

Self-supervised Multi-view Stereo via Effective Co-Segmentation and Data-Augmentation

Hongbin Xu^{1,3*}, Zhipeng Zhou^{1*}, Yu Qiao^{1,2†}, Wenxiong Kang³, Qiuxia Wu³

¹ShenZhen Key Lab of Computer Vision and Pattern Recognition,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

²Shanghai AI Lab, Shanghai, China

³South China University of Technology, Guangzhou, China
hongbinxu1013@gmail.com, {zp.zhou, yu.qiao}@siat.ac.cn, {auwxkang, qxwu}@scut.edu.cn

Abstract

Recent studies have witnessed that self-supervised methods based on view synthesis obtain clear progress on multi-view stereo (MVS). However, existing methods rely on the assumption that the corresponding points among different views share the same color, which may not always be true in practice. This may lead to unreliable self-supervised signal and harm the final reconstruction performance. To address the issue, we propose a framework integrated with more reliable supervision guided by semantic co-segmentation and data-augmentation. Specially, we excavate mutual semantic from multi-view images to guide the semantic consistency. And we devise effective data-augmentation mechanism which ensures the transformation robustness by treating the prediction of regular samples as pseudo ground truth to regularize the prediction of augmented samples. Experimental results on DTU dataset show that our proposed methods achieve the state-of-the-art performance among unsupervised methods, and even compete on par with supervised methods. Furthermore, extensive experiments on Tanks&Temples dataset demonstrate the effective generalization ability of the proposed method.

1 Introduction

Multi-view stereo (MVS) aims at recovering 3D scenes from multi-view images and calibrated cameras, which is an important problem and widely studied in computer vision community (Seitz et al. 2006). Recent success of deep learning has triggered the interest of extending MVS pipelines to end-to-end neural networks. The learning-based methods (Yao et al. 2018, 2019) adopt CNNs to estimate the feature maps and build a cost volume upon the reference camera frustum to predict a per-view depth map for reconstruction. With the help of large-scale 3D ground truth, they outperform traditional geometry-based approaches and dominate the leaderboard. Whereas the learning-driven approaches strongly depend on the availability of 3D ground truth data for training, which is not easy to acquire (Zhong, Li, and Dai 2018). Thus it drives the community to focus on unsupervised/self-supervised MVS approaches.

*H.Xu and Z.Zhou contributed equally.

†Corresponding author.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

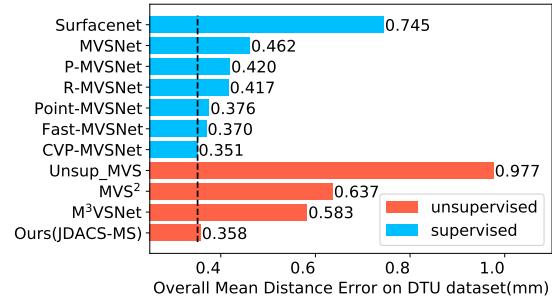


Figure 1: Comparison between SOTA supervised and unsupervised MVS methods.

Recently, there has been a surge in the number of self-supervised MVS methods that transform the depth estimation problem to an image reconstruction problem (Khot et al. 2019; Dai et al. 2019; Huang et al. 2020). The predicted depth map and the input image are used to reconstruct the image on another view, thus the self-supervision loss is built to estimate the difference between the reconstructed and realistic image on that view. However, as summarized in Figure 1, despite the impressive efforts in previous unsupervised methods, there still exists a clear gap between supervised and unsupervised results. In this paper, we suggest to rethink the task of self-supervision itself to improve the accuracy in MVS.

Previous self-supervised MVS methods largely rely on the same *color constancy hypothesis*, assuming the corresponding points among different views have the same color. However, as Figure 2 shows, in realistic scenarios, various factors may disturb the color distribution, such as light conditions, reflections, noise, etc. Consequently, the ideal self-supervision loss is susceptible to be confused by these common disturbances in color, leading to ambiguous supervision in challenging scenarios, namely *color constancy ambiguity*. To address the issues, we aim to incorporate the following extra priors of correspondence with the prior of color constancy in self-supervision loss: (1) *The prior of semantic correspondence can provide abstract matching*

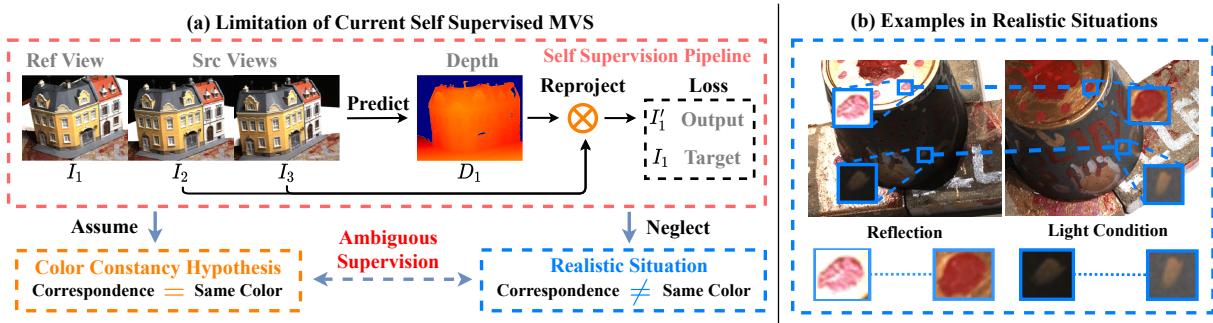


Figure 2: Illustration of the color constancy ambiguity problem in self-supervised MVS.

clues to guide the supervision. (2) The prior of data augmentation consistency can enhance the robustness towards color fluctuation. Hence, we propose a novel Joint Data-Augmentation and Co-Segmentation self-supervised MVS framework, namely JDACS.

