MVS2D: Efficient Multi-view Stereo via Attention-Driven 2D Convolutions

Zhenpei Yang\textsuperscript{1,}\textsuperscript{*} Zhihe Ren\textsuperscript{2} Qi Shan\textsuperscript{2} Qixing Huang\textsuperscript{1}
\textsuperscript{1}The University of Texas at Austin \textsuperscript{2}Apple

Abstract

Deep learning has made significant impacts on multi-view stereo systems. State-of-the-art approaches typically involve building a cost volume, followed by multiple 3D convolution operations to recover the input image’s pixel-wise depth. While such end-to-end learning of plane-sweeping stereo advances public benchmarks’ accuracy, they are typically very slow to compute. We present MVS2D, a highly efficient multi-view stereo algorithm that seamlessly integrates multi-view constraints into single-view networks via an attention mechanism. Since MVS2D only builds on 2D convolutions, it is at least \(2\times\) faster than all the notable counterparts. Moreover, our algorithm produces precise depth estimations and 3D reconstructions, achieving state-of-the-art results on challenging benchmarks ScanNet, SUN3D, RGBD, and the classical DTU dataset. Our code is released at https://github.com/zhenpeiyang/MVS2D

1. Introduction

Multi-view Stereo (MVS) aims to reconstruct the underlying 3D scene or estimate the dense depth map using multiple neighboring views. It plays a key role in a variety of 3D vision tasks. With high-quality cameras becoming more and more accessible, there are growing interests in developing reliable and efficient stereo algorithms in various applications, such as 3D reconstruction, augmented reality, and autonomous driving. As a fundamental problem in computer vision, MVS has been extensively studied [9]. Recent research shows that deep neural networks, especially convolutional neural networks (CNNs), lead to more accurate and robust solutions than traditional solutions. Several approaches [20, 55] report exceptional accuracy on challenging benchmarks like ScanNet [7] and SUN3D [45].

State-of-the-art CNN-based multi-view approaches typically fall into three categories: 1) Variants of a standard 2D UNet architecture with feature correlation [22, 26]. However, these approaches work best for rectified stereo pairs, and extending them to multi-view is nontrivial. 2) Constructing a differential 3D cost volume [12,14,15,28,51,52]. These algorithms significantly improve the accuracy of MVS, but at the cost of heavy computational burdens. Furthermore, the predicted depth map by 3D convolution usually contains salient artifacts which have to be rectified by a 2D refinement network [15]. 3) Maintain a global scene representation and fuse multi-view information through ray-casting features from 2D images [27]. This paradigm cannot handle large-scale scenes because of the vast memory consumption on maintaining a global representation.

Aside from multi-view depth estimation, we have also witnessed the tremendous growth of single-view depth prediction networks [21, 30, 46, 50, 54]. As shown in Table 3, Bts [21] has achieved impressive result on ScanNet [7]. Single-view depth prediction roots in learning feature representations to capture image semantics, which is orthogonal to correspondence-computation in multi-view techniques. A natural question is how to combine single-view depth cues and multi-view depth cues.

We introduce MVS2D that combines the strength of single-view and multi-view depth estimations. The core contribution is an attention mechanism that aggregates features along epipolar lines of each query pixel on the reference images. This module captures rich signals from the

\* Experiments are conducted by Z. Yang at The University of Texas at Austin. Email: yzp@utexas.edu
reference images. Most importantly, it can be easily inte-
gated into standard CNN architectures defined on the input
image, introducing relatively low computational cost.

Our attention mechanism possesses two appealing char-
acteristics: 1) Our network only contains 2D convolutions.
2) Besides relying on the expressive power of 2D CNNs,
the network seamlessly integrates single-view feature rep-
resentations and multi-view feature representations. Con-
sequently, MVS2D is the most efficient approach compared to
state-of-the-art algorithms (See Figure 1). It is 48× faster
than NAS [20], 39× faster than DPSNet [15], 10× faster
than MVSNet [51], 4.7× faster than FastMVSNet [55], and
almost 2× speed-up over the most recent fastest approach
PatchmatchNet [42]. In the mean-time, MVS2D achieves
state-of-the-art accuracy.

Intuitively, the benefit of MVS2D comes from the early
fusion of the intermediate feature representations. The out-
come is that the intermediate feature representations con-
tain rich 3D signals. Furthermore, MVS2D offers ample
space where we can design locations of the attention mod-
ules to address different inputs. One example is when the
input camera poses are inaccurate, and corresponding pix-
els deviate from the epipolar lines on the input reference
images. We demonstrate a simple solution, which installs
multi-scale attention modules on an encoder-decoder net-
work. In this configuration, corresponding pixels in down-
sampled reference images lie closer to the epipolar lines,
and MVS2D detect and rectify correspondences automatic-
ally.

