Compact Global Descriptor for Neural Networks

Xiangyu He   Ke Cheng   Qiang Chen   Qinghao Hu   Peisong Wang

Jian Cheng

NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China

`xiangyu.he, ke.cheng, qiang.chen, qinghao.hu, peisong.wang, jcheng@nlpr.ia.ac.cn`

Abstract

Long-range dependencies modeling, widely used in capturing spatiotemporal correlation, has shown to be effective in CNN dominated computer vision tasks. Yet neither stacks of convolutional operations to enlarge receptive fields nor recent nonlocal modules is computationally efficient. In this paper, we present a generic family of lightweight global descriptors for modeling the interactions between positions across different dimensions (e.g., channels, frames). This descriptor enables subsequent convolutions to access the informative global features with negligible computational complexity and parameters. Benchmark experiments show that the proposed method can complete state-of-the-art long-range mechanisms with a significant reduction in extra computing cost.

1. Introduction

Modeling long-range correlations, which commonly refers to enlarge receptive fields in deep convolutional neural networks [16, 29], is inherently designed in a stack of multiple convolution operators. This intuitive idea coincides with the empirical success of deeper networks [20, 41, 46]. However, repeating convolutional operations comes with unnecessary computing and the risk of overfitting [4, 44]. Besides that, though the respective fields of many deep CNNs have already covered the input image, the effective receptive field only occupies a fraction of the theoretical receptive field [35]. This insight implies that a series of convolutional layers, which aggregates spatial information within a local respective field (e.g., from a 3×3 region), may still lack the mechanism to model the long-range correlations. Capturing long-range dependencies (namely "context") in CNNs then becomes a hot topic of network designs.

The pioneering work is the Inception family [45, 46] with multi-scale processing. Dilated convolutions [57] shares the similar start-point of aggregating multi-scale contextual information without losing resolution. This idea has
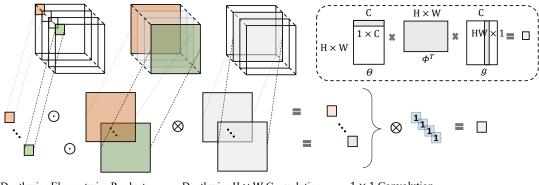
been widely discussed in recent semantic segmentation and detection works [30, 9, 11]. A different aspect of network design utilizes the relationship between channels [25, 24] and extend to spatial dimension [37, 52], yet the interactions between positions across channels or within a local respective field might not be fully extracted using a linear combination after global poolings (via fully connected layer or 3×3 convolution).

Inspired by the classical non-local algorithm for image denoising [5], nonlocal operations compute the response at a position using a weighted sum of the activations of all positions [50]. There is no doubt that non-local networks utilize the rich global information to greatly improve the performances of baseline methods, however, remains computationally expensive. To partially alleviate the above issue, a recent research proposes to reorder the computation between three input feature matrices [12]. [58] further splits the channels dimension into groups and reshape the large matrix into a long vector to avoid heavy matrix multiplications. As illustrated in Figure 1, there is another hidden factor might limit the learning capacity of non-local networks, the depthwise elementwise product without the cross-channel/spatial correlation. Though the depthwise $H \times W$ convolution helps to merge the information, it deserves further discussions on the optimal modeling at far less computing cost.

In this paper, we start by capturing the interactions between positions across all channels and the entire respective field. At first glance, this idea requires a significant amount of operations, while as illustrated in Figure 2 and 3, it saves *several orders of magnitude* in computation complexity. Since positions can be in space and time, our proposed Compact Global Descriptor (CGD) is an efficient module for both static image recognition and video classification.

The main contributions are as follows:

- We present a novel generic network module for long-range dependencies modeling, which considers the correlations among all positions across different dimensions, such as global spatiotemporal correlations.

Depthwise Elementwise Product   Depthwise H × W Convolution   1 × 1 Convolution

Figure 1. The illustration of non-local module [50]. We expand the matrix multiplication (upper right dotted box) in non-local networks into products and convolution operations to show the hidden weighted sum mechanism. The time and space complexity of non-local module is $\mathcal{O}(C(HW)^2)$ and $\mathcal{O}(C^2)$, where C, H and W indicate the number of channels, featuremap height, and featuremap width respectively. In this case, space complexity refers to weight parameters, e.g., 1×1 convolutions to produce Θ, Φ and g.

- The proposed Compact Global Descriptor (CGD) consistently outperforms baseline methods on the benchmark datasets. Compared with state-of-the-art long-range models, CGD gains significant decreases in space and time complexity.

- Our methods can be easily combined with the existing network components (e.g., no changes in size) relieving the dilemma of enhancing performance and reducing computation overhead.

We evaluate the generality of CGD on the tasks of ImageNet classification [14], Mini-Kinetics action recognition [53] and COCO object detection [32]. No matter with the strong ResNet-50 backbone or the lightweight MobileNet extractor, our CGD modules can promise a higher accuracy than baselines in all experiments.

