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

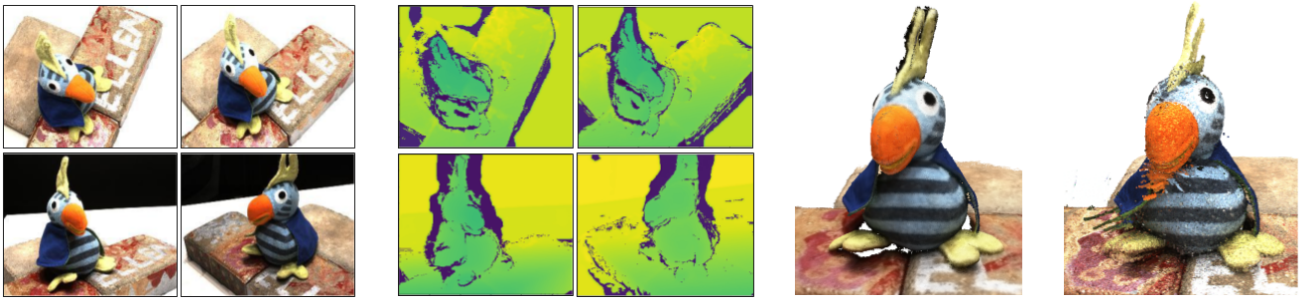

Figure 1: Our model consumes a collection of calibrated images of a scene from multiple views and produces depth maps for every such view. We show that this depth prediction model can be trained in an unsupervised manner using our robust photo consistency loss. The predicted depth maps are then fused together into a consistent 3D reconstruction which closely resembles and often improves upon the sensor scanned model. Left to Right: Input images, predicted depth maps, our fused 3D reconstruction, ground truth 3D scan.

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, and show that each component of our proposed robust loss results in a significant improvement. We qualitatively observe that our reconstructions are often more complete than the acquired ground truth, further showing the merits of this approach. Lastly, our learned model generalizes to novel settings, and our approach allows adaptation of existing CNNs to datasets without ground-truth 3D by unsupervised finetuning.

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.

To overcome these limitations, an emergent line of work has focused on learning based solutions for the MVS task, typically training CNNs to extract and incorporate information across views. While these methods yield impressive performance, they crucially rely on ground-truth 3D data during the learning phase. We argue that this form of supervision is too onerous, is not naturally available, and it is therefore of both practical and scientific interest to pursue solutions that do not rely on such 3D supervision.

We build upon these recent learning-based MVS ap-

proaches 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, 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.

While similar reprojection losses have been successfully used by recent approaches for other tasks *e.g*. monocular depth estimation, we note that naively applying them for learning MVS is not sufficient. This is because different available images may capture different visible aspects of the scene. A particular point (pixel) therefore need not be photometrically consistent with all other views, but rather only those where it is not occluded. Reasoning about occlusion explicitly to recover geometry, however, presents a chicken-and-egg problem, as estimates of occlusion depend on geometry and vice-versa. 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. Further, the lighting changes across views in an MVS setup are also significant, thereby making enforcing consistency only in pixel space undesirable, and our insight is to additionally enforce gradient-based consistency. We present a robust reprojection loss that enables us to capture these two insights, and allow learning MVS with the desired form of supervision. Our simple, intuitive formulation allows handling occlusions without ever explicitly modeling them. Our setup and sample outputs are depicted in Figure 1. 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:

- A framework to learn multi-view stereopsis in an unsupervised manner, 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 [31, 4] for excellent surveys. There are four main stages in an MVS pipeline: view selection, propagation scheme, patch matching and depth map fusion. Schemes for aggregating multiple views for each pixel have been studied in [20, 6, 41, 10, 28, 20], and our formulation can be seen as integrating some of these ideas via a loss function during training. The seminal work of PatchMatch[3] based stereo matching replaced the classical seed-and-expand[5, 10] propagation schemes. PatchMatch has since been used for multi-view stereo[41, 6, 28] in combination with iterative evidence propagation schemes, estimation of depth and normals. Depth map fusion[32, 28, 17, 14, 37] combines individual depth maps into a single point cloud while ensuring the resulting points are consistent among multiple views and incorrect estimates are removed. Depth representations continue to dominate MVS benchmarks [1, 29] and methods seeking depth images as output thus decouple the MVS problem into more tractable pieces.

Learning based MVS. The robustness of features learned using CNNs makes them a natural fit for the third step of MVS: matching image patches. CNN features have been used for stereo matching [12, 38] while simultaneously using metric learning to define the notion of similarity [11]. These approaches require a series of post-processing steps [13] to finally produce pairwise disparity maps. There are relatively fewer works that focus on learning all steps of the MVS pipeline. Volumetric representations encode surface visibility from different views naturally which has been demonstrated in [19, 21]. These methods suffer from the common drawbacks of this choice of representation making it unclear how they can be scaled to more diverse and large-scale scenes. In [22], a cost volume is created using CNN features and disparity values are obtained by regression using a differentiable soft argmin operation. Combining the merits of above methods and borrowing insights from classical approaches, recent works [36, 15, 34] produce depth images for multiple views and fuse them to obtain a 3D reconstruction. Crucially, all of the above 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 recent monocular [9, 7, 26] or binocular stereo based [42] depth prediction methods have leveraged photometric consistency losses. As supervision signal, these rely on images from stereo pairs [9, 7, 26] or monocular videos [35, 43] during training. As means for visibility reasoning, the network is made to predict an explainability [43], invalidation [40] mask or by incorporating a probabilistic model of observation confidence [24]. 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 – these parameters can typically be estimated via a preceding Structure-from-Motion (Sfm) step. While there are several formulations of the MVS problem focused on different 3D representations [21, 4, 5], we focus here on depth-based MVS setup. We therefore 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.

