Learning Descriptor, Confidence, and Depth Estimation in Multi-view Stereo

Sungil Choi Seungryong Kim Kihong park Kwanghoon Sohn
Yonsei University
khsohn@yonsei.ac.kr

Abstract

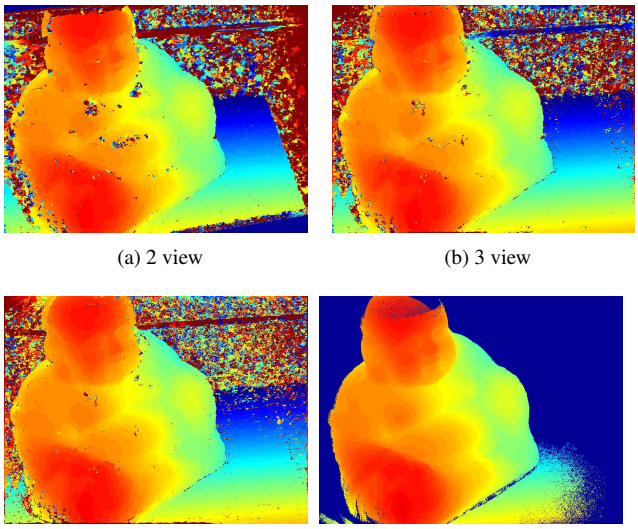
Depth estimation from multi-view stereo images is one of the most fundamental and essential tasks in understanding a scene imaginary. In this paper, we propose a machine learning technique based on deep convolutional neural networks (CNNs) for multi-view stereo matching. The proposed method measures the matching cost to extract depth values between two-view stereo images among multi-view stereo images using a deep architecture. Moreover, we present the confidence estimation network for incorporating the cost volumes along the depth hypothesis in multi-view stereo. Experiments show that our estimated depth map from multiple views shows the better performance than the other matching similarity measure on DTU dataset.

1. Introduction

Perceiving 3-D structure of a scene undoubtedly plays a fundamental role in understanding real-world imagery, and is essential for numerous computer vision and computational photography applications, such as image recognition [20] or reconstruction [7, 22].

To reliably estimate depth information from multi-view images, most of methods have tried to estimate dense correspondences between two-view stereo images or multi-view stereo images [3, 6]. First of all, in two-view stereo matching settings, most approaches compare the patches of given center pixel in the reference image and patches from corresponding pixel according to the disparity hypothesis, and their matching similarity is calculated [3]. To measure the similarities between patch candidates, many methods have been proposed, such as sum-of-squared difference and zero mean normalized cross-correlations [12, 5] that are invariant to radiometric changes and shadows. Nowadays, with the advent of deep convolutional neural networks (CNNs), which has succeeded in numerous computer vision tasks such as object detection, classification and semantic segmentation methods to learn the similarity measure are popularly proposed by leveraging CNNs [26, 18].

Unlike two-view stereo matching settings, depth information also can be estimated from multi-view stereo images [16]. It is advantageous that the occlusions can be effectively dealt with multi-view stereo images taken under various view points, which make the 3D reconstruction results more robust [2]. Figure 1 shows the improvement of matching quality. The number of outliers in depth map are reduced with increasing views in depth estimation. Similar to two-view stereo matching settings, given the depth hypothesis, we can measure the similarity of pixels from reference view and corresponding pixels from the other view by converting the 3D points by triangulation using depth, camera intrinsic and extrinsic parameters [24].



(a) 2 view (b) 3 view (c) 5 view (d) Ground truth

Figure 1. Comparison of depth estimated with different number of views are given: (a) 2 view, (b) 3 view, and (c) 5 view. In (d) ground truth, blue region represents non-depth values and background.

Even though numerous researches have been studied for leveraging the machine learning technique in stereo correspondence, learning a similarity measure for multi-view stereo images has not been studied a lot. As a pioneering work, Hartmann *et al.* [6] directly measures the multi-view similarities using CNNs that takes multi-view image

patches corresponding to the depth hypothesis and outputs the similarity score. However, during training, positive and negative samples are selected manually, and thus it can not compare all patches along depth hypothesis.

In this paper, we propose a novel deep architecture for multi-view stereo matching task, consisting of feature extraction network, confidence estimation network and depth regression network. Our method measures the matching costs between multi-view images by extracting CNN-based features for each image. Furthermore, given multiple view images, we improve the matching quality by estimating the confidence of matching cost in two-view stereo case. Finally, we refine the depth inferred from the cost volumes generated in multi-view stereo utilizing the additional regression network. Experimental results show the state-of-the-art performance of our method on multi-view stereo matching settings.