2. Related Work

Multi-scale Correlations: Inspired by the primate visual cortex model in neuroscience, the authors of [40] first propose the different size Gabor filters to handle multiple scales. Similarly, GoogLeNet [45] incorporates multi-scales convolutions to cover large patches. Considering the specific dense prediction problems in semantic segmentation, [57] uses dilated convolutions to aggregate multi-scale contextual information without losing resolution. To enlarge the respective field of neural networks, [60, 8, 9] also apply this technique to scene parsing and semantic segmentation.

Channel Correlations: Squeeze-and-Excitation networks [25] shows that reweighting feature channels by explicitly models the interdependencies of its spatial features can improve classification accuracy. Gather-Excite networks further aggregate feature responses across spatial

neighborhoods then produce an output with the same dimension as input [24]. Based on previous works, Convolution Block Attention [52] introduces global max pooling to gather another important clue. Our method learns the correlations between any two positions, which enjoys the dense clue to enhance the channel-wise relationships, and can be easily extended to spatiotemporal correlations. Besides that, CGD requires far less extra parameters and computing cost.

Spatial Correlations: The spatial correlations (also known as "attention" or "context") which coincide with human perception, focus on "where to look". In the tasks of visual captioning and question answering, spatial attention has been widely applied in structural prediction [54, 55, 7, 56]. Due to the limited channel-spatial modeling, SCA-CNN [10] proposes to first obtain the spatial attention weights, which are then multiplied in each spatial regions. For image classification, CBAM [52] and BAM [37] share a similar idea by using two sequential sub-modules: channel and spatial. Residual Attention Networks [49] presents the separate channel and spatial attention in the form of encoder-decoder. Recently, deformable convolutional networks [13, 62] introduce the spatial sampling with learned offsets to enhance geometric transformations modeling capability of CNNs. Our methods are compatible with these attention mechanisms, and can also be easily applied to the space dimension (detailed in section 3.2).

Spatio-temporal Correlations: Capturing spacetime dependencies is of central importance in video processing. [36, 15] intuitively utilize RNNs for sequences modeling and CNNs for images. Directly applying convolutions to spacetime leads to the 3D filters, known as C3D [27, 48].
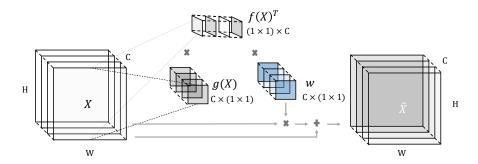
Figure 2. The overview of a Compact Global Descriptor (CGD) module, in the case of $T = 1$ (i.e., single frame) and applying to the channel dimension. The sub-descriptors $f(\cdot)$ and $g(\cdot)$ first maps features across spatial dimensions into a response vector. The outer product of g and f^T is then passed to capture the long-distance dependencies among different channels. The learned param $\mathbf{w}$ further aggregates feature responses from the correlations between the positions of any two channels.
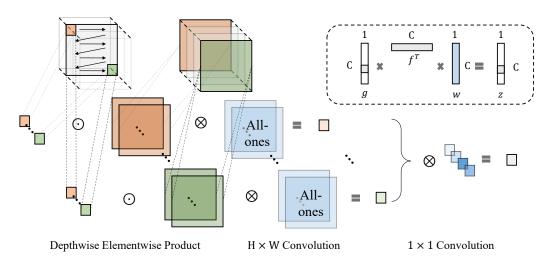

Figure 3. The illustration of the correlation between input features $\mathbf{X}$ and output global descriptions z in CGD-v1 module. The main body of this figure shows the operation illustrated in the upper right dotted box. Here, we use global average pooling as the sub-descriptor (i.e., $f(\cdot)$ and $g(\cdot)$) to obtain the C-dimensional response vectors. The time and space complexity of CGD are both $\mathcal{O}(C)$.

Another group of methods focus on high-level feature aggregations using clustering [2, 17] or building graphs [51] to connect space and time. The "relation" networks [39, 61] merge information among frames through multiple MLP and concat/elementwise sum operations. The non-local modules [12, 58, 47, 50] are different from fully-connected layers. These methods compute responses according to relations between different locations, which is also exploited in this paper.

3. Compact Global Descriptors:

In this section, we briefly review nonlocal networks from the point of extra computing costs and introduce a general formulation of the proposed lightweight global descriptors. Then, we give a further discussion about the specific instantiations.

As shown in Table 1, in the first several layers,

Methods	Embedding	$\Theta\Phi^T g$
Non-local [50]	$\mathcal{O}(C^2 HW)$	$\mathcal{O}(C(HW)^2)$
A^2 [12]	$\mathcal{O}(C^2 HW)$	$\mathcal{O}(C^2 HW)$
CGNL [58]	$\mathcal{O}(C^2 HW)$	$\mathcal{O}(CHW(P+1))$
CGD	–	$\mathcal{O}(C)$

Table 1. Time complexity comparsions of non-local operations ($T = 1$, i.e., single frame), where "Embedding" refers to 1×1 convolutions to halve input channels and $\Theta\Phi^T g$ is the correlation operation among positions. P in [58] indicates the order of Taylor expansion for kernel functions.

$\mathcal{O}(C(HW)^2)$ computing suffers from the unsatisfied time consumption. Recent works make it possible to place non-local modules in the first stage of ResNet, yet the computing cost of linear embeddings have been overlooked. We show that by utilizing the well-designed correlation operations,

we can still enjoy the global information with negligible computational complexity and parameters.

