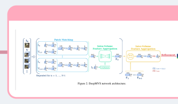
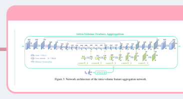
DeepMVS-Learning Multi-view Stereopsis

- DeepMVS: Learning Multi-view Stereopsis
  - Introduction
    - Multi-view Stereo (MVS) methods aim at reconstruct- ing disparity maps from a collection of images with known camera poses and calibration
    - Conventional MVS algorithms often estimate the dispar- ity map by computing plane-sweep volumes and optimizing photometric consistency with handcrafted error functions to measure similarity between patches
      - How- ever, designing algorithms that make explicit use of all these cues is a non-trivial task.
    - More recent work performs stereo reconstruction using end-to-end learning.
      - However, these methods either impose constraints on relative camera poses [17, 19] or the number of input images [5, 37], or produce a coarse volumetric reconstruction
    - In this paper, we present DeepMVS, a deep ConvNet for multi-view stereo that addresses these limitations
      - In summary, we make the following contributions: • We propose DeepMVS, a novel learning-based method for multi-view stereo.
        • Unlike existing work [5, 37, 19], DeepMVS can pro- cess an arbitrary number of input images. The dispar- ity estimation result is invariant to the order in which the inputs are processed.
        • Through extensive evaluation, we show that the in- corporation of semantic features, training on photore- alistic synthetic MVS-SYNTH dataset, and encoder-decoder architecture for aggregating features over large areas all contribute to the improved performance.
  - Learning Multi-view Stereopsis
    - Input
      - The input to our algorithm is a sequence of images and their camera poses and calibration
        - One of the input images is designated as the reference image, for which we seek to obtain a disparity map.
    - Plane-sweep Volume Generation
      - as- sume that the scene geometry is an infinite plane, fronto- parallel to the reference view, and at specific disparities: {0,δ,2δ,...,(D − 1)δ}
        - By doing this with all the neighbor images, we obtain a stack of plane-sweep volumes with N × D images
    - Network Architecture
      - 
      - Patch matching
        - The patch matching network takes a patch from the reference image IRand a single patch Vn,dfrom the plane-sweep volume
      - Intra-volume feature aggregation
        - 
        - add semantic features at each level of the decoder. We pass the reference image into the VGG- 19
      - Inter-volume feature aggregation
        - In this step, we take the N volumes, {F0,...,FN−1}, generated from each of the neighbor images and aggregate them using element-wise max-pooling.
        - The use of max-pooling enables the network to gather information from an arbitrary number of neighbor images, and also ensures that the results are invariant with respect to the order of the neighbor images
    - Training loss.
      - use the cross-entropy loss to train the network.
      - The predicted disparity map can be made by taking the disparity level at which the predicted probability is the highest for each pixel.
        - $\hat{d}_{\text{raw}} = \underset{d}{\operatorname{argmax}}\, y_d.$
    - Refinement
      - we ap- ply the Fully-Connected Conditional Random Field (Dense- CRF) [22] to our raw disparity predictions.
  - Experimental Results
    - Evaluation Metrics
      - Geometric errors.
        - We compute geometric error by tak- ing the L1 distance between the predicted disparity and the ground truth. Unavailable pixels are ignored.
      - Photometric errors
        - the L1 distance between the reference and the rephotography image
      - Completeness
        - e measure completeness using the percent- age of pixels whose errors are below a certain threshold.

DeepMVS-Learning Multi-view Stereopsis

DeepMVS: Learning Multi-view Stereopsis 1

Introduction 1

- Multi-view Stereo (MVS) methods aim at reconstruct- ing disparity maps from a collection of images with known camera poses and calibration 1
- Conventional MVS algorithms often estimate the dispar- ity map by computing plane-sweep volumes and optimizing photometric consistency with handcrafted error functions to measure similarity between patches 1
  - How- ever, designing algorithms that make explicit use of all these cues is a non-trivial task. 1
- More recent work performs stereo reconstruction using end-to-end learning. 2
  - However, these methods either impose constraints on relative camera poses [17, 19] or the number of input images [5, 37], or produce a coarse volumetric reconstruction 2
- In this paper, we present DeepMVS, a deep ConvNet for multi-view stereo that addresses these limitations 2
  - In summary, we make the following contributions: • We propose DeepMVS, a novel learning-based method for multi-view stereo.
    • Unlike existing work [5, 37, 19], DeepMVS can pro- cess an arbitrary number of input images. The dispar- ity estimation result is invariant to the order in which the inputs are processed.
    • Through extensive evaluation, we show that the in- corporation of semantic features, training on photore- alistic synthetic MVS-SYNTH dataset, and encoder- decoder architecture for aggregating features over large areas all contribute to the improved performance. 2