2. Related Work

Feature descriptors play an important role in the task of matching such as stereo matching[15], flow estimation[6], and dense correspondence[9] in matching task. In order for the reliable matching, feature descriptors are designed to have characteristics which are invariant to illumination and geometric variations. Before the advent of machine learning, hand-crafted feature descriptors are proposed such as SIFT[17] which utilizes histograms of the gradient orientation in the local patches and SURF[1] which reduces the computational complexity by utilizing integral images.

In multi-view stereo matching, hand-crafted descriptors are commonly used. T.Kanade *et al.*[12] and compute the sum of absolute distance (SAD) which is the simplest similarity measure along the epipolar line of the other images. Also, zero-mean-normalized cross-correlation (ZNCC) [5] which can tolerate brightness variations can be used. Recently, with the machine learning techniques, Zbontar *et al.*[26] proposed a convolutional neural networks (CNNs) based approach to compare patches for computing the matching cost in stereo problem and outperformed the conventional hand-crafted descriptors. Also, in multi-view stereo settings, Hartmann *et al.* [6] suggested the CNN-based the multi-view similarities measures using siamese network which takes multi-view image patches and outputs the similarity score along the depth hypothesis.

In order to improve the matching performance in stereo-matching, estimating confidence of the computed matching cost is becoming important issue. Confidence measures can be used for detecting occlusions [8], and improving overall depth map accuracy [19]. Haeusler et al. [27] proposed a random forest classifier utilizing features generated from matching cost volume to learn the confidence. With deep learning techniques, confidence prediction in stereo matching have been recently studied [21].

3. Proposed Method

3.1. Problem formulation and overview

Given multi-view images, the objective of multi-view stereo matching is to estimate depth from the possible depth candidates with the matching costs across multiple images. By leveraging a camera pose between views and intrinsic parameters, N warped image planes are first generated according to derive N depth hypothesis. The depth map is then acquired by choosing the best image plane for each pixel that has the best matching similarity score [3].

To define the matching costs, conventional methods such as sum of absolute differences (SAD) [12] and zero-mean-normalized cross-correlation (ZNCC) [5], are commonly used [28]. However, since they are formulated in a hand-crafted manner, they provide limited performances in measuring reliable patch similarities. To overcome this limitation, we leverage deep convolutional neural networks (CNNs) to extract the robust convolutional features and measure the similarity between multiple patches (Section 2.2). Moreover, to boost the matching quality, we also propose the confidence estimation network (Section 2.3). Finally, we design the depth regression network in order to post-process the depth gained from the cost volume in the multi-view stereo (Section 2.4).

3.2. Feature extraction network

To estimate matching costs for multi-view stereo images, we first present the feature extraction network. To deal with geometric variations across multi-view images, it consists of two major components; the image sampler in spatial transformer networks (STN) [10] and siamese feature extraction networks, whose inputs are warped images by the image sampler. The warping process is implemented by image sampler [10] from the STN. The grid of image sampler is determined by considering their camera extrinsic parameters R, T and intrinsic parameters K as follows:

$$[\hat{x}_n, \hat{y}_n, \hat{w}_n]^T = K(R(K^{-1}\mathbf{p}d_n) + T). \quad (1)$$

where $\mathbf{p} = [x, y, 1]^T$ is a pixel in uniform gird, d_n is the n-*th* depth value, and the location of corresponding n-*th* grid is calculated as $x_n = \hat{x}_n/\hat{w}_n$, $y_n = \hat{y}_n/\hat{w}_n$. For normalized patches, the siamese network is used to extract the feature to distinguish between similar and dissimilar patches of images. Formally, our siamese network consists of 5 consecutive convolutions of 3×3 filter size and ReLu operators. ReLu operator after the last convolution do not exist and the last layer normalizes the activations to have an unit norm for reliable comparison of descriptors. In addition, the network does not have pooling operator or convolution with more than 1 stride in order to make resolution of the output same as input.

