Multi-view Stereo 3D reconstruction based on MVSNet

Weiming Liu 23020231154152[1*], Ling Luo 33320231150356[2*], Guanqing Cui 23020231154175[1*], Yuxuan Liu 22320231151517[2*], Shuyu Cao 31520231154275[1*],

[1]School of Informatics, Xiamen University
[2]Institute of Artificial Intelligence, Xiamen University

Abstract

Multi-View Stereo (MVS) reconstruction is a technique in 3D modeling, which has garnered attention in various applications, from autonomous navigation to augmented reality and city planning. Traditional MVS methods face limitations in handling challenging scenes, while deep learning-based approaches like MVSNet have shown promise. Based on MVSNet, this study proposes the integration of the Convolutional Block Attention Module (CBAM) for cost volume regularization to enhance feature expression and introduces Squeeze-and-Excitation (SE) to enhance the network's ability to evaluate the importance of features and capture details. The method is evaluated using the DTU dataset, demonstrating its potential for enhanced 3D reconstruction performance.

Introduction

Multi-view stereo reconstruction, also known as MVS reconstruction, is a popular approach in 3D reconstruction that aims to reconstruct three-dimensional models of a scene based on a series of images captured from multiple viewpoints. This method is widely utilized in various domains such as autonomous driving, augmented reality, cultural heritage preservation, and smart cities. Compared to active 3D reconstruction methods that rely on devices like laser scanners and depth cameras, MVS as a passive image-based 3D reconstruction technique, offers several advantages, including high reconstruction accuracy, a wide field of view, low cost, and ease of widespread application.

Traditional MVS methods (Hirschmuller 2007; Furukawa and Ponce 2009; Schonberger and Frahm 2016) use handcrafted similarity metrics and regularizations to compute dense correspondences and recover 3D points. Although these methods have demonstrated great performance in ideal Lambertian conditions, they suffer from certain limitation. For example, low-textured, specular and reflective regions of the scene may make dense matching intractable, which leads to incomplete reconstructions.

Recent achievements in Deep Learning have sparked interest in improving MVS reconstruction as well. Compared with traditional methods, point clouds generated from multi view 3D reconstruction based on deep learning are more accurate and complete. Yao et al. (2018)proposed a deep learning-based stereo vision network called Multi-View Stereo Network (MVSNet). It decouples the problem of reconstructing a 3D scene by constructing a three-dimensional cost volume on the reference view and solving it as a single-view depth estimation task. Luo et al. (2019) introduced the Point Multi-View Stereo Network (Point-MVSNet), which is based on cost volumes by accumulating matching confidences. Yu et al. (2020) presented a new Fast Multi-View Stereo Network (Fast-MVSNet) that progresses from sparse to dense and from coarse to fine for rapid 3D reconstruction. Current state-of-the-art MVS reconstruction methods based on deep learning have demonstrated excellent performance (Sun et al. 2021; Zhang et al. 2023; Gu et al. 2020). They compute different resolution depth maps in a coarse-to-fine process and progressively narrow hypothesis plane guidance to reduce computational expense.

MVSNet propose an end-to-end deep learning architecture for depth map inference, which computes one depth map at each time. However, there is still room for improvement in MVSNet performance because it neglects the feature extract process of both 2D images and 3D cost volume regularization in which contains rich information. In this work, we propose a multi-view 3D reconstruction method based on MVSNet, which aims at improving accuracy in depth estimation and multi-view 3D reconstruction.

We have introduced the CBAM (Woo et al. 2018b) on the basis of MVSNet, extending the CBAM module to the extraction of three-dimensional features. CBAM applies attention based functional refinement through two different modules (channel and space), achieving significant performance improvements while maintaining minimal overhead, dynamically adjusting the weights of channel and spatial feature volume. Meanwhile, the SE attention module (Hu, Shen, and Sun 2018) is also embedded in the feature extraction network of MVSNet. SE helps networks better focus on important functional channels and extract more effective features. Through these steps, the accuracy and completeness of 3D reconstruction have been improved. The experiment using DTU dataset (Jensen et al. 2014) shows that this model achieves higher integrity and smoothness in 3D point cloud reconstruction compared with MVSNet, and its performance is also improved.

