SFDDE: Self-supervised Frequency Domain Depth Estimation in Stereoscopic Surgical Videos

Anonymous

Anonymous

Abstract. Dense depth estimation plays a crucial role in developing context-aware computer-assisted intervention systems. Considering the challenges in acquiring per-pixel ground truth depth data, self-supervised depth estimation using image reconstruction as the supervisory signals has been proposed. However, it is still challenging for existing methods to fully solve the problem and there are great potentials to improve the performance. To this end, we propose a novel frequency domain depth estimation framework, referred as SFDDE, for accurate depth estimation with fine-grained details and temporal consistency from stereoscopic surgical videos. Unlike previous works, our method takes full advantage of the complementary information of visual and temporal features learned from the stereoscopic video sequences. SFDDE can learn high-level representations that encode both visual features and temporal dependencies in an end-to-end architecture for improving the depth estimation accuracy. We further introduce high-frequency-based image reconstruction supervisory losses, which can help to preserve fine-grained details of the surgical scene. Results from experiments conducted on two publicly available datasets demonstrate the superior performance of SFDDE over other state-of-the-art methods.

Keywords: Depth estimation · Vision transformers · Frequency domain · Stereoscopic video sequences.

1 Introduction

Depth estimation from stereo laparoscopic images plays a crucial role in 3D surgical scene reconstruction, surgical navigation, and augmented reality (AR) visualization. There exist previous attempts to develop depth estimation methods based on bionocular stereo matching using image intensity or color information [1–3]. However, depth estimation from stereo matching is a challenging task due to multiple factors, such as tissue deformation, specular reflections, tool occlusion, and lack of photometric constancy across frames [4–6]. To meet the challenges, recent works have adopted deep learning-based methods, particularly convolutional neural networks (CNN).

Both supervised and self-supervised CNN methods have been introduced before. For example, Mayer *et al.* [7] introduced a supervised DispNet that can directly computes the correspondence field between two images. Their model was

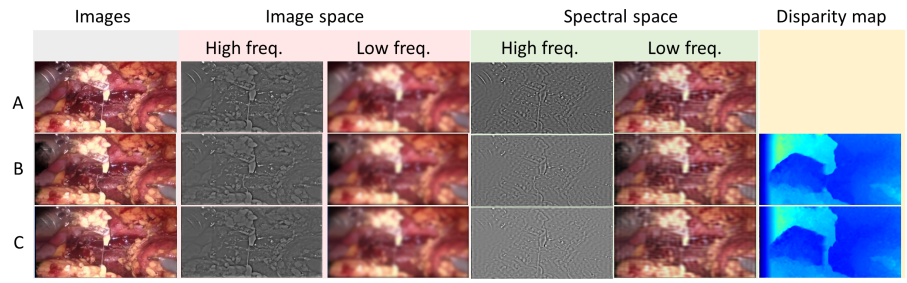

Fig. 1. Visualization of low and high-frequency image content for Row A: ground truth left image, B: reconstructed left image without high-frequency losses, and Row C: reconstructed left image using proposed frequency losses. For each image, we plotted the high-frequency and low-frequency image contents in both image space (red panel) and spectral space (green panel). Disparity maps are also shown for two reconstructed images (yellow panel), respectively.

trained by minimizing a regression training loss, thus requiring a large amount of accurate ground truth disparity data, which is challenging to obtain, especially for laparoscopic vision. This has motivated the development of self-supervised methods where image reconstruction is used as the supervisory signal and no ground truth disparity data is required for training [4, 8–10]. Taking a set of monocolar or stereo images as input, Godard *et al.* [8] proposed to generate disparity images by training their network with an image reconstruction loss, eliminating the requirement of ground truth depths. Along the same line, Ye *et al.* [9] proposed a deep learning framework consisting of an autoencoder for depth prediction, and a differentiable spatial transformer [11] for training the autoencoder on stereo image pairs without ground truth depths. In addition to the image reconstruction supervisory signal, Huang *et al.* [4] further introduced a supervisory signal based on generative adversarial learning [12]. Tukra and Giannarou [10] investigated randomly connected neural networks for self-supervised monocular depth estimation.