3.1. Formulation

Denote the dimension of interest (e.g., channel in Figure 2 and 3) as d, with $X = [X_1, X_2, ..., X_N]^T \in \mathbb{R}^{M \times N}$ being the features of an input sample represented along d where $X_n \in \mathbb{R}^M$ ($n = 1, \cdots, N$). A generic global descriptor is defined as:

$$z_n = \sum_{\forall i} \sum_{\forall j} \sum_{\forall k} \psi(X_{n,j}, X_{i,k}) w_i \qquad (1)$$

$$= \mathcal{F}(X_n, X; w). \qquad (2)$$

Here $n, i \in [1, N]$, $j, k \in [1, M]$, scalar z_n is the global description of X_n, and $w \in \mathbb{R}^N$ is the learnable weight. A pairwise function ψ produces a scalar between the position j in X_n and all possible position $\forall(i,k)$. It is clear that Eq.(1) considers the long-range dependencies since index (i, k) enumerates the whole featuremap.

When incorporating the global descriptor into neural networks, this module can be further formulated as

$$\widehat{X_n} := X_n + \phi\big(\mathcal{F}(X_n, X; w)\big) X_n \qquad (3)$$

$$= X_n\Big(1 + \phi\big(\mathcal{F}(X_n, X; w)\big)\Big), \qquad (4)$$

where ϕ is the nonlinear function, such as $tanh()$ in this work. Inspired by PreResNet [21], Eq.(3) also leads to nice backward propagation properties. According to the chain rule of backpropagation, we obtain:

$$\frac{\partial Loss}{\partial X_n} = \frac{\partial Loss}{\partial \widehat{X_n}} \frac{\partial \widehat{X_n}}{\partial X_n} \qquad (5)$$

$$= \frac{\partial Loss}{\partial \widehat{X_n}} \Big(1 + \phi\big(\mathcal{F}(X_n, X; w)\big) + \frac{\partial \phi}{\partial \mathcal{F}} \frac{\partial \mathcal{F}}{\partial X_n} X_n\Big). \qquad (6)$$

Eq.(6) not only guarantee a term of $\frac{\partial Loss}{\partial \widehat{X_n}}$ that propagates information directly without the influence of residual block, but reweight the gradient tensor based on the global descriptions $\mathcal{F}(X_n, X; w)$. The derivative of $tanh()$ also works as the gradient clipping technique to avoid the potential gradient exploding of $\frac{\partial \mathcal{F}}{\partial X_n}$.

3.2. Instantiations

Note that the dimension of interest d can be any dimensions in CNNs, not limited to channel dimension in Figure 2. Since it is more intuitive and easy to understand when applying to channel, we describe several instantiations CGD modules with the following notations: featuremap $X \in \mathbb{R}^{WH \times CT}$, where W, H, C and T indicate the featuremap width, featuremap height, number of channels and frame number respectively.
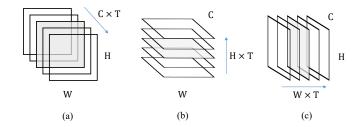

Figure 4. Different spacetime modelings by concatenating T frames $C \times H \times W$ feature tensors in channel dimension (a), height dimension (b) and width diemsion (c), which finally produce global descriptions $z \in \mathbb{R}^{CT}$, $z' \in \mathbb{R}^{HT}$ and $z'' \in \mathbb{R}^{WT}$ respectively.

To model the space-time correlations, we concatenate different frames in the channel dimension as illustrated in Figure 4(a). We can also focus on enhancing the "height-wise" or "width-wise" representational power in a similar way (Figure 4(b),(c)). For simplicity, this paper mainly considers the case of $z \in \mathbb{R}^{CT}$, more specifically, reweighting feature channels. Yet with $z' \in \mathbb{R}^{HT}$ and $z'' \in \mathbb{R}^{WT}$, one can easily obtain the spatial global description $Z \in \mathbb{R}^{HW \times T}$.

Elementwise Product: Following the dot product in nonlocal modules [50, 58, 12], a natural choice of ψ is the dot-product. Through a simple reformulation,

$$z_n = \sum_{\forall i} \sum_{\forall j} \sum_{\forall k} X_{n,j} X_{i,k} w_i = \sum_{\forall j} X_{n,j} \sum_{\forall i} w_i \sum_{\forall k} X_{i,k} \qquad (7)$$

$$= \Big(\frac{1}{WH} \sum_{\forall j} X_{n,j}\Big) \sum_{\forall i} \Big(\frac{1}{WH} \sum_{\forall k} X_{i,k}\Big) w_i' \qquad (8)$$

$$= AvePool(X_n) \sum_{i=1}^{CT} AvePool(X_i) w_i', \qquad (9)$$

we note that there is a compact representation of Eq.(1), which can be defined as

$$z = g(X) f(X)^T w. \qquad (10)$$

Here g and f are both global average pooling to produce response vectors ($\mathbb{R}^{CT}$) and $w \in \mathbb{R}^{CT}$ is the weight vector. Specifically, we apply the global average pooling operation to the masked region in Figure 4(a). In the case of single frame ($T = 1$), Figure 3 further shows the correlation between X and z_n. Compared with nonlocal blocks, we make $X_{n,j}$ access any position across all channels and the entire receptive field.

