A. Proof of Proposition 1

As mentioned in the main paper, with the aim to restore the sharp distribution in linear attention, we present our Focused Function $f_p$:

$$\text{Sim} (Q_i, K_j) = \phi_p (Q_i) \phi_p (K_j)^T,$$  \hspace{1cm} (1)

where $\phi_p(x) = f_p(\text{ReLU}(x))$, $f_p(x) = \frac{\|x\| \times x^{**p}}{\|x^{**p}\|}$,  \hspace{1cm} (2)

and $x^{**p}$ represents the power $p$ of $x$ bit by bit. We follow previous linear attention modules to use the ReLU function first to ensure the non-negativity of input. Therefore, when proving the effects of $f_p$, we only consider $x, y \geq 0$.

**Proposition 1** (Feature direction adjustment with $f_p$) Let $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n, x_i, y_j \geq 0$. Assume $0 < \langle x, y \rangle < \|x\| \|y\|$ and $x, y$ have the single largest value $x_m, y_n$, respectively.

For a pair of feature $\{x, y\}$ with $m = n$:

$$\exists p > 1, \text{ s.t. } \langle \phi_p(x), \phi_p(y) \rangle > \langle x, y \rangle.$$  \hspace{1cm} (3)

For a pair of feature $\{x, y\}$ with $m \neq n$:

$$\exists p > 1, \text{ s.t. } \langle \phi_p(x), \phi_p(y) \rangle < \langle x, y \rangle.$$  \hspace{1cm} (4)

**Proof.**

$$\phi_p(x) = f_p(\text{ReLU}(x)) = f_p(x),$$

$$\phi_p(y) = f_p(\text{ReLU}(y)) = f_p(y).$$  \hspace{1cm} (5)

Therefore, we have:

$$\langle \phi_p(x), \phi_p(y) \rangle = \langle f_p(x), f_p(y) \rangle$$

$$= \|f_p(x)\| \|f_p(y)\| \langle u, v \rangle$$

$$= \|x\| \|y\| \langle u, v \rangle,$$  \hspace{1cm} (7)

where

$$\langle u, v \rangle = \frac{\left\langle f_p(x), f_p(y) \right\rangle}{\|f_p(x)\| \|f_p(y)\|} = \frac{\sum_{i=1}^{n} x_i^p y_i^p}{\sqrt{\left(\sum_{i=1}^{n} x_i^{2p}\right) \left(\sum_{i=1}^{n} b_i^{2p}\right)}}$$  \hspace{1cm} (8)

and $a_i = x_i / \max_{1 \leq i \leq n} (x_i), b_i = y_i / \max_{1 \leq i \leq n} (y_i), a_i, b_i \in [0, 1]$. (9)

Based on our assumption, we have:

$$\exists! n, \text{ s.t. } a_m = 1, \exists! n, \text{ s.t. } b_n = 1.$$  \hspace{1cm} (10)

Therefore,

$$\lim_{p \to \infty} a_i^p = \begin{cases} 1, & i = m \\ 0, & i \neq m \end{cases}, \lim_{p \to \infty} b_j^p = \begin{cases} 1, & j = n \\ 0, & j \neq n \end{cases}.$$  \hspace{1cm} (11)

Then we consider the following two cases:

(1) $m = n$:

$$\lim_{p \to \infty} \langle u, v \rangle = \lim_{p \to \infty} \frac{\sum_{i=1}^{n} a_i^p b_i^p}{\sqrt{\left(\sum_{i=1}^{n} a_i^{2p}\right) \left(\sum_{i=1}^{n} b_i^{2p}\right)}}$$

$$= \frac{1 \times 1}{\sqrt{1 \times 1}} = 1.$$  \hspace{1cm} (12)

Eq. (7), Eq. (12) $\Rightarrow$

$$\lim_{p \to \infty} \langle \phi_p(x), \phi_p(y) \rangle = \lim_{p \to \infty} \|x\| \|y\| \langle u, v \rangle$$

$$= \|x\| \|y\| \langle u, v \rangle > \langle x, y \rangle.$$  \hspace{1cm} (13)

