Monocular Image-based 3D Detection and Reconstruction of Objects

Prof. Xiaoming Liu

Computer Vision Lab
Department of Computer Science and Engineering
Michigan State University
3D Perception

Image credit: Graham Murdoch/Popular Science
3D Perception
3D Perception
3D Detection and Reconstruction
3D Detection and Reconstruction
3D Detection and Reconstruction
Part 1

Monocular 3D Object Detection

Monocular 3D Region Proposal Network

Monocular 3D RPN

3D Object Detection
Prior Work: Monocular 3D Detection

Prior works:

- use external networks
- require multi-stage approach

Deep3DBox [1]

- IM
- CNN
- R-CNN
- Post Optim
- 3D Detection

Multi-Fusion [2]

- IM
- 2D RPN
- Depth
- Point Cloud
- R-CNN
- 3D Detection

M3D-RPN

- IM
- 2D-3D RPN
- Post Optim
- 3D Detection

Reference:

2D / 3D Anchor Definition

\[
\begin{bmatrix}
    x \cdot z \\
    y \cdot z \\
    z
\end{bmatrix}_P = P \cdot \begin{bmatrix}
    x \\
    y \\
    z
\end{bmatrix}_{3D}
\]

2D dimensions
3D dimensions, \( \theta \)
3D\( \rightarrow \)2D localization

observation angle \( \theta_{3D} \)
## 2D / 3D Anchor Generation

<table>
<thead>
<tr>
<th>$h_{2D}, w_{2D}$</th>
<th>$z_P$</th>
<th>$w_{3D}$</th>
<th>$h_{3D}$</th>
<th>$l_{3D}$</th>
<th>$\theta_{3D}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Car Image]</td>
<td>47m</td>
<td>0.98m</td>
<td>1.68m</td>
<td>2.18m</td>
<td>$-0.43r$</td>
</tr>
<tr>
<td>![Car Image]</td>
<td>42m</td>
<td>1.34m</td>
<td>1.60m</td>
<td>3.10m</td>
<td>$0.14r$</td>
</tr>
<tr>
<td>![Car Image]</td>
<td>37m</td>
<td>1.41m</td>
<td>1.59m</td>
<td>3.31m</td>
<td>$0.27r$</td>
</tr>
<tr>
<td>![Car Image]</td>
<td>6m</td>
<td>1.55m</td>
<td>1.55m</td>
<td>3.76m</td>
<td>$-0.04r$</td>
</tr>
</tbody>
</table>
Visualizing the 3D Default Anchors

Default 3D Anchors Projected in Image View ($n_a = 12$)
Visualizing the 3D Default Anchors

Sliding Window Default 3D Anchors in Image View \((n_a = 12)\)
2D / 3D Anchor Supervision

Each anchor box is matched to a ground truth when 2D IoU > 0.5

Default Anchor

\[
\begin{align*}
x_P' &= x + t_{x_P} \cdot w_{2D}, \\
y_P' &= y + t_{y_P} \cdot h_{2D}, \\
z_P' &= t_{z_P} + z_P, \\
\theta_3D' &= t_{\theta_3D} + \theta_{3D}.
\end{align*}
\]

\[
\begin{align*}
w_{3D}' &= \exp(t_{w_{3D}}) \cdot w_{3D}, \\
h_{3D}' &= \exp(t_{h_{3D}}) \cdot h_{3D}, \\
l_{3D}' &= \exp(t_{l_{3D}}) \cdot l_{3D},
\end{align*}
\]

Target Box

\[
\begin{align*}
x_{2D}' &= x + t_{x_{2D}} \cdot w_{2D}, \\
y_{2D}' &= y + t_{y_{2D}} \cdot h_{2D}, \\
h_{2D}' &= t_{h_{2D}} + h_{2D}, \\
w_{2D}' &= \exp(t_{w_{2D}}) \cdot w_{2D}, \\
h_{2D}' &= \exp(t_{h_{2D}}) \cdot h_{2D}.
\end{align*}
\]
2D / 3D Anchor Supervision

Each anchor box is matched to a ground truth when 2D IoU > 0.5

![Default Anchor](image1)  ![Target Box](image2)

