*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 14 Issue 3 ǁ March 2026 ǁ PP. 99-102*

# EfficientNet-Based Efficient Feature Extraction for Multi-View 3D Reconstruction

Houqi Lv Fujia Sun, Ping Liu, Wenxuan Song
*School of Materials and Chemistry, University of Shanghai for Science and Technology*
*Corresponding Author:Ping Liu*

***Abstract***
*This paper proposes an EfficientNet-based feature extraction framework for multi-view 3D reconstruction to improve matching robustness in weak-texture and illumination-varying scenes. EfficientNet-B0 is used to generate multi-scale deep features, and cosine-similarity-based multi-view aggregation is adopted to evaluate cross-view consistency for candidate 3D points. Experiments on the DTU benchmark show that the proposed method achieves an Overall error of 0.343 mm on point-cloud evaluation, outperforming representative baselines such as MVJTTSR (0.346 mm) and MVSNet (0.462 mm). Runtime analysis further indicates that EfficientNet-B0 requires only 5.3M parameters and 0.38G FLOPs with 26.0 ms inference time, compared with 25.6M/4.1G/48.0 ms for ResNet-50 and 11.7M/1.8G/31.0 ms for ResNet-18. These results demonstrate that the method provides a favorable balance between reconstruction quality and computational efficiency for practical multi-view reconstruction tasks.*

***Keywords:*** *Multi-view 3D reconstruction, EfficientNet-B0, deep feature extraction, feature matching, DTU benchmark, point cloud evaluation.*

---

Date of Submission: 05-03-2026 Date of Acceptance: 15-03-2026

---

## I. Introduction

Three-dimensional reconstruction plays an important role in many fields such as computer vision, medical imaging, virtual reality, and industrial inspection. By reconstructing the geometric structure of objects from multi-view images, it is possible to obtain accurate spatial information that can be used for measurement, navigation, simulation, and visualization. Among various reconstruction approaches, multi-view stereo (MVS) methods are widely used because they can generate dense point clouds and detailed surface models from multiple images captured at different viewpoints.

In traditional multi-view reconstruction pipelines, feature extraction and matching are fundamental steps that directly affect the accuracy and completeness of the reconstructed 3D model. Many classical methods rely on low-level image features such as RGB patches and similarity metrics like normalized cross-correlation (ZNCC). Although these approaches perform reasonably well in scenes with sufficient texture and stable illumination, they often fail in weak-texture regions, repetitive patterns, or areas with strong occlusions and lighting variations. As a result, the reconstructed point clouds may become sparse, contain mismatches, or lose important geometric structures.

In recent years, deep learning has shown great potential in feature representation learning. Compared with traditional hand-crafted features, deep neural networks are capable of extracting more discriminative and robust representations from images. These deep features can capture higher-level semantic information and contextual relationships, which significantly improves the reliability of feature matching in complex scenes. Therefore, incorporating deep learning-based feature extraction into the 3D reconstruction pipeline has become an effective strategy for improving reconstruction quality.

Among various deep convolutional neural network architectures, EfficientNet[1] has attracted considerable attention due to its high efficiency and strong feature representation capability. EfficientNet introduces a compound scaling strategy that simultaneously balances network depth, width, and resolution using a unified scaling coefficient. This design allows the network to achieve better performance while maintaining relatively low computational cost. In addition, EfficientNet employs mobile inverted bottleneck convolution (MBConv) modules and squeeze-and-excitation(SE) attention mechanisms to enhance feature extraction capability while keeping the model lightweight. These characteristics make EfficientNet particularly suitable for processing high-resolution images in large-scale multi-view reconstruction tasks.

Motivated by these advantages, this paper introduces EfficientNet as the backbone network for feature extraction in multi-view 3D reconstruction. Each input image is fed into the EfficientNet network to obtain multi-scale feature maps, which provide more discriminative representations than traditional pixel-level

features. For a candidate 3D point in space, its projections on different views are first computed using the camera projection model. The corresponding deep feature vectors are then sampled from the EfficientNet feature maps, and cosine similarity is used to measure the consistency of features across multiple views. By aggregating these similarity scores from different images, a confidence score for each candidate 3D point can be obtained.

The main idea of the proposed method is to replace traditional RGB patch features with deep multi-scale features extracted by EfficientNet. This strategy improves the robustness of feature matching in challenging scenarios such as weak-texture regions and complex illumination conditions. As a result, the density and accuracy of the reconstructed point clouds can be significantly improved. The proposed feature extraction framework is also computationally efficient and suitable for large-scale reconstruction tasks.

