Structure-Regularized Attention for Deformable Object Representation

Shenao Zhang
Georgia Institute of Technology
shenao@gatech.edu

Li Shen
Tencent AI Lab
lshen.lsh@gmail.com

Zhifeng Li
Tencent AI Lab
michaelzfli@tencent.com

Wei Liu
Tencent AI Lab
wl2223@columbia.edu

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 this goal by enabling unconstrained pairwise interactions between elements. In this work, we consider learning representations for deformable objects which can benefit from context exploitation by modeling the structural dependencies that the data intrinsically possesses. To this end, we provide a novel structure-regularized attention mechanism, which formalizes feature interaction as structural factorization through the use of a pair of light-weight operations. The instantiated building blocks can be directly incorporated into modern convolutional neural networks, to boost the representational power in an efficient manner. Comprehensive studies on multiple tasks and empirical comparisons with modern attention mechanisms demonstrate the gains brought by our method in terms of both performance and model complexity. We further investigate its effect on feature representations, showing that our trained models can capture diversified representations characterizing object parts without resorting to extra supervision.

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 on a wide range of tasks [14, 18, 51, 55]. Self-attention [51] 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. [55] 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 with respect 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 information into account [1, 43] but the structural information specific to tasks is not effectively exploited.

In this work, we aim to address the visual tasks of representing deformable objects, which may require or highly benefit from the modeling of structural dependencies the data intrinsically has. Inspired

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

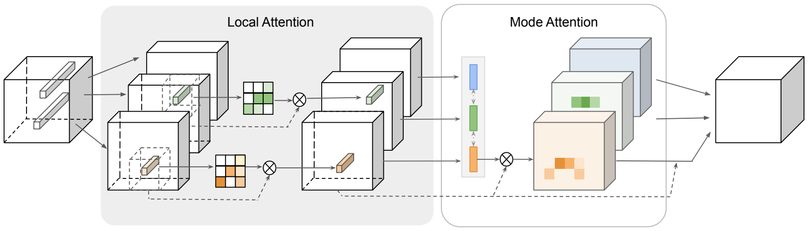

Figure 1: The illustration of the Structure-Regularized Attention Block.

from subspace algorithms [38, 49, 52], 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 components (e.g., associating with semantic concepts) of the data. A set of parameterized transformations project nodes into multiple modes and capture the correlation between nodes and modes, aiming to learn discriminative representations by effectively modeling structural dependencies.

To this end, we introduce a novel attention module which we term the “Structure-Regularized Attention" (StRAttention), formalized as the composition of two-level operations, namely local and mode attentions, to model feature correlations between nodes in a structure-regularized manner. 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 neighboring nodes. The higher-level contextual information can be accessed through the mode attention which allows diversified contextual information to be distributed. The mechanism enables each node to attend to (theoretically) global context in a structural manner.

In summary, the main contribution of the work is as follows. We first propose a novel attention mechanism which allows the capture of long-range dependencies effectively through a structural manner. This also facilitates the learning of structure-discributed representations. As another contribution, we present a formulation of local attention which is simple and efficient. Detailed descriptions for the method are presented in Section 2 and Section 3. We then conduct a series of experiments to validate the approach and compare it with the state-of-the-art attention methods across multiple tasks (including person re-identification, face recognition, and facial expression recognition) in Section 4, demonstrating that its effectiveness is not restricted to a certain task. 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 distinct patterns describing different parts automatically without resorting to extra supervision. The code is publicly available at `https://github.com/zsano1/StRAttention`.

2 Context Modeling by Structural Factorization

We will use “pixel" and“node" interchangeably, and “mode" and “group" 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 pixel x_i, where $i \in \mathcal{N}_G \equiv \{1, \cdots, HW\}$, and $\mathcal{N}_G$ denotes the set of spatial dimensions. The context feature is captured by virtue of the transformation $f : \mathbb{R}^C \times \mathbb{R}^C \to [0, 1]$ [51, 55]:

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

where u represents the unitary transformation on a single node and f captures the pairwise correlation between nodes within global context. f forms a complete graph in which 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 a quadratic computational complexity and memory overhead with respect to the size of $\mathcal{N}_G$ [16, 26]. More importantly, the structural prior is not fully exploited.

