Learning to Adapt Multi-View Stereo by Self-Supervision

Arijit Mallick[1]
arijit.mallick@uni-tuebingen.de
Jörg Stückler[2]
joerg.stueckler@tuebingen.mpg.de
Hendrik Lensch[1]
hendrik.lensch@uni-tuebingen.de

[1] Computer Graphics Group
University of Tübingen
Tübingen, Germany

[2] Embodied Vision Group
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Abstract

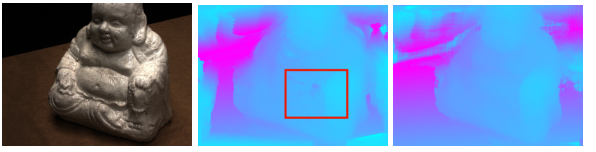
3D scene reconstruction from multiple views is an important classical problem in computer vision. Deep learning based approaches have recently demonstrated impressive reconstruction results. When training such models, self-supervised methods are favourable since they do not rely on ground truth data which would be needed for supervised training and is often difficult to obtain. Moreover, learned multi-view stereo reconstruction is prone to environment changes and should robustly generalise to different domains. We propose an adaptive learning approach for multi-view stereo which trains a deep neural network for improved adaptability to new target domains. We use model-agnostic meta-learning (MAML) to train base parameters which, in turn, are adapted for multi-view stereo on new domains through self-supervised training. Our evaluations demonstrate that the proposed adaptation method is effective in learning self-supervised multi-view stereo reconstruction in new domains.

1 Introduction

Dense 3D scene reconstruction based on images from multiple view points is one of the classical challenges in computer vision. It has widespread applications in areas such as computer aided design (CAD), virtual tours, augmented reality, cultural heritage preservation, construction maintenance and inspection, or robotics. Given the known view poses and camera intrinsics, multi-view geometry is typically used to find correspondences between pixels of reference along epipolar lines. Early approaches use handcrafted similarity measures for pixels or patches such as photometric similarity or normalized cross correlation. Deep learning has recently been demonstrated as a capable alternative for learning image features from data which can excel handcrafted measures [15, 16, 18, 23, 31, 34].

The state-of-the-art deep learning based methods for multi-view stereo reconstruction are supervised learning approaches which require immensive amounts of ground-truth 3D reconstruction data. Yet such data is tedious and difficult to obtain. Existing datasets such as [1, 17, 27] lack data diversity, come with calibration artifacts between the camera and the depth measuring device, or are synthetic. Hence, self-supervised learning methods which

© 2020. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.



(a) Buddha scan (b) Fine tuning (c) Ours

Figure 1: Example result for adaptive meta-learning for self-supervised domain transfer. a) 3D scan (from DTU dataset), b) Reconstruction result for pre-training on BlendedMVS dataset without meta-learning and fine-tuning on DTU training set. c) Reconstruction result for our approach with meta-learning on BlendedMVS and self-supervised fine-tuning on DTU. Note the depth artifacts in the red box by the naive fine-tuning approach which do not occur in our meta-learning approach.

can leverage large collections of camera images without the need of ground-truth 3D annotations are preferable. Apart from this, the given algorithm also needs to be robust against changes in environment or domains as it is not always possible to train a network with all possible environments in the training data. Hence, there needs to be a learning mechanism which can compensate for the changes in environment and quickly learn to adapt to different domains (indoors vs outdoors, low light vs bright light, building architecture scans vs object scans). A motivating example in our context is highlighted in Fig. 1. Recent developments in meta learning [8] demonstrated online adaptation to new tasks of supervised regression models which have been trained on a different set of tasks. In our approach, we propose a variant of model-agnostic meta-learning (MAML [8]) for training a multi-view stereo reconstruction network which facilitates self-supervised adaptation to new domains. We base our method on classical concepts from multi-view stereo (MVS) reconstruction and estimate dense depth in a reference view. Our model extends the network architecture of MVSNet [34] which has been demonstrated to yield state-of-the-art performance for supervised and self-supervised learning. In a first training stage, we use our meta-learning approach to train a network on a large dataset in several domains with ground-truth depth annotation. We train the network in such a way that it can better adapt to new domains through self-supervised training on data without ground-truth depth. In the second stage, we perform self-supervised fine-tuning on data from the new domain.