## II. Result And Discussion

### 2.1Experimental Setup

To evaluate the effectiveness of the proposed EfficientNet-based feature extraction method, benchmark experiments were conducted on the DTU evaluation dataset [2]. The DTU dataset contains 128 indoor scenes, each captured along a fixed camera trajectory under seven different lighting conditions. Each scene includes 49 images and the corresponding calibrated camera parameters. Following previous methods, the DTU dataset was divided into training, validation, and evaluation sets. Our method was trained on the DTU training set and evaluated on the DTU benchmark.

The proposed EGC-MVSNet was implemented using the PyTorch deep learning framework and trained and tested on public multi-view 3D reconstruction datasets, including DTU and BlendedMVS [3]. All input images were resized to 640 × 512. During training, three images from different viewpoints were randomly sampled in each group. EfficientNet-B0 [36] was used as the backbone network for feature extraction to achieve efficient multi-scale feature fusion. In addition, image normalization and data augmentation were applied before network input.

Due to GPU memory limitations, the batch size was set to 1. The Adam optimizer was adopted, and the initial learning rate was set to 0.001. When the validation loss stopped decreasing, the learning rate was dynamically decayed. The training process was performed for 30 epochs on a workstation equipped with a GeForce RTX 4060 Ti GPU with 8 GB memory.

In each training iteration, a confidence score for each candidate 3D point was first computed based on multi-view feature similarity, and a dynamic threshold was then applied to filter noisy points. The selected high-confidence points were used for mesh reconstruction. During the mesh connection stage, feature-map edge detection and normal consistency constraints were incorporated to suppress false edge connections effectively. The total loss was defined as the weighted sum of the depth L1 loss and the normal consistency loss, with weights of 1.0 and 0.5, respectively.

### 2.2Reconstruction Performance

The experimental results show that the proposed EfficientNet-based feature extraction method improves the robustness of feature matching in multi-view reconstruction tasks. Compared with traditional RGB patch matching methods, the deep features extracted by EfficientNet provide more discriminative representations and better adaptability to complex scene conditions.

In weak-texture regions where traditional patch-based methods usually fail to establish reliable correspondences, the EfficientNet features still maintain strong matching consistency. As a result, more valid correspondences can be obtained, which leads to denser reconstructed point clouds. In addition, the proposed method also reduces mismatches caused by repetitive textures and illumination variations.

From the reconstruction results, it can be observed that the point clouds generated using the proposed method exhibit higher completeness and smoother geometric structures. The improvement is particularly evident in areas with low texture or partial occlusion. These results indicate that deep feature representations extracted by EfficientNet significantly enhance the reliability of multi-view feature matching.

**Table 1: Quantitative result of the point cloud on the test set of DTU.**

| Method | Acc.(mm) | Comp.(mm) | Overall(mm) |
|---|---|---|---|
| Furu[4] | 0.613 | 0.941 | 0.777 |
| Gipuma[5] | 0.283 | 0.873 | 0.578 |
| COLMAP[6] | 0.400 | 0.664 | 0.532 |
| MVSNet[7] | 0.396 | 0.527 | 0.462 |
| R-MVSNet[8] | 0.385 | 0.459 | 0.422 |
| AA-RMVSNet[9] | 0.376 | 0.339 | 0.357 |
| Point-MVSNet[10] | 0.342 | 0.411 | 0.376 |
| Vis-MVSNet[11] | 0.369 | 0.361 | 0.365 |
| PatchmatchNet[12] | 0.427 | 0.277 | 0.352 |

| | | | |
|---|---|---|---|
| EPP-MVSNet[13] | 0.413 | 0.296 | 0.355 |
| CasMVSNet[14] | 0.325 | 0.385 | 0.355 |
| MVJTTSR[15] | 0.354 | 0.338 | 0.346 |
| MG-MVSNET[16] | 0.358 | 0.338 | 0.348 |
| GPGMVSNet[17] | 0.399 | 0.316 | 0.357 |
| TransMVSNet[18] | 0.385 | 0.329 | 0.357 |
| ours | 0.373 | 0.312 | 0.343 |

**Fig 6.Qualitative results of the point clouds on the DTU dataset**

**Table 2Runtime comparison of different backbone networks under the same experimental setting..**

| Backbone | Parameters (M) | FLOP (G) | Inference Time (ms) |
|---|---|---|---|
| ResNet-50 | 25.6 | 4.1 | 48.0 |
| ResNet-18 | 11.7 | 1.8 | 31.0 |
| EfficientNet-B0 | 5.3 | 0.38 | 26.0 |

Parameters and FLOPs are reported under the standard 224 × 224 input setting for backbone reference, while the inference time corresponds to the average forward-pass runtime of the neural network per sample in our implementation (including feature extraction and depth estimation, but excluding post-processing steps such as mesh reconstruction)

**2.3Discussion**

The performance improvement achieved by the proposed method can be attributed to several factors. First, EfficientNet adopts a compound scaling strategy that balances network depth, width, and input resolution, which enables the network to learn more expressive feature representations while maintaining computational efficiency.