We conduct extensive experiments on challenging
benchmarks ScanNet [7], SUN3D [45], RGBD [34] and
Scenes11 [34]. MVS2D achieves the state-of-the-art per-
formance on nearly all the metrics. Qualitatively, compared
to recent approaches [15,20,51,55], MVS2D helps generate
higher quality 3D reconstruction outputs.

2. Related Works

Recent advances of multi-view stereo. Multi-view stereo
algorithms can be categorized into depth map-based ap-
proaches, where the output is a per-view depth map, or
point-based approaches, where the output is a sparse re-
construction of the underlying scene (cf. [9]). Many tra-
ditional multi-view stereo algorithms follow a match-then-
reconstruct paradigm [10] that leverages the sparse-nature
of feature correspondences. Such a paradigm typically fails
to reconstruct textureless regions where correspondences
are not well-defined. Along this line, Zhontar et al. [56]
provided one of the first attempts to bring the power of fea-
ture learning into multi-view stereo. They proposed a sup-
ervised feature learning approach to find the correspondences.
Recently, researchers have found that depth map-based ap-
proaches [14,15,17,51,52] are more favorable than those
that follow the match-and-reconstruct paradigm. A key ad-
vantage of these approaches is that they can utilize the ef-

ciciency of regular tensor operations. [15,51] proposed an
end-to-end plane-sweeping stereo approach that constructs
learnable 3D cost volume. While MVSNet [51] focuses on
the reconstruction of 3D scene, DPSNet [15] focuses on
evaluating the per-view depth-map accuracy. Researchers
have also explored other 3D representations to regularized
the prediction, such as point clouds [4], surface normals
[20], or meshes [44]. There are also several benchmark
datasets for this task [1,7,34,40,45,53].

Cost volume for multi-view stereo. A recent line of
works on multi-view stereo utilizes the notion of cost vol-
ume, which contains feature matching costs for a pair of
images [13]. This feature representation has been suc-
scessfully implemented in various pixel-wise matching tasks
like optical flow [35]. Authors of MVSNet [51] and DP-
SNet [15] proposed to first construct a differentiable cost
volume and then use the power of 3D CNNs to regularize
the cost volume before predicting per-pixel depth or dispar-
ity. Most recent state-of-the-art approaches follow such a
paradigm [4,12,23,28,52,55]. However, the size of the cost
volume ($C \times K \times H \times W$) is linearly related to the number
of depth hypotheses $K$. These approaches are typically slow
in both training and inference. For example, DPSNet [15]
takes several days to train on ScanNet; NAS [20] takes
even longer because of its extra training of a depth-normal
consistency module. Recently, Murez et al. [27] proposed
to construct a volumetric scene representation from a cali-
ibrated image sequence for scene reconstruction. However,
their approach is very memory demanding due to the high
memory requirement of global volumetric representations.

Efficient multi-view stereo. Several recent works aim at
reducing the cost of constructing cost volumes. Duggal et al.
[8] prune the disparity search range during cost volume
construction. Xu et al. [47] integrate adaptive sampling
and deformable convolution into correlation-based meth-
ods [22,26] to achieve efficient aggregation. Several other
works [12,36,52] employ iterative refinement procedures.
The above approaches either only work for pairwise recti-
fied stereo matching tasks or have to construct a 3D cost-
volume. Alternatively, Poms et al. [29] learn how to merge
patch features for 3D reconstruction efficiently. Badki et al.
[2] convert depth estimation as a classification task, but
the resulting accuracy is not state-of-the-art. Recently, Yu et al.
[55] proposed constructing a sparse cost-volume through
regular sub-sampling and then applying Gauss-Newton iter-
ations to refine the dense depth map. Wang et al. [42] pro-
duced a highly-efficient Patchmatch-inspired approach for
MVS tasks. In contrast, we take an orthogonal approach
based on the attention-driven 2D convolutions.

Attention in 3D vision Attention mechanism has shown
prominent results on both natural language processing
(NLP) tasks [41] and vision tasks [43]. Recently, self-local
attention [31,33] has shown promising results compared
with the convolution-based counterparts. Several recent
works that build an attention mechanism in MVS [23,25,
Figure 2. Network architecture of MVS2D. We employ a 2D UNet structure $F$ to make the depth prediction on $I_0$, while injecting multi-view cues extracted using $G$ through the Epipolar Attention Module. Dashed arrows only exist in Ours-robust model (Section 3.4). We highlight that the proposed epipolar attention module can be easily integrated into most 2D CNNs.

3. Approach

We provide an overview of the network architecture of MVS2D in Fig. 2. We operate in a multi-view stereo setting (Sec. 3.1), and employ a 2D UNet structure in our network design (Sec. 3.2). Our core contribution is the epipolar attention module (Sec. 3.3–3.4), which is highly accurate and efficient (Sec. 3.5) for depth estimation (Sec. 3.6).

