Region-based Non-local Operation for Video Classification

Guoxi Huang, Adrian Bors

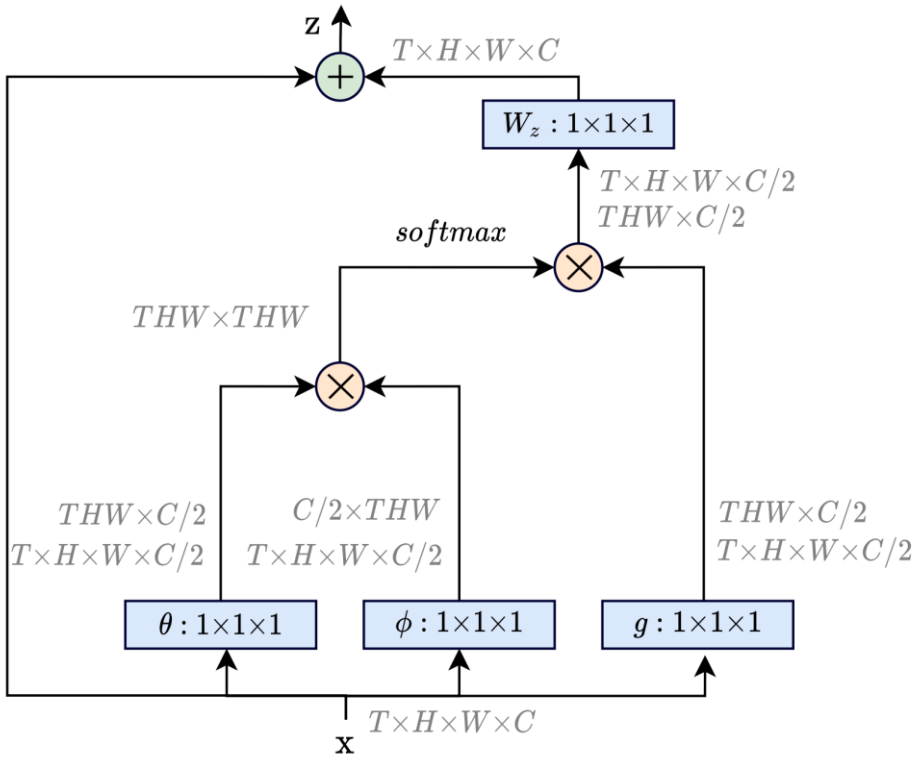
University of York

Problem description

- Convolution is a local operation.
- CNNs model long-range dependencies by deeply stacking convolution operation of small kernels.
- The interaction between two positions is not straightforward.
- Increasing the optimization difficulty.

Solution: We aim to design a non-local operation to directly capture long-range dependencies.

