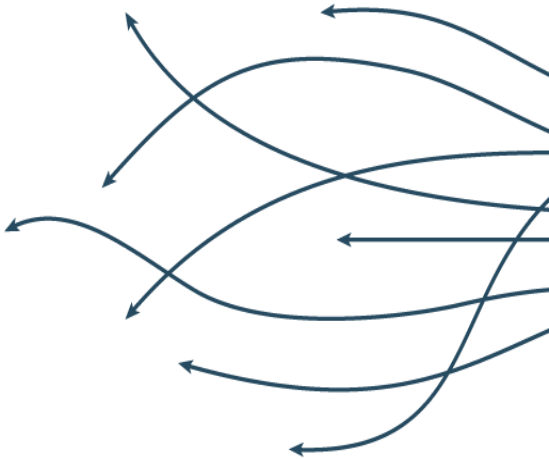
Spatial-Spectral Transforms for Non-Local Deep Neural Networks

Yadong Mu
Wangxuan Institute of Computer Technology
Peking University

Joint work with my graduate students
Lu Chi, Guiyu Tian, Yongzhi Li, Borui Jiang



This talk is a wrap-up of the following papers

- Lu Chi, Guiyu Tian, Yadong Mu*, Lingxi Xie, Qi Tian, Fast Non-Local Neural Networks with Spectral Residual Learning, ACM Multimedia 2019. code: https://github.com/1820366459/SRL-Pose-Estimation
- Lu Chi, Zehuan Yuan, Yadong Mu*, Changhu Wang, Non-Local Neural Networks with Grouped Bilinear Attentional Transforms, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020. (code is available upon request)
- Lu Chi, Borui Jiang, Yadong Mu*, Fast Fourier Convolution, Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS) 2020. (code will be released soon)

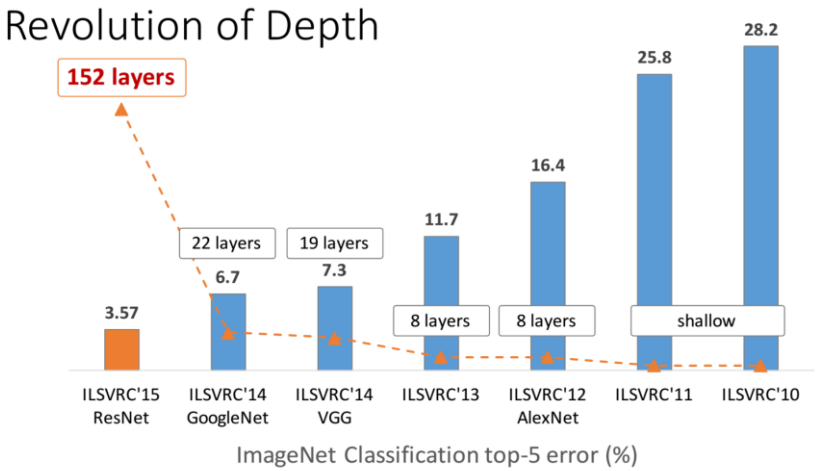
(* denotes the corresponding author.)

PDF-version manuscripts can be accessed at http://www.muyadong.com/publication.html

The Revolution of Deep Neural Networks

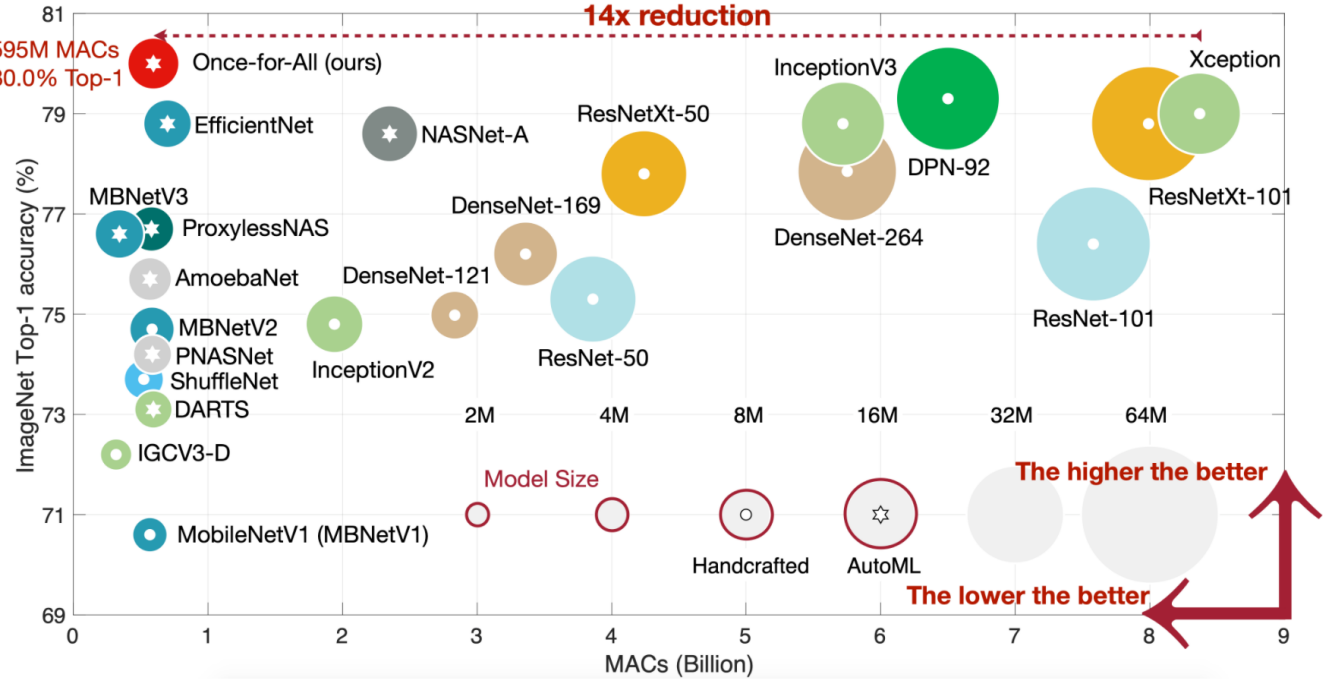
- Now most models in computer vision are named as X-Net

What is Deep Learning?
Photo by Kiran Foster, some rights reserved.
http://machinelearningmastery.com/what-is-deep-learning/



Necessity of Network Architecture Design

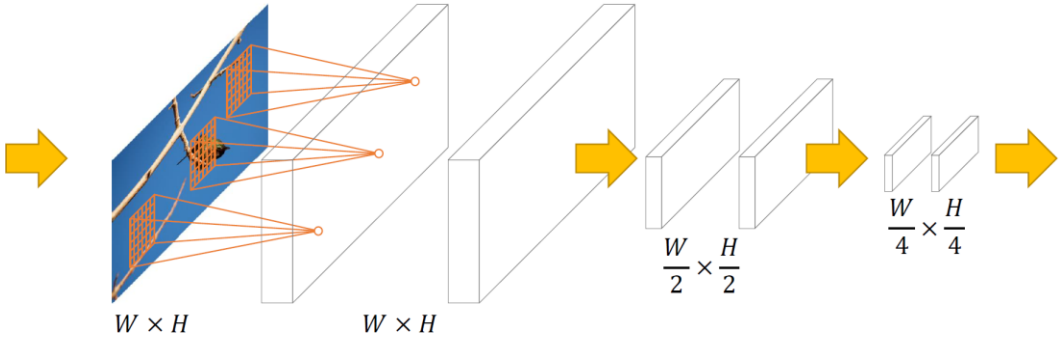
- DNNs (seemingly) can fit everything. Why work on network design?
- Accuracy-complexity tradeoff, data type, application task etc.



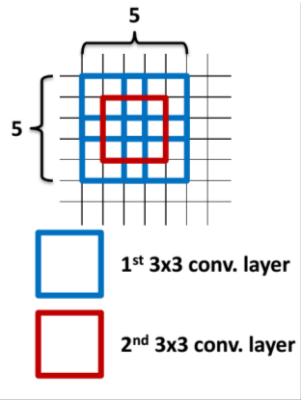
Cai et al., Once for All: Train One Network and Specialize it for Efficient Deployment, ICLR 2020

Receptive Field in Deep Networks

- Mainstream networks in CV adopt "tiny receptive field + large depth"



√ translation-invariant
√ shared parameters

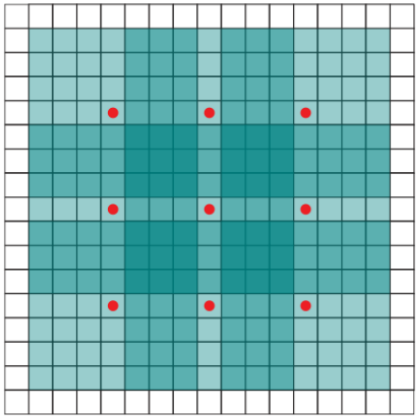


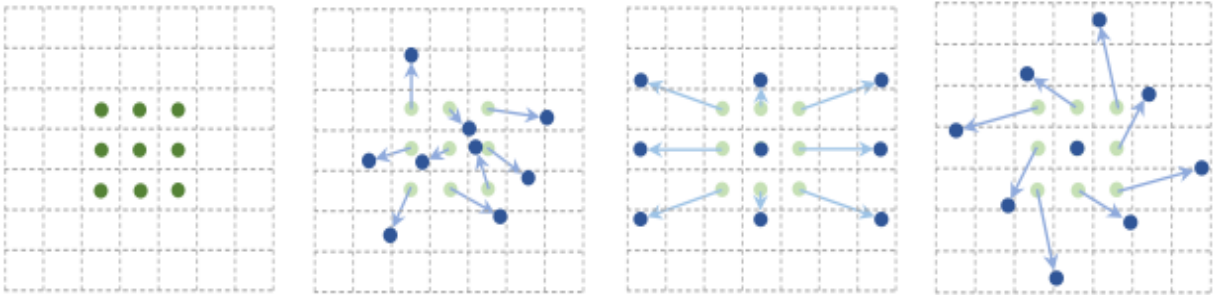
Receptive Field in Deep Networks

- Receptive field is crucial for context-sensitive tasks (e.g., image segmentation).