3.1. Problem Setup

We aim to estimate the per-pixel depth for a source image $I_0 \in R^{h \times w \times 3}$, given $n$ reference images $\{I_i\}_{i=1}^{n}$ of the same size captured at nearby views. We assume the source image and the reference images share the same intrinsic camera matrix $K \in R^{3 \times 3}$, which is given. We also assume we have a good approximation of the relative camera pose between the source image and each reference image $T_i = (R_i | t_i)$, where $R_i \in SO(3)$ and $t_i \in R^3$. $T_i$ usually comes from the output of a multi-view structure-from-motion algorithm. Our goal is to recover the dense pixel-wise depth map associated with $I_0$.

We denote the homogeneous coordinate of a pixel $p_0$ in the source image $I_0$ as $\overline{p}_0 = (\overline{p}_{0,1}, \overline{p}_{0,2}, 1)^T$. Given the depth $d_0 \in R$ of $p_0$, the unprojected 3D point of $p_0$ is

$$p_0(d_0) = d_0 \cdot (K^{-1}p_0).$$

Similarly, we use $p_i(d_0)$ and $\overline{p}_i(d_0)$ to denote respectively the 3D coordinates and homogeneous coordinates of $p_0(d_0)$ in the $i$-th image’s coordinate system. They satisfy

$$p_i(d_0) = R_ip_0(d_0) + t_i,$$

$$\overline{p}_i(d_0) = K\overline{p}_i(d_0).$$  \hspace{1cm} (1)

3.2. Network Design Overview

In this paper, we innovate developing a multi-view stereo approach that only requires 2D convolutions. Specifically, similar to most single-view depth prediction networks, our approach progressively computes multi-scale activation maps of the source image and outputs a single depth map. The difference is that certain intermediate activation maps combine both the output of a 2D convolution operator applied to the previous activation map and the output of an attention module that aggregates multi-view depth cues. This attention module, which is the main contribution of this paper, matches each pixel of the source image and corresponding pixels on epipolar lines on the reference images. The matching procedure utilizes learned feature activations on both the source image and the reference images. The output is encoded using learned depth codes compatible with the activation maps of the source image.

Formally speaking, our goal is to learn a feed-forward network $F$ with $L$ layers. With $F_j \in R^{h_j \times w_j \times m_j}$ we denote the output of the $j$-th layer, where $m_j$ is its feature dimension, $h_j$ and $w_j$ are its height and width. Note that the first layer $F_1 \in R^{h_1 \times w_1 \times 3}$ denotes the input, while the last layer $F_L \in R^{h_L \times w_L}$ denotes the output layer containing depth prediction. Between two consecutive layers are a general convolution operator $C_j$
We proceed to define \( A_j(p_0) \), which is the action of \( A_j \) on each pixel \( p_0 \). It consists of two parts:

\[
A_j(p_0) = \mathcal{A}^{e,p}_j(p_0, \{I_i\}_{i=1}^n) + A_j^0(F_j(p_0)). \tag{2}
\]

As we will define next, \( \mathcal{A}^{e,p}_j(p_0, \{I_i\}_{i=1}^n) \) uses trainable depth codes to encode the matching result between \( p_0 \) and the reference images. \( A_j^0 : \mathcal{R}^{m_j} \rightarrow \mathcal{R}^{m_j} \) is composed of an identity map and a trainable linear map that transforms the feature associated with \( p_0 \) in \( F_j \).

The formulation of \( \mathcal{A}^{e,p}_j(p_0, \{I_i\}_{i=1}^n) \) uses the epipolar context of \( p_0 \). It consists of samples on the epipolar lines of \( p_0 \) on the reference images. These samples are obtained from sampling the depth values \( d_0 \) of \( p_0 \) and then applying (1). With \( p_i^k \) we denote the \( k \)-th sample on the \( i \)-th reference image.

To match \( p_0 \) and \( p_i^k \), we introduce a feature extraction network \( G \) that has identical architecture (except the attention modules) as \( F_{\text{max}} \) where \( j_{\text{max}} \) is the maximum depth of any attention module of \( F \). With \( G_j(I_0, p_0) \in \mathcal{R}^{m_j} \) and \( G_j(I_i, p_i^k) \in \mathcal{R}^{m_j} \) we denote the extracted features of \( p_0 \) and \( p_i^k \), respectively. Following the practice of scaled-dot product attention [41], we introduce two additional trainable linear maps \( f^{ik} \) and \( f^{j_{\text{ref}}} \) to transform the extracted features. With this setup, we define the matching score between \( p_0 \) and \( p_i^k \) as

\[
w_{ik}^j = (f^{ik}_j(G_j(I_0, p_0)))^T (f^{j_{\text{ref}}}_j(G_j(I_i, p_i^k))). \tag{3}
\]