Thus we have,

$$\exists p > 1, \text{ s.t. } \langle \phi_p(x), \phi_p(y) \rangle > \langle x, y \rangle.$$  \hspace{1cm} (14)
(2) $m \neq n$:

$$\lim_{p \to \infty} \langle u, v \rangle = \lim_{p \to \infty} \frac{\sum_{i=1}^{n} a_i^p b_i^p}{\sqrt{\sum_{i=1}^{n} a_i^{2p} \left( \sum_{i=1}^{n} b_i^{2p} \right)}}$$

$$= \frac{1 \times 0 + 0 \times 1}{\sqrt{1 \times 1}} = 0. \quad (15)$$

Eq. (7), Eq. (15) $\Rightarrow$

$$\lim_{p \to \infty} \langle \phi_p(x), \phi_p(y) \rangle = \lim_{p \to \infty} \|x\| \|y\| \langle u, v \rangle$$

$$= 0 < \langle x, y \rangle . \quad (16)$$

Thus we have,

$$\exists p > 1, \ s.t. \ \langle \phi_p(x), \phi_p(y) \rangle < \langle x, y \rangle . \quad (17)$$

Therefore, with a proper $p$, our focused function $f_p(\cdot)$ practically achieves a more distinguished difference between similar query-key pairs (Eq. (3)) and dissimilar query-key pairs (Eq. (4)). Actually, $f_p$ divides the features into several groups according to their nearest axes, improving the similarity within each group while reducing the similarity between the groups, thus restoring the sharp attention distribution as the original Softmax function.

**Table 1.** Comparisons of focused linear attention with other vision transformer backbones on the ImageNet-1K classification task.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reso</th>
<th>#Params</th>
<th>Flops</th>
<th>Top-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeiT-T [3]</td>
<td>224²</td>
<td>5.7M</td>
<td>1.2G</td>
<td>72.2</td>
</tr>
<tr>
<td>FLatten-DeiT-T</td>
<td>224²</td>
<td>6.1M</td>
<td>1.1G</td>
<td>74.1 (+1.9)</td>
</tr>
<tr>
<td>PVT-T [4]</td>
<td>224²</td>
<td>13.2M</td>
<td>1.9G</td>
<td>75.1</td>
</tr>
<tr>
<td>FLatten-PVT-T</td>
<td>224²</td>
<td>12.2M</td>
<td>2.0G</td>
<td>77.8 (+2.7)</td>
</tr>
<tr>
<td>PVT-S</td>
<td>224²</td>
<td>24.5M</td>
<td>3.8G</td>
<td>79.8</td>
</tr>
<tr>
<td>FLatten-PVT-S</td>
<td>224²</td>
<td>21.7M</td>
<td>4.0G</td>
<td>81.7 (+1.9)</td>
</tr>
<tr>
<td>PVT-M</td>
<td>224²</td>
<td>44.2M</td>
<td>6.7G</td>
<td>81.2</td>
</tr>
<tr>
<td>FLatten-PVT-M</td>
<td>224²</td>
<td>37.2M</td>
<td>7.0G</td>
<td>83.0 (+1.8)</td>
</tr>
<tr>
<td>PVT-L</td>
<td>224²</td>
<td>61.4M</td>
<td>9.8G</td>
<td>81.7</td>
</tr>
<tr>
<td>FLatten-PVT-L</td>
<td>224²</td>
<td>50.6M</td>
<td>10.4G</td>
<td>83.4 (+1.7)</td>
</tr>
<tr>
<td>PVTv2-B0 [5]</td>
<td>224²</td>