Like MVSNet, our multi-view stereo reconstruction network compares image features in cost volumes. This volume is refined with a set of 3D convolutions and we infer a preliminary depth map by neural regression from this refined volume. Different to the probability map for the depth as in MVSNet, we learn a confidence mask which is utilised to weight pixels for the self-supervised loss in order to compensate for outliers such as occlusions.

We demonstrate our adaptive learning approach by training on the BlendedMVS [36] dataset which contains a large collection of outdoor scenes (e.g. views of buildings, architecture etc.) and indoor scenes. We fine-tune our pre-trained model using self-supervised training on the DTU dataset [17] which consists of high resolution close scans of objects with different environment and lighting conditions. We evaluate our method on the DTU evaluation split and compare our approach to state-of-the-art MVS approaches and variants of our method such as fine-tuning without meta-learning. We demonstrate that meta-learning indeed helps to improve accuracy of MVS over naive fine-tuning. Our approach improves reconstruction results over a self-supervised baseline method. In our experiments, it does even compare well with several previous supervised and classical methods in certain metrics.

In summary, our contributions are

- We propose a novel meta-learning scheme for adaptive learning of multi-view stereo reconstruction which improves self-supervised domain adaptation.
- We extend MVSNet to learn a confidence mask for per-pixel weighting for self-supervised learning which handles outliers such as occlusions.
- We demonstrate that our meta-learning approach can improve self-supervised domain adaptation performance over naive pre-training in a supervised way. Our domain-adapted self-supervised multi-view stereo reconstruction achieves improved performance over a self-supervised MVS baseline.

2 Related work

Optimization-based Approaches. Multi view stereo estimation is one of the classical problems in computer vision with copious amount of research literature (see for example [26] for a survey). State of-the-art systems such as COLMAP [25], MVE [9] or PMVS [11] perform sparse structure from motion from collections of images to estimate sparse point cloud reconstruction, camera view poses and calibration parameters. Dense reconstruction is typically performed in a subsequent step, for instance, using patch-based surface representations and region-growing [11, 13, 33] or energy-minimization methods [6, 24]. Dense 3D surface reconstruction can be obtained by fusing depth maps in a 3D representation such as volumetric signed distance functions [33] or extracting meshes using point-cloud based surface reconstruction techniques [3, 19, 30]. A major problem of conventional multiview stereo approaches is that they are texture dependent and handcrafting good patch similarity measures is difficult.

Supervised Learning Approaches. Early supervised deep learning methods learn similarity measures for patches from multiple views, for instance using Siamese network architectures [14, 37]. More recent architectures [15, 16, 34] integrate disparity plane sweeping directly into the deep neural network architecture and compare pixel locations based on learned deep feature representations. We also follow this approach with our architecture which is based on MVSNet [34]. We extend the architecture with the prediction of a confidence mask and use it for meta-learning a model for domain adaptation using self-supervised training. Recently, also methods have been proposed using recurrency [35], volumetric fusion [18, 23], or deep learning on point sets [5].

Self-supervised Learning Approaches. One of the major problems of supervised techniques is the unavailibility of sufficiently large scale multi view stereo datasets with accurate depth map ground truth. In order to compensate for that, self-supervised multi-view stereo techniques have been developed very recently [7, 20]. The basic idea behind these approaches is to use the predicted depth to synthesize stereo images or images in a temporal window and train the networks for photoconsistent estimates. The camera view poses are either known by different means or have to be estimated concurrently. One of the major problems of both supervised and self-supervised learning approaches is that they typically do not generalize well to novel domains.