For the prior of semantic consistency, most of the previous methods rely on the manually annotated semantic labels (Yang et al. 2018; Dovesi et al. 2019) restricted in fixed scenarios like autonomous driving with specified semantic classes. Whereas in the concern of MVS, on the one hand the semantic annotations are relatively expensive, on the other hand the huge variation in scenarios makes the semantic categories unfixed for segmentation which requires specified classes. Differently, we adopt non-negative matrix factorization (NMF) (Ding, He, and Simon 2005) to excavate the common semantic clusters among multi-view images dynamically for unsupervised co-segmentation (Collins, Achanta, and Susstrunk 2018). Then the semantic consistency is maximized among the re-projected multi-view semantic maps.

For the prior of data augmentation consistency, heavy data augmentation seldom appears in previous self-supervised MVS methods (Khot et al. 2019; Dai et al. 2019; Huang et al. 2020), because the natural color fluctuation in data augmentation will lead to the color constancy ambiguity in self-supervision. To preserve the reliability of self-supervision, we attach an additional data-augmentation branch with various transformations to the regular training branch. The output of regular training branch is taken as pseudo ground truth to supervise the output of augmented training branch.

In summary, our contributions are:

(1) We propose a unified unsupervised MVS pipeline called Joint Data-Augmentation and Co-Segmentation framework(JDACS) where extra priors of semantic consistency and data augmentation consistency can provide reliable guidance to overcome the color constancy ambiguity.

(2) We propose a novel self-supervision signal based on semantic consistency, which can excavate mutual semantic correspondences from multi-view images at unfixed scenarios in a totally unsupervised manner.

(3) We propose a novel way to incorporate heavy data augmentation into unsupervised MVS, which can provide regularization towards color fluctuation.

(4) The experimental results show that our proposed

method can lead to a leap of performance among unsupervised methods and compete on par with some top supervised methods.

2 Related Work

Supervised MVS: Recent advances in deep learning have interested a series of learnable systems for solving MVS problems (Huang et al. 2018; Ji et al. 2017). MVSNet (Yao et al. 2018) is an end-to-end MVS pipeline that builds a cost volume upon the reference camera frustum and learns the 3D regularization with CNNs. Many variants based on MVSNet have been proposed for improving the performance (Yao et al. 2019; Luo et al. 2019). Concurrently, along with the fervor for expanding the MVS framework to a coarse-to-fine manner, (Chen et al. 2019; Yu and Gao 2020; Yang et al. 2020; Cheng et al. 2020; Gu et al. 2019; Xu and Tao 2020) separate the single MVS pipeline into multiple stages, achieving impressive performances.

Unsupervised MVS: Under the assumption of photometric consistency (Godard, Mac Aodha, and Brostow 2017), unsupervised learning has been developed in multi-view systems. (Khot et al. 2019) inherit the self-supervision signal based on view synthesis and dynamically aggregates informative clues from nearby views. (Dai et al. 2019) predict the depth maps for all views simultaneously and filter the occluded regions. (Huang et al. 2020) further endow the depth-normal consistency into the MVS pipeline for improvement. Whereas all these methods share the assumption of color constancy, suffering from ambiguous supervision in challenging scenarios.

Segmentation Guided Algorithms: By assigning each pixel in the image to a specific class, semantic segmentation (Long, Shelhamer, and Darrell 2015) can provide an abstract representation. Several methods incorporate the scene parsing information with other tasks. SegStereo (Yang et al. 2018) enables joint learning for segmentation and disparity estimation simultaneously and (Cheng et al. 2017) utilize semantic clues to guide the training of optical flow estimation. These methods rely on annotated labels for segmentation in specific scenes like autonomous driving, whereas we differently concentrate on excavating semantics from dynamic scenarios. Co-segmentation methods aim at predicting foreground pixels of objects given an image collection (Joulin, Bach, and Ponce 2012). We apply unsupervised co-

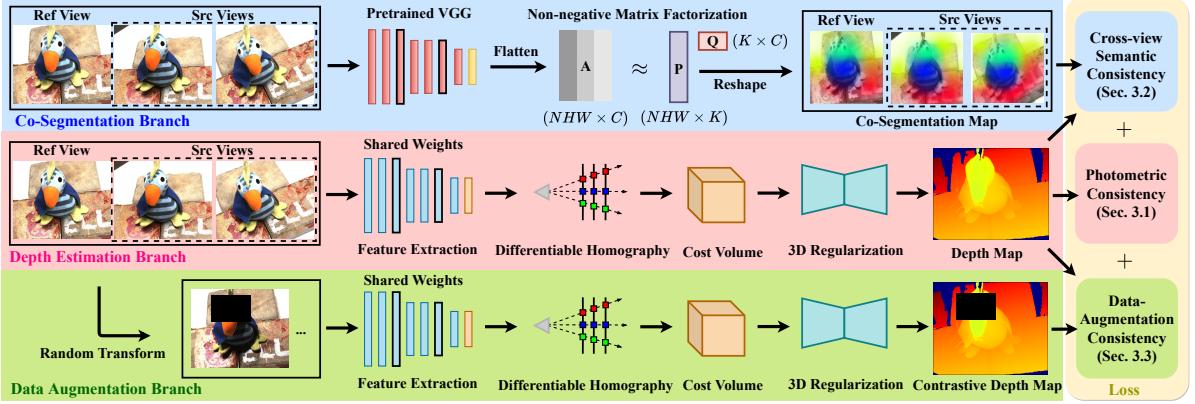


Figure 3: Illustration of our Joint Data-Augmentation and Co-Segmentation (JDACS) MVS framework.

segmentation (Casser et al. 2019) on the multi-view pairs to exploit the common semantics.

3 Method

In this section, we present Joint Data-Augmentation and Co-Segmentation framework(JDACS). To improve the reliability towards color constancy ambiguity, we incorporate extra priors of semantic consistency and data-augmentation consistency with a basic structure of deep MVS pipeline (Yao et al. 2018) in JDACS. As Figure 3 shows, the architecture of JDACS consists of Depth Estimation branch, Co-Segmentation branch and Data-Augmentation branch.

