Computer Methods and Programs in Biomedicine 244 (2024) 107937

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Spatio-temporal layers based intra-operative stereo depth estimation network via hierarchical prediction and progressive training

Ziyang Chen [a,*], Laura Cruciani [a], Elena Lievore [b], Matteo Fontana [b], Ottavio De Cobelli [b,c], Gennaro Musi [b,c], Giancarlo Ferrigno [a], Elena De Momi [a,b]

[a] Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milano, 20133, Italy
[b] European Institute of Oncology, Department of Urology, IRCCS, Milan, 20141, Italy
[c] University of Milan, Department of Oncology and Onco-haematology, Faculty of Medicine and Surgery, Milan, Italy

ARTICLE INFO

Keywords:
Robotic surgery
Intra-operative
Depth estimation
Deep learning
Stereo images

ABSTRACT

Background and Objective: Safety of robotic surgery can be enhanced through augmented vision or artificial constraints to the robotl motion, and intra-operative depth estimation is the cornerstone of these applications because it provides precise position information of surgical scenes in 3D space. High-quality depth estimation of endoscopic scenes has been a valuable issue, and the development of deep learning provides more possibility and potential to address this issue.

Methods: In this paper, a deep learning-based approach is proposed to recover 3D information of intra-operative scenes. To this aim, a fully 3D encoder-decoder network integrating spatio-temporal layers is designed, and it adopts hierarchical prediction and progressive learning to enhance prediction accuracy and shorten training time.

Results: Our network gets the depth estimation accuracy of MAE 2.55±1.51 (mm) and RMSE 5.23±1.40 (mm) using 8 surgical videos with a resolution of 1280×1024, which performs better compared with six other state-of-the-art methods that were trained on the same data.

Conclusions: Our network can implement a promising depth estimation performance in intra-operative scenes using stereo images, allowing the integration in robot-assisted surgery to enhance safety.

1. Introduction

Nowadays, Robot-Assisted Minimally Invasive Surgery (RAMIS) has gradually shown more advantages compared to traditional open surgery because it enhances the flexibility and accuracy of the operation, and reduces the bleeding rate and post-operative recovery time. Nonetheless, the safety of minimally invasive surgery remains well-researched due to the limited field of view of the endoscope and the lack of haptic feedback for surgeons [1,2]. Two emerging computer-aided technologies, Augmented Reality (AR) and Virtual Fixtures (VF), are mainstream research hotspots in the field of surgical robotics today. AR can provide surgeons with visual surgical guidance by real-time registration [3,4] between pre-operative models taken by Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) and intra-operative tissues, while VF can apply preemptive force feedback on the end of manipulator held by surgeons to avoid collisions with delicate tissues during operation [5,6]. To implement these auxiliary means for the enhancement of surgical safety, recovering 3D information of the intra-operative scene is a fundamental and significant problem since it directly determines the accuracy of AR registration and VF force feedback. Hence, an open technical challenge in the medical field is to accurately and fastly estimate the depth of intra-operative soft tissues.

The development potential of depth estimation using stereo endoscopic images can be foreseen because it does not require the movement of endoscope compared with the monocular depth estimation, which is more in line with surgical scenarios where the endoscope often remains stationary during the operation. Some famous robot-assisted surgical systems have been utilized in clinical treatment, such as the da Vinci Surgical System (dVSS, Intuitive Surgical Inc., USA) [7] which is one of the most used platforms in RAMIS. Surgeons can perceive depth information inside a patient's body by viewing the left and right endoscopic images simultaneously when remotely manipulating the robot, even though the endoscope remains fixed. Stereo endoscopic vision systems can recover the 3D shape of a scene surface by generating the dispar-

* Corresponding author.
E-mail address: ziyang.chen@polimi.it (Z. Chen).

https://doi.org/10.1016/j.cmpb.2023.107937
Received 7 November 2022; Received in revised form 18 November 2023; Accepted 19 November 2023
Available online 22 November 2023

ity of stereo images based on stereo matching [8]. It means searching for the corresponding left and right pixels along the epipolar line on the stereo image. Generally, image rectification is a necessary step before stereo matching because it can align the left and right epipolar lines horizontally, thereby reducing the complexity of the pixel search. After obtaining the disparity value of each pixel pair, the corresponding depth information is easily recovered by combining the camera focal length and baseline distance based on the triangular projection.

Generally speaking, classical stereo matching methods include matching cost calculation, cost aggregation, disparity calculation and refinement [9,10], and image pre-processing and post-processing are always integrated to improve the efficiency of stereo matching. The authors in [11] proposed an improved census transform to enhance the robustness of illumination following the above pipeline, and simple linear iterative clustering was implemented to fill the holes in the disparity map as a post-processing step. The experimental results showed that this strategy can reconstruct the surgical scene densely based on the Hamlyn phantom heart dataset [12], although it is time-consuming and loses some matched pixels. A sparse to semi-dense points reconstruction strategy for the stereo endoscopic domain was proposed in [13], which does not rely on specific feature matching approaches. They adopted the Zero Mean Normalized Cross Correlation (ZNCC) to measure the dissimilarity during structure propagation because it provides higher robustness for regions with poor illumination and texture. The experimental figures presented better accuracy compared with the three other existing methods, but it remains some holes in the generated disparity maps.

To improve the speed of stereo depth estimation, a GPU-based quasi-dense matching method was proposed in [14] to restore 3D surgical information in real time. The authors recovered a set of sparse feature points robustly, and then improved the matching accuracy by updating the disparity from semi-dense to dense levels using the same ZNCC evaluation. It can recover 3D surgical information at approximately 22 Frames Per Second (FPS) using the image pair with a resolution of 360×288, but the issue of missing reconstructed points remains to be solved. Similarly, after utilizing the SIFT descriptor to perform sparse matching, the authors in [15] performed dense correspondence by calculating the patch similarity of stereo image pairs based on the normalized cross-correlation metric, and three strict confidence criteria were added to enhance the robustness and post-processing was implemented to remove outliers that were significantly different from their neighborhood. Qualitative results showed a dense reconstruction effect compared to the other two methods, although the quantitative data were insufficient. In [16], the authors designed a novel cost function consisting of a data term and a local as well as a non-local smoothness term to search for the optimal disparity values globally, then the disparity map was upsampled based on an improved bilateral interpolation strategy, which achieved a promising performance on the reconstruction accuracy as well as the points of interest using an endoscopic phantom dataset. However, the globalized search strategy consumes high computing resources, which hinders real-time performance.

On the other hand, deep learning-based stereo correspondence has also been introduced in the past few years and achieved a more promising stereo depth estimation performance compared with the above parametric approaches. A Convolution Neural Network (CNN) integrating a Spatial Pyramid Pooling (SPP) module was utilized in [10] to extract the features of the stereo image and build the cost volume, and then regressed the disparity map through the 3D stacked hourglass architecture. This method can effectively predict the disparity value, but the stacked hourglass modules affect the inference speed, and the generalization also needs to be enhanced. Furthermore, an encoder-decoder framework with a coarse-to-fine prediction was designed in [17] to generate disparity values. They upsampled the 3D feature maps gradually in the decoder and concatenated the low-level feature maps from the encoder, which showed satisfactory performance on high-resolution image pairs. The authors in [18] streamlined hourglass modules by removing short connections, and the group-wise correlation layers were concatenated into the cost volume to contain more similarity-measure features. Quantitative evaluation based on two general datasets showed this approach could save inference time without diminishing accuracy. Following the standard pipeline for stereo matching, the authors [19] first introduced the Neural Architecture Search (NAS) strategy to select the optimal architectures for modules that contain trainable parameters, which could save computational resources and increase accuracy during the search process, and it achieved the top ranking in public datasets.

It was noticed that the component of cost volume heavily influences the final regression, so a cascade cost volume module was designed in [20]. The difference compared with the previous volume is that the hypothesis range and plane interval gradually reduced in a coarse-to-fine operation instead of keeping fixed, and the result showed that existing models integrating the new cost volume could improve both the accuracy and inference time. Next, the authors in [21] adopted a pyramid-shaped module to extract features, then constructed a fused cost volume to predict the coarsest disparity map, and finally a cascade cost volume was implemented to refine the disparity maps using the variance-based uncertainty estimation. More recently, some new operations continue to be proposed for the construction of cost volumes. Different from the classical feature concatenation, the authors in [22] constructed the cost volume using cosine similarity to enlarge pure similarity information and enhance generalization. Although the above methods can achieve promising performance on some natural datasets, the results in the field of medical images remain to be evaluated.

