# GSMNet: Gan-based Stereo Matching Networks

## 摘要

计算机视觉中的深度估计问题，按照算法的基本思想大体上可以分为两类[1]，一类是将深度估计任务看作是一个回归任务，直接对各个像素点的深度值进行预测，另一类是通过立体匹配的方式，对视差进行估计来间接地获得各个像素的深度值。随着强化学习的兴起，对抗生成网络[2]在生成图片方面取得了显著的效果。在本文提出的模型中，我们首先用对抗生成网络来生成双目视图，为了引入更显式的几何约束，输出更鲁棒的深度估计结果，我们接着用立体匹配的方法来获得视差图，最终得到基于单目图像的深度估计结果。

## 简介

随着深度学习的发展，通过神经网络对图像进行特征提取，接着用回归的方法直接对像素点的深度信息进行估计的方法取得了不错的效果，但是这样的方法很大程度上是利用了神经网络强大的表现力和记忆力，缺少显式的几何约束，整个网络模型更像是在执行记忆功能，更换场景之后，模型的泛化能力也会受到限制。除此之外，基于深度信息标注的数据集，在制作时也会面临传感器精度的问题，标注的深度信息并不一定是准确的，直接对不准确的深度信息进行回归，造成的误差也会相应地增大。

相机在成像的过程中，会丢失掉深度这个三维信息，而且这个过程是不可逆的，深度估计问题从原理上来说就是尝试从二维的图像中恢复出三维的信息，需要两张或两张以上的视图才能完成，所以为了得到更加鲁棒的结果，我们在设计模型时选择从基于立体匹配的方法入手。但是如何用单目视图来完成基于双目的立体匹配呢？很自然地，会想到要去构建基于当前视图的另一张视图,我们选择通过当前非常火的对抗生成网络[2]来实现这一功能。输入单目视图，由生成器构建基于当前视图的左右两张视图（这里我也没有完全想好是否需要生成两张视图，因为理论上生成另一个视图就能够实现数目视图的功能了，但是我觉得如果生成两张视图，对于生成器的训练可能会更加充分。这里存在的一个潜在问题是，在生成左右视图的时候，生成器对于生成左右视图的特征是否会混淆，这个需要通过后期实验的结果再做决定，目前先按照生成两张视图来写），接着用双目数据集去训练判别器鉴别真伪图像。我们希望经过训练之后的对抗生成网络能够构建出逼真的双目视图，这样就能够实现单目视图完成双目立体匹配进而得到深度估计结果的效果。在下面的内容中，将按照双目立体匹配和双目视图构建两部分来分别介绍。

## 相关工作

## 立体匹配

传统的立体匹配需要以下四个步骤[3]：

1. Matching cost computation
2. Cost aggregation
3. Disparity computation / optimization
4. Disparity refinement

对于第一点的计算匹配代价，我们在这里指的是局部的匹配代价。基本原理是给定在图像上的某一点，选取该像素点邻域内的一个子窗口，在另一幅图像中的一个区域中，根据某舟相似性判断依据，寻找与子窗口最为相似的子图，而其匹配的子图中对应的像素点就是该像素点的匹配点。传统的一种做法是直接根据像素点的强度来进行匹配，经典的算法有SSD、SAD等。匹配代价的计算是对于左右两幅图像的每一个像素点而言的，默认这里用到的图像是经过标定之后符合对极约束的图像对。第二点提到代价聚合的应用是为了减小噪声的影响，在计算匹配代价的时候，不是用单一的像素点来计算而是选择使用像素块来进行计算，相当于对匹配代价的计算进行了一个滤波的过程。而视差的计算则是建立在上述基础之上，在一定的视差范围之内，如果找到了最小的代价匹配值，就可以认为这个最小值对应的匹配点就是正确的匹配点，这个最小值对应的视差就是当前目标像素和对应匹配的参考像素之间的视差。