Depth Estimation Branch

As an unsupervised method, our proposed framework can be combined with arbitrary MVS networks. Here, we adopt MVSNet (Yao et al. 2018) as a representative backbone. The network firstly extracts features using a CNN from N input images. Then a variance-based cost volume is constructed via differentiable homography warping and a 3D U-Net is used to regularize the 3D cost volume. Finally, the depth map is inferred for every reference image. A sketch of the pipeline is shown in Figure 3.

Photometric Consistency: The key idea of photometric consistency (Barnes et al. 2009) is to minimize the difference between synthesized image and original image on the same view. Denote that the 1-st view is the reference view and the remaining $N - 1$ views as source views indexed by $i(2 \leq i \leq N)$. For a particular pair of images (I_1, I_i) with associated intrinsic and extrinsic parameters (K, T) . We can calculate the corresponding position p'_j in source view based on its coordinate p_j in reference view.

$$p'_j = KT(D(p_j)K^{-1}p_j) \quad (1)$$

where $j(1 \leq j \leq HW)$ is the index of pixels and D represents the predicted depth map.

The warped image I'_i can then be obtained by using the differentiable bilinear sampling from I_i .

$$I'_i(p_j) = I_i(p'_j) \quad (2)$$

Along with the warping, a binary validity mask M_i is generated simultaneously, indicating valid pixels in the novel view because some pixels may be projected to the external area of images. In a MVS system, we can warp all $N - 1$ source views to the reference view to calculate the loss.

$$L_{PC} = \sum_{i=2}^N \frac{\| (I'_i - I_1) \odot M_i \|_2 + \| (\nabla I'_i - \nabla I_1) \odot M_i \|_2}{\| M_i \|_1} \quad (3)$$

where ∇ denotes the gradient operator and \odot is dot product.

Co-Segmentation Branch

In previous methods (Yang et al. 2018; Casser et al. 2019), handcrafted semantic annotations are usually utilized to provide extra supervision to improve the performance. However, due to the huge variation of scenarios and the expensive cost for manual annotations in MVS, we differently choose to mine the implicit common segments from multi-view images via unsupervised co-segmentation. Co-segmentation aims at localizing the foreground pixels of the common objects given an image collection. It has been proven that non-negative matrix factorization (NMF) has an inherent clustering property in (Ding, He, and Simon 2005). Following a classical co-segmentation pipeline (Collins, Achanta, and Sussstrunk 2018), NMF applied to the activations of a pretrained CNN layer can be exploited to find semantic correspondences across images.

Non-negative Matrix Factorization: Non-negative matrix factorization(NMF) is a group of algorithms in multivariate analysis and linear algebra where a matrix A is factorized into two matrices P and Q . All the three matrices are with the property that having no negative elements. As (Ding, He, and Simon 2005) shows, NMF has an inherent clustering property that it automatically clusters the columns of matrix $A = (a_1, \dots, a_n)$. More specifically, if we impose an orthonormal constraint on $Q(QQ^T = I)$, then the approximation of A by $A \simeq PQ$ achieved by minimizing the following error function is equivalent to the optimization of K-means clustering.

$$\|A - PQ\|_F, P \geq 0, Q \geq 0 \quad (4)$$

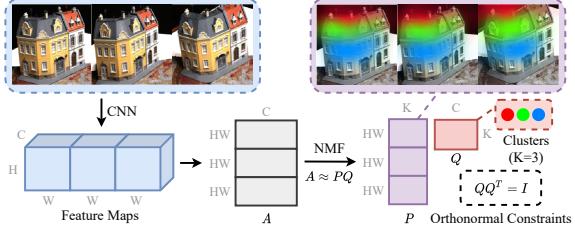


Figure 4: Brief illustration of the clustering effect of NMF.

where the subscript F means the Frobenius Norm.

Clustering on CNN Activations: ReLU is a common component for many modern CNNs, due to its desirable gradient properties. The CNN feature maps activated by ReLU result in non-negative activations, which naturally fit for the target of NMF. As shown in Figure 3, we apply a pretrained VGG network (Simonyan and Zisserman 2014) for feature extraction. Denote that the extracted feature map is of dimension (H, W, C) on each of the N views. Then the multi-view feature maps are concatenated and reshaped to a (NHW, C) matrix A . By utilizing multiplicative update rule in (Ding, He, and Simon 2005) to solve NMF, A is factorized into a (NHW, K) matrix P and a (K, C) matrix Q , where K is the NMF factors representing the number of semantic clusters. For a comprehensive understanding, we provide a brief interpretation of the results P, Q and the clustering effect of NMF in Figure 4.

The Q matrix: Due to the orthonormal constraints of NMF ($QQ^T = I$) (Ding, He, and Simon 2005), each row of the (K, C) matrix Q can be viewed as a cluster centroid of C dimensions, which corresponds to a coherent object among views.

The P matrix: The rows of the (NHW, K) matrix P correspond to the spatial positions of all pixels from N views. In general, the matrix factorization $A \approx PQ$ enforces the product between each row of P and each column of Q to best approximate the C dimensional feature of each pixel in A . As shown in Figure 4, $K = 3$ semantic objects are clustered in Q from the feature embeddings of all pixels in A , thus P contains the similarity between each pixel and each of the $K = 3$ clustered semantic objects. Consequently, P can further be reshaped into N heat maps of dimension (H, W, K) and fed into a softmax layer to construct the co-segmentation maps S .

Semantic Consistency Loss: With the co-segmentation maps S extracted from matrix P , we can design a self-supervision constraint based on semantic consistency. The key idea is to expand the photometric consistency across multiple views (Barnes et al. 2009) to the segmentation maps. Similar to the photometric consistency discussed in Section 3, we can calculate the corresponding position p'_j in source views with the pixel p_j in reference view according to Equation 1, given the predicted depth value $D(p_j)$ and the j -th pixel in the image. Then the warped segmentation map S'_i from the i -th source view can be reconstructed by bilinear sampling.

$$S'_i(p_j) = S_i(p'_j) \quad (5)$$

Finally, the semantic-consistency objective L_{SC} is measured by calculating the per-pixel cross-entropy loss between the warped segmentation map S'_i and the ground truth labels converted from reference segmentation map S_1 .

$$L_{SC} = - \sum_{i=2}^N \left[\frac{1}{\|M_i\|_1} \sum_{j=1}^{HW} f(S_{1,j}) \log(S'_{i,j}) M_{i,j} \right] \quad (6)$$

where $f(S_{1,j}) = \text{onehot}(\arg \max(S_{1,j}))$ and M_i is a binary mask indicating valid pixels from the i -th view to reference view.