Despite significant progress, however, it is still challenging for existing methods to fully solve the problem and there are great potentials to improve the performance for the following reasons. First, most of previous methods treat the dense depth estimation as a static stereo matching problem where stereo image pairs are taken as the input without considering temporal information, despite the fact that stereoscopic video sequences are available for training. Second, the introduction of image reconstruction supervisory signal helps to eliminate the requirement of ground truth depths but it also causes the problem of over-smoothing at image boundaries, leading to poor reconstruction of small structures and boundary edges (see Fig. 1-B for an example). Third, it has also been found in [13] that the lack of global context caused by the locality of estimation network in most of previous self-supervised depth estimation methods also leads to over-smoothing disparity.

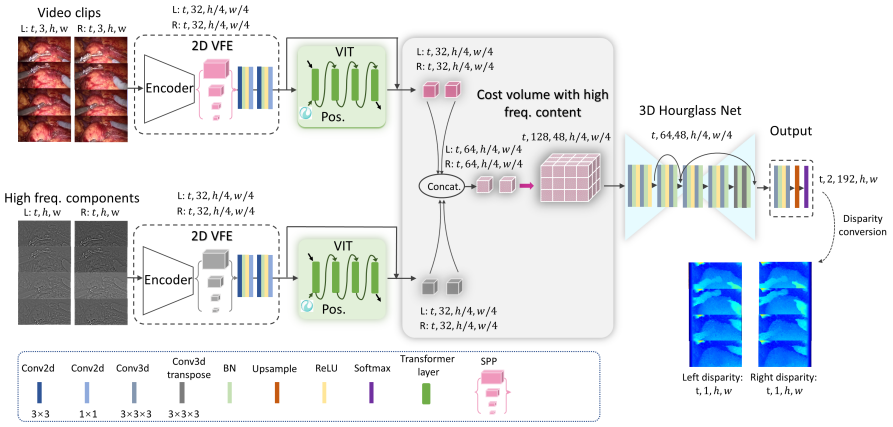

Fig. 2. The network architecture of SFDDE. See main text for a detailed explanation.

In this paper, to tackle the challenges, we propose a self-supervised frequency domain depth estimation method, referred as SFDDE, for fine-grained prediction of disparity maps in stereoscopic surgical videos. Different from previous methods which take static stereo images as the input, our method expects paired video clips and decomposed high-frequency components as the input to extract fine-grained visual features from each frame in the video clips. The extracted features are then patched into a sequence of spatio-temporal tokens and sent into vision transformers [14] for learning global spatio-temporal context via self-attention. The learned spatio-temporal features are then used to predict disparity maps for all images in the video clips. We further introduce high-frequency-based image reconstruction supervisory losses for self-supervised details-preserving depth estimation (see Fig.1-C for an example).

Our contribution can be summarized as follows: (1) We present a novel method, i.e., SFDDE, to accurately estimate depths from stereoscopic video sequences. Unlike previous works [4, 9, 10], our method takes full advantage of the complementary information of visual and temporal features learned from the stereoscopic video sequences. SFDDE can learn high-level representations that encode both visual features and temporal dependencies in an end-to-end architecture for improve the depth estimation accuracy; (2) We introduce high-frequency-based image reconstruction supervisory losses, which can help to preserve fine-grained details of the surgical scene; (3) We conduct comprehensive experiments on two publicly available datasets to demonstrate the efficacy of the present method.

2 Method

Given a rectified stereoscopic video sequences $\{I_i^l, I_i^r | 0 < i \leqslant T\}$ where T is the sequence length, we aim to determine two dense disparity maps $\{D_i^l; D_i^r\}$

for each each frame i, where point (x_i^l, y_i^l) on the left image I^l can find its correspondence on the right image I^r through wrapping $(x_i^l + d_i^r, y_i^l)$ and $d_i^r \in D^r$, and vice versa. Once the disparity maps are estimated, we can explicitly infer the depth map using camera parameters. The training objective of self-supervised depth estimation is usually designed to enforce consistency between original $\{I_i^l, I_i^r\}$ and reconstructed images $\{\tilde{I}_i^l, \tilde{I}_i^r\}$, where $\tilde{I}_i^l = wrap(I_i^r, D_i^l)$ and $\tilde{I}_i^r = wrap(I_i^l, D_i^r)$ [8]. Below we first present our network architecture, followed by the training objective.

