Appendix

Anonymous ICCV submission

Paper ID 5528

1. Details of the Proving for the relations

We give the details of the relationship between other nonlocal operators in the spectral view discussed in our paper (Sec.3.2). In the following proving, we assume that $X \in \mathbb{R}^{\mathbb{N}\times\mathbb{C}}$, $\boldsymbol{Z} = g(\boldsymbol{X}) = \boldsymbol{X}\mathbf{W}_Z$, $\boldsymbol{M}_{ij} = f(\boldsymbol{X}_i, \boldsymbol{X}_j)$. All the normalized term uses the inverse of the degree $1/d_i$ where $d_i = \sum_j f(\boldsymbol{X}_i, \boldsymbol{X}_j)$. We also merge the output of the operators with the weight kernel $\mathbf{W} \in \mathbb{R}^{N\times C}$ and defines it as $\boldsymbol{O}$ for consistency. Thus the target formulations in this section are a bit different with the definition in their own papers.

1.1. Nonlocal Block

The Nonlocal (NL) Block in the spectral view is the same as defining the graph $\mathcal{G} = (\mathbb{V}, \boldsymbol{D}^{-1}\boldsymbol{M}, \boldsymbol{Z})$ and then using the second term of the Chebyshev Polynomail to approximate the graph filter.

Proof. The NL operator defined in [9] can be formulated as:

$$\boldsymbol{O}_{i,:} = \frac{\sum_j \left[f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})g(\boldsymbol{X}_{j,:})\right]}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}\mathbf{W} \tag{1}$$

To unify it by our spectral view, we firstly define the graph $\mathcal{G} = (\mathbb{V}, \boldsymbol{A}, \boldsymbol{Z})$ to represent the graph structure of the NL operator, where the affinity matrix A is calculated by:

$$\boldsymbol{A} = \boldsymbol{D}_M^{-1}\boldsymbol{M}, \quad \boldsymbol{M} = f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:}) \tag{2}$$

Thus, each element of the affinity matrix $\boldsymbol{A}$ is:

$$\boldsymbol{A}_{ij} = (\boldsymbol{D}_M^{-1}\boldsymbol{M})_{ij} = \frac{f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})} \tag{3}$$

Based on Theorem.1, when using Chebyshev polynomails to approximate the generalized graph filter $\boldsymbol{\Omega}$ and only choosing the second term, it becomes :

$$F(\boldsymbol{A}, \boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{Z}\mathbf{W} \tag{4}$$

Then taking Eq. (3) into this equation, we can get the formulation of the NL operator:

$$F_{i,:}(\boldsymbol{A}, \boldsymbol{Z}) = \frac{\sum_j \left[f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})g(\boldsymbol{X}_{j,:})\right]}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}\mathbf{W} \tag{5}$$

□

1.2. Nonlocal Stage

The Nonlocal Stage (NS) in the spectral view is the same as defining the graph $\mathcal{G} = (\mathbb{V}, \boldsymbol{D}_M^{-1}\boldsymbol{M}, \boldsymbol{Z})$ and then using the 1_{st}-order Chebyshev Polynomail to approximate the graph filter with the condition $\mathbf{W}_1 = \mathbf{W}_2 = -\mathbf{W}$.

Proof. The NS operator given defined in [8] can be formulated as:

$$\boldsymbol{O}_{i,:} = \frac{\sum_j \left[f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})(\boldsymbol{Z}_{j,:} - \boldsymbol{Z}_{i,:})\right]}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}\mathbf{W} \tag{6}$$

Similar with the proof of NL, we can get each element of the affinity matrix $\boldsymbol{A}$ as:

$$\boldsymbol{A}_{ij} = (\boldsymbol{D}_M^{-1}\boldsymbol{M})_{ij} = \frac{f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})} \tag{7}$$

The graph filter $\boldsymbol{\Omega}$ on $\mathcal{G}$ is approximated by the Chebyshev polynomail. When using the 1_{st}-order Chebyshev Approximation, it becomes:

$$F(\boldsymbol{A}, \boldsymbol{Z}) = \boldsymbol{Z}\mathbf{W}_1 - \boldsymbol{A}\boldsymbol{Z}\mathbf{W}_2 \tag{8}$$