Enlarge Receptive Fields

- Typical large receptive field techniques:
  - dilation convolution
  - Deformable convolution
  - …
- Weakness
  - insufficient receptive field, still local

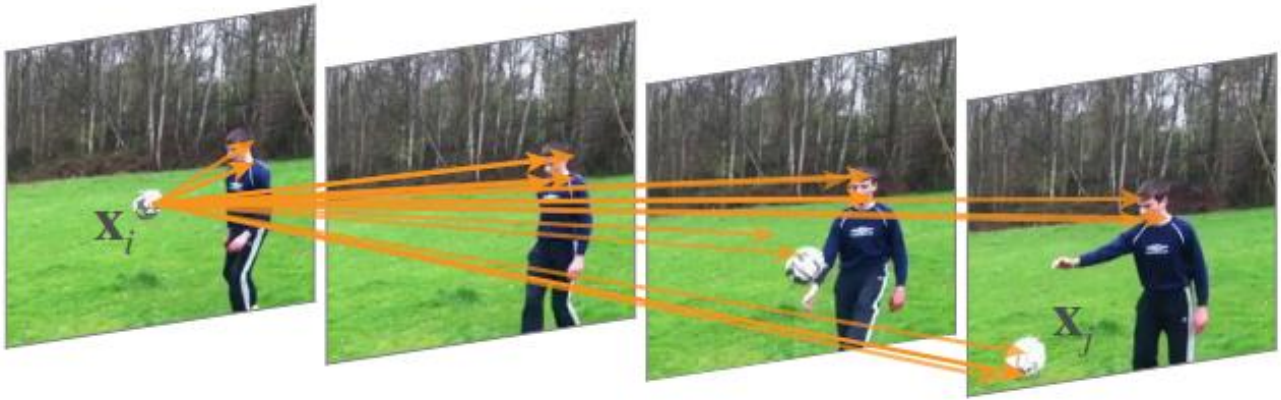




(a) (b) (c) (d)

Non-Local Neural Networks

- Connect a neuron with all others in the same layer
- Computations are based on self-attention
- Pros: full receptive field
- Cons: Slow

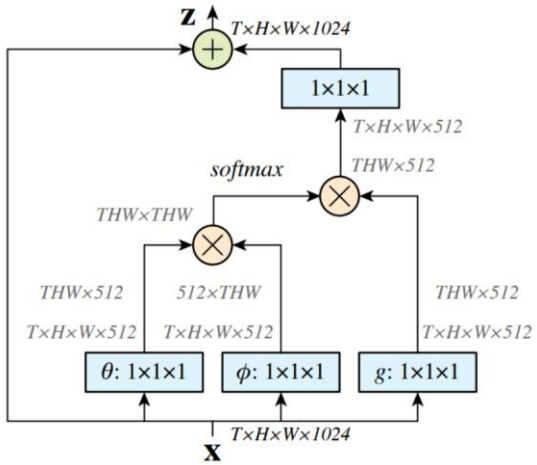


Xiaolong Wang et al., Non-Local Neural Networks, CVPR 2018

Non-Local Operator

- Following the non-local mean operation, a generic non-local operation in deep neural networks is as:

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j)$$



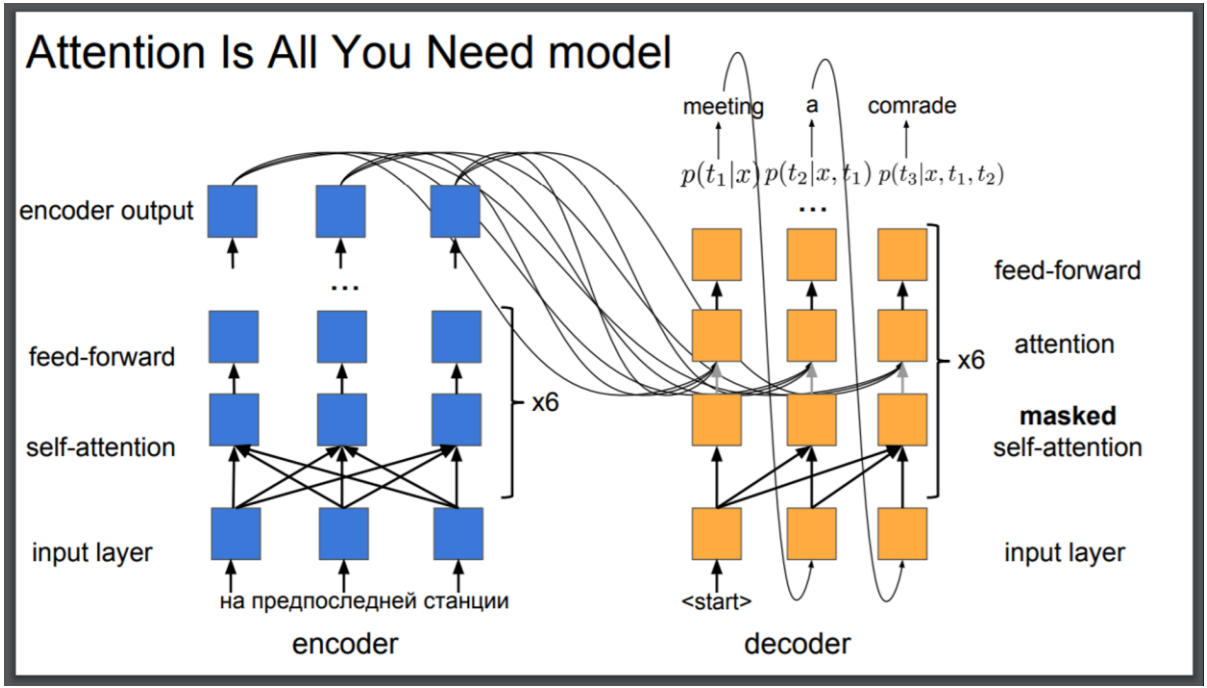
Embedded Gaussian. A simple extension of the Gaussian function is to compute similarity in an embedding space. In this paper we consider:

$$f(\mathbf{x}_i, \mathbf{x}_j) = e^{\theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j)}. \quad (3)$$

Here $\theta(\mathbf{x}_i) = W_\theta \mathbf{x}_i$ and $\phi(\mathbf{x}_j) = W_\phi \mathbf{x}_j$ are two embeddings. As above, we set $\mathcal{C}(\mathbf{x}) = \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j)$.

A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising, CVPR 2005

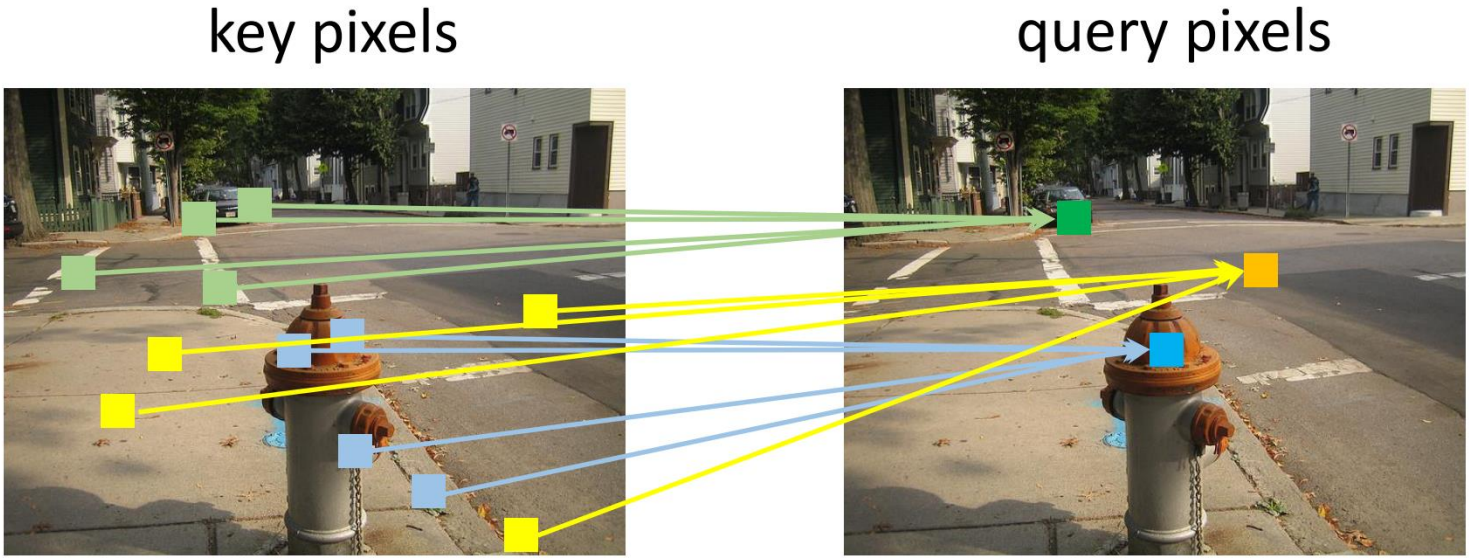
Non-Local Op. v.s. Transformer



A. Vaswni et al. Attention is all your need. NIPS'2017

Non-Local Neural Networks

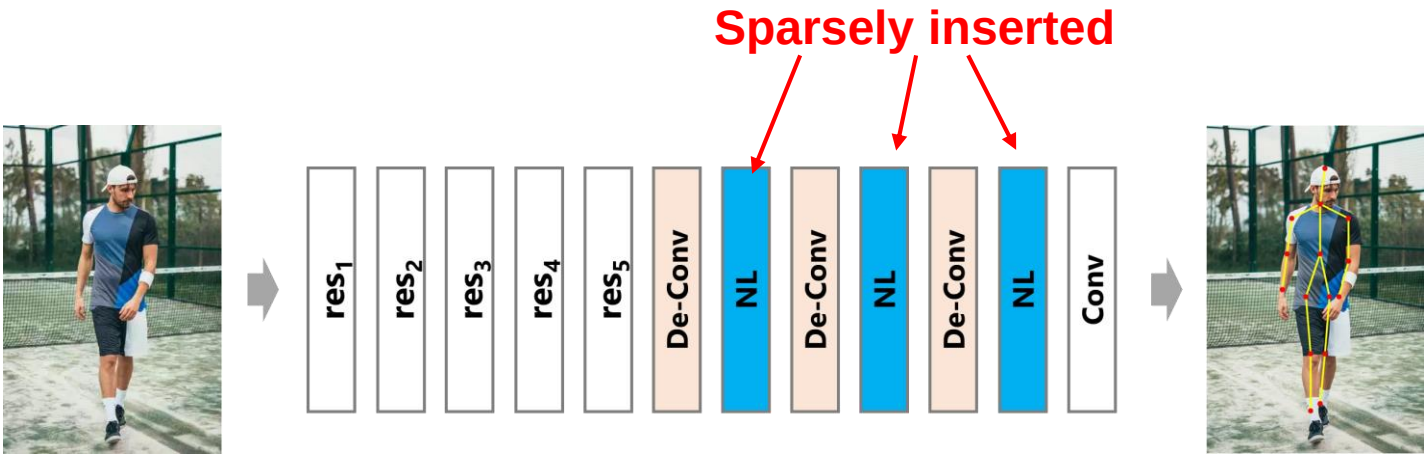
- Expectation: different query pixels impacted by different sets of key pixels



Xiaolong Wang et al., Non-Local Neural Networks, CVPR 2018

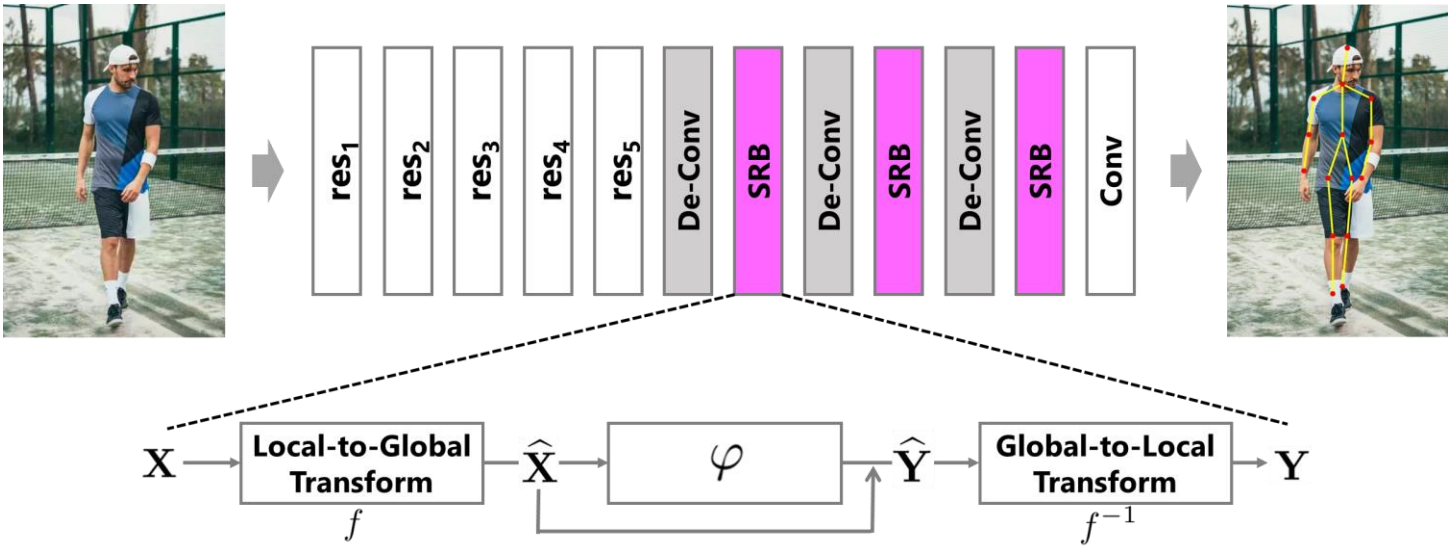
Non-Local Neural Networks

- Connect a neuron with all others in the same layer
- Computations are based on self-attention
- Pros: full receptive field
- Cons: Slow