We leverage a learning based system for the step of predicting a depth map, and learn a CNN that takes as input an image with associated neighboring views, and predicts a per-pixel depth map for the central image. Unlike previous learning based MVS methods which also adopt a similar methodology, we only rely on the available multi-view images as supervisory signal, but do not require a ground-truth 3D scene. Towards leveraging this supervision, we build upon insights from classical methods, and note that the accurate geometry prediction for a point (image pixel) should yield photometrically consistent predictions when projected onto other views. We operationalize this insight and use a photometric consistency loss to train our depth prediction CNN, penalizing discrepancy between pixel intensities in original and available novel views. However, we note that the assumption of photometric consistency is not always true. The same point is not necessarily visible across all views. Additionally, lighting changes across views would lead to further discrepancy between pixel intensities. To account for possible lighting changes, we add a first-order consistency term in the photometric loss and therefore also ensure that gradients match in addition to intensities. We then implicitly deal with possible occlusions by proposing a robust photometric loss, which enforces that a point should be consistent with *some*, but not necessarily all views.

We describe the architecture of the CNN used to infer depth in Section 3.1, and present in Section 3.2 the vanilla version of photometric loss that can be used to learn this CNN in an unsupervised manner. We then present our robust photometric loss in Section 3.3, and describe the overall learning setup, additional priors and implementation details in Section 3.4. While we primarily focus on the learning of the depth prediction CNN in this section, we briefly summarize how the learned CNN is integrated in a standard MVS setup at inference in Section 3.5.

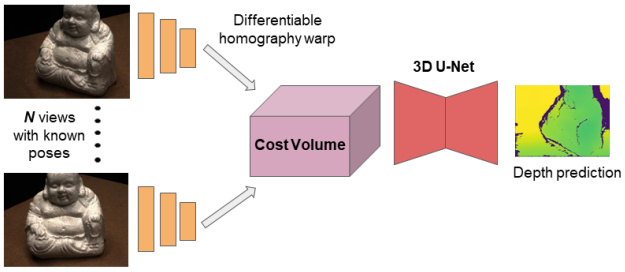

Figure 2: **Overview of our network.** We take as input N images of a scene. Image features are generated using a CNN. Using differentiable homography, a cost volume is constructed by warping image features over a range of depth values. The cost volume is then refined using a 3D U-Net style CNN. The final output is a depth map at a downsized resolution. Details of the network architecture can be found in the supplemental.

3.1. Network architecture

The unsupervised learning framework we propose is agnostic to network architecture. Here, we adopt the model proposed in [36] as a representative network architecture while noting that similar architectures have also been proposed in [15, 34]. 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 2. The emphasis of our work is on a way to train such CNNs in an unsupervised manner using a robust photometric loss, as described in the following sections.

3.2. Learning via Photometric Consistency

We now describe how our depth prediction network can be trained effectively without requiring ground truth depth maps. The central idea is to use a warping-based view synthesis loss, that has been quite effective in the stereo and monocular depth prediction tasks [43, 27] though hasn't been explored for unstructured multi-view scenarios. 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 using a spatial transformer network [16] followed by differentiable bilinear sampling to yield $\hat{I}_s^i$. For a pixel u in the source image I_s, we can obtain its coordinate in the novel view with the warp:

$$\hat{u} = K\, T(D_s(u)\, \cdot K^{-1}\, u) \tag{1}$$

The warped image can then be obtained by bilinear sam-

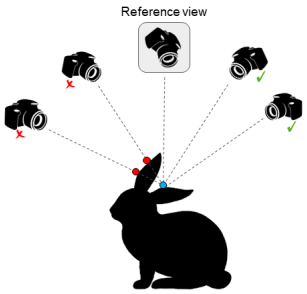

Figure 3: For a set of images of a scene, a given point in a source image may not be visible across all other views.

pling from the novel view image around the warped coordinates:

$$\hat{I}_s^m(u) = I_v^m(\hat{u}) \tag{2}$$

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. As previously done in context of learning monocular depth estimation [43], we can then formulate a photo-consistency objective specifying that the warped image should match the source image. In our scenario of a multi-view system, this can naively be extended to an inverse-warping of all M novel views to the reference view, with the loss being:

$$L_{photo} = \sum_{m}^{M} ||(I_s - \hat{I}_s^m) \odot V_s^m|| \tag{3}$$

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. While similar re-projection losses have been successfully used in datasets like KITTI[8] for monocular or stereo reconstruction, there is minimal disocclusion and lighting change across views in these datasets. However in MVS datasets, self-occlusion, reflection and shadows are a much bigger concern . We therefore extend this photometric loss and propose a more robust formulation appropriate for our setup.

3.3. Robust Photometric Consistency for MVS

Our proposed robust photometric loss formulation is based on two simple observations – image gradients are more invariant to lighting changes than intensities, and that a point need only be photometrically consistent with some (and not all) novel views.

The first modification we make in fact leverages insights developed over many years of MVS research [6], where a number of conventional approaches have found that a matching cost based on both the absolute image intensity

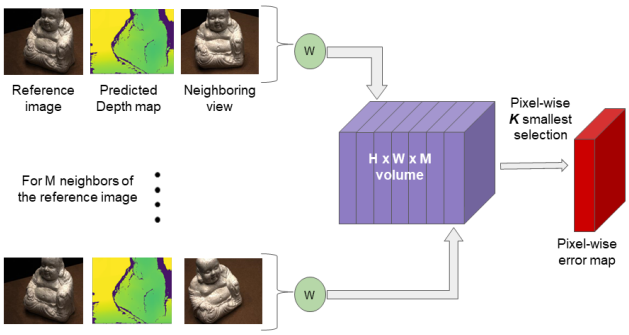

Figure 4: Visualization of the robust pixel-wise aggregation loss used for training. The predicted depth map from the network, along with the reference image are used to warp and calculate a loss map for each of M non-reference neighboring views, as given in eqn 4. These M loss maps are then concatenated into a volume of dimension $H \times W \times M$, where H and W are the image dimensions. This volume is used to perform a pixel-wise selection to pick the K "best" (lowest loss) values, along the 3rd dimension of the volume (i.e. over the M loss maps), using which we take the mean to compute our robust photometric loss.