When sharing the weight for $\mathbf{W}_1$ and $\mathbf{W}_2$, i.e $\mathbf{W}_1 = \mathbf{W}_2 = -\mathbf{W}$, we get:

$$F(\boldsymbol{A}, \boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{Z}\mathbf{W} - \boldsymbol{Z}\mathbf{W} \tag{9}$$

Then, taking it $\boldsymbol{Z} = g(\boldsymbol{X}) = \boldsymbol{X}\mathbf{W}_Z$ and Eq. (7) into this equation, it becomes:

$$F_{i,:}(\boldsymbol{A}, \boldsymbol{Z}) = \frac{\sum_j \left[f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})\boldsymbol{Z}_{j,:}\mathbf{W}\right]}{\sum_j f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})} - \boldsymbol{Z}_i\mathbf{W} \tag{10}$$

Due to the fact that $\frac{\sum_j f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}{\sum_j f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}=1$, we can get the formulation of the NS operator:

$$\begin{aligned} F_i(\boldsymbol{A},\boldsymbol{Z}) &= \frac{\sum_j \left[f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})\boldsymbol{Z}_{j,:}\mathbf{W}\right]}{\sum_j f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})} \\ &\quad - \frac{\sum_j f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}{\sum_{j,:} f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}\boldsymbol{Z}_{i,:}\mathbf{W} \\ &= \frac{\sum_j \left[f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})(\boldsymbol{Z}_{j,:}-\boldsymbol{Z}_{i,:})\right]}{\sum_j f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}\mathbf{W} \end{aligned} \tag{11}$$

□

1.3. Double Attention Block

The Double Attention Block in the spectral view is the same as defining the graph $\mathcal{G}=(\mathbb{V},\overline{\boldsymbol{M}},\boldsymbol{Z})$ and then using the second term of the Chebyshev Polynomail to approximate the graph filter, i.e $F(\boldsymbol{A},\boldsymbol{Z})=\overline{\boldsymbol{M}}\boldsymbol{Z}\mathbf{W}$:

Proof. The A^2 operator defined in [1] can be formulated as:

$$\boldsymbol{O}=\sigma(\theta(\boldsymbol{X}))\sigma(\phi(\boldsymbol{X})^T)g(\boldsymbol{X})=f^a(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})\boldsymbol{X}\mathbf{W} \tag{12}$$

The difference between the double A^2 operator and the NL operator is only the kernel function that calculating the affinity matrix [1]. Thus we can use the similar proving strategy to reformulate the A^2 operator into the spectral only by change the affinity matrix as:

$$\boldsymbol{A}=\overline{\boldsymbol{M}}=\sigma(\boldsymbol{X}\mathbf{W}_\phi)\sigma(\boldsymbol{X}\mathbf{W}_\psi) \tag{13}$$

□

1.4. Compact Generalized Nonlocal Block

When grouping all channels into one group, the Compact Generalized Nonlocal Block in the spectral view is the same as defining the graph $\mathcal{G}=(\mathbb{V}^f,\boldsymbol{D}_{M^f}^{-1}\boldsymbol{M}^f,\mathrm{vec}(\boldsymbol{Z}))$ and then using the second term of the Chebyshev Polynomail to approximate the graph filter, i.e $F(\boldsymbol{A},\boldsymbol{Z})=\boldsymbol{D}_{M^f}^{-1}\boldsymbol{M}^f\mathrm{vec}(\boldsymbol{Z})W$. Note that due to the dimension of the input feature $\mathrm{vec}(\boldsymbol{Z})\in\mathbb{R}^{NC\times 1}$ which is different with other nonlocal operators, here we uses $\boldsymbol{M}^f,\boldsymbol{A}^f\in\mathbb{R}^{NC\times NC}$ for clearity.

