Confident Multi-View Stereo

Xin Ma, Qiang Li, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*—Solving the Multi-View Stereo (MVS) problem is a cornerstone in computer vision, with depth map estimation and fusion being one of the most critical approaches. The depth confidence map is pivotal in ensuring the precision and completeness of the reconstruction outcomes. These algorithms frequently encounter a trade-off between completeness and accuracy in the confidence map, which can significantly impair the final reconstruction results. This paper analyzes the causes and phenomena of these issues, namely Confidence Jitter, Confidence Gap, and Confidence Disappearance. From these insights, a multi-view stereo network named CF-MVSNet is introduced, comprising three essential components. Firstly, the method mitigates the Confidence Jitter problem through two confidence fusion strategies. Secondly, it narrows the depth sampling space to near sub-pixel levels, addressing the Confidence Gap through neighborhood-average pooling. Lastly, the algorithm tackles the Confidence Disappearance problem resulting from multi-scale classification and regression with a loss function named CL. Our proposed method demonstrates superior performance across two critical metrics: the completeness of the depth map and the accuracy of the reconstructed point cloud, outperforming current state-of-the-art MVS methods.**

***Index Terms*—MVS, Depth Map, Point Cloud.**

I. Introduction

THE Multi-View Stereo matching task is pivotal in computer vision, entailing the generation of a dense point cloud from a scene captured by multiple images at varying perspectives, complemented by the intrinsic and extrinsic parameters of each image. This task has a wide range of applications [1] [2], including assisted driving, virtual reality [3], [4], historical building restoration [5], as well as indirect tasks like 3D object detection [6] and segmentation [7], and point cloud registration [8]. MVS is a crucial technology with broad prospects for application.

The MVS task encompasses a variety of approaches, each with its unique merits. The methods are primarily categorized into three domains: point cloud reconstruction, voxel-based reconstruction, and depth map-based reconstruction. Depth map-based reconstruction methods are distinguished by their ability to generate the depth map of the target view in a single computational pass. This characteristic ensures time efficiency, controllability, and reduced storage requirements. However, these methods have their limitations. They necessitate subsequent processing steps, such as point cloud triangulation and fusion, to achieve a complete and coherent 3D representation of the scene.

Xin Ma, Qiang Li, Yuan Yuna and Qi Wang are with the School of Artificial Intelligence, OPtics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. E-mails: machine@mail.nwpu.edu.cn, liqmges@gmail.com, y.yuan.ieee@nwpu.edu.cn, crabwq@gmail.com.

This work was supported by the National Natural Science Foundation of China under Grant U21B2041 and 62471394. Qi Wang is the corresponding author.

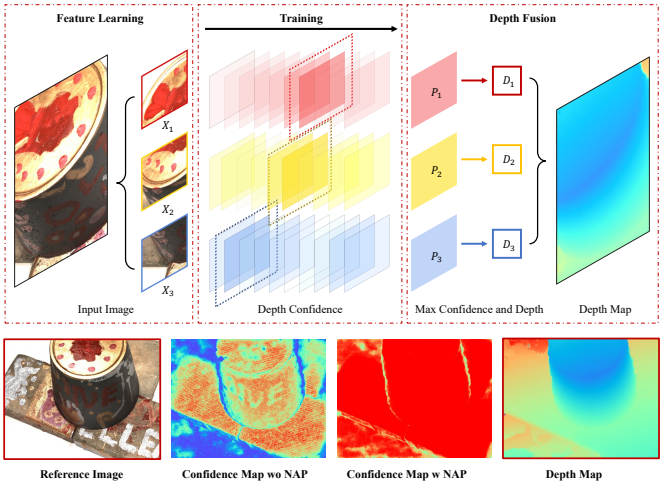


Fig. 1. The conceptual diagram illustrates our depth estimation method based on confidence fusion. In the diagram, the pixels processed in parallel within the same image are represented in red, yellow, and blue. The depth map in the center of the image shows varying shades to indicate the magnitude of confidence, with the dashed box indicating the selected maximum confidence. The middle two images among the four below depict the confidence maps before and after NAP processing, respectively.

From a training perspective, existing depth map-based MVS methods are commonly divided into two categories: those employing regression loss-based strategies and those utilizing classification loss-based approaches. Each category presents distinct advantages and challenges. Regression loss-based methods are known for their robustness in achieving depth completeness, yet they often fall short in accuracy, yielding a comprehensive but less precise outcome. Conversely, classification loss-based solutions excel in precision but are constrained by their limited depth completeness, producing accurate yet less exhaustive results.

Through simple scale expansion and sampling refinement, our proposed method based on CasMVSNet [9] classification loss achieves an accuracy of 0.172 on the DTU dataset, far surpassing geometric and deep learning-based MVS methods but exhibiting lower completeness. The network has learned the accurate depth projection patterns of shadow-free patches at a very early epoch. Subsequent epochs focus on learning challenging patches with different lighting conditions and repetitive textures.

Interpreting the experimental results, most regions across different epochs demonstrate consistent accuracy, but some areas exhibit varying confidence levels, referred to as Confidence Jitter. The comprehensive yet inaccurate issue is reflected in the confidence map, where most regions have extremely high confidence, leaving the task of handling incorrectly estimated areas to the point cloud fusion stage. The accurate but incomplete issue is evident in periodic confidence dips in regions

with accurate depth estimates, referred to as Confidence Gap. Additionally, the MVS method based on classification loss, during the process of narrowing the multi-scale depth range, directly discards pixels with incorrect estimates, leading to a further reduction in depth estimation completeness, referred to as Confidence Disappearance.

To address the abovementioned issues, we propose CF-MVSNet and optimize it from three aspects. Firstly, to tackle the Confidence Jitter issue, the paper suggests performing confidence fusion at different epoch levels and fine-tuning model levels, fully utilizing each stage of the model training process. Secondly, to address the Confidence Gap problem, we use multi-scale features to reduce depth sampling to 1/512 of the total sampling space, followed by neighborhood average pooling, obtaining an output that aligns the areas of accurate depth estimation with high confidence regions. Lastly, to tackle the Confidence Disappearance issue, we propose a joint loss function where the classification loss is primary and the regression loss is secondary. Compared with existing MVS models, the proposed method demonstrates significant improvements in the depth map's completeness and the accuracy of reconstructed point clouds.

The innovative contributions of this paper can be summarized as follows:

- Regarding the multi-scale depth estimation framework, we have identified and summarized three phenomena affecting the quality of reconstructed point clouds: Confidence Jitter, Confidence Gap, and Confidence Disappearance. These phenomena are characterized by fluctuations in depth confidence, discrepancies in depth estimation, and the complete absence of depth information. Furthermore, from an experimental perspective, we have analyzed the reasons behind these phenomena, pointing out the common deficiencies in this framework.
- To address the abovementioned phenomena, this paper proposes three effective solution strategies, including confidence fusion, neighborhood average pooling, and joint loss functions. The confidence fusion strategy involves both fusion during model training and fusion between fine-tuned models, employing a weighted ensemble learning approach to enhance the integrity of depth maps. Neighborhood average pooling explores the potential true confidence in Confidence Gap regions by jointly oscillating the probabilities in the depth direction for each pixel. The joint loss function combines primary and auxiliary losses to constrain network convergence while ensuring the iterative integrity of all pixel depth values.
- This paper embeds and combines the above three strategies into the CasMVSNet [9] model after structural fine-tuning named CF-MVSNet, achieving state-of-the-art results on the DTU and TAT datasets. It effectively suppresses the three phenomena mentioned above, validating the effectiveness of the proposed strategies.

II. Relate Works

The multi-view stereo matching task encompasses diverse reconstruction approaches, primarily categorized into three distinct representations: point cloud-based reconstruction [10], [11], voxel-oriented reconstruction [12], [13], and depth map-based reconstruction [9], [14]. The subsequent sections present an exhaustive overview, exploring geometric and learning models from various perspectives.

A. Geometric Models

The approach undertaken by Campbell *et al.* [15] to tackle outliers stemming from feature matching mismatches is ingenious, as it advocates for the concurrent retention of multiple depth values alongside the incorporation of pixel-wise uncertainty estimates. This strategy fortifies the robustness of the process against inaccuracies.

Furukawa *et al.* [10], on the other hand, introduces a groundbreaking solution that harnesses image patches. By partitioning the original imagery into uniform-sized patches, they facilitate depth diffusion and estimation at a more localized, patch-level granularity. This technique offers a fresh perspective on refining depth maps.

Drawing inspiration from PatchMatch [16], Zheng *et al.* [17] present a refined depth propagation method. Their approach streamlines the original slanted window assumption by diminishing the parameter count while enabling parallel propagation across rows and columns. Furthermore, the introduction of pixel-level visibility probability models for row-wise and column-wise visibility selection markedly elevates the reconstruction quality of the algorithm.

Galliani *et al.* [18] focus on enhancing the efficiency of depth propagation. They propose a red-black chessboard strategy that achieves pixel-level parallel depth propagation, thereby setting a primary objective to accelerate the process. This innovative approach underscores the ongoing pursuit of precision and speed in multi-view stereo matching.

B. Learning Models

Hartmann *et al.* [19] pioneered the concept that traditional MVS methods, rooted in geometric modeling, extend the computation of image patches across multiple perspectives akin to binocular views. They proposed enhancing the accuracy of patch matching by constructing a network that quantifies the similarity among multi-view patches. However, this advancement merely refines the traditional reconstruction process without a fundamental shift in the algorithmic framework.

Ji *et al.* [12] introduced an innovative end-to-end network framework, SurfaceNet, which utilizes a novel representation, CVC, to implicitly encode the geometric relationships between images and directly infer three-dimensional point clouds. Despite this, experimental results revealed a relatively high outlier error rate. In response, Huang *et al.* [20] and Yao *et al.* [14] independently proposed new frameworks, DeepMVS and MVSNet, respectively. DeepMVS emphasizes feature correlation during extraction, while MVSNet focuses on feature fusion and cost volume regularization. MVSNet [14] has since become the foundational framework for most MVS tasks.