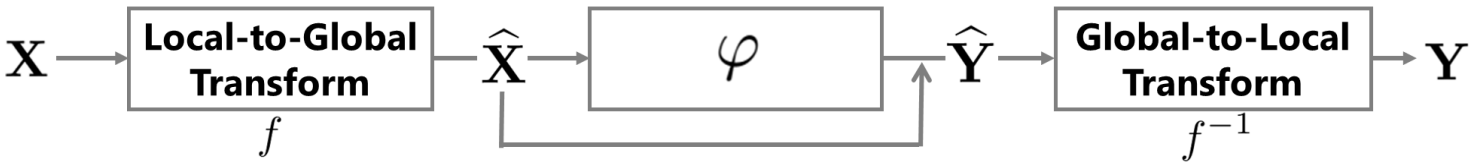
Our Idea: Use Spatial-Spectral Transforms

- Paired local-to-global (spatial->spectral) and global-to-local (spectral-to-spatial) transforms
- Wrapped as spectral residual block (SRB) and sparsely inserted

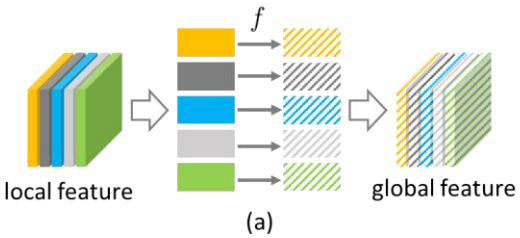


Lu Chi, Guiyu Tian, Yadong Mu(*), Lingxi Xie, Qi Tian, Fast Non-Local Neural Networks with Spectral Residual Learning, ACM Multimedia 2019.

Spectral Residual Block (SRB)



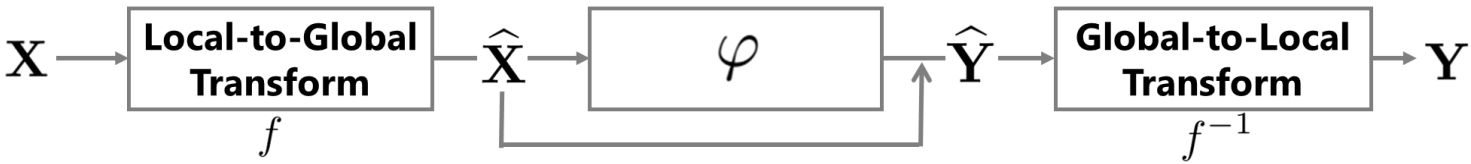
- Local / Global transforms (f / f^{-1})



$$f(\mathbf{X}) = concat\left(\mathbf{P}\mathbf{X}_1\mathbf{Q}, \ldots, \mathbf{P}\mathbf{X}_C\mathbf{Q}\right),$$
$$f^{-1}(\hat{\mathbf{Y}}) = concat\left(\mathbf{P}^*\hat{\mathbf{Y}}_1\mathbf{Q}^*, \ldots, \mathbf{P}^*\hat{\mathbf{Y}}_C\mathbf{Q}^*\right)$$

A complex square matrix $\mathbf{U}$ is unitary if its conjugate transpose matrix $\mathbf{U}^*$ is also its inverse, namely $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Spectral Residual Block (SRB)

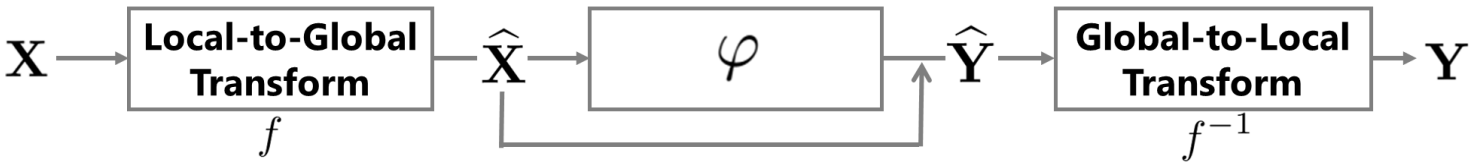


- Local / Global transforms (f / f^{-1})

$$
\begin{aligned}
f \cdot f^{-1}(\phi(\cdot)) &= \mathbf{P}\left[\mathbf{P}^{*}\ \phi(\cdot)\ \mathbf{Q}^{*}\right]\mathbf{Q} \\
&= \left(\mathbf{P}P^{*}\right)\phi(\cdot)\left(\mathbf{Q}^{*}\mathbf{Q}\right) = \phi(\cdot)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{Y} &= f^{-1}\left(\varphi(f(\mathbf{X})) + f(\mathbf{X})\right) \\
&= f^{-1}\left(\varphi(f(\mathbf{X}))\right) + \mathbf{X},
\end{aligned}
$$

Spectral Residual Block (SRB)



- Local / Global transforms (f / f^{-1})

$$\begin{aligned}\mathbf{Y} &= f^{-1}\left(\varphi(f(\mathbf{X})) + f(\mathbf{X})\right) \\ &= f^{-1}\left(\varphi(f(\mathbf{X}))\right) + \mathbf{X},\end{aligned}$$

$$\mathbf{Y} - \mathbf{X} = f^{-1}(\varphi(f(\mathbf{X}))).$$

$$\varphi(f(\mathbf{X})) = f(\mathbf{Y} - \mathbf{X}).$$

Conducting residual learning in the spectral (globalized) domain

L2G / G2L Transforms

- Fast Fourier Transform Matrix (FFT)

$$F_{2d}(X_c) = F_H \times X_c \times F_W$$
$$\mathrm{F}_{2d}^{-1}(\hat{Y}_c) = F_H^* \times \hat{Y}_c \times F_W^*$$

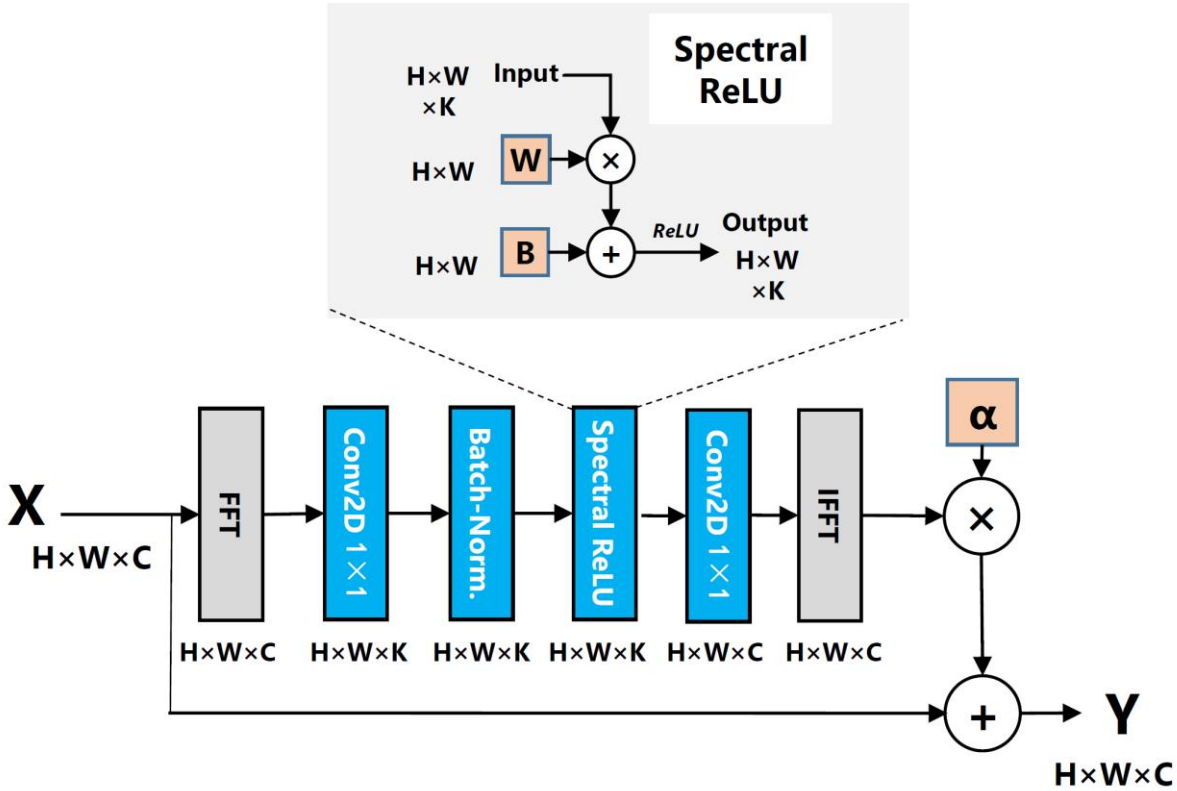
- Learnable Orthogonal Matrix (LO)
  - QR decomposition to re-orthogonalize P,Q

$$\begin{aligned} f(\mathbf{X}) &= concat\left(\mathbf{P}\mathbf{X}_1\mathbf{Q}, \ldots, \mathbf{P}\mathbf{X}_C\mathbf{Q}\right), \\ f^{-1}(\widehat{\mathbf{Y}}) &= concat\left(\mathbf{P}^*\widehat{\mathbf{Y}}_1\mathbf{Q}^*, \ldots, \mathbf{P}^*\widehat{\mathbf{Y}}_C\mathbf{Q}^*\right) \end{aligned}$$

$$P = \overline{P}R_1 \qquad Q = \overline{Q}R_2$$
$$P \leftarrow \overline{P} \qquad Q \leftarrow \overline{Q}$$