Non-local operation

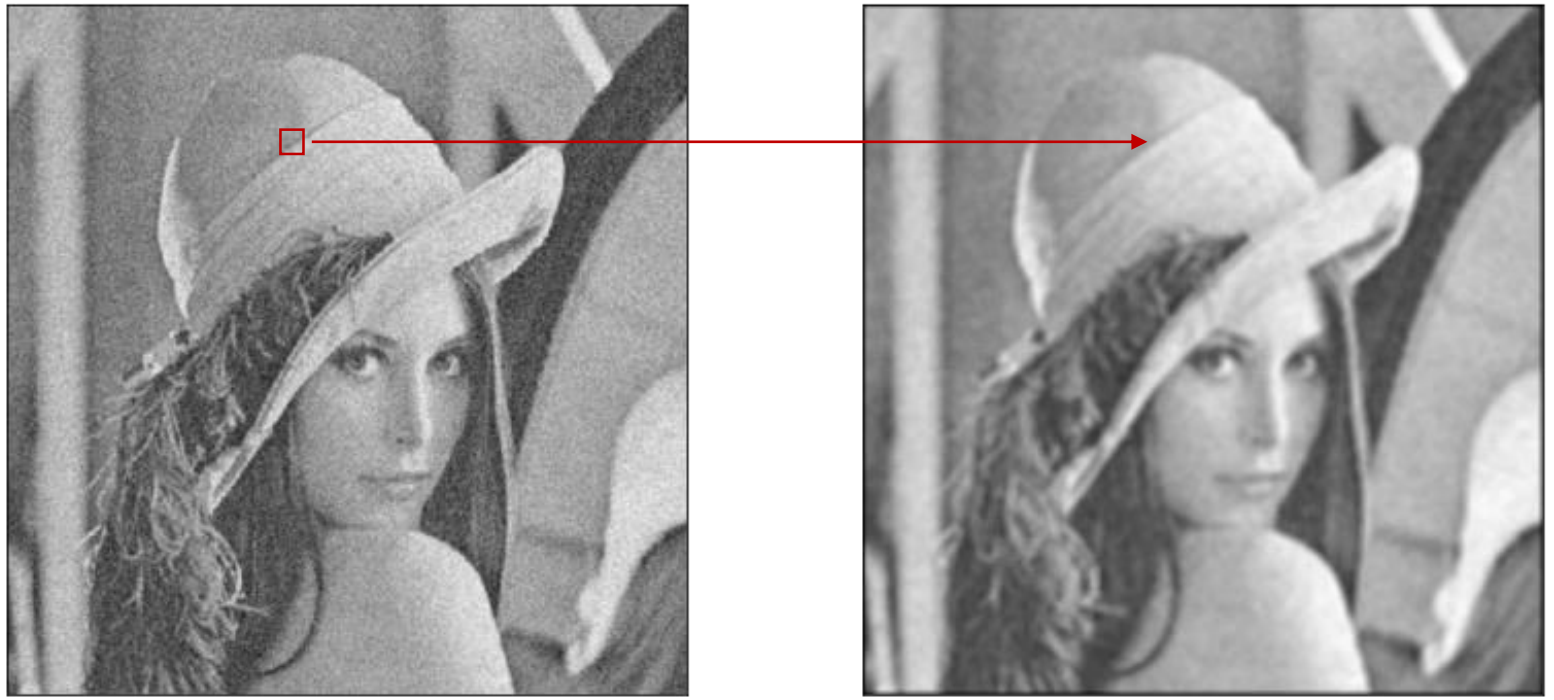


$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} w_{i,j} \mathbf{W}_g \mathbf{x}_j,$$

$$w_{i,j} = f(\mathbf{x}_i, \mathbf{x}_j),$$

- y_i – output features at position i
- x_i , x_j – input features at position j and j
- $w_{i,j}$ – weight computed by $f(,)$

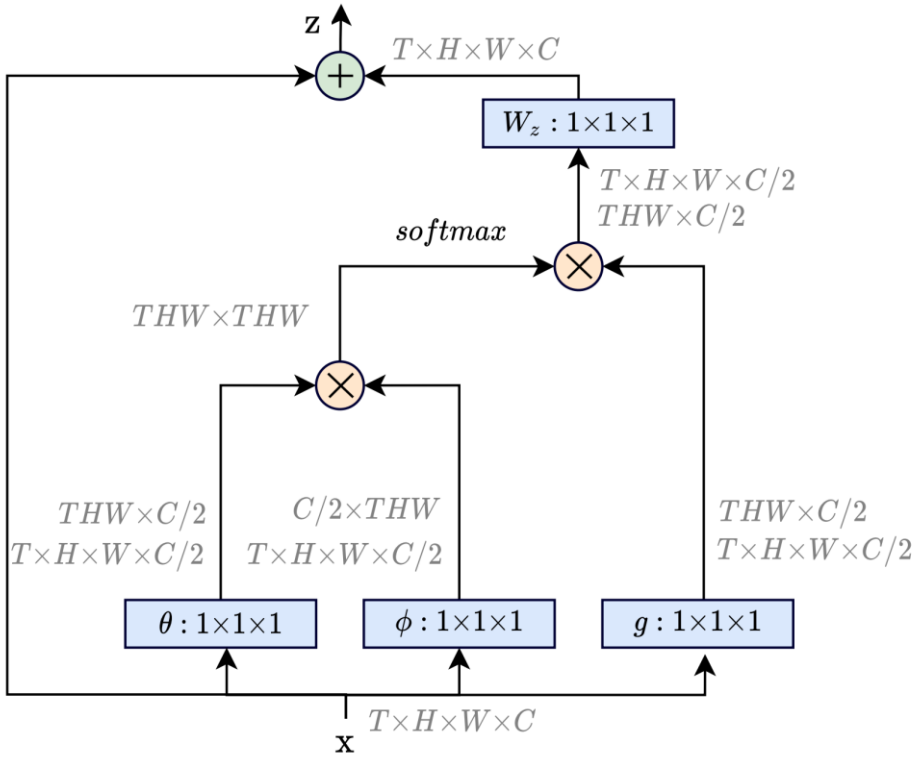
Simple Denoising with Avg



noisy

denoised with avg (9 x 9)