Meta-Learning. Recent developments in meta learning [8] have demonstrated methods that efficiently adapt to novel tasks for supervised regression and reinforcement learning. The main idea behind model agnostic meta learning (MAML) is to train the model parameters in such a way that the network can better generalise to a new task through fine-tuning. Previous work on adaptive learning of stereo disparity estimation [29] has utilised this meta-learning

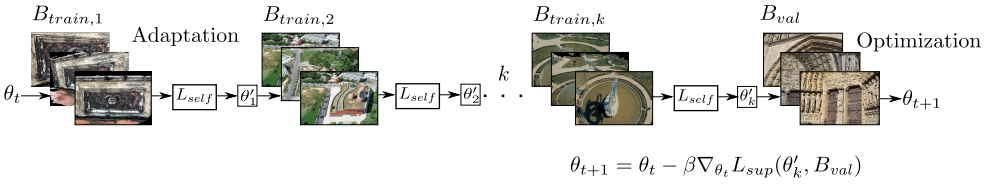


Figure 2: Meta-learning for self-supervised multi-view stereo. During a meta-learning iteration, adaptation is performed on k multi-view stereo reconstruction tasks with a self-supervised loss (L_{self}). The adapted parameters θ'_k are evaluated and the base model parameters θ_t are optimized on a validation set using a supervised loss L_{sup} to learn a better starting point for self-supervised parameter adaptation. For the new domain, the resulting base model trained through meta-learning is fine-tuned with self-supervised training.

and have shown how feature representations can be learned for self-supervised learning and improved generalization on new datasets. We propose to learn adaptive feature representations for self-supervised multi-view stereo reconstruction through meta-learning. We develop extensions to a network architecture based on MVSNet [34] with which the model learns to mask uncertain predictions due to outliers such as occlusions. This assists the self-supervised fine-tuning on new domain data.

3 Methodology

Meta learning aims at training a learning architecture for fast adaptability to new tasks. To this end, the model is trained on a set of different tasks during the meta learning phase. In our context, tasks correspond to self-supervised learning in different environments and conditions (i.e. domains).

The methodology can be summarized in two stages. The model is trained on a larger dataset with ground-truth depth in the first stage using meta-learning. In our experiments, we use the BlendedMVS dataset [36] which consists of indoor and outdoor scenes with varying environment conditions - making it ideal for domain adaptation. The model is trained on the training split by first updating cloned model parameters using the self-supervised photometric losses for k 'tasks', where a task refers to multi-view reconstruction of one of the k different scenes. The actual model parameters are then in turn updated using the supervised loss in Eq. (2) on the validation split, which involves the cloned and updated parameters from the previous step. The model trains network parameters to adapt well by self-supervised training. This is guided through the outer-loop supervised training (see Alg. 1). The second stage involves fine-tuning the model obtained in the first stage using self-supervised learning on the training data of the target domain dataset (DTU [17] in our experiments). We provide detailed explanation of the methodology in the following subsections.

3.1 Meta Learning for Self-supervised Multi-View Stereo

Our meta-learning algorithm for self-supervised multi-view stereo is summarized in Alg. 1 and illustrated in Fig. 2. We split the training dataset D into a training and a validation split D_{train} and D_{val}, the latter with m multi-view examples. Each example consists of a reference view (image with camera pose) and N neighbouring views.

We adapt the base model parameters θ for k multi-view examples $B_{train,i} \subset D_{train}$ each consisting of one reference view and N neighbouring views of a scene using a self-supervised loss (L_{self}, eq. (3)). Starting from the base parameters $\theta'_0 = \theta$, for each multi-view example i we perform the gradient update steps

$$\theta'_i = \theta'_{i-1} - \alpha \nabla_{\theta'_{i-1}} L_{self}(\theta'_{i-1}, B_{train,i}), \tag{1}$$

where α is a learning rate.

The base model parameters are optimized to improve the quality of the updated model parameters θ'_k with a supervised loss L_{sup} (eq. (6)) on a sampled multi-view example $B_{val} \subset D_{val}$ consisting of one reference view and N neighbouring views from the validation split,

$$min_{\theta}(L_{sup}(\theta'_k, B_{val})). \tag{2}$$

Note that θ'_k is a function of θ through the updates in Eq. (1). The supervised loss measures the discrepancy between the predicted and the ground-truth depth.