In summary, the main contributions are as follows.

[*]These authors contributed equally.

- By applying CBAM attention mechanism in both channel and spatial dimensions, more effective feature extraction can be achieved, resulting in more accurate and detailed 3D reconstruction.
- SE module's ability to adaptively recalibrate channel-wise feature responses enhances the network's focus on relevant features, improving the overall quality and fidelity of the reconstructed 3D models.
- The proposed method achieved better performance compared to MVSNet when evaluated on DTU dataset.

Related works

Traditional MVS Methods

Traditional MVS methods which represent the 3D geometry of objects or scene using voxels (De Bonet and Viola 1999; Sinha, Mordohai, and Pollefeys 2007), point cloud (Lhuillier and Quan 2005; Furukawa and Ponce 2009), meshes(Esteban and Schmitt 2004; Fua and Leclerc 1995) and depth maps(Tola, Strecha, and Fua 2012; Galliani, Lasinger, and Schindler 2015).

In the following, we mainly discuss about voxel-based MVS and depth maps-based MVS methods which have been integrated to learning-based framework recently. Voxel-based methods have the capability to represent a wide range of objects and scenes. These techniques do not impose constraints on the shape of the objects, but they require a significant amount of memory due to the discretization of space.

Comparatively, depth map-based methods are more concise and flexible. Galliani et al. (2015) present Gipuma, a massively parallel multi-view extension of Patchmatch stereo. It uses a red-black checkerboard pattern to parallelize message-passing during propagation. Sch¨onberger et al. (2016) present COLMAP, which jointly estimates pixel-wise view selection, depth map and surface normal.

Learning-based MVS Methods

Existing Learning-based MVS Methods can mainly be divided into two categories: voxel-based MVS (Sun et al. 2021; Ji et al. 2017) and depth maps-based MVS (Yao et al. 2018a; Zhang et al. 2023; Gu et al. 2020; Yang et al. 2020). Depth map-based methods such as MVSNet (Yao et al. 2018a) constructs the cost volume by aggregating deep features and camera parameters, and uses 3D CNN for regularization. And to reduce memory consumption and run-time, several subsequent studies have been developed (Zhang et al. 2023; Gu et al. 2020; Yang et al. 2020), which adopt cascade cost volumes or cost volume pyramid to estimate depth maps in a coarse-to-fine manner. To explicitly integrate geometric clues implied in coarse stages for delicate depth estimation, Zhang (2023) proposed a geometry awareness model termed GeoMVSNet, which achieves the state-of-the-art in Multi-view 3D reconstruction. Voxel-based method (Sun et al. 2021; Ji et al. 2017) uses a trained network to regress the occupancy rate of each voxel, but the volume representation method incurs significant memory consumption.

Method

Squeeze-and-Excitation Networks

Squeeze-and-Excitation Networks(SE Networks) adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels.

The structure of the SE building block is depicted in Figure 1. For any given transformation Ftr mapping the input X to the feature maps U where U ∈ RH×W×C$R^{H\times W\times C}$,e.g. a convolution, we can construct a corresponding SE block to perform feature recalibration. The features U are first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions (H × W). The function of this descriptor is to produce an embedding of the global distribution of channel-wise feature responses, allowing information from the global receptive field of the network to be used by all its layers. The aggregation is followed by an excitation operation, which takes the form of a simple self-gating mechanism that takes the embedding as input and produces a collection of per-channel modulation weights. These weights are applied to the feature maps U to generate the output of the SE block which can be fed directly into subsequent layers of the network.

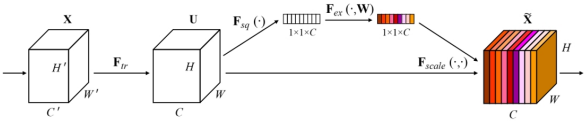

Figure 1: A Squeeze-and-Excitation block.

