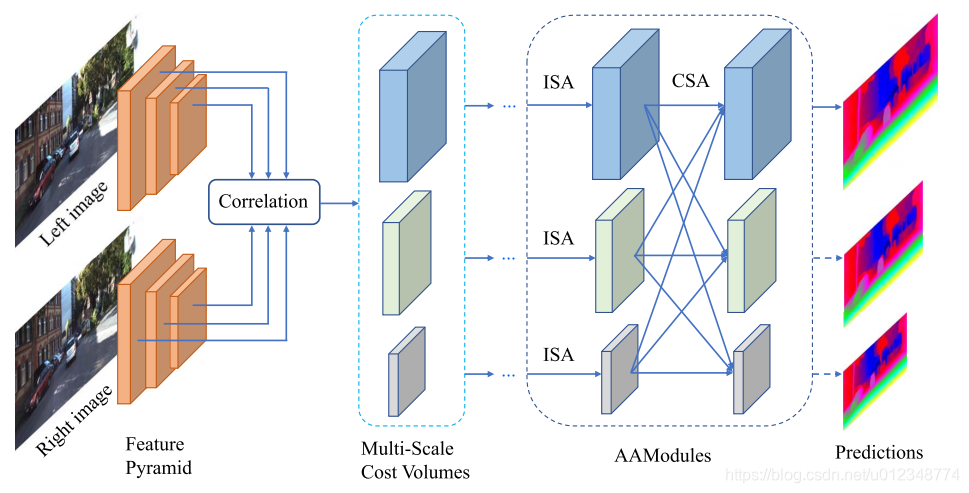
AANet:



input: 设置最大视差为192

left, right 1\*3\*312\*420

backbone out:

f1: 1\*128\*104\*140

f2: 1\*256\*52\*70

f3: 1\*512\*26\*35

主干网络得到3个尺寸,分别是原图的1/3, 1/6, 1/12

FPN out:

fp1: 1\*128\*104\*140

fp2: 1\*128\*52\*70

fp3: 1\*128\*26\*35

分辨率最小的放大后与上一层相加

fp3 = f3, fp2 = fp3\*2 + f2, fp3 = fp2 \* 2 + f1

代价计算out:

fp1: 1\*64\*104\*140

fp2: 1\*32\*52\*70

fp3: 1\*16\*26\*35

代价计算为左右特征图的相关运算结果, 通道维度为对应的视差, 64, 32, 12为原图最大视差缩放1/3, 1/6, 1/12到对应尺寸后的最大视差

代价聚合:

fp1: 1\*64\*104\*140

fp2: 1\*32\*52\*70

fp3: 1\*16\*26\*35

首先先对每个尺寸自身做卷积运算做自聚合, 然后再与其他两个尺寸做交叉聚合,即卷积再缩放再相加

视差计算:

对每个尺寸在通道/视差方向做softmax,然后再加权求和,权重就是softmax的结果,再乘上对应视差值

1\*26\*35

1\*52\*70

1\*104\*140

视差精炼:

用最大分辨率的视差(1\*104\*140)逐渐回归回原图尺寸(1\*312\*420)

如果两轮回归,则140->210->420.

第一轮: 原图先降采样为156\*210, 视差上采样为156\*210,然后进行优化

第二轮:第一轮优化后的视差继续上采样为312\*420, 与原图进一步优化.

STTR:

input: 1\*3\*376\*1241

backbone out: encoder

F0: [2N, 3, H, W]

F1: [2N, C0, H//4, W//4]

F2: [2N, C1, H//8, W//8]

F3: [2N, C2, H//16, W//16]

其中F0为原图, F1, F2为经过残差网络得到, F3为F2经过SPP(spatial pyramid pooling)网络得到.

tokenizer out: decoder

[2N,C,H,W]

从最小分辨率F3开始通过densenet逐步卷积放大再与上一层相加,迭代到F0分辨率.

pos\_encoder out:

[2W/3-1,C] 827\*128

downsample:

将特征图降采样到1/3

transformer:

input

left N, C, H/3, W/3

right N, C, H/3, W/3

pos\_encoder

output

以下的W为上面的W/3

self atten:

intput:

pos encoding [2W-1,HN,C]

[W,2HN,C] 414\*250\*128

output:

[W,2HN,C] 414\*250\*128

v\_out = v \* sofrmax(q\*k),此时q, k, v的输入都相同

cross atten:

input:

feat\_left: left image feature, [W,HN,C]

feat\_right: right image feature, [W,HN,C]

pos: pos encoding, [2W-1,HN,C]

分两步:

1 q为右特征图,k, v为左特征图,进行交叉注意力计算,将计算结果与原始右特征图相加得到新的右特征图

2 q为左特征图,k, v为新的右特征图

output:  
[N, H, 2W-1, 2W-1]1\*125\*414\*414

MSNET3D

INPUT:

LFET :1\*3\*256\*512

right: 1\*3\*256\*512

backbone:

1\*320\*64\*128

volume:

1\*40\*48\*64\*128

dres0:

