### Abstract
Transcribing content from structural images is a challenging task as not only the content objects should be recognized, but the internal structure should also be preserved. In our work, we propose a hierarchical Spotlight Transcribing Network (STN) framework followed by a two-stage “where-to-what” solution. We first decide “where-to-look” through a novel spotlight mechanism to focus on different areas of the original image following its structure. Then, we decide “what-to-write” by developing a GRU based network with the spotlight areas for transcribing the content accordingly.

### Problem Definition
**Definition 3.1. (Structural Image Transcription Problem).** Given a structural $W \times H$ image $x$, our goal is to transcribe the content from it as a sequence $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}$ as close as possible to the source code sequence $y$, where each $\hat{y}_i$ is the predicted token taking from the specific language corresponding to the image.

### Spotlight Mechanism
Given spotlight handle $e_{st} = (x_t, y_t, \sigma_t)^T$, assign weights to encoded vectors following Gaussian distribution.

### Spotlight Control
We provide two control modules:
- **Markovian** control module (as in STNM with Markov property)
- **Recurrent** control module (as in STNR with Recurrent modeling)

### Experimental Results
**Transcribing performance**
Outperforms traditional attention based methods.

**Validation loss**
Converges faster and achieves lower validation loss.

**Spotlight visualization**
- STNR finds a more reasonable reading path;
- STNR clearly distinguishes similar regions properly.

### Preliminaries
**Structural images**: printed graphics that organized in a complex structure.

**Characteristics**:
- Much semantics
- Larger output space
- Reversible

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image count</th>
<th>Token space</th>
<th>Taken count</th>
<th>Avg. tokens per image</th>
<th>Avg. image pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melody</td>
<td>4208</td>
<td>70</td>
<td>82,834</td>
<td>19.7</td>
<td>15,662.7</td>
</tr>
<tr>
<td>Formula</td>
<td>6164</td>
<td>127</td>
<td>607,661</td>
<td>9.7</td>
<td>1196.7</td>
</tr>
<tr>
<td>Multi-Line</td>
<td>4555</td>
<td>127</td>
<td>182,152</td>
<td>39.8</td>
<td>9,016.6</td>
</tr>
<tr>
<td>SVT</td>
<td>618</td>
<td>28</td>
<td>3,796</td>
<td>5.9</td>
<td>12,733.5</td>
</tr>
<tr>
<td>BITSK</td>
<td>3000</td>
<td>36</td>
<td>15,269</td>
<td>5.0</td>
<td>11,652.0</td>
</tr>
</tbody>
</table>

**Figure 7**: Comparison between attention and spotlight mechanism on Melody dataset.

**Figure 8**: Comparison between attention and spotlight mechanism on Formula dataset.