. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)

60. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., Jiang, W., Zhang, C., Sun, J.: Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
61. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR (2019)
62. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
63. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
64. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
65. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep **5** (2018)
66. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
67. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)
68. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)

Structure-Regularized Attention Networks
SUPPLEMENTARY MATERIAL

Anonymous ECCV submission

Paper ID 3180

A Module Implementation

The Structure-Regularized Attention (StRAttention) module is comprised of two operations, *local attention* and *mode attention*. The schemas of the two operations, local attention and mode attention, are shown in Figure 1. The input of the local attention operation is generated by a 1×1 convolution and the output of the module is consequently fed into the second 1×1 convolution when inserting the module into a bottleneck residual block [1]. As shown in Figure 2 in the main paper, input feature maps are dealt with local attention and then proceeded into mode attention. The output of two operation are added through the use of skip connection.

B Experimental Setup

Person re-identification. The backbone networks (including baselines and ours) are initialized with the parameters of models pre-trained on ImageNet expect the parameters of our modules are initialized randomly. When performing fine-turning, the parameters of all the models from stage 1 to 4 are frozen in the first 8 epochs that could facilitate convergence. Images are resized to 256×128 and simply augmented by randomly flipping, cropping and erasing. We use Adam [3] as the optimizer where the initial learning rate is set to 3e−4, and decayed by 0.2 every 20 epochs. Batch size is set to 32 and weight decay is 5e−4. We train all the model for 90 epochs on two NVIDIA Tesla P40 GPUs, based on the Pytorch [4] framework. The weight of the divergence loss $\mathcal{L}_d$ added to the objective is set to 1.0 in the StrAttention.

Face Recognition. All the models (including baselines and ours) are trained from scratch. Standard SGD with momentum is used as optimizer with the learning rate initially set to 0.1 and decayed by 0.1 at the 8th and the 12th epoch. All the models are trained for 20 epochs respectively. Batch size is set to 512 (on 8 GPUs, i.e., 64 per GPU) and weight decay is 5e−4. For both training and test data, face images are preprocessed with standard strategies that detect the face area and use five data points for similarity transformation [6, 5]. As an input, the patch of 112×112 pixels is cropped from the central region of the image whose shorter edge is first resized to 112, and normalized, i.e., each

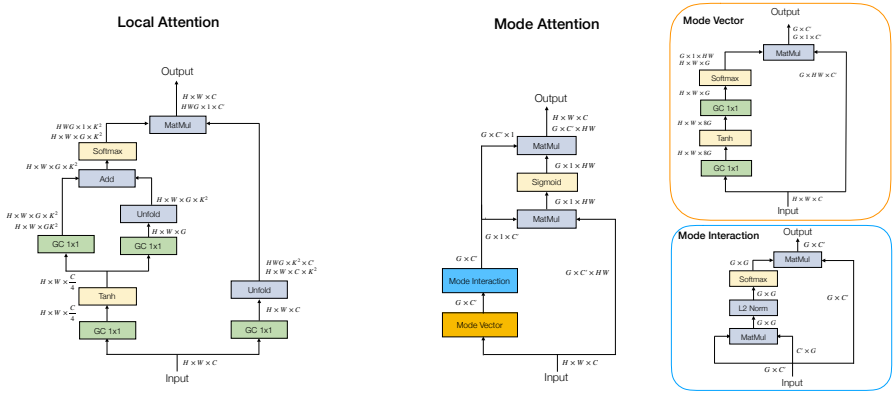


Fig. 1: The schema of an StRAttention module. **Left:** *local attention* operation (Eq. 4 in the main paper). **Right:** *mode attention* operation, which is comprised of mode vector unit and mode interaction unit. GC denotes group convolutions with the group number set to G. H and W denote spatial dimensions. C denotes channel dimension. G is mode/head number and $C' = C/G$ (assuming C is divisible by G). K is the local neighborhood size. Batch Normalization [2] is used after group convolutions by default. The implementation may also require reshape or permute operations, which are not explicitly illustrated in the figure.

pixel is subtracted 127.5 and divided by 128. Only randomly flipping is used for data augmentation during training. When applying the angular loss (ArcFace), the loss hyperparameters, angular margin m and the feature scale s, are set to 0.15 and 20, respectively. When applying the cosine loss (CosFace), the loss hyperparameters, the margin m and the feature scale s, are set to 0.2 and 20. The weight of the divergence loss $\mathcal{L}_d$ added to the objective is set to 0.1 in the StRAttention.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances

in Neural Information Processing Systems 32. pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
5. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
6. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)