Non-local operation



$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} w_{i,j} \mathbf{W}_g \mathbf{x}_j,$$

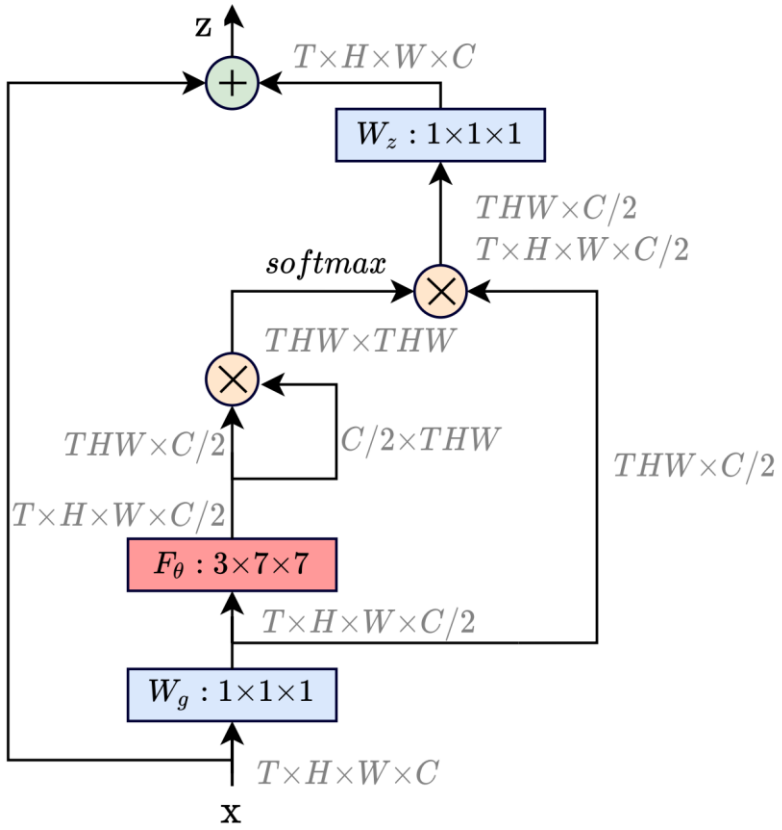
$$w_{i,j} = f(\mathbf{x}_i, \mathbf{x}_j),$$

replace

$$w_{i,j} = f(\theta(\mathcal{N}_i), \theta(\mathcal{N}_j))$$

- y_i – output features at position i
- x_i , x_j – input features at position j and j
- $w_{i,j}$ – weight computed by $f(,)$