Yao *et al.* [21] addressed the high GPU memory usage during MVSNet [14] training by proposing an RNN structure

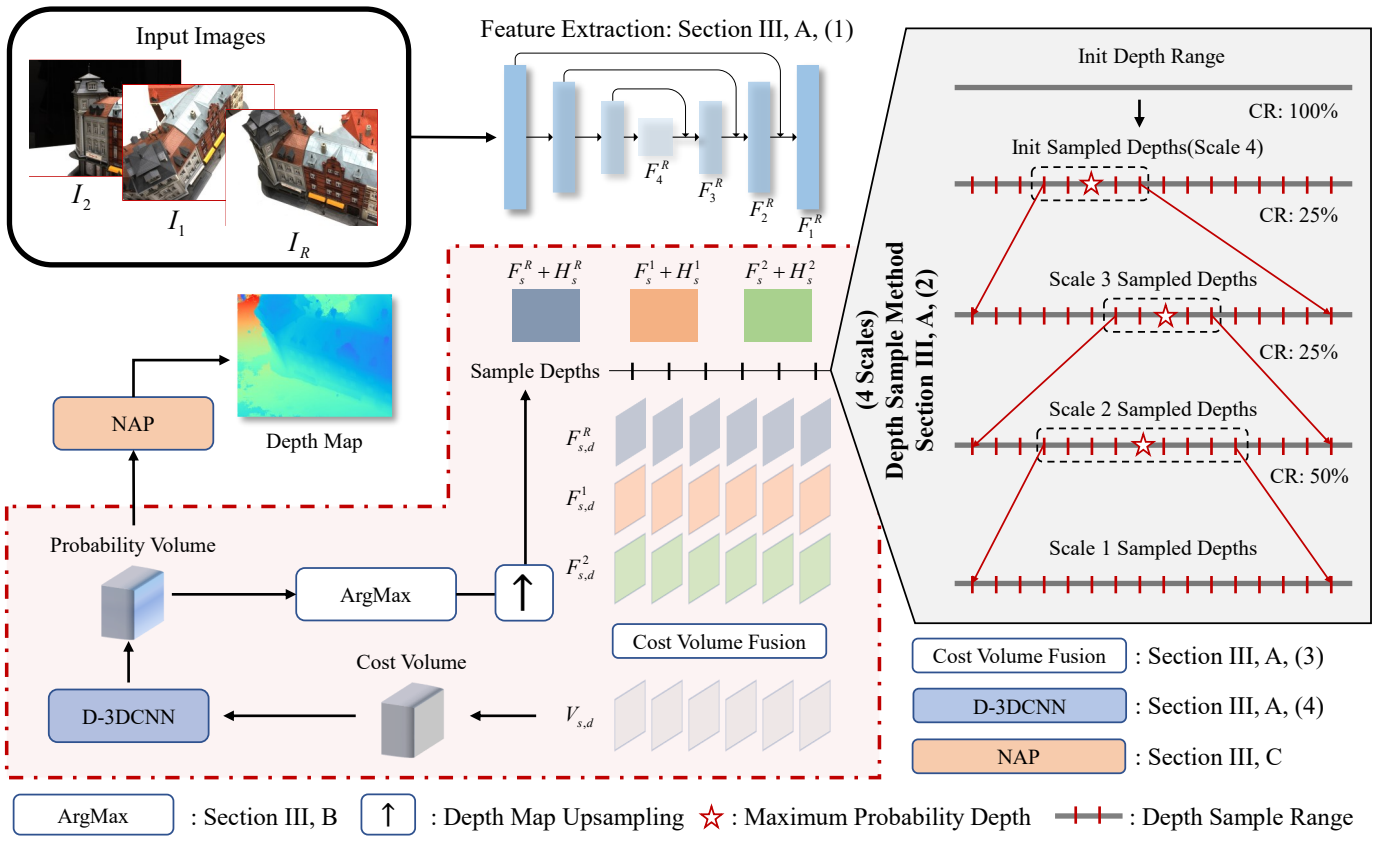


Fig. 2. The overall framework of CF-MVSNet. For ease of illustration, the number of input images is set to 3. Firstly, the multiple input images pass through a feature extraction network to obtain features at different scales. In order from 4 to 1 (as indicated by the subscript s in the diagram), features at a certain scale from the three images are fused based on sampled depth for cost volume, probability volume generation, and depth map upsampling. Iterating through the four scales (i.e., following the process outlined by the red dashed lines in the diagram), the probability volume at scale 1 is output, followed by Neighborhood Average Pooling (NAP) to obtain the final depth map. The locally magnified image on the right illustrates the relationship between depth sampling at different scales.

to replace the 3D-U-Net, effectively trading computational time for memory space. They also introduced a classification-based loss for training, which, while improving reconstruction metrics, increased training time. Chen *et al.* [22] attributed the high GPU memory usage to direct training in the original image pixel space and proposed a solution that generates a smaller coarse depth map based on MVSNet [14], followed by iterative optimization in point cloud space, significantly enhancing reconstruction completeness.

Hou *et al.* [23] presented a solution employing temporal non-parametric fusion, integrating a Gaussian prior during the multi-view feature aggregation to discern potential distances between viewpoints. Xue *et al.* [24] and Luo *et al.* [25] proposed enhanced models, MVSCRF and P-MVSNet, by incorporating specific steps in the regularization module. Gu *et al.* [9] introduced CasMVSNet, which achieved significant performance improvements through multi-scale processing on the MVSNet [14] foundation, establishing a benchmark for many methods, including the basic framework of this paper.

Ding *et al.* [26] and Cao *et al.* [27] further advanced the field by incorporating Vision Transformer (ViT) structures into the feature extraction module of TransMVSNet and MVSFormer, respectively. This facilitated global modeling to capture representations and latent correlations in multi-view images. Mi *et al.* [28] proposed GBi-Net [28], utilizing a novel binary depth partitioning strategy and training model, achieving state-of-the-art metrics on the DTU dataset, despite some limitations in depth map completeness.

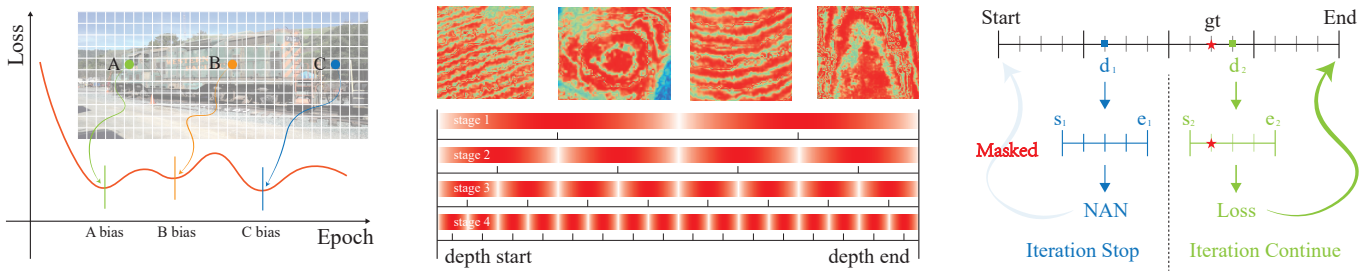
III. Proposed Method

A. Overall Framework

CF-MVSNet consists of two primary modules: a multi-view deep feature extraction network and a multi-scale cost volume fusion and depth map generation network, as shown in Figure 2. The following outlines the details of feature extraction, depth sampling, cost volume fusion, cost volume regularization, confidence, and depth map generation.

1) Feature Extraction: We use a U-Net network for feature extraction with 4 and 5 down-sampling steps, which means that the U-Net down-samples the size of the original images by factors of $1/2^4$ and $1/2^5$, respectively. The schematic diagram of the feature extraction process with four down-sampling steps is illustrated in the top left corner of Figure 2. For simplicity, the following description is based on the four down-sampling steps.

Given a reference image, R, for which the depth map needs to be estimated, the input to the feature extraction network



(a) Confidence Jitter (b) Confidence Gap (c) Confidence Disapperence

Fig. 3. The principle schematic diagram of three confidence issues. (a) Represents the Confidence Jitter problem, where different epochs exhibit varying degrees of confidence due to the diverse distribution characteristics across different locations within the same scene. (b) Illustrates the Confidence Gap issue in a multi-scale classification framework, wherein the confidence diminishes at the intermediate position between two sampling points, as evidenced by the four sub-figures above. (c) Depicts the Confidence Disappearance problem within the multi-scale classification framework, wherein some pixels in the sampling process fail to cover the true depth values and are consequently discarded using masks.

TABLE I
THE PROPOSED SAMPLING METHOD IN THIS ARTICLE IS COMPARED WITH THE DEPTH SAMPLING TECHNIQUES EMPLOYED IN THE MVSNET AND CASMVSNET MODELS.DN, CR, AND DS CORRESPOND TO THE NUMBER OF DEPTH SAMPLES, COVERAGE RATE OF THE CURRENT SCALE OVER THE PREVIOUS SCALE, AND THE STEP SIZE OF DEPTH SAMPLING RELATIVE TO THE ORIGINAL DEPTH SPACE, RESPECTIVELY. S1, S2, S3, AND S4 REPRESENT THE FOUR SCALE STAGES IN THE MULTISCALE APPROACH.

	MVSNet		**CasMVSNet**			**Ours**		
S4	256	D/256	48	100%	D/48	16	100%	D/16
S3	-	-	32	33.3%	D/96	16	25%	D/64
S2	-	-	8	12.5%	D/192	16	25%	D/256
S1	-	-	-	-	-	16	50%	D/512
	DN	**DS**	**DN**	**CR**	**DS**	**DN**	**CR**	**DS**

includes $N+1$ images. This process yields deep feature maps represented as:

$$F^n = \{F_1^n, F_2^n, F_3^n, F_4^n\}, \quad n \in \{0, \cdots, N\}, \tag{1}$$

where n denotes either the reference image R (where n = 0) or one of the source images indexed from 1 to N.