Considering the insufficiency of annotated medical image datasets, some unsupervised learning methods have also been introduced owing to the emergence of Generative Adversarial Networks (GAN) [23]. For instance, an adversarial depth estimation model was proposed in [24] to predict depth values without ground truth. They implemented a generator to predict the left and right disparity maps, and then reprojected them to the RGB images and input them to a discriminator to compare the difference between the original RGB images and the newly generated ones. Although these unsupervised learning-based methods [25] can overcome the limitation of insufficient labels in the medical domain, their depth estimation accuracy remains to be enhanced.

It can be seen that the above methods always consider the depth estimation performance of a single image pair, which means the spatial feature is utilized. However, medical images are continuous in the temporal domain, so the temporal attribute can also be processed and explored to measure the stereo depth estimation performance. Spatial-temporal layers, i.e., feature maps composed of spatial and temporal features by encoding video clips, were adopted to help improve deep learning-based prediction quality in some closed fields, including robotic instrument articulation detection [26], preterm infants' pose estimation [27] and inter-fetal membrane segmentation [28]. Applications in these similar image processing fields inspired us to explore the possibility of utilizing spatial-temporal layers in intra-operative stereo depth estimation, since it has not yet been implemented in this field. Furthermore, some end-to-end deep learning models have started to adopt a coarse-to-fine manner for their prediction instead of only outputting the final layer [17,18] because it is beneficial to the refinement of terminal features. We would also conduct this manner to explore its influence when inputting consecutive frames to the neural network instead of the traditional input of a single frame. Finally, progressive training using adjustable regularization conditions inspired by [29] was introduced in our stereo depth estimation network to speed up the training process. Here, adjustable regularization conditions refer to multi-scale image augmentation [30] of training datasets according to different sizes of training images because it could promote the learning performance of the model and reduce the risk of overfitting. Specifically, we input image pairs of three different sizes with data augmentation scales from weak to strong to train the network, to explore

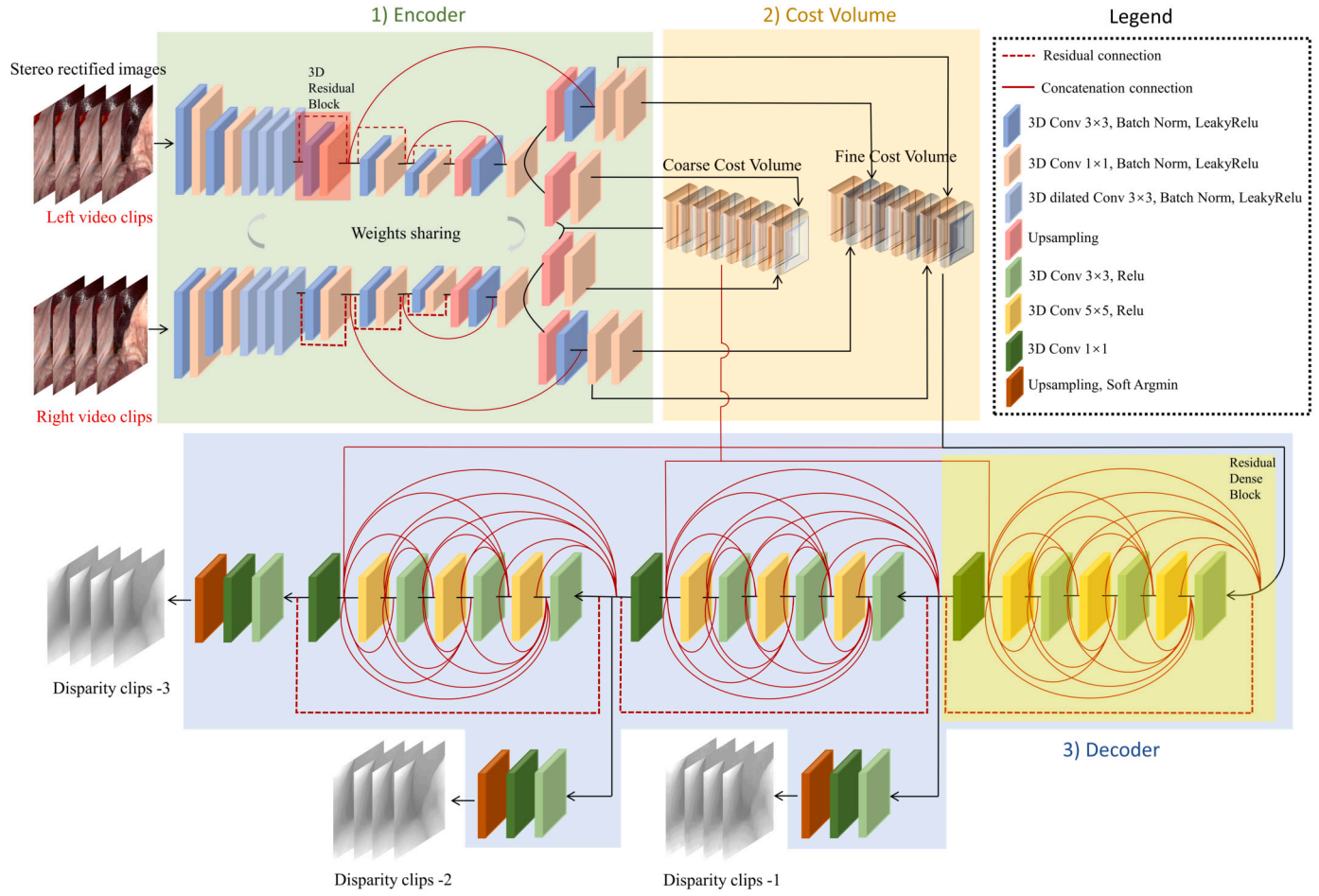

Fig. 1. The proposed FESDEnet architecture. Inputs are the rectified temporal clips, and outputs are the estimated disparity clips with three different levels in a coarse-to-fine manner. The black arcs show the transmitting direction of features, the dashed red arcs indicate the residual connection by the sum operation, and the solid red arcs denote the concatenation connection between two feature maps. A 3D U-shaped encoder is designed to extract high-level features from the RGB clips, and then two different levels of cost volumes are generated based on the extracted feature maps. Next, they are transmitted to the extended 3D residual dense blocks respectively, and finally the disparity clips are estimated after the soft argmin operation using hierarchical prediction. In the encoder, a transparent red mask is placed to indicate the composition of the 3D residual block, which consists of two convolutional layers and a residual connection. In the decoder, a transparent yellow mask is placed to show the components of a 3D residual dense block.

the effect of progressive training in medical depth estimation models. Based on the above considerations, we proposed three research hypotheses:

- Hypothesis 1 (H1): The spatio-temporal layers integrating into the convolutional neural network can enhance the intra-operative stereo depth estimation quality by predicting sequential frames in the temporal domain.
- Hypothesis 2 (H2): Hierarchical prediction in a coarse-to-fine manner can refine the final regression effect when the input of the network is consecutive frames.
- Hypothesis 3 (H3): Progressive training combined with regularization conditions at different scales can speed up the training time of the network, but the predictive ability will not be impaired.

