**Better to Follow, Follow to Be Better:**
Towards Precise Supervision of Feature Super-Resolution for Small Object Detection

Junhyug Noh Wonho Bae Wonhee Lee Jinhwon Seo Gunhee Kim

**Project Page:** [http://vision.snu.ac.kr/projects/better-to-follow](http://vision.snu.ac.kr/projects/better-to-follow)

---

### Summary

**Problem:** poor performance of a proposal-based detector using feature-level super-resolution on small objects

**Cause:** absence of direct supervision from target features

**Solution:** novel approach to "properly" extract target features as direct supervision

---

### Difficulty of Detecting Small objects

1. RoI pooled features do not contain detailed information due to its size.
2. In the process of RoI pooling, internal positions are distorted.

Then, Super-resolve features as large objects!

---

### Methods to Generate Super-Resolution

**Step 1.** Generate super-resolution features ($F_{SR}$) from an original image

1. To be similar to high-resolution features of a large object (Discriminator)
2. So that a class and box offsets of the small object are correctly predicted (Predictor)

> But without supervision, training of the super-resolution feature generator can be unstable

**Step 2.** Generate super-resolution features ($F_{SR}$) from a downsampled image

1. To be similar to the corresponding naïve targets ($F_{N}$)

**Step 3.** To be similar to the corresponding naïve targets ($F_{N}$)

Even with naïve supervision, it is hard to imitate target features due to high disparity between input ($F_{I}$) and target features ($F_{T}$)

---

### Quantitative Results

**Tsinghua-Tencent 100K**

1. Results on different backbones (input: 1600×1600)

<table>
<thead>
<tr>
<th>Model</th>
<th>AF</th>
<th>A50</th>
<th>A50-1.5</th>
<th>A50-2</th>
<th>A50-3</th>
<th>A50-3.5</th>
<th>A50-4</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-RCN</td>
<td>59.1</td>
<td>79.9</td>
<td>91.4</td>
<td>86.2</td>
<td>88.5</td>
<td>80.8</td>
<td>81.1</td>
<td>80.8</td>
</tr>
<tr>
<td>R-FCN</td>
<td>76.7</td>
<td>91.9</td>
<td>94.7</td>
<td>92.6</td>
<td>93.5</td>
<td>90.2</td>
<td>91.3</td>
<td>91.5</td>
</tr>
<tr>
<td>DSC (0)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
<td>80.8</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (1)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
<td>80.8</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (2)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
<td>80.8</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (3)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
<td>80.8</td>
<td>80.5</td>
</tr>
</tbody>
</table>

2. Comparison with SOTA models (input: 2048×2048)

<table>
<thead>
<tr>
<th>Model</th>
<th>AP (50)</th>
<th>AP (50–90)</th>
<th>AP (60–90)</th>
<th>AP (70–90)</th>
<th>AP (80–90)</th>
<th>AP (90–90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC (0)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (1)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (2)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>DSC (3)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
</tbody>
</table>

3. Comparison of super-resolution methods

<table>
<thead>
<tr>
<th>Base model</th>
<th>AP (50)</th>
<th>AP (50–90)</th>
<th>AP (60–90)</th>
<th>AP (70–90)</th>
<th>AP (80–90)</th>
<th>AP (90–90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base model</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>SR (Naive supervision)</td>
<td>67.0</td>
<td>92.6</td>
<td>95.3</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>SR (Naive supervision)</td>
<td>71.1</td>
<td>87.7</td>
<td>90.5</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
<tr>
<td>SR (Base)</td>
<td>82.5</td>
<td>95.7</td>
<td>98.7</td>
<td>83.7</td>
<td>82.6</td>
<td>80.5</td>
</tr>
</tbody>
</table>

---

### Our Approach

**Super-Resolution Feature Extractor**

Replace every layer that increases relative field of feature extractor with a layer that doubling it ($k$: kernel size, $s$: stride, $r$: dilatation rate)

- $k×k$ POOL ($k > 1$) $→$ $2k×2k$ POOL
- $k×k$ CONV ($k > 1$) $→$ $k×k$ Atrous CONV ($r = z - 1$)
- $k×k$ CONV ($k > 1$, $z = 2$) $→$ $k×k$ Atrous CONV ($r = z - 1 = 2$) + $2×2$ POOL ($z = 2$)

---

### PASCAL VOC & MS COCO

**Qualitative Results**

- **Visualization of Features**
- **Comparison of features from different extractors**
- **Comparison of different super-resolution methods**
- **Detection results on Tsinghua-Tencent 100K (C: TP; R: PP; B: FN)**

---

### References