Second, the MBConv structure used in EfficientNet integrates depthwise convolution and squeeze-and-excitation attention mechanisms, which enhances the ability of the network to capture important feature information. This design allows the extracted features to be more robust to changes in illumination, viewpoint, and texture distribution.

Third, the use of cosine similarity for feature matching provides a stable measure of similarity in the feature space. Compared with traditional intensity-based similarity measures, cosine similarity between deep features is less sensitive to local intensity changes and noise.

Overall, the experimental results demonstrate that integrating EfficientNet into the feature extraction stage of multi-view reconstruction can effectively improve matching robustness and reconstruction quality. The proposed approach provides a practical solution for improving 3D reconstruction performance in complex real-world scenes.

## III. Conclusion

In this paper, an EfficientNet-based feature extraction method for multi-view 3D reconstruction is presented and validated on the DTU benchmark. Traditional reconstruction methods mainly rely on low-level image features and often degrade in weak-texture or illumination-varying regions, while our approach introduces EfficientNet-B0 to extract robust multi-scale deep features for cross-view matching.

Quantitative results in Table 1 show that our method achieves an Overall error of 0.343 mm, which is the best among all listed methods (better than 0.346 mm for MVJTTSR and 0.462 mm for MVSNet). Although

some methods obtain lower single metrics in Acc. or Comp., our approach provides the best overall balance between accuracy and completeness, confirming more reliable reconstruction quality.

Runtime statistics in Table 2 further demonstrate efficiency: EfficientNet-B0 uses only 5.3M parameters and 0.38G FLOPs with 26.0 ms inference time, compared with 25.6M/4.1G/48.0 ms for ResNet-50 and 11.7M/1.8G/31.0 ms for ResNet-18. These results indicate that the proposed method improves reconstruction quality while maintaining a lightweight and fast inference pipeline suitable for practical multi-view 3D reconstruction.

In future work, the proposed method can be further extended by integrating geometric consistency constraints and depth estimation networks to improve reconstruction accuracy. In addition, the framework may also be applied to other reconstruction scenarios such as medical image reconstruction and endoscopic 3D reconstruction..

## References


[1]. Tan, M. and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. in International conference on machine learning, 2019. PMLR.

[2]. Jensen, R., et al. Large Scale Multi-view Stereopsis Evaluation. in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[3]. Yao, Y., et al. Mvsnet: Depth inference for unstructured multi-view stereo. in Proceedings of the European conference on computer vision (ECCV), 2018.

[4]. Furukawa, Y. and J. Ponce Accurate, Dense, and Robust Multiview Stereopsis IEEE. Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362-1376.

[5]. Yao, Y., et al. MVSNet: Depth Inference for Unstructured Multi-view Stereo. Computer Vision - ECCV 2018, Springer, 2018.

[6]. Gu, X., et al. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[7]. Wei, Z., et al. AA-RMVSNet: Adaptive Aggregation Recurrent Multi-View Stereo Network. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[8]. Schnberger, J.L., et al. ?Pixelwise View Selection for Unstructured Multi-View Stereo. Springer International Publishing, 2016.

[9]. Zhang, J., et al. Vis-MVSNet: Visibility-Aware Multi-view Stereo Network. International Journal of Computer Vision, 2022, 131(1): 199-214.

[10]. Ma, X., et al. EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo. 2021: 5712-5720.

[11]. Wang, F., et al. PatchmatchNet: Learned Multi-View Patchmatch Stereo. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[12]. Galliani, S., K. Lasinger, and K. Schindler Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[13]. Yao, Y., et al. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. 2019.

[14]. Chen, R., et al. Point-Based Multi-View Stereo Network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[15]. Ling, S., et al. Multi-View Jujube Tree Trunks Stereo Reconstruction Based on UAV Remote Sensing Imaging Acquisition System. 2024, 14(4): 1364.

[16]. Zhang, X., et al. MG-MVSNet: Multiple Granularities Feature Fusion Network for Multi-View Stereo. Neurocomputing, 2023, 528(C): 35-47.

[17]. Liu, L., et al. Geometric Prior-Guided Self-Supervised Learning for Multi-View Stereo. 2023, 15(8): 2109.

[18]. Ding, Y., et al. TransMVSNet: Global Context-Aware Multi-View Stereo Network with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.