DPSNet End-to-end Deep Plane Sweep Stereo

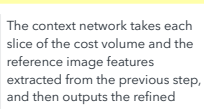
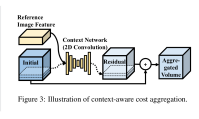
- 1 DPSNET: END-TO-END DEEP PLANE SWEEP STEREO
- 1 INTRODUCTION
  - 1 conventional tech- niques employ photometric consistency constraints on local image patches
    - 1 Such photo-consistency constraints, though effective in many instances, can be unreliable in scenes containing textureless and reflective regions
  - 1 Recently, convolutional neural networks (CNNs) have demonstrated some capacity to address this issue by leveraging semantic information inferred from the scene
    - 1 follow the plane-sweep approach, but require plane-sweep volumes as input to their networks.
      As a result, they are not end-to-end systems that can be trained from input images to disparity maps.
  - 1 In this paper, we present Deep Plane Sweep Network (DPSNet), an end-to-end CNN framework for robust multiview stereo.
    - 1 With the proposed network, plane-sweep stereo can be learned in an end-to-end fashion
    - 1 Additionally, we introduce a cost aggregation module based on
    - 2 local cost-volume filtering Rhemann et al. (2011) for context-aware refinement of each cost slice.
- 3 APPROACH
  - 3 overall framework
    - 4 
  - 3 feature extraction
    - 3 first pass a reference image and target images through seven convolutional layers
    - 3 and extract hierarchical contextual information from these images using a spatial pyramid pooling (SPP) module
    - 3 After upsampling the hierarchical contextual information to the same size as the original feature map, we concatenate all the feature maps and pass them through 2D convolutional layers.
  - 3 cost volume generation
    - 3 We propose to generate cost volumes for the multiview images by adopting traditional plane sweep stereo
      - 4 first set the number of virtual planes perpendicular to the z-axis of the reference viewpoint [0, 0, 1]?and uniformly sample them in the inverse-depth space as
      - 4 $d_l = \frac{(L \times d_{\min})}{l},\ (l = 1, .., L),$
      - 4 warp all the paired features Fi, (i = 1, .., N), where i is an index of viewpoints and N is the total number of input views, into the coordinates of the reference feature
      - 4 concatenating the reference image features and the warped image features for all of the depth labels.
    - 4 our DPSNet learns a cost volume generation of size W × H × L by using a series of 3D convolutions on the concatenated features.
  - 3 cost aggregation
    - 5 The context network takes each slice of the cost volume and the reference image features extracted from the previous step, and then outputs the refined cost slice. We run the same process for all the cost slices.
    - 5 

      Figure 3: Illustration of context-aware cost aggregation.
  - 3 depth map regression
    - 5 The predicted label'l is computed as the sum of each label l weighted by its probability.
    - 5 $\tilde{d} = \frac{L \times d_{\min}}{\tilde{l}}, \quad \tilde{l} = \sum_{l=1}^{L} l \times \sigma(c_l).$
    - 5 TRAINING LOSS
    - 5 $\mathcal{L}(\theta) = \sum_{\mathbf{x}} \lambda |\tilde{d}_{\mathbf{x}}^{\theta} - d_{\mathbf{x}}^{gt}|_{\mathbf{H}} + |\tilde{d}_{\mathbf{x}}^{\theta} - d_{\mathbf{x}}^{gt}|_{\mathbf{H}},$