Network Outputs (per box)

\[
\begin{align*}
\mathbf{t}_x, \mathbf{t}_y, \mathbf{t}_z & \quad \mathbf{p}_c \\
[t_w, t_h, t_t, t_\theta]_{3D} & \quad b_{3D} \\
[t_x, t_y, w, h]_{2D} & \quad b_{2D}
\end{align*}
\]

\[
L_c = -\log \left( \frac{\exp(\hat{c})}{\sum_i^n \exp(c_i)} \right)
\]

\[
L_{b_{3D}} = \text{SmoothL}_1(b_{3D}, \hat{b}_{3D})
\]

\[
L_{b_{2D}} = -\log (\text{IoU}(b_{2D}, \hat{b}_{2D}))
\]
M3D-RPN Architecture

Network Architecture

- **DenseNet**
  - $h = 32$, $w = 110$

- **$F_{\text{global}}$**
  - $k = 3 \times 3$
  - $b \approx 1$

- **$O_{\text{global}}$**
  - $k = 1 \times 1$
  - $b \approx 1$

- **$F_{\text{local}}$**
  - $k = 3 \times 3$
  - $b = 32$

- **$O_{\text{local}}$**
  - $k = 1 \times 1$
  - $b = 32$

- **3D→2D $\theta$ Optimization**

- **Input Feature Map** $h \times w$

Depth-Aware Convolution

- **Kernel**
  - $3 \times 3$
  - $k_1$
  - $3 \times 3$
  - $k_2$
  - $3 \times 3$
  - $k_3$
  - $3 \times 3$
  - $k_4$
  - $b$

- **Input Feature Map** $h \times w$

- **Output Feature Map** $b$
Post-optimization Algorithm

Initialize: $b_{2D}$, $b_{3D}$, $\theta$
$\eta = \text{IoU}(b_{3D}, \theta, b_{2D})$

- **update $\eta$ and $\theta$**
- **local neighborhood**
  $\Phi_+ = \text{IoU}(b_{3D}, \theta + \sigma, b_{2D})$
  $\Phi_- = \text{IoU}(b_{3D}, \theta - \sigma, b_{2D})$

- **compare**
  $\Phi_+ > \eta$
  $\Phi_- > \eta$

- **continue**
  $\sigma > \beta$

- **no**
- **yes**

- **iter 1**

- **yes**
  $b_{3D}$, $\theta$, $b_{2D}$

- **no**
Experiments

<table>
<thead>
<tr>
<th>Type</th>
<th>Easy</th>
<th>IoU ≥ 0.7 [val1 / val2 / test]</th>
<th>Mod</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3D-RPN</td>
<td>Mono</td>
<td><strong>25.94 / 26.29 / 21.29</strong></td>
<td><strong>21.18 / 19.19 / 15.23</strong></td>
<td><strong>17.90 / 16.83 / 13.16</strong></td>
</tr>
</tbody>
</table>

Table 1. Bird’s Eye View. Comparison of our method to image-only 3D localization frameworks on the BEV task (AP_{BEV}).

<table>
<thead>
<tr>
<th>Type</th>
<th>Easy</th>
<th>IoU ≥ 0.7 [val1 / val2 / test]</th>
<th>Mod</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono3D [1]</td>
<td>Mono</td>
<td>2.53 / - / -</td>
<td>2.31 / - / -</td>
<td>2.31 / - / -</td>
</tr>
<tr>
<td>Deep3DBox [3]</td>
<td>Mono</td>
<td>- / 5.85 / -</td>
<td>- / 4.10 / -</td>
<td>- / 3.84 / -</td>
</tr>
<tr>
<td>Multi-Fusion [4]</td>
<td>Mono</td>
<td>10.53 / 7.85 / 7.08</td>
<td>5.69 / 5.39 / 5.18</td>
<td>5.39 / 4.73 / 4.68</td>
</tr>
<tr>
<td>M3D-RPN</td>
<td>Mono</td>
<td><strong>20.27 / 20.77 / 15.52</strong></td>
<td><strong>17.06 / 16.68 / 11.44</strong></td>
<td><strong>15.21 / 13.42 / 9.62</strong></td>
</tr>
</tbody>
</table>

Table 2. 3D Detection. Comparison of our method to image-only 3D localization frameworks on the 3D Detection task (AP_{3D}).

