Recurrent Scene Parsing with Perspective Understanding In the Loop

Shu Kong
CS, ICS, UCI
1. Background
2. Attention to Perspective: Depth-aware Gating
3. Recurrent Refining
4. Attentional Mechanism
5. Conclusion and Future Work
1. Background
Semantic Segmentation with Deep Convolutional Neural Networks

Keywords: skip connection, multi-scale, upsampling
DeepLab is a strong baseline (based on ResNet architecture), yet simple and straightforward.

It sums up feature maps at different scales using atrous convolution, i.e. convolution with various dilate rates.

1. a trous (French) -- holes (English)

2. Atrous convolution (skipping/inserting zero)

\[ y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k] \]

(a) Sparse feature extraction

fusing responses with multiple atrous kernels of different rates.
Background

That's all about the baseline.

The fusion of multi-scale feature maps exhibits some degree of scale invariance; but it's not obvious this invariance covers the range scale variation existing in perspective images.
Large Perspective Image

large range scale variation in perspective images.

car

pole

white/black board

charis
1. Background

2. Attention to Perspective: Depth-aware Gating
disparity, or depth, conveys the scale information.

pooling region size modulated by scene depth
image with example pooling regions
ground-truth

prediction w/o depth
prediction w/ depth
Depth-aware pooling module

select the right scale with depth
quantize the disparity into five scales with dilate rates \{1, 2, 4, 8, 16\}
Alternatively, learning depth estimator, and testing without depth

\[ \ell_{\text{depth Reg}}(\mathbf{D}, \mathbf{D}^*) = \frac{1}{|M|} \sum_{(i,j) \in M} \| \log(D_{ij}) - \log(D_{ij}^*) \|^2 \]

(a) depth-aware gating module using ground-truth depth map

(b) depth-aware gating module using predicted depth map
Alternatively, learning depth estimator, and testing without depth reliable monocular depth estimation

Table 1: Depth prediction on NYU-depth-v2 dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>0.542</td>
<td>0.614</td>
<td>0.614</td>
<td>0.769</td>
<td>0.811</td>
<td>0.809</td>
<td>0.816</td>
</tr>
<tr>
<td>1.25$^2$</td>
<td>0.829</td>
<td>0.883</td>
<td>0.888</td>
<td>0.950</td>
<td>0.953</td>
<td>0.945</td>
<td>0.950</td>
</tr>
<tr>
<td>1.25$^3$</td>
<td>0.940</td>
<td>0.971</td>
<td>0.972</td>
<td>0.988</td>
<td>0.988</td>
<td>0.986</td>
<td>0.989</td>
</tr>
</tbody>
</table>

Figure 3: Examples of monocular depth predictions. First row: the input RGB image; second row: ground-truth; third row: our result. In our visualizations, all depth maps use the same fixed (absolute) colormap to represent metric depth.
more configurations to compare --

1. sharing the parameters in this pooling module (multiPool)
more configurations to compare --

1. sharing the parameters in this pooling module (multiPool)
2. averaging the feature vs. depth-aware gating
more configurations to compare --

1. sharing the parameters in this pooling module (multiPool)
2. averaging the feature vs. depth-aware gating
3. MultiPool vs. MultiScale (input)
more configurations to compare --

1. sharing the parameters in this pooling module (multiPool)
2. averaging the feature vs. depth-aware gating
3. MultiPool vs. MultiScale (input)
Qualitative Results -- street images
Qualitative Results -- panorama images

Depth-aware pooling module
1. Background

2. Attention to Perspective: Depth-aware Gating

3. Recurrent Refining
Recurrent Refinement Module

Recurrently refining the results by adapting the predicted depth

![Diagram of Recurrent Refinement Module]

- **Input Image**
- **CNN Backbone**
- **Feed-forward Pathway**
- **Depth-aware Gating Module**
- **Recurrent Module**

- **Loop-0, IoU=0.418**
- **Loop-1, IoU=0.427**
- **Loop-2, IoU=0.431**

- Output Difference
unrolling the recurrent module during training
adding a loss to each unrolled loop
embedding the depth-aware gating module in the loops
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IoU</td>
<td>pixel acc.</td>
<td>IoU</td>
<td>pixel acc.</td>
</tr>
<tr>
<td>baseline</td>
<td>0.406</td>
<td>0.703</td>
<td>0.402</td>
<td>0.776</td>
</tr>
<tr>
<td>w/ gt-depth</td>
<td>0.413</td>
<td>0.708</td>
<td>0.422</td>
<td>0.787</td>
</tr>
<tr>
<td>w/ pred-depth</td>
<td>0.418</td>
<td>0.711</td>
<td>0.423</td>
<td>0.789</td>
</tr>
<tr>
<td>loop1 w/o depth</td>
<td>0.419</td>
<td>0.706</td>
<td>0.432</td>
<td>0.793</td>
</tr>
<tr>
<td>loop1 w/ gt-depth</td>
<td>0.425</td>
<td>0.711</td>
<td>0.439</td>
<td>0.798</td>
</tr>
<tr>
<td>loop1 w/ pred-depth</td>
<td>0.427</td>
<td>0.712</td>
<td>0.440</td>
<td>0.798</td>
</tr>
<tr>
<td>loop2</td>
<td>0.431</td>
<td>0.713</td>
<td>0.443</td>
<td>0.799</td>
</tr>
<tr>
<td>loop2 (test-aug)</td>
<td>0.445</td>
<td>0.721</td>
<td>0.451</td>
<td>0.803</td>
</tr>
<tr>
<td>DeepLab [6]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LRR [13]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Context [28]</td>
<td>0.406</td>
<td>0.700</td>
<td>0.423</td>
<td>0.784</td>
</tr>
<tr>
<td>PSPNet [38]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefineNet-Res50 [27]</td>
<td>0.438</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefineNet-Res101 [27]</td>
<td>0.447</td>
<td>-</td>
<td>0.457</td>
<td>0.804</td>
</tr>
<tr>
<td>RefineNet-Res152 [27]</td>
<td>0.465</td>
<td>0.736</td>
<td>0.459</td>
<td>0.806</td>
</tr>
</tbody>
</table>

Recurrently refining the results by adapting the predicted depth
Recurrent Refinement Module

Qualitative Results -- NYU-depth-v2 indoor dataset
Recurrent Refinement Module

Qualitative Results -- Cityscapes

yellow --> closer --> larger pooling size
Qualitative Results -- Stanford-2D-3D (panoramas)

blue --> closer --> larger pooling size
Outline

1. Background
2. Attention to Perspective: Depth-aware Gating
3. Recurrent Refining
4. Attentional Mechanism
Some slides from this point are removed due to research conflicts. They will be disclosed in the future.
Attention to Scale Again

Cityscapes

Stanford-2D-3D

baseline 0.738
MultiPool
- tied weights
- untied weights
  - average 0.747
  - depth-gating 0.748
  - average attention 0.751
  - depth-gating 0.754
  - gt-depth 0.753
  - pred-depth 0.759
Outline

1. Background
2. Attention to Perspective: Depth-aware Gating
3. Recurrent Refining
4. Attentional Mechanism
5. Conclusion and Future Work
1. Attentional module is powerful.
1. Attentional module is powerful.

2. Such attentional module should be also useful in various pixel-level tasks, e.g. pixel embedding for instance grouping, depth estimation, surface normal estimation, etc.
Thanks