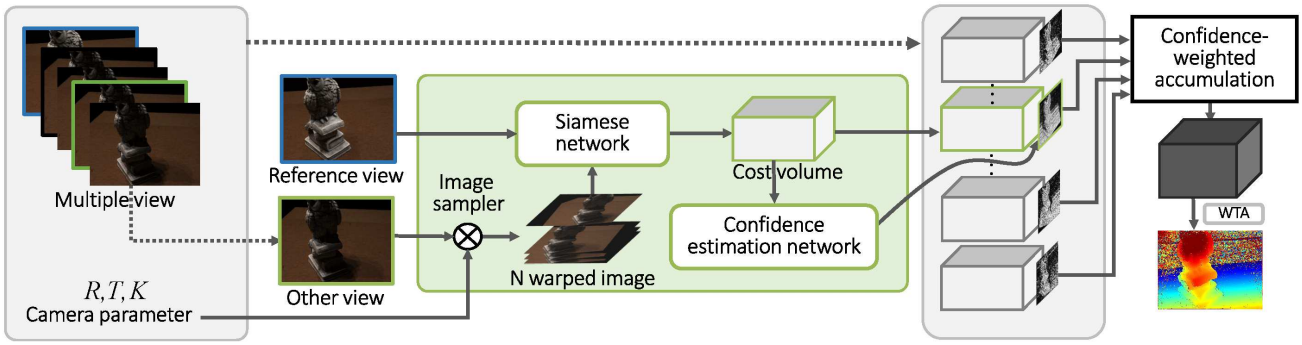


Figure 2. Overall architecture of generating cost volume in multi-view stereo. In order to train descriptors in multi-view stereo, descriptor network is proposed which consists of image sampler and siamese network. Also, for aggregating cost volumes generated by each two-views, we propose confidence estimation network. In multi-view stereo, the final matching cost in multiple view is made by confidence weighted sum of each cost volume.

After passing through the siamese networks with parameters $\mathcal{W}_s$, the output $\mathcal{F}(I_{ref}, I_{oth}; \mathcal{W}_s)$ is $H \times W \times D \times (N+1)$ descriptor volumes generated for N warped images from another view image I_{oth} according to single reference image I_{ref}, where H, W are the image size and D is the size of the descriptor. Here, we denote $\mathbf{F}_n(\mathbf{p}) = [f_n(\mathbf{p}, 1), ..., f_n(\mathbf{p}, D)]$ as the feature representation of pixel $\mathbf{p}$ from n-*th* warped image planes which has D dimensions. Similarly, $\mathbf{F}_{\text{ref}}$ is denoted as the feature representation from reference image. For each pixel $\mathbf{p}$, n-*th* cost volume is then constructed such that

$$\mathcal{C}_n(\mathbf{p}) = \sum_{k=1}^{D} |f_{\text{ref}}(\mathbf{p}, k) - f_n(\mathbf{p}, k)|_1. \qquad (2)$$

Across these cost volumes $\mathbf{C}(\mathbf{p}) = [\mathcal{C}_1(\mathbf{p}), ..., \mathcal{C}_N(\mathbf{p})]$, the depth map can be estimated by selecting the best matching similarity score among N depth hypothesis. The training procedure and the loss are explained in Section 2.4

3.3. Confidence Estimation Network

In traditional two-view stereo settings, there can exist the limitations on estimated depth maps due to occlusion or noise. Given multi-view images, estimating depth information can be divided into multiple two-view stereo settings, and thus occlusion problems can be solved by regarding the other view cost volumes. In other words, we can complement the matching similarity estimated from a two-view stereo within occlusion region by considering the cost volume from another two-view stereo. Thus, in order to enhance the depth estimation quality, multiple images can be explicitly used for selecting the best depth hypothesis. In [6], multiple view images go into the input to the multi-patch similarity network for inferring the best matching depth planes.