Experiments (ablations)

<table>
<thead>
<tr>
<th></th>
<th>Bird’s Eye View (AP_{BEV})</th>
<th>3D Detection (AP_{3D})</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Mod</td>
</tr>
<tr>
<td>Car</td>
<td>21.29</td>
<td>15.23</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>12.80</td>
<td>11.66</td>
</tr>
<tr>
<td>Cyclist</td>
<td>2.19</td>
<td>4.55</td>
</tr>
</tbody>
</table>

**Table 3. Multi-class 3D Localization**

<table>
<thead>
<tr>
<th>b</th>
<th>Post-Optim</th>
<th>AP_{2D}</th>
<th>AP_{3D}</th>
<th>AP_{BEV}</th>
<th>RT (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>82.16</td>
<td>10.99</td>
<td>12.99</td>
<td>118</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>82.16</td>
<td>15.08</td>
<td>17.47</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>✓</td>
<td>82.88</td>
<td>12.87</td>
<td>17.91</td>
<td>133</td>
</tr>
<tr>
<td>16</td>
<td>✓</td>
<td>84.15</td>
<td>14.46</td>
<td>19.14</td>
<td>134</td>
</tr>
<tr>
<td>32</td>
<td>✓</td>
<td>83.86</td>
<td>16.04</td>
<td>20.99</td>
<td>143</td>
</tr>
</tbody>
</table>

**Table 4. Ablation Experiments**
Qualitative Results
Qualitative Results
Video Demo
Part 2

Monocular 3D Object Reconstruction

Tran Luan*, Feng Liu*, Xiaoming Liu, “Intrinsic 3D Decomposition and Modeling for Genetic Objects via Colored Occupancy Field,” under review
Monocular 3D Reconstruction
Prior Works

- Limited quality of 3D shapes or 2.5D results
- Unable to reconstruct in-the-wild images
- Lack of appearance modeling

PointSetGen [CVPR’17]
AtlasNet [CVPR’18]
ShapeHD [ECCV’18]
Motivation

- Train and test on real images (w/o GT shapes)
- Complete generic object model

- Efficient image formation for generic object
Proposed Framework

Model fitting

Model learning

Input image $I$

Output $E$

Output $D_S$

Output $D_A$

Occupancy field

Color Field

Normal

Shading

Albedo

Reconstruction

Self-supervised Loss
Shape Representation

(a) Voxel

[Maturana et al. 2015]
[Wu et al. 2015] (GAN)
[Qi et al. 2016]
[Liu et al. 2016]
[Wang et al. 2017] (O-Net)
[Tatarchenko et al. 2017] (OGN)

(b) Point

[Qi et al. 2017] (PointNet)
[Fan et al. 2017] (PointSetGen)

(c) Mesh

[Defferard et al. 2016]
[Henaff et al. 2015]
[Yi et al. 2017] (SyncSpecCNN)

(d) Occupancy Field

[Park et al. 2019 (Deep SDF)]
[Chen et al. 2019]
[Mescheder et al. 2019]
Shape Representation

\[ \mathcal{D}_S : \mathbb{R}^3 \times \mathbb{R}^{d_S} \to V \]

\[ V = [0, 1] \]

Probability of occupancy

0

\[ \mathcal{D}_S \]

128

f_s

x

y

z

3

Shape Decoder
Albedo Representation

$D_A \colon \mathbb{R}^3 \times \mathbb{R}^{d_T} \times \mathbb{R}^{d_S} \to \mathbb{R}^3$

Color value (RGB)

$[0.19 \ 0.19 \ 0.19]$
Rendering

Surface point \( x_j \)

\[
I_j = A_j \cdot \sum_{b=1}^{B^2} \gamma_b H_b(n_j)
\]

\[
n_j = \sigma \left( \frac{\delta D_S(x_j)}{\delta x_j} \right)
\]

\[
A_j = D_A(f_A, f_S, x_j)
\]
Model Training

- Supervised prior learning with synthetic images
  - Learning shape decoder
  - Learning albedo decoder and encoder
- Unsupervised joint modeling

\[ \mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{sil}} \mathcal{L}_{\text{sil}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} \]
Experiments - Expressiveness

Latent Space

3D Shape
Experiments - Expressiveness
Experiments - Synthetic images

Input image

AtlasNet

IM-SVR

Proposed

GT

Experiments - Real Images

(ShapeHD) Wu et al. Learning shape priors for single-view 3D completion and reconstruction. In ECCV 2018
## Experiments - Quantitative Evaluation

Chamfer Distance (CD) on ShapeNet synthetic image reconstruction. The CD is multiplied by $10^3$

