Lab CudaVision
Learning Vision Systems on Graphics Cards (MA-INF 4308)

Lab Project

20.07.2021

PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

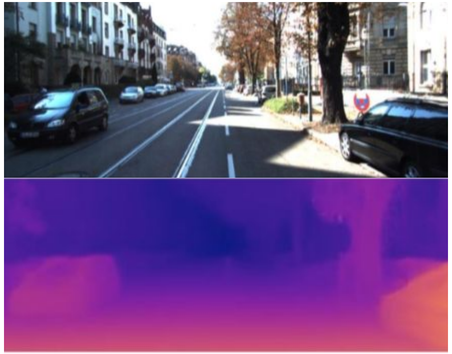
Contact: villar@ais.uni-bonn.de

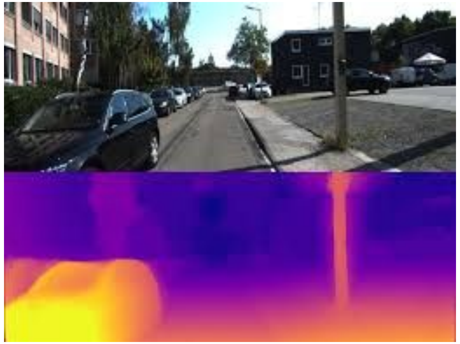
Stereo Depth Estimation

Depth Estimation

- Estimating distance and depth
  - Trivial for humans
  - Very challenging for machines

- Approaches
  - Radar/Lidar
  - Epipolar geometry
  - Deep Learning (mono or stereo)





Stereo Depth Estimation

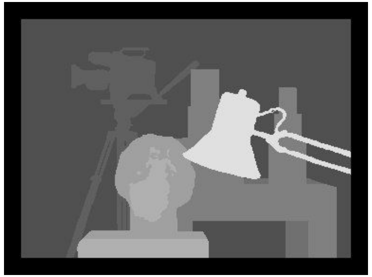
- Using information from left and right cameras
  - Imitates human physiology

- **Disparity:** distance between two corresponding points in the left and right image of a stereo pair

$$Depth\ (mm) = \frac{focal\ length(pixels)\ x\ Baseline(mm)}{Disparity\ (pixels)}$$

Left Image

Right Image



Model

Model Inspiration

End-to-End Learning of Geometry and Context for Deep Stereo Regression

Alex Kendall Hayk Martirosyan Saumitro Dasgupta Peter Henry
Ryan Kennedy Abraham Bachrach Adam Bry
Skydio Inc.
{alex,hayk,saumitro,peter,ryan,abe,adam}@skydio.com

Abstract

We propose a novel deep learning architecture for regressing disparity from a rectified pair of stereo images. We leverage knowledge of the problem's geometry to form a cost volume using deep feature representations. We learn to incorporate contextual information using 3-D convolutions over this volume. Disparity values are regressed from the cost volume using a proposed differentiable soft argmin operation, which allows us to train our method end-to-end to sub-pixel accuracy without any additional post-processing or regularization. We evaluate our method on the Scene Flow and KITTI datasets and on KITTI we set a new state-of-the-art benchmark, while being significantly faster than competing approaches.

1. Introduction

Accurately estimating three dimensional geometry from stereo imagery is a core problem for many computer vision applications, including autonomous vehicles and UAVs [2]. In this paper we are specifically interested in computing the disparity of each pixel between a rectified stereo pair of images. To achieve this, the core task of a stereo algorithm is computing the correspondence of each pixel between two images. This is very challenging to achieve robustly in real-world scenarios. Current state-of-the-art stereo algorithms often have difficulty with textureless areas, reflective surfaces, thin structures and repetitive patterns. Many stereo algorithms aim to mitigate these failures with pooling or gradient based regularization [15, 23]. However, this often requires a compromise between smoothing surfaces and detecting detailed structures.

