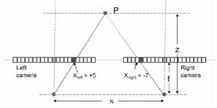
Joint stereo matching and semantic segmentation with CNN

WANG Zhicong & CHEN Ran

Contact Information:
Information Engineering
Chinese University of Hong Kong
Shatin, NT, Hong Kong SAR

Introduction

Convolutional neural networks have become the method of choice in many yields of computer vision. There have been a lot of works done on semantic segmentation for perpixel predictions. Depth estimation from joint stereo is another interesting topic which can widely be implemented in the networks.In this paper, we try to combine these two tasks with the idea of multi-task to improve the accuracy of the segmentation.



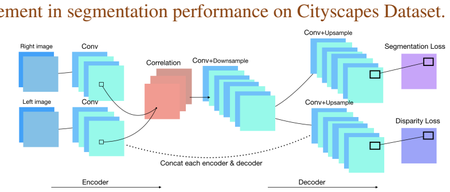
Dataset

The Cityscapes Dataset is a recently released dataset for semantic urban scene understanding.It focuses on semantic understanding of urban street scenes. It provide about 3000 image pairs for training and 500 for validation.

Dataset	Cityscapes
Complexity	19 classes
Volume	5000 annotated images with fine annotations
Type of annotations	Semantic ,Instance-wise Dense pixel annotations
Origin Size	1024*2048

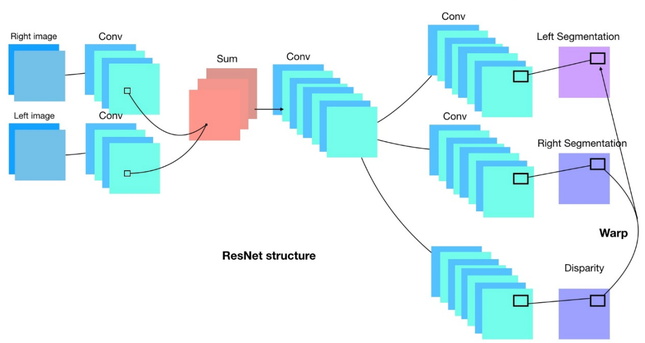
Review on last semester

Last semester, we worked on FlowNet[1]. In order to use the features for stereo matching to do semantic segmentation, we simply added a decoder for segmentation on origin network and got around 8% improvement in segmentation performance on Cityscapes Dataset.



Model

The main idea of our model is based on ResNet and hard parameter sharing for multi-task learning.



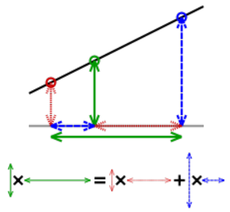
Warp method

Reasons:

- No segmentation ground truth for the right image in the dataset
- Make full use of the origin segmentation network
- As a feature learning task, right image features may provide further confidence on result

Principle of warp:

- linear interpolation
- $W(x,y) = R(floor(D(x,y)),y) * (ceil(D(x,y)) - D(x,y)) + R(ceil(D(x,y)),y) * (D(x,y) - floor(D(x,y)))$



In training phase(Caffe)

- Implement Warp Layer in Caffe
- Follow the previous structure when training
- Try different settings, like loss weights and learning rate
- No improvement

network	settings	mIOU
PSPNet	left only(baseline)	79.40%
	warp network fused output	76.79%
	warp network single left output	77.36%
ResNet	left only(baseline)	69.91%
	warp network fused output	69.24%

In testing phase(Matlab)

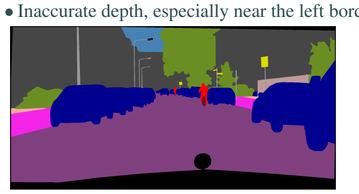
- Use matlab to test and evaluate the performance
- Run the origin PSPNet or ResNet separately for left image and right image
- Warp and fuse the images after generated by the origin network
- No improvement
- Some problems

network	settings	mIOU
PSPNet	left only(baseline)	79.40%
	fused result with no dilation	78.48%
	fused result + 8*8 dilation + mask 60	79.08%
ResNet	left only(baseline)	69.91%
	right only	67.13%
	bilinear warp on right image	69.55%
	warp network fused output	69.77%
	warped right + 8*8 dilation + mask 60	70.72%
	fused output	69.88%

Problems & solutions

Here is a sample set of images about the warpped result. There are two main reasons for lower performance

- Wrong depth at left border of the image, which is mainly caused by the view difference of two cameras
- Inaccurate depth, especially near the left border of an object, which is caused by occlusion

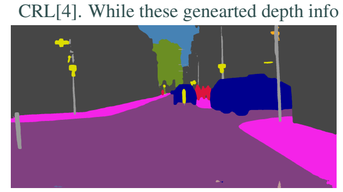


To solve the problems, we use the following methods:

- Mask first 60 columns
- Use dilation to expand the ignore points

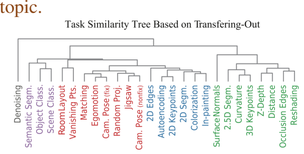
Possible reasons for the failure:

- Confliction on object border → We want to get improvement on segmentation from right image features especially on borders, while because of the occlusion problem, we have to use dilation method to ignore these pixels on borders.
- Inaccuracy of the depth map → We try other depth map generators like CPM[2], SVS[3] and CRL[4]. While these genearted depth info are not robust enough →No solid improvement



Future direction

Although warp method performs not well in this experiment, there may be some other tricks to combine two tasks together smartly. We should learn more from recent papers on multi task learning, especially those related to our topic.



A recent work[5] considered these two tasks as a transfer learning task and get the conclusion that segmentation and depth estimation are not quite related, which means maybe from other views, it's not reasonable to make them connected.

References

[1] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[2] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single View Stereo Matching. *ArXiv e-prints*.

[4] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCV Workshop on Geometry Meets Deep Learning*, Oct 2017.

[5] Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.