The use of hierarchical structure is believed to play a critical role in capturing the statistics in images independent of learnable parameters [50]. 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 end, we formalize the problem as a form of structural factorization. We want to learn a set of transformations to project data onto multiple diversified subspaces, $\Phi := \{\Phi_g\}_{g=1}^G$, where $\Phi_g : \mathcal{X} \to \mathcal{S}_g$ corresponds to the projection from the universal feature space (i.e., input feature maps) onto the g-th subspace which we call "mode" here. The corresponding output of node x_i is represented as $\mathbf{s}_i^g$. Each mode is expected to represent a certain factor the data consists of (e.g., discriminative parts), which is represented by modal vectors $\mathbf{Z} = \{\mathbf{z}_g\}_{g=1}^G$ for input $\mathbf{X}$. The modal vectors are generated by integrating the projections through a parameterized function $\xi_g : \mathbf{S}_g \mapsto \mathbf{z}_g$ where $\mathbf{S}_g = \{\mathbf{s}_i^g\}_{i \in \mathcal{N}_G}$. Let $r_{ig} \in [0, 1]$ indicate the matching degree of node x_i with respect to the g-th mode, which we term *attention coefficients*. Then the context for 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 vectors and propagates such higher-level context to each node.

3 Structure-Regularized Attention

We instantiate the above definition concretely with a two-level attention mechanism, comprised of local attention and mode attention. Local attention projects inputs onto a set of feature subspaces and simultaneously captures local correlation within neighborhoods. Mode attention is responsible for modeling the relation between nodes and modes, as well as the relation between modes.

3.1 Local Attention

The input features are first projected to multiple modes through the use of 1×1 convolutions. For simplification, we omit the index g in the section as the operation performs independently on each mode here. It has been shown that self-attention performing on a local neighborhood is a comparative alternative of convolution [40]. We propose a simplified variant inspired by the dynamic convolution in [57] which is designed for sequence-to-sequence modeling,

$$\begin{aligned} \mathbf{s}_i &= \sum_{j \in \mathcal{N}_K(i)} a_{ij} u(\mathbf{x}_j), \\ a_{ij} &= \sigma_m \left(\omega(\mathbf{x}_i)_j + \nu(\mathbf{x}_j) \right), \end{aligned} \tag{3}$$

where σ_m denotes the softmax function, u denotes a unitary transformation. The transformations $\omega : \mathbb{R}^C \to \mathbb{R}^{K \cdot K}$ and $\nu : \mathbb{R}^C \to \mathbb{R}$ (C is input channels of the transformations) 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 the spatial layout. The contribution from the neighbor itself is derived from ν. The affinity matrix $A_i = \{a_{ij}\}_{j \in \mathcal{N}_K(i)} \in [0, 1]^{K \times K}$ 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. Compared to [40], in which the affinity matrix is computed as,

$$a_{ij} = \sigma_m \left(q(\mathbf{x}_i)^T k(\mathbf{x}_j) + q(\mathbf{x}_i)^T \mathbf{r}_{j-i} \right), \tag{4}$$

where q and k denote the unitary transformations (e.g., linear mappings). The affinity matrix is produced by the sum of two terms. The first captures the relation between 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 the learnable relative position embedding $\mathbf{r}_{j-i}$, where $j - i$ denotes the relative distance of x_i with respect to x_j. We use a more efficient way to model the local relation with the attention mechanism.

Discussion. Compared to linear summation over neighboring nodes in convolutions, local attention expresses patterns in a second-order manner. Unlike convolutions, the operation generates data-dependent weights at each position which introduce dynamics into networks. This motivation shares some similarity with dynamic filters [20] which generate weights conditioned on extra inputs.