In contrast, deep learning models have been successful in learning powerful representations directly from the raw data in object classification [28], detection [17] and semantic segmentation [31, 3]. These examples demonstrate that deep convolutional neural networks are very effective for understanding semantics. They excel at classification tasks when supervised with large training datasets. We observe that a number of these challenging problems for stereo algorithms would benefit from knowledge of global semantic context, rather than relying solely on local geometry. For example, given a reflective surface of a vehicle's windshield, a stereo algorithm is likely to be erroneous if it relies solely on the local appearance of the reflective surface to compute geometry. Rather, it would be advantageous to understand the semantic context of this surface (that it belongs to a vehicle) to infer the local geometry. In this paper we show how to learn a stereo regression model which can be trained end-to-end, with the capacity to understand wider contextual information.

Stereo algorithms which leverage deep learning representations have so far been largely focused on using them to generate unary terms [48, 32]. Applying cost matching on the deep unary representations performs poorly when estimating pixel disparities [32, 48]. Traditional regularization and post processing steps are still used, such as semi global block matching and left-right consistency checks [23]. These regularization steps are severely limited because they are hand-engineered, shallow functions, which are still susceptible to the aforementioned problems.

This paper asks the question, can we formulate the entire stereo vision problem with deep learning using our understanding of stereo geometry? The main contribution of this paper is an end-to-end deep learning method to estimate per-pixel disparity from a single rectified image pair. Our architecture is illustrated in Figure 1. It explicitly reasons about geometry by forming a cost volume, while also rea-

Pyramid Stereo Matching Network

Jia-Ren Chang Yong-Sheng Chen
Department of Computer Science, National Chiao Tung University, Taiwan
{followwar.cs00g, yschen}@nctu.edu.tw

Abstract

Recent work has shown that depth estimation from a stereo pair of images can be formulated as a supervised learning task to be resolved with convolutional neural networks (CNNs). However, current architectures rely on patch-based Siamese networks, lacking the means to exploit context information for finding correspondence in ill-posed regions. To tackle this problem, we propose PSM-Net, a pyramid stereo matching network consisting of two main modules: spatial pyramid pooling and 3D CNN. The spatial pyramid pooling module takes advantage of the capacity of global context information by aggregating context in different scales and locations to form a cost volume. The 3D CNN learns to regularize cost volume using stacked multiple hourglass networks in conjunction with intermediate supervision. The proposed approach was evaluated on several benchmark datasets. Our method ranked first in the KITTI 2012 and 2015 leaderboards before March 18, 2018. The codes of PSMNet are available at: https://github.com/JiaRenChang/PSMNet.

1. Introduction

Depth estimation from stereo images is essential to computer vision applications, including autonomous driving for vehicles, 3D model reconstruction, and object detection and recognition [4, 31]. Given a pair of rectified stereo images, the goal of depth estimation is to compute the disparity d for each pixel in the reference image. Disparity refers to the horizontal displacement between a pair of corresponding pixels on the left and right images. For the pixel (x, y) in the

using CNNs treated the problem of correspondence estimation as similarity computation [27, 30], where CNNs compute the similarity score for a pair of image patches to further determine whether they are matched. Although CNN yields significant gains compared to conventional approaches in terms of both accuracy and speed, it is still difficult to find accurate corresponding points in inherently ill-posed regions such as occlusion areas, repeated patterns, textureless regions, and reflective surfaces. Solely applying the intensity-consistency constraint between different viewpoints is generally insufficient for accurate correspondence estimation in such ill-posed regions, and is useless in textureless regions. Therefore, regional support from global context information must be incorporated into stereo matching.

One major problem with current CNN-based stereo matching methods is how to effectively exploit context information. Some studies attempt to incorporate semantic information to largely refine cost volumes or disparity maps [8, 13, 27]. The Displets [8] method utilizes object information by modeling 3D vehicles to resolve ambiguities in stereo matching. ResMatchNet [27] learns to measure reflective confidence for the disparity maps to improve performance in ill-posed regions. GC-Net [13] employs the encoder-decoder architecture to merge multiscale features for cost volume regularization.