Quasi-Kernelized: A simple extension of the outer-product $g(X) f(X)^T$ is to adopt other mappings, g' and f'

from $\mathbb{R}^{WH \times CT}$ to $\mathbb{R}^{CT}$. In this paper, we consider max average pooling and softmax wrapped poolings. Since kernel methods are commonly defined on proper inner-product, $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the input space, we just borrow the insight of measuring the similarity in the kernel space.

Cascaded: Inspired by the reproducing kernel [3], in particular $\langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x')$, we consider using the output z of a previous CGD module as an input to replace $g(X)$ and $f(X)$. In this case, we denote $\langle , \rangle$ as the outer-product, x and x' refer to different f and g. As shown in Table 2, when we directly model $k \langle g_{ave}, f_{max}^T \rangle$ as module v4, the result is unsatisfying. Combining $k \langle \cdot, f_{ave}^T \rangle$ (module v3) and $k \langle \cdot, f_{max}^T \rangle$ (module v6), the cascaded module v10 outperforms other candidates. It is shown that we still can utlize the relationship between $g_{ave}(X)$ and $f_{max}^T(X)$ implicitly.

3.3. Block Setting

A global descriptor in Eq.(3) can be easily added in any residual block without the burden of high computing cost. When incorporated with convolutional layers (directly after the convolutions), it allows us to access a richer hierarchy that combines both local and global information. Based on Eq.(6) and Eq.(10), we define the specific channel-oriented global descriptor as:

$$\widehat{X} = X \left(1 + \text{Tanh}\big(g(X)f(X)^T w\big) \right). \tag{11}$$

Compared with Sigmoid$(\cdot)$, the residual connection with Tanh$(\cdot)$ is more compatible with pre-trained models, if w is initialized as zero then $\widehat{X} = X$. Also, we make a more aggressive assumption that some channels can be omitted $(1 + \text{Tanh}(-\infty) = 0)$, which shares the similar idea with channel prunings [22, 34, 63].

The inner computation of a global descriptor is lightweight when we compute the inner-product $s = f^T w$ first and then calculate $z = gs$ as a *scal* operation. We further discuss the following implementations to make it more efficient.

Implementation of CGD blocks: Unless specified, for lightweight networks such as MobileNet [23], SqueezeNet [26] and Cifar-100 networks, we set each CGD module right after the first convolution layer in every block. This benefits from the bottleneck design to reduce the channel number. Besides that, we can further merge the global description z with the batch-normalization parameters to avoid the *axpby* operation at inference time. For ResNet-style networks with several stages, we still add a CGD module to the first convolution layer in every residual block, since this setting will enlarge the receptive field of the following two convolution layers at a minimum cost.

In spatio-temporal modeling, $w \in \mathbb{R}^{CT}$ is still T times larger than static image recognition models. We consider evaluate a summation form of z:

$$z = \sum_{\forall t} g(X^{\{t\}}) f(X^{\{t\}})^T w, \tag{12}$$

where $X^{\{t\}} \in \mathbb{R}^{WH \times C}$. With weight-sharing, the above equation reduces the space complexity from $\mathcal{O}(CT)$ to $\mathcal{O}(C)$. This simplification bases on the assumption that the main object remains unchanged for most of neighbor frames, which implies the potential similarity in channels of interest between adjacent frames.

4. Experiments:

In this section, we first conduct extensive ablation studies to obtain the optimal CGD module on ImageNet dataset. Then we evaluate the proposed CGD method using different network architectures on both image recognition and video recognition tasks, and compare it with state-of-the-art approaches. We use PyTorch [1] in all our experiments, specifically, mmdetection [6] on MS COCO objection detection.

Ablation Studies: For the ablation studies on ImageNet, we replace the basic building block in ResNet-18 with "bottleneck" building block and change the block setting from [2,2,2,2] to [1,2,2,1], which builds a 20 layers residual network, namely ResNet-20. We use 2 GPUs per experiment with a batch size of 256 training from scratch. The ImageNet-1K [14] dataset consists of 1.2 million training images and 50K for validation. We adopt the same data augmentation scheme as described in [52, 37, 58, 12] for training and report the single-crop results with input size 224×224 at inference time. The learning rate starts from 0.1 which is reduced with a factor of 0.1 at the 30th and 60th epoch. We train the networks for 90 epochs using a weight decay of 0.0001 and momentum of 0.9. Following [52, 37, 58, 12], we report Top-1 and Top-5 classification error rates on the validation set.

Table 2 shows that, in most cases, global average pooling (first row) outperforms other variants. This result indicates that CGD with elementwise product can exploit the second order information, which is similar to non-local operations [50]. The cascaded scheme further utilizes the interactions with global max pooling to obtain the optimal long-range dependences modeling.