## Full correlation cost volume

传统的立体匹配按照上述的四个步骤来计算，但是这样的方法并不能通过端到端的方式来进行优化。随着深度学习的发展，越来越多的工作在计算匹配代价的时候，不再用像素点的强度来进行，而是选择通过对特征图进行处理来计算匹配程度或是一致性。FlowNet[4]的提出是为了解决光流的问题，同样也需要对视差进行估计。作者提出了两种不同的网络结构，第一种是直接将输入的两张图像联结起来作为一个六通道的输入，经过卷积神经网络之后直接输出预测的视差结果。这样的网络结构并没有显式地计算两张图之间的相关性，所以作者又提出了第二种网络结构，两张图像分别经过相同的网络层进行特征的提取，在后面的阶段，将两个特征进行融合。通过这样的方式，能够相对独立地计算出有意义的特征，然后在更高层进行特征的合并。这样找到两组图像的特征向量之后进行比较的操作类似于标准的匹配操作，作者通过提出corr layer网络层来解决匹配的问题。

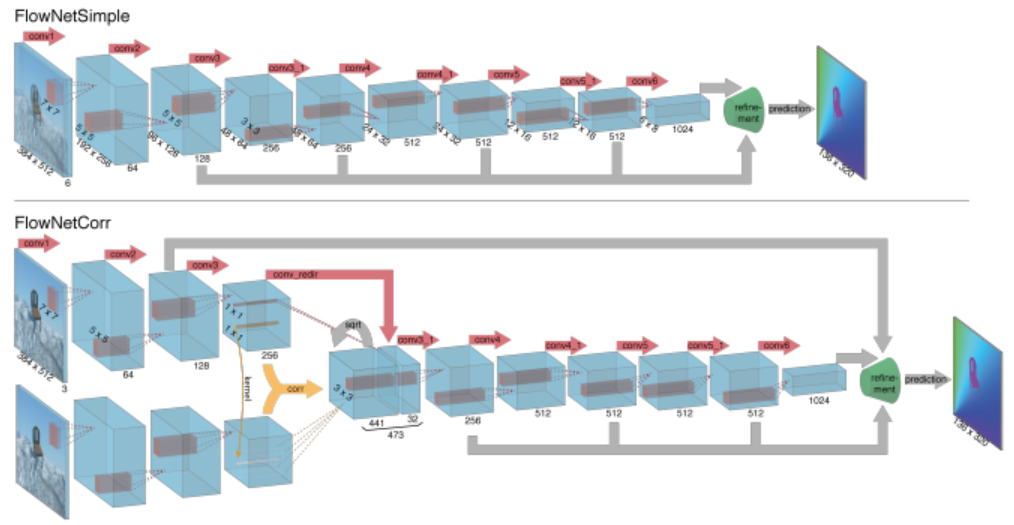


图3.1 FlowNet Architecture

corr layer所执行的计算是，在以x1和x2为中心的两个图像块中，像素点x1和x2之间的一致性计算如下：

|  |  |  |
| --- | --- | --- |
|  |  |  |

经过corr layer的计算之后能够将两张图像特征的相关性结合起来，用于后面进一步的卷积处理，最终得到估计的视差。DispNetC[5]在FlowNet的基础上，在上采样的过程中添加了一些额外的卷积层，获得了更加平滑的视差输出，在模型设计的思想上并没有本质的改变。以FlowNet和DispNetC为代表的方法在使用2D卷积神经网络提取出来左右视图各自的特征之后，直接计算full correlation cost volume，接着直接回归输出预测的视差结果。虽然这样的方法能够得到最终的视差估计，而且也添加了立体匹配的思想有几何的约束，但是这样的处理模式在之后的研究中被认为可能会损失一定的信息，于是就有了下面的方法。