and the difference of image gradients works much better than just the former. We also found that due to the large variations in pixel intensities between images, it is important to take a huber loss for the absolute image difference term. The inverse-warping based photometric loss of eqn 3 is therefore modified to reflect this:

$$L_{photo} = \sum_{m=1}^{M} ||(I_s - \hat{I}_s) \odot V_s^m||_\epsilon + ||(\nabla I_s - \nabla \hat{I}_s^m) \odot V_s^m|| \tag{4}$$

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. As shown in Fig 3, this is undesirable as a particular point may only be visible in a subset of novel views due to occlusion. 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:

$$L_{photo} = \sum_{u} \min_{\substack{m_1,\cdots m_K \\ m_i \neq m_j \\ V_s^{m_k}(u)>0}} \sum_{m_k} L^{m_k}(u) \tag{5}$$

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 illustra-

tion of this is shown in Fig 4. 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". All M loss-maps are then stacked up into a 3D loss volume of dimensions $W \times H \times M$. For each pixel, we find the K least value entries with valid mask, and sum them to obtain a pixel-level consistency loss.

3.4. Learning Setup and Implementation Details

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 ($M = 6, K = 3$). This allows us to extract supervisory signal from a larger set of images, while only requiring a smaller set at inference.

In addition to the robust photometric losses above, we add structured similarity (L_{SSIM}) and depth smoothness (L_{Smooth}) objectives suggested by [27] for monocular depth prediction task. The smoothness loss enforces an edge-dependent smoothness prior on the predicted disparity maps. The SSIM loss is a higher order reconstruction loss on the warped images, but as it is based on larger image patches, we do not apply our pixel-wise selection approach for the robust photometric loss here. Instead, the two neighboring views with the highest view selection score are used to calculate SSIM loss. We describe the formulation in more detail in the appendix.

Our final end-to-end unsupervised learning objective is a weighted combination of the losses described previously:

$$L = \sum \alpha L_{photo} + \beta L_{SSIM} + \gamma L_{Smooth} \qquad (6)$$

For all our experiments, we use $\alpha = 0.8$, $\beta = 0.2$ and $\gamma = 0.0067$. The network is trained with ADAM[23] optimizer, learning rate of 0.001 and a 1st moment decay factor of 0.95. We use Tensorflow [2] to implement our learning pipeline. As also noted by [36], the high GPU memory requirements of the network imply that it is efficient to use a smaller image resolution and coarser depth steps at training, while a higher setting can be used for evaluation. We note the image resolutions used in the Experiments section.

3.5. Inference using Learned Depth Prediction

At test time, we take a set of images of a 3D scene, and predict the depth map of each image through our network. This is done by passing one reference image and 2 neighboring images through the network, which are chosen on the basis of the camera baselines or a view selection score if available. The set of depth images are then fused to form the point cloud. We use Fusibile [6], an open source utility, for the point cloud fusion.

4. Experiments

We now describe the evaluation of our proposed models. The primary dataset of evaluation is the DTU MVS dataset [18]. In section 4.1 we describe the DTU dataset and our training and evaluation setup, and discuss our results, qualitatively and quantatively. Next, we perform rigorous ablation studies on the effects of various components of the robust loss function we propose (Section 4.2). We also show in Section 4.3 that our method can allow us to adapt pretrained models to datasets without using ground-truth, by finetuning using our robust photometric consistency loss. Lastly, we demonstrate the generalization of our model to another dataset without finetuning (Section 4.4).

4.1. Benchmarking on DTU

The DTU MVS dataset contains scans of 124 different scenes with 3D structure and high-resolution RGB images captured using a robotic arm. For each scene, there are 49 images whose camera poses are known with high accuracy. We use the same train-val-test split as used in SurfaceNet [19] and MVSNet [36].

As in MVSNet, for a given reference image of one scan, its neighboring images for input to the network (N) are selected using a view-selection score [39], which uses the sparse point cloud and camera baselines to pick the most suitable neighboring views for a given reference view. We similarly use neighboring M views during training for self-supervision with the top-K loss.

4.1.1 Training setup

For training, we scale the DTU images to 640x512 resolution. All of our networks are trained with $N = 3$, such that during each iteration, one reference view and 2 novel views are used for predicting a depth map. For our top-K aggregation based robust photometric loss, we use $M = 6$ and $K = 3$. Thus, 6 neighboring views are used to calculate the photometric loss volume, and per pixel the best 3 are selected. We later discuss the effect of varying K.

For evaluation on the test set, depth maps are generated at image resolution 640x512. The d_{min} and d_{max} for the plane sweep volume generation in the network is set to 425mm and 935mm respectively.

4.1.2 Results on DTU

We evaluate our models on the test split of the DTU dataset[18] using the officially prescribed metrics:

- *Accuracy* : The mean distance of the reconstructed points from the ground truth.
- *Completion* : The mean distance of the ground truth points to the reconstruction.
- *Overall* : the mean of accuracy and completion.