In this work, we propose a novel pyramid stereo matching network (PSMNet) to exploit global context information in stereo matching. Spatial pyramid pooling (SPP) [9, 32] and dilated convolution [2, 29] are used to enlarge the receptive fields. In this way, PSMNet extends pixel-level features to region-level features with different scales of receptive fields; the resultant combined global and local feature

On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach

Nikolai Smolyanskiy Alexey Kamenev Stan Birchfield
NVIDIA
{nsmolyanskiy, akamenev, sbirchfield}@nvidia.com

Abstract

We revisit the problem of visual depth estimation in the context of autonomous vehicles. Despite the progress on monocular depth estimation in recent years, we show that the gap between monocular and stereo depth accuracy remains large—a particularly relevant result due to the prevalent reliance upon monocular cameras by vehicles that are expected to be self-driving. We argue that the challenges of removing this gap are significant, owing to fundamental limitations of monocular vision. As a result, we focus our efforts on depth estimation by stereo. We propose a novel semi-supervised learning approach to training a deep stereo neural network, along with a novel architecture containing a machine-learned argmax layer and a custom runtime that enables a smaller version of our stereo DNN to run on an embedded GPU. Competitive results are shown on the KITTI 2015 stereo dataset. We also evaluate the recent progress of stereo algorithms by measuring the impact upon accuracy of various design criteria.[1]

1. Introduction

Estimating depth from images is a long-standing problem in computer vision. Depth perception is useful for scene understanding, scene reconstruction, virtual and augmented reality, obstacle avoidance, self-driving cars, robotics, and other applications.

Traditionally, multiple images have been used to estimate depth. Techniques that fall within this category include stereo, photometric stereo, depth from focus, depth from defocus, time-of-flight,[2] and structure from motion. stereo cameras); and 2) multiple images provide geometric constraints that can be leveraged to overcome the many ambiguities of photometric data.

The alternative is to use a single image to estimate depth. We argue that this alternative—due to its fundamental limitations—is not likely to be able to achieve high-accuracy depth estimation at large distances in unfamiliar environments. As a result, in the context of self-driving cars we believe monocular depth estimation is not likely to yield results with sufficient accuracy. In contrast, we offer a novel, efficient deep-learning stereo approach that achieves compelling results on the KITTI 2015 dataset by leveraging a semi-supervised loss function (using LIDAR and photometric consistency), concatenating cost volume, 3D convolutions, and a machine-learned argmax function. The contributions of the paper are as follows:

- Quantitative and qualitative demonstration of the gap in depth accuracy between monocular and stereoscopic depth.
- A novel semi-supervised approach (combining lidar and photometric losses) to training a deep stereo neural network. To our knowledge, ours is the first deep stereo network to do so.[3]
- A smaller version of our network, and a custom runtime, that runs at near real-time (~20 fps) on a standard GPU, and runs efficiently on an embedded GPU. To our knowledge, ours is the first stereo DNN to run on an embedded GPU.
- Quantitative analysis of various network design choices, along with a novel machine-learned argmax layer that yields smoother disparity maps.

Wasserstein Distances for Stereo Disparity Estimation

Divyansh Garg[1] **Yan Wang**[1] **Bharath Hariharan**[1]
Mark Campbell[1] **Kilian Q. Weinberger**[1] **Wei-Lun Chao**[2]
[1]Cornell University, Ithaca, NY [2]The Ohio State University, Columbus, OH
{dg595, yw763, bh497, mc288, kqw4}@cornell.edu chao.209@osu.edu

Abstract

Existing approaches to depth or disparity estimation output a distribution over a set of pre-defined discrete values. This leads to inaccurate results when the true depth or disparity does not match any of these values. The fact that this distribution is usually learned indirectly through a regression loss causes further problems in ambiguous regions around object boundaries. We address these issues using a new neural network architecture that is capable of outputting arbitrary depth values, and a new loss function that is derived from the Wasserstein distance between the true and the predicted distributions. We validate our approach on a variety of tasks, including stereo disparity and depth estimation, and the downstream 3D object detection. Our approach drastically reduces the error in ambiguous regions, especially around object boundaries that greatly affect the localization of objects in 3D, achieving the state-of-the-art in 3D object detection for autonomous driving. Our code will be available at https://github.com/Div99/W-Stereo-Disp.

1 Introduction

Depth estimation from stereo images is a long-standing task in computer vision [30, 36]. It is a key component of many downstream problems, ranging from 3D object detection in autonomous vehicles [8, 21, 33, 42, 54] to graphics applications such as novel view generation [23, 55]. The importance of this task in practical applications has led to a flurry of recent research. Convolutional networks have now superseded more classical techniques and led to significant improvements in accuracy [4, 27, 43, 57].