A Squeeze-and-Excitation block is a computational unit which can be built upon a transformation F_{tr} mapping an input$X \in R^{H'\times W'\times C'}$to feature maps $U \in R^{H\times W\times C}$ .In the notation that follows we take F_{tr} to be a convolutional operator and use V = [v1, v2, . . . , vC] to denote the learned set of filter kernels, where vc refers to the parameters of the c-th filter. We can then write the outputs as U = [u1, u2, . . . , uC], where $\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s$. Here denotes convolution, vc = $[V_c^1, V_c^2, ..., V_c^{C\prime}]$,X =$[x^1, x^2, ..., x^{c\prime}]$ and $u_c \in R^{H\times W}\cdot v_c^s$ is a 2D spatial kernel representing a single channel of vc that acts on the corresponding channel of X. To simplify the notation, bias terms are omitted. Since the output is produced by a summation through all channels, channel dependencies are implicitly embedded in v_c,but are entangled with the local spatial correlation captured by the filters.

3D convolutional block attention module

Convolutional Block Attention Module (CBAM),a simple yet effective attention module for feed-forward convolutional neural networks.Different aspect of 2D convolutional network is that 3D convolutional network has one more deep dimension, The specific integration mode is shown in Figure Figure 2.When extracting spatial and spatial features, the variation in depth parameters need to be taken into account. For an intermediate 3D convolutional layer: $F_{3D} \in$

$R^{W\times H\times D\times C}$,3D-CBAM will deduce the channel attention feature map in order: $M_{c3D} \in R^{1\times 1\times 1\times C}$,and the spatial attention features of , Fig: $M_{s3D} \in R^{1\times H\times W\times D}$,the whole process formula is shown as follows:

$$\begin{aligned} F'_{3\mathrm{D}} &= M_{c_{3D}}(F_{3\mathrm{D}}) \otimes F_{3\mathrm{D}}. \\ F''_{3\mathrm{D}} &= M_{s_{3D}}(F'_{3\mathrm{D}}) \otimes F'_{3\mathrm{D}}. \end{aligned} \tag{1}$$

The channel attention module of 3D-CBAM focuses on which channels serve the final classification result of the fused 3D network, i. e., selecting the features that are decisive for the prediction, and the specific steps are shown in Figure 3. First, the input feature diagram $F_{3\mathrm{D}}$ through the maximum pooling and mean pooling based on width W, depth H and depth D, and then through the features of the MLP, and then activate the generated channel feature diagram $M_{\mathrm{C3D}}(F_{3\mathrm{D}})$and the input feature diagram $F_{3\mathrm{D}}$ to generate the final channel feature diagram $F_{3\mathrm{D}}$, formula:

$$\begin{aligned} M_{c_{3\mathrm{0}}}(F_{30}) &= \sigma(\mathrm{MI}P(\mathrm{AvgPool3D}(F_{30})) \\ &\quad + \mathrm{MI}P(\mathrm{MaxPool3D}(F_{30}))) \\ &= \sigma(W_1(W_0(F^{\infty}_{\mathrm{arg}})) + W_1(W_0(F^{\infty}_{\max}))) \end{aligned} \tag{2}$$

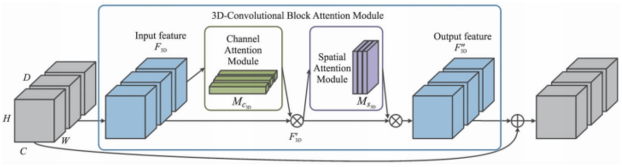

Figure 2: The structure of 3D-CBAM.

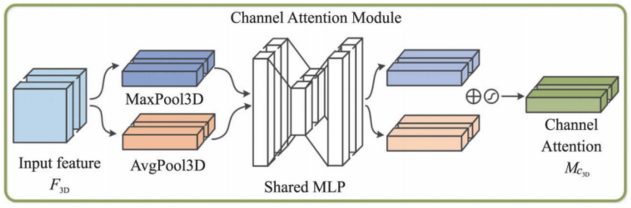

Figure 3: Channel attention module of 3D-CBAM.

