Learning Unsupervised Multi-View Stereopsis via Robust Photometric Consistency

Tejas Khot* [1], Shubham Agrawal* [1], Shubham Tulsiani [2], Christoph Mertz [1], Simon Lucey [1], Martial Hebert [1]
[1]Carnegie Mellon University, [2] Facebook AI Research
[1]{tkhot, sagrawal, cmertz, slucey, mhebert}@andrew.cmu.edu, [2]shubhtuls@fb.com

Abstract

We present a learning based approach for multi-view stereopsis (MVS). While current deep MVS methods achieve impressive results, they crucially rely on ground-truth 3D training data, and acquisition of such precise 3D geometry for supervision is a major hurdle. Our framework instead leverages photometric consistency between multiple views as supervisory signal for learning depth prediction in a wide baseline MVS setup. However, naively applying photo consistency constraints is undesirable due to occlusion and lighting changes across views. To overcome this, we propose a robust loss formulation that: a) enforces first order consistency and b) for each point, selectively enforces consistency with some views, thus implicitly handling occlusions. We demonstrate our ability to learn MVS without 3D supervision using a real dataset. We qualitatively observe that our reconstructions are often more complete than the acquired ground truth, showing the merits of this approach. Project webpage: https://tejaskhot.github.io/unsup_mvs.

1. Introduction

Recovering the dense 3D structure of a scene from its images has been a long-standing goal in computer vision. Several approaches over the years have tackled this multi-view stereopsis (MVS) task by leveraging the underlying geometric and photometric constraints – a point in one image projects on to another along the epipolar line, and the correct match is photometrically consistent. While operationalizing this insight has led to remarkable successes, these purely geometry based methods reason about each scene independently, and are unable to implicitly capture and leverage generic priors about the world *e.g.* surfaces tend to be flat, and therefore sometimes perform poorly when signal is sparse *e.g.* textureless surfaces.

We build upon recent learning-based MVS approaches that present CNN architectures with geometric inductive biases, but with salient differences in the form of supervision used to train these CNNs. Instead of relying on ground-truth 3D supervision, which is onerous, we present a framework for learning multi-view stereopsis in an *unsupervised* manner, relying only on a training dataset of multi-view images. Our insight that enables the use of this form of supervision is akin to the one used in classical methods – that the correct geometry would yield photometrically consistent reprojections, and we can therefore train our CNN by minimizing the reprojection error.

Applying the reprojection losses for learning MVS is difficult because different available images may capture different visible aspects of the scene. To circumvent this, we note that while a correct estimate of geometry need not imply photometric consistency with all views, it should imply consistency with at least *some* views. We present a robust reprojection loss that enables us to reason about occlusions implicitly and allow learning MVS with the desired form of supervision.hem. Our model, trained without 3D supervision, takes a collection of images as input and predicts per-image depth maps, which are then combined to obtain a dense 3D model. In summary, our key contributions are:

- An unsupervised learning framework for MVS using only images from novel views as supervisory signal
- A robust multi-view photometric consistency loss for learning unsupervised depth prediction that allows implicitly overcoming lighting changes and occlusion across training views.

2. Related Work

Multi-view Stereo Reconstruction. There is a long and rich history of work on MVS. We only discuss representative works here and refer the interested readers to [11, 2] for excellent surveys. Combining the merits of CNNs and borrowing insights from classical approaches, recent works [13] produce depth images for multiple views and fuse them to obtain a 3D reconstruction. Crucially, all of the above learning based methods have relied on access to 3D supervision and our work relaxes this requirement.

Unsupervised depth estimation. With a similar motivation of reducing the requirement of supervision, several re-

* The first two authors procrastinated equally on this work.

cent monocular [6, 5] or binocular stereo based [15] depth prediction methods have leveraged photometric consistency losses. As supervision signal, these rely on images from stereo pairs or monocular videos during training. As means for visibility reasoning, the network is made to predict an explainability [16], invalidation [14] mask or by incorporating a probabilistic model of observation confidence [9]. These methods operate on a narrow baseline setup with limited visual variations between frames used during training, and therefore do not suffer significantly due to occlusions and lighting changes. As we aim to leverage photometric losses for learning in an MVS setup, we require a robust formulation that can handle these challenges.