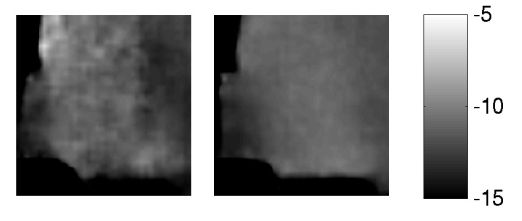


图3.2 DispNetC对视差平滑的处理

## Concatenation-based feature volume

Cost volume这个概念最早应该是GC-Net[6]中提出来的，作者在设计网络结构的时候尝试将对立体几何的认知融入到算法的设计中，使用深度特征的表征来生成cost volume，对应的翻译应该是成本容积，简单的理解就是将各个视差等级上的匹配代价存放起来。

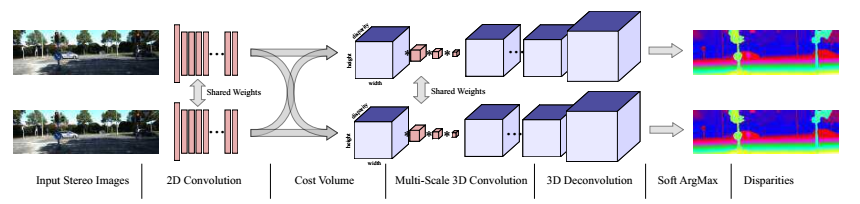


图3.3 GC-Net Architecture

首先由共享参数的2D卷积网络对左右视图进行特征的提取，得到特征之后，对特征图进行级联，也就是concatenation操作。假设经过2D卷积网络层得到的左右视图的特征图尺寸为（1,32,96,312），其中1代表batch的大小，32为channel通道数，96\*312是特征图的分辨率大小，对应的cost volume的尺寸为（1,64,48,96,312），1同样指代的是batch的大小，第二维64的前32层对应的左视图，后32层对应右视图，第三维的48指代视差的等级，即视差的等级从0到47个像素点，最后两个维度依然是特征图分辨率的大小。以第三维度等于10为例，也就是cost volume的第三个维度是10的时候，第二维前32个通道上取x轴上10以后的像素点，后32个通道上去掉x轴上后10个点，没有被选上的像素点的值等于0。得到了四维的cost volume之后，接下来使用三维卷积在cost volume上进一步处理。网络在最后的输出是D\*H\*W尺寸大小的结果，其中D表示全部的视差等级，然后根据各个视差等级上的得分，应用softmax计算出各个视差的可能性，对各个视差进行加权处理，得到最后各个像素点的视差值。之所以使用使用三维的卷积是为了更好地利用语义信息，作者在论文中展示了使用三维卷积在学习语义信息和提高立体匹配结果方面的效果。视差的加权计算公式如下，其中cd代表计算得到各个视差等级上的得分，σ代表的是softmax运算，即对得分进行取负之后用softmax计算得到概率：

|  |  |  |
| --- | --- | --- |
|  |  |  |

损失函数方面，用的就是基本的L1损失函数：

在GC-Net的基础上，为了进一步解决病态区域如重叠区域、低纹理等区域等视差估计精度不高的问题，更好地利用上下文的语义信息，PSMNet[7]引入了空间金字塔池化层，用来融合不同尺度的特征。