Where $V_c \in R^{c/r\times C}$, $W^{C\times C/r}$, σ is the sigmoid operation, and W_0 needs to be activated by the Relu function. In this paper, the value of the reduction rate r is 8, that is, the channel C is transformed to C / 8 when the maximum pool and the mean pool are transformed, the number of parameters is reduced, and finally the full connection is transformed to the original channel C.

The spatial attention model of 3D-CBAM focuses on which pixels in the RGB image play a decisive role in the prediction of the network, and the specific attention feature extraction process is shown in Figure 4. First, the feature map F'_{3D} of the channel attention module was taken as the input feature map of the spatial attention module, Do a

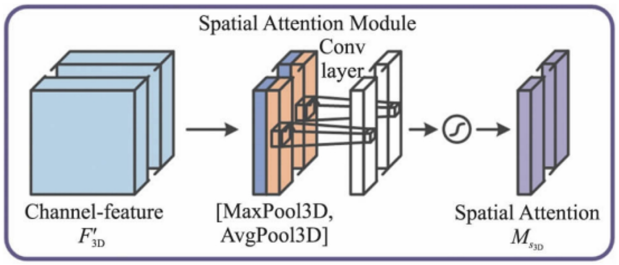

Figure 4: Spatial attention module of 3D CBAM.

channel-based maximum pooling and mean pooling operation, Both the extracted features F^s_{avg} and F^s_{max} are then subjected to the channel-based merging operation, Then, a convolution operation of 77 reduces its dimension into a channel and then goes through the sigmoid activation function to generate the spatial attention feature map, Finally, use the feature map and F'_{3D} of the module to multiply the final generated feature F''_{3D},The formula is given as follows:

$$\begin{aligned} M_{s_{30}}(F'_{30}) &= \sigma(f^{3\times 7}([\mathrm{AvgPool3D}(F'_{30}); \mathrm{MaxPool3D}(F'_{30})])) \\ &= \sigma(f^{7\times 7}([F^s_{avg}; F^s_{max}])) \end{aligned} \tag{3}$$

It is proved that the convolution operation of 77 is better than the convolution of 33. Because it is applied to 3D convolution and the channel sorting format of video sequence frames is channel-last, it is necessary to string the channels of axis = 4 in the tensor during the merging operation, and then conduct the convolution operation to ensure that the number of features of the axis = 4 is 1.

Proposed Solution

MVSNet is a deep learning architecture designed for inferring depth maps from multi-view images. It operates by initially extracting intricate visual features from images, then generating a 3D cost volume based on a reference camera frustum through differentiable homography warping.A distinguishing feature is its adaptability to arbitrary N-view inputs via a variance-based cost metric, consolidating multiple features into a single cost feature. Then, an initial depth map can be regressed by utilizing 3D convolutions, which is subsequently refined by the reference image to produce the final output.

However, acknowledging potential limitations in feature extraction, a proposed enhancement involves incorporating a squeeze-and-excitation (SE) module to dynamically adjust feature map channel weights, aiming to improve the net work's adaptability and performance in capturing more relevant features. Besides,the 3D cost volume regression part applies the 3D-CBAM module to effectively extract complex 3D features. The full architecture of proposed method is shown in Figure 5.

Cost Volume MVSNet constructs a 3D cost volume utilizing extracted feature maps and input camera data. Denoting the reference image as I_1, source image as$\{\mathrm{I}_i\}_{i=1}^{\tilde{N}}$ from 2 to