It remains to 1) model samples that are occluded in the reference images, and 2) bridge the weights \( w_{ik}^j \), defined in (3) and the input to the convolution operator \( C_j \). To this end, we first introduce trainable mask codes \( c_{ijk} \) that correspond to the \( k \)-th depth sample. We then introduce \( v_{in}^j \in \mathcal{R}^{m_j} \) and \( v_{out}^j \in \mathcal{R}^{m_j} \), which are trainable codes for inside and outside samples, respectively. Define

\[
v_{ik}^j = \begin{cases} v_{in}^j & 0 \leq p_{i,1}^k < w, 0 \leq p_{i,2}^k < h, p_{i,3}^k \geq 0, \\ v_{out}^j & \text{otherwise} \end{cases}
\]

where \( p_{i}^k = (p_{i,1}^k, p_{i,2}^k, 1)^T \), \( p_{i}^k = (p_{i,1}^k, p_{i,2}^k, p_{i,3}^k)^T \). To enhance the expressive power of \( G_j \), we further include a trainable linear map \( A_j^0 \) that depends only on feature of \( p_0 \) and not on the matching results. Combing with (3) and (4), we define

\[
A_j^{e,p}(p_0, \{I_i\}_{i=1}^n) = A_j^0(F_j(p_0)) + \sum_{i=1}^n \sum_{k=1}^K N \left( \frac{w_{ik}^j}{\sqrt{m_j}} \right) (v_{ik}^j \odot c_k) \tag{5}
\]

where \( N \) is the softmax normalizing function over \( \frac{w_{ik}^j}{\sqrt{m_j}} \), \( 1 \leq k \leq K \). Substituting (5) into (2), the final attention module is given by

\[
A_j(p_0) = A_j^0(F_j(p_0)) + A_j^0(G_j(p_0)) + \sum_{i=1}^n \sum_{k=1}^K N \left( \frac{w_{ik}^j}{\sqrt{m_j}} \right) (v_{ik}^j \odot c_k).
\]

Note that the attention modules at different layers have different weights. Eq. 3 can be viewed as a similarity score between source pixel and correspondence candidates. In Fig. 3, we visualize the learned attention scores for query pixels. The true corresponding pixels on reference images have larger learned weights along epipolar lines.

![Figure 3. Visualization of attention scores. Left: source view with query pixels. Right: reference view with candidate pixels, where opacity is learned attention scores.](image328x212 to 526x362)

#### 3.4. Attention Design for Robust Multi-View Stereo

Since the attention module assumes that the corresponding pixels lie on the epipolar lines, the accuracy of MVS2D depends on the relative poses’ accuracy between the reference images and the source image. When input poses are accurate, our experiments suggest a single attention module
Table 1. Quantitative comparison on computational efficiency. FPS (V) only applies to multi-view methods [15,20,51,55] and means we use V images to make the prediction. Note that numbers under the AbsRel metric are identical to those in Table 3 for ease of comparison. We use a single Nvidia V100 GPU for measuring FPS. Please refer to section 4.4 for additional discussions.

| Method                | FPS (3) | FPS(7) | FPS(11) | Param (M) | AbsRel |
|-----------------------|---------|--------|---------|-----------|--------|
| MVSNet [51]           | 4.1     | 2.4    | 1.6     | 1.1       | 0.094  |
| DPSSNet [15]          | 1.1     | 0.7    | 0.5     | 4.2       | 0.094  |
| FastMVS [53]          | 9.0     | 6.0    | 4.3     | 0.4       | 0.089  |
| PatchmatchNet [42]    | 21.8    | 11.6   | 8.5     | 0.2       | 0.133  |
| NAS [20]              | 0.9     | 0.6    | 0.4     | 18.0      | 0.086  |
| Ours-mono             | 94.7    | -      | -       | 12.3      | 0.145  |
| Ours-robust           | 17.5    | 10.1   | 7.1     | 24.4      | 0.059  |
| Ours                  | 42.9    | 29.1   | 21.8    | 13.0      | 0.059  |

Table 2. Quantitative metrics for depth estimation. The training set consists of three data sources, SUN3D [45], RGBD [34], and Scenes11 [40]. SUN3D and RGBD contain real indoor scenes, while Scenes11 is synthetic. In total, there are 79577 training pairs for SUN3D, 16786 for RGBD, and 71820 for Scenes11.

4.1. Datasets

ScanNet [7] The ScanNet dataset contains 807 unique scenes with image sequences captured from different camera trajectories. We sample 86324 triple images (one source image and two reference images) for training and 666 triple images for testing. Our setup ensures the scene corresponding to test images is not included in the training set.

DeMoN [40] We further validate our method on DeMoN, which is a dataset introduced by [40] for multi-view depth estimation. The training set consists of three data sources, SUN3D [45], RGBD [34], and Scenes11 [40]. SUN3D and RGBD contain real indoor scenes, while Scenes11 is synthetic. In total, there are 79577 training pairs for SUN3D, 16786 for RGBD, and 71820 for Scenes11.