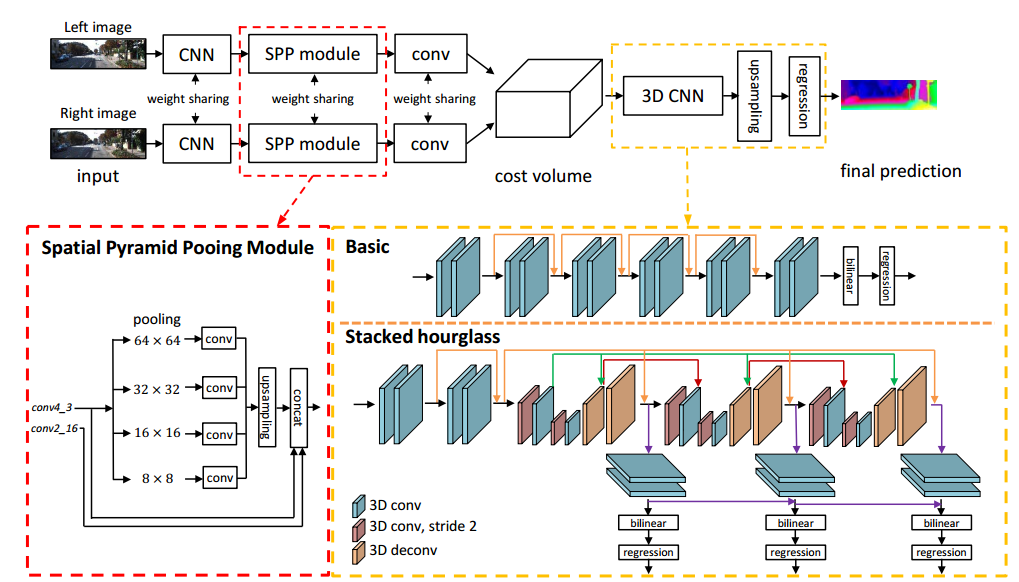


图3.4 PSMNet Architecture

由于当前的方法没有很好地利用上下文的语义信息，所以引入了空间金字塔池化层，SPP模型的引入是为了聚合多尺度信息。除了金字塔池化模型之外，作者还用了空洞卷积的方法来扩大感受野，这些操作的引入都是为了能更好的利用上下文的信息。通过引入上下文的语义信息来改善图像理解问题是一个有效的方法，在解决语义分割问题时，PSPNet[8]便引入了金字塔池化的思想，通过引入不同尺度的特征，有效改善了语义分割的效果。

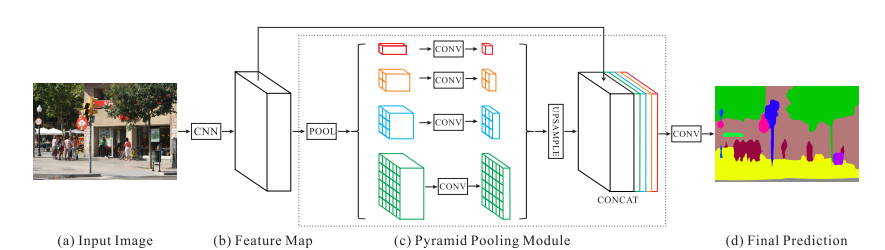


图3.5 PSPNet Architecture

除了引入空间金字塔池化模型之外，PSMNet[7]在网络结构上还进行进行了额外的改进，设计了漏斗式的网络结构，但是复杂的模型带来的代价就是运算成本的增加。

2019年CVPR的另一篇论文GWCNet[9]，参考了上述的两个工作，对cost volume的计算提出了新的方法。因为full correlation cost volume在计算的过程中会损失部分的信息，而concatenation-based cost volume的计算又需要太多的参数，所以作者在计算cost volume的时候，两种cost volume各计算了一部分，然后将二者合并之后得到最终的cost volume，在网络结构上去掉了PSMNet的空间金字塔池化。作者对实验结果进行总结的时候写道，GWCNet和其他的网络模型相比，在同时对模型参数减少的条件下，其性能并不会降低很多，这一点要比PSMNet表现出色。

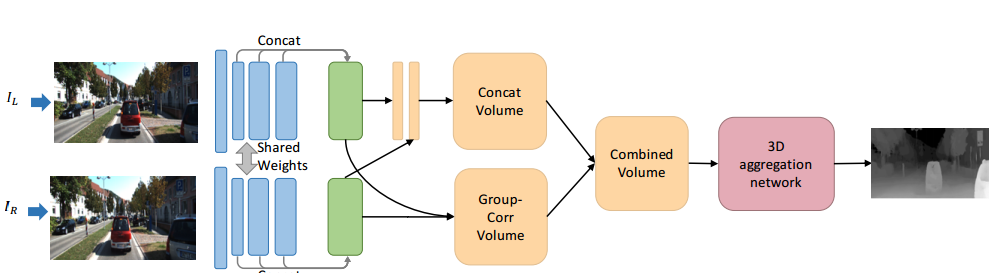
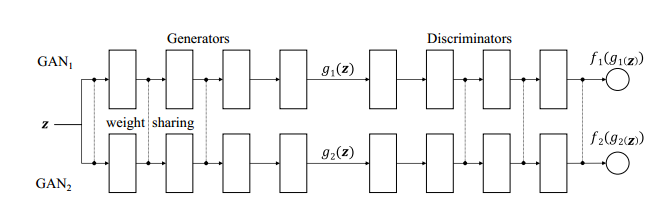


