SYGNet: A SVD-YOLO based GhostNet for Real-time Driving Scene Parsing

Hewei Wang\textsuperscript{1}, Bolun Zhu\textsuperscript{1}, Yijie Li\textsuperscript{1}, Kaiwen Gong\textsuperscript{1}, Ziyuan Wen\textsuperscript{1}, Shaofan Wang\textsuperscript{2}, and Soumyabrata Dev\textsuperscript{3,4}

\textsuperscript{1}Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China
\textsuperscript{2}Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
\textsuperscript{3}The ADAPT SFI Research Centre, Dublin D04V1W8, Ireland
\textsuperscript{4}School of Computer Science, University College Dublin, Dublin D04V1W8, Ireland
Problem statement

- Motivation
- Contribution of this paper
The usage of autonomous driving has been realized for years but exhibiting the two shortcomings on many occasions:

1. Bad usage accuracy in high-speed driving. Autonomous driving has a higher failure rate for high-speed targets.

2. In close traffic jams, autonomous driving cannot judge the complicated road conditions ahead. Even the above conditions are different, but all of them have got a common point which can be viewed as the inaccuracy of the algorithm.

To solve the aforementioned issues, we propose SYGNet for improving scene parsing accuracy.
Contributions of this paper

1. A novel autonomous driving recognition model named SYGNet is proposed wherein we introduce feature extraction component and combine SVD-YOLO and GhostNet as a subsequent component.

2. SYGNet produces promising results in recognition accuracy, loss value, train and test condition and qualitative figures on KITTI dataset.

3. The model, diagrams, dataset, and the three types of experiment's results in this paper are available at: https://github.com/WangHewei16/SYGNet-for-Real-time-Driving-Scene-Parsing
Proposed Method

- Feature Extraction Component
- SVD-YOLO GhostNet Component
This component is used to extract important perceptual scene features includes two branches: LiDAR Stream and Camera Stream, which extract point cloud features and image features respectively.
Firstly, similar to VoxelNet, we divide the original point cloud into equal voxel grids, and then uses the spatial coordinates and the relative offset of random sampling points as the representation of each voxel. The relative offset of the points in the j-th voxel grid is the offset of each point from its centroid, the centroid of the j-th with the format:

\[
v_x^j = \frac{\sum_{i=1}^{M} x_i^j}{M}, \quad v_y^j = \frac{\sum_{i=1}^{M} y_i^j}{M}, \quad v_z^j = \frac{\sum_{i=1}^{M} z_i^j}{M}
\]

where M is the number of point clouds in the j-th voxel. Although this representation can capture the global spatial information of the point cloud, it ignores the local structure information of the midpoint of each voxel. In order to capture local structural information, this paper designs local directional features, which are calculated by the following equation:

\[
d^j = \frac{\sum_{i=1}^{M} \arctan \left( \frac{y_i^j - y_x^j}{x_i^j - x_x^j} \right)}{M}
\]

The final representation of the i-th point in the j-th voxel grid can be rewritten as:

\[
V_{in} = \left\{ x_i^j, y_i^j, z_i^j, r_i^j, x_x^j - v_x^j, y_y^j - v_y^j, z_x^j - v_z^j, d^j \right\}
\]
**SVD**
decompose image data to obtain data with stronger spatial and environmental characteristics.

**YOLOv3**
predict the bounding box, determines anchor box through size clustering.

**GhostNet**
reduces the computational costs of deep neural networks by utilizing fewer filters to generate some intrinsic feature maps. To achieve real-time
Experiment and Discussion
Experiment and Discussion

Ablation study

Quantitative evaluation

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Cars</th>
<th>People</th>
<th>Edge</th>
<th>Side</th>
<th>Light</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv3</td>
<td>78.9</td>
<td>76.4</td>
<td>77.1</td>
<td>73.4</td>
<td>79</td>
</tr>
<tr>
<td>RCNN</td>
<td>82.8</td>
<td>79.9</td>
<td>81.1</td>
<td>77.3</td>
<td>82.6</td>
</tr>
<tr>
<td>SVD-YOLOv2</td>
<td>86.7</td>
<td>83.4</td>
<td>85.1</td>
<td>81.2</td>
<td>86.2</td>
</tr>
<tr>
<td>SVD-RCNN</td>
<td>90.2</td>
<td>87.4</td>
<td>89</td>
<td>84.8</td>
<td>86.2</td>
</tr>
<tr>
<td>SVD-YOLOv3</td>
<td>94.1</td>
<td>90.9</td>
<td>93</td>
<td>88.7</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Qualitative evaluation
Conclusion
Conclusion

Summary of content

- Study the autonomous driving scene parsing technology.
- Analyze some inherent problems related to training time of the neural nets, and proposes SYGNet.
- In the feature extraction component, we propose an algorithm based on VoxelNet to extract point cloud features and image features.
- In SVD-YOLO GhostNet component, SVD decompose the image data and then use YOLOv3 to obtain the future map, then convert to GhostNet, which is used to realize the real-time scene parsing by utilizing fewer filters to generate some intrinsic feature maps.
- The experimental results show that SYGNet can effectively and significantly improve the scene parsing and recognition ability of autonomous driving under traffic jams or complex road conditions.

Future work

- We intend to focus on the reuse and fusion of visual transformer and combination with DeepBillboard so as to greatly improve the accuracy of autonomous driving technology.