3.2 Mode Attention

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 modal vectors which can be realized by mean features (averaging over the local attention output $\mathbf{S}_g$) or centroid features (averaging over representative nodes). The attention coefficient r_{ig} in (2) is then measured by inner product between the corresponding feature vector for node x_i and the modal vector:

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

σ denotes the gating which can be defined with either softmax or sigmoid. The first assumes that the coefficients satisfy a multinomial distribution and will encourage node assignment in a mutually exclusive way. The second measures over different modes independently. Both forms can achieve the goal that nodes will share some context induced by modes to enhance their desired representations. Furthermore, mode interaction $\rho : \mathcal{Z} \to \mathcal{Z}$ can be conveniently achieved by,

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

The updated $\mathbf{Z}$ of across-mode interactions can substitute that 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 [34] which learn clusters by updating central vectors and node assignments iteratively through Expectation-Maximization algorithm [7]. Such an 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.

3.3 Module Instantiation

The module can be instantiated as the replacement of a bottleneck residual block [12], which is comprised of a 1×1 convolution, a 3×3 convolution and another 1×1 convolution, where the 3×3 convolutions are replaced by the operations of local and mode attention.

Local Attention. The formula in (3) 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 Appendix A.

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

4 Experiments

To validate the effectiveness of the proposed Structure-Regularized Attention (StRAttention or StRA for short), we conduct extensive experiments on two types of widely studied deformable objects:

Table 1: Comparison on Market1501. Models ending with *StRA* stand for integrating StRAttention.

Network	mAP	Rank1	FLOPs
ResNet50 [12]	77.1	90.6	4.05G
SASA [40]	79.5	92.3	3.19G
SENet [14]	80.7	93.3	4.49G
Non-Local [55]	80.2	91.9	7.28G
ResNet50_StRA	**84.1**	**93.8**	**3.17G**
MNetV2 [41]	71.7	88.7	**370M**
MNetV2_StRA	74.2	89.3	**370M**
MNetV2(1.4) [41]	72.5	89.0	680M
MNetV2(1.4)_StRA	**74.6**	**89.9**	720M

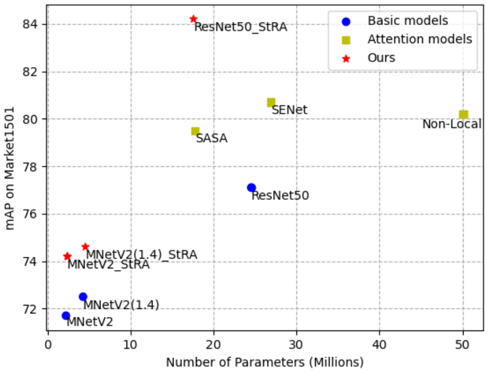

Figure 2: **Model size** vs **mAP**. StrAttention can boost mAP while keeping models small.

human body and human face. We will focus on three tasks: person re-identification (ReID), face recognition and facial expression recognition. All the ablation studies and comparisons are conducted in a fair manner with same training schedule and hyperparameters (except the settings following the original papers). The detailed experimental configurations can be found in the Appendix B.

4.1 Human Body: Person ReID

We evaluate our method mainly on the Market1501 dataset[65] (more comparisons are shown in Appendix B.1). Mean average precision (mAP) and rank-1 accuracy are used as metrics.

Comparison with modern attention networks. We compare our method with three modern attention networks. For all the methods, ResNet-50 [12] is used as backbone architecture. Non-local networks [55] and SENets [14] are two widely-used attention models which integrate the modeling of global information, in which network configuration follows the original papers. SASA [40] is an instantiation of local attention in Eq. (4)) which has shown good results on general object recognition. The results are the mean of 3 runs by default.

Model comparison is conducted in terms of both performance and model complexity. The results in Table 1 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 parameter size versus mAP of different methods in Fig.2. 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 in this task.

Incorporating into efficient architecture. To verify the generalization on backbone architectures, we integrate our modules into MobileNetV2 [41] which is a representative architecture for mobile setting and test on the backbone configuration with the width multiplier setting of 1 and 1.4. The StRA variant is implemented by inserting the modules at the last stage of the backbone.