In this paper, we address the problem of dense depth estimation of intra-operative scenes to enhance surgical safety because it is a solid cornerstone for the potential integration with AR and VF. More specifically, we proposed a fully 3D encoder-decoder architecture to perform stereo matching accurately and robustly. Different from the previous work in which a single image pair was always regarded as the input, we chose surgical video clips for encoding more inter-frame information. Hence, we designed a 3D U-shaped encoder to extract the high-level features of video clips, and the third dimension processed the temporal information between consecutive frames. 3D U-Net is an extension from the classic 2D U-Net architecture [31] by replacing all 2D oper-

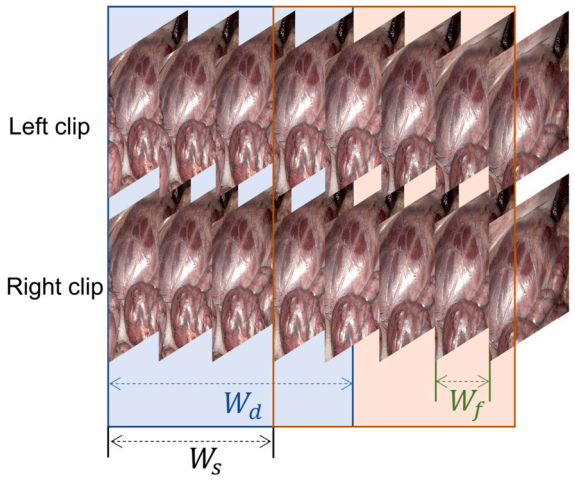

Fig. 2. The implementation of sliding window algorithm based on the stereo images. W_d is the number of frames of the temporal sliding window, and W_s is the skipping stride of the window. It overlaps one frame, as shown in the figure, when W_d and W_s are equal to 4 and 3, respectively. W_f is the number of frames between two sampled frames, and the sampled frames are consecutive when W_f is 0.

Table 1
Parameters of the proposed FESDEnet. The components of each module and the corresponding outputs are provided. The stereo image pairs share weights from layers 2 to 16. H is the height of the feature maps, W is the width, W_d is the number of frames of the temporal sliding window, and D denotes the maximum disparity hypothesis value. The symbol "∼" is the abbreviation to contain multiple consecutive nodes. Here, Conv3D 3×3 means the kernel size is 3×3×3, Conv3D 1×1 means the kernel size is 1×1×1, and Conv3D 5×5 means the kernel size is 5×5×5. The first RDB module consists of layers 18 to 24, the second RDB module contains layers 25_2, 26_2 to 31, and the third one contains layers 32_2, 33_2 to 38.

ID	Layer	Output	Connected to
1	Input	H×W×Wd×3	2
2	Conv3D 3×3, BN, LeakyRelu	H/2×W/2×Wd×16	3
3	Conv3D 1×1, BN, LeakyRelu	H/2×W/2×Wd×32	4
4	Conv3D 3×3, BN, LeakyRelu	H/4×W/4×Wd×32	5
5	Conv3D 1×1, BN, LeakyRelu	H/4×W/4×Wd×32	6
6	Dilated Conv3D 3×3, BN, LeakyRelu	H/4×W/4×Wd×32	7
7	Dilated Conv3D 3×3, BN, LeakyRelu	H/4×W/4×Wd×32	8
8	Dilated Conv3D 3×3, BN, LeakyRelu	H/4×W/4×Wd×32	9
9	3D Residual block	H/8×W/8×Wd×64	10, 15
10	3D Residual block	H/16×W/16×Wd×128	11, 13
11	3D Residual block	H/32×W/32×Wd×128	12
12	Upsampling, Conv3D 3×3, BN, LeakyRelu	H/16×W/16×Wd×64	13
13	Conv3D 1×1, BN, LeakyRelu	H/16×W/16×Wd×120	14_1 ∼14_3
14_1	Upsampling	H/8×W/8×Wd×120	17_1
14_2	Upsampling, Conv3D 1×1, BN, LeakyRelu	H/8×W/8×Wd×12	17_1
14_3	Upsampling, Conv3D 3×3, BN, LeakyRelu	H/8×W/8×Wd×64	15
15	Conv3D 1×1, BN, LeakyRelu	H/8×W/8×Wd×120	16, 17_2
16	Conv3D 1×1, BN, LeakyRelu	H/8×W/8×Wd×12	17_2
17_1	Coarse Cost Volume	H/8×W/8×(Wd×D/8)×32	24, 31
17_2	Fine Cost Volume	H/8×W/8×(Wd×D/8)×32	18 ∼25_2
18	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×48	19 ∼24
19	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×64	20 ∼24
20	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×80	21 ∼24
21	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×96	22 ∼24
22	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×112	23,24
23	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×128	24
24	Conv3D 1×1	H/8×W/8×(Wd×D/8)×32	25_1, 25_2, 26_2 ∼32_2
25_1	Conv3D 3×3, Relu, Conv3D 1×1	H/8×W/8×(Wd×D/8)×1	26_1
25_2	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×48	26_2 ∼31
26_1	Upsampling, Soft Argmin	H×W×Wd	Output_1
26_2	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×64	27 ∼31
27	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×80	28 ∼31
28	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×96	29 ∼31
29	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×112	30, 31
30	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×128	31
31	Conv3D 1×1	H/8×W/8×(Wd×D/8)×32	32_1, 32_2, 33_2 ∼39
32_1	Conv3D 3×3, Relu, Conv3D 1×1	H/8×W/8×(Wd×D/8)×1	33_1
32_2	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×48	33_2 ∼38
33_1	Upsampling, Soft Argmin	H×W×Wd	Output_2
33_2	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×64	34 ∼38
34	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×80	35 ∼38
35	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×96	36 ∼38
36	Conv3D 3×3, Relu	H/8×W/8×(Wd×D/8)×112	37 ∼38
37	Conv3D 5×5, Relu	H/8×W/8×(Wd×D/8)×128	38
38	Conv3D 1×1	H/8×W/8×(Wd×D/8)×32	39
39	Conv3D 3×3, Relu, Conv3D 1×1	H/8×W/8×(Wd×D/8)×1	40
40	Upsampling, Soft Argmin	H×W×Wd	Output_3

ations with their 3D counterparts, and it has been widely adopted in medical image segmentation by inputting 3D data such as CT slices and magnetic resonance images [32,33]. Also, 3D U-Net has started to be utilized in more image processing fields, including depth estimation [34]. On the one hand, 3D U-Net maintains a U-shape architecture with a skip connection to fuse multi-scale features for fine prediction. On the other hand, it provides the possibility of encoding 3D data such as video clips since the 3D convolution operation can slide along the temporal dimension. Furthermore, we extended the original 2D residual dense block [35] into a 3D version as a basic module in the decoder, and the third dimension processed the disparity information. To the best of our knowledge, this is the first work to explore the potential of a fully 3D network with spatio-temporal layers, hierarchical prediction and progressive training in the intra-operative stereo depth estimation domain. 8 videos (3016 image pairs) containing different surgical scenes were used to evaluate our network and verify the proposed hypotheses based on a comprehensive comparison study and an ablation study. Statistical tests were also performed to calculate significant differences in these studies.

2. Methodology

A Fully 3D Endoscopic Stereo Depth Estimation network (FESDEnet) is proposed to perform dense depth estimation, and the overall framework is illustrated in Fig. 1 and consists of an encoder-decoder architecture with long-short skip connections. Specifically, the encoder is established using a 3D U-shaped network to generate high-level feature representations, then the extracted feature maps are combined to foster cost volumes as the input of the decoder, and finally the decoder consisting of 3D residual dense blocks is implemented to predict final results. Table 1 presents the detailed parameters in our network. To encode the temporal information within inter-frame similarity, we choose to input the video clip instead of the traditional single frame. A sliding window algorithm [26] is integrated to generate temporal clips based on the

stereo image pairs, as shown in Fig. 2. Starting from the first stereo clip consisting of W_d frames, the window will slide to the next clip along the temporal sequence with a stride of W_s frames. W_f refers to the number of frames between two sampled frames along the temporal direction in the sliding window, and $W_f = 0$ means the two sampled frames are consecutive. Hence, the input of the network can be extended to 4D data volume ($H \times W \times W_d \times 3$).

1) A 3D U-shaped encoder is designed to extract high-level feature maps from the video clips. We adopt the 3D convolutional layer with the size of $3 \times 3 \times 3$ and $1 \times 1 \times 1$ alternately to learn the inter-frame information within a video clip and limit the parameter complexity. Also, we add three dilated convolutional layers to expand the receptive fields, and then three residual blocks [36] with the striding operation are exploited for unary feature extraction inspired by [10]. Next, the upsampling layer combined with the convolutional block is utilized to increase the size of feature maps. Here, the twin network is adopted through weight sharing to extract left and right high-level feature maps, respectively.