DTU [1] While our approach is designed for multi-view depth estimation, we additionally validate our method on the DTU dataset, which has been considered as one of the main test-beds for multi-view reconstruction algorithms.

4.2. Evaluation Metrics

Efficiency. We benchmark our methods against baseline methods on the frame per second (FPS) during inference. We additionally compare the FPS when increasing the number of reference views.

Depth Accuracy. We use the conventional metrics of depth estimation [21] (See Table 2). Note that in contrast to monocular depth estimation evaluation, we do not factor out the depth scale before evaluation. The ability to correctly predict scale will render our method more applicable.

Scene Reconstruction Quality. We further apply MVS2D for scene reconstruction. We follow PatchmatchNet [42] to fuse the per-view depth map into a consistent 3D model. Please refer to supp. material for quantitative and more qualitative comparisons.

Robustness under Noisy Input Pose. We perturb the input relative poses $T_j$ during training and report the model performance on ScanNet test set in Table 8. Please refer to the supp. material for details of the pose perturbation procedures.

4.3. Baseline Approaches

MVSNet [51] is an end-to-end plane sweeping stereo approach based on 3D-cost volume.

4.4. Computational Complexity

For the sake of simplicity in notation, we assume the feature channel dimension $C$ is the same in both input and output. We denote the feature height and width as $H$ and $W$ respectively and denote the kernel size of convolution layers as $k$. Suppose there are $K$ depth samples, the complexity of 3D convolution is $O(C^2HW/K/k^3)$.

For our approach, the computational complexity for executing one layer of $A(a)$ is in total $O(CHW(Ck^2 + K))$. Since $K$ is usually less than $Ck^2$, our module leads to a $KK$ times reduction in computation. The actual runtime can be found in Table 1.

3.6. Training Details

Our implementation is based on PyTorch. For ScanNet and DeMoN, we simply optimize the $L_1$ loss between predicted and ground truth depth. For DTU, we introduce a simple modification, as was done in [18], to simultaneously train a confidence prediction. We use Adam [19] optimizer with $\varepsilon = 10^{-8}$, $\beta = (0.9, 0.999)$. We use a starting learning rate $2e^{-4}$ for ScanNet, $8e^{-4}$ for DeMoN and $2e^{-4}$ for DTU. Please refer to supp. material for more training details.
DPSNet [15] shares similar spirit of MVSNet [51] but focus on accurate depth map prediction.

NAS [20] is a recent work that jointly predicts consistent depth and normal, using extra normal supervision.

FastMVSNet [55] is a recent variant to MVSNet which accelerates the computation by computing sparse cost volume.

Bts [21] is a state-of-the-art single view depth prediction network. It incorporates planar priors into network design. Additionally, we use an asterisk sign ‘∗’ to denote an oracle version Bts∗, where we use the ground truth depth map to factor out the global scale.

PatchmatchNet [42] is one of the most recent state-of-the-art efficient MVS algorithm.

Ours-mono is our method without the epipolar attention module, thus equivalent to single-view depth estimation. Similar to Bts∗, we also report the results factoring out the global scale for Ours-mono∗.

Ours-robust is our method with multi-scale epipolar attention module applied on F’s second layer.

Ours is our method with epipolar attention module applied only in F’s second layer.

### 4.4. Result Analysis

**Comparison on Efficiency.** We compare against both single-view methods [21] and multi-view methods [15, 20, 51, 55]. The inference speed of our method is comparable to single-view methods [21] and significantly outperforms other multi-view methods [15, 20, 51, 55]. Evaluations are done on ScanNet [7]. Our method is 48× faster than NAS, 39× faster than DPSNet, 10× faster than MVSNet, and 4×

| Method   | AbsRel ↓ | SqRel ↓ | log10 ↓ | RMSE ↓ | RMSELog ↓ | δ < 1.25 ↑ | δ < 1.25^2 ↑ | δ < 1.25^3 ↑ |
|----------|----------|---------|---------|--------|------------|------------|-------------|-------------|
| Bts [21] | 0.117    | 0.052   | 0.049   | 0.270  | 0.151      | 0.862      | 0.966       | 0.992       |
| Bts∗ [21]| 0.088    | 0.035   | 0.038   | 0.228  | 0.128      | 0.916      | 0.980       | 0.994       |
| MVSNet [51]| 0.094  | 0.042   | 0.040   | 0.251  | 0.135      | 0.897      | 0.975       | 0.993       |
| FastMVSNet [55]| 0.089 | 0.038   | 0.038   | 0.231  | 0.128      | 0.912      | 0.978       | 0.993       |
| DPSNet [15]| 0.094  | 0.041   | 0.043   | 0.258  | 0.141      | 0.883      | 0.970       | 0.992       |
| NAS [20]| 0.086   | 0.032   | 0.038   | 0.224  | 0.122      | 0.917      | 0.984       | 0.996       |
| PatchmatchNet [42]| 0.133 | 0.075   | 0.055   | 0.320  | 0.175      | 0.834      | 0.955       | 0.987       |
| Ours-mono| 0.145   | 0.065   | 0.061   | 0.300  | 0.173      | 0.807      | 0.957       | 0.990       |
| Ours-mono∗| 0.103   | 0.037   | 0.044   | 0.237  | 0.135      | 0.892      | 0.984       | 0.996       |
| Ours-robust| 0.059  | 0.016   | 0.026   | 0.159  | 0.083      | 0.965      | 0.996       | 0.999       |
| Ours     | 0.059   | 0.017   | 0.026   | 0.162  | 0.084      | 0.963      | 0.995       | 0.999       |