Table 2: Ablation Studies on module components.

Model	Component			mAP	Rank1
	Local Attn.	Mode Attn.	Mode Interact.		
Resnet50				77.1	90.6
StRA	✓			79.9	92.3
	✓	✓		83.3	93.4
	✓	✓	✓	**84.1**	**93.8**

Model complexity is the critical factor for evaluating efficient architectures. The results in Table 1 demonstrate that our attention module is capable of boosting performance whilst keeping efficiency. Figure 2 verifies the superiority of our model, which is able to achieve higher performance in a computationally efficient and light-weight manner. It further validates the motivation of our module design, i.e., taking advantage of structural dependencies, that deformable objects intrinsically have, into representation learning enables features to interact in an effective manner, facilitating contextual information propagation and further discriminative feature extraction in the network.

Table 3: Performance (%) of different gatings for attention coefficients in *mode attention*.

Gating	mAP	Rank1
Sigmoid	**84.1**	**93.8**
Softmax	82.4	93.7
Tanh	82.7	93.4

Table 4: Performance (%) of using different kernel sizes in *local attention*.

Kernel	mAP	Rank1
3×3	**84.1**	**93.8**
5×5	83.1	93.0
7×7	83.1	93.0

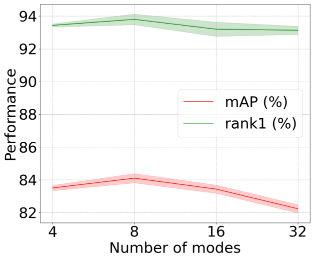

Figure 3: Effects of mode number.

Ablation study on module configuration. The StRAttention block is comprised of two components, i.e., local attention which integrates information over spatially-adjacent regions, and mode attention which models the longe-range contextual relationships in a node-to-mode manner where mode interaction (in Eq. (6)) is exploited to allow mode interactions.

Module component. We first assess each component and the results are shown in Table 2. Using our Local Attention (in Eq.(3)) 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.

We next investigate key configurations, including gating functions for distributing contextual information in Mode Attention Eq. (5) (in Tab. 3), kernel size of Local Attention (in Tab. 4) and mode number (in Fig. 3).

Gating function. Sigmod function works best than the softmax and tanh functions. 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. 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. Kernels with the size 3×3 achieve the best result with least parameters, which is used by default.

Mode numbers. Intuitively we expect each mode to represent a certain component of the object. The results show that mode number must be neither too large that may hamper valid arrangement, nor too small that may reduce effective feature subspace projection. We use the best setting of 8 by default.

4.2 Face: Face Recognition and Expression Recognition

We next conduct the experiments on face data, including face recognition and facial expression recognition, to demonstrate the effectiveness is not confined to certain tasks.

Face Recognition. Challenges of face recognition may come from various factors, e.g., variations in pose, expression and illumination. We conduct the experiments to assess the scalability of the method, based on standard classification loss, i.e., softmax loss, for training. We use a medium scale training set with 1.5 million images [63, 32, 4, 37] and a larger scale dataset [3] with 3.3 million images for model training.

ResNet-50 is used as backbone architecture. The attention modules are used at the last stage of the architecture. All the models are trained from scratch. The evaluation is performed on verification sets including LFW [15], CPLFW [66], CALFW [67], CFP-FP [42] and AgeDB-30 [36]. When testing, pairs of normalized output vectors are compared with Euclidean distance, and the final verification accuracy is conducted with the best threshold in the range of $[0, 4]$ following the protocols [15, 8]. The results in Table 5 show that the scalability of the method, i.e., it can consistently enhance the representational power of networks benefiting from increased dataset scale. We also study on face-oriented loss functions (in the suppl. material) which further demonstrate the gain achieved by the method is complementary to the advance of loss functions [54, 8].

Table 5: Scalability on face recognition (performance %).