2) Depth Sampling: We introduce a multi-stage uniform sub-depth partitioning strategy that diverges from the depth sampling methodologies utilized in CasMVSNet [9]. Our objective is to preserve the continuity of the depth sampling range across stages while achieving pixel-level sub-depth estimation for each stage. As observed from the coverage rate in Table I, the method proposed herein demonstrates enhanced stability. Furthermore, the depth space step size introduced exhibits superior accuracy.

The depth partitioning strategy delineated in this paper is markedly different from that of the baseline model, CasMVSNet [9], with significant distinctions in several pivotal aspects. These distinctions encompass the number of scales employed (varying from S1 to S4), the uniformity in the quantity of depth samples (DN) at each scale, the consistency of the coverage rate (CR) across different scales, and the precision of the interval size in relation to the original space. A comparative analysis of these parameters is presented in Table I and depicted in Figure 2. Notably, our approach introduces an increased number of scales, ensures a more uniform distribution of DN across scales, maintains a more stable CR, and refines the interval size to more closely approximate the sub-pixel domain. For example, while the interval in the original space for the DTU dataset is set at 192, our refined intervals are specifically designed to augment the granularity of depth estimation.

3) Cost Volume Fusion: Employing the methodologies above for feature extraction and depth sampling, we get multi-scale deep features, denoted as F, for a comprehensive set of $N+1$ images. By the projection approach outlined by MVSNet [14], the deep features derived from the N source images are harmoniously integrated with the sampled depths through the employment of the projection matrix H. This strategic integration guarantees that the features are accurately mapped onto depths that correspond with the perspective of the reference image.

Following the projection transformation, the deep features from the source images are meticulously aligned to the viewpoint of the reference image at each depth level sampled. The convergence of these deep features—emanating from the $N+1$ unique viewpoints and pertinent to each sampled depth—results in constructing the cost volume V, which encapsulates the collective information for depth estimation.

4) Cost Volume Regularization: In a significant departure from the uniform treatment of the dimensions D, H, and W as employed in CasMVSNet [9], we introduce a novel D-3DCNN, a depth-biased cost volume regularization network crafted to retain essential depth dimension information. Contrary to conventional methods, our D-3DCNN maintains the integrity of the depth dimension during the down-sampling phase.

The 3DCNN architecture proposed by MVSNet utilizes convolution kernels of size (5,5,5) and strides of (2,2,2) for down-sampling across the (D, H, W) dimensions. Our approach diverges by acknowledging the inherent symmetry in the height (H) and width (W) dimensions, advocating for a distinct treatment of the depth (D) dimension. Consequently, the proposed D-3DCNN employs an innovative down-sampling strategy that preserves the depth dimension. This strategy leverages convolution kernels of size (1,5,5) with strides of (1,2,2), thereby ensuring the preservation of depth information throughout the network's processing stages. This innovative design, coupled with multi-scale uniform sampling, not only

bolsters performance but also streamlines the parameter count of the original network.

5) Confidence and Depth Map Generation: The cost volume V is fed into the D-3DCNN, a depth-biased cost volume regularization network that outputs a probability volume. This volume signifies the probability of each pixel's depth value corresponding to a specific depth. In this paper, we generate a confidence map for the depth prediction of the reference image, leveraging a fusion strategy. Additionally, we impose constraints on the network's convergence process by applying Confidence Loss function. For a comprehensive understanding, readers are directed to subsections B, C, and D for detailed discussions.

B. Confidence Jitter

From a machine learning standpoint, multi-view depth estimation can be categorized as either a pixel-wise classification or a regression problem. During training and inference, pixels are considered discrete entities, and the ideal scenario is to achieve accurate depth estimation for all pixels. However, the models are prone to biased confidence levels due to various factors, such as object movement during data capture or variations in lighting conditions. As illustrated in Figure 33(a), the biases in different pixel positions across epochs are influenced by complex environmental factors, leading to macroscopic oscillations in the absolute depth error of test scenes. On a microscopic level, certain regions may exhibit confidence and depth error fluctuations during the later stages of model training, particularly for challenging examples. Typically, epochs with higher confidence levels are associated with more minor depth errors, as shown in Figure 9. This phenomenon is termed as Confidence Jitter.

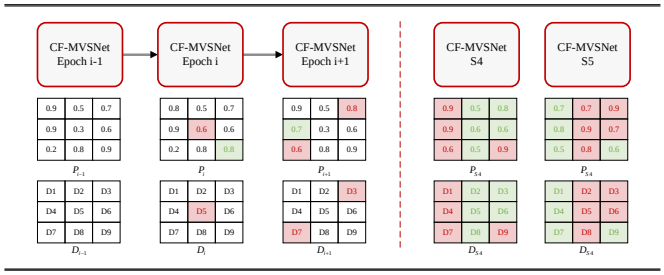


Fig. 4. Confidence fusion EF at the epoch level and SF at the model fine-tuning level. Red digits indicate pixels that need updating, green indicates pixels to be ignored, and black indicates pixels to remain consistent.

To capitalize on the strengths of different training epochs—where the conventional approach is to select the epoch with the lowest loss—and to harness the fine-tuning characteristics of models, we propose a weighted ensemble learning strategy known as "Confidence Fusion." This method assigns weights to various output results based on their confidence values. The process of confidence fusion is divided into two primary modules, Epoch Fusion (EF) and Scale Fusion (SF), designed to enhance confidence and depth information for both unambiguous and contentious pixel regions. The detailed procedure is depicted in Figure 4.

In the EF module, the network is trained to output the confidence P and the corresponding depth map D of the target image as the reference output starting from a specific intermediate epoch. If the confidence in subsequent epochs surpasses that of the reference output P to particular pixels, the confidence and depth of those pixels in the reference image are updated accordingly. This iterative process continues until the final training epoch is reached.

$$
\begin{aligned}
P^t(x) &= \begin{cases} P^{i+1}(x), \text{ if } P^{i+1}(x) > P^t(x), \\ P^t(x), \text{ if } P^{i+1}(x) \leq P^t(x), \end{cases} \\
D^t(x) &= \begin{cases} D^{i+1}(x), \text{ if } P^{i+1}(x) > P^t(x), \\ D^t(x), \text{ if } P^{i+1}(x) \leq P^t(x). \end{cases}
\end{aligned} \tag{2}
$$

In contrast to EF, which integrates information across multiple training epochs, SF focuses on confidence fusion from distinct trained models. The S4 and S5 versions presented in this paper yield confidence values P_{S4}, P_{S5}, and their respective depth maps D_{S4}, D_{S5} post EF fusion. During the SF process, a pixel-wise comparison is conducted between P_{S4} and P_{S5}, and for each pixel, the higher confidence value, along with its corresponding depth, is chosen. This selection culminates in the generation of the final confidence and depth maps.

$$
\begin{aligned}
P(x) &= \max\left(P_{s4}(x), P_{s5}(x)\right), \\
D(x) &= \text{gather}\left(D_{s4}(x), D_{s5}(x) \mid P(x)\right).
\end{aligned} \tag{3}
$$

The experimental outcomes demonstrate that the fused results exhibit enhanced integrity in confidence and improved accuracy in the depth mapping compared to their pre-fusion counterparts. Please refer to Figures 8 and 9 for an experimental comparison.

C. Confidence Gap

Compared to MVSNet [14], CasMVSNet [9] has notably reduced the computational complexity of single-stage depth estimation by incorporating a multi-scale network architecture. This approach enhances the overall stability of the algorithm by progressively narrowing the depth range at each sampling step. However, this hierarchical reduction introduces a significant challenge: pixels between two depth samples exhibit oscillations in their confidence values. Figure 3(b) visually represents this phenomenon, where simplified depth vertical black lines denote samples, and the red-white transition bars signify the confidence values for pixels at the current depth stage, with white indicating areas of uncertainty. Empirical experimental results corroborate this theoretical observation.

The experimental results reveal that the confidence maps display distinct bar-like structures along the depth axis at specific depths, termed the Confidence Gap. Ideally, these structures should exhibit a uniform distribution. However, the upper examples in Figure 3(b) deviate from this theoretical expectation. This discrepancy arises from the multi-scale framework's depth resampling process, including pixel-space up-sampling, causing post up-sampled depth values to deviate from the initial depth samples, resulting in perturbed bar-like structures.

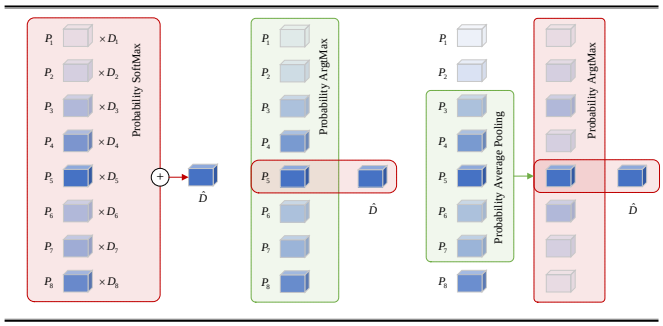


Fig. 5. Comparison of neighborhood average pooling method with two other common techniques. The red and blue regions represent the depth ranges involved in each step of the operation.

To address this issue, we analyzed the regions significantly impacted by the Confidence Gap in the model's output. We compared them with the ground truth depth values, as depicted in Figure 10. Upon meticulous examination, we observed that Confidence Gaps frequently occur at the boundaries of depth pre-sampling. Nonetheless, the depth estimates at these boundaries meet the criteria for 3D reconstruction. The presence of these Confidence Gaps suggests that the probability values along the depth direction for each pixel are fluctuating. Consequently, by considering the depth probability values in the vicinity of each pixel, the reliability of these Confidence Gaps can be substantially enhanced. In light of this, we integrated existing softmax and argmax algorithms with a novel module called Neighborhood Average Pooling (NAP), which fortifies the integrity of the confidence map while ensuring the precision of depth estimation.