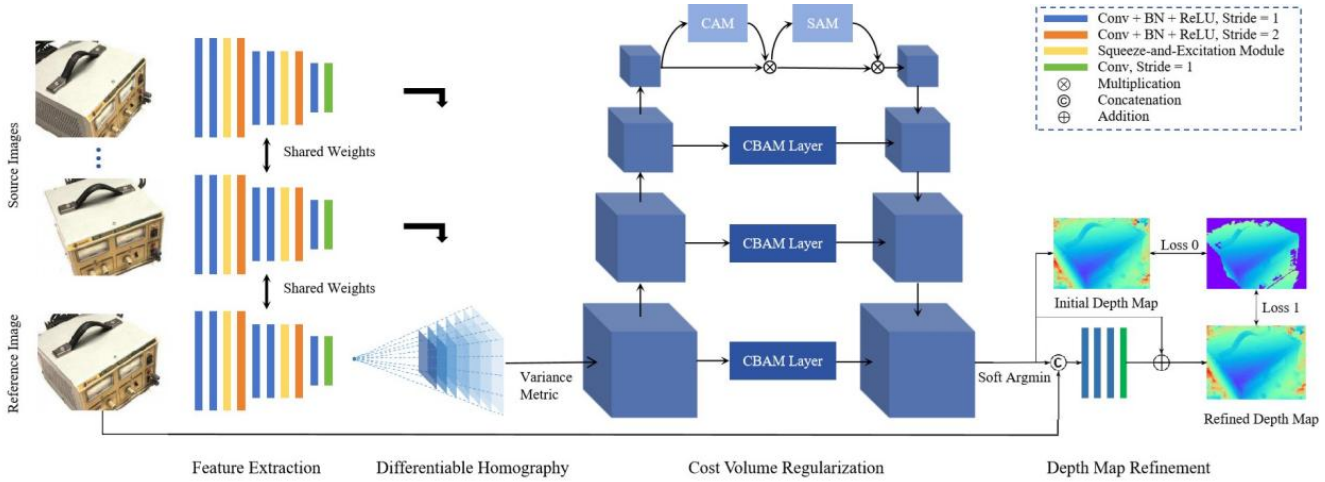

Figure 5: The overview of the proposed network.

N, and the corresponding camera intrinsics, rotations, and translations as $\{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}_{i=1}^{N}$ for each feature map. The network employs a differentiable homography to warp feature maps into frontoparallel planes aligned with the reference camera, as shown in (1):

$$\mathbf{H}_i(d) = \mathbf{K}_i \cdot \mathbf{R}_i \cdot \left(\mathbf{I} - \frac{(\mathbf{t}_1 - \mathbf{t}_i) \cdot \mathbf{n}_1^T}{d} \right) \cdot \mathbf{R}_1^T \cdot \mathbf{K}_1^T. \tag{4}$$

$H_i(d)$ denotes the homography matrix between the i-th feature map and the reference feature map at depth d. The matrix is represented as a 3 × 3 matrix, where 'n_1' signifies the principal axis of the reference camera, facilitating the coordinate mapping.

The warping process, serving as a pivotal step in connecting 2D feature extraction to 3D regularization networks, is executed in a differentiable manner, facilitating the end-to-end training of depth map inference.

The aggregation of multiple feature volumes $\{\mathbf{V_i}\}_{i=1}^{\tilde{N}}$ into a single cost volume C follows the proposal of a variance-based cost metric M, facilitating N-view similarity measurement. The metric accounts for the input image dimensions (W, H), depth sample number (D), and feature map channel count (F) to define a mapping : $\underbrace{R^V \times \cdots \times R^V}_{N} \rightarrow R^V$, as shown in (2):

$$\mathbf{C} = \mathcal{M}(\mathbf{V}_1, \cdots, \mathbf{V}_N) = \frac{\sum_{i=1}^{N} (\mathbf{V}_i - \overline{\mathbf{V}_i})^2}{N} \tag{5}$$

Here, $\overline{V_l}$ denotes the average volume among all feature volumes, and the operations are performed element-wise. This metric design is rooted in the idea that each view should contribute equally to the matching cost without favoring the reference image. Diverging from prior work's mean operation[11 et al.], MVSNet adopts a 'variance' operation, offering explicit measurement of multi-view feature differences.

Additionally, recognizing the challenge in extracting intricate 3D features efficiently, the cost regularization network incorporates a convolutional block attention module (CBAM) adapted for three-dimensional feature extraction. This CBAM dynamically adjusts the weighting of feature quantities across both channel and spatial dimensions, enhancing the network's ability to capture complex features effectively.