The intuition behind this two-step update scheme is that the base model parameters are changed to a better starting point for learning model parameters on different domains with the self-supervised loss. For a new dataset, we use the base parameters θ for fine-tuning to the new domain (i.e an entirely unseen dataset with different conditions and environment) using self-supervised training.

Algorithm 1: Adaptive learning for self-supervised multi-view stereo.

Data: Dataset split D_{train}, D_{val}, hyperparameters k, α, β
Initialize base model parameters θ ;
while *not converged* **do**
  Sample k multi-view examples $B_{train,i} \subset D_{train}$;
  Initialize model parameters $\theta_0 = \theta$;
  for $i \in [1, \ldots, k]$ **do**
    Compute adapted model parameters $\theta'_i = \theta'_{i-1} - \alpha \nabla_{\theta'_{i-1}} L_{self}(\theta'_{i-1}, B_{train,i})$;
    `// Adaptation`
  Sample batch $B_{val} \subset D_{val}$;
  Perform gradient descent step on base model parameters θ to minimize $L_{sup}(\theta'_k, B_{val})$ with learning rate β ; `// Optimization`

3.2 Network Architecture

While our adaptation module is model-agnostic, we base our network architecture on the MVSNet [34] model. MVSNet has demonstrated state-of-the-art performance for both supervised and self-supervised [20] training. Besides changing the training schemes with our meta-learning approach, we also augment the network with predicting confidence masks which are in turn used for self-supervised fine-tuning on novel domains (additional details can be found in the supplementary material). For details on the base network, readers are encouraged to refer to MVSNet [34].

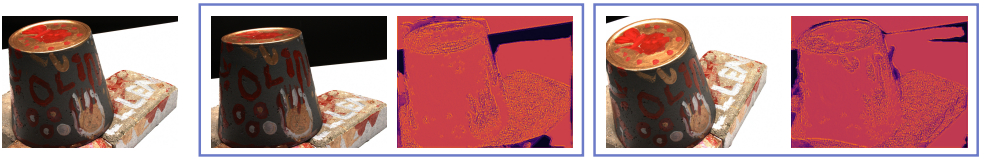


Figure 3: *From left to right*: reference image frame, first neighbouring frame, predicted confidence mask for first frame, second neighbouring frame and predicted confidence mask for second view. Different pairs of reference and neighbouring frames have different outliers such as occlusions, reflections, etc. We learn a confidence mask during meta-learning to downweight uncertain pixels for the self-supervised training and fine-tuning on new domains (darker color correspond to lower confidence in the visualization).

3.3 Learning Confidence Masks for Self-supervised Domain Adaptation

A major problem in learning multi-view stereo is to handle (dis-)occlusions and out-of-image projections correctly when quantifying the loss on the predicted depth maps. We take inspiration from [29] to learn a confidence mask during meta-learning which is used to improve the fine-tuning of the network on the new domain. While the approach in [29] has been proposed for learning dense reconstruction from stereo images of a constant-baseline stereo rig, multi-view stereo poses additional challenges due to the varying baselines between reference image and neighbouring frames.

Our network learns a confidence mask for each pair of reference image and neighboring frame in a multi-view training set. Fig. 3 provides an example of the confidence masks learned by our approach for different neighbouring views. The out-of-image projection mask $C_{proj} : \Omega \rightarrow [0,1]$ can be directly determined from the relative camera pose between the views and the predicted depth map. We train an additional component of our network architecture during meta-learning which predicts a confidence mask $C_\tau : \Omega \rightarrow [0,1]$ with learnable parameters τ for learning to downweight pixels in the loss at occlusions and other outliers. The final per-pixel mask is obtained by the product of the two masks at each pixel. The masks are used for the self-supervised loss to compensate for occlusions due to view pose changes. Note that the parameters τ are included into θ and updated during the meta-learning stage. They are held fixed when fine-tuning on a new domain dataset in the second stage.

The confidence mask network is a 4-layer CNN with sigmoid activation at the end to generate values between 0 and 1. The photometric warping error between reference I_{ref} and neighboring image I^i, and the out-of-image projection mask are concatenated and used as an input to the network that predicts the confidence mask. Please refer to the supplementary material for further details on the confidence mask subnetwork details and how it is integrated in our network architecture.