3. Approach

The goal in the MVS setup is to reconstruct the dense 3D structure of a scene given a set of input images, where the associated intrinsics and extrinsics for these views are known. We focus here on depth-based MVS setup wherein we infer the per-pixel depth map associated with each input, and the dense 3D scene is then obtained via back-projecting these depth maps into a combined point cloud.

The unsupervised learning framework we propose is agnostic to network architecture. Here, we adopt the model proposed in [13] as a representative network architecture. The network takes as input N images, extracts features using a CNN, creates a plane-sweep based cost volume and infers a depth map for every reference image. A sketch of the architecture is given in Figure 1. The emphasis of our work is on a way to train such CNNs in an unsupervised manner using a robust photometric loss

Robust Photometric Consistency for MVS. The central idea is to use a warping-based view synthesis loss. Given an input image I_s, and additional neighboring views, our CNN outputs a depth map D_s. During training, we also have access to M additional novel views of the same scene $\{I_v^m\}$, and use these to supervise the predicted depth D_s.

For a particular pair of views (I_s, I_v^m) with associated intrinsic/relative extrinsic (K, T) parameters, the predicted depth map D_s allows us to "inverse-warp" the novel view to the source frame to yield $\hat{I}_s^i$.

Alongside the warped image, a binary validity mask V_s^m is also generated, indicating "valid" pixels in the synthesized view as some pixels project outside the image boundaries in the novel view. We can then formulate a photo-consistency objective specifying that the warped image should match the source image. This loss allows us to learn a depth prediction CNN without ground-truth 3D, but there are several issues with this formulation *e.g.* inability to account for occlusion and lighting changes.

Our proposed robust photometric loss formulation is based on two simple observations – image gradients are more invariant to lighting changes than intensities, and that

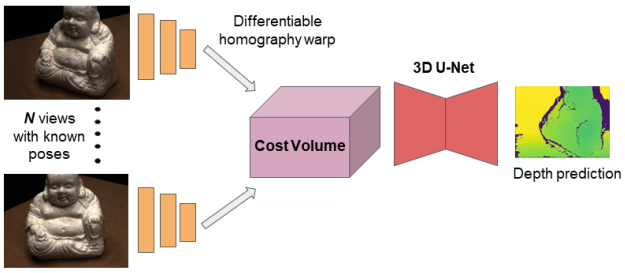

Figure 1: **Overview of our network.** Using differentiable homography, a cost volume is constructed by warping image features over a range of depth values. The cost volume is then refined using a 3D U-Net style CNN. The final output is a depth map at a downsized resolution.

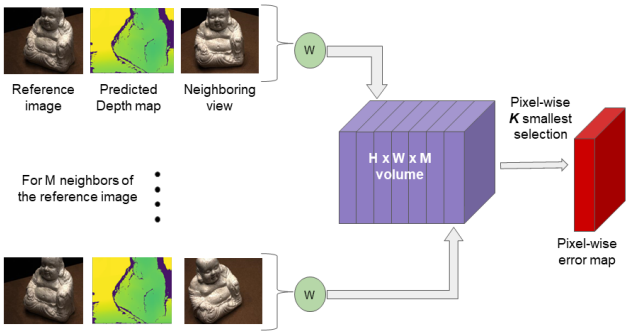

Figure 2: **Robust loss used for training.** The volume of M loss maps is used to perform a pixel-wise selection to pick the K "best" (lowest loss) values using which we take the mean to compute our robust photometric loss.

a point need only be photometrically consistent with some (and not all) novel views.

The inverse-warping based photometric loss is thus:

$$L_{photo} = \sum_{m=1}^{M} ||(I_s - \hat{I}_s) \odot V_s^m||_\epsilon + ||(\nabla I_s - \nabla \hat{I}_s^m) \odot V_s^m|| \tag{1}$$

We refer to this as a first-order consistency loss. We next address the issues raised by occlusion of the 3D structure in the different images. The loss formulations discussed above enforce that each pixel in the source image should be photometrically consistent with *all* other views. Our key insight is to enforce per-pixel photo-consistency with only top-K (out of M) views. Let $L^m(u)$ denote the first-order consistency loss for a particular pixel u w.r.t a novel view I_m^i. Our final robust photometric loss can be formulated as:

$$L_{photo} = \sum_{u} \min_{\substack{m_1,\cdots m_K \\ m_i \neq m_j \\ V_s^{m_k}(u)>0}} \sum_{m_k} L^{m_k}(u) \tag{2}$$

The above equation simply states that for each pixel u,

among the views where the pixel projection is valid, we compute a loss using the best K disjoint views. An illustration of this is shown in Fig 2. To implement this robust photometric loss, we inverse-warp the M novel-view images to the reference image and compute a per-pixel first order consistency "loss-map".