Depth Map Rather than using a pixel-wise winner-take-all (argmax) approach[5 et al.], MVSNet computes the expectation value across depth dimensions using a probability weighted sum over all hypotheses, as shown in (3):

$$\mathbf{D} = \sum_{d=d_{min}}^{d_{max}} d \times \mathbf{P}(d) \tag{6}$$

P(d) is the probability estimation for all pixels at depth d. This operation, known as the soft argmin[17 et al.], enables continuous depth estimation and differentiability for training.

Despite the 3D CNN's strong regularization abilities, for inaccurately matched pixels, the probability distributions tend to scatter, preventing concentration into a single peak. To measure estimation quality, the method defines it as the probability that the ground truth depth falls within a small range near the estimation. By summing probabilities over the nearest depth hypotheses, this approach evaluates estimation quality, emphasizing outliers' thresholding control for better depth map filtering.

Leveraging the reference image's boundary information, MVSNet utilizes a depth residual learning network inspired by image matting algorithms[37 et al.]. This network incorporates the initial depth map and a resized reference image as a 4-channel input, processing it through convolutional layers to learn depth residuals. The refined depth map emerges after adding the learned depth residuals back to the initial depth map.

Loss Our training loss is determined by the mean absolute difference calculated between the estimated depth map and the ground truth depth map. We only consider those pixels with valid ground truth labels:

$$Loss = \sum_{p \in \mathbf{p}_{valid}} \underbrace{\|d(p) - \hat{d}_i(p)\|}_{Loss0} + \lambda \cdot \underbrace{\|d(p) - \hat{d}_r(p)\|_1}_{Loss1} \quad (7)$$

Where p_{valid} denotes the set of valid ground truth pixels, d(p) the ground truth depth value of pixel p, $\hat{d}_i(p)$ the initial depth estimation, $\hat{d}_i(p)$ the refined depth estimation.

Experiments

DTU Dataset

The DTU dataset (Jensen et al. 2014) is a large-scale dataset widely used for deep learning in 3D reconstruction and multi-view stereo reconstruction. It comprises multi-view image sequences and corresponding accurate 3D point cloud models for 128 indoor scenes. Each scene is captured using a fixed camera from 49 viewpoints under 7 different lighting conditions, resulting in RGB images with a resolution of 1200×1600 pixels. The camera viewpoints for all scenes in the DTU dataset surround the objects, sampled at fixed angular intervals, ensuring sufficient overlap in the acquired images for reconstruction.

Implementation Details In experiments, the DTU dataset is partitioned into training, validation, and test sets following the dataset division method of MVSNet(Yao et al. 2018b).

Training The depth sampling range for the experiment is 425mm 935mm and sampling frequency is set to 192 which indicates that each depth assumption represents 2.67mm. During the training phase, the input of the model is the 640 x 512 resolution images from DTU dataset, including one reference image and five source images(N=5) with the corresponding camera parameters.We use PyTorch for implementation and train the model with the Adam(Kingma and Ba 2014) optimizer for 20 epochs from a start learning rate of 0.001 on NVIDIA 4070 GPU, and the learning rate is divided by 2 at the 10th, 13th, and 16th epoch during training.

Evaluation We used images of the original resolution size and crop the images to 1600 × 1152 for the DTU evaluation. Other settings are consistent with the training process. Our model consumes 0.75s and 11.5G memory for the full-resolution DTU depth estimation. As for depth fusion, we us the fusion algorithm(Merrell et al. 2007) to integrate depth maps from different views to a unified point cloud representation.

Metrics

For point cloud evaluation, we follow the standard evaluation protocol as in MVSNet(Yao et al. 2018b). The accuracy, completeness and overall score of the reconstructed point clouds are adopted.

