0.1 VIL Net

Based on above implementations, we propose a light and image guided network called Vertex-Image-Light Network (VIL-Net). The structure is shown in Figure 1. As mentioned in the name, the network utilizes vertex map, light map and image map to accomplish the normal inference task.

The network can be consider in two parts. The first part extracts the feature map encoders. It deals with two kinds of input, the vertex map $X_1 \in \mathbb{R}^{w \times h \times 3}$ and the concatenation of light and image map $X_2 \in \mathbb{R}^{w \times h \times 4}$. The network extracts the geometry features X_v from vertex map X_1 (represented as a regression function v)

$$v : X_1 \to X_v$$

meanwhile, the network extracts the photometric features X_l from image and the light map X_2 (represented as a regression function l)

$$l : X_2 \to X_l$$

where the two encoders have the same network architecture based on GCNN model.The second part is basically identify to the original GCNN architecture, but before the start of upsampling part, 2 extra layers are added. First, a concatenate layer is added to fuse the vertex feature and the image and light map feature getting from encoder. Second, a fused feature map is predicted from all the feature maps base on a single gated convolution layer. (represented as a regression function m)

$$m : [X_v X_l] \to X_f$$

Then the network interpolates the feature maps X_f using 3 times upsamling with gated convolution layers. Meanwhile, the skip connections from UNet have been kept. Thus the feature maps $X_{df_1, df_2, \ldots}$ of each down-sampling time in the regression model v are considered in the upsamling part.

$$n : X_f, X_{df_1, df_2, \ldots} \to N$$

With the help of an extra image-light encoder, the network gained more information of the object surface, which is supposed to predict the surface normal more accurate. In this scenario, the output is still the surface normal, thus the training loss can be the same as GCNN model.

0.2 An3 Net

The first branch (shown above) takes a light map introduced in **??** as the input, the structure is the same as GCNN architecture except that the last two standard convolution layers, the skip connections are kept to connect the 3 down/up samplings. The second branch (shown below) takes image as the input, the structure is the same as the first branch other than the input image is 1 channel but not 3 channels. The third branch takes the 3D vertex map as

the input. The structure is based on GCNN architecture. However, in order to merge the other two branches, the vertex branch equips 4 times fusions in the up sampling part. Specifically, the first fusion locates immediately after the last gconv layer of the last down sampling, the second fusion after the second gconv layer of first up sampling, the third fusion after the second gconv layer of second up sampling, the fourth fusion after the second gconv layer of the third up sampling. Each fusion follows by an interpolation layer, a gconv layer to reduce the channel back to 32, a skip connection concatenate layer and another gconv layer to reduce the channel back to 32. After the fourth fusion, a gconv layer used for channel reduction, 2 standard conv layer for output prediction.

0.3 Loss Function

For the case of normal output, the loss function is the same as Mask-L2 loss as introduced in **??**. For the case of the product of albedo and normal, the loss function utilized a scaled Mask-L2 loss, which gives the range of inliers between $[0, 255]$.

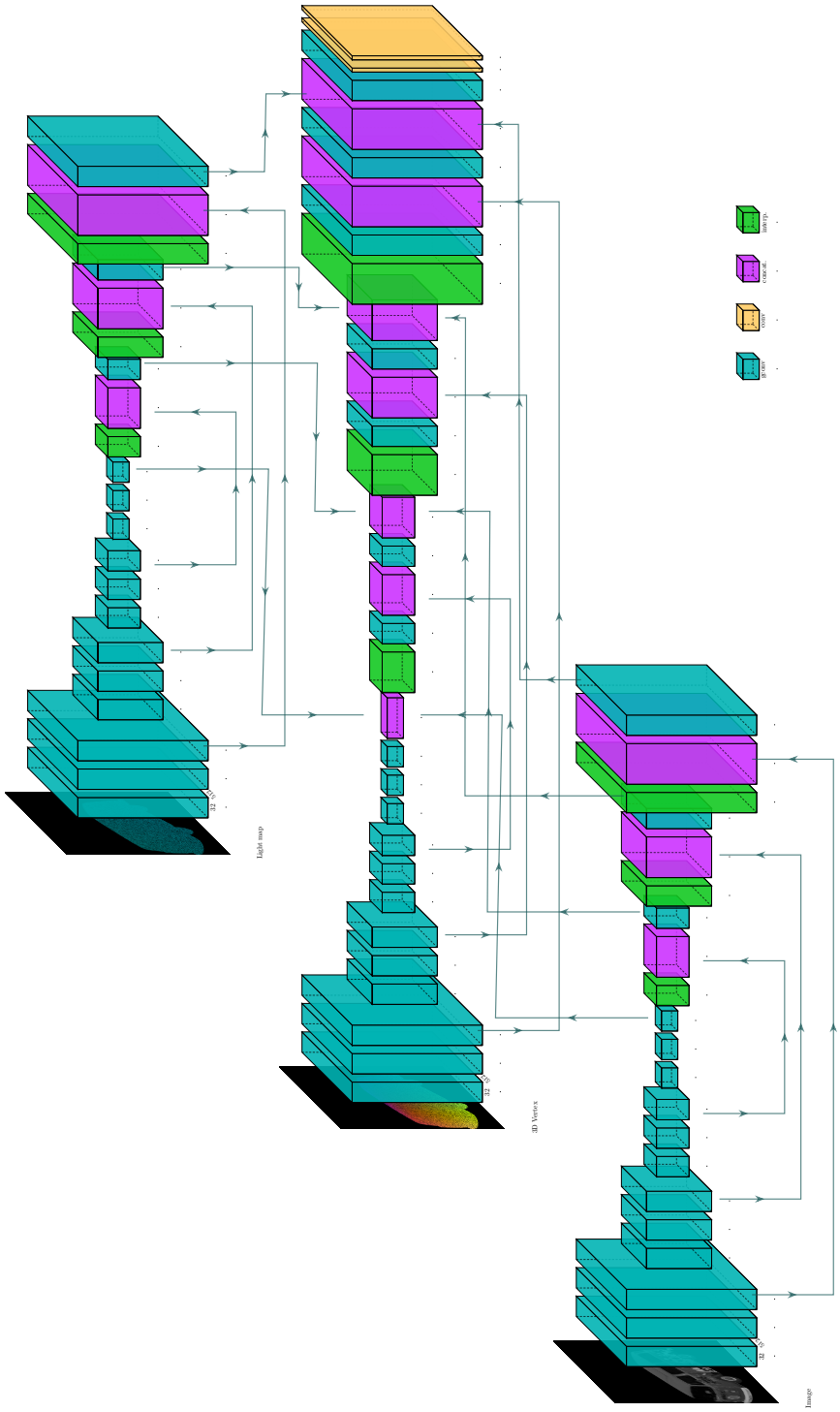

Figure 1: The architecture of TriGNet