4.1. Experiments on Video Classification:

In this subsection, we evaluate the proposed method on two benchmark datasets, Mini-Kinetics [53] and UCF-101 [43]. The Mini-Kinetics dataset consists of 200 action categories. Due to some online videos are missing, we down-

Module	$g(\mathbf{X})$	$f(\mathbf{X})$	Top-1 Error (%)	Top-5 Error (%)
CGD-v1	Global-Ave-Pooling	Global-Ave-Pooling	28.79	9.63
CGD-v2	Global-Ave-Pooling	Softmax(Global-Ave-Pooling)	31.28	11.10
CGD-v3	Softmax(Global-Ave-Pooling)	Global-Ave-Pooling	29.88	10.44
CGD-v4	Global-Ave-Pooling	Global-Max-Pooling	32.25	12.15
CGD-v5	Global-Max-Pooling	Global-Max-Pooling	45.15	21.73
CGD-v6	Softmax(Global-Ave-Pooling)	Global-Max-Pooling	30.48	10.84
CGD-v7	Softmax(Global-Max-Pooling)	Global-Ave-Pooling	28.80	10.23
CGD-v8	v3 output	v7 output	28.51	9.60
CGD-v9	v1 output	v7 output	28.78	9.65
CGD-v10	v3 output	v6 output	**28.46**	**9.48**

Table 2. Ablation studys on the selection of sub-descriptor: $f(\cdot)$ and $g(\cdot)$. We use ResNet-20 (detailed in section 4) as the development and validation architecture and apply the best setting to all the other experiments. Single-crop classification error rates on ImageNet validation set are reported. CGD-v8/v9/v10 refer to the cascaded CGD modules where the output of previous CGD will be used as f and g in the following module.

Model	#Params	Δ_NParams	MACs	Δ_NMACs	Top-1(%)	Top-5 (%)	Dataset
ResNet-50 [20]	25.56M	–	17.87G	–	81.62	94.62	UCF-101
ResNet-50 [20] + Non-local [50]	+2.10M	$105\times$	+1.80G	$11^4\times$	82.88	**95.74**	UCF-101
ResNet-50 [20] + CGNL [58]	+2.10M	$105\times$	+1.64G	$11^4\times$	83.38	95.42	UCF-101
ResNet-50 [20] + CGD	**+0.02M**	$1\times$	**+0.095M**	$1\times$	**84.06**	95.51	UCF-101
ResNet-50 [20]	25.56M	–	17.87G	–	75.54	92.16	Mini-Kinetics
ResNet-50 [20] + Non-local [50]	+2.10M	$105\times$	+1.80G	$11^4\times$	76.53	92.90	Mini-Kinetics
ResNet-50 [20] + CGNL [58]	+1.64M	$82\times$	+1.29G	$10^4\times$	**77.76**	93.18	Mini-Kinetics
ResNet-50 [20] + CGD	**+0.02M**	$1\times$	**+0.095M**	$1\times$	77.56	**93.20**	Mini-Kinetics

Table 3. Comparsions on the extra computing cost between CGD and other long-range modules on video classification tasks. 1 non-local or CGNL block is added into res_4 of ResNet. In #Params, "M" refers to one million parameters and "K" is one thousand parameters. In MACs, "G" and "M" equals to one billion and one million Multiply-Accumulate operations respectively, i.e., gigaMACs and megaMACs. Δ_N means normalized extra costs. We assume the input size is $32 \times 3 \times 224 \times 224$ at inference time.

load 79551 videos for training and 4784 videos for validation. UCF-101 contains 13320 videos from 101 action categories, we use the official train/test split in our experiments.

Experiments on UCF-101: To make a fair comparison, we keep the same architecture configuration with non-local networks [50] and CGNL [58], .i.e., C2D ResNet-50. Following the setting in [50], we use ImageNet pre-trained models to initialize the weights, then fine-tune our models using 32-frame input clips. For w in CGD modules, we adopt the method in [18] to initialize weights and set biases to zeros. We first randomly crop out 64 consecutive frames from the full-length video then drop every other frame. Following [58], the input size is 224×224, first randomly cropped between 0.75 and 1 of the original image. Due to the limited computing resources, we train our models on 4 GPUs with a mini-batch size of 44 clips. As described in CGNL [58], the strategy of gradual warmup is used. Then we train our models for 100 epochs in total, starting with a learning rate of 0.01 and reducing it by a factor of 10 at 40th and 80th epoch. We use a weight decay of 0.0001 and momentum of 0.9 in default. Specifically, we halve the weight decay of w in CGD modules since gf^T is rather small, a larger w is required. A dropout layer with

Method	Baseline	Conv1 + CGD	Conv2 + CGD	dataset
Top-1 (%)	81.62	**84.06**	83.67	UCF-101
Top-5 (%)	94.62	**95.51**	95.40	UCF-101
Top-1 (%)	75.54	**77.56**	77.18	Mini-Kinetics
Top-5 (%)	92.16	**93.20**	93.00	Mini-Kinetics

Table 4. Comparsions on CGD module added into different positions in every residual block. All entries are with ResNet-50 on both UCF-101 and Mini-Kinetics. Top-1 and Top-5 accuracies are reported.

0.5 dropout rate is adopted after the global pooling layer. BatchNorm (BN) is also enabled to reduce overfitting [50].

At inference time, we follow the implementation in [50, 42, 58] to perform infernce on videos with shorter side 256. We report the spatial single crop results in Table 3, without temporal domain multiple clips. Since Conv1 and Conv2 in residual blocks has the same channel number, we discuss a little bit about this setting in Table 4. Conv1 + CGD performs slightly better, yet both settings considerably exceed the baseline result.