2.1 Network architecture

The network architecture of the SFDDE is presented in Fig. 2. The network expects input of paired video clips as well as decomposed high-frequency components (how to obtained them will be presented below). The inputs are passed through two parallel branches, where each branch is composed of a 2D visual feature extractor (VFE), follow by a vision transformer (ViT) module to extract multi-scale spatial-temporal features. Next, features produced by the two branches are concatenated together to build an enhanced cost volume with high-frequency image content. To simply the description, below we illustrate each component using the left video clips.

Visual feature extractor: Let us define $C^l = \left\{I_i^l \in \mathbb{R}^{h \times w \times 3} | 0 < i \leq t\right\}$ be the left video clip with t frames, and $H^l = \left\{h_i^l \in \mathbb{R}^{h \times w} | 0 < i \leq t\right\}$ is the image space high-frequency component. We design a 2D U-Net-like encoder as the visual feature extractor, which generates features at $1/2, 1/4, 1/8$ and $1/16$ down-sampling levels, as shown in Fig. 2. Features of different sizes are sent into a spatial pyramid pooling (SPP) module [15], followed by 2D convolutions. Finally, each branch generates a sequence of feature maps $\left\{f_i \in \mathbb{R}^{32 \times \frac{h}{4} \times \frac{w}{4}} | 0 < i \leq t\right\}$, where t is the frame length and 32 is the number of channels.

Vision transformer: Feature f_i^l of the i^{th} image frame is splitted into N vectorized 3×3 patches. Then, we collapse spatial dimension of the patches and represent f_i^l as a sequences of patch vectors $\left\{p^j \in \mathbb{R}^{C \times 1} | 0 < j \leq N\right\}$, where $N = \frac{h}{4} \cdot \frac{w}{4} \cdot \frac{1}{3^2}$ is the patch number and $C = 32 \cdot 3^2$ is the channel dimension. We can represent the clip as a $(t \times N)$-length sequence of 288-dimensional features, which servers as the input to transformers. We adopt a 4-layer ViT module with a hidden dimension of 288, where sinusoidal position embeddings are added to the first transformer layer. Thereafter, each transformer layer calculates a $(t \times N) \times (t \times N)$ spatial-temporal attention map. Finally, the outputs of ViT are reshaped back to have the same dimension as f_i^l. Skip connections are added from the input feature f_i^l to the reshaped feature for cost volume construction.

Cost volume: To build the cost volume, as shown in Fig. 2, the top branch spatio-temporal features are concatenated with the bottom branch high-frequency features first. We then follow [16] to construct the cost volume by concatenating the left and the right features on locations corresponding to different disparity levels. A 3D hourglass network is then used to regress a continuous map $\{D_i^l | 0 < i \leqslant t\}$ from the cost volume. Similarly, we can also compute a con-

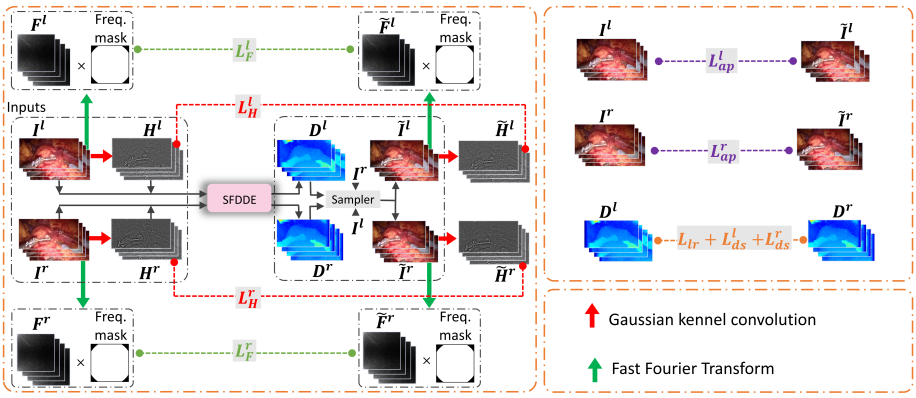

Fig. 3. Overview of the proposed SFDDE framework.

tinuous map $\{D_i^r | 0 < i \leqslant t\}$ from the right video clips and the decomposed high-frequency components.

2.2 Training objectives

Although the input to our network are paired video clips and high-frequency components, we compute the loss for each pair of image (left and right image pair) in the video clips separately. The overall loss to train our model is then computed by adding the losses computed on all image pairs in the video clips. Below to simplify description, we only describe the losses for one pair of images and drop out the time index.