Specifically, we initiate by performing average pooling on the probability volume of the final stage along the depth direction for each pixel, followed by the application of the argmax algorithm to the pooled probability volume. The radius of the pooling operation delineates the neighborhood scope for each pixel, which is presumed to be the range of probability fluctuations for that pixel. Empirical comparisons have indicated that a pooling radius of 5 yields the most favorable outcomes. A comparative analysis with the other two methods is presented in Figure 5, where the left and middle subplots represent the softmax and argmax algorithms converting the probability volume into depth values, respectively, and the right subplot illustrates the NAP approach. MVSNet [14] shares a similar design philosophy but primarily addresses the multi-modalities in the depth direction with softmax. The experimental results demonstrate that the NAP module significantly mitigates the Confidence Gap phenomenon. Furthermore, the NAP module is universally applicable to all multi-scale MVSNet architectures.

D. Confidence Disappearance and Loss Function

The phenomenon known as Confidence Disappearance occurs in the context of multi-stage depth estimation, where a pixel is inadvertently excluded from subsequent loss calculations due to an erroneous depth estimation at an intermediate stage. This exclusion happens when the revised depth range fails to include the actual depth value, effectively discarding the pixel. This exclusion compromises the completeness of the depth estimation. Figure 3(c) illustrates this principle, with the left blue region showing a pixel range that is prematurely terminated from further iterations because the adjusted range does not include the true depth value. In contrast, the right side illustrates a normal iteration process. It is important to note that subsequent process stages may possess enhanced features capable of accurately identifying and estimating the depth of such pixels.

To counteract the above issue, we introduce a novel joint loss function, denoted as the Confidence Loss (CL), which integrates both the cross-entropy loss (L_c) and the error classification regression loss (L_r). The formulation of this joint loss function is delineated as follows:

1) Class Loss L_c: In alignment with other classification-driven depth estimation methodologies, the cornerstone of our loss function is the cross-entropy loss applied to pixel-wise probability distributions. The formula for this loss is articulated as follows:

$$L_c = -\sum_{s=1}^{4}\sum_{x\in\Omega_1}\sum_{d=1}^{D} y_d \cdot \log\left(P_d^x\right), \tag{4}$$

In this equation, s denotes the stage within the CF-MVSNet framework. x signifies the effective pixels that fall within the current sampling bounds. d is the index for depth sampling. y_d is a binary indicator, representing the true label that signifies whether the pixel at x is associated with depth d (1 for association, 0 otherwise). P_d^x is the predicted probability that the pixel x corresponds to the depth value d.

2) Regre Loss L_r: For pixels at each stage where the depth range estimation is not precise, our approach employs a direct computation of loss based on the discrepancy between the estimated and true depths. The formula for this loss is delineated as follows:

$$L_r = \sum_{s=1}^{4}\sum_{x\in\Omega_2} |\hat{D}(x) - D(x)|, \tag{5}$$

Here, s represents the stage in the multi-scale network. x refers to the pixels with inaccurate depth range estimation. $\hat{D}(x)$ and $D(x)$ are the predicted and actual depth values for the pixel x, respectively.

3) Joint Loss CL: The final CL joint loss is derived by amalgamating the L_c and L_r losses, moderated by a weighting coefficient λ, as expressed by the following equation:

$$CL = L_c + \lambda \times L_r, \tag{6}$$

In our experiments, λ is empirically set to 0.01. The first term of the CL function serves as the primary constraint, driving the model's convergence, while the second term complements this by ensuring the iterative integrity across all pixels.

TABLE II
THE PROPOSED METHOD IS COMPARED WITH STATE-OF-THE-ART (SOTA) APPROACHES IN TERMS OF RECONSTRUCTION PERFORMANCE ON THE TAT DATASET. HERE, RED, GREEN, AND YELLOW RESPECTIVELY INDICATE THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS.

Methods	Mean	Family	Francis	Horse	Lightous	M60	Panther	Playground	Train
Colmap [32]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
R-MVSNet [21]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
Point-MVSNet [22]	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
ACMH [33]	54.82	69.99	49.45	45.12	59.04	52.64	52.37	58.34	51.61
P-MVSNet [25]	55.63	70.04	44.64	40.22	65.20	55.08	55.17	60.37	54.29
MVSNet [14]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
CasMVSNet [9]	56.43	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
Ours	59.78	76.89	63.52	50.62	58.75	59.10	54.31	59.70	55.31

IV. EXPERIMENTS

This section introduces the dataset used to validate the proposed method CF-MVSNet, provides quantitative and qualitative comparisons with existing methods for depth estimation, and assesses point cloud reconstruction quality. Additionally, ablation experiments are conducted to validate the effectiveness of different modules.

A. Dataset

This study leverages three benchmark datasets for validation: DTU [29], BlendedMVS [30], and TanksAndTemples [31]. The DTU dataset, a cornerstone for comparative analysis, is a comprehensive collection specifically curated for 3D reconstruction tasks. It encompasses 128 diverse scenes, each exhibiting a distinct yet consistent scale. The meticulous construction of this dataset involves the utilization of controlled laboratory lighting conditions, ensuring a spectrum of seven varied illumination scenarios per scene. The robotic arm's precise motion programming allows for the accurate capture of camera viewpoints across all scenes. For uniformity in experimental design, the camera poses for each scene are sampled from a standardized set of positions. Our training set aligns with that of CasMVSNet [9], consisting of 79 scenes, each with 49 distinct viewpoints, while the remaining 22 scenes constitute the test set.

The BlendedMVS dataset stands as a substantial resource for 3D reconstruction, featuring 113 scenes with an extensive collection of over 17,000 viewpoints. It diverges from the DTU dataset by offering a broader range of scene content, from common items such as shoes to historical villages, and presents a more liberal sampling of viewpoints. In contrast to the controlled settings of DTU, BlendedMVS captures a wider variety of real-world conditions. The TanksAndTemples dataset further expands the scope with its inclusion of diverse elements such as statues, military tanks, and civic structures, providing a rich tapestry of scene types for evaluation.

B. Quantitative Comparison

This subsection presents a comparative analysis of the proposed CF-MVSNet method with existing geometric and learning-based methods on the DTU dataset, focusing on reconstruction accuracy, completeness, and average error. Additionally, the F-score on the TanksAndTemples dataset is examined, as detailed in Table III and Table II. Geometric

TABLE III
THE PROPOSED METHOD IS COMPARED WITH STATE-OF-THE-ART (SOTA) APPROACHES IN TERMS OF RECONSTRUCTION PERFORMANCE ON THE DTU DATASET. HERE, RED, GREEN, AND YELLOW RESPECTIVELY INDICATE THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS.

Methods		Acc.	Comp.	Average
Geo	**Gipuma [18]**	0.283	0.873	0.578
	Colmap [32]	0.400	0.664	0.532
Learning	**SurfaceNet [12]**	0.450	1.040	0.745
	MVSNet [14]	0.396	0.527	0.462
	P-MVSNet [25]	0.406	0.434	0.420
	R-MVSNet [21]	0.383	0.452	0.418
	MVSCRF [24]	0.371	0.426	0.399
	Point-MVSNet [22]	0.342	0.411	0.377
	CVP-MVSNet [34]	0.296	0.406	0.351
	CasMVSNet [9]	0.325	0.385	0.355
	MSCVP-MVSNet [35]	0.379	0.278	0.328
	ATLAS-MVSNet [36]	0.278	0.377	0.327
	UniMVSNet [37]	0.352	0.278	0.315
	TransMVSNet [26]	0.321	0.289	0.305
	GBi-Net [28]	0.327	0.268	0.298
Ours-acc		0.172	-	-
Ours		0.315	0.277	0.296

methods, exemplified by Colmap [17] and Gipuma [18], prioritize accuracy, whereas learning-based approaches emphasize completeness and average error.

Colmap [17] and Gipuma [18] are quintessential geometric reconstruction algorithms. Gipuma enhances the depth propagation approach of Colmap, achieving near pixel-wise parallelism but at the cost of reduced reconstruction accuracy. Within the domain of MVS algorithms, Gipuma is noted for its lower reconstruction accuracy. In contrast, our non-confidence fusion method denoted as Ours-acc in Table III, significantly improves reconstruction accuracy when completeness is not a factor. Upon incorporating confidence fusion within a single model, our proposed algorithm exhibits substantial reductions in reconstruction completeness and average error relative to Gipuma and Colmap, as illustrated in Table III.

Point-MVSNet [22], CVP-MVSNet [34], CasMVSNet [9], TransMVSNet [26], and GBi-Net [28] are methodologies that extend the foundational MVSNet [14] framework. Among these, CasMVSNet [9] is the baseline for our proposed method, which surpasses it in terms of reconstruction completeness and average error, as denoted by Ours in Table III. Moreover, our method achieves an accuracy that rivals the performance of existing learning-based models.