These techniques estimate depth by finding accurate pixel correspondences and estimating the *disparity* between their X-coordinates, which is inversely proportional to depth. Because pixels have integral coordinates, so does the estimated disparity — causing even the resulting depth estimates to be discrete. This introduces inaccuracy, as the ground truth disparity and depth are naturally real-valued. This discrepancy is typically addressed by predicting a *categorical distribution* over a fixed set of discrete values, and then computing the *expected* depth from this distribution, which can in theory be any arbitrary real value (within the range of the set) [4, 14, 43, 54, 57].

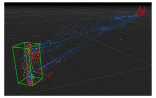
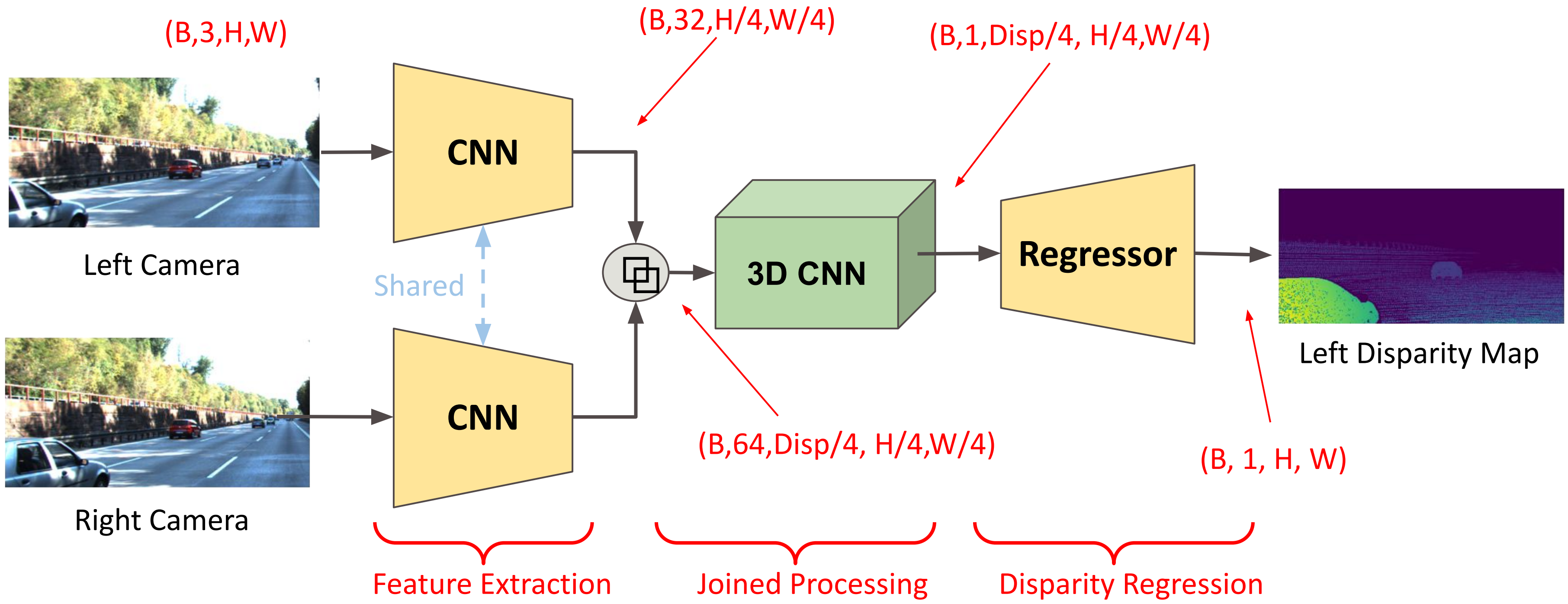