图3.6 GWCNet Architecture

## 对抗生成网络

关于对抗生成网络，这部分的内容因为还没有看很多，先简单写一个对我们的模型有帮助的论文。对抗生成网络有很多的变形，相关的论文有很多，但是绝大部分的gan都是学习单一域的数据分布，CoupledGAN[10]则提出了一种权重共享的网络，使用无监督的方法来学习多个域图像的联合分布。

图3.7 CoupledGAN Architecture

为什么想选择coupled gan作为对抗生成网络的基本模型呢？因为我们希望能够通过对抗生成网络生成左右两张视图而不是单一的某一张视图，所以我们希望能够利用coupled gan能够学习到多个于联合分布的特性来学习左右视图之间的联合特征分布，将生成的两张视图作为立体匹配的输入使用。

## 本章小结

根据上面的内容，我们认为要想解决深度估计的问题还是要从立体匹配入手，因为深度估计的问题从本质上来说就不是单目图像能解决的问题。那么在上面提到的两种不同种计算立体匹配的方法中，concatenation-based cost volume的方法相比full correlation会有更强的几何约束，也符合传统方法中对深度估计的计算流程，所以我们会选择使用计算cost volume然后用3D卷积的方式来进行立体匹配去估计视差。在进行立体匹配之前，还需要至少两张的视图，这部分通过coupled gan来实现。

目前也有一些研究实现了使用gan来直接进行视差的估计，GANVO[11]和SGANVO[12]都是使用了相对基础的gan结构，基本的思想就是由单张图像先生成视差图，然后再由另一个分支卷积网络去生成六自由度的位姿信息，由位姿信息和视差信息一同去合成新的视图，通过训练判别器和生成器来对视差图进行优化。

除此之外，以coupled gan作为对抗生成模型的基本模型配合条件随机场进行单目深度估计。然而所有这些基于gan的方法更多地是通过利用对抗生成网络在生成图像方面的优势来训练得到最终的深度图，也没有引入明显的几何约束。所以我们在尝试用单目相机进行深度估计的时候，一方面需要利用对抗生成网络强大的“造假”能力，另一方面也要将几何约束作为我们网络模型的重要参考内容。