Regarding the reconstruction F-score metric from the intermediate subset of the TanksAndTemples, our proposed method

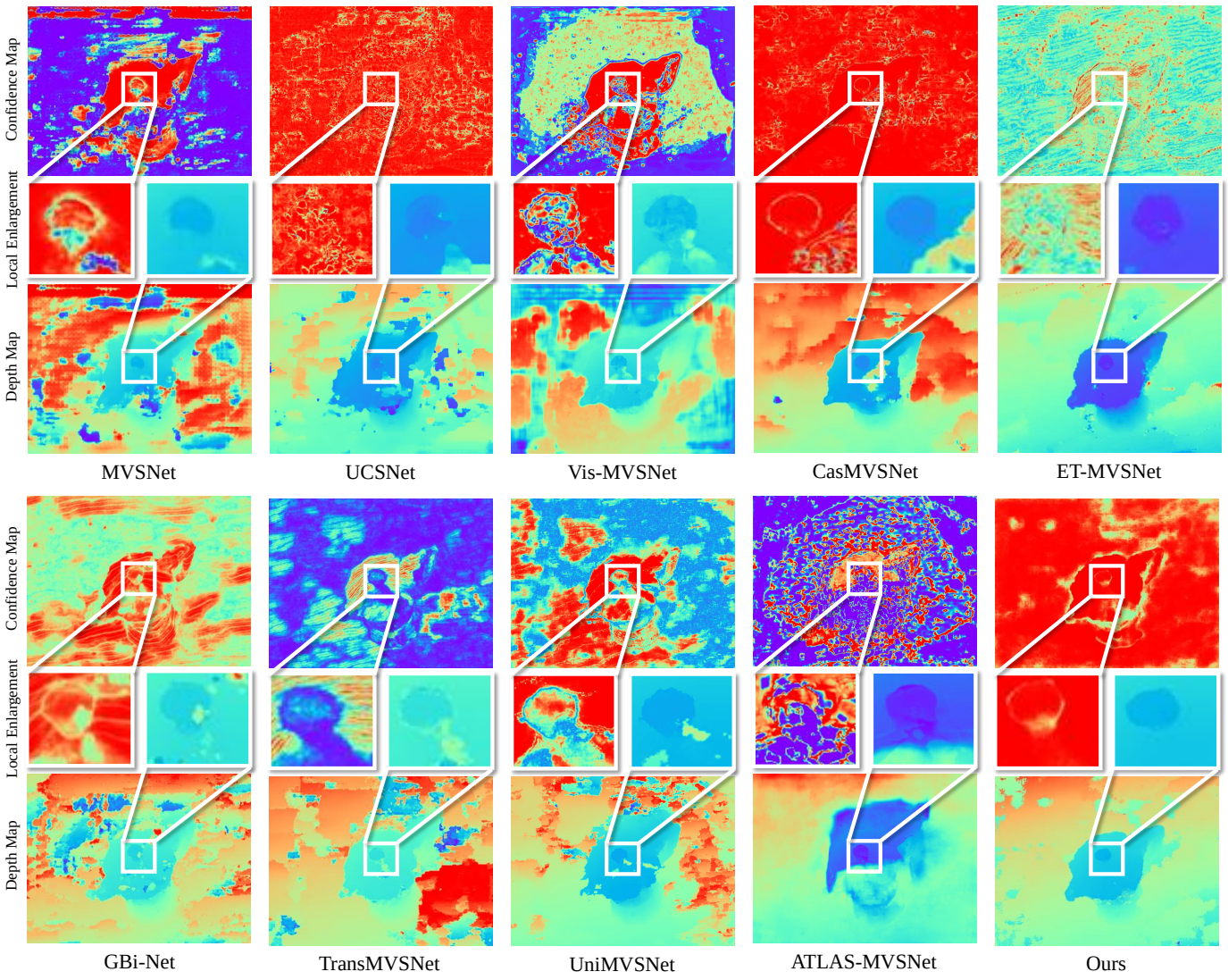


Fig. 6. The our proposed method is compared with existing state-of-the-art (SOTA) methods using depth maps and confidence maps. This includes classical approaches such as MVSNet [14], UCSNet [38], Vis-MVSNet [39], and CasMVSNet [9], as well as newer methods incorporating attention mechanisms such as ET-MVSNet [40], GBi-Net [28], TransMVSNet [26], UniMVSNet [37], and ATLAS-MVSNet [36]. The middle subfigures in the first and second columns represent representative local comparison subfigures.

secured the top position in five scenes, the runner-up spot in one scene, and ranked third in the remaining two scenes. Collectively, the proposed approach demonstrates superior performance across the board, as shown in Table II.

C. Qualitative Comparison

This section aims to substantiate the reliability of the depth estimation in our proposed method by juxtaposing it with other state-of-the-art MVS methods, focusing on the visualization of depth estimation and point cloud fusion. It is imperative to recognize that while the accuracy metrics for point cloud estimation presented in the preceding section offer clear insights, they may not entirely encapsulate the performance spectrum of various methods. There are two principal rationales for this limitation. Firstly, the completeness and accuracy of the depth map do not invariably correspond to the fidelity and comprehensiveness of the reconstructed point clouds. For instance, as delineated in Table II, GBi-Net [28] demonstrates commendable reconstruction accuracy, yet the depth map is marred by several conspicuous error patches. Secondly, the reconstructed point cloud frequently encompasses regions absent in the target point cloud, which can adversely affect the accuracy of the reconstructed point cloud. However, these unlabeled regions may constitute more exhaustive segments than the target point cloud. Consequently, a visual assessment of the effectiveness of different methods emerges as an essential adjunct to the quantitative metrics discussed in the previous section.

1) Depth Map and Confidence Map: In all reconstructed scenes, the presence of weakly textured areas and reflective materials poses a significant challenge for depth estimation tasks. A quintessential example of such a challenging scenario is the Scan77 scene from the DTU dataset. Consequently, we have conducted a comparative analysis between our proposed method and other state-of-the-art approaches, focusing on

the completeness of depth estimation and the reliability of confidence maps, with a particular emphasis on the Scan77 scene. This comparative evaluation is depicted in Figure 6.

MVSNet [14] is the progenitor of many subsequent methods, with CasMVSNet [9] further evolving it into a multi-scale framework. When scrutinizing the depth maps, the latter shows a marked improvement in overall accuracy over the former. Nevertheless, discrepancies are observed between regions with accurately estimated depths and areas of high confidence, which may diminish the reliability of the confidence map. Similar challenges are encountered in UCSNet [38] and Vis-MVSNet [39], which are contemporaneous with the aforementioned methods.

Emerging methodologies such as GBi-Net [28], Trans-MVSNet [26], ATLAS-MVSNet [36], and UniMVSNet [37] endeavor to bolster final performance by incorporating innovative feature learning strategies. However, these methods exhibit a decline in the accuracy of depth maps and the consistency of confidence maps. Although ET-MVSNet [40] achieves completeness in depth map estimation comparable to our proposed method, it experiences significant information loss in its confidence map, heavily depending on geometric consistency during the depth map fusion phase.

In conclusion, our proposed method is distinguished by a depth map of high completeness, and the congruence between areas of high confidence in the confidence map and regions of accurately estimated depth is notably superior. Collectively, our method showcases a distinct advantage in the quality of depth map estimation outcomes.

2) Point Cloud: In practice, when the accuracy of point cloud estimation dips below the 0.3 mm threshold, evaluating the overall accuracy across millions of points becomes exceedingly difficult. Consequently, comparisons should concentrate on the completeness of depth estimation provided by various methods. For this purpose, we have chosen the scan33 scene from the DTU dataset as a reference point for comparative analysis, as depicted in Figure 7.

SurfaceNet [12] diverges from other MVSNet-based approaches by directly estimating the point cloud. However, it falls short in terms of point cloud completeness, as it not only misestimates in regions devoid of point clouds but also omits certain areas within the point cloud regions. In contrast to other MVSNet [14] framework-based methods such as UCSNet [38], Vis-MVSNet [39], and CasMVSNet [9], our proposed method excels in achieving superior completeness, particularly in the local edge regions of the point cloud.

D. Confidence Statistical Analysis

1) Relationship between Depth Error and Confidence: To substantiate the phenomenon of Confidence Jitter observed during the model training process, as previously discussed, this section examines the scan11 example from the DTU test set, renowned for its challenging reconstruction characteristics. Six distinctive pixel points with varied depth profiles have been identified from the depth map generated for this test instance. Given that most probability values asymptote to approximately 1.0 in the later stages of training, selecting pixel positions where the complexity is most pronounced is crucial. As illustrated in Figure 8, panels (a) through (f) correspond to these six distinct pixel locations. Epochs exhibiting higher confidence are often associated with the lowest depth value errors, signifying optimal depth estimations. Different pixel positions tend to manifest optimal depth values across various epochs.

To ascertain the synergistic relationship between the 4-stage and 5-stage models in their output results, we select scan11 and scan48 from the test samples for comparative analysis. However, with their large pixel dimensions, the original images hinder direct visual comparison. The original depth and confidence maps are segmented into multiple subplots, each with a base dimension of 32x32, facilitating calculating average confidence and average depth error within each subplot. This approach enhances the comparability of the results. As depicted in Figure 8, panels (a) and (b) display the confidence and depth error maps for scan11, while panels (c) and (d) present the analogous maps for scan48. Figures 1 and 2 correspond to the outputs of the 4-stage and 5-stage models, respectively.

The subplots on the right side of (a-1) and (a-2) reveal that the 4-stage model exhibits heightened confidence in the lower regions of the subplots. In contrast, the 5-stage model's output compensates for the 4-stage model's deficiency in confidence within the central area. The depth error plots (b-1) and (b-2) corroborate this complementary dynamic.

Similarly, the locally magnified subplots (c-1) and (c-2) reaffirm the observations above. These experiments provide an intuitive validation of the rationality behind the proposed confidence fusion approach and further underscore its holistic efficacy, as evidenced in the subsequent ablation studies.

2) Relationship between Confidence Gap and Depth Sampling: This section delves into a comparative and statistical analysis of depth and confidence maps to elucidate the previously discussed Confidence Gap phenomenon. Specifically, as depicted in Figure 10, subplot (f) presents a locally magnified depth estimation map, where each color block corresponds to an estimated depth value, with uniform depth values denoted by identical color blocks. A juxtaposition with the locally magnified confidence map in subplot (d) reveals the systematic emergence of Confidence Gaps at the peripheries of distinct color blocks, that is, at the transition points of depth pre-sampling intervals.

Further analysis involves correlating the confidence values from subplot (d) with those in subplot (f) to ascertain the effective depth range for statistical scrutiny. It is evident that the mean confidence in subplot (c) displays a trough at a position that aligns with the ring structure observed in (d), thereby directly substantiating the root cause of the Confidence Gap.