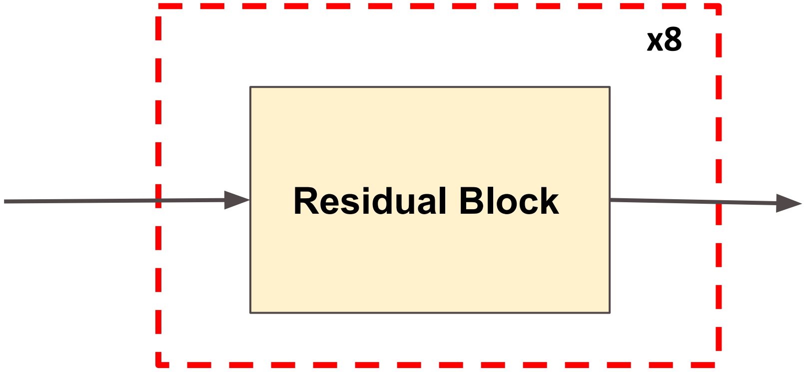
Figure 1: **The effect of our continuous disparity network (CDN).** We show a person (green box) in front of a wall. The blue 3D points are obtained using PSMNet [4]. The red points from our CDN model are much better aligned with the shape of the objects: they do not suffer the streaking artifacts near edges. Yellow points are from the ground truth LiDAR. (One floor square is 1m×1m.)

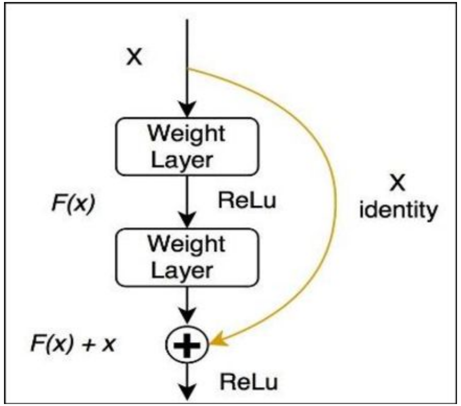
Baseline Model Pipeline



Feature Extraction

- Siamese residual CNN
    - Shared weights for left and right images
- Cascade of residual blocks (e.g., 8 blocks)

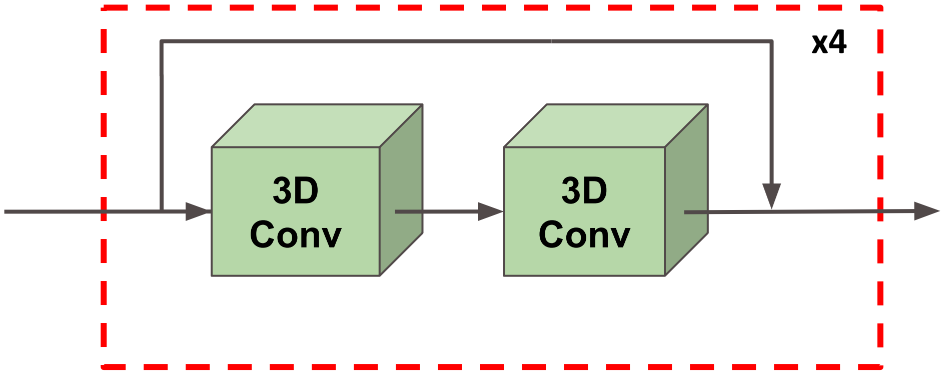




Joined Processing

- Concatenate features from both views into a cost volume
  - (B, 2*C, Disp/4, H/4, W/4)