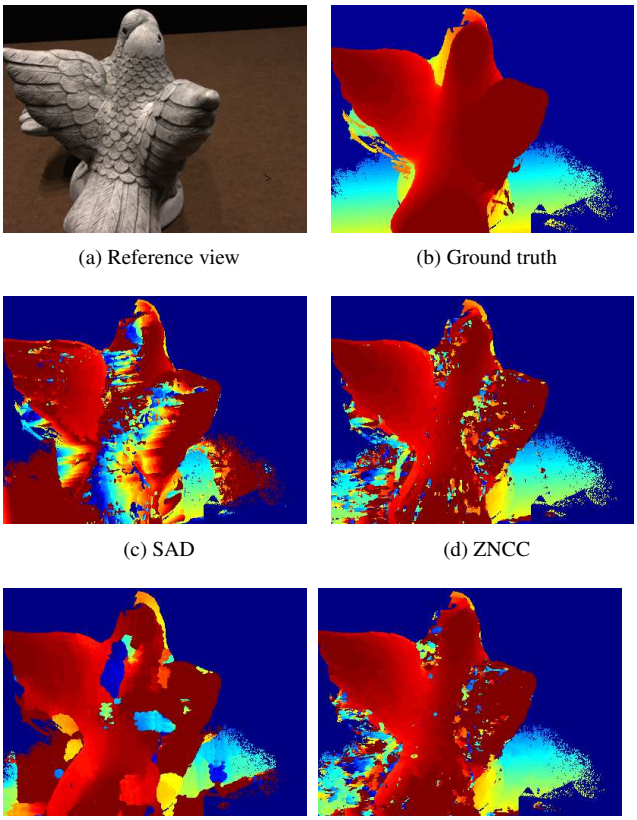
In this paper, we compare the two-view patch similarity. By measuring the confidence of the cost volume constructed by the process described in the previous section, we can aggregate the cost volumes multiplied by the confidence. It should be noted that estimating confidence has been studied especially in stereo matching for a long time in order to interpolate correspondence [14, 23]. We use the confidence map to improve the performance of estimating depth from multiple views by detecting outliers of the cost volume.

The confidence estimation network has 15 consecutive convolutions of 5×5 filter size and ReLu operators for having large receptive fields which enable the network consider large size of cost volume. The last layer of the network is a fully-connected layers which has two nodes and the output is denoted by $\mathcal{F}(\mathbf{C}; \mathcal{W}_c) = [F_1(\mathbf{C}; \mathcal{W}_c), F_2(\mathbf{C}; \mathcal{W}_c)]$ where $\mathcal{W}_c$ is the parameters of confidence estimation network. The estimated confidence c is the softmax normalized of the value of true confidence $F_1(\mathbf{C}; \mathcal{W}_c)$ which results in continuous value between 0 and 1. Given multiple images, we can aggregate the cost volumes generated by each two-view by weighting the confidence c.

3.4. Depth Regression Network

In the plane-sweeping stereo setting, the depth map is induced by choosing the minimum cost of the cost volume constructed by measuring the difference between features from reference image and the corresponding depth image planes along the depth candidates. However, deriving the depth from the cost volume directly limit the quality of depth due to noise of the cost volumes caused by the uncorrected matching in multi-view stereo. Also, the quality of depth estimated by conducting argmin operation can be degraded by the reason that the depth hypothesis are discrete. In order to solve the problem as described above, we suggest the depth regression network to aggregate the cost volume in multi-view stereo.

The depth regression network takes the input of the confidence-weighted sum of multiple cost volumes constructed by two-view stereo. The network architecture is



(a) Reference view (b) Ground truth

(c) SAD (d) ZNCC

(e) SIFT (f) Ours

Figure 3. For comparison of qualitative evaluations of descriptors, results of plane sweeping method in (a) reference view are presented: (c) SAD, (d) ZNCC, and (e) SIFT, and (f) our learned descriptor. In (b) ground truth, blue region represents non-depth values and background.

described in Table 1. From Layer 1 to 5, the cost volume is down-sampled by convolution with stride 2 in order to consider large receptive field of cost volume in the network. After series of up-convolutional and convolutional operations, the network generates refined cost volume, denoted as $\mathbf{C_{refined}}$, in which matching outliers are removed. The depth is estimated by soft-argmin operation proposed by GC-net[13] which enables to output the smooth the depth. The soft-argmin operation is defined as follows:

$$\sum_{d=d_1}^{d_N} d \cdot (e^{-t_i} / \sum\nolimits_j e^{-t_j}), \tag{3}$$

where t_i is i-*th* value of refined cost volume. The refined cost volume $\mathbf{C_{refined}}$ is converted to depth probability distribution normalized by soft-max operation of negative cost value.