E. Impact of Image Quantity on Point Cloud Reconstruction

It is a common assumption in point cloud reconstruction algorithms that augmenting the number of input images enhances the quality of the reconstructed point cloud. However, there exists a need for more analytical inquiry into

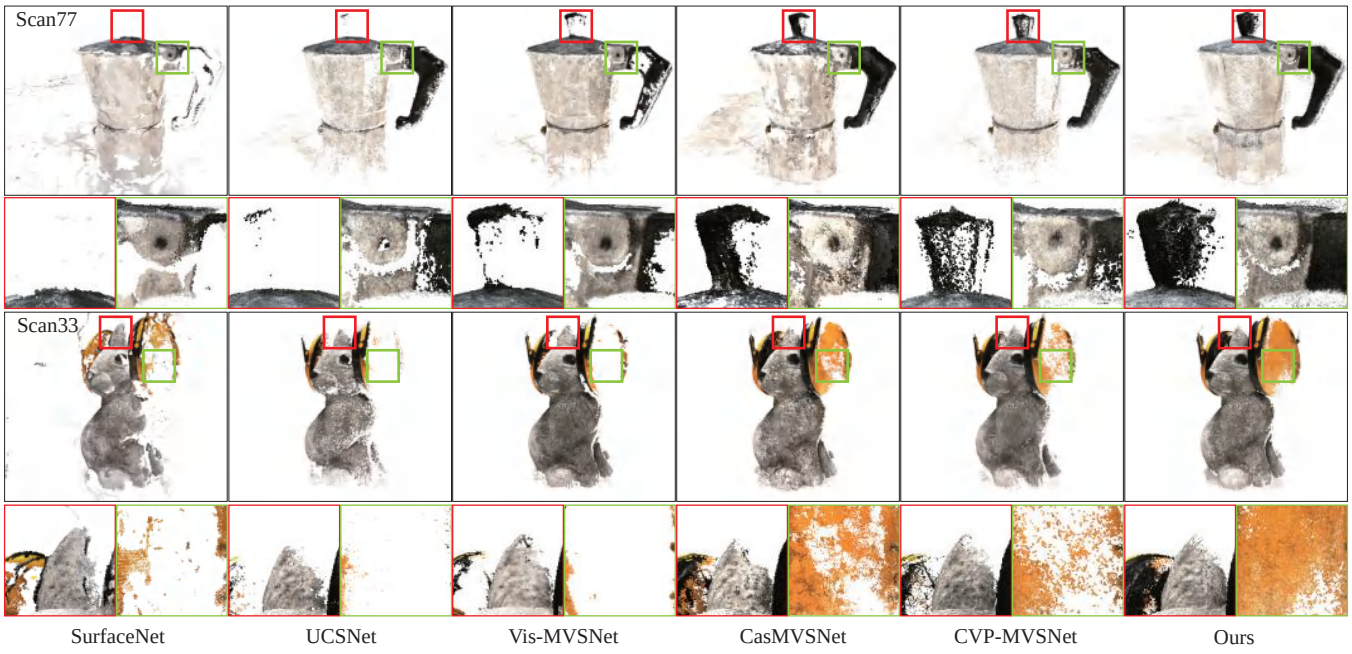


Fig. 7. The method proposed in this paper is compared with existing state-of-the-art (SOTA) methods for the point cloud. From left to right, the methods compared include SurfaceNet [12], UCSNet [38], Vis-MVSNet [39], CasMVSNet [9], CVP-MVSNet [34], and our proposed method. The red and green subregions in each subfigure correspond to representative comparison regions. The magnified display below shows these two sets of regions.

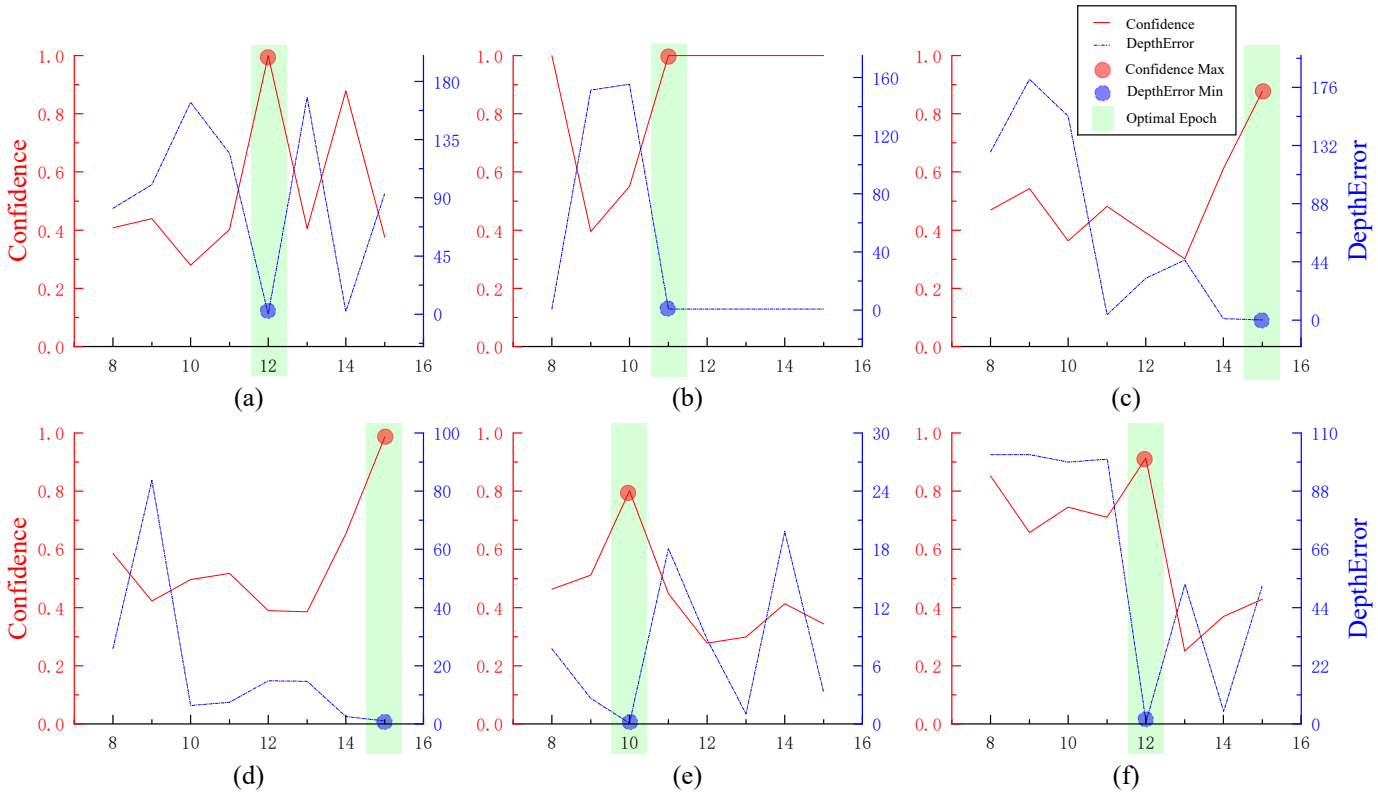


Fig. 8. On the DTU dataset, the variation trends of confidence and depth values at different pixel positions across different epochs are illustrated. In the six subplots denoted by (a) to (f), random samples of pixel positions exhibit distinct trends. In each subplot, the solid red line represents confidence, while the dashed blue line represents the absolute error between predicted and ground truth depth values. Red dots denote the highest confidence, while blue dots represent the minimum absolute depth error.

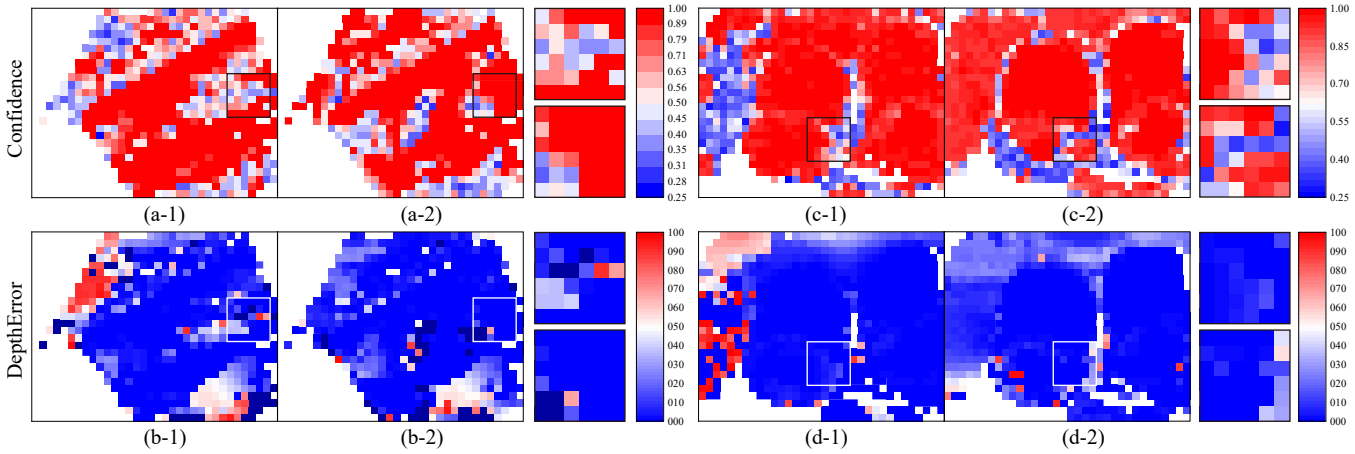


Fig. 9. Depth estimation confidence values and depth error heatmaps output by the 4-scale and 5-scale models. Where a to b, c to d represent confidence values and depth error values from the same viewpoint, 1 and 2 respectively denote outputs from the 4-scale and 5-scale models. The small images on the right side of each subplot from a to d represent zoomed-in views of 1 and 2.