Proof. The CGNL operator defined in [10] can be formulated as:

$$\mathrm{vec}(\boldsymbol{O})=f(\mathrm{vec}(\boldsymbol{X}),\mathrm{vec}(\boldsymbol{X}))\mathrm{vec}(\boldsymbol{Z})\mathbf{W} \tag{14}$$

For simplicity, we use $\boldsymbol{x}$ to represent $\mathrm{vec}(\boldsymbol{X})$, thus the target becomes:

$$\boldsymbol{o}=f(\boldsymbol{x},\boldsymbol{x})\boldsymbol{z}\boldsymbol{W} \tag{15}$$

Then, we define the graph $\mathcal{G}=(\mathbb{V}^f,\boldsymbol{A}^f,\boldsymbol{z})$, where the set $\mathbb{V}^f$ contains each index (including position and channel) of the vector $\boldsymbol{x}$. The affinity matrix $\boldsymbol{A}$ is calculated by:

$$\boldsymbol{A}^f=\boldsymbol{M}^f,\quad \boldsymbol{M}^f=f(\boldsymbol{x},\boldsymbol{x}) \tag{16}$$

The graph filter $\boldsymbol{\Omega}$ on $\mathcal{G}$ is approximated by the Chebyshev polynomials. When only choosing the second term, we can get the formulation of the CGNL operator:

$$F(\boldsymbol{A}^f,\boldsymbol{z})=\boldsymbol{A}^f\boldsymbol{z}\mathbf{W}=f(\boldsymbol{x},\boldsymbol{x})\boldsymbol{z}\mathbf{W} \tag{17}$$

□

1.5. Criss-Cross Attention Block

The Criss-Cross Attention Block in the spectral view is the same as defining the graph $\mathcal{G}=(\mathbb{V},\boldsymbol{D}_{\boldsymbol{C}\odot\boldsymbol{M}}^{-1}\boldsymbol{C}\odot\boldsymbol{M},\boldsymbol{X})$ and then using the second term of the Chebyshev Polynomail to approximate the graph filter with node feature $\boldsymbol{X}$:

Proof. The criss-cross attention operator defined in [4] can be formulated as:

$$\boldsymbol{O}_{i,:}=\sum_{j\in\mathbb{V}^i}\boldsymbol{A}_{ij}\boldsymbol{\Phi}_{j,:}=\frac{\sum_{j\in\mathbb{V}^i}f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})\boldsymbol{X}_{j,:}}{\sum_{j\in\mathbb{V}^i}f(\boldsymbol{X}_{i,:},\boldsymbol{X}_{j,:})}\mathbf{W}$$

where the set $\mathbb{V}^i$ is collection of feature vector in $\mathbb{V}$ which are in the same row or column with position u.

Then, we define the graph $\mathcal{G}=(\mathbb{V},\widetilde{\boldsymbol{A}},\boldsymbol{X})$ to represennt the criss-cross attention operator in the spectral view. The affinity matrix $\widetilde{\boldsymbol{A}}$ is calculated by:

$$\widetilde{\boldsymbol{A}}=\boldsymbol{D}_{\boldsymbol{C}\odot\boldsymbol{M}}^{-1}\boldsymbol{C}\odot\boldsymbol{M},\quad \boldsymbol{M}=f(\boldsymbol{X}_i,\boldsymbol{X}_j)$$
$$\boldsymbol{C}_{ij}=\begin{cases}1 & j\in\mathbb{V}^i\\ 0 & \text{else}\end{cases},$$

We use $\widetilde{\boldsymbol{M}}$ to represent $\boldsymbol{C}\odot\boldsymbol{M}$, i.e. $\widetilde{\boldsymbol{M}}=\boldsymbol{C}\odot\boldsymbol{M}$. Thus, each element of the affinity matrix $\widetilde{\boldsymbol{M}}$ is:

$$\widetilde{\boldsymbol{M}}_{ij}=\begin{cases}\boldsymbol{M}_{ij} & j\in\mathbb{V}^i\\ 0 & \text{else}\end{cases},$$

Thus, we can get the definition of each element in the affinity matrix $\widetilde{\boldsymbol{A}}$:

$$\widetilde{\boldsymbol{A}}_{ij}=\begin{cases}\frac{f(\boldsymbol{X}_i,\boldsymbol{X}_j)}{\sum_{j\in\mathbb{V}^i}f(\boldsymbol{X}_i,\boldsymbol{X}_j)}, & j\in\mathbb{V}^i\\ 0, & \text{else}\end{cases},$$