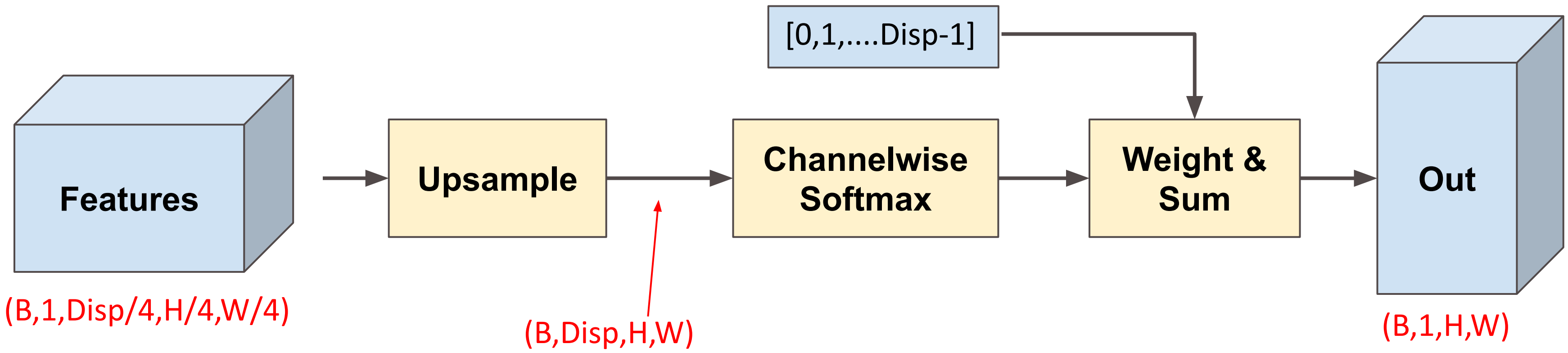
- Process volume with 3D-Convolutions



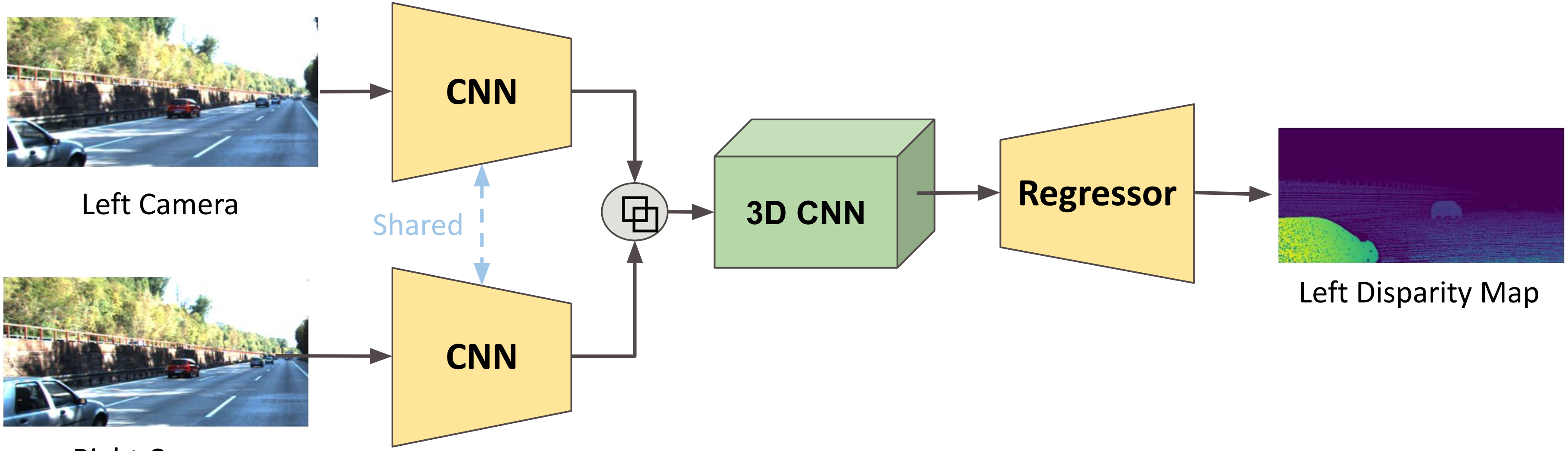

```python
"""
making cost volume: concatenating features across channel dimension
for each disparity level
"""
cost = torch.Tensor(B, C*2, self.max_disp//4, H//4, W//4).to(device)
for i in range(self.max_disp // 4):
    if(i == 0):
        cost[:, :C, i, :,:] = left_feats
        cost[:, C:, i, :,:] = right_feats
    else:
        cost[:, :C, i, :, i:] = left_feats[:,:,:,i:]
        cost[:, C:, i, :, i:] = right_feats[:,:,:,:-i]
```

Disparity Regression

- Upscales the volumetric features
- Performs a **soft-regression** of the disparity values