3.4 Training Losses

Self-supervised Losses. Self-supervised losses are used for adaptation during meta-learning and for fine-tuning on the new domain. The self-supervised loss comprises two components,

$$L_{self}(\theta, B) = L_{\text{recon}}(\theta, B) + \gamma_{\text{smooth}} L_{\text{smooth}}(\theta, B), \tag{3}$$

a reconstruction loss L_{recon} and a smoothness loss L_{smooth}, where θ are network parameters and B is a data example consisting of a reference frame and N neighbouring frames.

The reconstruction loss measures the image-based consistency between the reference and the N neighbouring views given their relative camera pose and the predicted depth map,

$$L_{\text{recon}}(\theta,B) = \sum_{i=1}^{N} \gamma_{\text{photo}} \left\| C^i_\tau(\theta,B) \odot C^i_{proj} \odot \left(I_{ref} - I^i_{warped}(\theta,B)\right) \right\|_1 + \gamma_{\text{ssim}} \left\| 1 - SSIM(C^i_{proj} \odot I_{ref}, C^i_{proj} \odot I^i_{warped}(\theta,B)) \right\|_1, \quad (4)$$

where γ_{photo} and γ_{ssim} are weighting factors and $\odot$ denotes pixel-wise multiplication. We use a combination of a photoconsistency measure and the structural similarity index (SSIM [32]). The reference image is I_{ref}, $I^i_{warped}(\theta,B)$ is the i^{th} neighbouring frame warped to the reference frame given the predicted depth by the network and known camera parameters. C^i_{proj} is the out-of-image projection mask which excludes the out of bound pixels while warping and $C^i_\tau(\theta,B)$ is the predicted confidence mask for the i^{th} frame. The structural similarity index [32] quantifies the similarity between I_{ref} and I_{warped} in patches centered at the pixels, and has been used in the literature [12, 20] since it measures texture similarity while being more robust to lighting changes than the photometric L1-loss.

An edge-dependent smoothness prior on the predicted depth maps with respect to the reference image is applied in order to encourage smoothness of the depth map. The smoothness loss for the predicted depth map $D(\theta,B)$ is

$$L_{\text{smooth}}(\theta,B) = \sum_{(x,y)} \left|\partial_x D_{x,y}(\theta,B)\right| e^{-\left\|\partial_x I_{x,y}\right\|_2} + \left|\partial_y D_{x,y}(\theta,B)\right| e^{-\left\|\partial_y I_{x,y}\right\|_2}, \quad (5)$$

where x, y range over the pixels in the reference frame.

Supervised Loss. For evaluation during meta-training, we use an L1 supervised loss on the depth map $D(\theta,B)$ predicted by the network to compare it with the ground truth D_{gt},

$$L_{sup}(\theta,B) = \left\| D(\theta,B) - D_{gt} \right\|_1. \quad (6)$$

4 Experiments

We evaluate our approach on a large-scale MVS dataset with ground-truth for the meta-learning stage and demonstrate domain adaptation on a smaller-scale MVS dataset from a different domain. For meta-learning, we use the BlendedMVS dataset [36] which has a mix of outdoor and indoor scenes. The dataset contains over 17k high-resolution images covering a variety of scenes, including cities, architectures, sculptures and small objects. The dataset is divided into training and validation sets which we use for the meta-learning. Domain adaptation is tested on the DTU [17] dataset, where we fine-tune the model on the training split and evaluate its final performance on the test split. The DTU scans consist of different objects in a different indoor environment with varied lighting conditions.

4.1 Training Details

The number of neighbouring frames N is 2 for meta-learning and fine-tuning. The model is tested with $N = 4$ frames. The number of depths ($d = 256$), input resolution ($H = 512, W =$

640) and output depth resolution ($H = 128, W = 160$) are initialized as in the original MVS-Net setup for fair comparison [34]. Learning rates are selected as $\alpha = 10^{-4}$ and $\beta = 10^{-4}$. The model is fine-tuned with a learning rate of 10^{-7} and a batch size of 4 multi-view examples with one reference frame and N neighbouring frames each. The self-supervised loss weights are set to $\gamma_{\text{photo}} = 5, \gamma_{\text{ssim}} = 1$ and $\gamma_{\text{smooth}} = 0.01$. For meta-learning we use $k = 3$ multi-view examples in each update cycle. The meta-training and testing have been performed on the same hardware configuration (4 NVidia Titan RTX GPUs) using a PyTorch implementation. We used the Learnable [2] library for implementing first-order MAML.