Table 1: 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 [25] (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 [5]	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 [33]	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 (Photometric+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	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[19]	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[36]	0.444	0.741	0.592	82.93	62.71	71.42	88.58	68.70	77.38	89.85	70.11	78.76

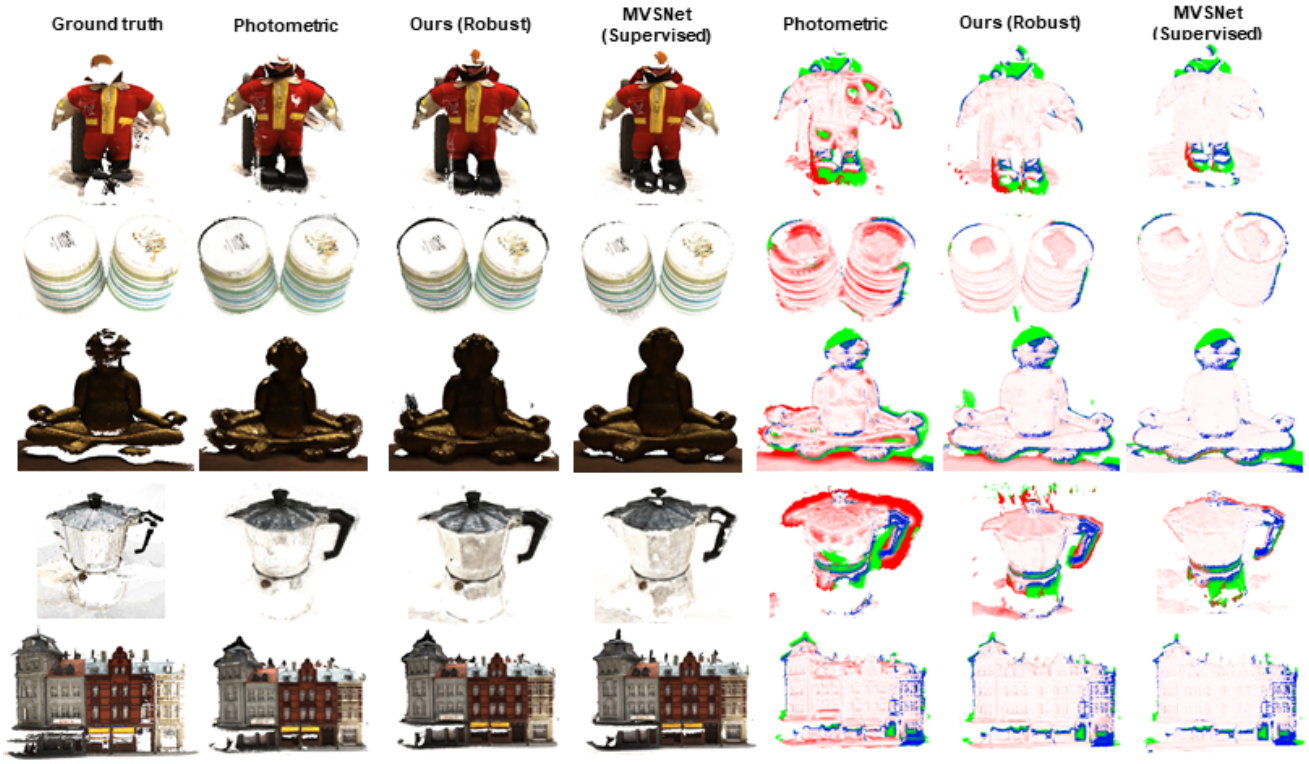

Figure 5: Left to right: a) Ground truth 3D scan, b) result with baseline photo-loss, c) result with robust photo-loss, d) result using a supervised approach (MVSNet [36]), e-g) corresponding error maps. For the error maps, points marked blue/green are masked out in evaluation. Magnitude of error is 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.

Additionally, we report the percentage metric and f-score (which measures the overall accuracy and completeness of the point cloud) as used in the Tanks and Temples benchmark [25].

We quantitatively evaluate three unsupervised models, namely:

- *Photometric*: This model uses a combination of the naive photometric image reconstruction loss as in Equation 3, along with SSIM and Smooth loss.
- *Photometric + first order loss*: We replace the naive photometric loss with our proposed first order gradient consistency loss of Equation 4, which makes the

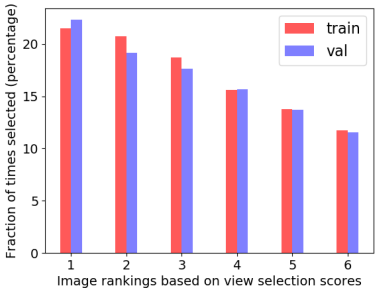

Figure 6: Frequency with which pixels from differently ranked images are picked as valid contributors to the top-K photo-loss. The input images are ranked based on the view selection scores as detailed in Section 4.1.1.

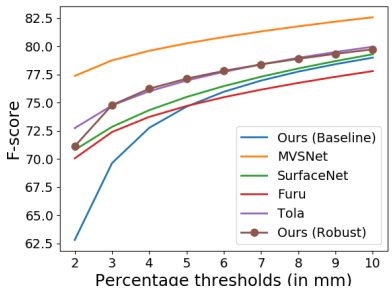

Figure 7: Comparison of different models on the *DTU*'s evaluation set [1] using the F-score metric proposed in [25]. We see that our model trained with robust loss consistently outperforms the baseline and several classical methods.

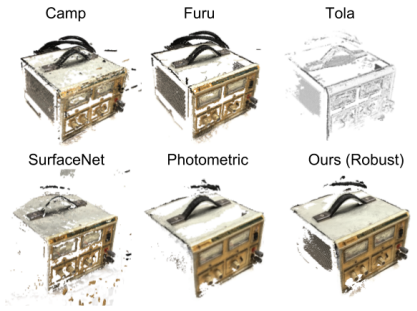

Figure 8: An example of how our proposed technique improves completeness over other methods in low-texture regions. Our result is a smooth dense reconstruction with significantly fewer holes or missing regions.

network much more robust to local lighting variations (denoted as Photometric + G in Table 1).

- *Robust:* Our best model, which combines both the first order gradient consistency loss and the top-K view aggregation scheme.

To place our results in context, in addition to the unsupervised photometric setting, we compare our models against two classical methods (Furukawa et. al. and Tola et al.) [5, 33] , and two more recent deep learning methods *that are fully-supervised*, SurfaceNet [19] and MVSNet [36]. 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. In order to characterize the performance of our model, we further compute the percentage metrics for distance thresholds up to $10mm$ and report the f-score plot in Figure 7. As reported in Table 1, 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 1 and qualitative visualizations of the errors in Figure 5 show that our robust model leads to higher quality reconstructions. Figure 8 shows superior performance of our model in low-texture regions.