Baseline Model Pipeline



Left Camera

Right Camera

Left Disparity Map

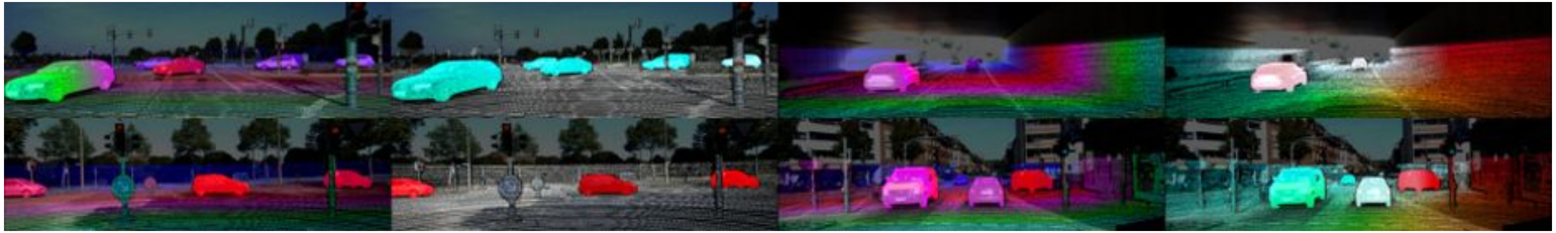
Datasets

SceneFlow Dataset

- Synthetic dataset
  - 35,454 training images
  - 4,370 test images
  - Dense disparity maps
- Use for pretraining model
- https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html

KITTI-2015 Dataset

- Real-world autonomous driving dataset
- Small size for stereo depth:
  - 200 stereo pairs with sparse disparity maps
  - 150/50 training and evaluation split
- http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo



Training & Evaluation

Train/Eval on KITTI

- Train with image crops of size: (3, 256, 512)
- Evaluation on original size: ≅(3, 376, 1240) with batch size of 1
- Predict left disparity map with occluded pixels (*dips_occ*)
- Dataset splits:
  - First 150 image pairs for training
  - Final 50 images for evaluation
- Maximum disparity: Disp=192
- *SmoothL1* loss function for training
  - Considering only pixels with non-zero disparity

Evaluation Metrics

- Only evaluate pixels where true disparity is non-zero
- **SmoothL1** regression loss on evaluation set (β=1)
- **3-pixel error (3PE)**: Percentage of pixels where
  - disparity error is less than 3 pixels
  - error is less than 5% of the true disparity

$$\text{Smooth}\,\ell 1 = \frac{1}{N}\sum_{i}^{N} l_i$$

$$li = \begin{cases} \frac{1}{2\cdot\beta}(\mathbf{X}_i - \hat{\mathbf{X}}_i)^2 & |\mathbf{X}_i - \hat{\mathbf{X}}_i| \leq \beta \\ |\mathbf{X}_i - \hat{\mathbf{X}}_i| - 0.5\cdot\beta & \text{otherwise} \end{cases}$$

$$3PE(\mathbf{X}, \hat{\mathbf{X}}) = 1 - \frac{1}{N}\sum_{i=1}^{N} \text{CorrectDisp}(\hat{x}_i, x_i)$$

$$\text{CorrectDisp}(\hat{x}_i, x_i) = \begin{cases} 1 & ; \text{if } |x_i - \hat{x}_i| < 3 \\ 1 & ; \text{if } |x_i - \hat{x}_i| < 0.05\cdot x_i \\ 0 & ; \text{otherwise} \end{cases}$$

Project Goals and Deliverables

Passing Requirements

1. Implement model, pipelines and utils
2. Beat a weak and simple baseline
    a. No pretraining
    b. Little parameter optimization
3. Create overview notebook
4. Write project report