Data-Augmentation Branch

Some recent works (Xie et al. 2019; Chen et al. 2020) in contrastive learning demonstrate the benefits of data augmentation in self-supervised learning. The intuition is that data augmentation brings challenging samples which bust the reliability of unsupervised loss and hence provides robustness towards variations.

Briefly, a random vector θ is defined to parameterize an arbitrary augmentation $\tau_\theta : I \rightarrow \bar{I}_{\tau_\theta}$ on image I . However, data augmentation has seldom been applied in self-supervised methods (Khot et al. 2019; Dai et al. 2019; Huang et al. 2020), because natural color fluctuation in augmented images may disturb the color constancy constraint of self-supervision. Hence, we enforce the unsupervised data augmentation consistency by contrasting the output of original data and augmented samples as a regularization, instead of optimizing the original objective of view synthesis.

Data Augmentation Consistency Loss: Specifically, as shown in Figure 3, the prediction of a regular forward pass for original images I in Depth Estimation branch is denoted as D . Accordingly, the prediction of augmented images \bar{I}_{τ_θ} is denoted as \bar{D}_{τ_θ} . In a contrastive manner, the data-augmentation consistency is ensured by minimizing the difference between D and \bar{D}_{τ_θ} :

$$L_{DA} = \frac{1}{\|M_{\tau_\theta}\|_1} \sum \| (D - \bar{D}_{\tau_\theta}) \odot M_{\tau_\theta} \|_2 \quad (7)$$

where M_{τ_θ} represents the unoccluded mask under transformation τ_θ . Due to the epipolar constraints among different views, the integrated augmentation methods in our framework should not change the spatial location of pixels. We will show some augmentation methods used in our method as follows:

Cross-view Masking: To simulate the occlusion hallucination among the multi-view situations, we randomly generate a binary crop mask $1 - M_{\tau_{\theta_1}}$ to block out some regions on reference view. Then the occlusion mask is projected to other views to mask out the corresponding area in images. Following the assumption that the remaining regions $M_{\tau_{\theta_1}}$ should be immune to the transformation, we can contrast the validity regions between the results of original and augmented samples.

Gamma Correction: Gamma correction is a nonlinear operation used to adjust the illuminance of images. To simulate various illuminations, we integrate random gamma correction τ_{θ_2} parameterized by θ_2 to challenge the unsupervised loss.

	Method	Acc.	Comp.	Overall
Geo.	Furu	0.613	0.941	0.777
	Tola	0.342	1.190	0.766
	Camp	0.835	0.554	0.694
	Gipuma	0.283	0.873	0.578
Sup.	Surfacenet	0.450	1.040	0.745
	MVSNet	0.396	0.527	0.462
	P-MVSNet	0.406	0.434	0.420
	R-MVSNet	0.383	0.452	0.417
	Point-MVSNet	0.342	0.411	0.376
	Fast-MVSNet	0.336	0.403	0.370
	CVP-MVSNet	0.296	0.406	0.351
UnSup.	Unsup_MVS	0.881	1.073	0.977
	MVS ²	0.760	0.515	0.637
	M ³ VSN	0.636	0.531	0.583
	JDACS	0.571	0.515	0.543
JDACS-MS		0.398	0.318	0.358

Table 1: Quantitative results on DTU evaluation benchmark. Geo. represents traditional geometric methods. Sup. represents supervised methods. UnSup. represents unsupervised methods.

Color Jitter and Blur: Many transformations can attach color fluctuation to images, such as random color jitter, random blur, random noise. The color fluctuation makes the unsupervised loss in MVS unreliable, because the photometric loss requires the color constancy among views. In contrast, these transformations denoted as τ_{θ_3} can create challenging scenes and regularize the robustness towards color fluctuation in self-supervision.

The overall transformation τ_{θ} can be represented as a combination of the aforementioned augmentations: $\tau_{\theta} = \tau_{\theta_3} \circ \tau_{\theta_2} \circ \tau_{\theta_1}$, where \circ represents function composition.

Overall Architecture and Loss

As shown in Figure 3, the overall framework has three components: Depth Estimation branch, Co-Segmentation branch and Data-Augmentation branch. In our paper, we aim to handle the color constancy ambiguity problem in self-supervised MVS, as discussed in Section 1. Apart from the basic self-supervision signal based on photometric consistency L_{PC} (Equation 1), we add two extra self-supervision signals of semantic consistency L_{SC} and data-augmentation consistency L_{DA} to the framework. In addition to the aforementioned loss, some common regularization terms suggested by (Mahjourian, Wicke, and Angelova 2018; Khot et al. 2019) for depth estimation are applied, such as structured similarity L_{SSIM} and depth smoothness L_{Smooth} .

The final objective can be constructed as follows:

$$\begin{aligned} L = & \lambda_1 L_{PC} + \lambda_2 L_{SC} + \lambda_3 L_{DA} \\ & + \lambda_4 L_{SSIM} + \lambda_5 L_{Smooth} \end{aligned} \quad (8)$$

where the weights are empirically set as: $\lambda_1 = 0.8$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.2$, $\lambda_5 = 0.0067$.