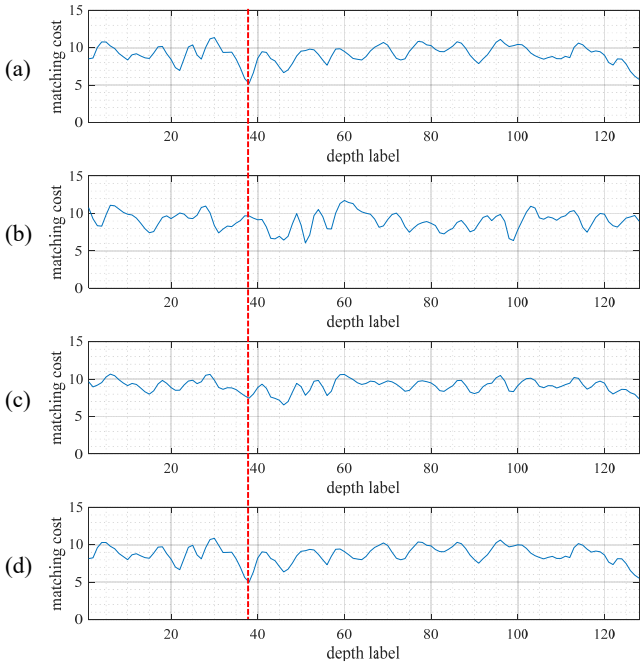


Figure 4. Explanation of effect of confidence weighting. (a) matching cost (correct), (b) matching cost (discorrect), (c) matching cost averaged, and (d) confidence weighted sum of matching cost. The red line shows the correct depth index.

Name	Description	Dimension
Cost Volume		300x400x256
Layer 1	conv 5x5, stride 2, relu	150x200x512
Layer 2	conv 3x3, stride 1, relu	150x200x512
Layer 3	conv 5x5, stride 2, relu	75x100x1024
Layer 4	conv 3x3, stride 1, relu	75x100x1024
Layer 5	conv 3x3, stride 1, relu	75x100x1024
Layer 6	upconv 4x4, stride 2	150x200x512
Layer 7	conv 3x3, stride 1, relu	150x200x512
Layer 8	upconv 4x4, stride 2	300x400x256
Layer 9	conv 3x3, stride 1, relu	300x400x256
Layer 10	conv 3x3, stride 1	300x400x256
Prediction	Soft-argmin	300x400x1

Table 1. Illustration of depth regression network. It consists of series of convolutions (conv), upconvolutions(upconv). All layers have batch normalization operation. The output of the depth regression network is the depth by conducting soft-argmin operation on refined cost-volumes (Layer 10).

3.5. Training

We used the training dataset as the DTU dataset [11] which captures the object in 49 different scenes. In DTU dataset, We sampled 10 3D points clouds for generating training labels. First, we convert the 3D points to the depth map in each image using the camera intrinsic and its corre-

sponding camera parameters in the DTU dataset. We train the descriptor network by end-to-end learning where the input of the network is two input images (reference and another image) and the training label is the 128-level discretized depth map corresponding to reference image. In training the confidence estimation network, the training label for confidence is obtained by comparing the depth from cost volume and the ground truth depth. If the training label is positive, depth value from cost volume is correct compared to the ground truth depth value.

However, we have to distinguish whether each pixel is suitable for training the descriptor network. During training, in order to the descriptor network only compare the training-possible region, we perform the ray consistency check. This consistency check is usually used in training loss [4] and post-processing in stereo method[26]. In ray consistency check in multi-view stereo method, given depth map of the reference image, we can map the the pixel to the 3D point and find the corresponding pixel in the other image with camera poses and intrinsic parameter. In the next step, we check the depth of corresponding pixel and map the pixel to the 3D point. If these two 3D points are very close, we regard the pixel as valid for training. Also, since the depth map is sparse and there do not exist the depth values in non-object regions, we train the descriptor network only in the region where the depth value exists.

We train the descriptor network and confidence estimation network by minimizing the cross-entropy loss using stochastic gradient descent

$$\mathcal{L} = -\sum_i p(l_i) \log \frac{e^{-s_i}}{\sum_j e^{-s_j}}. \tag{4}$$

In training the descriptor network, s_i is $-\mathcal{C}_i$ i.e. matching similarity of *i-th* depth value. In the case of confidence estimation network, s_i is $F_i(\mathbf{C}; \mathcal{W}_c)$ i.e. the output vector elements of confidence estimation network. $p(l_i)$ is delta function, it is 1 when the l_i is the ground truth label.

The depth regression network is trained with supervised learning with the ground truth depth. We train our network by reducing the absolute error between estimated depth, d_n, and ground truth depth ,$\widehat{d}_n$, for all valid pixel n. The loss for depth regression is defined as follows:

$$\mathcal{L} = \sum_n \left\| d_n - \hat{d}_n \right\|_1. \tag{5}$$

We train the descriptor network for $60,000$ iterations with the learning rate 0.01 and a momentum of 0.9 and the confidence estimation network for $10,000$ iterations with the learning rate 0.0001 and a momentum of 0.9. Also, the depth regression network is trained for $30,000$ iterations with the learning rate 0.001 and a momentum of 0.9.