Region-based Non-local operation



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

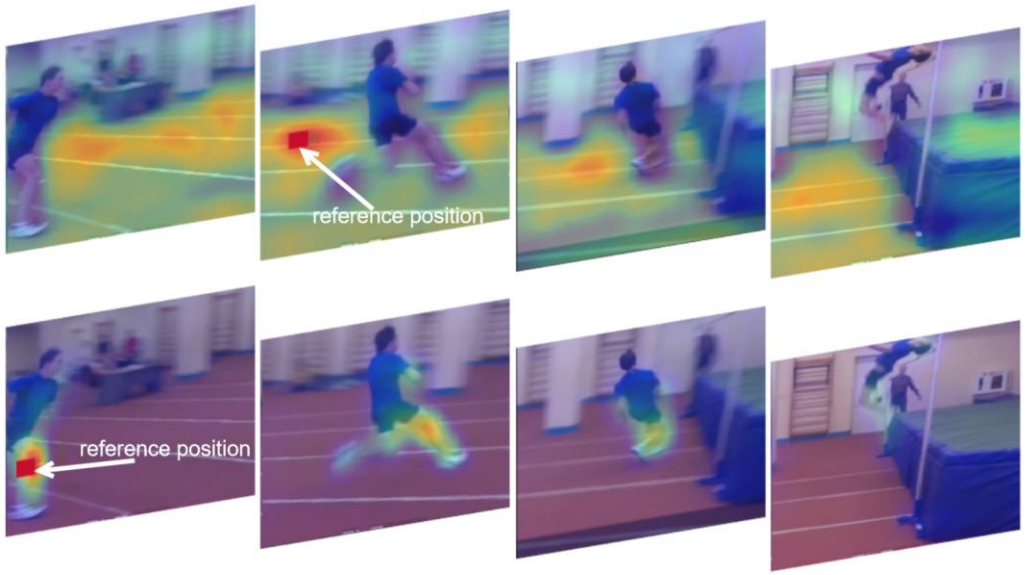
$$w_{i,j} = f(\theta(\mathcal{N}_i), \theta(\mathcal{N}_j))$$

$$\theta(\mathcal{N}_i) = \sum_{k\in\mathcal{N}_i} \mathbf{u}_k \odot \mathbf{x}_k$$

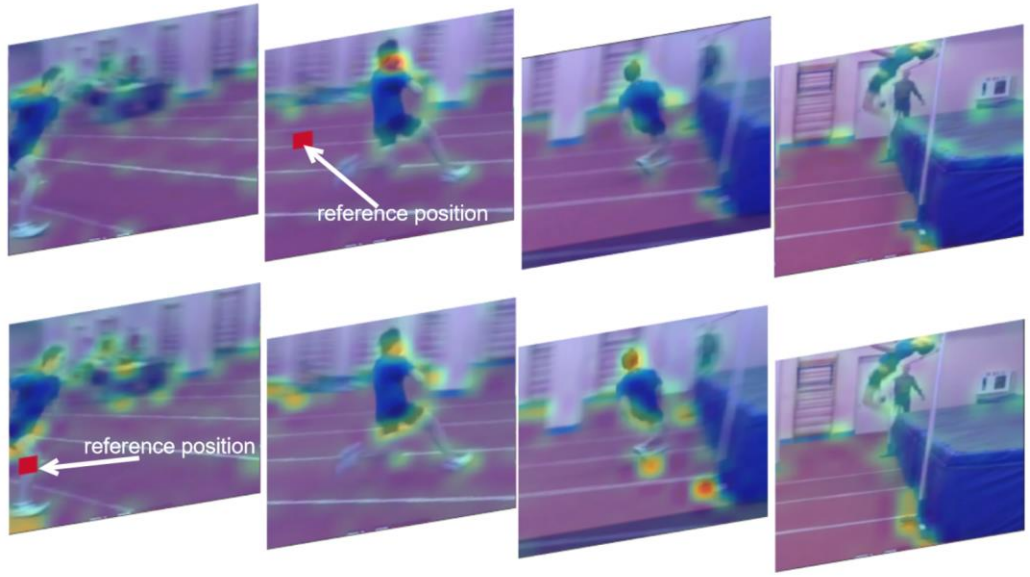
calculating the relationship $w_{i\ j}$ of position i and j by utilizing the information from the neighboring regions N_i and N_j.

- N_i – cuboid region of fixed sized centered at position i
- $\theta(,)$ – information aggregation function