2) The 3D stacked cost volume is constructed in a coarse-to-fine manner. More specifically, we consider both feature difference measure [17] and group-wise correlation [18] (i.e., dividing the feature maps into some groups along the channel dimension after the inner product operation, and generating the correlation maps by the mean operation in each group) to form a stacked cost volume C^i_{stacked}, and temporal features are integrated using layer concatenation. The final cost volume C^i_{stacked} can be formulated as

$$\begin{aligned} C^i_{1\_k}\left(d^i,x,y,f\right) &= \left|f^i_{l\_k}(x,y) - f^i_{r\_k}\left(x-d^i,y\right)\right| \\ C^i_{2\_k}\left(d^i,x,y,g\right) &= \frac{1}{N_c/N_g}\left\langle f^{ig}_{l\_k}(x,y), f^{ig}_{r\_k}\left(x-d^i,y\right)\right\rangle \\ C^i_{\text{stacked}} &= \left\| C^i_{1\_k}, C^i_{2\_k}\right\|_{k=1}^{K} \end{aligned} \tag{1}$$

Where $C^i_{1\_k}$ is calculated by measuring the difference between left and right feature maps, while $C^i_{2\_k}$ considers their correlation. $||$ is the absolute value symbol, $\langle,\rangle$ represents the inner product, and $\|,\|$ denotes the vector concatenation operation. f_l and f_r are the left and right feature maps, N_c is the number of extracted feature channels, while N_g represents the number of feature groups ($N_g = 20$ in this work). g represents the feature groups (i.e., group-wise feature maps), which means the division of feature maps in the channel dimension after the inner product so that each feature group contains N_c/N_g feature maps. After adopting the mean operation along the channel dimension in each group, we can generate the corresponding group-wise correlation maps. i is the different stage of cost volume (coarse or fine), k denotes the frame k in the video clip, and d is the specific disparity value from the full disparity hypothesis range (i.e., from 0 to the maximum disparity value).

3) To estimate the disparity map at a high-accurate pixel level, we design a 3D Residual Dense Block (RDB) as the basic module in the decoder, so the module can refine features of cost volumes in height, width, as well as the third dimension containing the disparity values in the full hypothesis range. Larger convolution kernels ($3 \times 3 \times 3$ and $5 \times 5 \times 5$ adding biases) are alternately introduced in the residual dense block to expand the receptive field when extracting features, and batch normalization layers are removed to decrease computational resources. The concatenated feature maps are further transited to a $1 \times 1 \times 1$ convolutional layer (with a bias) to squeeze the number of channels, and the residual connection is introduced to improve the network representation ability [35]. Here, we concatenate the stacked coarse cost volume to the first two RDB modules, and the fine one is concatenated to the start and end of the RDB backbone to integrate multi-level features in the decoder. Finally, the feature maps are upsampled to the original resolution to generate the matching cost, and the soft argmin operation [37] is adopted to regress the disparity maps (i.e., converting the matching cost to a probability volume along the disparity dimension using the softmax function, and taking the sum by multiplying all the disparity hypothesis values with their corresponding probabilities). In this case, we could compare the difference between the predicted disparity map and the ground truth since they have the same resolution. The estimated disparity value $\tilde{d}_k$ of frame k can be defined as [37],

$$\tilde{d}_k = \sum_{d=0}^{D_{\max}} d \times \sigma\left(-c_{d\_k}\right) \tag{2}$$

Where $D_{\max}$ denotes the maximum disparity hypothesis value, c_{d_k} is the predicted matching cost, and $\sigma(\cdot)$ represents the softmax operation. Here, c_{d_k} means the predicted cost when the disparity value is d in the k-th frame, and it is generated from the decoder before the softmax operation.

Finally, the "smooth L1" loss function is adopted to train the network since it has strong robustness and low sensitivity to outliers reported in [38]. Also, the hierarchical prediction is implemented to estimate the disparity maps in three different levels, and the formula can be denoted as,

$$L(d,\tilde{d}) = \sum_{l=1}^{3} \lambda_l \left(\sum_{k=1}^{K} \sum_{n=1}^{N} \text{smooth}_{L_1}\left(d_{n\_k} - \tilde{d}_{n\_k}\right)\right) \tag{3}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{4}$$

Where λ_l is the loss weight in different prediction levels, and λ_1 is the coarsest estimation while λ_3 is the finest one. N is the total pixel number, d_{n_k} is the ground truth of the disparity value in the n-th pixel of the k-th frame, and $\tilde{d}_{n\_k}$ is the estimated one.

3. Experimental protocol

3.1. Dataset

Three annotated datasets with consecutive frames were used to train our network, including Scene Flow [39], Sintel [40] as well as a self-made dataset. Specifically, there are three scenes in the Scene Flow dataset, and two of them (Driving and Monkaa) were selected for the training after checking manually, since the remaining one (FlyingThings3D) has an apparent difference between two consecutive frames. Furthermore, all frames from the Sintel dataset were utilized to train the model. To enhance the generalization of our network, we also created a dataset based on Vision Blender [41] consisting of five different phantom scenes. 5000 consecutive image pairs (640×480) with the ground truth were collected in 25 FPS. This self-made dataset is now available through this link: https://drive.google.com/file/d/1DaNDHde2fk21CoP1iWoCpm_-lptDQays/view?usp=sharing. The information about how to make this dataset was introduced by the authors of Vision Blender: https://github.com/Cartucho/vision_blender.

To conduct a comprehensive test, we used the public SCARED test dataset [42] to perform the qualitative and quantitative evaluation. It was captured using porcine cadavers based on the da Vinci Xi surgical robot, and it provides 8 surgical videos (1280×1024) with the ground truth of consecutive sparse point clouds. After checking the ground truth, we adopted the images with the ground truth containing valid points over 30 percent to avoid the possible shifting error in the ground truth, so we have a total of 3016 stereo images for the evaluation. Considering that the SCARED dataset is not rectified, we rectified the stereo images and remapped the predicted depth maps to the original coordinates by calling the cv2.stereoRectify and cv2.remap functions, and compared them with the original ground truth. The details of the training and test datasets are given in Fig. 3.

3.2. Training strategy

Our network was trained from scratch using these three datasets, i.e., Scene Flow, Sintel and the self-made synthetic dataset. Temporal

Table 2
Quantitative evaluation based on 8 different surgical videos. For instance, "V1 (392)" means video 1 with 392 consecutive stereo images, and 'ALL' is the final result using all videos. Smaller values are better, and the value with the best performance in each metric is bold.

	Videos	Ref. [9]	Ref. [10]	Ref. [18]	Ref. [19]	Ref. [21]	Ref. [22]	FESDEnet
MAE (mm)	V1 (392)	6.73±2.39	8.16±2.09	12.28±5.40	6.78±1.60	5.30±1.56	8.17±1.55	**5.11±1.53**
	V2 (470)	2.93±0.90	3.32±1.05	5.74±2.53	2.87±0.77	2.22±0.57	3.60±0.97	**2.16±0.58**
	V3 (559)	1.88±0.66	3.11±2.32	5.70±4.14	2.36±1.34	1.86±0.96	2.27±1.23	**1.61±0.70**
	V4 (752)	2.32±1.06	3.53±2.01	6.02±5.27	2.90±1.27	2.19±0.73	2.16±1.10	**1.95±0.67**
	V5 (476)	4.14±0.57	4.57±0.70	4.15±0.63	4.04±0.54	4.06±0.63	**3.59±0.38**	3.92±0.64
	V6 (243)	1.30±0.21	1.29±0.21	1.24±0.18	1.34±0.23	1.08±0.19	1.46±0.21	**1.04±0.17**
	V7 (81)	2.14±0.39	2.29±0.24	1.97±0.16	1.75±0.21	1.60±0.12	1.61±0.16	**1.47±0.14**
	V8 (43)	2.09±0.09	1.65±0.10	1.64±0.08	1.94±0.19	1.56±0.08	1.83±0.10	**1.45±0.07**
	ALL (3016)	3.10±1.98	3.94±2.48	5.88±4.81	3.31±1.87	2.72±1.55	3.34±2.25	**2.55±1.51**
RMSE (mm)	V1 (392)	12.72±6.35	14.69±4.50	22.23±8.39	13.30±5.65	7.72±1.43	14.27±3.41	**7.42±1.55**
	V2 (470)	9.71±2.76	9.13±3.67	12.58±4.81	7.71±4.04	4.72±0.86	9.87±4.53	**4.50±0.87**
	V3 (559)	8.82±1.48	9.40±4.00	13.83±6.29	9.02±4.31	5.70±1.42	9.60±3.92	**5.16±0.81**
	V4 (752)	9.30±3.32	9.39±3.67	12.87±7.54	8.18±3.28	5.19±0.56	7.22±4.09	**4.82±0.50**
	V5 (476)	9.56±1.39	11.17±3.68	6.72±0.51	8.21±2.07	6.48±0.49	7.57±0.98	**6.16±0.46**
	V6 (243)	8.07±0.98	4.58±1.15	3.82±0.98	5.85±2.70	3.33±0.73	6.35±0.83	**3.18±0.69**
	V7 (81)	11.33±1.72	11.95±1.94	5.54±0.57	5.98±1.15	4.30±0.69	6.84±1.20	**4.12±0.65**
	V8 (43)	12.89±0.53	5.54±0.79	5.26±0.42	7.49±2.36	5.03±0.43	6.40±0.45	**4.82±0.39**
	ALL (3016)	9.77±3.44	9.95±4.42	12.21±7.77	8.68±4.25	5.57±1.50	8.95±4.20	**5.23±1.40**
Inference Time (s)	ALL(3016)	0.39±0.04	0.88±0.04	0.65±0.09	0.99±0.01	0.45±0.09	0.61±0.01	*1.12±0.00