<td>3.4M</td>
<td>0.6G</td>
<td>70.5</td>
</tr>
<tr>
<td>FLatten-PVTv2-B0</td>
<td>224²</td>
<td>3.6M</td>
<td>0.6G</td>
<td>71.1 (+0.6)</td>
</tr>
<tr>
<td>PVTv2-B1</td>
<td>224²</td>
<td>13.1M</td>
<td>2.1G</td>
<td>78.7</td>
</tr>
<tr>
<td>FLatten-PVTv2-B1</td>
<td>224²</td>
<td>12.9M</td>
<td>2.2G</td>
<td>79.5 (+0.7)</td>
</tr>
<tr>
<td>PVTv2-B2</td>
<td>224²</td>
<td>25.4M</td>
<td>4.0G</td>
<td>82.0</td>
</tr>
<tr>
<td>FLatten-PVTv2-B2</td>
<td>224²</td>
<td>22.6M</td>
<td>4.3G</td>
<td>82.5 (+0.5)</td>
</tr>
<tr>
<td>PVTv2-B3</td>
<td>224²</td>
<td>45.2M</td>
<td>6.9G</td>
<td>83.2</td>
</tr>
<tr>
<td>FLatten-PVTv2-B3</td>
<td>224²</td>
<td>38.3M</td>
<td>7.3G</td>
<td>83.7 (+0.5)</td>
</tr>
<tr>
<td>PVTv2-B4</td>
<td>224²</td>
<td>62.6M</td>
<td>10.1G</td>
<td>83.6</td>
</tr>
<tr>
<td>FLatten-PVTv2-B4</td>
<td>224²</td>
<td>51.8M</td>
<td>10.7G</td>
<td>84.0 (+0.4)</td>
</tr>
<tr>
<td>Swin-T [2]</td>
<td>224²</td>
<td>29M</td>
<td>4.5G</td>
<td>81.3</td>
</tr>
<tr>
<td>FLatten-Swin-T</td>
<td>224²</td>
<td>29M</td>
<td>4.5G</td>
<td>82.1 (+0.8)</td>
</tr>
<tr>
<td>Swin-S</td>
<td>224²</td>
<td>50M</td>
<td>8.7G</td>
<td>83.0</td>
</tr>
<tr>
<td>FLatten-Swin-S</td>
<td>224²</td>
<td>51M</td>
<td>8.7G</td>
<td>83.5 (+0.5)</td>
</tr>
<tr>
<td>Swin-B</td>
<td>224²</td>
<td>88M</td>
<td>15.4G</td>
<td>83.5</td>
</tr>
<tr>
<td>FLatten-Swin-B</td>
<td>224²</td>
<td>89M</td>
<td>15.4G</td>
<td>83.8 (+0.3)</td>
</tr>
<tr>
<td>Swin-B</td>
<td>384²</td>
<td>88M</td>
<td>47.0G</td>
<td>84.5</td>
</tr>
<tr>
<td>FLatten-Swin-B</td>
<td>384²</td>
<td>91M</td>
<td>46.5G</td>
<td>85.0 (+0.5)</td>
</tr>
<tr>
<td>CSwin-T [1]</td>
<td>224²</td>
<td>23M</td>
<td>4.3G</td>
<td>82.7</td>
</tr>
<tr>
<td>FLatten-CSwin-T</td>
<td>224²</td>
<td>21M</td>
<td>4.3G</td>
<td>83.1 (+0.4)</td>
</tr>
<tr>
<td>CSwin-S</td>
<td>224²</td>
<td>35M</td>
<td>6.9G</td>
<td>83.6</td>
</tr>
<tr>
<td>FLatten-CSwin-S</td>
<td>224²</td>
<td>35M</td>
<td>6.9G</td>
<td>83.8 (+0.2)</td>
</tr>
<tr>
<td>CSwin-B</td>
<td>224²</td>
<td>78M</td>
<td>15.0G</td>
<td>84.2</td>
</tr>
<tr>
<td>FLatten-CSwin-B</td>
<td>224²</td>
<td>75M</td>
<td>15.0G</td>
<td>84.5 (+0.3)</td>
</tr>
<tr>
<td>CSwin-B</td>
<td>384²</td>
<td>78M</td>
<td>47.0G</td>
<td>85.4</td>
</tr>
<tr>
<td>FLatten-CSwin-B</td>
<td>384²</td>
<td>78M</td>
<td>46.4G</td>
<td>85.5 (+0.1)</td>
</tr>
</tbody>
</table>
B. More Visualizations