Table 3. Depth evaluation results on ScanNet [7]. We compare against both multi-view depth estimation methods [15, 20, 42, 51, 55] and a state-of-the-art single-view method [21]. Our approach achieve significant improvements over top-performing method NAS [20] on AbsRel. The improvements are consistent across all metrics.

| Method   | AbsRel ↓ | AbsDiff ↓ | SqRel ↓ | RMSE ↓ | RMSELog ↓ | δ < 1.25 ↑ | δ < 1.25^2 ↑ | δ < 1.25^3 ↑ |
|----------|----------|------------|---------|--------|------------|------------|-------------|-------------|
| COLMAP [32]| 0.623  | 1.327      | 3.236   | 2.316  | 0.661      | 0.327      | 0.554       | 0.718       |
| DeMoN [40]| 0.214  | 2.148      | 1.120   | 2.421  | 0.206      | 0.733      | 0.922       | 0.963       |
| DeepMVS [14]| 0.282 | 0.604      | 0.435   | 0.944  | 0.363      | 0.562      | 0.739       | 0.895       |
| DPSNet-U [15]| 0.147 | 0.336      | 0.117   | 0.449  | 0.196      | 0.781      | 0.926       | 0.973       |
| NAS [20]| 0.127   | 0.288     | 0.085   | 0.378  | 0.170      | 0.830      | 0.944       | 0.978       |
| Ours-robust| 0.100  | 0.231      | 0.057   | 0.313  | 0.140      | 0.895      | 0.966       | 0.991       |
| Ours     | 0.099   | 0.224      | 0.055   | 0.304  | 0.137      | 0.893      | 0.970       | 0.993       |

Table 4. Depth evaluation results on SUN3D, RGBD, and Scenes11 datasets(synthetic). The numbers for COLMAP, DeMoN, DeepMVS, DPSNet, and NAS are obtained from [20]. We achieve significant improvements on SUN3D and RGBD. We show the best number in bold and the second best with underline.
Comparison on Depth Estimation. MVS2D achieves considerable improvements in the depth prediction accuracy (see Table 3). On ScanNet, our approach outperforms MVSNet by large margins, reducing AbsRel error from 0.094 to 0.059. The improvements are consistent across most other metrics. Remarkably, our approach also outperforms NAS, which uses more parameters and runs 48 times slower. We visualize some depth predictions in Figure 4.

Our approach yields significant improvements over single-view baselines. Adding multi-view cues improves the AbsRel of ours-mono from 0.145 to 0.059 on ScanNet. Since single-view has scale-ambiguity, we further investigate whether our methods will still be favorable when factoring out the scale. The results show that when eliminating the scale, ours-mono∗ has AbsRel 0.103, which is still a significant improvement. This means our approach does not simply infer the global scaling factor from multi-view cues. Compared with the single-view model, our model only incurs a 5.8% increase in parameters. Such efficiency will enable multi-view methods to embrace a much larger 2D convolutional network which is not possible before.

On other datasets, MVS2D also performs favorably (see Table 4). We achieve the AbsRel error of 0.078 on the RGBD dataset, while the next best NAS only achieves 0.131. Although MVS2D excels at adapting to the scene prior, it is encouraging that it also performs well on the Scenes11 dataset, a synthetic scene with randomly placed objects. We ranked second on the Scenes11 dataset on AbsRel. Please refer to the supp. material for experiments on our generalization ability to novel datasets.

Evaluations on DTU. We evaluate on DTU dataset following the practice of [42]. We use 4 reference views and 96 depth samples uniformly placed in the inverse depth space ([1/935, 1/1425]). The quantitative results can be found in Table 7.
5. Conclusions and Limitations

Conclusions. We proposed a simple yet effective method for multi-view stereo. The core of our method is to integrate single-view and multi-view cues during the prediction jointly. Such a design not only improves the performance but also has the appealing factor of being efficient. Furthermore, we have demonstrated the trade-off between input pose accuracy and network complexity. When the input pose is exact, we can leverage minimum additional computation to inject more multi-view information through the epipolar attention. We can see that learning the codes end-to-end leads to noticeable performance gains. One explanation is that these learned codes can adapt to the single-view feature representations of the source image.