RNL vs NL

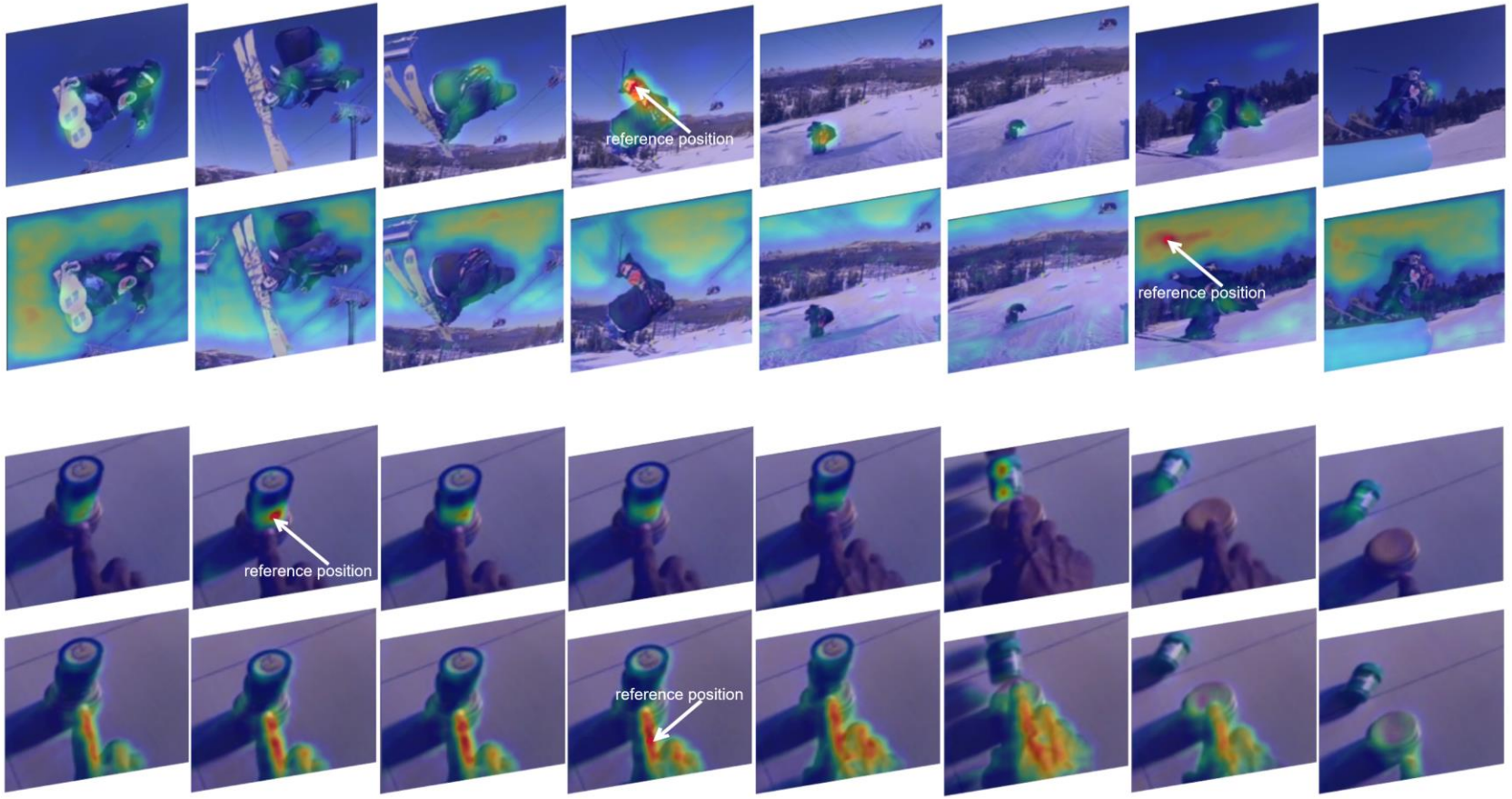


(a) RNL attention maps

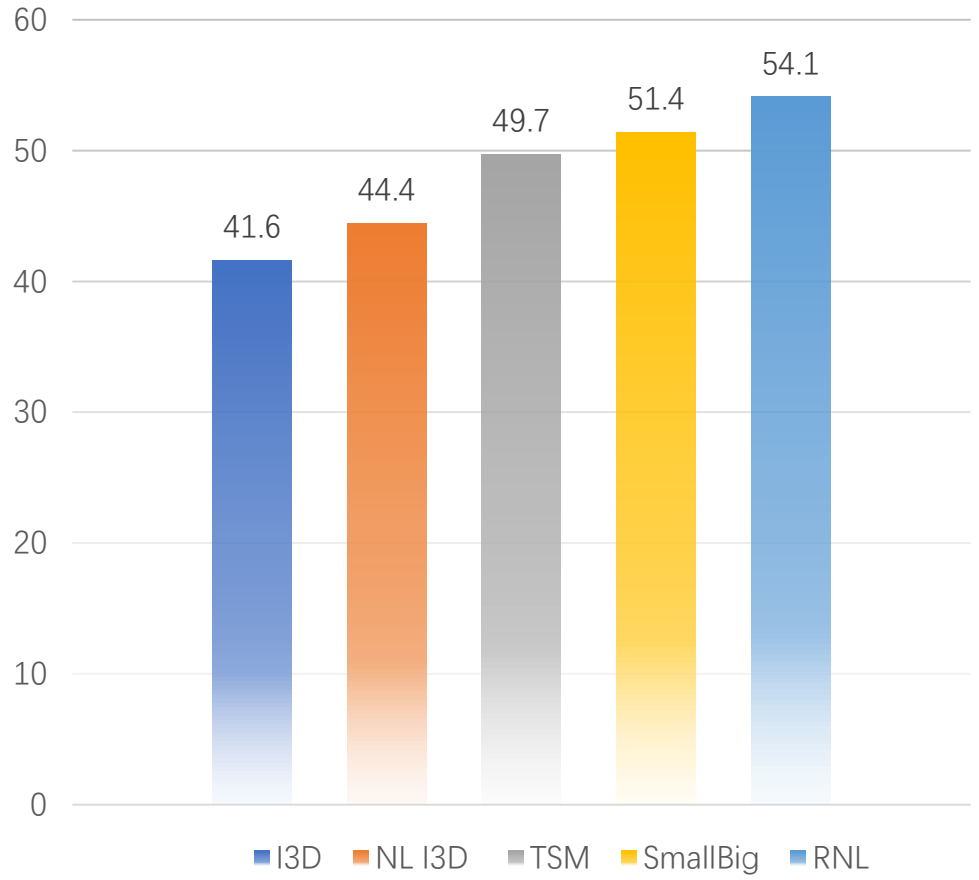