4 Experiments

In this section, we conduct comprehensive experiments to evaluate the proposed JDACS framework. First, we intro-

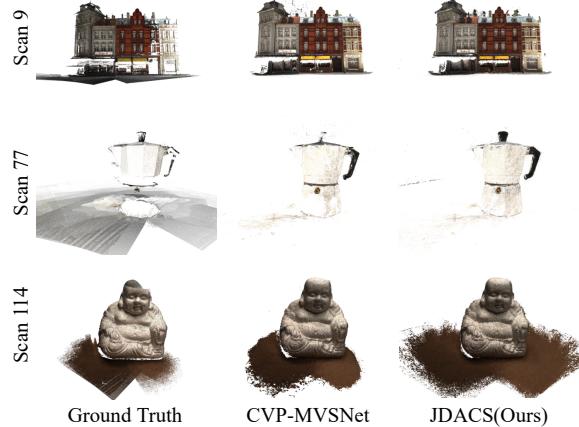


Figure 5: Qualitative comparison in 3D reconstruction between our JDACS and SOTA supervised method(CVP-MVSNet) on DTU dataset. From left to right: ground truth, results of supervised CVP-MVSNet, our results.

duce the implementation details. Then, we evaluate the proposed method on *DTU benchmark* (Aanæs et al. 2016) and further conduct ablation studies to analyze the significant components. At last, we test the proposed method on *Tanks&Temples benchmark* (Knapitsch et al. 2017) to verify the generalization ability.

Implementation Details

Backbone: In default, the most concise MVSNet (Yao et al. 2018) is applied as backbone in our JDACS framework. We denote the framework as JDACS-MS if a multi-stage MVSNet like CVP-MVSNet (Yang et al. 2020) is selected as backbone.

Training and Testing: During the training phase, we only use the training set of DTU without any ground truth depth maps. Our proposed JDACS¹ is implemented in Pytorch and trained on 4 NVIDIA RTX 2080Ti GPUs. In default, the hyper-parameters during training and testing phase follow the same setting of Unsup_MVS (Khot et al. 2019). With a pattern of data-parallel, the batch size is set to 1 per GPU for JDACS and 4 per GPU for JDACS-MS, which consume no more than 10G memories in each GPU. We use Adam optimizer with a learning rate of 0.001 which decreases by 0.5 times for every two epochs. JDACS is trained for 10 epochs as MVSNet (Yao et al. 2018) and JDACS-MS is trained for 27 epochs as CVP-MVSNet(Yang et al. 2020).

Error Metrics: In the DTU benchmark, *Accuracy* is measured as the distance from the result to the ground truth, encapsulating the quality of reconstruction; *Completeness* is measured as the distance from the ground truth to the result, encapsulating how much of the surface is captured; *Overall* is a the average of *Accuracy* and *Completeness*, acting as a compositive error metric. In the Tanks&Temples bench-

¹The code is released at: <https://github.com/ToughStoneX/Self-Supervised-MVS>

Method	Supervised	Input Size	Depth Map Size	Acc.	Comp.	Overall
MVSNet	✓	1152 × 864	288 × 216	0.456	0.646	0.551
JDACS	✗	1152 × 864	288 × 216	0.571	0.515	0.543
CVP-MVSNet	✓	1600 × 1152	1600 × 1152	0.296	0.406	0.351
JDACS-MS	✗	1600 × 1152	1600 × 1152	0.398	0.318	0.358

Table 2: Comparison between the backbone networks with same settings trained by supervision and our JDACS self-supervision framework. Due to the GPU memory limitation, we decrease the resolution of MVSNet to 1152 × 864 as (Chen et al. 2019).

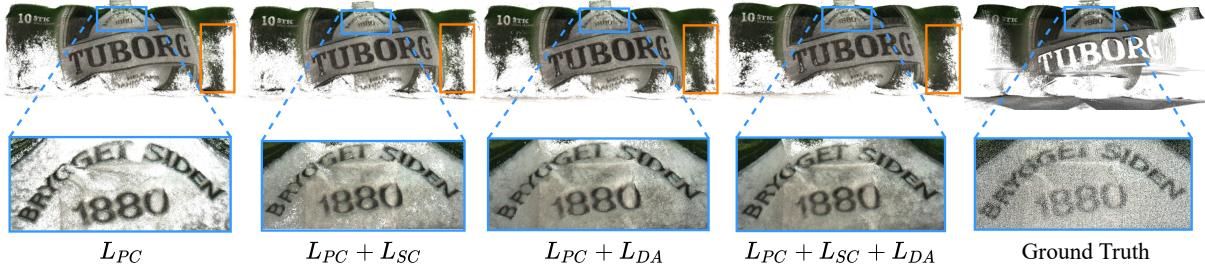


Figure 6: Qualitative results JDACS on *scan12* of the DTU dataset. Top row: Overview of generated point clouds with different combinations of self-supervision components. Bottom row: zoomed local areas. L_{PC} : Photometric-Consistency Loss; L_{SC} : Semantic-Consistency Loss; L_{DA} : Data-Augmentation-Consistency Loss.

L_{PC}	L_{SC}	L_{DA}	Acc.	Comp.	Overall
✓			0.7215	0.6339	0.6777
✓	✓		0.6134	0.5771	0.5953
✓		✓	0.5908	0.5887	0.5898
✓	✓	✓	0.5713	0.5146	0.5429

Table 3: Ablation Study of different components in our JDACS self-supervision network.

L_{PC}	L_{SC}	L_{DA}	Acc.	Comp.	Overall
✓			0.4645	0.4092	0.4369
✓	✓		0.4433	0.3892	0.4163
✓		✓	0.4330	0.3373	0.3851
✓	✓	✓	0.3977	0.3177	0.3577

Table 4: Ablation Study of different components in our JDACS-MS self-supervision network.

mark, *F-score* in each scene is calculated following the official evaluation process.