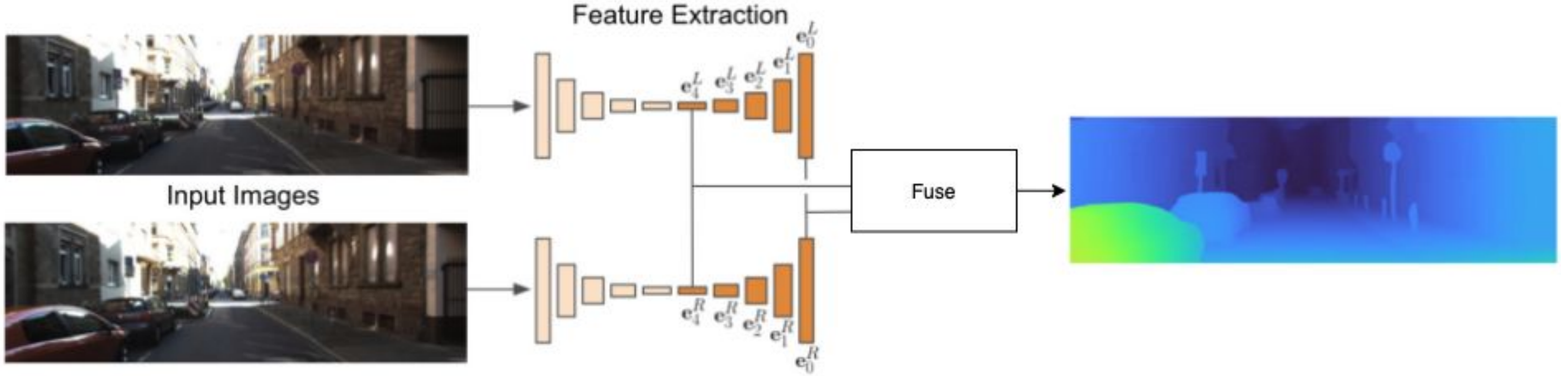
	Mean Loss	Mean 3PE
Baseline	1.1	9.0%

Improvement Ideas

- Test and debug your code
- KITTI dataset is very small. I strongly recommend:
  - Pretrain on SceneFlow
  - Fine-tune on KITTI with data augmentation
- Tweak the model
  - Use a pretrained feature extractor (e.g., ResNet18)
  - Change modules (num. layers, num. kernels, ...)
  - Slightly change the building blocks
- Hyper-parameter and training optimization
  - Optimize for LR, optimizer, batch size, ...
  - Scheduling, learning rate warmup

If Motivated :)

- Implement a model with the following architecture
  - Shared Hourglass/UNet backbone
  - Feature fusion
- Try to obtain the best results possible



Deliverables

- Complete codebase
  - Clean and structured
  - Not just a notebook!
- Trained model checkpoint and tensorboard logs
- Overview notebook (.ipynb & .html) showing main functionalities:
  - Load data
  - Load pretrained model
  - Display some results
- Project report

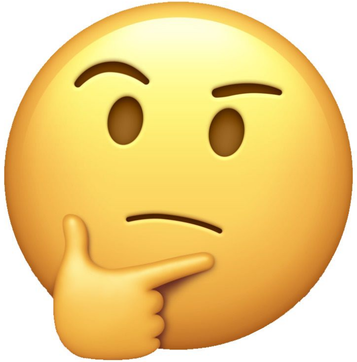
Project Report

- Document your work in the project report
- Try to be brief, but readable and informative
- Include figures and tables
- Use *BibTex* for the references
- I expect 6-10 pages
- Use the following template
    - https://www.overleaf.com/read/tmnvhrsdmjrp

Important Dates

- **20.07**: Starting date
- **03.09**: Draft submission due
- **30.08-15.09**: Revision session
- **30.09**: Final submission:

Questions?



References

1. http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo
2. https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html
3. Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." The International Journal of Robotics Research 32.11 (2013): 1231-1237.
4. Mayer, Nikolaus, et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
5. Luo, Wenjie, Alexander G. Schwing, and Raquel Urtasun. "Efficient deep learning for stereo matching." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
6. Kendall, Alex, et al. "End-to-end learning of geometry and context for deep stereo regression." Proceedings of the IEEE International Conference on Computer Vision. 2017.
7. Chang, Jia-Ren, and Yong-Sheng Chen. "Pyramid stereo matching network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

References

8. Smolyanskiy, Nikolai, Alexey Kamenev, and Stan Birchfield. "On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018.
9. Garg, Divyansh, et al. "Wasserstein distances for stereo disparity estimation." arXiv preprint arXiv:2007.03085 (2020).
10. Menze, Moritz, and Andreas Geiger. "Object scene flow for autonomous vehicles." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.