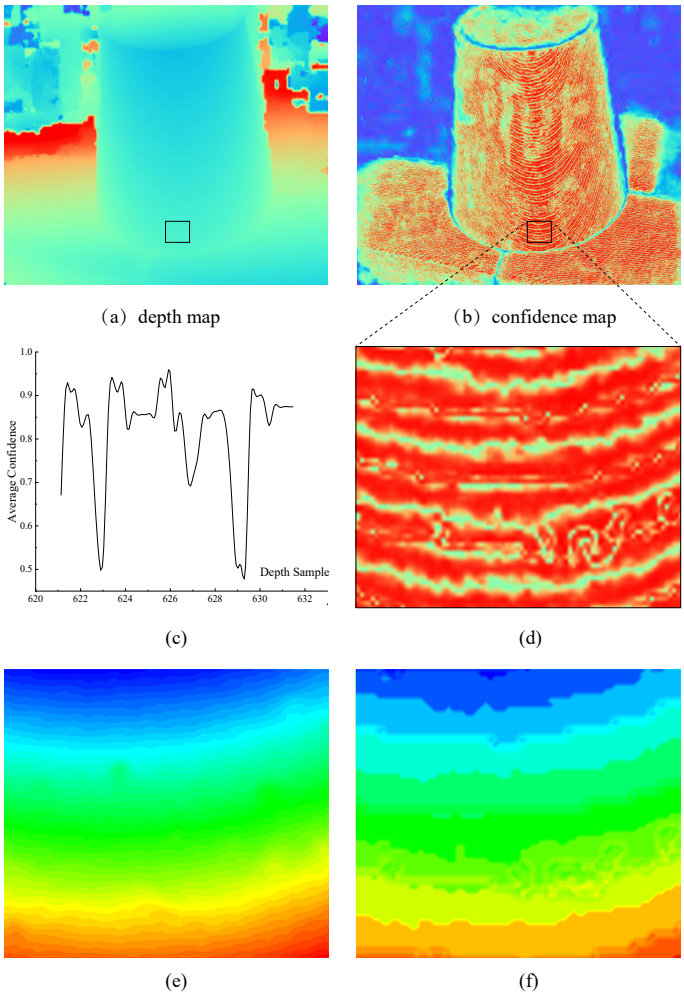


Fig. 10. Comparison and statistical analysis of Confidence Gap. (a) and (b) represent the depth map and confidence map respectively. (d) denotes local regions with significant features of Confidence Gap. (c) represents the confidence values over depth statistics mean for the region in (d). (e) and (f) represent locally magnified real depth map and predicted depth map respectively. Among them, the same color represents the same depth value. Colors ranging from warm to cool indicate depth increasing from small to large.

TABLE IV
THE COMPARISON BETWEEN DIFFERENT MODULES PROPOSED IN THIS PAPER, WHERE: NOF, EF, EF+N3, E2+N3+S5 REPRESENT DIFFERENT CONFIDENCE FUSION STRATEGIES. THE LAST COLUMN INDICATES THE AVERAGE METRIC REDUCTION COMPARED TO THE NO-FUSION APPROACH FOR EACH FUSION STRATEGY.

Fusion	Acc.	Comp.	Overall	*100↓
NoF	0.344	0.346	0.345	-
EF(11∼15)	0.318	0.324	0.321	2.4↓
EF+N3	0.317	0.302	0.310	3.5↓
EF+N3+S5	0.315	0.277	0.396	4.9↓

precisely how the quantity of images influences the accuracy and completeness of the reconstruction process. In practical applications, the intricacies of the objects being reconstructed mean that an overabundance of images could potentially amplify computational demands and, to some extent, introduce additional errors. To probe this phenomenon, we undertake an incremental comparative analysis using the scan1 test instance from the DTU dataset.

The specific outcomes are delineated in Figure 11, wherein the numerals beneath each subplot indicate the number of reference images culled from a total of 49 available images alongside the corresponding accuracy and completeness metrics of the reconstructed point cloud. A discernible trend from subplots (a) through (e) is that an increment in the number of images correlates with a decrement in point cloud accuracy while simultaneously enhancing completeness. Utilizing 30 images appears to strike an equilibrium between accuracy and completeness, thereby underscoring the rationale for judiciously reducing the image count in real-world projects.

F. Ablation Studies

This subsection delves into the pivotal components of our proposed methodology, elucidating their respective roles within the framework of confidence fusion, the Confidence Loss function, and Neighborhood Average Pooling.

1) Confidence Fusion: Given the inherent diversity and unpredictability of environmental conditions, a multitude of

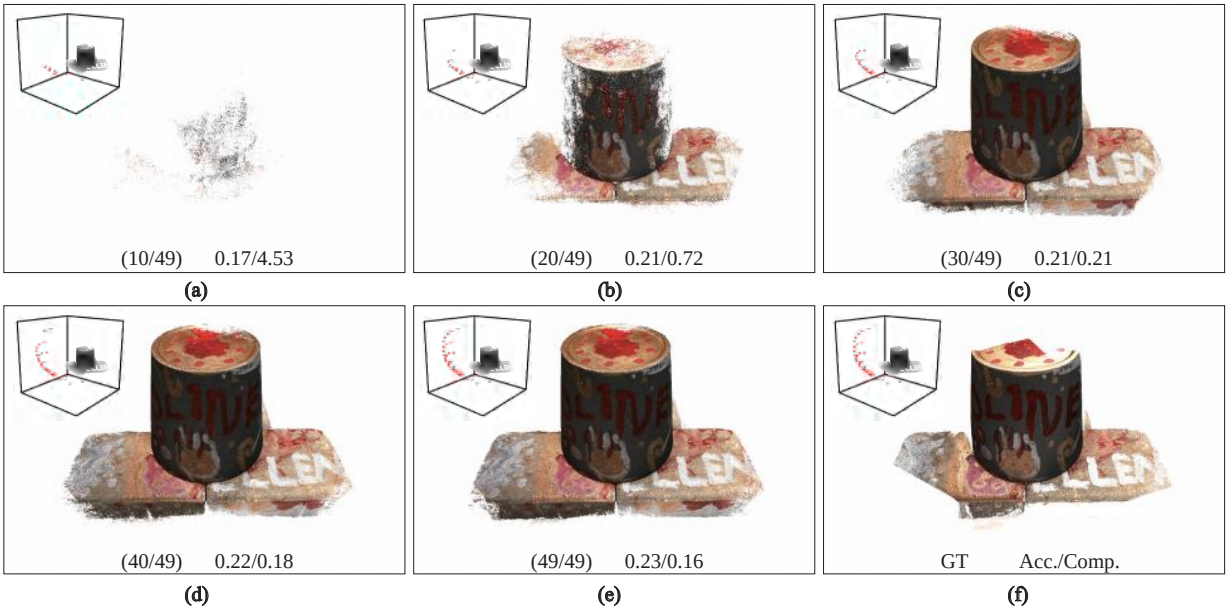


Fig. 11. Visualization of the reconstruction accuracy and completeness of the same scene with different numbers of reference images. From subfigure (a) to (e), the top-left corner of each subfigure represents the spatial position of the reference image relative to the target scene, with the main body of the subfigure showing the reconstructed point cloud. The numbers below each subfigure indicate the quantity of reference images used and the metrics for accuracy and completeness of the reconstructed point cloud. Subfigure (f) depicts the ground truth point cloud obtained from radar scanning.

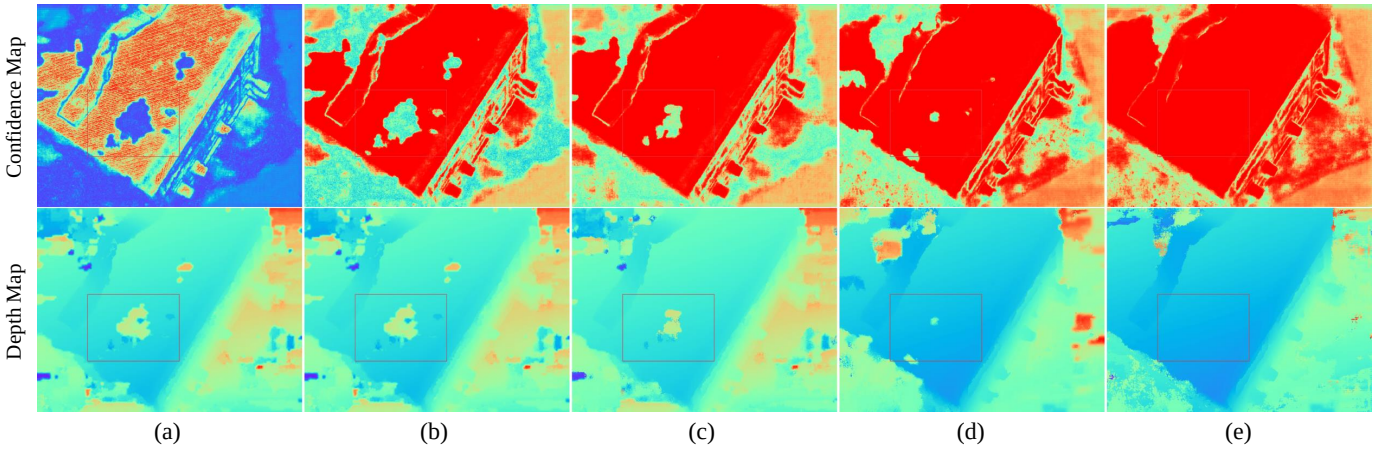
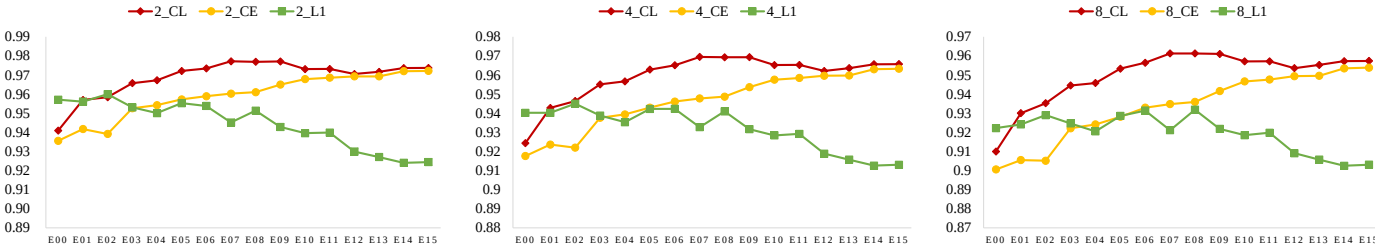


Fig. 12. Comparison of confidence maps and depth maps for different confidence fusion methods. From left to right: AP1(a), AP5(b), E2(c), S5(d), and the proposed fusion strategy(e).

comparative schemes can be devised. Our extensive experimental findings indicate that confidence fusion is particularly effective when applied to a single network across various training epochs and to the outputs of diverse fine-tuned models. Consequently, our comparative analysis primarily targets the enhancement in point cloud accuracy under these two distinct confidence fusion paradigms. As illustrated in Table IV, the terms NoF and EF correspond to scenarios utilizing five source images, where a single network is either unfused or fused from the 11th to the 15th epochs, respectively. The term N3 signifies scenarios with three source images, while S5 refers to the model scaling from four to five reference images. By amalgamating depth information from both five and three reference images, coupled with confidence fusion across the 11th to the 15th epochs and subsequent post-fine-tuning confidence fusion, we achieve a notable increase in overall precision, albeit with a minor trade-off in point cloud estimation accuracy.