Benchmark Results on DTU

Comparison with SOTA: The official metrics of the DTU dataset (Aanæs et al. 2016) are: *Accuracy*, *Completeness* and *Overall*. These metrics are used to compare our proposed methods with other methods. The comparison includes traditional methods such as Furu (Furukawa and Ponce 2009), Tola (Tola, Strecha, and Fua 2012), Camp (Campbell et al. 2008), Gipuma (Galliani, Lasinger, and Schindler 2015). For the supervised methods, single stage networks such as Surfacenet (Ji et al. 2017), MVSNet (Yao et al. 2018), P-MVSNet (Luo et al. 2019), R-MVSNet (Yao et al. 2019), and multi-stage networks such as Point-MVSNet (Chen et al. 2019), Fast-MVSNet (Yu and Gao

Clusters	Acc.	Comp.	Overall
$K = 2$	0.6166	0.5752	0.5959
$K = 4$	0.6134	0.5771	0.5953
$K = 6$	0.6207	0.5827	0.6017
$K = 8$	0.6224	0.6030	0.6127

Table 5: Ablation Study of different numbers of semantic clusters K .

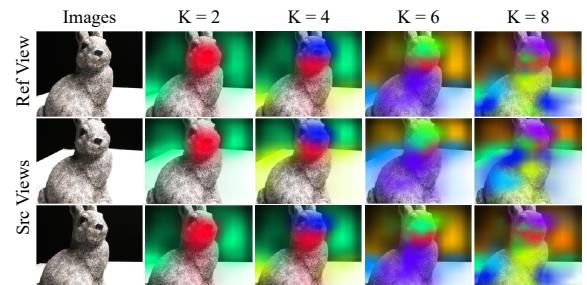


Figure 7: Visualization of the co-segmentation results with different number of segmentation parts K .

2020), CVP-MVSNet (Yang et al. 2020) are included. Furthermore, the current state-of-the-art unsupervised methods such as Unsup_MVS (Khot et al. 2019), M²VS (Dai et al. 2019) and M³VSN (Huang et al. 2020) are compared.

The quantitative results are shown in Table 1. From Table 1, we can conclude that our proposed method outperforms previous unsupervised methods in all official metrics. Furthermore, our proposed method can reconstruct better point cloud than traditional methods and some supervised methods in the metric of *Overall*. The supervised methods tend to have better performance in the metric of *Accuracy*, while

Method	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
MVS ²	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72
M ³ VSNet	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
Ours	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16

Table 6: Quantitative comparison with previous unsupervised methods without finetuning on Tanks&Temples dataset.

unsupervised methods usually achieve better performance in the metric of *Completeness*. The qualitative comparisons in Figure 5 demonstrate that our proposed method is comparable with some of the SOTA supervised methods.

Supervised vs Self-Supervised: From Table 1, we can find that there still exists a clear gap of performance between SOTA supervised methods and previous unsupervised methods. To provide a fair comparison without extra components, we compare our proposed self-supervision framework with supervised methods in the same network settings. The only difference is that our model is trained without any ground truth depth maps. The comparison is provided in Table 2. The supervised baselines are borrowed from previous papers(MVSNet from (Chen et al. 2019), CVP-MVSNet from (Yang et al. 2020)). The results in Table 2 demonstrate that our proposed framework can compete on par with the supervised opponents in the same network settings.

Ablation Studies

Effect of Different Prior Components: To evaluate the effect of our proposed prior of semantic consistency and data augmentation consistency, we train the networks with different combinations of these self-supervised signals. The quantitative results with different components in our proposed JDACS framework are summarized in Table 3 and Table 4. The model settings of JDACS in Table 3 and JDACS-MS in Table 4 is the same as the ones in Table 2. The qualitative visualization of the results of different components in JDACS-MS is provided in Figure 6. The experimental results demonstrate that endowing these extra priors into the self-supervision training can promote the performance in MVS. For example, as illustrated in Table 3, the *Overall* error metric decreases from 0.6777mm to 0.5953mm by including the prior of semantic consistency, from 0.6777mm to 0.5898mm with the help of involving data augmentation based branch.

Effect of Semantic Cluster Numbers: Different from manual semantic annotations in supervised learning, the semantic concepts excavated in an unsupervised manner are ambiguous. The number of semantic clusters K is a significant hyper-parameter for determining the categories of common semantic concepts among different views. Hence we conduct experiments about the effect of different semantic cluster numbers K and the results are reported in Table 5. Furthermore, a brief visualization of these semantic clusters is provided in Figure 7. From the visualization and the table, we can conclude that when the semantic clusters are more than 4, the localization of the semantic parts becomes less accurate than the ones with less than 4 clusters. As a result, we select $K = 4$ clusters as a default setting in our proposed method.



Figure 8: Visualization of the generated 3D point clouds without any finetuning on Tanks&Temples dataset.

Generalization

In this section, we compare our proposed JDACS with previous unsupervised methods on Tanks&Temples dataset. Due to the requirement of more than 20G memories in GPU using the original post-processing tool provided by (Yao et al. 2018), instead, we use an open simplified version² which can be deployed on a GPU with 11G memories like RTX 2080Ti. We follow the same hyper-parameter settings as MVS² (Dai et al. 2019). The quantitative comparison with previous unsupervised methods is provided in Table 6 and the visualization of the reconstructed dense point clouds is shown in Figure 8. Our proposed JDACS has better performance by the mean score of 8 scenes than previous unsupervised methods, which is the best unsupervised MVS method until September 9, 2020.

5 Conclusion

In this paper, we have proposed a novel unsupervised learning based MVS framework, JDACS, aiming at alleviating the gap between supervision and self-supervision caused by the coarse hypothesis of color constancy. On the one hand, our proposed method can enforce cross-view data-augmentation consistency into self-supervision with challenging variations. On the other hand, we can excavate the implicit common semantic clusters among different views and enforce the cross-view semantic consistency to provide a semantic-level correspondence metric. Experimental results on multiple benchmarks demonstrate the effectiveness of our proposed self-supervised framework.