	BUDDHA	BIRD	FLOWER	Avg.
2 view - matching accuracy				
SAD	0.6174	0.5415	0.4591	0.5393
ZNCC	0.6169	0.5642	0.4812	0.5541
SIFT	0.6201	0.5534	0.4641	0.5459
Ours	0.6717	0.5651	0.4751	0.5706
5 view - matching accuracy				
SAD	0.6748	0.6215	0.4926	0.5936
ZNCC	0.6842	0.6156	0.5094	0.6027
SIFT	0.6934	0.6214	0.4975	0.5654
Ours(wo/c)	0.7142	0.6387	0.5012	0.6180
Ours(w/c)	**0.7359**	**0.6794**	**0.5381**	**0.6511**

Table 2. Comparison of quantitative evaluations on the DTU dataset [11]. Matching accuracy of SAD, ZNCC, SIFT and our descriptor is compared in 2-view case. Also, we compare matching quality of the averaged cost volume of SAD, ZNCC, SIFT and ours(wo/c) and confidence weighted sum of cost volume of ours(w/c) given 5 views.

4. Experimental Results

We evaluated the performance of our proposed method compared to other methods in multi-view stereo matching, and analyzed the effect of confidence estimation in estimating depth given multiple views on DTU dataset [11].

Specifically, first, for measuring the matching performance of descriptors, we evaluated our proposed feature representation with SAD [12], ZNCC [5] similarity measure and SIFT[17] descriptor with the plane-sweeping algorithm [3]. It should be noted that we did not apply any post-processing scheme for measuring the matching quality of descriptors. In experiments, we take the reference image and another image from closest view to the reference view. In this experiment on DTU data set, the image size was set to 300×400 and the depth ranges are set to $0.38m$ to $0.9m$. Also, the discrete depth intervals were set to have 128 uniform intervals in the epipolar line. For a fair comparison, since the receptive field size of our descriptor network is 11×11, we set the patch size of SAD and ZNCC to have size of 11×11. We tested three scenes containing different objects: BIRD, BUDDHA, and FLOWER.

The quantitative results are shown in the Table 2. We measured the accuracy of cost volumes generated from all methods by comparing the estimated depth of a pixel is correct or not with respect to the ground truth label, where the accuracy is the ratio of the correct pixel over the whole pixel. Plane sweeping with our learned descriptors achieve the best accuracy in all scenes, especially in the scene containing BUDDHA, BIRD and FLOWER.

Figure 3 shows the plane-sweeping stereo results using SAD, ZNCC, SIFT, and our learned descriptor when only