Dataset	Network	LFW	CFP-FP	CPLFW	CALFW	AgeDB-30
Medium	ResNet50	99.1	94.4	82.0	89.1	93.1
	ResNet50-StRA	99.1	95.0	82.4	89.5	93.3
Large	ResNet50	99.5	97.3	87.2	89.9	93.9
	ResNet50-StRA	99.7	97.5	88.0	91.8	94.4

Facial Expression Recognition. Facial expressions explicitly correspond to the deformation of discriminative part/landmarks [6]. FER2013 [11] is used for training and evaluation, which includes 7 categories defined as the combination of action units [6]. ResNet is used as backbone and StRAttention variant is constructed by replacing the original blocks at the last stage and omitting the last 1×1 convolution which typically fuses feature maps across channels in order to facilitate the understanding of behavior among modes (in Section 5). Our model can achieve 73.2% accuracy on test set (the groundtruths are not publicly accessible) and 71.5% on the public validation set, outperforming the baseline performance, 71.5% and 69.9%, by a large margin (1.7% and 1.6%) respectively.

5 Interpretation and Discussion

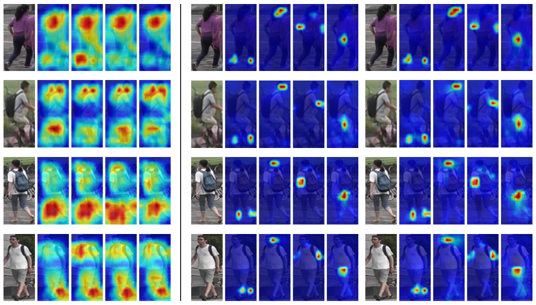

Figure 4: Visualization for the activation of local attention variant, attention coefficients in *mode operation* and the module outputs.

Activation of structure-distributed representations. In order to further understand the behavior of the module, we present the examples of activation visualization (i.e., pixel-wise magnitude on feature maps) for the four modes at the stage 5-1 of ResNet50_StRA in the figure 4. We compare three types of activation in the 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. Although the multi-head transformations are assumed to detect distinct patterns, the diversity between groups is still difficult to achieve in practice. In contrast, incorporating the regularization of the mode attention unit can diversify feature learning on different groups and encourage exciting features corresponding to discriminative parts of objects, realizing the learning of effective structure-distributed representations.

We also provide the spatial distribution of the four modes (i.e., statistics of attention coefficients with respect to spatial locations across a set of samples) in Fig.5, showing that nodes (pixels) tend to be projected and highly correlated with certain modes. It demonstrates that such structural regularization mechanism encourages capturing structure-distributed features for deformable objects in a factorized manner, which potentially provides interpretable features.

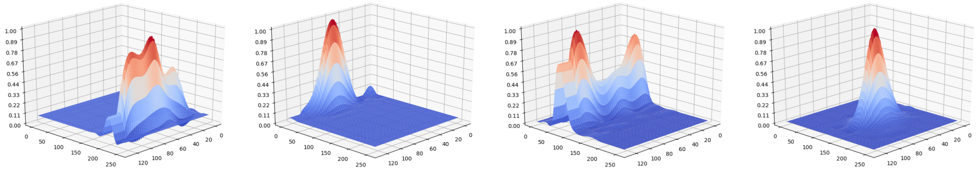

Figure 5: Spatial distributions of high activations (i.e., attention coefficients) on the four modes. Higher peaks indicate that more samples are focusing on the corresponding locations.

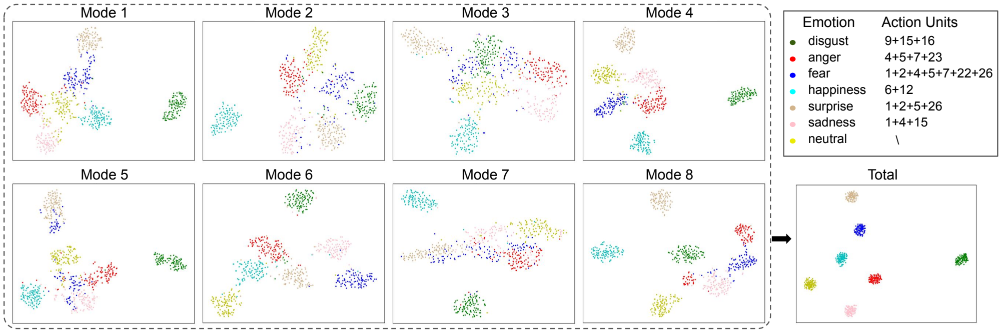

Figure 6: Visualization of features on the 7 facial expression classes.