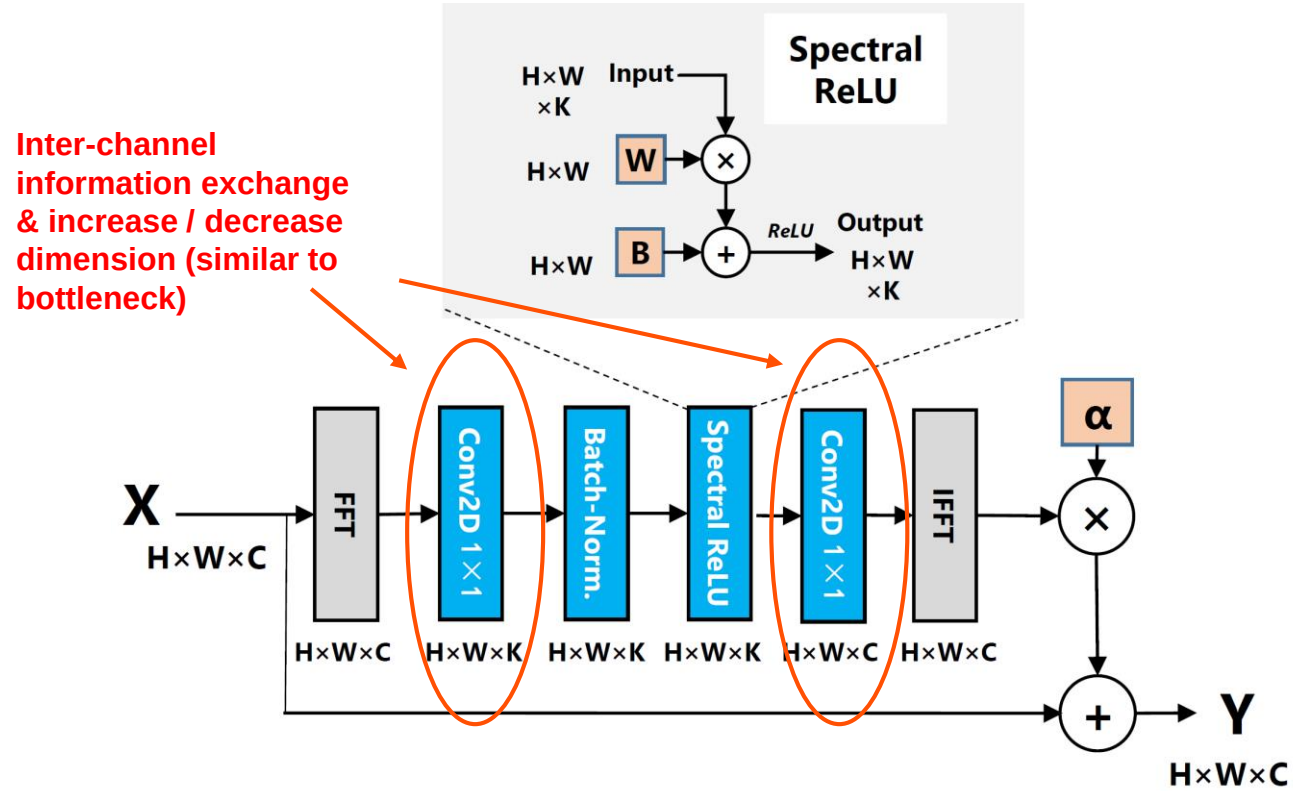
Instances of Phi

- Dependent on task



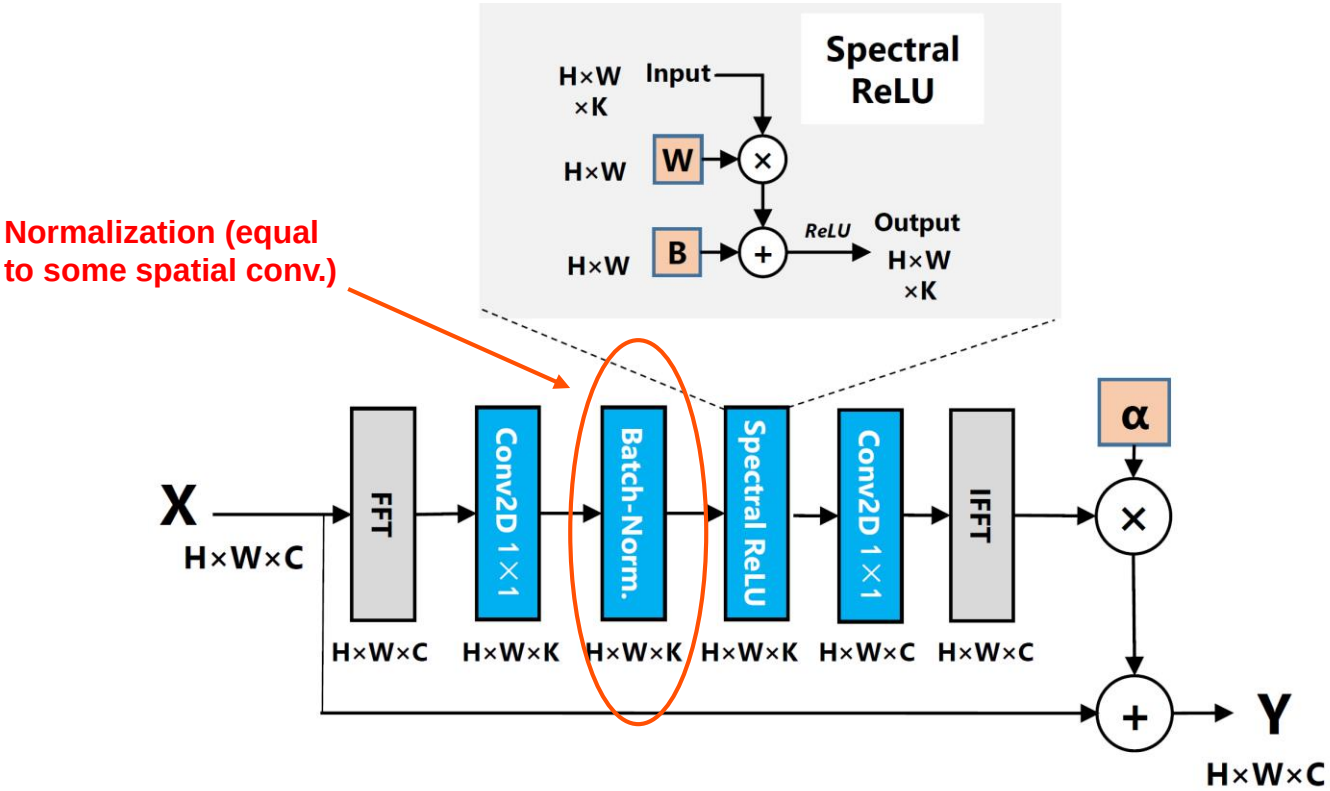
Instances of Phi

- Dependent on task



Instances of Phi

- Dependent on task

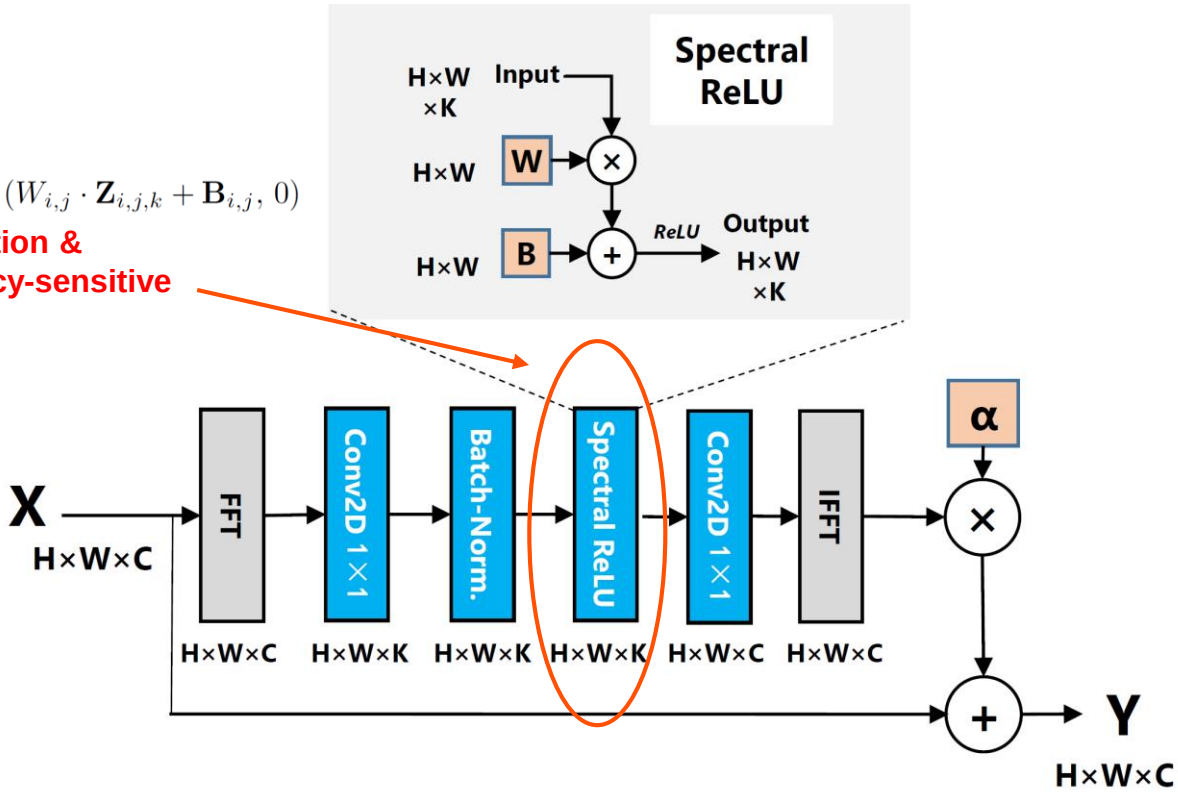


Instances of Phi

- Dependent on task

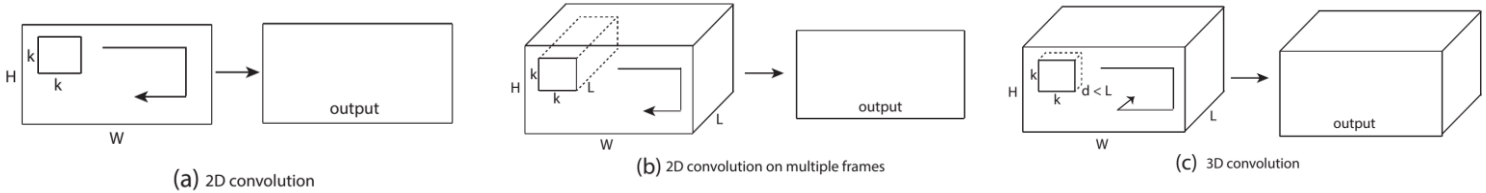
$$\mathbf{Z}_{i,j,k} \leftarrow \max\left(W_{i,j} \cdot \mathbf{Z}_{i,j,k} + \mathbf{B}_{i,j},\, 0\right)$$

Convolution & Frequency-sensitive filtering



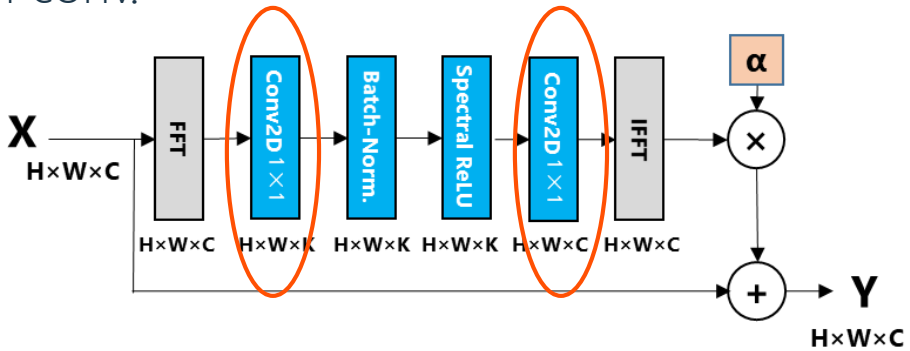
Extension to Spatio-Temporal Convolutions

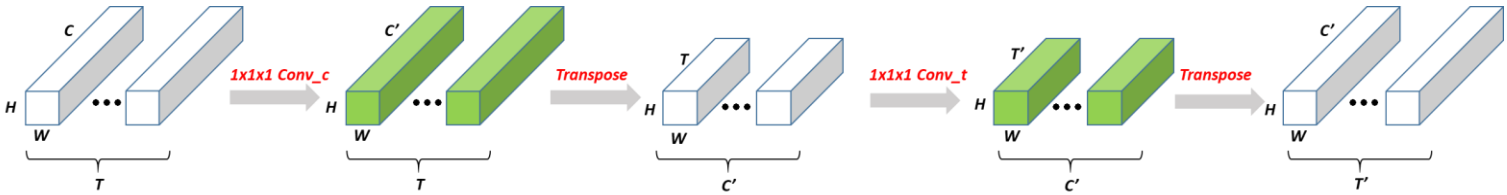
- 3D convolution + pooling



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

- Replace 1*1 conv.