* The inference time of FESDEnet includes the prediction of four consecutive frames, while the inference time of other methods includes the prediction of a single frame.

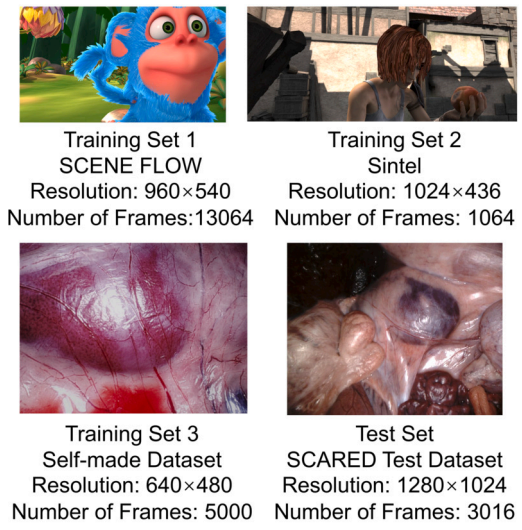

Fig. 3. The details of the training and test datasets.

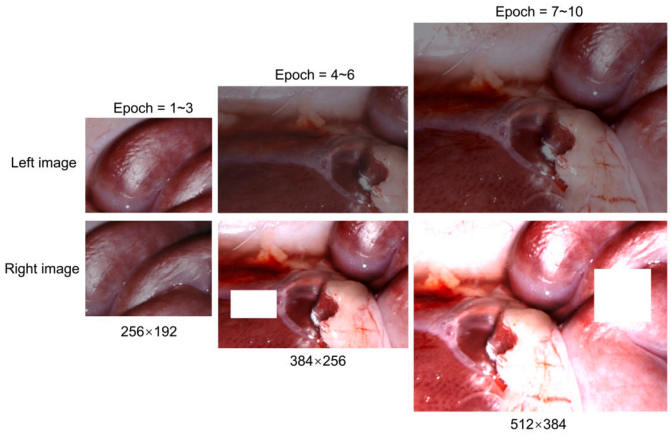

Fig. 4. The implementation of progressive training. Image pairs of three different sizes are sequentially fed into the network with progressively growing data augmentation.

clips as inputs were made using the mentioned window sliding algorithm with $Wd = 4$ to pack consecutive four frames as a clip, because it was found to have the best performance based on the following ablation study. To maximize the usage of datasets for the training, we set W_s as 1 to skip only one frame, while it was set as 4 in the test part to save the inference time. In this way, it will not predict the repeated frames during the evaluation. W_f was always set as 0 because the training datasets provide different scales between consecutive two frames by checking inter-frame similarity, which enhances the adaptability of our network.

Progressive training [29] was introduced to train our network, as shown in Fig. 4. Three different sizes of image pairs (256×192, 384×256, 512×384, respectively) were adopted combining with weak-to-strong data augmentation (gradually increasing the threshold and probability) which includes adjusting brightness, gamma, contrast, random scaling and cropping. Here, we chose the cropping operation instead of downsampling to enlarge the training datasets. To enhance the robustness of our network to illumination changes, we randomly added white rectangles to occlude some pixel values. Furthermore, we trained the network with epochs = 10 and the maximum disparity hypothesis value $D_{\max} = 191$. For the batch size, we set 24 during the first three epochs with the smallest image pairs, 12 during the middle three epochs, and 4 for the last four epochs with the largest image pairs. The learning rate was set as 0.001 in the first 9 epochs while it was reduced to 0.0001 in the last epoch. The training process was implemented on a Red Hat Linux server with 4 NVIDIA V100 GPUs (16 GB).

3.3. Performance metrics

Two common evaluation metrics for depth estimation were adopted to measure the prediction accuracy, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [43],

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\bar{d}_i(x,y) - d_i'(x,y)\right| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\bar{d}_i(x,y) - d_i'(x,y)\right|^2} \tag{6}$$

Where N is the total number of pixel points (x, y), $\bar{d}_i$ is the ground truth of the depth value, while d_i' denotes the predicted depth value. It can be seen that MAE is able to measure the predicted depth performance comprehensively, while RMSE is more sensitive to outliers, since