Accuracy Accuracy is measured as the distance from estimated point clouds to the ground truth ones in millimeter, which can be computed as:

$$\begin{aligned} \text{Acc} &= \frac{1}{|\mathrm{S}_1|} \sum_{\substack{\mathrm{x} \in \mathrm{S}_1 \\ \mathrm{y} \in \mathrm{S}_2}} \min||\mathrm{x} - \mathrm{y}||^2 \\ &= \frac{1}{|\mathrm{S}_1|} \sum_{\substack{\mathrm{x} \in \mathrm{S}_1 \\ \mathrm{y} \in \mathrm{S}_2}} \min(\|\mathrm{x} - \mathrm{y}\|^2) \end{aligned} \quad (8)$$

Where S_1 represents the set of all spatial points in the reconstructed 3D point cloud, S_2 represents the set of all spatial points in the ground truth point cloud.

Completeness Completeness is defined as the distance from ground truth point clouds to the estimated ones, which can be computed as:

$$\begin{aligned} \text{Comp} &= \frac{1}{|\mathrm{S}_1|} \sum_{\substack{\mathrm{x} \in \mathrm{S}_1 \\ \mathrm{y} \in \mathrm{S}_2}} \min\|\mathrm{y} - \mathrm{x}\|^2 \\ &= \frac{1}{|\mathrm{S}_1|} \sum_{\substack{\mathrm{x} \in \mathrm{S}_1 \\ \mathrm{y} \in \mathrm{S}_2}} \min(\|\mathrm{y} - \mathrm{x}\|^2) \end{aligned} \quad (9)$$

Overall Score The overall score is the average of accuracy and completeness, which is taken as the comprehensive evaluation metric.It can be computed as:

$$\text{Overall} = \frac{\text{Acc} + \text{Comp}}{2} \quad (10)$$

Performance on DTU Dataset The qualitative results are shown in Figure 6 and Figure 7. As shown in Figure 6, our proposed methods successfully predicts the depth map of the scans and filters the generated depth map through a probability map mask. Proposed method captures more details of the scenes and estimates the depth map significantly accurate, complete and smooth.

After fusing depth maps of different viewpoints to generate 3D point clouds, we use MeshLab to visualize the 3D models. As shown in Figure 7, we compare our results with baseline method MVSNet(Yao et al. 2018b). The 3D point clouds generated by our method is more complete, especially for the geometry structures of the subject. Meanwhile, the good performance on scans with drastic illumination changes and reflections also proves the robustness of our method.

For quantitative evaluation, we report accuracy, completeness and overall score by using official MATLAB codes (Jensen et al. 2014) as shown in table 1. Our approach outperforms the baseline method MVSNet in completeness and raises the accuracy metrics to a new altitude.

Ablation Studies

To verify the effectiveness of the introduced SE(Hu, Shen, and Sun 2018) and 3D-CBAM(Woo et al. 2018a) modules,

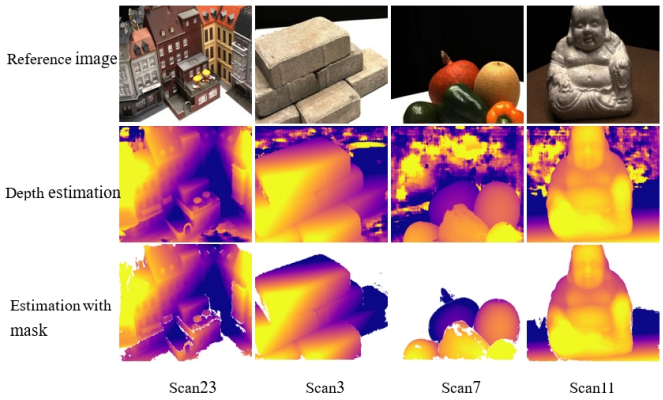

Figure 6: Estimated depth maps and depth maps with probability mask of scans 23,34,75,114.

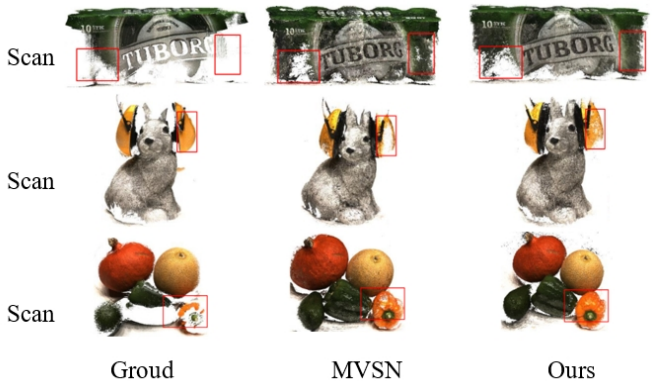

Figure 7: Visualized 3D models of scans 1, 3 and 75.