Time Complexity

- H, W – image resolution
- C - #channel
- T - #frames in a stack (typically 8)

Method	Time Complexity	
	2D Convolution	3D Convolution
NL [8]	$\mathcal{O}(CH^2W^2)$	$\mathcal{O}(CH^2W^2T^2)$
A^2 [22]	$\mathcal{O}(C^2HW)$	$\mathcal{O}(C^2HWT)$
CGNL [21]	$\mathcal{O}(CHW(P+1))$	$\mathcal{O}(CHWT(P+1))$
RCCA [19]	$\mathcal{O}(CHW(H+W))$	N/A
SRB-LO	$\mathcal{O}(CHW(H+W))$	$\mathcal{O}(CHWT(H+W)+CHWT^2)$
SRB-FFT	$\mathcal{O}(CHW\log(HW))$	$\mathcal{O}(CHWT\log(HW)+CHWT^2)$

Experiments: Human Pose Estimation

- Conducted on MS-COCO benchmark

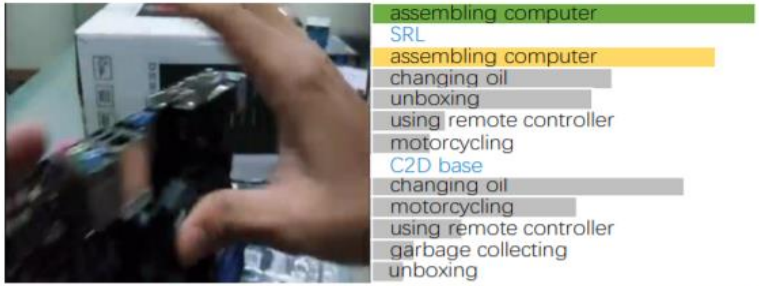
input size	Non-Local Block	GFLOPs(Δ)	# Params((MΔ))	AP
256 × 192	-	0	0	70.4
	NL [60]	2.90	0.39	70.6
	A^2 [9]	0.28	0.20	70.5
	CGNL [67]	0.39	0.31	70.6
	RCCA [21]	1.20	0.92	70.6
	SRL-FFT	**0.09**	**0.10**	**70.9**
384 × 288	-	0	0	72.2
	NL [60]	13.26	0.39	72.8
	A^2 [9]	0.62	0.20	72.8
	CGNL [67]	0.88	0.31	72.4
	RCCA [21]	2.76	0.92	72.8
	SRL-FFT	**0.20**	**0.10**	**73.3**

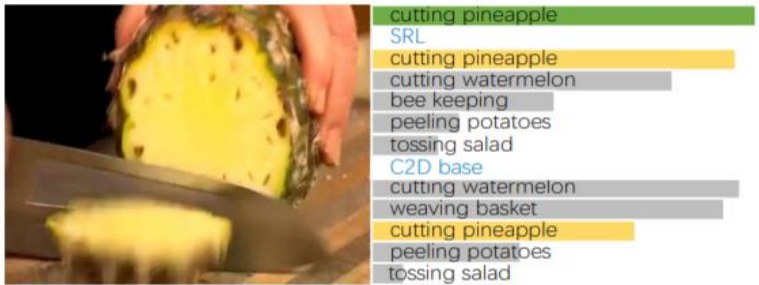
Table 2: Comparisons with other non-local blocks on COCO keypoint dataset.

Experiments: Video Classification

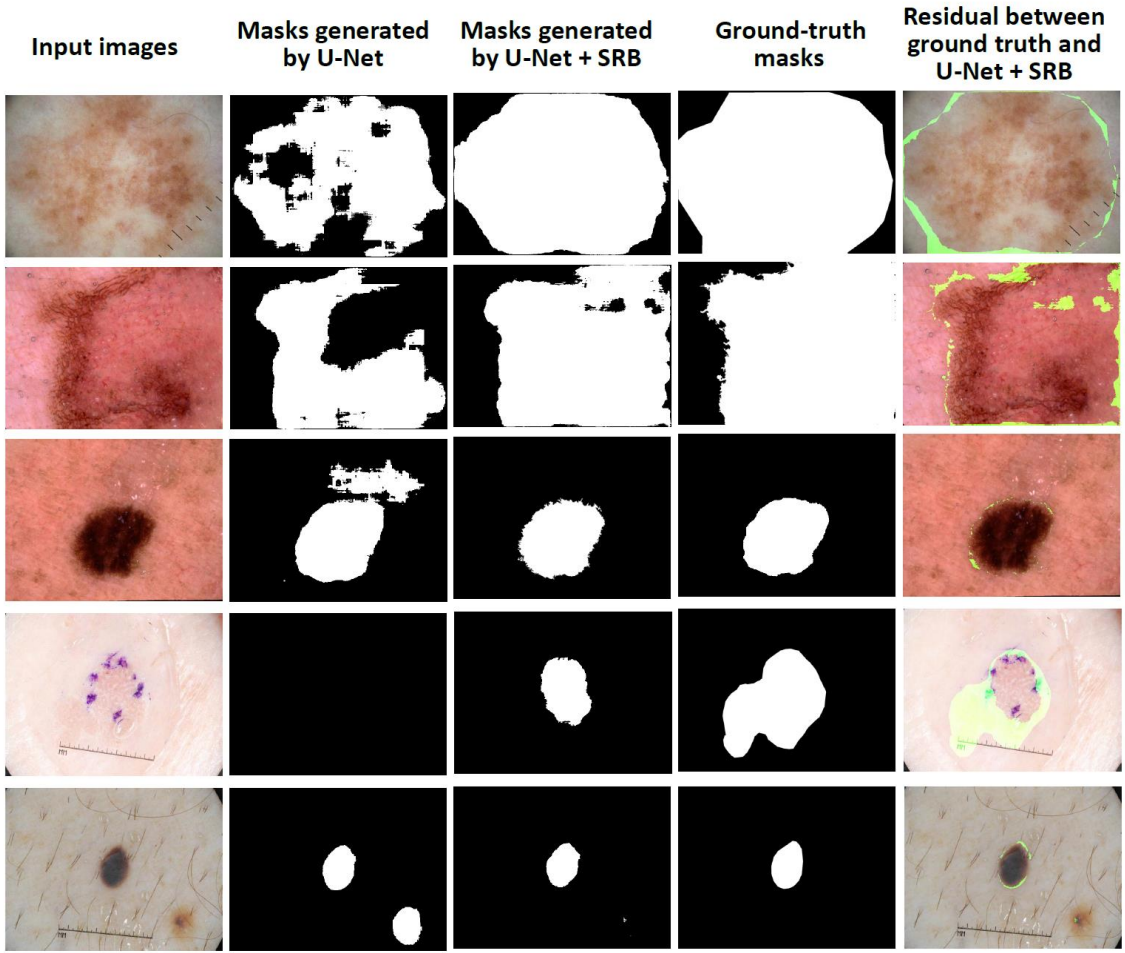
Method	#Frames	Top-1 Accuracy	Top-5 Accuracy
I3D [5]	64	71.1	89.3
ARTNet [58]	16	70.7	89.3
S3D [66]	64	72.2	90.6
R(2+1)D [54]	-	72.0	90.0
C2D base	8	70.5	89.4
C2D base + SRL-FFT	8	71.9	90.3
C2D base + SRL-FFT*	8	**72.7**	**90.9**

Table 1: Comparisons with state-of-the-art methods on the Kinetics video

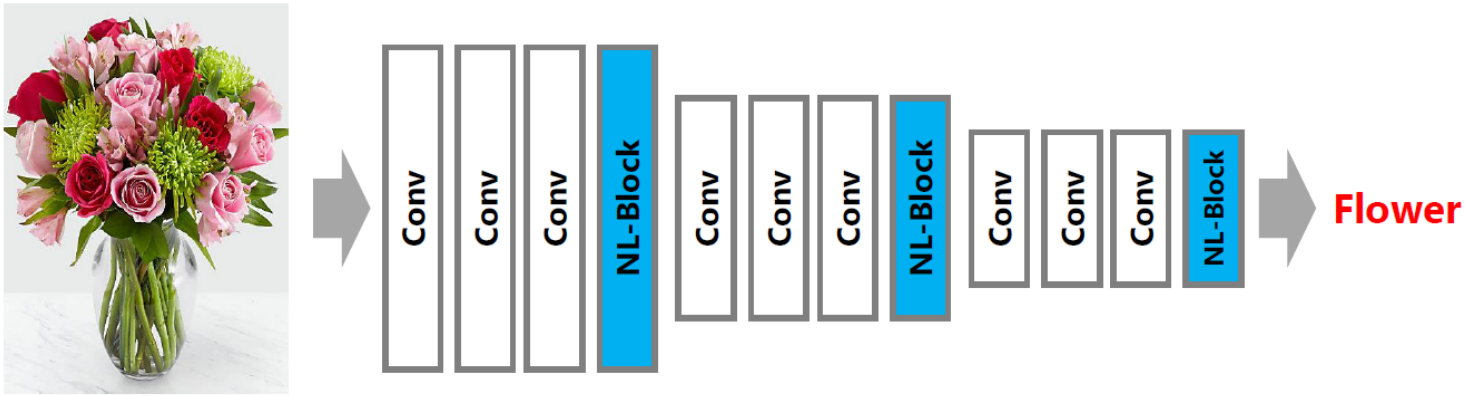




Experiments: Medical Image Analysis



Ablation Study



- Location to insert SRBs
- #inserted SRBs

Ablation Study

+N Blocks	Top-1	Top-5
C2D base	62.75	83.82
+1 block	64.23	85.16
+2 blocks	**64.79**	**85.71**
+4 blocks	64.65	85.52

Table 3: Effect of key factors in SRB for video classification. Left: Different number of SRL blocks are added.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
baseline	72.2	89.3	78.9	68.1	79.7	77.6
baseline*	66.3	87.1	71.4	61.3	74.6	72.0
dilated conv	72.3	**89.4**	79.1	68.1	80.1	77.6
+3 NL [60]	72.7	88.9	79.1	68.9	79.9	77.9
+3 SRL-FFT	73.4	89.3	79.8	69.5	80.7	78.8
+4 SRL-FFT	**73.7**	**89.4**	**80.3**	**69.7**	**81.2**	**78.9**

Table 4: Performance of human key point detection by varying the inserting number of SRBs.

