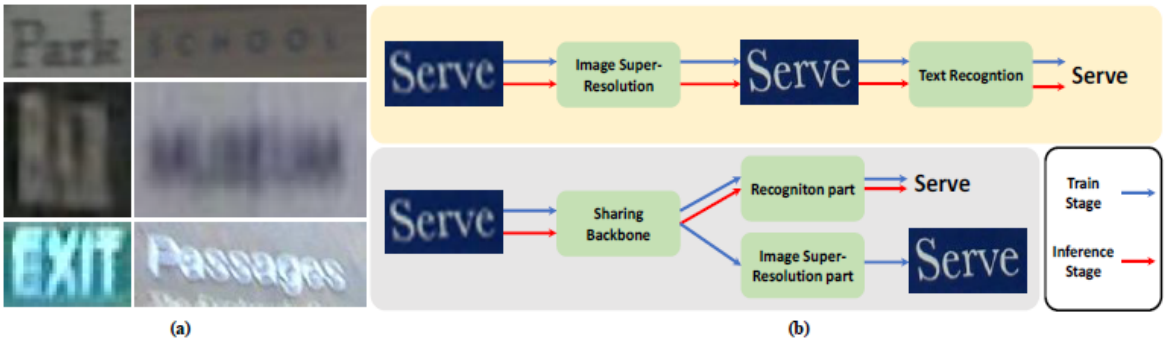
# PlugNet: Degradation Aware Scene Text Recognition Supervised by a Pluggable Super-Resolution Unit

## 论文理论

**研究问题：**低分辨率，高模糊，抖动等原因产生的低质量图片，影响文本识别的效果

**研究思路：**1）在图片层用强校正的方式，通过引入超分辨模块（SR）作为预处理，如19年出的TexSR和ESRGAN-Aster，但增加了大量的计算；2）在特征层用弱监督的方式，将SR和文本识别组成多任务学习，用SR分支辅助"共享的CNN-backbone"对低质量图片的特征表示。




### 模型设计

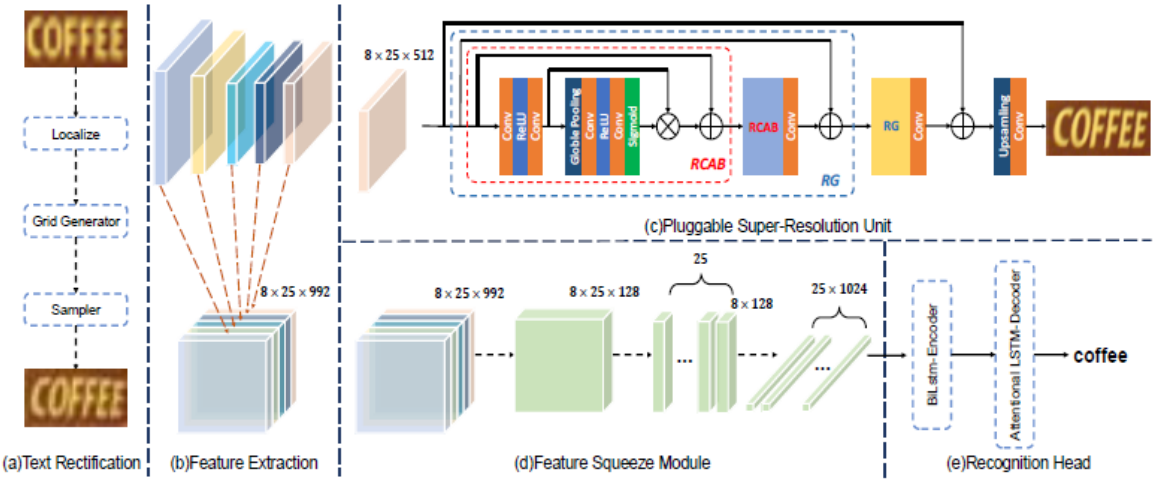
PlugNet包括校正模块，**CNN-backbone**，超分辨分支，识别分支四个部分。校正模块采用TPS方法（参考Aster），超分辨分支采用SR网络（查考RCAN），识别分支采用Seq2Seq

- **CNN-backbone**的设计

  1）采用大分辨率的特征图。具体做法是去除后三个res-block的下采样层，最终输出的特征图是原图的1/4。

  2）FEM特征组合模块。将CNN的多层特征图按通道组合在一起

  3）FSM特征压缩模块。通过1x1卷积降维，再沿高和通道方向"铺平"，构造1D向量适配LSTM的输入要求




**训练学习：**

PlugNet只在训练时加入超分辨单元，推理时不会增加计算量。训练时要求输入高低分辨率的图片样本对，采用的方法是通过对高分辨率图片**加模糊，噪声和下采样**来模拟低分辨率图片（如何得到真实样本是SR的重要问题）