4.2. Ablation studies

This section analyzes the influence of several design choices involved in our system, and further highlights the importance of the robust losses in our training setup.

Top-K Selection Frequency. In order to characterize the top-K choice selection, we visualize the frequency with which pixels from different views are selected for photo-consistency. We run the trained model on the training and validation datasets for 50 iterations and store frequency counts of top-K operations which are shown in Figure 6. We can observe two things: 1) A view's selection frequency is directly proportional to its view selection score. This validates that the view-selection criterion used for picking image sets for training corresponds directly to photo-consistency, 2) More than 50% of selections are from views ranked lower than 2 which explains why adding the flexibility of accumulating evidence from additional images leads to better performance.

Top-K Selection Threshold. We ablate the effect of varying K in our robust loss formulation. As can be seen from Table 3, 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%.

Impact of loss terms. We perform ablations to analyze the different components of our robust photometric loss. Although our models are trained in an unsupervised manner, we use the ground truth depth maps of the validation set of DTU to evaluate their performances. In particular, we evaluate the methods on 3 metrics : 1) Absolute difference between predicted and ground truth depths (in mm, lower is better) 2) Percentage of predicted depths within $1mm$ of ground truth (higher is better) 3) Percentage of predicted depths within $3mm$ of ground truth. The detailed quantitative results are provided in Table 2. We observe that both the proposed modifications over the naive baseline yield significant improvements.

Table 2: Ablation study of models with various combinations of loss functions, in terms of validation accuracy against ground truth depth maps of DTU MVS datasets. B signifies the naive baseline photometric loss, as given in eqn. 3. G is the first order gradient consistent loss. Our robust model is a combination of G, SSIM, Smooth and top-K aggregation.

Loss used	L1 error	$\% < 1mm$	$\% < 3mm$
B + SSIM	6.57	30.93	55.07
B + Smooth	5.92	42.52	63.05
B + Smooth + SSIM (our baseline)	4.98	49.37	72.92
G + Smooth + SSIM	5.33	61.92	77.29
G+Smooth+SSIM w/ top k (our robust)	**4.06**	**65.33**	**81.08**

Table 3: 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

4.3. Fine-tuning on ETH3D

We also test the effectiveness of the robust consistency formulation as a means of fine-tuning pretrained models on unseen datasets without available annotations. On the low-res many view dataset of the ETH3D benchmark[30], we compare results of a pretrained MVSNet model with one that is fine-tuned on the train split of ETH3D. For fusion of depth maps, we run Fusibile[6] with identical hyperparameters for each. The results in Table 4 demonstrate that fine-tuning with our proposed loss, in the absence of available 3D ground truth annotations, improves performance.

Table 4: Effect of fine-tuning on the low-res many view ETH3D dataset. All metrics are represented as $(\%)$ and higher is better.

Method	F1 score	Accuracy	Completeness
Pretrained MVSNet	16.91	17.51	19.59
Fine-tuned MVSNet	**17.31**	**18.31**	**19.68**

4.4. Generalization on Tanks and Temples

We perform an experiment to check the generalization ability of our network. Since the network has explicitly been trained to match image correspondences rather than memorize scene priors, our hypothesis is that it should generalize well to completely unseen data. We select the Tanks and Temples dataset for this purpose, which contains high-res images of outdoor scenes of large objects. We use our model trained on DTU on images from this dataset, without any fine-tuning. We downscale the images to 832x512 resolution and use 256 depth intervals for the plane-sweep volume. The results are visualized in Figure 9. More extensive results are provided in the supplemental. However, we do note that the very high depth range of scenes in open-world datasets like Tanks and Temples are not amenable to the current deep architectures for MVS, as they all rely on some sort of volume formulation. Thus to sample depths at a finer resolution for higher quality reconstructions becomes extremely computationally expensive, and is perhaps a promising direction for future work.

Figure 9: Generalization result of our robust model on the Tanks and Temples[25] dataset. Without any finetuning, our robust model provides reasonable reconstructions.

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. In addition, as our experiments demonstrated, these unsupervised methods can be used in conjunction with, and further allow us to improve over supervised methods, thereby allowing us to leverage both, the benefits of supervision along with the scalability of unsupervised methods. Lastly, we also hope that the proposed robust photometric loss formulation would be more broadly applicable for unsupervised 3D prediction approaches.

References

[1] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 2, 6, 7

[2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: a system for large-scale machine learning. 5

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24, 2009. 2

[4] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2, 3

[5] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010. 2, 3, 6, 7

[6] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 2, 4, 5, 8

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

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 4

[9] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 2

[10] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. 2007. 2

[11] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3279–3286, 2015. 2

[12] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler. Learned multi-patch similarity. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1595–1603. IEEE, 2017. 2

[13] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008. 2

[14] X. Hu and P. Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 531–538. IEEE, 2012. 2

[15] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 2, 3

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. 3

[17] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3121–3128. IEEE, 2011. 2

[18] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 5

[19] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. Surfacenet: an end-to-end 3d neural network for multiview stereopsis. *arXiv preprint arXiv:1708.01749*, 2017. 2, 5, 6, 7

[20] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001. 2

[21] A. Kar, C. Häne, and J. Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017. 2, 3

[22] 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. 2

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[24] M. Klodt and A. 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

[25] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017. 6, 7, 8

[26] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2215–2223, 2017. 2

[27] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2011. 3, 5

[28] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2

[29] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017, 2017. 2

[30] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[31] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *null*, pages 519–528. IEEE, 2006. 2

[32] S. Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE transactions on image processing*, 22(5):1901–1914, 2013. 2

[33] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2011. 6, 7

[34] K. Wang and S. Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018. 2, 3

[35] J. Xie, R. B. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, 2016. 2

[36] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *CoRR*, abs/1804.02505, 2018. 2, 3, 5, 6, 7