Ablation Study

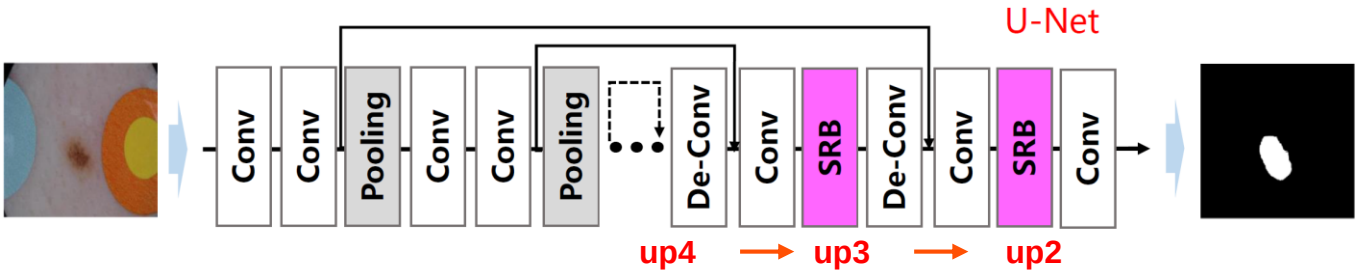
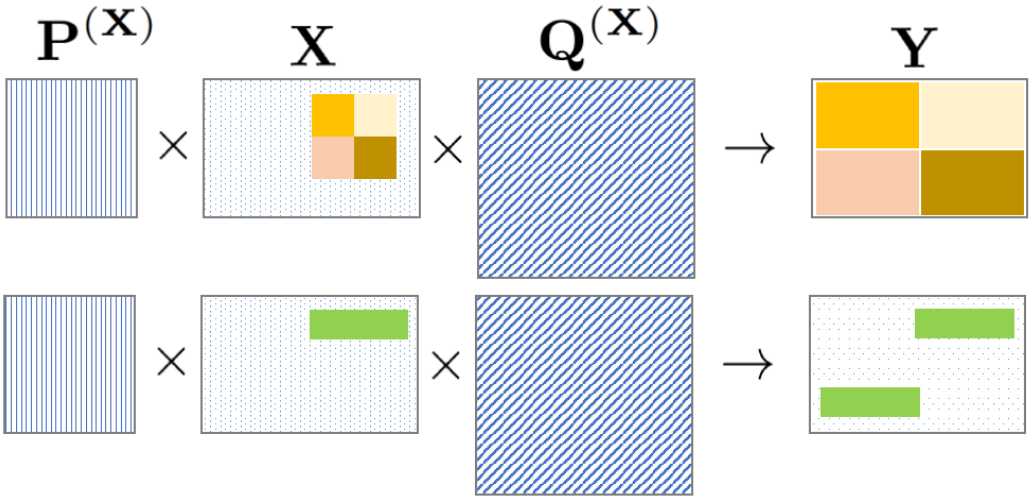


TABLE 13: Ablation study about the inserting locations of SRBs in skin lesion segmentation. U-Net* denotes the model removed the last maxpooling layer from the original U-Net. upN denotes the N_{th} up-sampling layer in U-Net's decoder. Among all, up4 is closest to the encoder (*i.e.*, most shallow one). 'a/', 'b/' mean 'after', 'before' respectively.

Model	Mean Acc	Over Acc	Mean IoU	FW IoU
U-Net*	90.60	93.79	83.10	88.65
+1 a/ up2	92.14	95.05	86.12	90.77
+1 a/ up3	91.87	95.48	86.99	91.44
+1 a/ up4	91.87	95.54	87.14	91.55
+1 b/ up4	**92.38**	**95.61**	**87.42**	**91.70**
+1 NL [8]	91.92	95.33	86.69	91.22
+1 A^2 [22]	91.36	95.09	86.03	90.78

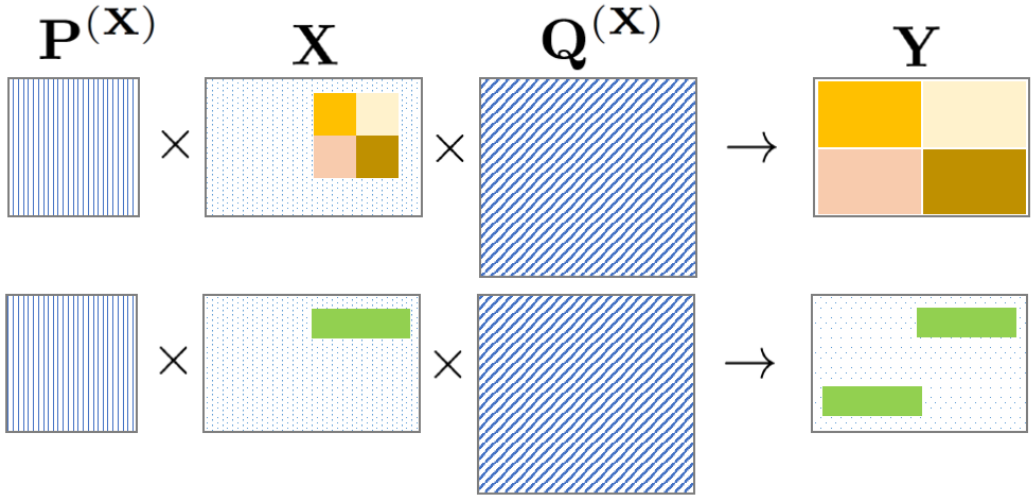
From SRNet to Bilinear Attentional Transform

- Previous transforms are image-independent
  - LO variant learns parameters from data, but the bilinear matrices do not change for specific input image or video
- Let's rethink what bilinear transform actually does



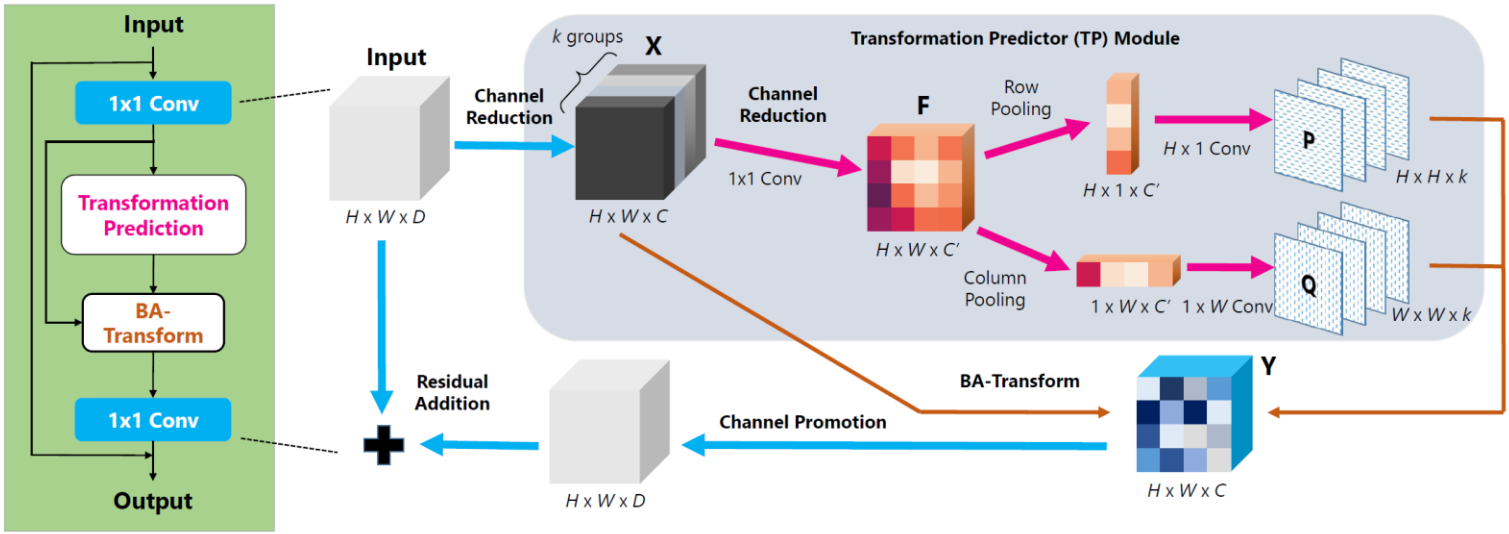
Lu Chi, Zehuan Yuan, Yadong Mu(*), Changhu Wang, Non-Local Neural Networks with Grouped Bilinear Attentional Transform, under review

From SRNet to Bilinear Attentional Transform

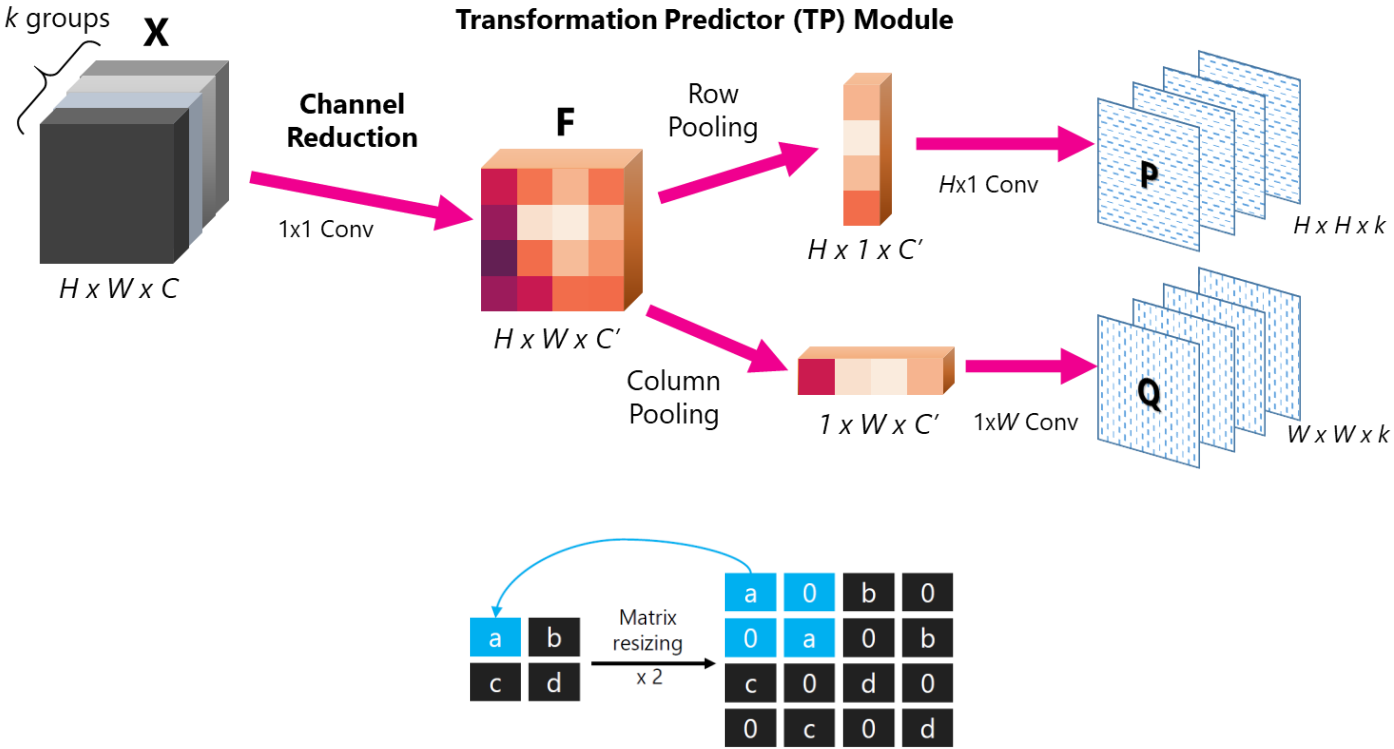


- Allow P, Q to be (almost) general matrices
- Make it input data (i.e., X) dependent
- Single (rather than paired) transform at the entrance of a block

Bilinear Attentional (BA) Transform



Bilinear Attentional (BA) Transform



Complexity

- Comparable to classic NL, but performances are much better.