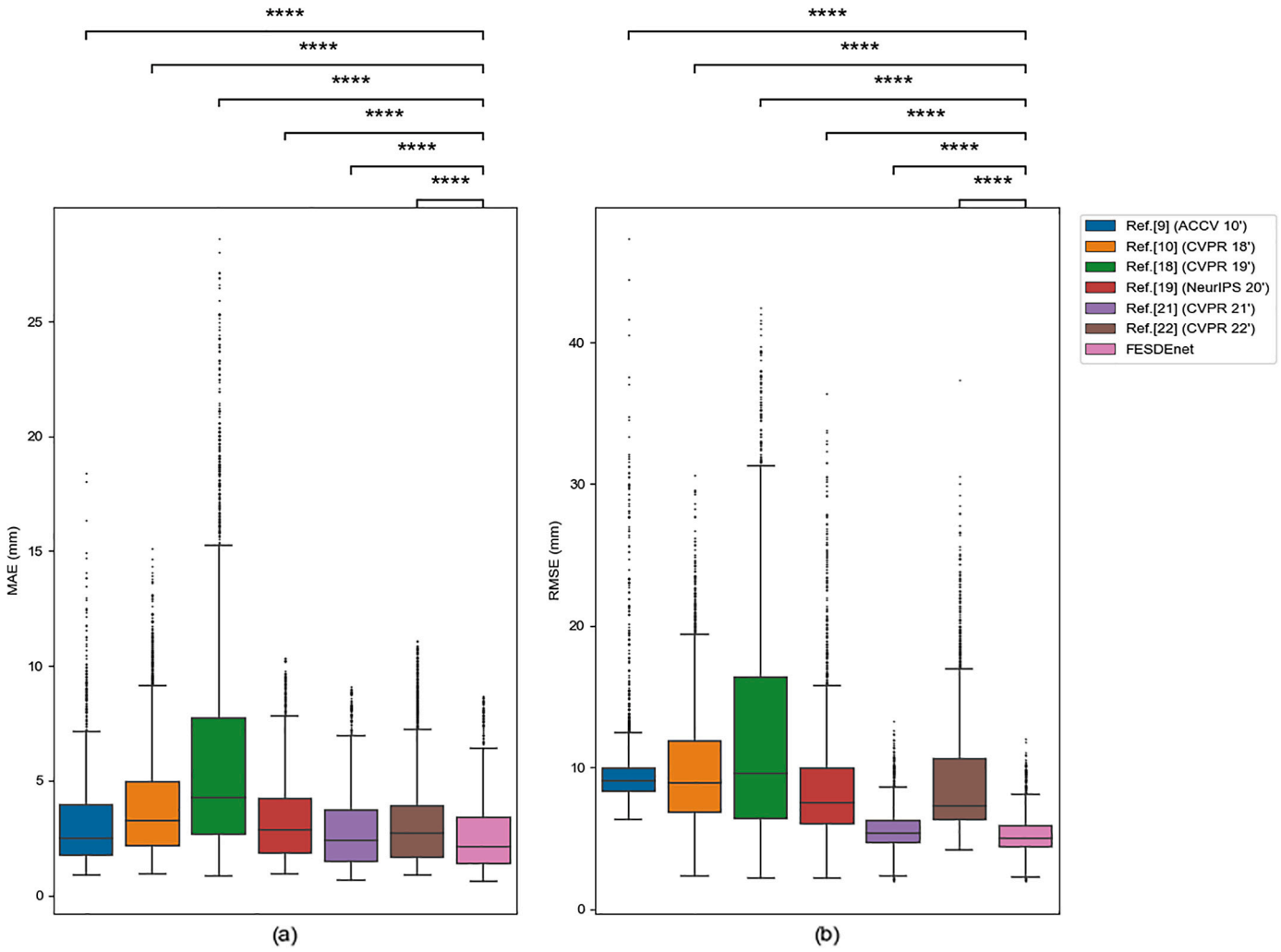

Fig. 5. The box plots related to MAE (a) and RMSE (b) using 8 surgical videos consisting of 3016 image pairs. The Mann-Whitney U test was conducted to check if they are significantly different between other methods and our FESDEnet, and the result is shown as ns : $0.05 < p \leq 1$, $*$: $0.01 < p \leq 0.05$, $**$: $0.001 < p \leq 0.01$, $***$: $0.0001 < p \leq 0.001$, and $****$: $p \leq 0.0001$.

it will be more affected when outliers exist compared with MAE. Hence, it is necessary to evaluate both errors to see the robust performance related to accuracy. Moreover, the inference time was also calculated as an important metric to evaluate the prediction speed.

3.4. Ablation study

As our final model, we chose to set W_d as 4, which means the temporal clip was composed of four consecutive frames. Also, the weight combination λ_l in the loss function was set as $\lambda_1 = 0.3, \lambda_2 = 0.5, \lambda_3 = 1$. To evaluate the proposed three hypotheses and find out the best configuration for the network, we conducted the ablation study based on four groups of experiments:

Group 1: To evaluate the integration performance of spatio-temporal layers, we designed three experiments with different W_d values, specifically, $W_d = 3$ (G1E1), $W_d = 2$ (G1E2), and $W_d = 1$ (G1E3). Here, the network will become a 2D encoder when W_d is equal to 1 because the input is a single frame.

Group 2: To check the influence of the progressive training, three experiments were implemented with different sizes of image pairs as the input. Firstly, we only input the image pairs of 256×192 with the batch size of 24 (G2E1); then, the stereo images of 384×256 were selected with the batch size of 12 in the second experiment (G2E2); finally, the stereo images of 512×384 with the batch size of 4 were input to the network for training (G2E3).

Group 3: To explore the effect of hierarchical prediction in our network, we set different weight combination in the loss function, i.e., $\lambda_1 = 0.5, \lambda_2 = 0.7, \lambda_3 = 1$ (G3E1); $\lambda_1 = 0.3, \lambda_2 = 0.7, \lambda_3 = 1$ (G3E2); $\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 1$ (G3E3). It can be seen that the hierarchical prediction was removed in the last experiment.

Group 4: To assess the influence of different components in our network, three experiments were also implemented later. First of all, we modified the dilated convolutional layers and residual blocks as the normal convolutional layers (G4E1); next, we added the batch normalization in the decoder (G4E2); finally, we removed the group-wise correlation in the cost volume (G4E3).

4. Results

The SCARED test dataset, consisting of 8 different surgical videos with 3016 image pairs, was used to perform a comprehensive evaluation of our proposed network. Six state-of-the-art methods, including one local-optimization-based method [9] and five learning-based models [10,18,19,21,22], were chosen to conduct the comparison study with our network. To perform a fair comparison study, we adopted the same three datasets (Scene Flow, Sintel, and our self-made synthetic dataset) to train the existing models from scratch except for [9] since it is a parametric method that does not need to be trained. We retained the original hyperparameters (such as learning rate) recommended by the authors for other learning-based methods. For those methods that

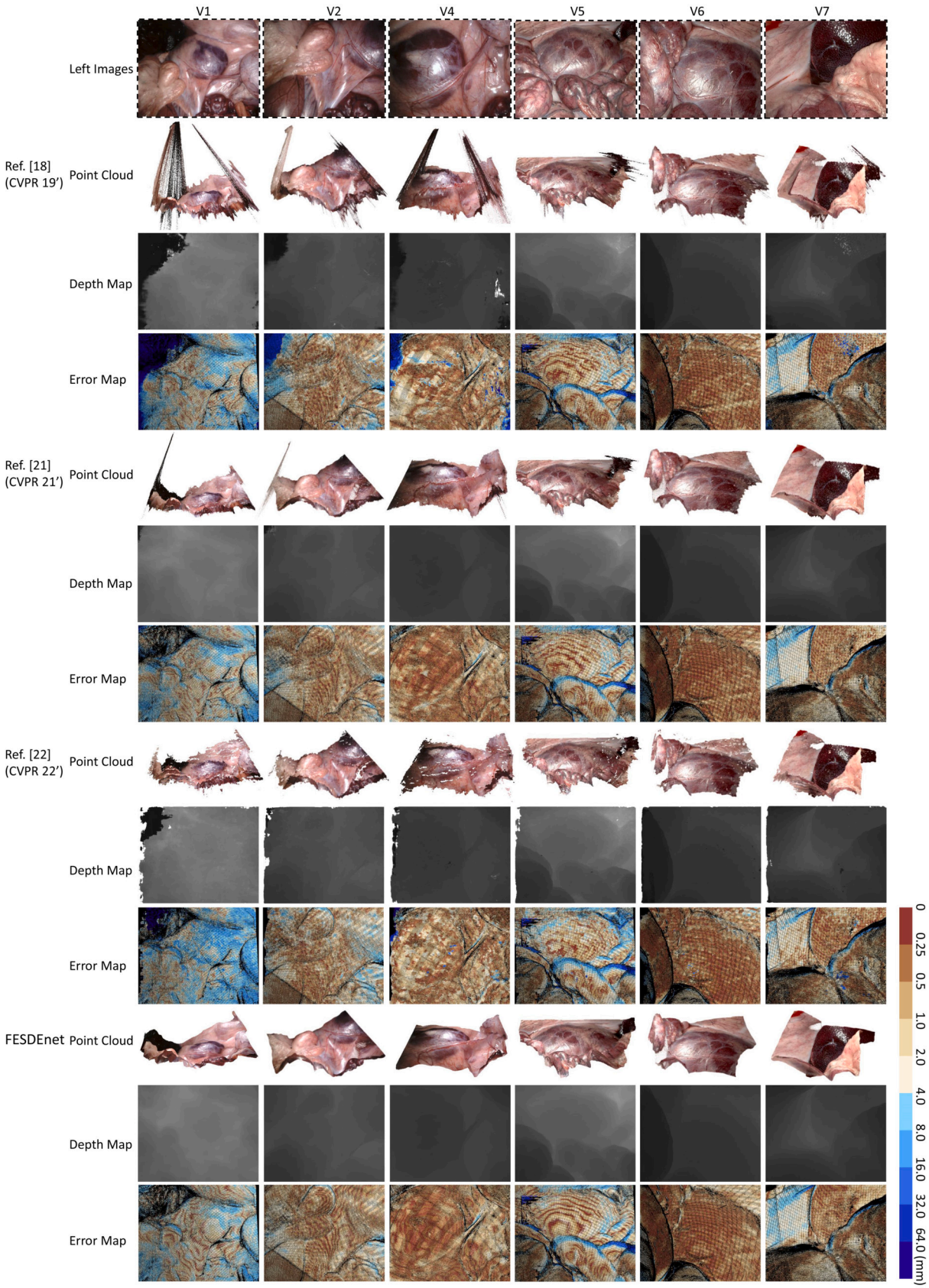

Fig. 6. Qualitative depth estimation results using 6 random samples from different surgical videos. The error maps were also given for a comprehensive understanding of the prediction error at each pixel level. For each model, the first row presents the reconstructed point cloud, the second row shows the estimated depth map, and the third row shows the error map.