Learning Multi-view Stereopsis 3

- Input 3

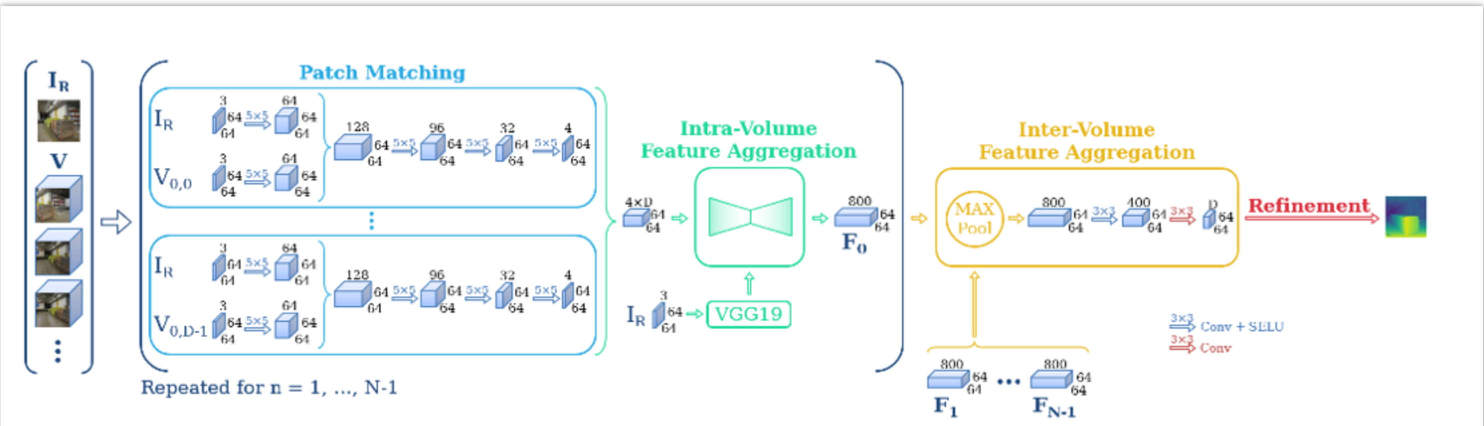
- The input to our algorithm is a sequence of images and their camera poses and calibration
  - One of the input images is designated as the reference image, for which we seek to obtain a disparity map.
- Plane-sweep Volume Generation
  - as- sume that the scene geometry is an infinite plane, fronto- parallel to the reference view, and at specific disparities: {0,δ,2δ,...,(D − 1)δ}
    - By doing this with all the neighbor images, we obtain a stack of plane-sweep volumes with N × D images
- Network Architecture
  - 

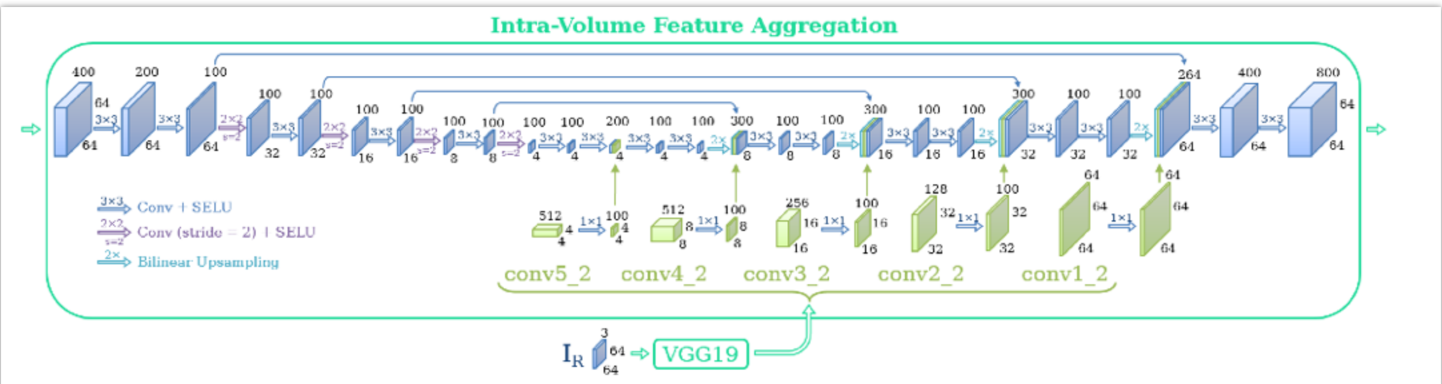
Figure 2. DeepMVS network architecture.
  - Patch matching
    - The patch matching network takes a patch from the reference image IRand a single patch Vn,dfrom the plane-sweep volume
  - Intra-volume feature aggregation
    - 

Figure 3. Network architecture of the intra-volume feature aggregation network.

- add semantic features at each level of the decoder. We pass the reference image into the VGG- 19 (4)

- Inter-volume feature aggregation (4)
  - In this step, we take the N volumes, {F0,...,FN−1}, generated from each of the neighbor images and aggregate them using element-wise max-pooling. (4)
  - The use of max-pooling enables the network to gather information from an arbitrary number of neighbor images, and also ensures that the results are invariant with respect to the order of the neighbor images (4)
- Training loss. (4)
  - use the cross-entropy loss to train the network. (4)
  - The predicted disparity map can be made by taking the disparity level at which the predicted probability is the highest for each pixel. (4)
    - $$\hat{d}_{\text{raw}} = \underset{d}{\operatorname{argmax}}\, y_d.$$ (5)
- Refinement (5)
  - we ap- ply the Fully-Connected Conditional Random Field (Dense- CRF) [22] to our raw disparity predictions. (5)

- Experimental Results (5)
  - Evaluation Metrics (6)
    - Geometric errors. (6)
      - We compute geometric error by tak- ing the L1 distance between the predicted disparity and the ground truth. Unavailable pixels are ignored. (6)

- Photometric errors 6
  - the L1 distance between the reference and the rephotography image 6
- Completeness 6
  - e measure completeness using the percent- age of pixels whose errors are below a certain threshold. 6