²https://github.com/xy-guo/MVSNet_pytorch

6 Acknowledgments

This work was supported in part by the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800), in part by the National Natural Science Foundation of China under Grant (61876176, U1713208), and in part by the Shenzhen Basic Research Program (CXB201104220032A), Guangzhou Research Program (201803010066). This work was done during his internship at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

References

- Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision* 120(2): 153–168.
- Barnes, C.; Shechtman, E.; Finkelstein, A.; and Goldman, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* 28(3): 24.
- Campbell, N. D.; Vogiatzis, G.; Hernández, C.; and Cipolla, R. 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, 766–779. Springer.
- Casser, V.; Pirk, S.; Mahjourian, R.; and Angelova, A. 2019. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.
- Chen, R.; Han, S.; Xu, J.; and Su, H. 2019. Point-based multi-view stereo network. In *Proceedings of the IEEE International Conference on Computer Vision*, 1538–1547.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.
- Cheng, J.; Tsai, Y.-H.; Wang, S.; and Yang, M.-H. 2017. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, 686–695.
- Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L. E.; Ramamoorthi, R.; and Su, H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2524–2534.
- Collins, E.; Achanta, R.; and Susstrunk, S. 2018. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 336–352.
- Dai, Y.; Zhu, Z.; Rao, Z.; and Li, B. 2019. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*, 1–8. IEEE.
- Ding, C.; He, X.; and Simon, H. D. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, 606–610. SIAM.
- Dovesi, P. L.; Poggi, M.; Andraghetti, L.; Martí, M.; Kjellström, H.; Pieropan, A.; and Mattoccia, S. 2019. Real-time semantic stereo matching. *arXiv preprint arXiv:1910.00541*.
- Furukawa, Y.; and Ponce, J. 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence* 32(8): 1362–1376.
- Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 873–881.
- Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 270–279.
- Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2019. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching.
- Huang, B.; Huang, C.; He, Y.; Liu, J.; and Liu, X. 2020. M⁺ 3VS-Net: Unsupervised Multi-metric Multi-view Stereo Network. *arXiv preprint arXiv:2005.00363*.
- Huang, P.-H.; Matzen, K.; Kopf, J.; Ahuja, N.; and Huang, J.-B. 2018. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2821–2830.
- Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; and Fang, L. 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2307–2315.
- Joulin, A.; Bach, F.; and Ponce, J. 2012. Multi-class cosegmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 542–549. IEEE.
- Khot, T.; Agrawal, S.; Tulsiani, S.; Mertz, C.; Lucey, S.; and Hebert, M. 2019. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*.
- Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)* 36(4): 1–13.
- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.
- Luo, K.; Guan, T.; Ju, L.; Huang, H.; and Luo, Y. 2019. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, 10452–10461.
- Mahjourian, R.; Wicke, M.; and Angelova, A. 2018. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5667–5675.
- Seitz, S. M.; Curless, B.; Diebel, J.; Scharstein, D.; and Szeliski, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, 519–528. IEEE.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Tola, E.; Strecha, C.; and Fua, P. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications* 23(5): 903–920.
- Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Xu, Q.; and Tao, W. 2020. Learning Inverse Depth Regression for Multi-View Stereo with Correlation Cost Volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12508–12515.

Yang, G.; Zhao, H.; Shi, J.; Deng, Z.; and Jia, J. 2018. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 636–651.

Yang, J.; Mao, W.; Alvarez, J. M.; and Liu, M. 2020. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4877–4886.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 767–783.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5534.

Yu, Z.; and Gao, S. 2020. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1949–1958.

Zhong, Y.; Li, H.; and Dai, Y. 2018. Open-world stereo video matching with deep rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 101–116.

7 Supplementary Materials

Supplementary Details for Implementation

Implementation of NMF NMF plays an important role in the Co-Segmentation branch of our JDACS framework. We use the multiplicative update rule to calculate the solution of NMF iteratively, as shown in Algorithm 1.

Algorithm 1 Multiplicative Update Rule Based NMF

```

Set the number of segments as  $K$ ;
Set the number of maximum iterations as  $\text{ite}_{max}$  and the
tolerance constant as  $\text{tol}$ ;
Initialize non-negative matrices  $P$  and  $Q$  such that  $P \geq
0, Q \geq 0$ ;
for each iterative step  $v, 1 \leq v \leq \text{ite}_{max}$  do
     $Q^{v+1}_{[i,j]} \leftarrow Q^v_{[i,j]} \frac{((P^v)^t A)_{[i,j]}}{((P^v)^t P^v Q^v)_{[i,j]}}$ 
     $P^{v+1}_{[i,j]} \leftarrow P^v_{[i,j]} \frac{(A(Q^{v+1})^t)_{[i,j]}}{(P^v Q^{v+1} (Q^{v+1})^t)_{[i,j]}}$ 
    if  $\|A - P^{v+1} Q^{v+1}\|_F \leq \text{tol}$  then
         $P = P^{v+1}, Q = Q^{v+1}$ , stop the iterative process
    end if
end for

```

Implementation of JDACS-MS As mentioned in the main paper, if a multi-stage MVS-Net is applied in the Depth Estimation branch of JDACS, the framework is denoted as JDACS-MS. The predicted depth maps on all stages are utilized to calculate the self-supervision loss, as shown in Figure 9. In default, we adopt CVP-MVSNet as the backbone network.

Similar to the loss function of JDACS (Equation 8 in the main paper), the final objective of JDACS-MS can be constructed as follows:

$$L_{JDACS-MS} = \sum_{s=1}^5 (\lambda_1 L_{PC}^s + \lambda_2 L_{SC}^s + \lambda_3 L_{DA}^s + \lambda_4 L_{SSIM} + \lambda_5 L_{Smooth}) \quad (9)$$

where s represent each stage of the multi-stage MVSNet which is separated into 5 stages in default. The weights are empirically set as: $\lambda_1 = 0.8, \lambda_2 = 0.1, \lambda_3 = 0.1, \lambda_4 = 0.2, \lambda_5 = 0.0067$.

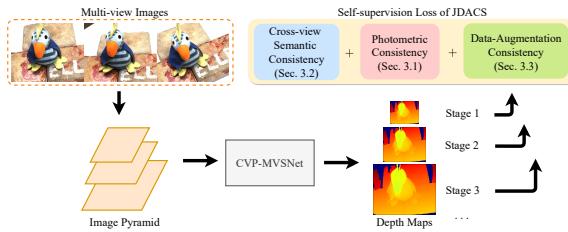


Figure 9: Brief illustration of JDACS-MS.