Experiments on Mini-Kinetics: We further use Mini-Kinetics to evaluate the generality of CGD modules. In this experiment, we adopt the same training setting as in UCF-

Methods	#Params	Δ_N Params	MACs	Δ_N MACs	Top-1 Error (%)	Top-5 Error (%)
MobileNet [23]	4.23M	–	0.569G	–	31.39	11.51
MobileNet [23] + SE [25]	+0.49M	16×	+0.480M	16×	29.97	10.63
MobileNet [23] + BAM [37]	+0.09M	3×	+20.56M	685×	30.58	10.90
MobileNet [23] + CBAM [52]	+0.49M	16×	+3.082M	103×	29.01	9.99
MobileNet [23] + CGD	**+0.03M**	1×	**+0.030M**	1×	**27.44**	**9.08**
MobileNet $\alpha = 0.7$ [23]	2.30M	–	0.283G	–	34.86	13.69
MobileNet $\alpha = 0.7$ [23] + SE [25]	+0.24M	12×	+0.235M	11×	32.50	12.49
MobileNet $\alpha = 0.7$ [23] + BAM [37]	+0.04M	2×	+10.10M	505×	33.09	12.69
MobileNet $\alpha = 0.7$ [23] + CBAM [52]	+0.24M	12×	+2.592M	130×	31.51	11.48
MobileNet $\alpha = 0.7$ [23] + CGD	**+0.02M**	1×	**+0.021M**	1×	**29.89**	**10.56**
SqueezeNet [26]	1.24M	–	0.716G	–	43.09	20.48
SqueezeNet+BAM [37]	+0.02M	7×	+14.00M	7000×	41.83	19.58
SqueezeNet+CGD	**+3.2K**	1×	**+0.002M**	1×	**39.63**	**17.52**

Table 5. Comparsions on the extra computing cost between CGD and other long-range modules on resource constrained architectures, MobileNet-v1 and SqueezeNet-v1.1. We report the single-crop results on ImageNet validation set after 90 epochs.

Methods	#Params	MACs	Top-1(%)	Top-5 (%)
ResNet-50 [20]	25.56M	3.858G	23.85	7.13
+ SE [25]	+2.52M	+2.52M	23.29	6.62
+ BAM [37]	+0.36M	+82.2M	23.14	6.53
+ CBAM [52]	+2.53M	+6.39M	22.38	**6.05**
+ A^2 [12]	+33.0M	+2.65G	23.00	6.50
+ CGNL [58]	+1.64M	+321M	**22.31**	6.36
+ CGNLx [58]	+1.64M	+321M	22.68	6.54
+ CGD	**+0.02M**	**+0.023M**	23.10	6.49

Table 6. Top-1 and Top-5 single crop classification error rates (%) on the ImageNet-1K validation set.

Methods	Error (%)
WideResNet28 (w=8) [59]	20.40
WideResNet28 (w=8) [59] + SE [25]	19.85
WideResNet28 (w=8) [59] + BAM [37]	19.06
WideResNet28 (w=8) [59] + CGD	**18.72**
PreResNet110 [21]	22.22
PreResNet110 [21] + SE [25]	21.85
PreResNet110 [21] + BAM [37]	21.96
PreResNet110 [21] + CGD	**20.96**
PreResNet56 [21]	26.57
PreResNet56 [21] + Nonlocal Modeling [47]	25.29
PreResNet56 [21] + CGD	**24.30**

Table 7. Classification error rates (%) on the CIFAR-100 validation set. Top-1 error rates are reported.

101. Following [58], the testing video clips are resized to shorter side 256 and 3 crops are used to cover the entire spatial size along the longer side. For the temporal domain, we evenly sample 10 clips from a full-length video [58, 50]. The bottom section of Table 3 shows that CGD can complete recent long-range mechanisms with far less extra cost.

4.2. Experiments on Image Recognition:

In this subsection, we further report the results of our CGD modules on the large-scale ImageNet dataset and the benchmark Cifar-100 dataset. ImageNet [14] contains 1.2 million training images and 50K images for validation with 1K object classes. Cifar-100 [28] consists of 60,000 color images with 32×32 pixels drawn from 100 classes. The training and test sets contain 50,000 and 10,000 images respectively.

Experiments on ImageNet: Unless specified, we keep the same experiment settings as dscribed in ablation studies. For lightweight networks such as MobileNet and SqueezeNet, we set the weight decay to 4×10^{-5} including w in CGDs. For ResNets, we still keep the weight decay as 10^{-4} except for 5×10^{-5} on w. Since [12] decreases learning rate when training accuracy is saturated, [25, 52] report their performances after 100 epochs and [58] with warmup,

we train ResNet-50 for 100 epochs. Note that our approach notably surpass other methods on lightweight architectures (Table 5) with negligible computational complexity and parameters. However, CGD perform poorly on ResNet-50 as shown in Table 6, which may be caused by the overfitting problem using rich global information across different dimensions. To verify the effectiveness of CGD on deep neural networks, we further conduct COCO object detection experiments with ResNet-50 in the next subsection.