[37] C. Zach. Fast and high quality fusion of depth maps. In *Proceedings of the international symposium on 3D data processing, visualization and transmission (3DPVT)*, volume 1. Citeseer, 2008. 2

[38] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016. 2

[39] R. Zhang, S. Li, T. Fang, S. Zhu, and L. Quan. Joint camera clustering and surface segmentation for large-scale multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2084–2092, 2015. 5

[40] Y. Zhang, S. Khamis, C. Rhemann, J. P. C. Valentin, A. Kowdle, V. Tankovich, M. Schoenberg, S. Izadi, T. A. Funkhouser, and S. R. Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. *CoRR*, abs/1807.06009, 2018. 2

[41] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 2

[42] Y. Zhong, Y. Dai, and H. Li. Self-supervised learning for stereo matching with self-improving ability. *CoRR*, abs/1709.00930, 2017. 2

[43] T. Zhou, M. Brown, N. Snavely, and D. 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, 3, 4

Supplementary

A. Overview

In this document we provide additional quantitative results, technical details and more qualitative examples of the results on both DTU[4] and Tanks and Temples dataset[6].

In Section B, we present details of the model architecture used for all our experiments. Section C describes the mathematical formulation of the various loss terms used. In Section D we show some elaborate quantitative results on the DTU dataset and in Section E we visualize the qualitative output results for the remaining instances in the DTU dataset. Higher resolution qualitative results on the Tanks and Temples dataset are included in Section F.

B. Model Architecture

We adapt the same network architecture as MVSNet[9] to emphasize that our key contribution is the loss objective used for training a standard network with sufficient capacity for this task.

Every input image is first passed through a feature extraction network having shared weights for all images. For this network, we use use an 8-layer CNN having batch-normalization and ReLU after every convolution operation till the penultimate layer. The last layer produces a 32 channel downsized feature map for every image. Using the differentiable homography formulation, the feature maps are warped into different fronto-parallel planes of the reference camera at 128 depth values to form one cost volume per non-reference image. All such cost volumes are aggregated into a single volume using a variance-based cost metric. Note that MVSNet uses 256 depth values for the cost volume during training. Since this setup does not fit in our 12GB GPU memory, we use only 128 depth values. This reduction does play a role in the output reconstruction quality, but we leave this optimization for future work since our contributions hold nonetheless.

In order to refine the cost volume and incorporate smooth variations of the depth values, we use a three layer 3D U-Net. An initial estimate of the predicted depth map can be obtained by performing a soft $argmin$ operation along the depth channel. Unlike the $winner - take - all$ approach which requires the non-differentiable $argmax$ operation, such a soft aggregation of volumes allows for sub-pixel accuracies while being amenable to training due to its differentiability. Thus, in spite of the discretization of depth value for constructing the cost volume, the resulting depth map follows a continuous distribution.

The resulting probability distribution which the output volume represents is likely to be containing outliers and would not necessarily contain a single peak. To account for this, a notion of depth estimate quality is established wherein the quality of estimate at any pixel is defined to be the sum of the probabilities over the four nearest depth hypotheses. This estimate is then filtered at a threshold of 0.8 and applied as a mask to the output volume. The predicted depth map is then concatenated to the reference image and passed through a four-layer CNN to output a depth residual map. The final depth map is obtained by adding the residual map to the initial estimated depth map. For complete details of the hyperparameters, we refer the reader to the MVSNet[9] paper and it's corresponding supplemental.

C. Loss functions

While minimizing the photometric consistency loss obtained by view synthesis is our primary objective, we make use of two additional ingredients in the loss objective to improve model performance. We augment our robust photometric loss with two additional losses, namely image-patch level structured similarity loss and an image-aware smoothness loss on the depth map's gradients.

SSIM: We take cues from recent works[10, 3, 7] showing the effectiveness of perceptual losses for evaluating the quality of image predictions. Similarly, we also use the structured similarity (SSIM) as a loss term for training. The SSIM similarity betwen two image patches is given by :

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)} \tag{1}$$

Here, μ and σ are the local mean and variance respectively. We compute μ_x, μ_y, σ_x, σ_y using average pooling and set $c_1 = 0.01^2$ and $c_2 = 0.03^2$. Since higher values of SSIM are desirable, we minimize its distance to 1 which is the highest attainable similarity value. The SSIM loss for an image pair then becomes :

$$L_{SSIM} = \sum_{ij} \left[1 - SSIM(I_s^{ij}, \hat{I_s^{ij}})\right] M_s^{ij} \tag{2}$$

Here, M_s^{ij} is a mask which excludes all pixels whose projections after inverse warping lie outside the source image. As observed in [7], ignoring such regions improves depth predictions around the boundaries. We apply the SSIM loss only between the reference image and two nearest images ranked by view selection score.

Depth Smoothness loss: In order to encourage smoother gradient changes and allow sharp depth discontinues at pixels corresponding to sharp changes in the image, it is important to regularize the depth estimates. Similar to [7], we add an l_1 penalty on the depth gradients.

$$L_{Smooth} = \sum_{ij} ||\nabla_x D^{ij}||e^{-||\nabla_x I^{ij}||} + ||\nabla_y D^{ij}||e^{-||\nabla_y I^{ij}||} \tag{3}$$

D. Quantitative Results

The DTU dataset's evaluation script measures the reconstruction quality in terms of accuracy and completeness while also reporting their median values and variances. We list these results for two classical (top), two supervised learning based (bottom), and three unsupervised learning (middle) methods in Table 1. As noted by [9], SurfaceNet[5] used their own script for evaluation. However, we use the released DTU evaluation benchmark scheme for reporting results from all the methods.

Additionally, we show the comparison of two components of the percentage metric, precision and recall, that make up the f-score reported in the main paper, in Figure 1.

E. Qualitative Results

Qualitative results for the remaining 17 instances of the DTU test set are presented in Figures 2, 3.

