INCORPORATING LUMINANCE, DEPTH AND COLOR INFORMATION BY A FUSION-BASED NETWORK FOR SEMANTIC SEGMENTATION

Shang-Wei Hung, Shao-Yuan Lo, Hsueh-Ming Hang

National Chiao Tung University, UC San Diego
Outline

- Introduction
- Method
- Experiments
- Conclusion
- References
Introduction
Introduction

Road Scene
Semantic Segmentation
Method
Method

- RGB Encoder and Decoder
- D&Y Encoder
- Fusion Mechanism
RGB Encoder and Decoder

- Use ERFNet [Romera et al.] as our backbone network.
- Reach good balance between accuracy and complexity.
- Use three downsampler block as encoder.
- Use deconvolution filter as decoder.
D&Y Encoder

- Adapt from FuseNet [Hazirbas et al.].
- Adopt dense connectivity from DenseNet [Gao et al.].
- Add shallow dense block to extract boundary information.
- Stack luminance images into depth maps to suppress noises.
Fusion Mechanism

- Direct stacking cannot effectively exploit the depth information.
- Conduct fusion operation on different scale.
- Use element-wise summation for each fusion.
- 1×1 convolution layer is used for matching the number of channels.
Experiments
Experiments

- Implementation Details
  - Optimizer: Adam
  - Learning rate initialization: 0.0005
  - Learning rate policy: Poly
  - Weight decay: 0.0001
  - Use class weighting: \( \omega_{\text{class}} = \frac{1}{\ln(c+p_{\text{class}})} \)
Experiements

- Datasets: Cityscapes
Ablation Study

- Simply stacking RGB and D channels cannot benefit from the additional depth information.
- Our fusion mechanism is a more effective design for depth information extraction.

| Method       | RGB Inputs | Depth Maps | Y Info. | Shallow Block | Dense Connects | mIoU (%) | Params |
|--------------|------------|------------|---------|---------------|----------------|----------|--------|
| ERFNet-Depth | ●          | ●          |         |               |                | 47.48    | 1.97M  |
| ERFNet-RGB   | ●          | ●          |         |               |                | 65.59    | 1.97M  |
| ERFNet-Stack | ●          | ●          |         |               |                | 65.06    | 1.97M  |
| LDFNet       | ●          | ●          | ●       | ●             | ●              | 68.48    | 2.31M  |
Ablation Study

- Adopting dense connectivity can obtain a higher mIoU score with fewer parameters.

| Method       | RGB Inputs | Depth Maps | Y Info. | Shallow Block | Dense Connects | mIoU (%) | Params |
|--------------|------------|------------|---------|---------------|----------------|----------|--------|
| LDF-non-Dense| ●          | ●          | ●       |               |                | 66.53    | 2.95M  |
| LDFNet       | ●          | ●          | ●       | ●             | ●              | 68.48    | 2.31M  |
Ablation Study

- Depth information has a strong correlation to the object edge, contour, and boundary information, so placing Shallow Block at the early stage is beneficial to extract these desired low-level features.

| Method              | RGB Inputs | Depth Maps | Y Info. | Shallow Block | Dense Connects | mIoU (%) | Params |
|---------------------|------------|------------|---------|---------------|----------------|----------|--------|
| LDF-w/o-Shallow     | ●          | ●          | ●       |               |                | 66.54    | 2.20M  |
| LDF-58-w/o-Shallow  | ●          | ●          | ●       |               |                | 65.93    | 2.42M  |
| LDFNet              | ●          | ●          | ●       | ●             | ●              | 68.48    | 2.31M  |
Incorporating luminance information achieves a great improvement.

| Method     | RGB Inputs | Depth Maps | Y Info. | Shallow Block | Dense Connects | mIoU (%) | Params |
|------------|------------|------------|---------|---------------|----------------|----------|--------|
| LDF-w/o-Y  | ●          | ●          | ●       | ●             | ●              | 65.72    | 2.31M  |
| LDFNet     | ●          | ●          | ●       | ●             | ●              | 68.48    | 2.31M  |
Ablation Study

- The increased parameters indeed provide some improvements, but our fusion mechanism of incorporation multi-modal information contributes significantly more.

| Method        | RGB Inputs | Depth Maps | Y Info. | Shallow Block | Dense Connects | mIoU (%) | Params |
|---------------|------------|------------|---------|---------------|----------------|----------|--------|
| ERFNet-RGB    | ●          | ●          | ●       |               |                | 65.59    | 1.97M  |
| LDF-RGB-RGB   | ●          | ●          | ●       | ●             | ●              | 67.79    | 2.31M  |
| LDFNet        | ●          | ●          | ●       | ●             | ●              | 68.48    | 2.31M  |
Comparison

Table 2: Evaluation results on the Cityscapes test set, comparing LDFNet with the other RGB-D methods.

| Method                        | mIoU (%) | Speed (fps) |
|-------------------------------|----------|-------------|
| MultiBoost                    | 59.3     | 4.0         |
| Pixel-level Encoding [16]     | 64.3     | n/a         |
| Scale invariant CNN+CRF [10]  | 66.3     | n/a         |
| RGB-D FCN                     | 67.4     | n/a         |
| LDFNet (ours)                 | 71.3     | 18.4        |

Table 3: Comparison of model efficiency with RGB methods. Sub: the amount of subsampling used by the method at test time.

| Method               | Parameters | Sub | Speed (fps) |
|----------------------|------------|-----|-------------|
| DeepLabv2 [2]        | 44.0M      | no  | n/a         |
| PSPNet [20]          | 65.7M      | no  | n/a         |
| Dilation10 [19]      | 140.8M     | no  | 0.25        |
| FCN-8s [12]          | 134.5M     | no  | 2.0         |
| SegNet [1]           | 29.5M      | 4   | 16.7        |
| LDFNet (ours)        | 2.31M      | 2   | 18.4        |
### Results

| RGB Image | Depth Map | Ground Truth | LDFNet |
|-----------|-----------|--------------|--------|
| ![RGB Image](image1) | ![Depth Map](image2) | ![Ground Truth](image3) | ![LDFNet](image4) |
| ![RGB Image](image5) | ![Depth Map](image6) | ![Ground Truth](image7) | ![LDFNet](image8) |
| ![RGB Image](image9) | ![Depth Map](image10) | ![Ground Truth](image11) | ![LDFNet](image12) |
| ![RGB Image](image13) | ![Depth Map](image14) | ![Ground Truth](image15) | ![LDFNet](image16) |
## Results

| RGB Image | Depth Map | Ground Truth | LDFNet |
|-----------|-----------|--------------|--------|
| ![RGB Image 1](image1.jpg) | ![Depth Map 1](image2.jpg) | ![Ground Truth 1](image3.jpg) | ![LDFNet 1](image4.jpg) |
| ![RGB Image 2](image5.jpg) | ![Depth Map 2](image6.jpg) | ![Ground Truth 2](image7.jpg) | ![LDFNet 2](image8.jpg) |
| ![RGB Image 3](image9.jpg) | ![Depth Map 3](image10.jpg) | ![Ground Truth 3](image11.jpg) | ![LDFNet 3](image12.jpg) |

Note: Images are placeholders and should be replaced with actual images.
Conclusion
We propose a novel solution named LDFNet, which incorporates Luminance, Depth and Color information by a fusion-based network. It includes a sub-network to process depth maps and employs luminance images to assist the depth information in processes. LDFNet outperforms the other state-of-art systems on the Cityscapes dataset, and its inference speed is faster than most of the existing networks. The experimental results show the effectiveness of the proposed multi-modal fusion network and its potential for practical applications.
The End

Thank you for your attention