Table 3
The results of the ablation study are based on the whole dataset, and they are divided into four groups. Both MAE and RMSE were given as well as the corresponding results of the Wilcoxon signed-rank test. The specific description of this ablation study can be read in Section 3.4.

Groups	Experiments	MAE (mm) / Signed-rank Test		RMSE (mm) / Signed-rank Test	
G1: Spatio-temporal layers	G1E1: Wd = 3	2.62±1.56	****	5.42±1.44	****
	G1E2: Wd = 2	2.58±1.57	***	5.41±1.43	****
	G1E3: Wd = 1	2.59±1.56	****	5.34±1.50	****
	Proposed: Wd = 4	2.55±1.51		5.23±1.40	
G2: Progressive training	G2E1: 256×192	2.67±1.50	****	5.50±1.57	****
	G2E2: 384×256	2.58±1.52	***	5.51±1.71	****
	G2E3: 512×384	2.57±1.54	ns	5.30±1.55	ns
	Proposed: Mixed	2.55±1.51		5.23±1.40	
G3: Hierarchical prediction	G3E1: (0.5, 0.7, 1)	2.64±1.63	****	5.52±1.51	****
	G3E2: (0.3, 0.7, 1)	2.62±1.53	****	5.38±1.44	****
	G3E3: (0, 0, 1)	2.65±1.57	****	5.52±1.46	****
	Proposed: (0.3, 0.5, 1)	2.55±1.51		5.23±1.40	
G4: Different components	G4E1: modify conv layers	2.59±1.57	***	5.30±1.41	****
	G4E2: add BN in decoder	2.54±1.52	**	5.27±1.45	****
	G4E3: modify cost volume	2.59±1.56	****	5.39±1.61	**
	Proposed	2.55±1.51		5.23±1.40	

provide multiple hyperparameter configurations for different training datasets, we chose the configuration of Scene Flow since it is also the main dataset in our case. It should be noticed that the optimization-based method [9] outputs sparse depth maps while other learning-based models predict dense depth maps. All assessments were based on an Ubuntu server with an NVIDIA A100 GPU (40 GB), and the quantitative results were shown in Table 2. When calculating all the image pairs, our FESDEnet got the best performance with 2.55±1.51 mm in MAE and 5.23±1.40 mm in RMSE. Our method also got the shortest inference time of 0.28 s in one frame with a resolution of 1280×1024, but please note that our network predicts four consecutive frames simultaneously and the inference time is 1.12 s in practice. Specifically, our network got the 7 best reconstruction performances related to MAE in these videos. Also, our network got the lowest error in all videos when calculating RMSE, which showed that our approach has stronger robustness to the outliers when reconstructing the surgical scene. Furthermore, the box plots of MAE and RMSE using all 8 videos were shown in Fig. 5. We conducted the Mann-Whitney U test to evaluate the significant difference between our network and the state-of-the-art methods. It can be seen that our methods are significantly different from other methods when calculating both MAE and RMSE. Finally, the qualitatively reconstructed scenes were given in Fig. 6. We selected three typical methods [18,21,22] from the comparison study to present their reconstruction results compared with ours. The motivation for choosing these three methods is that [18] performs the worst in accuracy, while [21] performs the best in the existing methods, and [22] is the latest model. In particular, we also visualized the error maps at the pixel level to demonstrate the error distribution. Our network could generate a smoother surface with fewer outliers compared with other approaches, which showed promising reconstruction performance in the medical domain.

Then, the ablation study was conducted to explore our proposed hypotheses and find out the best configuration, as shown in Table 3. We divided the ablation study into four groups and calculated both MAE and RMSE. Furthermore, we performed the Wilcoxon signed-rank test [44] to check the statistical difference. In group 1, it can be seen that the reconstruction accuracy is the best when the temporal clips consist of four consecutive frames. It should be noted that the performance growth is not proportional to the number of consecutive frames, since the performance when $W_d = 3$ is even worse than the single input. Furthermore, the progressive training gets the best performance compared with the input of fixed resolution, although there is no significant difference when inputting the maximum resolution of 512×384. Nevertheless, this strategy speeds up the training time. Next, we found that the hierarchical prediction with the weights of $\lambda_1 = 0.3, \lambda_2 = 0.5, \lambda_3 = 1$ performs best, while it would be the worst when removing the hierarchical prediction. Finally, we can see that our proposed network gets the best performance compared with those that modify some components in Group 4.

5. Discussion and conclusions

In this paper, we proposed a fully 3D encoder-decoder network with stacked cost volumes, and we integrated spatio-temporal layers, hierarchical prediction and progressive training into our network. The comparison study based on the quantitative and qualitative evaluation showed promising performance by comparing with the state-of-the-art methods using 8 surgical videos. More importantly, we explored the proposed three hypotheses using the ablation study. We found that for Hypothesis 1, spatio-temporal layers with the proper consecutive frames could enhance the performance of intra-operative stereo depth estimation. However, we cannot assume that the performance of the model will improve as the number of consecutive frames increases, because the temporal clip composed of 3 consecutive frames will even perform worse than the input of a single frame in our case. Furthermore, the progressive training (Hypothesis 2) strategy with the weak-to-strong data regularization could shorten the training time while not deteriorating the performance of the network. We can also observe that input with higher resolution could enhance accuracy although the training time will be slower. Finally, the hierarchical prediction (Hypothesis 3) with the proper weights performs better compared with the traditional terminal output, which shows the potential for future regression models with the input of video clips.

A possible limitation of the network is that it requires annotated datasets for training. We understand that the datasets with ground truth in the intra-operative stereo depth estimation community are always insufficient. It is also not easy to make the annotation manually by ourselves, since the ground truth belongs to the pixel level and relies on external devices to capture it. Thanks to the contributions of previous scholars in making annotated medical datasets such as Endoabs [45], SERV-CT [46], and the toolkit of Vision Blender [41], which could promote the development of supervised learning in the medical community. A possible improvement is to train our network without the requirement of annotated datasets, for instance, by introducing GAN to train the network in an adversarial way [8,47], which may release the burden of requiring annotated medical datasets.

Another limitation is that the temporal information may be difficult to be utilized well in clinical applications. The sliding window algorithm was adopted to encode the temporal information in this work, and the parameters are fixed, such as the frame number of video clips W_d, the number of frames between two sampled images W_f, and the

skipping stride along the temporal sequence W_s. However, the clinical environment is always complex, which means the fluctuation of the inter-frame difference is not constant. Many factors may influence the performance when encoding the temporal information, such as the movement of endoscope and the variation of light. For instance, the temporal information is different in two cases: when the endoscope moves rapidly and the surgical scene remains static. A possible solution is to consider robotic kinematics and enhance the perception of surgical semantics so that the temporal information can be encoded dynamically.

Our model predicts the disparity maps of four frames simultaneously, so the prediction time is relatively slower than other models based on single-frame prediction. To perform a real-time application using our model, downsampling the original images to a smaller resolution can significantly reduce the inference time, which needs more evaluations in the future.

To conclude, the proposed network suggests an effective and promising performance in intra-operative depth estimation based on the stereo endoscope. It provides the foundation to integrate AR and VF in robot-assisted surgery for safety. In the future, we will extend the proposed network to perform a multi-task estimation, such as the segmentation of point clouds [48,49], which is also an important topic today.

Declaration of competing interest

The authors declare no conflict of interest relevant to this work.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpb.2023.107937.

References

[1] A. Marzullo, S. Moccia, M. Catellani, F. Calimeri, E. De Momi, Towards realistic laparoscopic image generation using image-domain translation, Comput. Methods Programs Biomed. 200 (2021) 105834.

[2] S. Bano, F. Vasconcelos, M. Tella-Amo, G. Dwyer, C. Gruijthuijsen, E. Vander Poorten, T. Vercauteren, S. Ourselin, J. Deprest, D. Stoyanov, Deep learning-based fetoscopic mosaicking for field-of-view expansion, Int. J. Comput. Assisted Radiol. Surg. 15 (11) (2020) 1807–1816.