F. Tanks and Temples

We show qualitative results of our robust models on scenes from the Intermediate set of the Tanks and Temples dataset in Figure: 4. The model has not been finetuned on the dataset but produces reasonable reconstructions demonstrating the learned photo-consistency behavior.

References

[1] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 3

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

[3] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017. 1

[4] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 1

[5] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. Surfacenet: an end-to-end 3d neural network for multiview stereopsis. *arXiv preprint arXiv:1708.01749*, 2017. 2, 3

[6] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017. 1, 3

[7] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, 2011. 1

[8] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2011. 3

[9] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *CoRR*, abs/1804.02505, 2018. 1, 2, 3, 4, 5

[10] T. Zhou, M. Brown, N. Snavely, and D. 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. 1

Table 1: Quantitative results on the *DTU*'s evaluation set [1]. We evaluate two classical MVS methods (top), two learning based MVS methods (middle) and three variants of our unsupervised method using the distance metrics. For all columns, lower is better.

	Accuracy			Completeness			Overall
	Mean	Median	Variance	Mean	Median	Variance	
Furu [2]	0.612	0.324	1.249	0.939	0.463	3.392	0.775
Tola [8]	0.343	0.210	0.385	1.190	0.492	5.319	0.766
Ours (Baseline)	1.565	1.041	3.683	1.378	0.694	4.964	1.472
Ours (Baseline+G)	1.069	0.759	1.883	1.020	0.595	2.779	1.045
Ours (Robust)	0.881	0.673	1.075	1.073	0.617	3.418	0.977
SurfaceNet[5]	0.450	0.254	1.270	1.043	0.285	5.594	0.746
MVSNet[9]	0.444	0.307	0.436	0.741	0.399	2.501	0.592

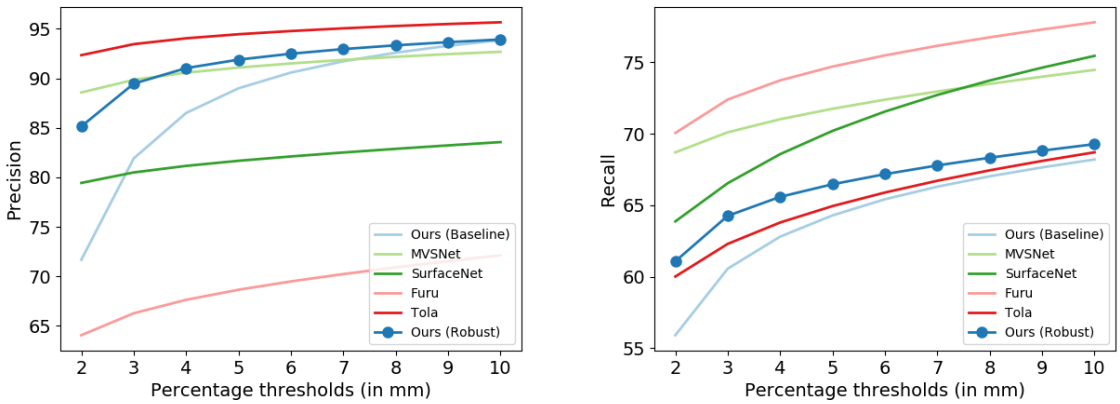

Figure 1: Percentage metrics on the DTU datasets as proposed in [6]. The f-scores reported in the main paper are computed using precision and recall which are displayed here.

Figure 2: Predictions for the remaining instances of the DTU test set. 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 [9]), 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. Best viewed in color.