Limitations. One limitation of our approach is that the network is trained in a way that adapted to data distribution well, which might makes it less generalizable to out-of-distribution testing data. In the future, we propose to address this issue by developing robust training losses. Another limitation is that the proposed attention mechanism does not explicitly model the consistency between different

| Methods          | Acc.(mm) | Comp.(mm) | Overall(mm) |
|------------------|----------|-----------|-------------|
| Camp [3]         | 0.835    | 0.554     | 0.695       |
| Furu [10]        | 0.613    | 0.941     | 0.777       |
| Tola [38]        | 0.342    | 1.190     | 0.766       |
| Gipuma [11]      | 0.283    | 0.873     | 0.578       |
| SurfaceNet [16]  | 0.450    | 1.040     | 0.745       |
| MVSNet [51]      | 0.396    | 0.527     | 0.462       |
| R-MVSNet [52]    | 0.383    | 0.452     | 0.417       |
| CIDER [48]       | 0.417    | 0.437     | 0.427       |
| P-MVSNet [24]    | 0.406    | 0.434     | 0.420       |
| Point-MVSNet [4] | 0.342    | 0.411     | 0.376       |
| Fast-MVSNet [55] | 0.336    | 0.403     | 0.370       |
| CasMVSNet [12]   | 0.325    | 0.385     | 0.355       |
| UCS-Net [6]      | 0.338    | 0.349     | 0.344       |
| CVP-MVSNet [49]  | 0.296    | 0.406     | 0.351       |
| PatchMatchNet [42]| 0.427  | 0.277     | 0.352       |
| MVS2D (Ours)     | 0.394    | 0.290     | 0.342       |

Table 5. Quantitative results on the evaluation set of DTU [1]. We bold the best number and underline the second best number.

Table 6. Speed benchmark on DTU dataset. We show FPS (frame per second) on three input resolutions. We use one source image and 4 reference images.

| Metric                      | FPS@640 × 480 ↑ | FPS@1280 × 640 ↑ | FPS@1536 × 1152 ↑ |
|-----------------------------|-----------------|------------------|-------------------|
| PatchMatchNet               | 16.5            | 6.30             | 4.57              |
| MVS2D (Ours)                | 36.4            | 10.9             | 7.3               |

Table 7. Depth evaluation on DTU validation set. We show the root mean square error and the percentage of errors fall below 0.2/0.5/1.0mm thresholds.

5. MVS2D is the best on overall score and the second-best completeness score. Such performance is encouraging since our method is quite simple: it is just a single-stage procedure without using any multi-stage refinement as commonly used in recent MVS algorithms ([12, 42, 52]). We show some qualitative results of 3D reconstruction on DTU objects in Figure 5. Qualitatively, our reconstruction is typically more complete on flat surface areas. The behavior is reasonable because our approach utilizes strong single-view priors. We also compare the inference speed with the recent SOTA PatchmatchNet [42]. Our approach yield around 2x speed up as shown in Table 6. Lastly, as our method was mainly designed for multi-view depth estimation, we additionally examine the depth evaluation metrics. Since DTU does not have ground truth depth for the test set, we report the depth evaluation results on the validation set. As expected, MVS2D is better than PatchmatchNet in terms of depth metrics, and the performance gap there is wider than in 3D Reconstruction. The results can be found in Table 7. Comparison on Robustness under Noisy Pose. As shown in Table 7, Ours-robust (multi-scale cues) and Ours (single-scale cues) perform similarly when the input poses are accurate. However, as shown in Table 8, multi-scale aggregation is preferred when the input poses are noisy. It suggests that when having inaccurate training data, it is necessary to incorporate multi-scale cues, though at a cost of increased computations (as shown in Table 1).

Table 8. Different methods’ performance under noisy input poses on ScanNet [7]. We notice that most methods suffer from significant performance drops. Our method with multi-scale epipolar aggregation shows notable robustness.

Ablation Study on Depth Encoding. The ablation study of our depth code design can be found in Table 9. We tested four code types. ‘Uniform’ serves as a sanity check, where we use the same code vector for all depth hypotheses. In other words, the network does not extract useful information from the reference images. ‘Linear’ improves on uniform encoding by scaling a base code vector with the corresponding depth value. ‘Cosine’ codes are identical to the one used in [41]. ‘Learned’ codes are optimized end-to-end. We can see that learning the codes end-to-end leads to noticeable performance gains. One explanation is that these learned codes can adapt to the single-view feature representations of the source image.

Table 9. Ablation study on different depth encodings. We can see that jointly training depth encodings gives the best performance.
pixels on the same epipolar line. We plan to address this issue by developing novel attention mechanisms to explicitly enforce those constraints.