We visualize more examples of attention weights in Fig. 1. To better show the contribution of our focused function and DWC, we start from the vanilla linear attention and introduce $f_p$ and DWC separately. As demonstrated in the last three rows, DWC improves local focus ability but cannot focus on any position, while $f_p$ practically enhances model’s focus ability, helping model focus on more informative regions. Combining $f_p$ and DWC, our focused linear attention module restores the sharp distribution as the original Softmax attention.

C. Full Classification Results

Due to the page limit, we only present representative ImageNet classification results in Figure 6 of main paper. Here, we give all the classification results when applying our focused linear attention module on various sizes of the five baseline models in Tab.1.

D. Model Architectures

We summarize the architectures of five Transformer models adopted in the main paper, including DeiT [3], PVT [4], PVTv2 [5], Swin Transformer [2], CSwin Transformer [1] in Tab.2-8. In practice, we substitute the original self-attention blocks at all stages of the DeiT, PVT and PVTv2 with the focused linear attention block, but only adopt our module at early stages of Swin and CSwin. The model structure (width and depth) are kept unchanged, except for CSwin-T and CSwin-B, where we increase the depth of the first and second stages and correspondingly reduce the depth of the third stage to better reflect our module’s advantage of enlarged receptive field.

References


<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-DeiT-T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten</td>
</tr>
<tr>
<td>res1</td>
<td>14 × 14</td>
<td>win 14 × 14</td>
</tr>
</tbody>
</table>

Table 2. Architectures of FLatten-DeiT models.

<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-PVT-M</th>
<th>FLatten-PVT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten</td>
<td>PVT Block</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FLatten</td>
</tr>
<tr>
<td>res1</td>
<td>56 × 56</td>
<td>win 56 × 56</td>
<td>dim 64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 1</td>
<td></td>
</tr>
<tr>
<td>res2</td>
<td>28 × 28</td>
<td>win 28 × 28</td>
<td>dim 128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 2</td>
<td></td>
</tr>
<tr>
<td>res3</td>
<td>14 × 14</td>
<td>win 14 × 14</td>
<td>dim 320</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 5</td>
<td></td>
</tr>
<tr>
<td>res4</td>
<td>7 × 7</td>
<td>win 7 × 7</td>
<td>dim 512</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 8</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Architectures of FLatten-PVT models (Part1).

<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-PVT-M</th>
<th>FLatten-PVT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten</td>
<td>PVT Block</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FLatten</td>
</tr>
<tr>
<td>res1</td>
<td>56 × 56</td>
<td>win 56 × 56</td>
<td>dim 64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 1</td>
<td></td>
</tr>
<tr>
<td>res2</td>
<td>28 × 28</td>
<td>win 28 × 28</td>
<td>dim 128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 2</td>
<td></td>
</tr>
<tr>
<td>res3</td>
<td>14 × 14</td>
<td>win 14 × 14</td>
<td>dim 320</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 5</td>
<td></td>
</tr>
<tr>
<td>res4</td>
<td>7 × 7</td>
<td>win 7 × 7</td>
<td>dim 512</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 8</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Architectures of FLatten-PVT models (Part2).
<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-PVTv2-B0</th>
<th>FLatten-PVTv2-B1</th>
<th>FLatten-PVTv2-B2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten PVTv2 Block</td>
<td>FLatten PVTv2 Block</td>
<td>FLatten PVTv2 Block</td>
</tr>
<tr>
<td>res1</td>
<td>56 x 56</td>
<td>win 56 x 56 dim 32 head 1 x 2 None</td>
<td>win 56 x 56 dim 64 head 1 x 2 None</td>
<td>win 56 x 56 dim 64 head 1 x 3 None</td>
</tr>
<tr>
<td>res2</td>
<td>28 x 28</td>
<td>win 28 x 28 dim 64 head 2 x 2 None</td>
<td>win 28 x 28 dim 128 head 2 x 2 None</td>
<td>win 28 x 28 dim 128 head 2 x 3 None</td>
</tr>
<tr>
<td>res3</td>
<td>14 x 14</td>
<td>win 14 x 14 dim 160 head 5 x 2 None</td>
<td>win 14 x 14 dim 320 head 5 x 2 None</td>
<td>win 14 x 14 dim 320 head 5 x 6 None</td>
</tr>
<tr>
<td>res4</td>
<td>7 x 7</td>
<td>win 7 x 7 dim 512 head 8 x 2 None</td>
<td>win 7 x 7 dim 512 head 8 x 2 None</td>
<td>win 7 x 7 dim 512 head 8 x 3 None</td>
</tr>
</tbody>
</table>

Table 5. Architectures of FLatten-PVTv2 models (Part1).

<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-PVTv2-B3</th>
<th>FLatten-PVTv2-B4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten PVTv2 Block</td>
<td>FLatten PVTv2 Block</td>
</tr>
<tr>
<td>res1</td>
<td>56 x 56</td>
<td>win 56 x 56 dim 64 head 1 x 3 None</td>
<td>win 56 x 56 dim 64 head 1 x 3 None</td>
</tr>
<tr>
<td>res2</td>
<td>28 x 28</td>
<td>win 28 x 28 dim 128 head 2 x 3 None</td>
<td>win 28 x 28 dim 128 head 2 x 8 None</td>
</tr>
<tr>
<td>res3</td>
<td>14 x 14</td>
<td>win 14 x 14 dim 320 head 5 x 18 None</td>
<td>win 14 x 14 dim 320 head 5 x 27 None</td>
</tr>
<tr>
<td>res4</td>
<td>7 x 7</td>
<td>win 7 x 7 dim 512 head 8 x 3 None</td>
<td>win 7 x 7 dim 512 head 8 x 3 None</td>
</tr>
</tbody>
</table>