Data-augmentation Consistency Various transformations are adopted for generating challenging samples, such as occluding mask, Gaussian noise, blur, random jitter in brightness, color and contrast. In JDACS with a backbone of single stage MVSNet, the input is the original multi-view images and differently randomized transformation is applied to each view. In JDACS-MS with a backbone of multiple stage MVSNet, the input is an image pyramid of multiple scales and different transformation is added to different level of the image pyramid on each view.

How to Avoid Overflow in GPU Memory? The proposed framework possesses three parallel branches which may lead to an overflow in GPU memory during training. It may be impracticable to train the model on a GPU with 11G memory, such as GTX 1080 Ti or RTX2080 Ti. Hence, we conduct a simple trick to avoid the memory overflow by *trading the GPU memory with time*. In the training process, each step comprised of these three parallel branches and loss functions is separated into two sequential training step. For example, we can propagate the Depth Estimation branch and Co-segmentation branch in the first step, and save the estimated depth map as pseudo depth label. Then the Photometric consistency loss and Co-segmentation consistency loss are calculated, and the gradient is calculated during back-propagation. After updating the gradients, the cached memories are cleared. In the second step, the forward-propagation in the Data-augmentation branch is conducted and the weights are updated during the back-propagation.

How to Adjust the Weights for Self-supervised Loss? In practice, the convergence is sensitive to the attribution of weights for each term in self-supervision loss. If inappropriate weights are applied, it is likely to result in trivial solution in self-supervision. Hence, it is important to balance the weights. For the photometric consistency loss term, the weight is assigned following an open implementation³. For the co-segmentation consistency loss term, the weight is set according to the scale of the loss, which can be selected from 0.01 to 0.1. For the data-augmentation consistency loss term, it is mentioned that the data-augmentation consistency is actually a strong regularization to the self-supervised framework. In the starting phase of the training process, it may corrupt the convergence of self-supervision. Hence, we set the weight of data-augmentation consistency loss to 0.01 as an initial value, and increase it by 2 times after each 2 epochs, acting as a warming up process.

Visualization

In addition to the qualitative comparison in 3D reconstruction shown in Figure 6 of the main paper, we further provide more comparisons in Figure 10. Furthermore, we present the visualization of the reconstruction results on DTU dataset (Figure 11) and Tanks&Temples dataset (Figure 12). Please refer to Table 1 and Table 6 in the main paper for quantitative results on the datasets.

³https://github.com/tejaskhot/unsup_mvs



Figure 10: More qualitative comparison in 3D reconstruction between our JDACS and SOTA supervised method (CVP-MVSNet) on DTU dataset. From left to right: ground truth, results of CVP-MVSNet, our results.

Limitation and Discussion

Restriction of Coarse-grained Semantic Feature As shown in Table 5 and Figure 8 in the main paper, the co-segmentation results can only provide coarse-grained semantic feature with no more than 4 semantic clusters. The reason is that the semantic centroids are clustered from the feature space of a pretrained VGG specialized for classification task, where only the coarse-grained semantics are enough to construct distinguishable clues. However, in intuition, fine-grained semantics can provide more effective priors of correspondence for self-supervision. In the future, more accurate and refined semantic features are required for further improving the performance of self-supervision.

Restriction of Texture-less Region Although our proposed method can handle challenging cases with huge variation in color, it still fails to generalize to texture-less regions. The convergence of all self-supervision reconstruction loss is only effective on colorful regions. Because any pixels in texture-less regions share the same color intensity,

leading to the fact that self-supervision loss is fixed to 0 and becomes meaningless. However, the texture-less regions often appear in realistic scenarios, where self-supervision may be confused and fail to generalize. Exploration of handling texture-less regions may provide a potential direction in the future.



Figure 11: Point cloud reconstruction results on the DTU test set.

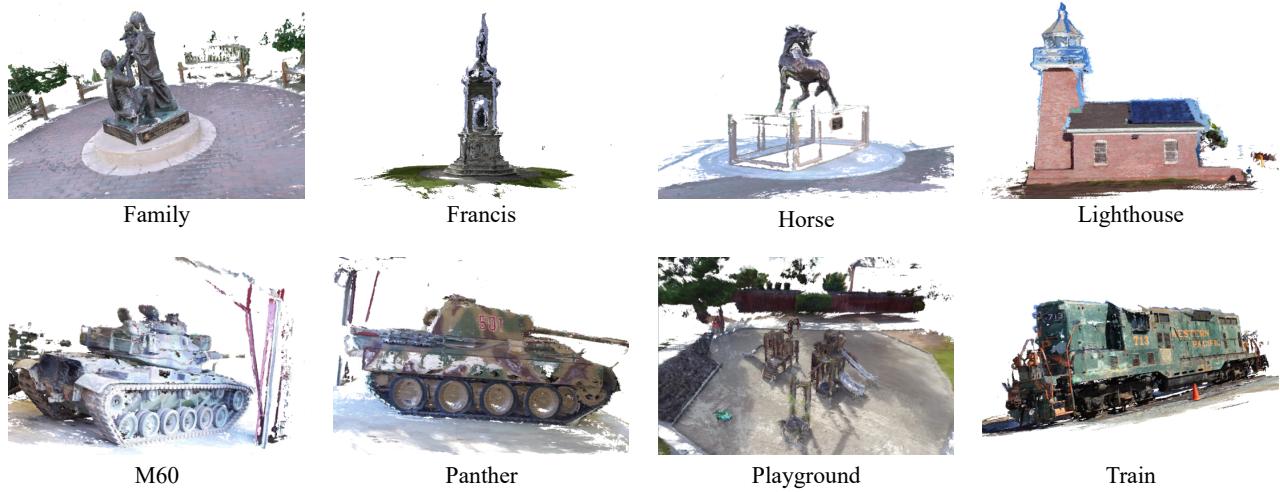


Figure 12: Point cloud reconstruction results on the Tanks&Temples test set without fine-tuning.