4.2 Depth Map Fusion

Similar to MVSNet [34], we fuse the predicted depth maps into point cloud reconstructions using [22][1]. The method determines a subset of the images using the view selection score of COLMAP [25]. Their depth maps are projected to 3D points in a common coordinate frame. Matches of points in neighboring views are found through reprojection into the images. Points with reprojection distance error with threshold < 1 and relative depth difference with threshold < 0.01 are averaged to obtain the final point cloud. We reconstruct point clouds by fusing the generated depth maps for those pixels with confidence above threshold > 0.8.

4.3 Quantitative Results

The fine-tuned model is evaluated on the DTU test split [18, 34]. We use the evaluation metrics as in [1]. The *accuracy* distance metric is measured as the distance from the estimated reconstruction to the ground-truth, encapsulating the accuracy of the estimated points. The *completeness* is measured as the distance from the ground-truth reconstruction points to the estimated reconstruction, encapsulating how much of the surface is captured by the MVS reconstruction. *Overall* is the mean of *accuracy* and *completion* (see Table 1). Additionally, we report the *overall F-score* metric [21] at inlier thresholds of 1 mm and 2 mm. We utilize [39] for calculating the precision and recall (see Table 1: %-age (percentage) columns). The F-score is the harmonic mean of precision and recall.

The results in Table 1 demonstrate that our method can improve results over its self-supervised baseline MVSNet in [20]. It is second to [2] in terms of overall metric among self-supervised methods. Filtering with the confidence mask can lead to higher accuracy in favor of lower completeness. Note that our method achieves state-of-the-art results in the overall F-score measures at 1 mm and 2 mm inlier threshold compared to self-supervised and classical methods. Remarkably it fares similar to one of the supervised methods (SurfaceNet) in several metrics.

4.4 Qualitative Results

Figure 4 display the qualitative evaluation of our proposed method with respect to supervised methods ([18, 34]). Our method provides a superior completeness and as it can be observed from the reconstruction, some surrounding structures are also reconstructed which are not present in the ground truth.

[1] We use the open-source implementation at https://github.com/xy-guo/MVSNet_pytorch with its default parameter setting

method	acc.	comp.	over.	prec.	rec.	over. F	prec.	rec.	over. F
				(1 mm) in %			(2mm) in %		
Camp [3] (C)	0.835	0.554	0.695	71.75	64.94	66.31	84.93	69.93	74.36
Furu [11](C)	0.612	0.939	0.775	69.55	61.52	63.26	77.3	64.06	70.06
Tola [23](C)	0.343	1.19	0.766	90.49	57.83	68.07	92.35	60.01	72.75
MVSNet [20](Sup DTU)	0.396	0.527	0.462	86.46	71.13	75.69	91.06	75.70	80.25
Ours (Sup PT bMVS, Sup FT DTU)	0.441	0.387	0.414	83.55	74.25	76.93	88.56	77.63	81.09
Surfacenet [13](Sup DTU)	0.450	1.043	0.746	83.8	63.38	69.95	87.44	67.87	74.81
MVSNet [20] (Self DTU)	0.881	1.073	0.977	61.54	44.98	51.98	85.15	61.08	71.13
MVS2 [3](Self DTU)	0.760	**0.515**	**0.633**	70.56	**66.12**	68.27	-	-	-
Ours (Meta PT bMVS, Self FT DTU)	**0.5942**	0.7787	0.6865	**80.18**	63.58	**68.67**	**90.95**	**69.08**	**76.22**