Effect of modeling structural dependencies. To give a picture for the effect of modeling such structural dependencies, we present the feature distribution on each mode (using the feature layer adjacent to classifiers) of 7 expression classes with t-SNE [33] in Figure 6. The figure at the lower right corner shows the distribution of global features. Emotion relation is typically described by action units [6], which are shared between classes. In other words, there are structural correlations between classes. Although the subspace projections on modes are unsupervised (without extra supervision), we can still observe distinct class correlation reflected on different modes, which function as a kind of latent descriptors for action units and finally strengthen the representational power of networks.

6 Related Work

Convolutional neural networks [23, 12, 47, 58] have proven to be powerful models for representation learning. Graph convolutions have been successfully used for capturing long range relations between the nodes defined by parts [60], objects [29] or distinct regions [5] in a wide range of vision tasks. Recent advance have shown that attention mechanisms are effective to boost discriminative ability of neural networks by enabling model learning to mostly focus on the important components of data. Attention mechanism on spatial [19, 27] or channel [14] or joint [56] dimensions encourages the use of the information from informative regions of inputs. Self attention mechanisms have shown the effectiveness of modeling long-range relationships for sequence learning [51]. Nonlocal operations [55] capture the pair-wise relationship between positions. Some works are proposed to address the issue of computational cost and memory overhead when performing on the complete graph [16, 26] which show benefits on semantic segmentation. Self attention operations are adapted into local regions as the substitute of convolutions in [40]. The works [69, 28] extend such mechanism on group selection. Compared to them, our method is capable of modeling structural dependencies within and between groups in an effective and interpretable manner.

Deformable objects are typically described by the composition of discriminative parts. Multiple part-oriented regions are accordingly detected to extract discriminative features for prediction [64, 59, 24]. Deformable Parts Model [9] 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 [61], human body [25, 31]) where extra manual annotations or rules induced from prior knowledge are typically introduced for supervision. [48] proposes to capture distinct landmarks in an unsupervised way. Different from these methods, we expect to facilitate (intermediate) feature learning by regularizing the information flow through networks with structural dependencies so that stacking such blocks can boost network ability for describing deformable objects.

7 Conclusion

In this work we introduced a novel attention module which can effectively capture the longe-range dependency for deformable objects 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. The extensive experiments on multiple tasks have shown the effectiveness of the proposed method.

The proposed mechanism encourages learning structure-distributed representations which are realized by regularizing information flow conditioned on feature space factorization. The structure prior is assumed to be spatial factorization in the work, which would be interesting to generalize to disentangle factors (e.g., describing the factors of age and emotions for face perception) and may be helpful for representation learning in generative models.

Broader Impact

The work presents an insight that representation learning for deformable objects can strongly benefit from exploiting the prior knowledge the data consists of, though recent progress on a wide range of tasks has shown that the advance of general network architectures is capable of being transferred to such tasks with structured data. The experiments on a variety of deformable object perception tasks verify the value that properly modeling the structural dependencies is not confined to a certain task, while instead it can benefit a wide range of related tasks. We hope that the work can encourage research on network architectures and representation learning for better modeling structural dependencies which will potentially facilitate research on disentangling and interpretable features. At the same time, the effectiveness of the method may be obstructed by the given subspace number (analogous to subspace number for subspace segmentation methods), especially when the number is difficult to estimate in advance or needs gradually increase during training, while the issue may be addressed from the direction of incorporating network evolution in a nonparametric manner.

Appendix

A Module Implementation

The Structure-Regularized Attention (StRAttention) block 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 7. 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, 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.