Learning Setup. During training, the input to our depth prediction network comprises of a source image and $N = 2$ additional views. However, we enforce the photometric consistency using a larger set of views ($M = 6, K = 3$). In addition, we add structured similarity (L_{SSIM}) and depth smoothness (L_{Smooth}) objectives. Our final end-to-end unsupervised learning objective is a weighted combination of the losses described previously:

$$L = \sum \alpha L_{photo} + \beta L_{SSIM} + \gamma L_{Smooth} \qquad (3)$$

At test time, we take a set of images of a 3D scene, and predict the depth map of each image through our network. The set of depth images are then fused to form the point cloud. We use Fusibile [4] for the point cloud fusion.

4. Experiments

We follow the same experimental setup as described in [13] and benchmark on the DTU MVS dataset [7]. The results are shown in Table 2 and Figure 3. We evaluate the results with adding first-order consistency (G) with and without top-K aggregation. To the best of our knowledge, we are not aware of any other existing deep-learning based models that learn this task in an unsupervised manner.

We find that for our model, the one with the robust loss, significantly outperforms the variants without it across all metrics. As reported in Table 2, while our model struggles at a high resolution ($< 1mm$), we outperform all other methods (except the fully-supervised MVSNet model) on increasing resolutions. This indicates that while some classical methods are more accurate compared to ours in very low thresholds, our approach produces fewer outliers. The quantitative results from Table 2 and qualitative visualizations of the errors in Figure 3 show that our robust model leads to higher quality reconstructions. We ablate the effect of varying K in our robust loss formulation. As can be seen from Table 1, using $K = 3$ i.e. 50% of the non-reference images has a substantially better validation accuracy. Note that for validation, we use accuracies against the ground truth depth maps. We report percentages of pixels where the absolute difference in depth values is under 3%.

5. Discussion

We presented an unsupervised learning based approach for multi-view stereopsis, and proposed robust photometric losses to learn effectively in this setting. This is however, only an initial attempt, and further efforts are required to realize the potential of unsupervised methods for this task. We are however optimistic, as an unsupervised approach is more scalable as large amounts of training data can be more easily acquired. Lastly, we also hope that the proposed robust photometric loss formulation would be more broadly applicable for unsupervised 3D prediction approaches.

Table 1: Performance comparison as the K in our robust photo-consistency loss varies. Results for using best 25%, 50% and 100% of warping losses per-pixel.

Method (M=6)	K=1	K=3	K=6
Validation Accuracy ($\%$)	75.59	**81.08**	77.99

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 4

[2] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[3] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 4

[4] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 3

[5] Ravi Garg, G VijayKumarB., and Ian D. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 2

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 2

[7] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 3

[8] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: an end-to-end 3d neural network for multiview stereopsis. *arXiv preprint arXiv:1708.01749*, 2017. 4

[9] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: Learning SFM from SFM. In *ECCV (10)*, volume 11214 of *Lecture Notes in Computer Science*, pages 713–728. Springer, 2018. 2

[10] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017. 4

[11] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006*

Table 2: Quantitative results on the *DTU*'s evaluation set [1]. We evaluate two classical MVS methods (top), two learning based MVS methods (bottom) and three unsupervised methods (naive photometric baseline and two variants of our robust formulation) using both the distance metric [1] (lower is better), and the percentage metric [10] (higher is better) with respectively $1mm$, $2mm$ and $3mm$ thresholds.