Table 1: Evaluation scores for reconstruction metrics (C: classical, Sup: supervised, Self: self-supervised, Meta: meta-learning). PT: pre-trained, FT: fine-tuned. bMVS: trained on Blended MVS. DTU: trained on DTU. Lower score is better for accuracy (acc.), completeness (comp.) and overall (over.) metrics. Higher score is better for precision (prec.), recall (rec.) and overall F-score (over. F) metric. Blue indicates best among all methods. Best results among methods trained self-supervised on DTU are shown in bold. Our approach demonstrates improved results over its self-supervised baseline MVSNet [20]. Our method achieves state-of-the-art results in the overall F-score measures at 1 mm and 2 mm inlier threshold compared to self-supervised and classical methods. We even fare similar to a supervised approach (SurfaceNet) in several metrics.

method	acc.	comp.	over. F	prec.	rec.	over. F
				(1 mm) in %		
Self DTU(d=128)	0.881	1.073	0.977	61.54	44.98	51.98
Self DTU (d=256)	1.159	**0.6083**	0.8837	64.85	**64.68**	63.57
Self PT bMVS, Self FT DTU	0.9448	0.6345	0.7896	68.43	63.38	64.42
Sup PT bMVS, Self FT DTU	0.7808	0.6769	0.7288	74.54	64.35	67.49
Ours (Meta PT bMVS, Self FT DTU, no conf. mask)	0.7242	0.8422	0.7832	75.22	60.25	65.31
Ours (Meta PT bMVS, Self FT DTU)	**0.5942**	0.7787	**0.6865**	**80.18**	63.58	**68.67**

Table 2: Ablation study (bold shows best results). Acronyms follow Table 1. Our meta learning approach achieves better overall scores than the other training variants.

4.5 Ablation Studies

We perform ablation studies on the following training conditions:

- Self-supervised MVSNet setup (*Self DTU (d=256)*) similar to [20], with twice the depth discretization level (d=256). It was trained on DTU train split, and has different loss hyperparameters (such as reprojection loss weights as proposed in [20]).
- Similar as the previous setup, but pre-trained (PT) on BlendedMVS using self-supervised learning (*(Self PT bMVS, Self FT DTU)*) and supervised learning (*Sup PT bMVS, Self FT DTU)*). The model is fine-tuned (FT) on DTU using self-supervised learning.
- Our meta-training setup without the confidence mask training (*Ours (Meta PT bMVS, Self FT DTU, no conf. mask)*).
- Our proposed meta-training setup with the confidence mask training (*Ours (Meta PT bMVS, Self FT DTU)*).

Table 2 shows the results for these variations of our model. The *overall* scores highlight that meta learning outperforms the straightforward fine-tuning strategy (PT bMVS) with the same sequence of datasets, even if it is pre-trained supervised. Confidence weight masks are effective for decreasing the effect of outliers during learning which improves performance.

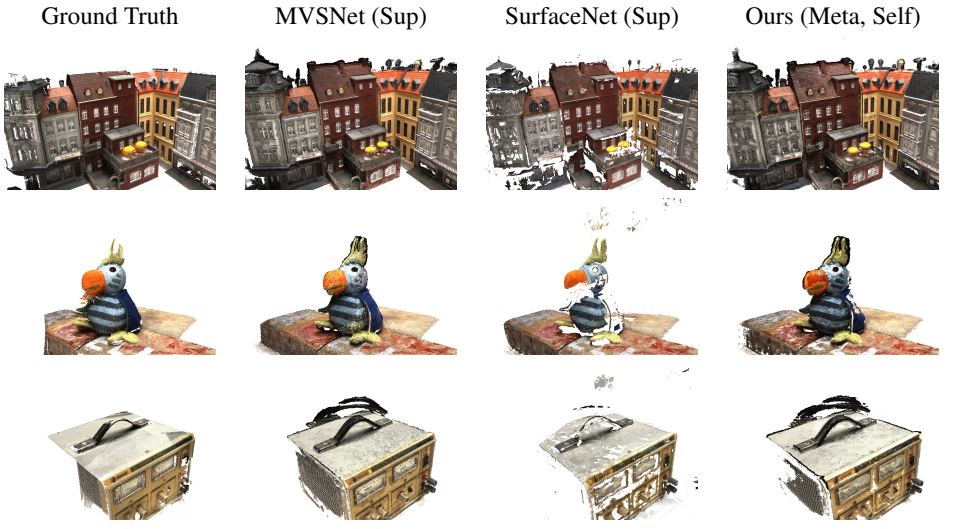


Figure 4: Point cloud reconstructions. From left to right: ground truth, MVSNet, SurfaceNet and ours. Our reconstruction results provide a better completeness than SurfaceNet and appear similar to the supervised MVSNet results.