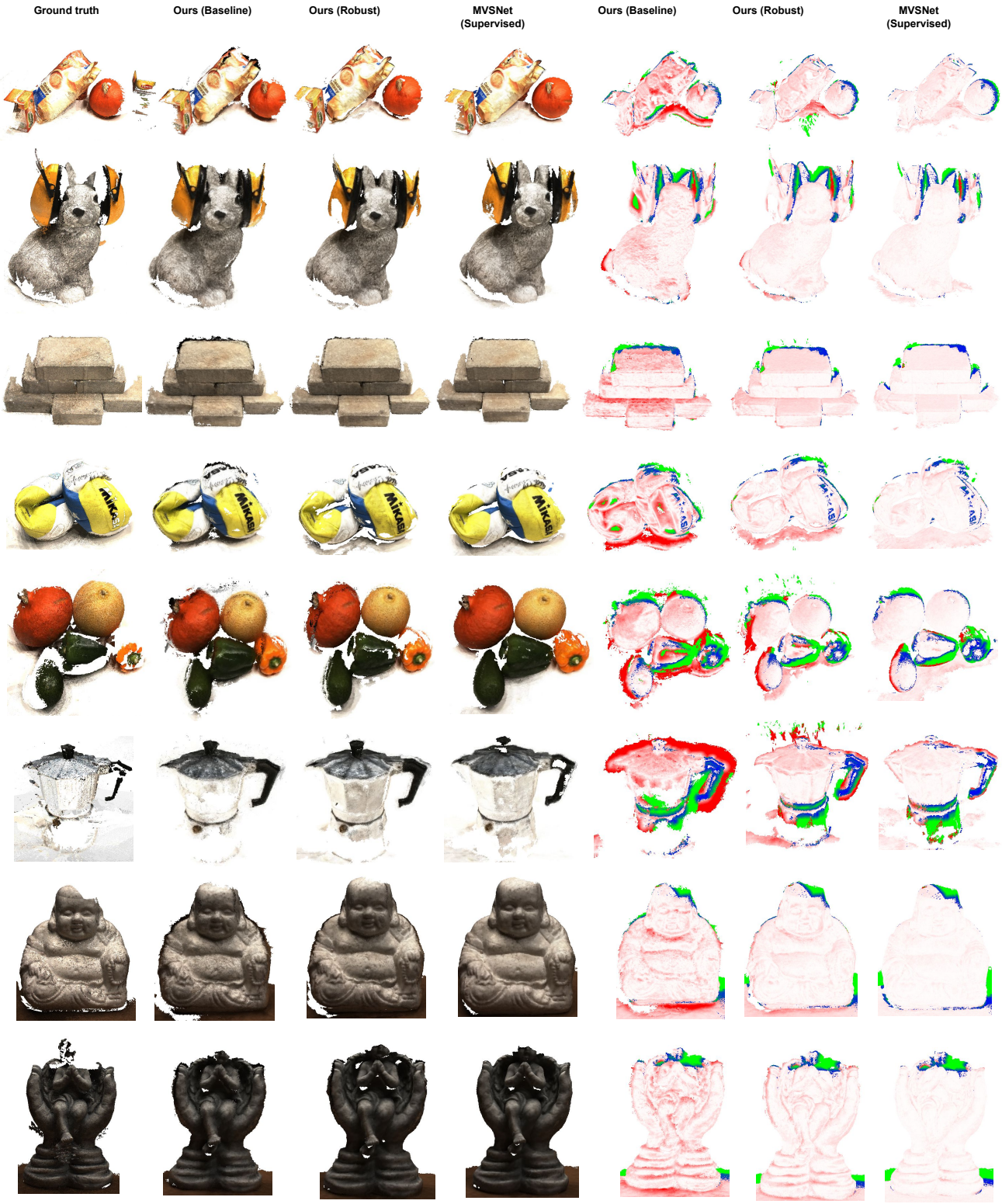

Figure 3: Predictions for the remaining instances of the DTU test set. Left to right: a) Ground truth 3D scan, b) result with baseline photo-loss, c) result with robust photo-loss, d) result using a supervised approach (MVSNet [9]), e-g) corresponding error maps. For the error maps, points marked blue/green are masked out in evaluation. Magnitude of error is represented by variation from white-red in increasing order. Best viewed in color.

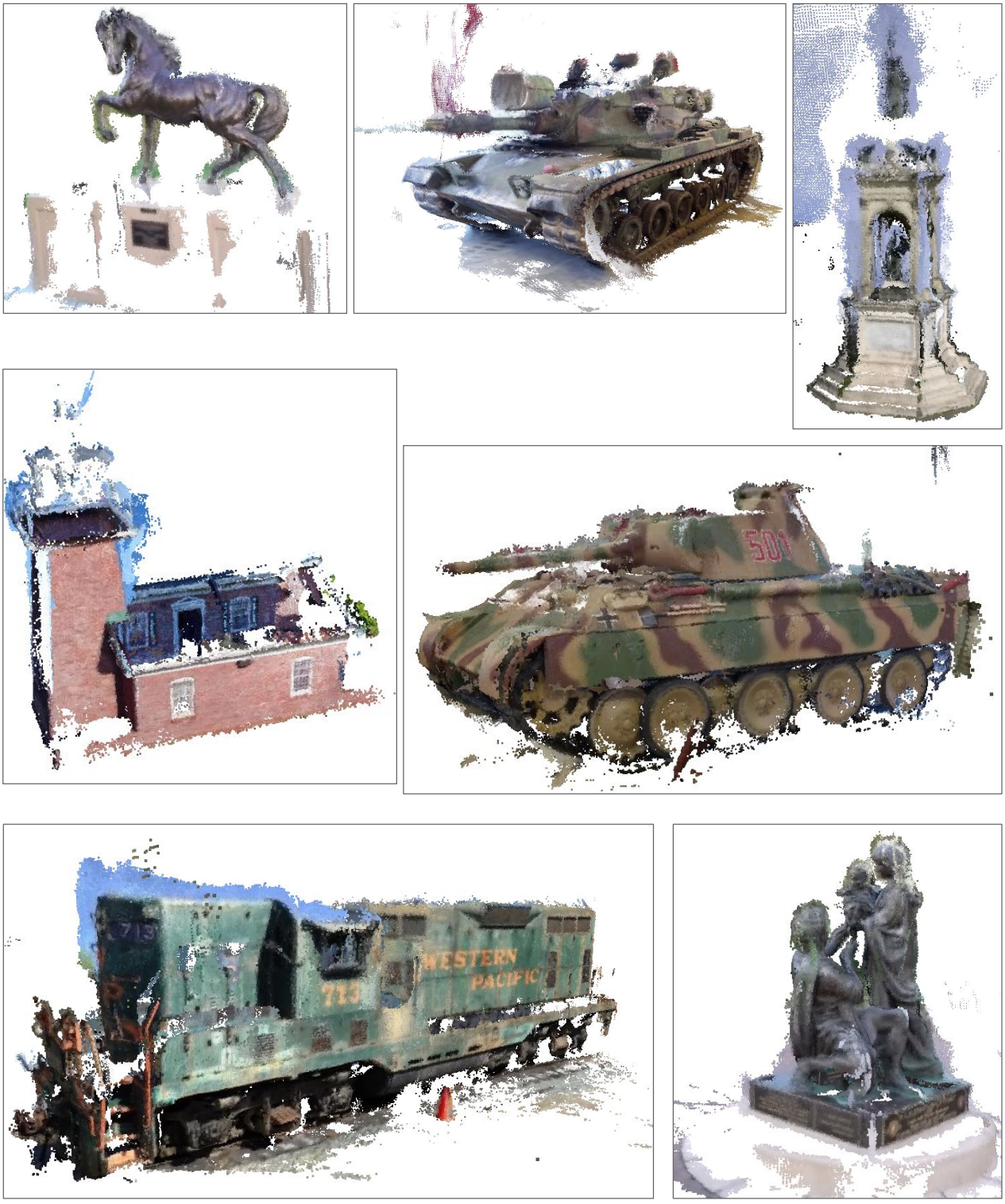

Figure 4: Generalization results of our robust model on the Tanks and Temples dataset without any finetuning. The results shown are from the Intermediate level instances of the dataset. The scene names from top left to bottom right: Horse, M60, Francis, Lighthouse, Panther, Train, Family. Best viewed in color.