When using the Chebyshev polynomials to approximate the generalized graph filter $\boldsymbol{\Omega}$ on $\mathcal{G}$ and choose the second term, it becomes:

$$F(\widetilde{\boldsymbol{A}}, \boldsymbol{X}) = \widetilde{\boldsymbol{A}}\boldsymbol{X}\mathbf{W} \tag{18}$$

When taking Eq.18 into this formulation, we can get the formulation of CC operator:

$$\begin{aligned} F_{i,:}(\widetilde{\boldsymbol{A}}, \boldsymbol{X}) &= (\frac{\sum_{j\in\mathbb{V}^i} f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})\boldsymbol{X}_{j,:}}{\sum_{j\in\mathbb{V}^i} f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})} + \sum_{j\notin\mathbb{V}^i} 0\boldsymbol{X}_{j,:})\mathbf{W} \\ &= \frac{\sum_{j\in\mathbb{V}^i} f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})\boldsymbol{X}_{j,:}}{\sum_{j\in\mathbb{V}^i} f(\boldsymbol{X}_{i,:}, \boldsymbol{X}_{j,:})}\mathbf{W} \end{aligned} \tag{19}$$

□

2. Defining Novel Nonlocal with Caylay Filter

As discussed in our paper, the five existing nonlocal-based blocks uses the Chebyshev filter as the generalized graph filter defined in our spectral view. However the Chebyshev filter need to map the eigenvalue to range $[-1, 1]$ in a linear manner, which makes its spectral smooth or lost the interest signal on a small frequency band [6]. This makes the result feature map tend to ignore the large object (having high correlation to a large amount of positions). Cayley filter solves this problem by an additional spectral zoom parameter h, which can help to select interest frequency band.

To defines novel nonlocal block based on Caylay filter with the help of on our proposed framework, we can replace the generalized graph filter $\boldsymbol{g}(\Lambda)$ in Theorem. 1 by the k^{th}-order Caylay filter, which is formulated as:

$$\boldsymbol{g}_{\theta,h}(\Lambda) = \theta_0 + 2Re\{\sum_{j=1}^{k} \theta_1(h\boldsymbol{\Lambda} - i\boldsymbol{I})^j(h\boldsymbol{\Lambda} + i\boldsymbol{I})^{-j}\} \tag{20}$$

where "Re" means the real part of an imaginary number, i is the imaginary unit, h and θ are parameters that can be learned by SGD. Taking this into Theorem. 1, we get the CaylaySNL operator with the help of Jacobi approximation [6]:

$$F_{cl}(\boldsymbol{A}, \boldsymbol{Z}) = \theta_0\boldsymbol{Z} + 2\sum_{j=1}^{k} \theta_1(h\mathbf{L} - i\boldsymbol{I})^j(h\mathbf{L} + i\boldsymbol{I})^j\boldsymbol{Z} \tag{21}$$

where $\boldsymbol{L}$ is the graph Laplacian of $\boldsymbol{A}$. Note that, eigenvalues of $\boldsymbol{L}$ are all real, so we can remove the "Re" and take $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ into this polynomial form equation. Finally, we uses the 1^{st}-order for Eq. 21 for high computational efficiency and extend it to multi-channel condition:

$$F_{cl}(\boldsymbol{A}, \boldsymbol{Z}) = \boldsymbol{Z}\mathbf{W}_1 + 2(\boldsymbol{L}^2\boldsymbol{Z}\mathbf{W}_{h^2} + \boldsymbol{Z})\mathbf{W}_2 \tag{22}$$

Similar with other nonlocal-based operator, when adding residual connection, the CaylaySNL operator becomes the CaylaySNL block. Note that the Caylay filter do not require the affinity matrix A is normalized, so we just uses the sigmoid to make the elements of affinity matrix higher than 0. The CaylaySNL block can be formulated as:

$$\begin{aligned} &\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}\mathbf{W}_1 + 2(\boldsymbol{L}^2\boldsymbol{Z}\mathbf{W}_{h^2} + \boldsymbol{Z})\mathbf{W}_2 \\ &\text{s.t.} \quad \boldsymbol{A} = \boldsymbol{D}_{\hat{M}}^{-\frac{1}{2}}\hat{\boldsymbol{M}}\boldsymbol{D}_{\hat{M}}^{-\frac{1}{2}}, \quad \hat{\boldsymbol{M}} = (\boldsymbol{M} + \boldsymbol{M}^\top)/2 \end{aligned} \tag{23}$$