Table 6. Architectures of FLatten-PVTv2 models (Part2).
<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-Swin-T</th>
<th>FLatten-Swin-S</th>
<th>FLatten-Swin-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten Swin Block</td>
<td>FLatten Swin Block</td>
<td>FLatten Swin Block</td>
</tr>
<tr>
<td>res1</td>
<td>56 × 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 56×56, 96, LN</td>
<td>win 56×56, 96, LN</td>
<td>win 56×56, 128, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 96 × 2</td>
<td>dim 96 × 2</td>
<td>dim 128 × 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 3 None</td>
<td>head 3 None</td>
<td>head 3 None</td>
<td></td>
</tr>
<tr>
<td>res2</td>
<td>28 × 28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 28×28, 96, LN</td>
<td>win 28×28, 96, LN</td>
<td>win 28×28, 256, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 192 × 2</td>
<td>dim 192 × 2</td>
<td>dim 256 × 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 6 None</td>
<td>head 6 None</td>
<td>head 6 None</td>
<td></td>
</tr>
<tr>
<td>res3</td>
<td>14 × 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>win 7×7, 192, LN</td>
<td>win 7×7, 192, LN</td>
<td>win 7×7, 512, LN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dim 384 × 6</td>
<td>dim 384 × 18</td>
<td>dim 512 × 18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 12 None</td>
<td>head 12 None</td>
<td>head 12 None</td>
</tr>
<tr>
<td>res4</td>
<td>7 × 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>win 7×7, 384, LN</td>
<td>win 7×7, 384, LN</td>
<td>win 7×7, 1024, LN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dim 768 × 2</td>
<td>dim 768 × 2</td>
<td>dim 1024 × 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>head 24 None</td>
<td>head 24 None</td>
<td>head 24 None</td>
</tr>
</tbody>
</table>

Table 7. Architectures of FLatten-Swin models.

<table>
<thead>
<tr>
<th>stage</th>
<th>output</th>
<th>FLatten-CSwin-T</th>
<th>FLatten-CSwin-S</th>
<th>FLatten-CSwin-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FLatten CSwin Block</td>
<td>FLatten CSwin Block</td>
<td>FLatten CSwin Block</td>
</tr>
<tr>
<td>res1</td>
<td>56 × 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conv7×7, stride=4, 64, LN</td>
<td>Conv7×7, stride=4, 96, LN</td>
<td>Conv7×7, stride=4, 96, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 3×3, 64 × 2</td>
<td>win 3×3, 96 × 2</td>
<td>win 3×3, 128 × 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 64</td>
<td>dim 96</td>
<td>dim 128</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 2</td>
<td>head 3</td>
<td>head 4</td>
<td></td>
</tr>
<tr>
<td>res2</td>
<td>28 × 28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conv7×7, stride=4, 128, LN</td>
<td>Conv7×7, stride=4, 192, LN</td>
<td>Conv7×7, stride=4, 192, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 3×3, 128 × 4</td>
<td>win 3×3, 192 × 4</td>
<td>win 3×3, 256 × 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 128</td>
<td>dim 192</td>
<td>dim 256</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 4</td>
<td>head 6</td>
<td>head 8</td>
<td></td>
</tr>
<tr>
<td>res3</td>
<td>14 × 14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conv7×7, stride=4, 256, LN</td>
<td>Conv7×7, stride=4, 384, LN</td>
<td>Conv7×7, stride=4, 384, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 3×3, 256 × 18</td>
<td>win 3×3, 384 × 32</td>
<td>win 3×3, 512 × 29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 256</td>
<td>dim 384</td>
<td>dim 512</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 8</td>
<td>head 16</td>
<td>head 24</td>
<td></td>
</tr>
<tr>
<td>res4</td>
<td>7 × 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conv7×7, stride=4, 512, LN</td>
<td>Conv7×7, stride=4, 768, LN</td>
<td>Conv7×7, stride=4, 768, LN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>win 7×7, 512 × 1</td>
<td>win 7×7, 768 × 2</td>
<td>win 7×7, 768 × 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dim 512</td>
<td>dim 768</td>
<td>dim 768</td>
<td></td>
</tr>
<tr>
<td></td>
<td>head 16</td>
<td>head 32</td>
<td>head 32</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Architectures of FLatten-CSwin models.