	NL block [56]	BAT-Block
#Params	$2C^2$	$\frac{5}{4}C^2 + \frac{1}{2}CC' + 2C'ks^3$
FLOPs	$2C^2HW + CH^2W^2$	$\frac{5}{4}C^2HW$ $+\frac{1}{2}CHW(H+W)$ $+\frac{1}{2}CC'HW + 2C'ks^3$

Table 1: **Complexity analysis.** For brevity, here we set $s_h = s_w = s$ and $C = D/2$ for BAT-Block, which is also consistent with experiments in Section 4.

Experiments

Backbone	Method	GFLOPs(Δ)	#Params(Δ)	Top-1
ResNet-18	baseline	-	-	70.2
	+NL	0.23	0.17	70.9
	+BAT	0.03	0.13	**71.3**
ResNet-50	baseline	-	-	76.3
	+NL	3.55	7.36	77.5
	+BAT	1.30	4.67	**78.3**

Table 5: **Comparisons with NL block on ImageNet.**

Method	3D-Conv	GFLOPs	#Params	Top-1
A^2-Net [7]	Yes	40.8	-	74.6
Oct-I3D [6]	Yes	25.6	-	74.6
TSM [55]	No	32.8	24.3	74.1
GloRe [8]	Yes	28.9	-	75.1
C2D	No	19.6	24.3	71.9
I3D	Yes	28.4	28.4	72.6
C2D + NL	No	30.7	31.7	73.8
I3D + NL	Yes	39.5	35.4	73.5
C2D + BAT	No	24.8	29.2	74.5
I3D + BAT	Yes	33.6	32.9	74.8
C2D + 3D-BAT	No	24.8	29.2	75.2
C2D + 3D-BAT†	No	24.8	29.2	**75.9**

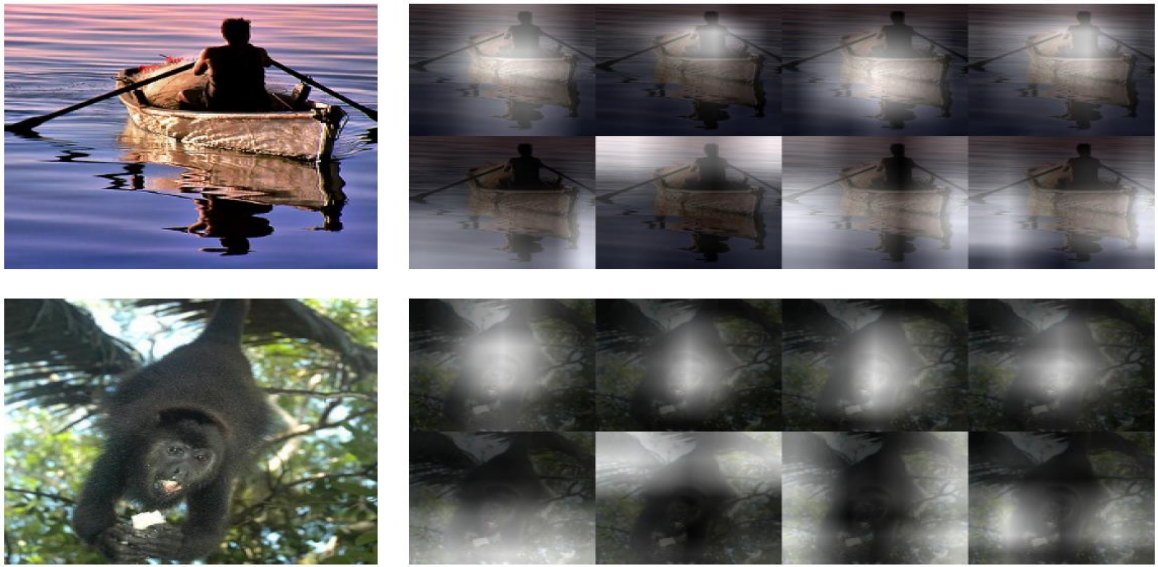
Table 8: **Results on Kinetics-400.** The first set is recent state-of-the-art, the second set is our re-implemented models, and the last set is our methods. The group number of spatial attention is 8 and that of the temporal attention is set to 4. All the models use ResNet-50 as backbone and 8 frames as input. "†" represents finetuning with TSN framework [55].

Visualization

- Which part of images are highlighted?

$$\mathbf{W} = \mathbf{P}^\intercal \mathbf{A} \mathbf{Q}^\intercal \tag{8}$$

Here $\mathbf{A} \in \mathbb{R}^{H \times W}$ is an all-ones matrix. $\mathbf{W}$ is the re-projected attention weight with shape $H \times W$. The results are normalized between 0 and 255 for visualization.

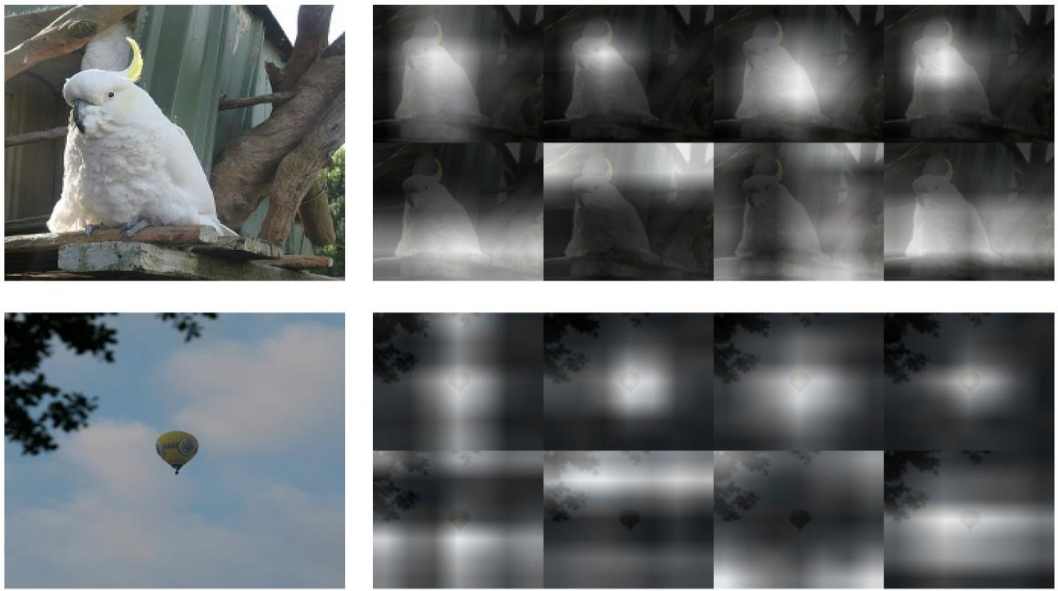


Visualization

- Which part of images are highlighted?

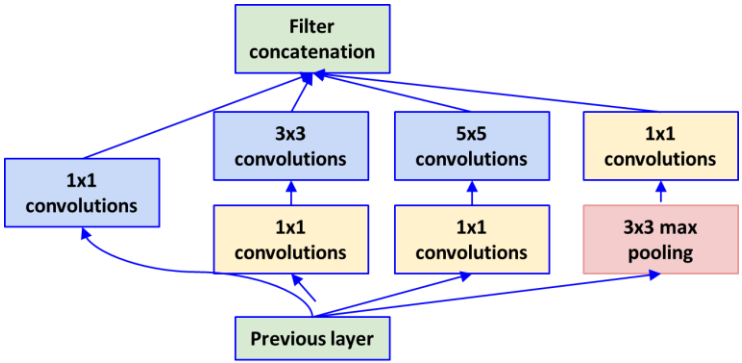
$$\mathbf{W} = \mathbf{P}^\intercal \mathbf{A} \mathbf{Q}^\intercal \tag{8}$$

Here $\mathbf{A} \in \mathbb{R}^{H \times W}$ is an all-ones matrix. $\mathbf{W}$ is the re-projected attention weight with shape $H \times W$. The results are normalized between 0 and 255 for visualization.

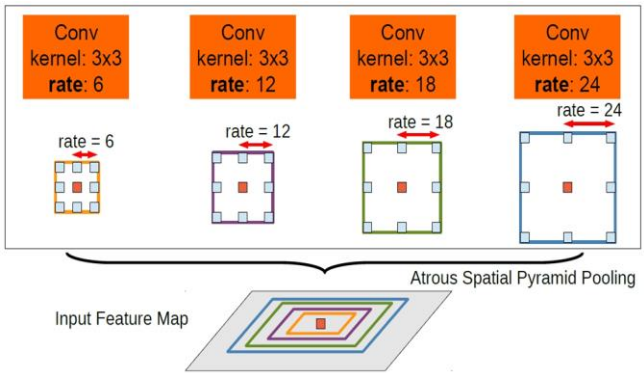


Fast Fourier Convolution

- Dense insertion into a neural pipeline (or, use a plug-and-play "convolution")
- Ensemble of multiple receptive fields



Inception Module in GoogleNet

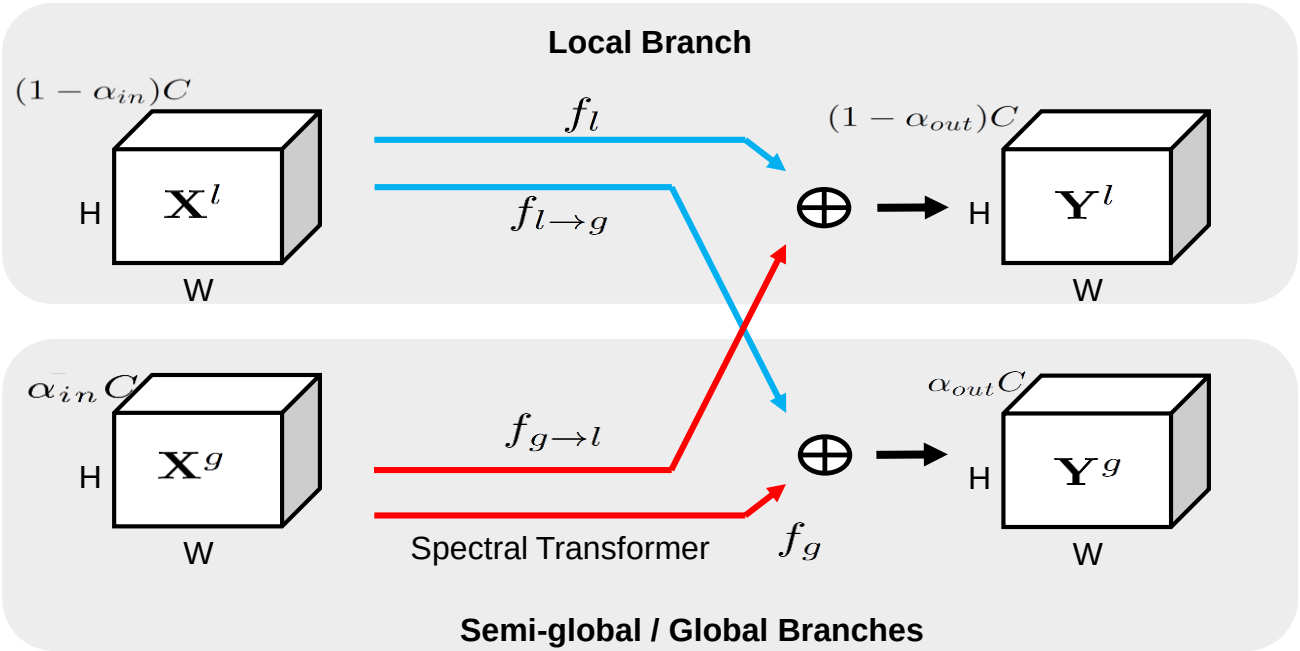


Atrous Spatial Pyramid Pooling (ASPP)