Remark 1. *Different with all existing nonlocal-based blocks that derivated from Chebyshev filter, the proposed ChebySNL utilizes the Caylay filter which can better concerns the small frequency band, i.e. object with large scale.*

We validate the proposed CaylaySNL on CIFAR-100 shown in Table. 1. It can be see that, benefited form concerning the spectral zoom, the Caylay can generate the highest performance (0.31 higher than the proposed SNL). This shows the efficient that using other type graph filter to define nonlocal-based blocks.

Table 1: The Performances of Caylay SNL on CIFAR-100

Models	Top1 (%)	Top5 (%)
PreResNet56	$75.33^{\uparrow 0.00}$	$93.97^{\uparrow 0.00}$
+ NL	$75.29^{\downarrow 0.04}$	$94.07^{\uparrow 0.10}$
+ NS	$75.39^{\uparrow 0.06}$	$93.00^{\downarrow 0.97}$
+ A^2	$75.51^{\uparrow 0.18}$	$92.90^{\downarrow 1.07}$
+ CGNL	$74.71^{\downarrow 0.62}$	$93.60^{\downarrow 0.37}$
+ SNL	$76.41^{\uparrow 1.08}$	$94.38^{\uparrow 0.41}$
+ CaylaySNL	$\underline{\mathbf{76.72}}^{\uparrow \mathbf{1.39}}$	$\underline{\mathbf{94.54}}^{\uparrow 0.57}$

3. External Experiments

3.1. Action Recognization

We also conduct the experiments on UCF-101 dataset with other state-of-the-art action recognition models in our supplementary materials including the P3D [7], the MARS [2], and the VTN [5]. For Pseudo 3D Convolutional Network (P3D) and Motion-augmented RGB Stream (MARS), our SNL block are inserted into the P3D right before the last residual layer of the $res3$. For the Video Transformer Network (VTN), we replace its multi-head self-attention blocks (paralleled-connected NL blocks) into our SNL blocks. We use the model pre-trained on Kinetic dataset and fine-tuning on the UCF-101 dataset. Other setting such as the learning rate and training epochs are the same as the experiment on I3D in our paper. We can see that all the performance are improved when adding our proposed SNL model especially when training end-to-end on the small-scale dataset. In sum, our SNL blocks have shown superior

Table 2: Experiments on Video Person Re-identification

Mars			ILID-SVID			PRID-2011		
Models	Rank1(%)	mAP(%)	Models	Rank1(%)	mAP(%)	Models	Rank1(%)	mAP(%)
ResNet50	$82.30^{\uparrow 0.00}$	$75.70^{\uparrow 0.00}$	ResNet50	$74.70^{\uparrow 0.00}$	$81.60^{\uparrow 0.00}$	ResNet50	$86.50^{\uparrow 0.00}$	$90.50^{\uparrow 0.00}$
+ NL	$83.21^{\uparrow 0.91}$	$76.54^{\uparrow 0.84}$	+ NL	$75.30^{\uparrow 0.60}$	$83.00^{\uparrow 1.40}$	+ NL	$85.40^{\downarrow 1.10}$	$89.70^{\downarrow 0.80}$
+ SNL	$\mathbf{83.40}^{\uparrow \mathbf{1.10}}$	$\mathbf{76.80}^{\uparrow \mathbf{1.10}}$	**+ SNL**	$\mathbf{76.30}^{\uparrow \mathbf{1.60}}$	$\mathbf{84.80}^{\uparrow \mathbf{3.20}}$	**+ SNL**	$\mathbf{88.80}^{\uparrow \mathbf{2.30}}$	$\mathbf{92.40}^{\uparrow \mathbf{1.90}}$

results across three SOTAs (the VTN and MARS) in the action recognition tasks (0.30% improvement with VTN, 0.50% improvement with MARS).