References

[1] Henrik Aanæs, Rasmus Ramsbol Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multi-view stereopsis. *International Journal of Computer Vision (IJCV)*, 120(2):153–168, 2016. 2, 5, 8

[2] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3D: Stereo depth estimation via binary classifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1600–1608, 2020. 2

[3] Neil DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 766–779. Springer, 2008. 8

[4] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1538–1547, 2019. 2, 8

[5] Ricson Cheng, Ziyan Wang, and Katerina Fragkiadaki. Geometry-aware recurrent neural networks for active visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5086–5096, 2018. 3

[6] Shuo Cheng, Zexiang Xu, Shulin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2524–2534, 2020. 8

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5, 6, 8

[8] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. DeepPruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4384–4393, 2019. 2

[9] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Found. Trends. Comput. Graph. Vis.*, 9(1-2):1–148, June 2015. 1, 2

[10] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(8):1362–1376, 2009. 2, 8

[11] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 873–881, 2015. 8

[12] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhao Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2020. 1, 2, 8

[13] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(2):504–511, 2012. 2

[14] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. DeepMVS: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2821–2830, 2018. 1, 2, 6

[15] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. DPSNet: End-to-end deep plane sweep stereo. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 5, 6

[16] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacedenet: An end-to-end 3d neural network for multi-view stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2307–2315, 2017. 8

[17] Abhishek Kar, Christian Hane, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 364–375, 2017. 2

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 5

[20] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 5, 6, 7

[21] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1, 5, 6, 7

[22] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2811–2820, 2018. 1, 2

[23] Xiaoxiao Long, Lingjie Liu, Wei Li, Christian Theobalt, and Weinping Wang. Multi-view depth estimation using epipolar spatio-temporal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[24] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10452–10461, 2019. 8

[25] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1590–1599, 2020. 2

[26] Nikolaus Mayer, Eddy Ilg, Philip Hauser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A
large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4040–4048, 2016. 1, 2

[27] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3D scene reconstruction from posed images. In Proceedings of the European Conference on Computer Vision (ECCV), 2020. 1, 2

[28] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3283–3291, 2019. 1, 2

[29] Alex Poms, Chenglei Wu, Shou-I Yu, and Yaser Sheikh. Learning patch reconstructability for accelerating multi-view stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3041–3050, 2018. 2

[30] Xiaojuan Qi, Renjie Liao, Zhenghe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 283–291, 2018. 1

[31] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. Advances in Neural Information Processing Systems (NeurIPS), 2019. 2

[32] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4104–4113, 2016. 6

[33] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 464–468, 2018. 2

[34] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012. 2, 5

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8934–8943, 2018. 2

[36] Vladimir Tankovich, Christian Hane, Sean Fanello, Yinda Zhang, Shahram Izadi, and Sofien Bouazizi. HITNet: Hierarchical iterative tile refinement network for real-time stereo matching. arXiv preprint arXiv:2007.12140, 2020. 2

[37] Josh Tobin, OpenAI Robotics, and Pieter Abbeel. Geometry-aware neural rendering. Advances in Neural Information Processing Systems (NeurIPS), 2019. 3

[38] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra-high-resolution image sets. Machine Vision and Applications, 23(5):903–920, 2012. 8

[39] Hsiao-Yu Fish Tung, Ricson Cheng, and Katerina Fragkiadaki. Learning spatial common sense with geometry-aware recurrent networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2595–2603, 2019. 3

[40] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Niko- laus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5622–5631, 2017. 2, 5, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 2017. 2, 4, 8

[42] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 14194–14203, 2021. 2, 5, 6, 7, 8

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7794–7803, 2018. 2

[44] Yuesong Wang, Tao Guan, Zhuo Chen, Yawei Luo, Keyang Luo, and Lili Ju. Mesh-guided multi-view stereo with pyramid architecture. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2039–2048, 2020. 2

[45] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 1625–1632, 2013. 1, 2, 5

[46] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1

[47] Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1959–1968, 2020. 2

[48] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In AAAI Conference on Artificial Intelligence (AAAI), pages 12508–12515, 2020. 8

[49] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4877–4886, 2020. 8

[50] Zhenpei Yang, Li Erran Li, and Qixing Huang. Strumononet: Learned stereo matching network for stereo depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 14175–14184, 2021. 1

[51] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2595–2603, 2019. 3

[52] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), pages 767–783, 2018. 1, 2, 5, 6, 8
multi-view stereo depth inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5525–5534, 2019. 1, 2, 8

[53] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2

[54] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 5684–5693, 2019. 1

[55] Zehao Yu and Shenghua Gao. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1949–1958, 2020. 1, 2, 5, 6, 7, 8

[56] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1592–1599, 2015. 2

[57] Xudong Zhang, Yutao Hu, Haochen Wang, Xianbin Cao, and Baochang Zhang. Long-range attention network for multi-view stereo. In IEEE Winter Conference on Applications of Computer Vision (WACV), pages 3782–3791, 2021. 2