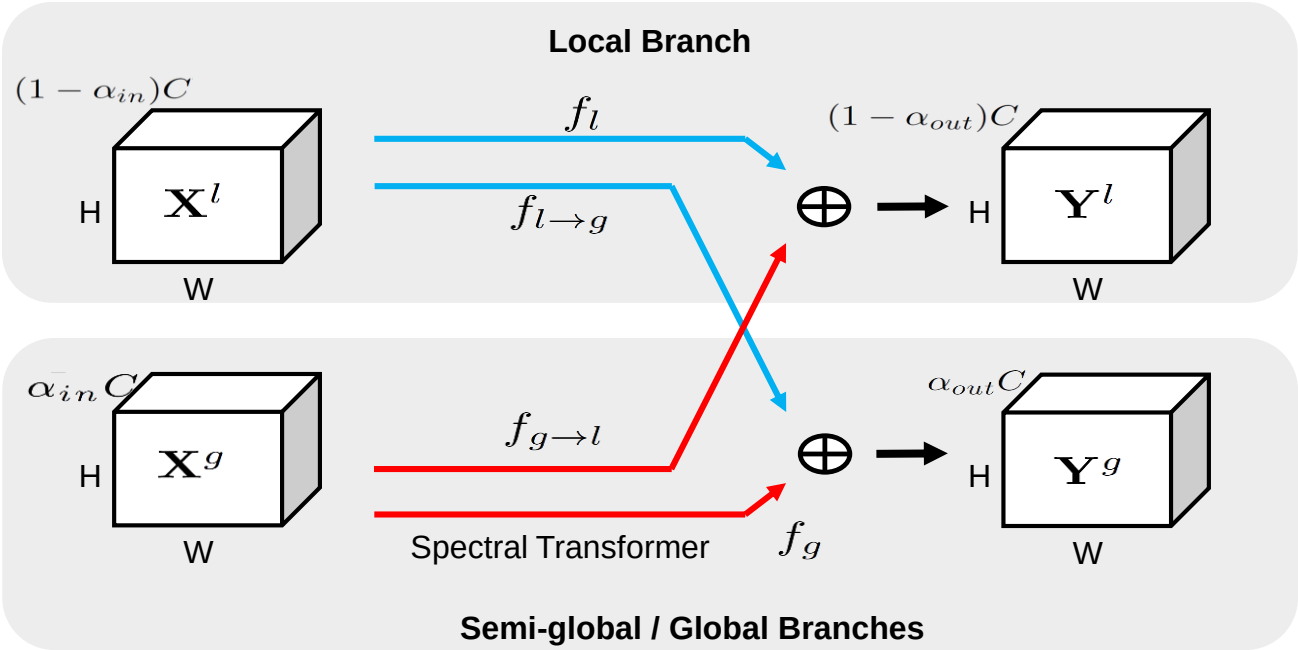
Lu Chi, Borui Jiang, Yadong Mu(*), Fast Fourier Convolution, to be submitted

Fast Fourier Convolution

- two inter-connected paths
  - a spatial (or local) path – vanilla convolutions on spatial domain
  - a spectral (or global) path - spectral operations



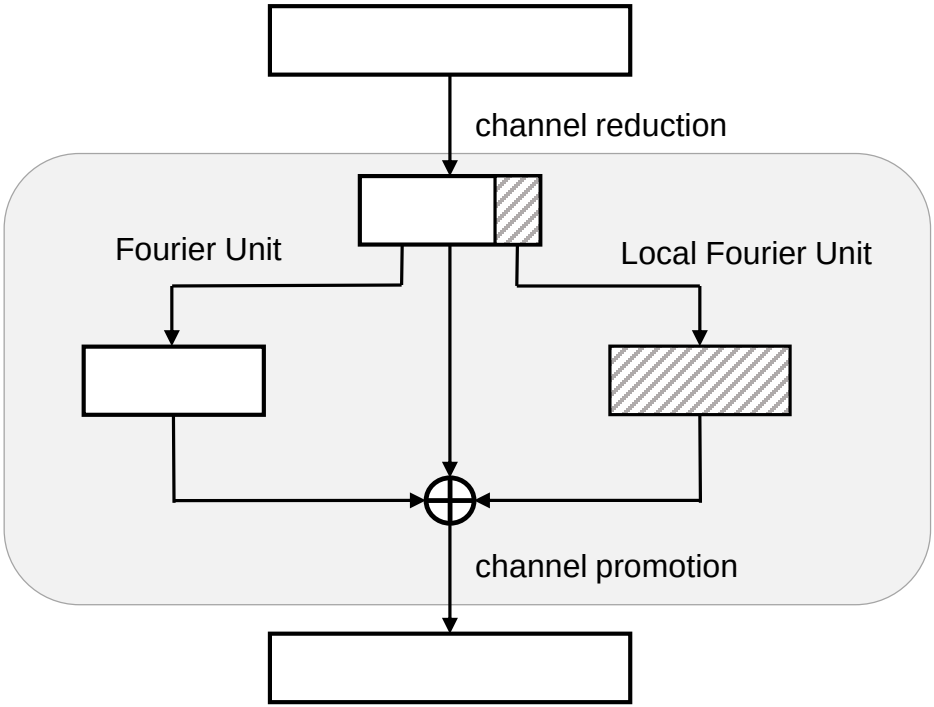
Fast Fourier Convolution



$$\mathbf{Y}^l = \mathbf{Y}^{l\to l} + \mathbf{Y}^{g\to l} = f_l(\mathbf{X}^l) + f_{g\to l}(\mathbf{X}^g),$$
$$\mathbf{Y}^g = \mathbf{Y}^{g\to g} + \mathbf{Y}^{l\to g} = f_g(\mathbf{X}^g) + f_{l\to g}(\mathbf{X}^l).$$

Fast Fourier Convolution

- Design of spectral transform f_g



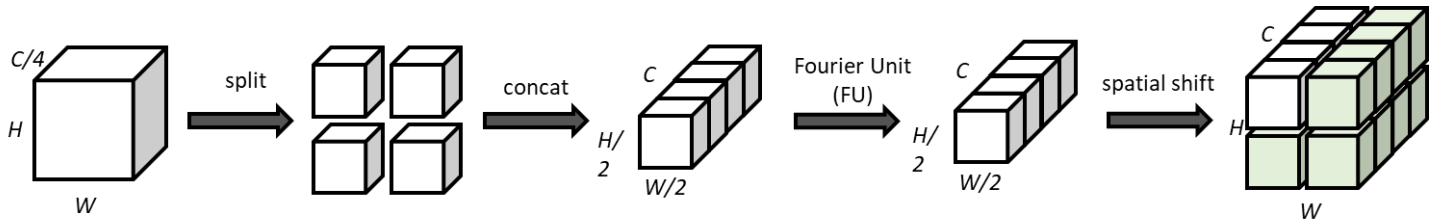
Global: Fourier Unit (FU)

```
def FU(x):
    # x: input features with shape [N,C,H,W]

    # y_r / y_i is the real / imaginary part of the results of FFT, respectively
    y_r, y_i = FFT(x) # y_r/y_i: [N,C,H,⌊W/2⌋+1]
    y = Concatenate([y_r, y_i], dim=1) # [N,C*2,H,⌊W/2⌋+1]
    y = ReLU(BN(Conv(y))) # [N,C*2,H,⌊W/2⌋+1]
    y_r, y_i = Split(y, dim=1) # y_r/y_i: [N,C,H,⌊W/2⌋+1]
    z = iFFT(y_r, y_i) # [N,C,H,W]

    return z
```

Semi-global: Local Fourier Unit (LFU)



Complexity Analysis

- Comparable to vanilla convolutions

	#Params	FLOPs
vanilla	$C_1C_2K^2$	$C_1C_2K^2HW$
$Y^{l\to l}$	$(1-\alpha)^2C_1C_2K^2$	$(1-\alpha)^2C_1C_2K^2HW$
$Y^{g\to g}$	$\frac{\alpha^2}{2}C_2(C_1+3C_2)$	$\frac{\alpha^2}{2}C_1C_2HW+\frac{13\alpha^2}{16}C_2^2HW$
$Y^{l\to g}$	$\alpha(1-\alpha)C_1C_2K^2$	$\alpha(1-\alpha)C_1C_2K^2HW$
$Y^{g\to l}$	$\alpha(1-\alpha)C_1C_2K^2$	$\alpha(1-\alpha)C_1C_2K^2HW$
FFC	$(1-\alpha^2)C_1C_2K^2+\alpha^2C_2(\frac{1}{2}C_1+\frac{3}{2}C_2)$	$(1-\alpha^2)C_1C_2K^2HW+\alpha^2C_2HW(\frac{1}{2}C_1+\frac{13}{16}C_2)$

Empirical Observations

Experimental Results on Kinetics-400

Method	GFLOPs	#Params	Top-1
TSM [17]	32.8	24.3	74.1
A^2-Net [5]	40.8	-	74.6
Oct-I3D [4]	25.6	-	74.6
GloRe [6]	28.9	-	75.1
C2D	19.6	24.3	71.9
I3D	28.4	28.4	72.6
I3D + NL	39.5	35.4	73.5
C2D + NL	30.7	31.7	73.8
FFC-C2D	20.2	24.9	73.5
FFC-C2D + NL	31.4	32.2	74.9
FFC-I3D + NL	40.2	35.9	75.1
FFC-I3D + NL †	40.2	35.9	**76.1**

Accuracy (FFC ~= NL), but Complexity (FFC << NL)

FFC and NL are complementary

Additionally, on ImageNet, using same parameters (*e.g.*, $\alpha = 0.25$), FFC with all cross-scale fusion achieves a top-1 accuracy of 77.6%. Removing global-to-local fusion or local-to-global fusion reduces the accuracy to 76.6%, 76.2% respectively. Removing $f_{l\to g}$ and $f_{g\to l}$ in Figure 1 only strikes an accuracy of 75.6%. This well validates the value of inter-path transitions.

Concluding Remarks

- Spatial-spectral transforms lead to efficient and effective non-local convolutions
- Critical engineering tricks: data-dependent transforms, multi-scale fusion
- Essentially, no theoretic analysis yet: denoising, low-rank analysis, spectral-spatial transform
- Neural attention on transformed domains?

Questions?

Email: myd@pku.edu.cn
Website: http://www.muyadong.com