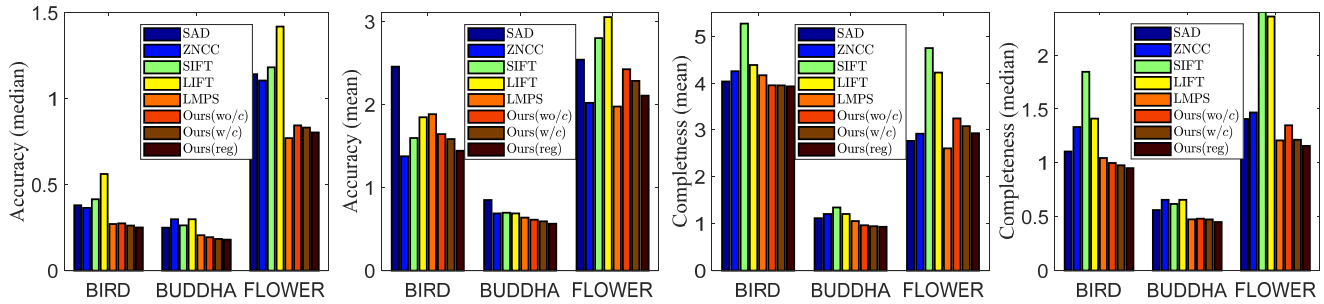
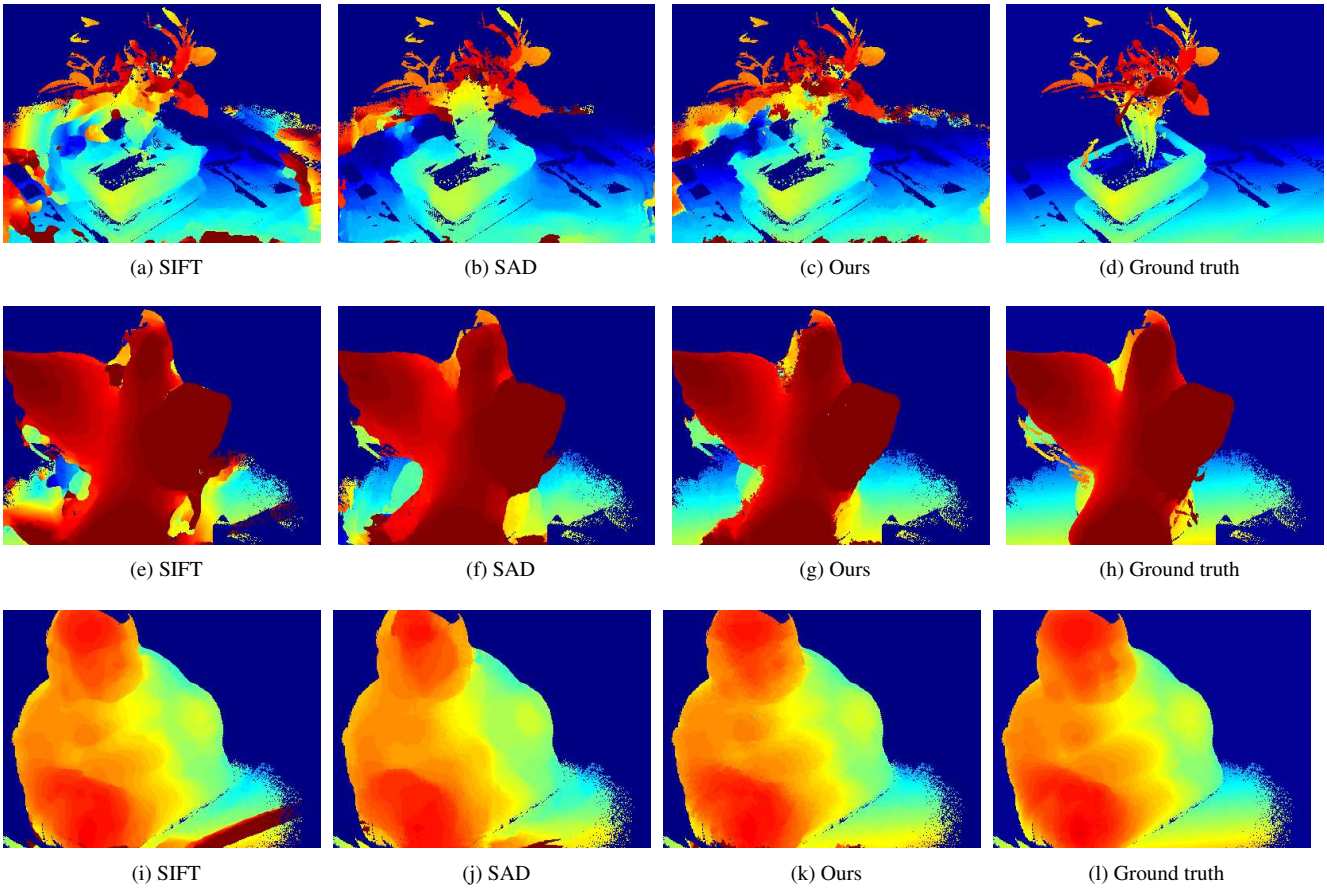


Figure 5. Quantitative results in DTU dataset[11]. Values other than ours are from [6].



(a) SIFT (b) SAD (c) Ours (d) Ground truth

(e) SIFT (f) SAD (g) Ours (h) Ground truth

(i) SIFT (j) SAD (k) Ours (l) Ground truth

Figure 6. For comparison of qualitative evaluations of plane sweeping method using SIFT (1st column), SAD (2nd column), and our depth estimation method (3rd column), and ground truth (4th column). In ground truth, blue region represents non-depth values and background.

two-views are given. We tested matching similarity when the illumination changes are large in order to confirm the robustness of illumination variation for all descriptors. The results of our learned descriptors showed less noise on the wings of BIRD than SAD, ZNCC and SIFT. Our descriptor is more robust to photometric variations because the shadows exist on the right wing.