B Experimental Setup

B.1 Person ReID

Database: We evaluate our method on widely used person ReID datasets including Market1501 [65] and DukeMTMC-ReID [68]. 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. 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.

Configuration: We conduct all the experiments in the single-query setting without a re-ranking algorithm. For Market1501, 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 [46, 35, 13] and a dimensionality-reduction layer is used after average pooling layer, leading to a 512-D feature vector, analogous to [46, 25].

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

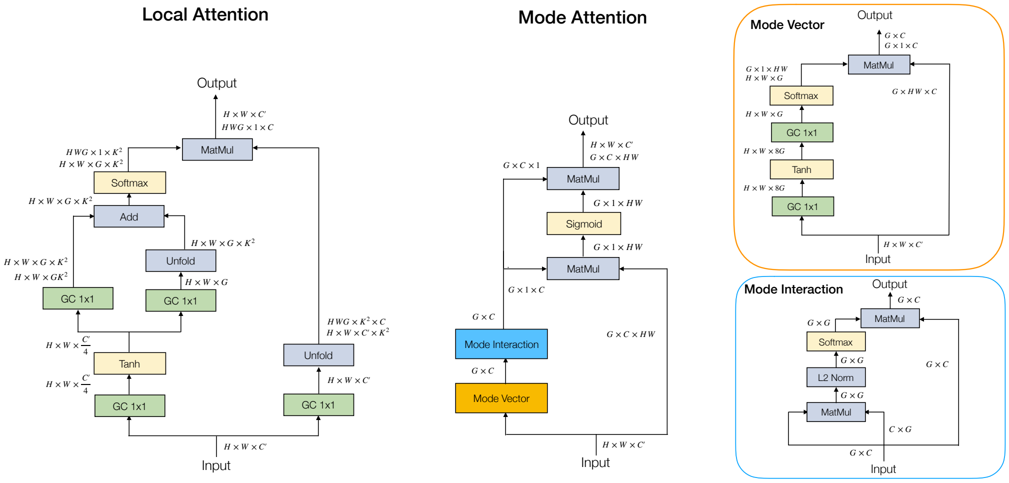

Figure 7: The schema of an StRAttention module. **Left:** *local attention* operation (Eq. 3). The two branches performing with “Add" operation correspond to the transformation ω and ν, respectively. The normalized local softmasks (i.e., affinity matrix) multiply with the output of transformation u and produce the operation output. **Right:** *mode attention* operation, which is comprised of mode vector unit and mode interaction unit. As explained in the section “ Module Instantiation - Mode Attention", 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. 5. 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 [17] 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 DukeMTMC and performing fine-tuning, the parameters of the models from stage 1 to 4 are frozen at the first 10 epochs that could facilitate convergence. When training on Market1501, similar strategies are used at the first 8 epochs. Images are resized to 256×128 and simply augmented by random flipping, cropping, and erasing. We use Adam [22] 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 on Market1501 for 100 epochs, and train models on DukeMTMC for 120 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 [63, 32, 4, 37] as the medium-size training set, and use VGGFace2 [3] 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 [15] 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 [15]. CFP-FP [42] dataset aims to evaluate the models when pose variation is high and extreme pose exists. AgeDB-30 [36] contains face images with high age variance. CPLFW [66] and CALFW [67] 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 [12] is adopted as the backbone network, where conventional global average pooling is replaced by an BN [17]-Dropout [45]-FC-BN module following [8], and finally produces a 512-D feature vector. 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 [54, 8].

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 $5\text{e}{-4}$. 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 [3], 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 [54], *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 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 $5\text{e}{-4}$. The learning rate is initially set to 0.01 and decayed by 0.9 every 5 epochs after 80 epochs, following [39]. Models are trained for 190 epochs in total.

C Comparison with SOTAs of Person ReID

We compare our method with the state-of-the-art (SOTA) methods of person ReID task. These approaches can be categorized into three groups, *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 results in Table 6 show that our method can achieve competitive performance compared to these approaches.