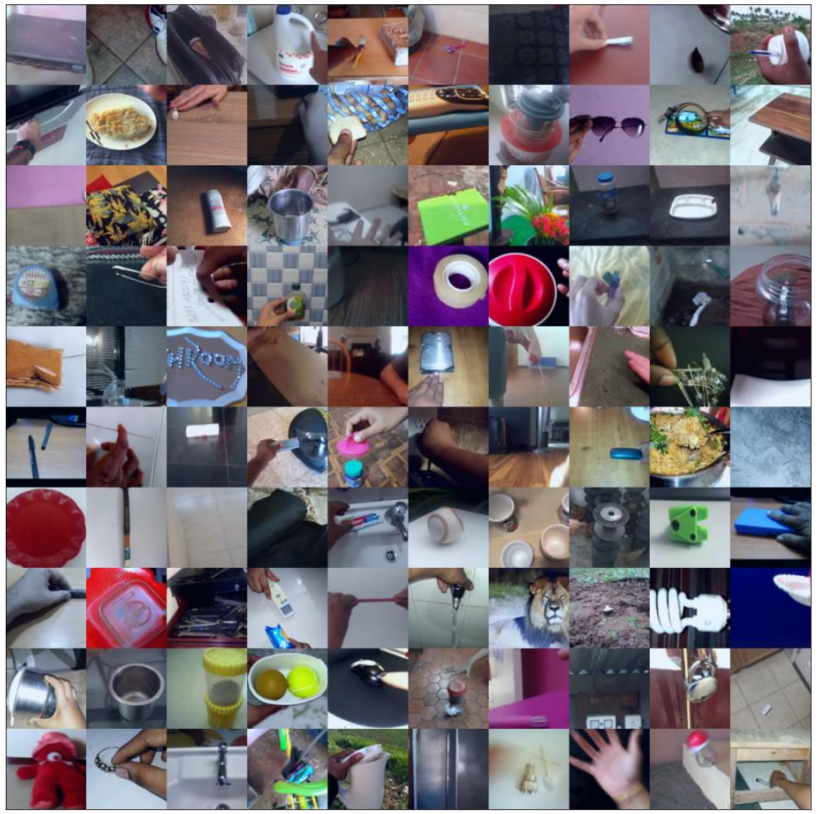
(b) NL attention maps

Attention Map of RNL



Comparison on Something-V1





Thank you