Table 3: Experiments with state-of-the-art backbone

Models	Top1(%)
P3D[7]	$81.23^{\uparrow 0.00}$
P3D + **Ours**	$82.65^{\uparrow 1.42}$
VTN[5]	$90.06^{\uparrow 0.00}$
VTN + **Ours**	$90.34^{\uparrow 0.30}$
MARS[2]	$92.29^{\uparrow 0.00}$
MARS + **Ours**	$92.79^{\uparrow 0.50}$

3.2. Person Re-identification

The experiments for video person re-identification are conducted on Mars, ILID-SVID and PRID-2011 datasets. For the backbone, we follow the strategy of [3] that use the temporal pooling to fuse the spatial-temporal features. Note that the models are totally trained on ilidsvid and prid2011 rather than fine-tuning the pretrained model on Mars. From Table. 2 (Mars), we can see that our SNL can generate 1.10% improvement both on Rank1 and mAP, which are both higher than the original nonlocal block (0.91% on Rank1, 0.84% on mAP). We also generate experiments on two relatively small datasets: ILID-SVID datasets which contains 300 pedestrians captured by two cameras with 600 tracklets; PRID-2011 dataset which contains 200 pedestrians captured by two cameras with 400 tracklets. In Table. 2 (ILID-SVID), our model can generate 1.60% and 3.20% improvements on the Rank1 and mAP respectively for the ILID-SVID dataset. Moreover, on PRID-2011, we get a significant improvement (2.30% on Rank1, 1.90% on mAP) as shown in Table. 2 (PRID-2011).

3.3. Fine-grained Image Classification

The experiments for the fine-grained classification are generated on the Birds-200-2011 (CUB-200) dataset which contains 11,788 images of 200 categories of different birds. We use 5,994 images as the training set and 5,794 images as the testing set [10]. We use the ResNet50 model pre-trained on ImageNet as the backbone and train the models for 110 epochs with the initial learning rate 0.1 which is subsequently divided by 10 at 31, 61, 81 epochs. Table 4 (CUB-200) shows that our model can generate (0.59%) improvement. Compared with the CGNL block concerning channel-wise relations, our SNL is just a bit lower in Top-1 (0.12%). That is because the dependencies among channels play an important role in the fine-grained classification. However, these channel dependencies of CGNL can impede the practical implementations, which needs elaborate preparations for the number of channels per block, the number of blocks and their positions as shown in our main paper. Compared with the other nonlocal block with non-channel concerned, our SNL has improvements with a large margin.

Table 4: Experiments for Nonlocal-based Blocks Added into ResNet50 on CUB-200 Datasets

Models	Top-1 (%)	Top-5 (%)
ResNet50	$85.43^{\uparrow 0.00}$	$96.70^{\uparrow 0.00}$
+ NL	$85.34^{\downarrow 0.09}$	$\mathbf{96.77}^{\uparrow \mathbf{0.07}}$
+ NS	$85.54^{\uparrow 0.11}$	$96.56^{\downarrow 0.14}$
+ A^2	$85.91^{\uparrow 0.48}$	$96.56^{\downarrow 0.14}$
+ CGNL	$\mathbf{86.14}^{\uparrow \mathbf{0.71}}$	$96.34^{\downarrow 0.36}$
+ Ours	$86.02^{\uparrow 0.59}$	$96.65^{\downarrow 0.05}$

References

[1] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *Neural Information Processing Systems (NeurIPS)*, pages 352–361, 2018. 2

[2] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[3] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018. 4

[4] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 603–612, 2019. 2

[5] Alexander Kozlov, Vadim Andronov, and Yana Gritsenko. Lightweight network architecture for real-time action recognition. *arXiv preprint arXiv:1905.08711*, 2019. 3, 4

[6] Ron Levie, Federico Monti, Xavier Bresson, and Michael M Bronstein. Cayleynets: Graph convolutional neural networks

with complex rational spectral filters. *IEEE Transactions on Signal Processing (TSP)*, 67(1):97–109, 2018. 3

[7] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5533–5541, 2017. 3, 4

[8] Yunzhe Tao, Qi Sun, Qiang Du, and Wei Liu. Nonlocal neural networks, nonlocal diffusion and nonlocal modeling. In *Neural Information Processing Systems (NeurIPS)*, pages 496–506, 2018. 1

[9] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 1

[10] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu. Compact generalized non-local network. In *Neural Information Processing Systems (NeurIPS)*, pages 6510–6519, 2018. 2, 4