D Ablation Study on Face-Oriented Loss Functions

We conduct the ablation study on face recognition to validate the method with the face-oriented loss functions. Two large margin loss functions, *i.e.*, the cosine loss (CosFace) [54] and the angular loss (ArcFace) [8], are used in the experiments. Models are trained on the medium-scale dataset by following the above setup. For the cosine loss (CosFace), the loss hyperparameters, *i.e.*, the margin m and the feature scale s, are set to 0.2 and 20. For the angular loss (ArcFace), the loss hyperparameters, *i.e.*, the angular margin m and the feature scale s, are set to 0.15 and 20.

Table 6: Performance (%) comparison with the state-of-the-art person ReID methods. The listed methods are categorized into 3 groups: **group 1** contains methods using additional data/supervision; **group 2** represents part based methods and methods in **group 3** adopt attention mechanisms. The three groups are divided by horizontal lines. The last row shows our results.

		Market1501		DukeMTMC-ReID	
	Backbone	mAP	Rank-1	mAP	Rank-1
SPReID [21]	Inception-V3	76.6	90.8	63.3	80.5
PGFA [35]	ResNet-50	76.8	91.2	65.5	82.6
FD-GAN [10]	ResNet-50	77.7	90.5	64.5	80.0
VCFL[30]	GoogLeNet	74.5	89.3	–	–
PCB+RPP [46]	ResNet-50	81.6	93.8	69.2	83.3
IANet [13]	ResNet-50	83.1	93.4	73.4	87.1
OSNet [69]	OSNet	84.9	94.8	73.5	88.6
HA-CNN [25]	Inception-V4	75.7	91.2	63.8	80.5
MltB [62]	ResNet-50	79.0	91.6	65.8	80.7
Mancs [53]	ResNet-50	82.3	93.1	71.8	84.9
SGGNN [44]	ResNet-50	82.8	92.3	68.2	81.1
Ours	ResNet-50	84.1	93.8	73.2	86.1

Table 7: Face recognition performance (%) across different loss functions.

Loss	Network	LFW	CFP-FP	CPLFW	CALFW	AgeDB-30
CosFace [54]	ResNet50	99.0	94.4	82.1	90.1	93.7
	ResNet50-StRA	**99.2**	**95.8**	**83.6**	90.5	94.1
ArcFace [8]	ResNet50	99.1	94.9	82.0	90.6	93.8
	ResNet50-StRA	99.1	95.7	83.1	**91.1**	**94.4**

The results are shown in Table 7. We can see that StRAttention blocks can consistently boost the performance across a range of different losses, suggesting that the gains induced by the blocks could be complementary to the advance of loss functions.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, 2019.

[2] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[3] 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.

[4] 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.

[5] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Yan Shuicheng, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2019.

[6] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007.

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[8] 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.

[9] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in neural information processing systems*, pages 1222–1233, 2018.

[11] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

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

[13] 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.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[15] 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.

[16] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

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

[18] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2001.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, 2015.

[20] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016.

[21] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018.

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

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

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

[26] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9167–9176, 2019.

[27] Xiang Li, Xiaolin Hu, and Jian Yang. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. *arXiv preprint arXiv:1905.09646*, 2019.

[28] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 510–519, 2019.

[29] Yin Li and Abhinav Gupta. Beyond grids: Learning graph representations for visual recognition. In *Advances in Neural Information Processing Systems*, pages 9225–9235, 2018.

[30] Fangyi Liu and Lei Zhang. View confusion feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6639–6648, 2019.

[31] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

[32] 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.

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[34] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. M. Dekker New York, 1988.

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

[36] 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.

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

[38] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 2004.

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

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

[41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[42] 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.

[43] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[44] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 486–504, 2018.

[45] 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.

[46] 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.

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[48] 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.

[49] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[50] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, 2018.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[52] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[53] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, 2018.

[54] 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.

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[56] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018.

[57] Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.

[58] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[59] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *ECCV*, 2018.

[60] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[61] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3676–3684, 2015.

[62] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 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*, pages 1389–1398, 2019.

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

[64] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

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

[66] 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.

[67] 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.

[68] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[69] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3702–3712, 2019.