Furthermore, in order to prove the performance of the depth estimation improvement with confidence, we compared the depth quality of averaged cost volume from different views generated by our descriptor network, and cost volume generated by confidence-weighted sum operation. The quantitative results were shown in Table 2. The depth quality is improved with given multiple views compared to given only two-views. Our averaged cost volume shows more accuracy than other methods since it has shown prior performance only given in two views shown in Table 2. Specifically, the confidence weighted-averaged

cost-volume generated by our method shows better accuracy than averaged cost volume. Figure 4 represents costs a pixel along the depth indices. Figure 4.(c) shows the averaged cost of Figure 4.(a) and Figure 4.(b), and the result shows false matching cost due to the negative effect of false matching cost in Figure 4.(b). However, Figure 4.(d) which are made by confidence weighted sum of matching costs represent correct matching cost since it eliminates the negative effect of Figure 4. (b) by estimating confidence of it.

We output the final depth map with depth regression network. The refined depth are mapped to 3D point clouds by using camera parameters and compared with the evaluation from [6] . The results are shown in Figrue 5. It should be noted that our results and other results in Figure 5 are experimented with different conditions: post processing method. Our method shows more accuracy than other hand-crafted measure and LIFT[25] especially in BUDDHA, and BIRD.

5. Conclusion

In this paper, the technique for measuring the matching similarity in multi-view stereo matching was proposed, including the learning descriptor in two-view stereo matching and the learning confidence of matching of two-view stereo case. With the estimated confidence, the matching similarity was improved by solving the problem existed in two-view stereo matching such as occlusion and matching failure. Furthermore, the depth network enables refining the cost volume and estimating smooth the depth values. Experimental results have demonstrated the effect of learned descriptor, matching improvement using confidence and depth enhancement on multi-view stereo databases.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, 2008.

[2] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. *In Proc. of ECCV*, 2008.

[3] R. T. Collins. A space-sweep approach to true multi-image matching. *In Proc. of CVPR*, 1996.

[4] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *In Proc. of CVPR*, 2017.

[5] M. J. Hannah. Computer matching of areas in stereo images. *PhD thesis, Stanford University*, 1974.

[6] W. Hartmann, S. Galliani, M. Havlena, K. Schindler, and L. Van Gool. Learned multi-patch similarity. *In Proc. of ICCV*, 2017.

[7] C. Hernández, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. *In Proc. of CVPR*, 2007.

[8] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on PAMI*, 30(2):328–341, 2008.

[9] J. Hur, H. Lim, C. Park, and S. Chul Ahn. Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation. 2015.

[10] M. Jaderberg, K. Simonyan, and A. Zisserman. Spatial transformer networks. *In Proc. of NIPS*, 2015.

[11] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. *In Proc. of CVPR*, 2014.

[12] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka. A stereo machine for video-rate dense depth mapping and its new applications. *In Proc. of CVPR*, 1996.

[13] 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. *CoRR, vol. abs/1703.04309*, 2017.

[14] S. Kim, D. Min, B. Ham, and K. Sohn. Deep stereo confidence prediction for depth estimation. *In Proc. of ICIP*, 2017.

[15] D. Kong and H. Tao. A method for learning matching errors for stereo computation.

[16] Y. Liu, X. Cao, Q. Dai, and W. Xu. Continuous depth estimation for multi-view stereo. *In Proc. of CVPR*, 2009.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. *In Proc. of CVPR*, 2016.

[19] P. Mordohai. The self-aware matching measure for stereo. 2009.

[20] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmose, T. Moeslund, and S. Escalera. Real-time human pose recognition in parts from single depth images. *ACM*, 56(1):126–124, 2013.

[21] M. Poggi and M. S. Learning from scratch a confidence measure. 2016.

[22] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, and P. Merrell. Detailed real-time urban 3d reconstruction from video. *IJCV*, 78(2):143–167, 2008.

[23] A. Spyropoulos, N. Komodakis, and P. Mordohai. Learning to detect ground control points for improving the accuracy of stereo matching. *In Proc. of CVPR*, 2014.

[24] A. Torabi and G. Bilodeau. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002.

[25] K. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. *In Proc. of ECCV*, 2016.

[26] J. Zbontar and Y. LeCun. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. JMLR*, 17(1).

[27] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. 2015.

[28] E. Zheng, E. Dunn, V. Jojic, and J. Frahm. Patchmatch based joint view selection and depthmap estimation. *In Proc. of CVPR*, 2014.