<table>
<thead>
<tr>
<th>Method</th>
<th>Car</th>
<th>Chair</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtlasNet25</td>
<td>3.45</td>
<td>3.15</td>
<td>3.30</td>
</tr>
<tr>
<td>AtlasNetO</td>
<td>4.67</td>
<td>4.27</td>
<td>4.47</td>
</tr>
<tr>
<td>IM-SVR</td>
<td>3.92</td>
<td>3.51</td>
<td>3.71</td>
</tr>
<tr>
<td>Proposed</td>
<td>3.24</td>
<td>3.19</td>
<td>3.22</td>
</tr>
</tbody>
</table>

CDs on PASCAL 3D+ database (in-the-wild image reconstruction)

<table>
<thead>
<tr>
<th>Method</th>
<th>Car</th>
<th>Chair</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeHD</td>
<td>5.82</td>
<td>6.17</td>
<td>6.00</td>
</tr>
<tr>
<td>Proposed</td>
<td>3.24</td>
<td>3.38</td>
<td>3.31</td>
</tr>
</tbody>
</table>

(ShapeHD) Wu et al. Learning shape priors for single-view 3D completion and reconstruction. In ECCV 2018
Experiments --- Overfitting Problem Study

Input

Rec.

Closest Shape
Experiments --- Overfitting Problem Study

t-SNE visualization of the shape latent features

Blue: training samples
Red: testing samples
3DMM for ImageNet?
Face Modeling and 3D Recon

Tran and Liu, CVPR 2018, CVPR 2019, PAMI 2019
Part 3  Autonomous Driving Efforts

High Level

Middle Level

Low Level
Depth Completion

Depth Completion

Vision at Night

Transfer a night image into daytime-like!

Input

Output

Image Super Resolution

Scale factor 4 super-resolution results
Contributions

Shortcomings of very deep networks

- Too much parameters were introduced!
- Both storage/memory and overfitting issues

Monocular Pedestrian Detection

- Simultaneous detection and segmentation:
  - Pedestrian classification, Bounding box regression, Semantic segmentation
  - Segmentation masks “illuminate” pedestrians in shared feature maps
Autoregressive Pedestrian Detection

- Efficient de-encoder modules to learn new features (diversification $\uparrow\downarrow$)
- Autoregressive phases to iteratively improve predictions (refinement $\rightarrow$)

<table>
<thead>
<tr>
<th></th>
<th>Caltech Reas.</th>
<th>Caltech Occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$MR_{10}^O$</td>
<td>$MR_{10}^O$</td>
</tr>
<tr>
<td>MS-CNN</td>
<td>9.95</td>
<td>22.45</td>
</tr>
<tr>
<td>RPN+BF</td>
<td>9.58</td>
<td>18.60</td>
</tr>
<tr>
<td>F-DNN</td>
<td>8.65</td>
<td>19.92</td>
</tr>
<tr>
<td>SDS-RCNN</td>
<td>7.36</td>
<td>17.82</td>
</tr>
<tr>
<td>GDFL</td>
<td>7.85</td>
<td>19.86</td>
</tr>
<tr>
<td>DSSD</td>
<td>10.85</td>
<td>18.20</td>
</tr>
<tr>
<td>AR-RPN (ours)</td>
<td>8.01</td>
<td>21.62</td>
</tr>
<tr>
<td>AR-Ped (ours)</td>
<td><strong>6.45</strong></td>
<td><strong>15.54</strong></td>
</tr>
</tbody>
</table>

State-of-the-Art Pedestrian Detection
Traffic Monitoring
Object Forecasting

At 30 miles per hour, vehicle can travel 2 meters in 0.15 second!

Adam Terwilliger, Garrick Brazil, Xiaoming Liu, “Recurrent Flow-Guided Segmentation Prediction,” in WACV 2019
Object Forecasting

t

t+0.15
Object Forecasting
Trailer Coupler Detection

Develop an automatic computer vision (CV) system utilizing a rear-view camera to detect/track the coupler of the trailer.

Automatic control

Results

Distance error: 0.129 meters
Conclusions

• 3D detection from 2D imagery is promising and only at its infancy.

• Category-specific 3D reconstruction can leverage strong prior.

• Many interesting and challenging CV problems in autonomous driving.
Thanks

Sponsors:

[Logos of various sponsors]
Questions?