Table 1: Quantitative comparison of our method and MVSNet(D=192) on DTU dataset

Method	Acc(mm)	Comp (mm)	Overall (mm)
MVSNet	0.449	0.379	0.414
Ours	0.393	0.371	0.382

we first build a baseline model MVSNet with the depth hypotheses of 192 layer. Then, the SE and 3D-CBAM modules are added one by one.

The squeeze and excitation mechanism allows the model to pay more attention to the features of important channels and to extract rich 2D image features. As shown in table 2, the introduction of SE module has improved the accuracy and overall score of the model. The 3D-CBAM module also has a greater improvement on the model's performance because it has a stronger ability to extracts complex 3D features by selectively emphasizing relevant spatial and channel-wise information through its Channel Attention Module (CAM) and Spatial Attention Module (SAM).

Table 2: Ablation study on the DTU dataset. Components are added one by one in the upper part.

Method	Acc(mm)	Comp (mm)	Overall (mm)
Baseline	0.449	0.379	0.414
+SE	0.426	0.382	0.404
+3D-CBAM	0.393	0.371	0.382

In conclusion, both SE and CBAM modules contribute to the improvement of model performance. They enable the model to extract more comprehensive geometric information, resulting in a more complete reconstruction of the 3D point cloud.

Conclusion

In this paper, we propose an learning based end-to-end multi-view depth estimation architecture for 3D reconstruction. Specifically, we introduce the SE module into the feature extraction network of MVSNet to adaptively adjusting the channel weights of the feature map and achieving more effective feature extraction. In addition, to bridges the 2D feature extraction and 3D cost regularization,we encode the camera parameters as the differentiable homography to build the cost volume upon the reference camera frustum. Due to the original MVSNet is difficult to extract rich and complex 3D features in the cost volume, in cost regularization network, we extend the application of the CBAM module to 3D feature extraction, which can dynamically adjusts the weights of the feature volume in both channel and spatial dimensions. It has been demonstrated on DTU dataset that our model can reconstruct more complete and smoother 3D point cloud and achieves better performance than MVSNet.

References

De Bonet, J. S.; and Viola, P. 1999. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, volume 2, 2. Citeseer.

Esteban, C. H.; and Schmitt, F. 2004. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding*, 96(3): 367–392.

Fua, P.; and Leclerc, Y. G. 1995. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16(1): 35–56.

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.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Hirschmuller, H. 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2): 328–341.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanaes, H. 2014. Large Scale Multi-view Stereopsis Evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lhuillier, M.; and Quan, L. 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3): 418–433.

Merrell, P.; Akbarzadeh, A.; Wang, L.; Mordohai, P.; Frahm, J.-M.; Yang, R.; Nister, D.; and Pollefeys, M. 2007. Real-time visibility-based fusion of depth maps. In *International Conference on Computer Vision (ICCV)*.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Sinha, S. N.; Mordohai, P.; and Pollefeys, M. 2007. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *2007 IEEE 11th international conference on computer vision*, 1–8. IEEE.

Sun, J.; Xie, Y.; Chen, L.; Zhou, X.; and Bao, H. 2021. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15598–15607.

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: 903–920.

Woo, S.; Park, J.; Lee, J.-Y.; and et al. 2018a. CBAM: Convolutional Block Attention Module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018b. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

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. 2018a. 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.; Fang, T.; and Quan, L. 2018b. MVS-Net: Depth Inference for Unstructured Multi-View Stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Zhang, Z.; Peng, R.; Hu, Y.; and Wang, R. 2023. GeoMVS-Net: Learning Multi-View Stereo With Geometry Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21508–21518.