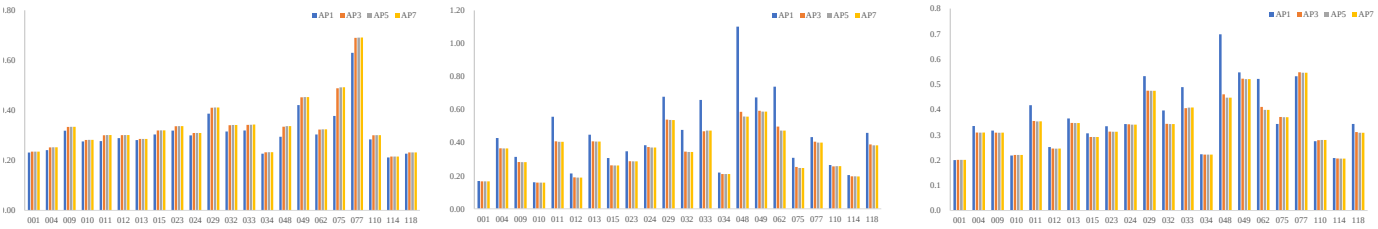
To provide a visual comparison of the efficacy of different confidence fusion strategies, we have chosen a scene with low texture for evaluation, as shown in Figure 12. The accuracy of the depth maps and the reliability of the confidence maps substantiate the superior performance of our proposed confidence fusion strategies.

Furthermore, to attest to the versatility of the EF strategy, we have extended its application to the TransMVSNet method, as delineated in Table VI. The integration of EF has led to a significant enhancement in the average completeness of the TransMVSNet outputs, with the average point cloud error being reduced by 0.006 mm.



(a) 2 mm (b) 4 mm (c) 8 mm

Fig. 13. The accuracy of the our proposed method with other loss functions across various training epochs.



(a) accuracy (b) completeness (c) average

Fig. 14. Comparison of accuracy, completeness, and reconstruction mean under different neighborhood average pooling ranges across all scenes.

TABLE V
COMPARISON OF METRICS FOR DIFFERENT CONFIDENCE NEIGHBORHOOD AVERAGE POOLING STRATEGIES ON THE DTU DATASET.

	Acc.	Comp.	Overall	*100↓
AP1	0.306	0.818	0.562	-
AP3	0.341	0.356	0.349	21.3↓
AP5	0.344	0.346	0.345	21.7↓
AP7	0.345	0.343	0.344	21.8↓

TABLE VI
COMPARISON OF METRICS FOR TRANSMVSNET W/O EF ON THE DTU DATASET.

	Acc.	Comp.	Overall	*100↓
TransMVSNet	0.338	0.295	0.317	-
TransMVSNet + EF	0.339	0.284	0.311	0.6↓

2) CL Loss Function: The optimization of network parameters during the model training process is directed by the loss function, which is pivotal in ensuring that the model's output closely matches the actual depth map. In this section, we undertake a comparative analysis of the performance of our proposed CL function across the training epochs. We scrutinize the average confidence scores of three distinct loss functions, including our CL, under a range of error threshold conditions, as depicted in Figure 13. Throughout the intermediate phases of the training epochs, the average confidence of our CL function consistently surpasses that of the other two functions. This trend underscores the superior performance of our proposed method in the generation of depth maps, especially when juxtaposed with the traditional L1 loss function.

3) Near Average Pooling: In this section, we delve into the comparative efficacy of various NAP radii by conducting experiments. Specifically, we explore the impact of pooling neighborhoods with radii of 1, 3, 5, and 7, as detailed in Table V. To ensure a fair comparison, the depth maps utilized in this analysis are devoid of confidence fusion. The results, as presented in Table V, reveal a trade-off between the accuracy and completeness of the reconstructed point cloud. With an increase in the neighborhood radius, there is a discernible decline in accuracy yet a concomitant enhancement in completeness.

Further elucidating the performance of these NAP radii, Figure 14 presents the outcomes of the experiments conducted on the DTU test set. The visualization underscores the robustness of average neighborhood pooling, particularly in scenarios characterized by high complexity. A comparative assessment of the overall accuracy of the reconstructed point clouds indicates that a confidence neighborhood radius of 5 strikes an optimal balance. This radius offers a commendable point cloud accuracy without incurring excessive computational overhead.

V. CONCLUSION

This paper introduces a novel multi-view, multi-scale stereo matching network, CF-MVSNet, designed to mitigate Confidence Jitter, Confidence Gap, and Confidence Disappearance issues. The innovative methodology harnesses a fusion of confidence at multiple levels, employs neighborhood average pooling, and integrates a Confidence Loss (CL) joint loss function. These enhancements contribute to enhanced precision and completeness in the estimation of depth.

The CF-MVSNet addresses common challenges encountered within MVS frameworks by providing robust solutions. It stands out from existing MVS techniques by exhibiting exceptional performance in the reliability of depth estimation and the fidelity of reconstructed point clouds. The adaptability of the proposed approach ensures that it can tackle the nuances associated with varying scales and views, thereby delivering a more accurate and reliable outcome.

REFERENCES

[1] W. Nie, W. Jia, W. Li, A. Liu, and S. Zhao, "3d pose estimation based on reinforce learning for 2d image-based 3d model retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 1021–1034, 2021.

[2] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng, "3d face reconstruction from a single image assisted by 2d face images in the wild," *IEEE Transactions on Multimedia*, vol. 23, pp. 1160–1172, 2021.

[3] F. Bruno, S. Bruno, G. De Sensi, M.-L. Luchi, S. Mancuso, and M. Muzzupappa, "From 3d reconstruction to virtual reality: A complete methodology for digital archaeological exhibition," *Journal of Cultural Heritage*, vol. 11, no. 1, pp. 42–49, 2010.

[4] P. Shan and W. Sun, "Research on landscape design system based on 3d virtual reality and image processing technology," *Ecological Informatics*, vol. 63, p. 101287, 2021.

[5] O. Soto-Martin, A. Fuentes-Porto, and J. Martin-Gutierrez, "A digital reconstruction of a historical building and virtual reintegration of mural paintings to create an interactive and immersive experience in virtual reality," *Applied Sciences*, vol. 10, no. 2, p. 597, 2020.

[6] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: A survey," *Pattern Recognition*, vol. 130, p. 108796, 2022.

[7] A. Ahmed, A. Jalal, and K. Kim, "Rgb-d images for object segmentation, localization and recognition in indoor scenes using feature descriptor and hough voting," in *2020 17th international Bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2020, pp. 290–295.

[8] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[9] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "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*, 2020, pp. 2495–2504.

[10] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376.

[11] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 418–433, 2005.

[12] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "Surfacenet: An end-to-end 3d neural network for multiview stereopsis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2307–2315.

[13] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[15] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo." Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 766–779.

[16] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo - stereo matching with slanted support windows," in *PROCEEDINGS OF THE BRITISH MACHINE VISION CONFERENCE 2011*, 2011.

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

[18] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 873–881.

[19] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler, "Learned multi-patch similarity," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1586–1594.

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

[21] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.

[22] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1538–1547.

[23] Y. Hou, J. Kannala, and A. Solin, "Multi-view stereo by temporal nonparametric fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2651–2660.

[24] Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, and J. Bao, "Mvscrf: Learning multi-view stereo with conditional random fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4312–4321.

[25] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 451–10 460.

[26] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8585–8594.

[27] C. Cao, X. Ren, and Y. Fu, "Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth," *Transactions on Machine Learning Research*, 2022.

[28] Z. Mi, C. Di, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 991–13 000.

[29] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, pp. 153–168, 2016.

[30] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[32] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 501–518.

[33] Q. Xu and W. Tao, "Multi-view stereo with asymmetric checkerboard propagation and multi-hypothesis joint view selection," *arXiv preprint arXiv:1805.07920*, 2018.

[34] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[35] S. Gao, Z. Li, and Z. Wang, "Cost volume pyramid network with multi-strategies range searching for multi-view stereo," in *Computer Graphics International Conference*. Springer, 2022, pp. 157–169.

[36] R. Weilharter and F. Fraundorfer, "Atlas-mvsnet: Attention layers for feature extraction and cost volume regularization in multi-view stereo," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 3557–3563.

[37] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[38] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2524–2534.

[39] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "Vis-mvsnet: Visibility-aware multi-view stereo network," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 199–214, 2023.

[40] T. Liu, X. Ye, W. Zhao, Z. Pan, M. Shi, and Z. Cao, "When epipolar constraint meets non-local operators in multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 088–18 097.

Xin Ma received the B.E. degree in Architecture and the M.E. degree in Computer Science from Northwestern Polytechnical University, Xi'an, China, in 2020 and 2023 respectively. He is currently working toward the Ph.D. degree in the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include 3D reconstruction and image enhancement.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Yuan Yuan (Senior Member, IEEE) is currently a Full Professor with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image / video.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition and remote sensing.