1\*32\*48\*64\*128

dres1:

cost0: 1\*32\*48\*64\*128

encoder\_decoder1(hourglass3D)

out1: 1\*32\*48\*64\*128

encoder\_decoder2(hourglass3D)

out2: 1\*32\*48\*64\*128

encoder\_decoder3(hourglass3D)

out3: 1\*32\*48\*64\*128

classif0:

in: cost0 out:(cost0)1\*1\*48\*64\*128

classif1:

in: out1 out:(cost0)1\*1\*48\*64\*128

classif2:

in: out2 out:(cost0)1\*1\*48\*64\*128

classif3:

in: out3 out:(cost0)1\*1\*48\*64\*128

训练阶段:

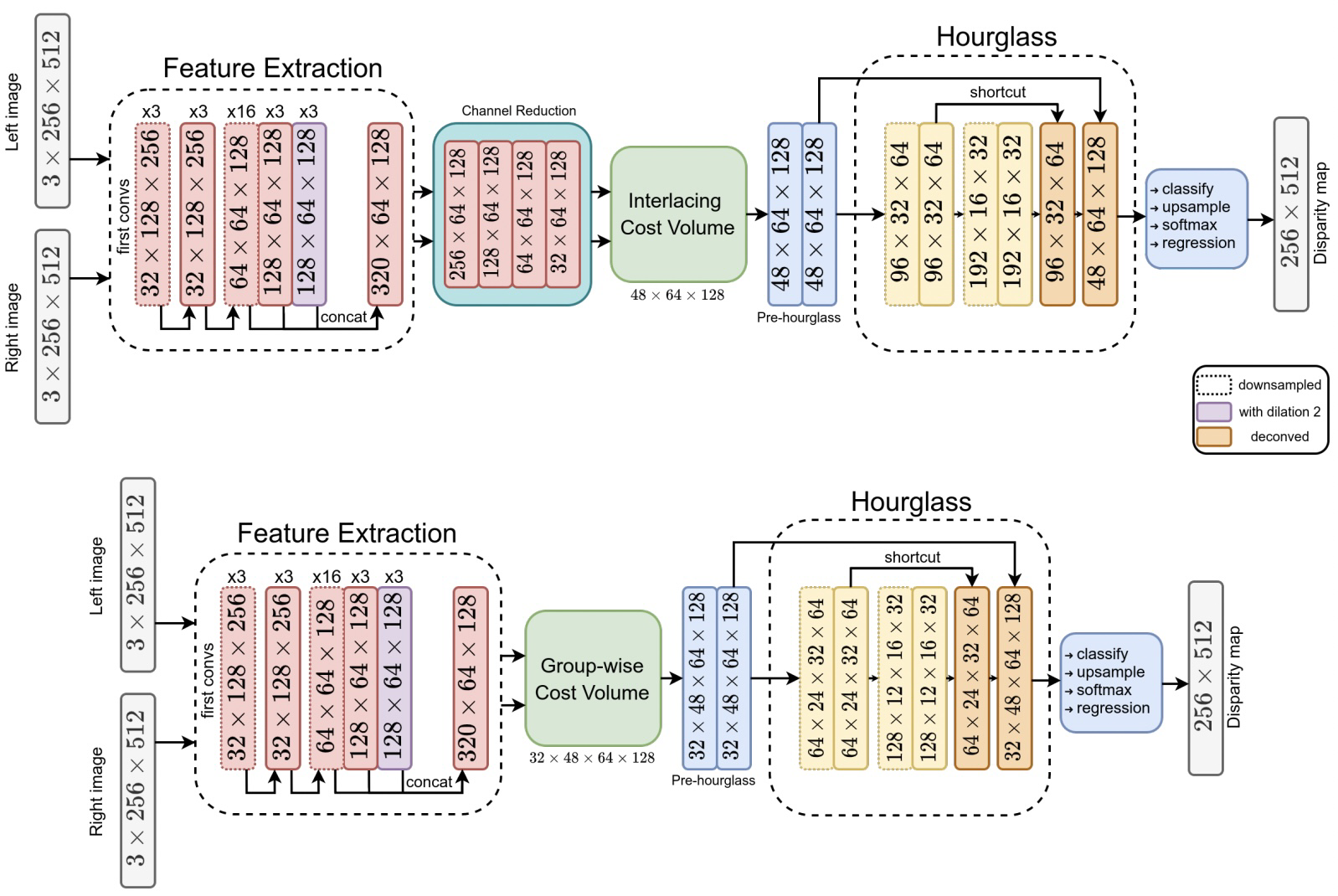
cost0 - cost3都先resize到1\*192\*256\*512然后再回归视差,其中192为设定的最大视差, h\*w为原图分辨率

推理阶段:

只对cost3进行视差回归预测

模型输出:

1\*256\*512



图中,上图为mobile2d,下图为3d

MSNET2D

INPUT:

LFET :1\*3\*256\*512

right: 1\*3\*256\*512

backbone:

1\*320\*64\*128

preconv11:

1\*32\*64\*128