[3] G. Zampokas, G. Peleka, K. Tsiolis, A. Topalidou-Kyniazopoulou, I. Mariolis, D. Tzovaras, Real-time stereo reconstruction of intraoperative scene and registration to preoperative 3d models for augmenting surgeons' view during Ramis, Med. Phys. 49 (10) (2022) 6517–6526.

[4] Z.M. Baum, Y. Hu, D.C. Barratt, Real-time multimodal image registration with partial intraoperative point-set data, Med. Image Anal. 74 (2021) 102231.

[5] F. Chen, X. Cui, B. Han, J. Liu, X. Zhang, H. Liao, Augmented reality navigation for minimally invasive knee surgery using enhanced arthroscopy, Comput. Methods Programs Biomed. 201 (2021) 105952.

[6] M. Selvaggio, G.A. Fontanelli, F. Ficuciello, L. Villani, B. Siciliano, Passive virtual fixtures adaptation in minimally invasive robotic surgery, IEEE Robot. Autom. Lett. 3 (4) (2018) 3129–3136.

[7] Z. Chen, S. Terlizzi, T. Da Col, A. Marzullo, M. Catellani, G. Ferrigno, E. De Momi, Robot-assisted ex vivo neobladder reconstruction: preliminary results of surgical skill evaluation, Int. J. Comput. Assisted Radiol. Surg. (2022) 1–9.

[8] H. Luo, C. Wang, X. Duan, H. Liu, P. Wang, Q. Hu, F. Jia, Unsupervised learning of depth estimation from imperfect rectified stereo laparoscopic images, Comput. Biol. Med. 140 (2022) 105109.

[9] A. Geiger, M. Roser, R. Urtasun, Efficient large-scale stereo matching, in: Asian Conference on Computer Vision, Springer, 2010, pp. 25–38.

[10] J.-R. Chang, Y.-S. Chen, Pyramid stereo matching network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.

[11] V. Penza, J. Ortiz, L.S. Mattos, A. Forgione, E. De Momi, Dense soft tissue 3d reconstruction refined with super-pixel segmentation for robotic abdominal surgery, Int. J. Comput. Assisted Radiol. Surg. 11 (2) (2016) 197–206.

[12] S. Giannarou, M. Visentini-Scarzanella, G.-Z. Yang, Probabilistic tracking of affine-invariant anisotropic regions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 130–143.

[13] D. Stoyanov, M.V. Scarzanella, P. Pratt, G.-Z. Yang, Real-time stereo reconstruction in robotically assisted minimally invasive surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2010, pp. 275–282.

[14] G. Zampokas, K. Tsiolis, G. Peleka, I. Mariolis, S. Malasiotis, D. Tzovaras, Real-time 3d reconstruction in minimally invasive surgery with quasi-dense matching, in: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, 2018, pp. 1–6.

[15] S. Bernhardt, J. Abi-Nahed, R. Abugharbieh, Robust dense endoscopic stereo reconstruction for minimally invasive surgery, in: International MICCAI Workshop on Medical Computer Vision, Springer, 2012, pp. 254–262.

[16] C. Wang, F.A. Cheikh, M. Kaaniche, O.J. Elle, Liver surface reconstruction for image guided surgery, in: Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, vol. 10576, SPIE, 2018, pp. 576–583.

[17] G. Yang, J. Manela, M. Happold, D. Ramanan, Hierarchical deep stereo matching on high-resolution images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5515–5524.

[18] X. Guo, K. Yang, W. Yang, X. Wang, H. Li, Group-wise correlation stereo network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3273–3282.

[19] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, Z. Ge, Hierarchical neural architecture search for deep stereo matching, Adv. Neural Inf. Process. Syst. 33 (2020) 158–22 169.

[20] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, 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.

[21] Z. Shen, Y. Dai, Z. Rao, Cfnet: cascade and fused cost volume for robust stereo matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13 906–13 915.

[22] B. Liu, H. Yu, G. Qi, Graftnet: towards domain generalized stereo matching with a broad-spectrum and task-oriented feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 012–13 021.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[24] B. Huang, J.-Q. Zheng, A. Nguyen, D. Tuch, K. Vyas, S. Giannarou, D.S. Elson, Self-supervised generative adversarial network for depth estimation in laparoscopic images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 227–237.

[25] Z. Yang, R. Simon, Y. Li, C.A. Linte, Dense depth estimation from stereo endoscopy videos using unsupervised optical flow methods, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2021, pp. 337–349.

[26] E. Colleoni, S. Moccia, X. Du, E. De Momi, D. Stoyanov, Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers, IEEE Robot. Autom. Lett. 4 (3) (2019) 2714–2721.

[27] S. Moccia, L. Migliorelli, V. Carnielli, E. Frontoni, Preterm infants' pose estimation with spatio-temporal features, IEEE Trans. Biomed. Eng. 67 (8) (2019) 2370–2380.

[28] A. Casella, S. Moccia, D. Paladini, E. Frontoni, E. De Momi, L.S. Mattos, A shape-constraint adversarial framework with instance-normalized spatio-temporal features for inter-fetal membrane segmentation, Med. Image Anal. 70 (2021) 102008.

[29] M. Tan, Q. Le, Efficientnetv2: smaller models and faster training, in: International Conference on Machine Learning, PMLR, 2021, pp. 096–10 106.

[30] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 702–703.

[31] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[32] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Proceedings, Part II 19, Athens, Greece, October 17-21, 2016, Springer, 2016, pp. 424–432.

[33] P. Ghosal, T. Chowdhury, A. Kumar, A.K. Bhadra, J. Chakraborty, D. Nandi, Mhuri: a supervised segmentation approach to leverage salient brain tissues in magnetic resonance images, Comput. Methods Programs Biomed. 200 (2021) 105841.

[34] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, H. Su, Mvsnerf: fast generalizable radiance field reconstruction from multi-view stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 124–14 133.

[35] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[37] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, A. Bry, End-to-end learning of geometry and context for deep stereo regression, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 66–75.

[38] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[39] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. 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, 2016, pp. 4040–4048.

[40] D.J. Butler, J. Wulff, G.B. Stanley, M.J. Black, A naturalistic open source movie for optical flow evaluation, in: European Conference on Computer Vision, Springer, 2012, pp. 611–625.

[41] J. Cartucho, S. Tukra, Y. Li, D.S. Elson, S. Giannarou, Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery, Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 9 (4) (2021) 331–338.

[42] M. Allan, J. Mcleod, C. Wang, J.C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K.X. Fu, T. Zeffiro, W. Xia, et al., Stereo correspondence and reconstruction of endoscopic data challenge, arXiv preprint, arXiv:2101.01133, 2021.

[43] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, Adv. Neural Inf. Process. Syst. 27 (2014).

[44] S. Montaha, S. Azam, A.R.H. Rafid, M.Z. Hasan, A. Karim, K.M. Hasib, S.K. Patel, M. Jonkman, Z.I. Mannan, Mnet-10: a robust shallow convolutional neural network model performing ablation study on medical images assessing the effectiveness of applying optimal data augmentation technique, Front. Med. 9 (2022).

[45] V. Penza, A.S. Ciullo, S. Moccia, L.S. Mattos, E. De Momi, Endoabs dataset: endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms, Int. J. Med. Robot. Comput. Assist. Surg. 14 (5) (2018) e1926.

[46] P.E. Edwards, D. Psychogyios, S. Speidel, L. Maier-Hein, D. Stoyanov, Serv-ct: a disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction, Med. Image Anal. 76 (2022) 102302.

[47] G.P. Cannata, L. Migliorelli, A. Mancini, E. Frontoni, R. Pietrini, S. Moccia, Generating depth images of preterm infants in given poses using gans, Comput. Methods Programs Biomed. 225 (2022) 107057.

[48] S. Bano, F. Vasconcelos, L.M. Shepherd, E.V. Poorten, T. Vercauteren, S. Ourselin, A.L. David, J. Deprest, D. Stoyanov, Deep placental vessel segmentation for fetoscopic mosaicking, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 763–773.

[49] M. Kamari, Y. Ham, Vision-based volumetric measurements via deep learning-based point cloud segmentation for material management in jobsites, Autom. Constr. 121 (2021) 103430.