5 Conclusions

Adaptability to new domains through self-supervision is a powerful property, especially for a multi-view stereo learning module where dense ground-truth depth data is tedious and difficult to obtain. We propose a meta learning approach which trains a network for self-supervised adaptation to a novel data domain with changes in environment and conditions. Our approach learns a loss confidence mask for self-supervised learning. In our experiments, we demonstrate that our meta-learning helps to train the network for adapting to new domains using self-supervision. Our approach can improve self-supervised domain adaptation performance over naive pre-training using depth supervision. It achieves reconstruction results which well compare with a previous supervised method and classical methods, and can improve performance over a self-supervised baseline.

Meta learning and multi-view stereo learning is a popular topic in the field of machine learning and computer vision. In the future, we will investigate architectures for high resolution images for an improved and more detailed reconstruction.

Acknowledgements

This work has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC number 2064/1 - project number 390727645 and SFB 1233 - project number 276693517. It was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and Cyber Valley. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Arijit Mallick.

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*, pages 1–16, 2016.

[2] Sebastien M.R. Arnold, Praateek Mahajan, Debajyoti Datta, and Ian Bunner. learn2learn, September 2019. URL https://github.com/learnables/learn2learn.

[3] Fatih Calakli and Gabriel Taubin. SSD: Smooth signed distance surface reconstruction. *Comput. Graph. Forum*, 30:1993–2002, 2011.

[4] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 766–779, 2008.

[5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1161–1174, 2011.

[7] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. MVS^2: Deep unsupervised multi-view stereo with multi-view symmetry. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2019.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135, 2017.

[9] Simon Fuhrmann, Fabian Langguth, Nils Moehrle, Michael Waechter, and Michael Goesele. MVE - an image-based reconstruction environment. *Computers & Graphics*, 53:44 – 53, 2015.

[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, Aug 2010. doi: 10.1109/TPAMI.2009.161.

[11] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, Aug 2010. doi: 10.1109/TPAMI.2009.161.

[12] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[13] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[14] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *Proceedings of International Conference on Learning Representations (ICLR)*, December 2018.

[17] 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.

[18] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2315, 2017.

[19] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, pages 61–70, 2006.

[20] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *CoRR*, abs/1905.02706, 2019. URL http://arxiv.org/abs/1905.02706.

[21] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 36:78:1–78:13, 2017.

[22] Paul Merrell, Amir Akbarzadeh, Liang Wang, Jan-Michael Frahm, Ruigang Yang, and David Nister. Real-time visibility-based fusion of depth maps. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[23] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. RayNet: Learning volumetric 3D reconstruction with ray potentials. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[25] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2006.

[27] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[28] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Mach. Vision Appl.*, 23(5):903–920, 2012.

[29] Alessio Tonioni, Oscar Rahnama, Tom Joy, Luigi Di Stefano, Ajanthan Thalaiyasingam, and Philip Torr. Learning to adapt for stereo. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[30] B. Ummenhofer and T. Brox. Global, dense multiscale reconstruction for a billion points. *International Journal of Computer Vision*, pages 1–13, 2017.

[31] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. DeMoN: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861. URL http://dx.doi.org/10.1109/TIP.2003.819861.

[33] Jian Wei, Benjamin Resch, and Hendrik P.A. Lensch. Dense and occlusion-robust multi-view stereo for unstructured videos. In *13th Conference on Computer and Robot Vision (CRV)*, 2016.

[34] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 785–801, 2018.

[35] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *IEEE/CVJ Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[37] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17:1–32, 2016.

[38] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017.

[39] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.