	Mean Distance (mm)			Percentage ($<1mm$)			Percentage ($<2mm$)			Percentage ($<3mm$)		
	Acc.	Comp.	*overall*	Acc.	Comp.	*f-score*	Acc.	Comp.	*f-score*	Acc.	Comp.	*f-score*
Furu [3]	0.612	0.939	0.775	69.37	57.97	63.16	77.30	64.06	70.06	79.77	66.27	72.40
Tola [12]	0.343	1.190	0.766	88.96	53.88	67.12	92.35	60.01	72.75	93.46	62.29	74.76
Photometric	1.565	1.378	1.472	46.90	42.16	44.40	71.68	55.90	62.82	81.92	60.56	69.64
Ours (Photo + G)	1.069	1.020	1.045	55.98	45.24	50.04	81.11	60.70	69.43	87.03	64.36	74.00
Ours (Photo + G + top-K)	0.881	1.073	0.977	61.54	44.98	51.98	85.15	61.08	71.13	89.47	64.26	74.80
SurfaceNet[8]	0.450	1.043	0.746	75.73	59.09	66.38	79.44	63.87	70.81	80.50	66.54	72.86
MVSNet[13]	0.444	0.741	0.592	82.93	62.71	71.42	88.58	68.70	77.38	89.85	70.11	78.76

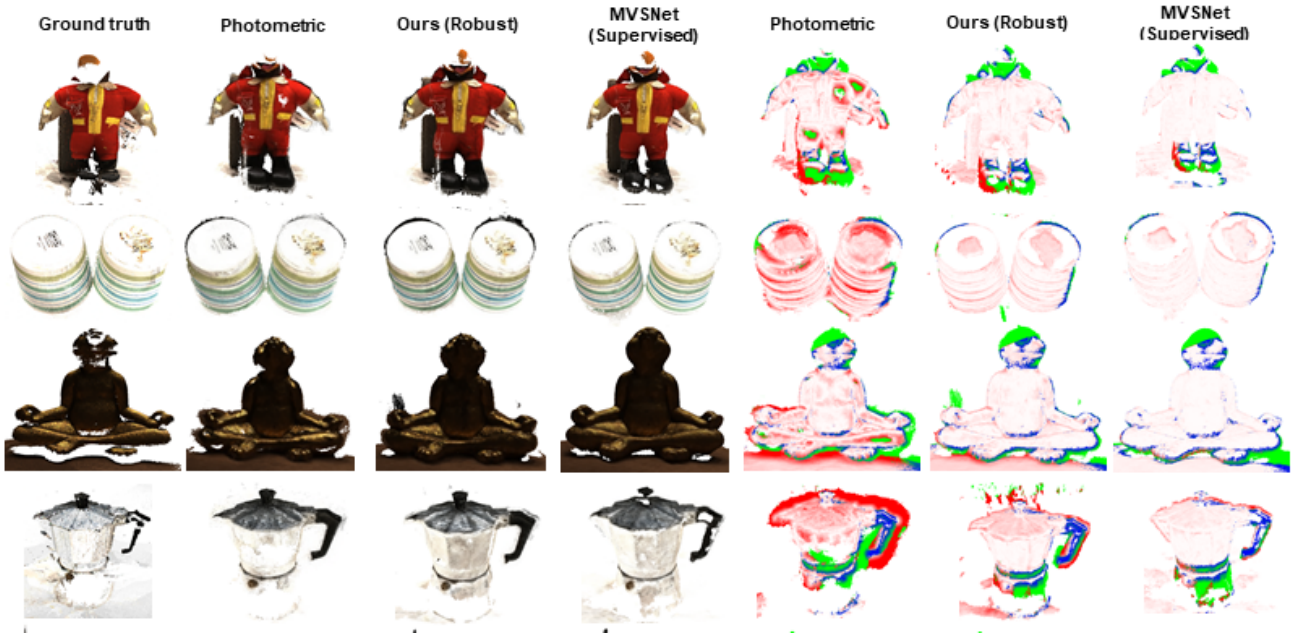

Figure 3: Left to right: a) Ground truth 3D scan, b) result with baseline photo-loss, c) result with robust photo-loss, d) result using a supervised approach (MVSNet [13]), e-g) corresponding error maps. For the error maps, points marked blue/green are masked out in evaluation. Magnitude of error is represented by variation from white-red in increasing order. Note how our method reconstructs areas not captured by the ground truth scan- doors and walls for the building in last row, complete face of statue in third row. Best viewed in color.

IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 1, pages 519–528. IEEE, 2006. 1

[12] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2011. 4

[13] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *CoRR*, abs/1804.02505, 2018. 1, 2, 3, 4

[14] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien P. C. Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas A. Funkhouser, and Sean Ryan Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. *CoRR*, abs/1807.06009, 2018. 2

[15] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *CoRR*, abs/1709.00930, 2017. 2

[16] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619, 2017. 2