Experiments on Cifar-100: Following [20, 21], we adopt a standard data augmentation method of random cropping with 4-pixel padding and horizontal flipping. We train all models for 300 epochs with initial learning rate 0.1, and is divided by 10 at 150th and 225th epoch [37]. The weight decay is set to 10^{-4} including w in CGDs. As shown in Table 7, CGD can be easily combined with no matter deeper networks or wider networks to yield higher performance.

Experiments on VOC 2007: We conduct object detection on the PASCAL VOC 2007 test set. The union set of VOC 2007 trainval and VOC 2012 trainval, namely

Evaluation	Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
Bounding Box	Faster R-CNN$_{RoI}$ [19]	ResNet-50-FPN [31]	36.4	58.4	39.1	21.6	40.1	46.6
Bounding Box	Faster R-CNN$_{RoI}$ [19]	ResNet-50-FPN [31]+CGD	**37.9**	**60.3**	**40.7**	**23.0**	**42.0**	**48.0**
Bounding Box	Mask R-CNN [19]	ResNet-50-FPN [31]	37.3	59.1	40.3	22.0	40.9	48.2
Bounding Box	Mask R-CNN [19]	ResNet-50-FPN [31]+CGD	**38.8**	**60.9**	**42.3**	**23.4**	**42.8**	**49.2**
Bounding Box	Mask R-CNN [19]	ResNet-50-C4 [19]	35.6	–	–	–	–	–
Bounding Box	Mask R-CNN [19]	ResNet-50-C4 [19]+CGNL [58]	36.3	–	–	–	–	–
Segmentation	Mask R-CNN [19]	ResNet-50-FPN [31]	34.2	55.9	36.3	18.2	37.5	46.5
Segmentation	Mask R-CNN [19]	ResNet-50-FPN [31]+CGD	**35.4**	**57.6**	**37.6**	**19.4**	**39.2**	**47.4**
Segmentation	Mask R-CNN [19]	ResNet-50-C4 [19]	31.5	–	–	–	–	–
Segmentation	Mask R-CNN [19]	ResNet-50-C4 [19]+CGNL [58]	32.1	–	–	–	–	–
Segmentation	Mask R-CNN [19]	ResNet-50-FPN [19]+Train$_{end-to-end}$ [38]	34.6	56.4	36.5	–	–	–
Segmentation	Mask R-CNN [19]	ResNet-50-FPN [19]+Train$_{end-to-end}$+Non-local [50]	35.5	58.0	37.4	–	–	–

Table 8. COCO object detection single-model results on val2017. The detectors are Faster R-CNN with RoIAlign and Mask R-CNN using different backbones for training and testing. Due to the gap between baseline results, we reach the similar AP^{mask} as non-local networks. However, it should be noted that the gain of our method is +1.2AP with +0.02M MACs, compared with non-local +0.9AP with at least +6.84G MACs. Our method saves over $\mathbf{12^5}\times$ extra computing cost.

"07+12", is used as the training dataset. We evaluate the latest model on the VOC 2007 test after training 240 epochs. The detailed settings are the same as VGG-SSD300 in [6] except that the backbone is replaced by MobileNet. We use a weight decay of 4×10^{-5} including w in CGDs and a momentum of 0.9.

It is noteworthy that CGD hugely improves baseline results (+6.7 mAP) with only 30K extra parameters and Mult-Adds operations (Table 9). This comparison shows that CGD with position-spatial modeling significantly contributes to capturing long-range dependencies.

Experiments on MS COCO: We further investigate whether CGD is complementary to going deeper in standard ways. The models are trained 12 epochs on COCO train2017 (i.e., trainval35k in 2014) and tested on val2017 (i.e., minival in 2014), *without any bells and whistles*. We still follow the setting in [6] except that CGD is added into backbone networks and the weight decay of w in CGD is set to 5×10^{-5}. For MobileNet-SSD300, we adopt the same setting as in VOC except for training 120 epochs.

In our experiments, all models are fine-tuned from ImageNet pre-training. Table 10 shows that MobileNet-SSD with CGD consistently outperforms baselines (+2.1 points). When applying CGDs to deep neural networks such as ResNet-50, Table 8 shows a steady improvement (+1.5 AP^{bbox}) on both Faster R-CNN and Mask R-CNN. Besides that, the extra computing cost of CGD will not change along the input size, since the channel dimension remains the same for different input resolutions. This characteristic leads to the several orders of magnitude reduction in computation complexity. Combining Table 8 and Table 11, CGD has shown to be efficient and effective to model the long-range dependencies. It can complete even outperform the state-of-the-arts [50, 58] on both performance and cost.

Backbone	Detector	mAP@0.5	Training Dataset
MobileNet [23]	SSD [33]	68.1	07+12
MobileNet [23]+CGD	SSD [33]	**74.8**	07+12
MobileNet [23]	SSD [33]	72.7	07+12+COCO

Table 9. Object detection mAP(%) on the VOC 2007 test set.

Backbone	Detector	mAP	MAC	#Params
MobileNet [23]	SSD [33]	19.3	1.2G	6.8M
MobileNet [23] +CGD	SSD [33]	**21.4**	1.2G	6.8M
Deeplab-VGG [23]	SSD [33]	21.2	34.9G	33.1M

Table 10. COCO validation set object detection results comparison using different network architectures. mAP (%) is AP (%) at IoU=0.50:0.05:0.95.