- Loss损失=识别损失 + SR损失。识别损失用交叉熵损失，SR损失用像素点的L1损失，再通过权重系数调整。

$$L = L_{rec} + \lambda L_{sr}$$

由于SR分支只是辅助CNN-backbone更好的提取特征，导致损失的权重系数对结果比较敏感。

**核心认知：**

- 设计了PSU单元，构建SR分支进行多任务学习，辅助CNN-Backbone更好的提取低质量的图片特征
- 采用大分辨率的特征图，增加空间信息，改善识别效果

## 论文的实验效果

- 验证特征分辨率：大的特征分辨率在各个数据集上效果都有提升

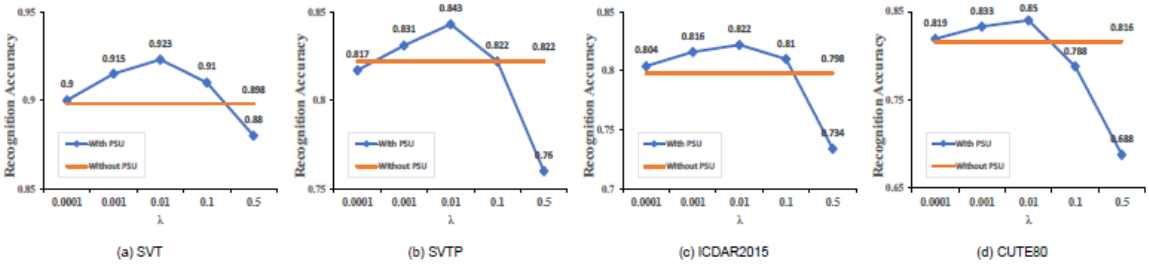
| Resolution | Data | SVT | SVTP | IIIT5K | IC03 | IC13 | IC15 | CUTE80 |
|---|---|---|---|---|---|---|---|---|
| 1×25 | *90K* | 85.2 | 76.1 | 80.7 | 91.8 | 89.3 | 69.3 | 66.3 |
| 2×25 | *90K* | $87.0_{\uparrow 1.8}$ | $78.1_{\uparrow 2.0}$ | $82.7_{\uparrow 2.0}$ | $92.3_{\uparrow 1.5}$ | $89.4_{\uparrow 0.1}$ | $69.3_{\uparrow 0}$ | $68.4_{\uparrow 2.1}$ |
| 4×25 | *90K* | $87.9_{\uparrow 0.9}$ | $79.5_{\uparrow 1.4}$ | $82.2_{\downarrow 0.5}$ | $92.7_{\uparrow 0.4}$ | $89.8_{\uparrow 0.4}$ | $71.4_{\uparrow 2.1}$ | $69.1_{\uparrow 0.7}$ |
| 8×25 | *90K* | $\mathbf{89.0}_{\uparrow 2.1}$ | $\mathbf{82.0}_{\uparrow 2.5}$ | $\mathbf{85.3}_{\uparrow 3.1}$ | $\mathbf{94.3}_{\uparrow 1.6}$ | $\mathbf{91.0}_{\uparrow 1.2}$ | $\mathbf{73.6}_{\uparrow 2.2}$ | $\mathbf{69.1}_{\uparrow 0}$ |

- 模块消融实验:PSU单元平均提升2.5个点，FEM单元平均提升1.1个点

| Methods | FSM | FEM | Data_Aug | ESRGAN | PSU | SVT | SVTP | IC15 | CUTE80 |
|---|---|---|---|---|---|---|---|---|---|
| *Baseline(R)* [33] | ✗ | ✗ | ✗ | ✗ | ✗ | 89.5 | 78.5 | 76.1 | 79.5 |
| *PlugNet(R)* | ✓ | ✗ | ✗ | ✗ | ✗ | $90.0_{\uparrow 0.5}$ | $80.8_{\uparrow 2.3}$ | $78.2_{\uparrow 2.1}$ | $82.6_{\uparrow 3.1}$ |
| *PlugNet(R)* | ✓ | ✓ | ✗ | ✗ | ✗ | $90.6_{\uparrow 0.6}$ | $81.6_{\uparrow 0.8}$ | $80.2_{\uparrow 2.0}$ | $83.7_{\uparrow 1.1}$ |
| *PlugNet* | ✓ | ✓ | ✓ | ✗ | ✗ | $89.8_{\downarrow 0.8}$ | $82.2_{\uparrow 0.6}$ | $79.8_{\downarrow 0.4}$ | $81.6_{\downarrow 2.1}$ |
| *SR-PlugNet* | ✓ | ✓ | ✓ | ✓ | ✗ | $90.6_{\uparrow 0.8}$ | $80.8_{\downarrow 1.4}$ | $79.4_{\downarrow 0.4}$ | $82.6_{\uparrow 1.0}$ |
| *PlugNet* | ✓ | ✓ | ✓ | ✗ | ✓ | $\mathbf{92.3}_{\uparrow 1.7}$ | $\mathbf{84.3}_{\uparrow 3.5}$ | $\mathbf{82.2}_{\uparrow 2.8}$ | $\mathbf{85.0}_{\uparrow 2.4}$ |

| | | | | |
|---|---|---|---|---|
| Groud Truth | school | arts | for | the |
| Aster | scrool | ar_ | row | till |
| PlugNet | school | arts | for | the |

- 对损失权重系数的实验：0.01效果最好




(a) SVT (b) SVTP (c) ICDAR2015 (d) CUTE80

权重系数的递增，会减少噪声和模糊有利，但也会让CNN更聚焦SR任务而不利于识别。


- 整体模型效果的对比：相比baseline在各类数据集上平均提升4个点左右

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Aster(Baseline)[33] | *90K, ST* | 93.4 | 89.5 | 94.5 | 91.8 | 76.1 | 78.5 | 79.5 |
| TextSR(SR-Aster)[39] | *90K, ST* | 92.5 | 87.2 | 93.2 | 91.3 | 75.6 | 77.4 | 78.9 |
| **Ours** | *90K, ST* | **94.4** | **92.3** | **95.7** | 95.0 | **82.2** | **84.3** | 85.0 |