## GSMNet

我们构想的网络结构如图所示：

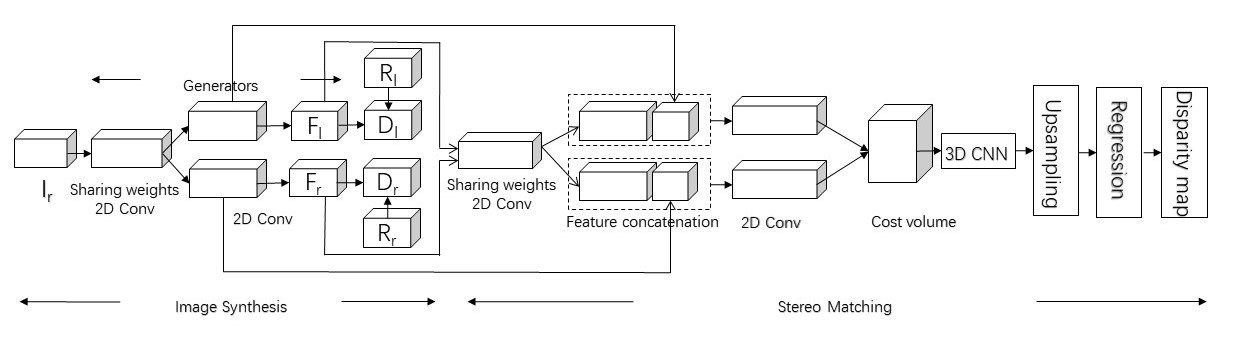


图4.1 GSMNet Architecture

整个网络模型由两部分组成，第一部分由coupled gan结构组成，用来合成左右视图，第二部分是基本的立体匹配计算网络，我们认为在进行图像合成时得到特征可能既包含了左右视图的联合分布信息同时也包含了视差的信息，因为从左视图生成右视图的过程中从理论上来讲会生成一定的视差信息，所以将这部分的特征进行concatenation之后，一同用于生成cost volume。

网络的最终效果要通过实验来进行验证。对于使用对抗生成网络生成器部分的特征是否有帮助现在还不确定，因为用左视图生成右视图可能会有视差信息生成，但是左视图生成左视图应该是没有视差信息的，所以这个信息是否能利用呢？网络结构还有第二种方案就是利用视图合成中共享权重部分得到的特征，因为这部分特征可能包含更多联合分布的特征，对最终是用来生成左视图还是生成右视图不会那么敏感。

# 参考文献

1. H. Laga. A Survey on Deep Learning Architectures for Image-based Depth Reconstruction. arXiv:1906.06113v1 [cs.CV] 14 Jun 2019.
2. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.
3. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision, 47(1-3):7–42, 2002. 1,2.
4. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazırbas¸, ¨ V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In ICCV, 2015. 1, 2, 3, 5, 7, 12.
5. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene  
   flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4040–4048, 2016. 1, 2, 3, 4, 5, 6, 7, 8.
6. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision, pages 66–75, 2017. 1, 2, 4, 5, 8.
7. J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5410–5418, 2018. 1, 2, 3, 4, 5, 6, 8.
8. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. (2017).
9. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise Correlation Stereo Network. In: CVPR.(2019)

1. [Ming-Yu Liu](https://dblp.uni-trier.de/pers/hd/l/Liu_0001:Ming=Yu), [Oncel Tuzel](https://dblp.uni-trier.de/pers/hd/t/Tuzel:Oncel): Coupled Generative Adversarial Networks. [NIPS 2016](https://dblp.uni-trier.de/db/conf/nips/nips2016.html#LiuT16): 469-477.

1. [Yasin Almalioglu](https://dblp.uni-trier.de/pers/hd/a/Almalioglu:Yasin), [Muhamad Risqi U. Saputra](https://dblp.uni-trier.de/pers/hd/s/Saputra:Muhamad_Risqi_U=), [Pedro Porto Buarque de Gusmão](https://dblp.uni-trier.de/pers/hd/g/Gusm=atilde=o:Pedro_Porto_Buarque_de), [Andrew Markham](https://dblp.uni-trier.de/pers/hd/m/Markham:Andrew), [Niki Trigoni](https://dblp.uni-trier.de/pers/hd/t/Trigoni:Niki): GANVO: Unsupervised Deep Monocular Visual Odometry and Depth Estimation with Generative Adversarial Networks. [ICRA 2019](https://dblp.uni-trier.de/db/conf/icra/icra2019.html#AlmaliogluSGMT19): 5474-5480.

1. [Tuo Feng](https://dblp.uni-trier.de/pers/hd/f/Feng:Tuo), [Dongbing Gu](https://dblp.uni-trier.de/pers/hd/g/Gu:Dongbing): SGANVO: Unsupervised Deep Visual Odometry and Depth Estimation with Stacked Generative Adversarial Networks. [CoRR abs/1906.08889](https://dblp.uni-trier.de/db/journals/corr/corr1906.html#abs-1906-08889) (2019)