Method	ΔMACs	Δ_NMACs	ΔParams	Δ_NParams
res_4+1 Non-local [50]	+6.84G	$12^5\times$	+2.10M	$92\times$
res_4+1 CGNL [58]	+4.10G	$11^5\times$	+1.64M	$72\times$
$res_{2,3,4,5}$+CGDs	**+0.02M**	$1\times$	**+0.02M**	$1\times$

Table 11. Comparsions on the extra computing cost between Non-local, CGNL and CGD on ResNet-50. We assume the input size is $3 \times 800 \times 800$, since images are resized such that their scale (shorter edge) is 800 pixels. Δ_N refers to normalized extra costs.

5. Conclusion

We present a novel network module which models the long-range dependencies between positions across different dimensions (e.g., channels, frames), namely, global descriptors. Our compact global descriptor can be easily combined with lightweight architectures and deeper networks without the burden of high computing costs. On all tasks, the proposed method significantly contribute to the solid improvement over baselines, and comes with impressive reduction in extra computing cost. Benchmark results show that this module is almost free and could be applied in a wide range of applications.

References

[1] S. C. Adam Paszke, Sam Gross and G. Chanan. Pytorch. https://github.com/pytorch/pytorch, 2017. 5

[2] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1437–1451, 2018. 3

[3] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950. 5

[4] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662, 2014. 1

[5] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 60–65, 2005. 1

[6] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. mmdetection. https://github.com/open-mmlab/mmdetection, 2018. 5, 8

[7] K. Chen, J. Wang, L. Chen, H. Gao, W. Xu, and R. Nevatia. ABC-CNN: an attention based convolutional neural network for visual question answering. *CoRR*, abs/1511.05960, 2015. 2

[8] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 2

[9] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 1, 2

[10] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6298–6306, 2017. 2

[11] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 833–851, 2018. 1

[12] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng. Aˆ2-nets: Double attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 350–359. Curran Associates, Inc., 2018. 1, 3, 4, 5, 7

[13] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 764–773, 2017. 2

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 2, 5, 7

[15] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):677–691, 2017. 2

[16] K. Fukushima and S. Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In S.-i. Amari and M. A. Arbib, editors, *Competition and Cooperation in Neural Nets*, pages 267–285, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg. 1

[17] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3165–3174, 2017. 3

[18] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 249–256, 2010. 6

[19] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017. 8

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 6, 7

[21] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645, 2016. 4, 7

[22] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1398–1406, 2017. 5

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 5, 7, 8

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9423–9433. Curran Associates, Inc., 2018. 1, 2

[25] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 7132–7141, 2018. 1, 2, 7

[26] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016. 5, 7

[27] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2013. 2

[28] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 7

[29] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 1

[30] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Detnet: Design backbone for object detection. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, pages 339–354, 2018. 1

[31] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944, 2017. 8

[32] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755, 2014. 2

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 21–37, 2016. 8

[34] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell. Rethinking the value of network pruning. *CoRR*, abs/1810.05270, 2018. 5

[35] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4898–4906, 2016. 1

[36] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4694–4702, 2015. 2

[37] J. Park, S. Woo, J. Lee, and I. S. Kweon. BAM: bottleneck attention module. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 147, 2018. 1, 2, 5, 7

[38] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Sys-*

tems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 91–99, 2015. 8

[39] A. Santoro, D. Raposo, D. G. T. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4974–4983, 2017. 3

[40] T. Serre, L. Wolf, S. M. Bileschi, M. Riesenhuber, and T. A. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):411–426, 2007. 2

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[43] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 5

[44] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 1

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9, 2015. 1, 2

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826, 2016. 1

[47] Y. Tao, Q. Sun, Q. Du, and W. Liu. Nonlocal neural networks, nonlocal diffusion and nonlocal modeling. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 494–504, 2018. 3, 7

[48] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497, 2015. 2

[49] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6450–6458, 2017. 2

[50] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City,*

UT, USA, June 18-22, 2018, pages 7794–7803, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[51] X. Wang and A. Gupta. Videos as space-time region graphs. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, pages 413–431, 2018. 3

[52] S. Woo, J. Park, J. Lee, and I. S. Kweon. CBAM: convolutional block attention module. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, pages 3–19, 2018. 1, 2, 5, 7

[53] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning for video understanding. *CoRR*, abs/1712.04851, 2017. 2, 5

[54] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, pages 451–466, 2016. 2

[55] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057, 2015. 2

[56] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 21–29, 2016. 2

[57] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015. 1, 2

[58] K. Yue, M. Sun, Y. Yuan, F. Zhou, E. Ding, and F. Xu. Compact generalized non-local network. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 6511–6520, 2018. 1, 3, 4, 5, 6, 7, 8

[59] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016. 7

[60] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6230–6239, 2017. 2

[61] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 831–846, 2018. 3

[62] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. *CoRR*, abs/1811.11168, 2018. 2

[63] Z. Zhuang, M. Tan, B. Zhuang, J. Liu, Y. Guo, Q. Wu, J. Huang, and J. Zhu. Discrimination-aware channel pruning for deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 883–894, 2018. 5