An overview of the training objective of the proposed method is shown in Fig. 3. The key idea is to preserve fine-grained details by applying restrictions directly on decomposed high-frequency contents of the inputs $\{I^l, I^r\}$ and the reconstructed images $\{\tilde{I}^l, \tilde{I}^r\}$. Losses are calculated for both left and right images. To simplify description, below we only present the losses computed for the left image.

Image space high-frequency matching. We extract high-frequency contents in image space by removing the low-frequency content, which is obtained using 2D Gaussian kernel convolution. Let us denote a left image $I^l \in \mathbb{R}^{h \times w \times 3}$, where h ,w, 3 are height, width and channel, respectively. To remove the influence of illumination, we convert I^l to grayscale image $I_g^l \in \mathbb{R}^{h \times w}$. Next, we compute the high-frequency component H^l in image space as:

$$H^l(x,y) = I_g^l(x,y) - \sum_{i,j \in G_k} \left(\frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}} \right) \cdot I_g^l(x+i, y+j) \tag{1}$$

where (x, y) is position on image I_g^l, (i, j) is index on Gaussian kernel G_k. k is the kernel size which is linearly correlated with variance σ [17].

Image space high-frequency loss. The high-frequency loss L_H^l in image space is estimated by minimizing the L1 difference between high-frequency content of the input and the reconstructed images:

$$L_H^l = \frac{1}{hw}\left(\sum_{(x,y)\in I_g^l}\left\|H^l(x,y) - \tilde{H}^l(x,y)\right\|_1\right) \tag{2}$$

Spectral space high-frequency matching. We transform input images from spatial space to spectral space using Fast Fourier Transformation (FFT):

$$fft^l(k,m) = \frac{1}{hw}\sum_{x=0}^{h-1}\sum_{y=0}^{w-1} I_g^l(x,y)\cdot e^{-2\pi j(\frac{xk}{h}+\frac{ym}{w})} \tag{3}$$

where we have $k = 0, \cdots, h-1$, $m = 0, \cdots, w-1$. Please note that we do not shift zero-frequency component to center of spectrum, thus the center region of $fft(\cdot)$ represents high-frequency content while the corner region corresponds to that of low-frequency. We then extract high-frequency content F^l in spectral domain by multiplying the log magnitude of the Fourier spectrum with mask M.

$$F^l(k,m) = \log\left(1 + \sqrt{\mathrm{Re}(fft^l(k,m))^2} + \sqrt{\mathrm{Im}(fft^l(k,m))^2} + 10^{-8}\right) * M \tag{4}$$

where:

$$M(k,m) = \begin{cases} 1, & if\ D(k,m) \le r \\ 0, & if\ D(k,m) > r \end{cases} \tag{5}$$

$Re(\cdot)$ and $Im(\cdot)$ denotes the real and imaginary part of the spectrum, respectively. r is the radius to separate out high-frequency region. $D(k,m)$ is the L2 distance from (k,m) to the mask center.

High-frequency loss in spectral space. The spectral space loss is enforced by directly minimizing the L1 difference of the high-frequency spectral map for the input and the reconstructed images.

$$L_F^l = \frac{1}{hw}\left(\sum_{(k,m)\in F^l}\left\|F^l(k,m) - \tilde{F}^l(k,m)\right\|_1\right) \tag{6}$$

Image loss functions. In addition to high-frequency losses, we employ the appearance matching loss, disparity smoothness loss and left-right disparity consistency loss as proposed in [8]. The appearance matching loss L_{ap} is defined as:

$$L_{ap}^l = \tfrac{1}{hw}\sum_{(x,y)\in I^l}(1-\alpha)\left\|I^l(x,y) - \tilde{I}^l(x,y)\right\|_1 + \alpha\frac{1-SSIM\left(I^l(x,y),\tilde{I}^l(x,y)\right)}{2} \tag{7}$$

where $\alpha = 0.85$ and $SSIM(\cdot)$ is a 3×3 block filter [8].

The disparity smoothness loss is defined as:

$$L_{ds}^{l} = \frac{1}{hw} \sum_{(x,y)\in I^{l}} \left|\partial_x d^{l}(x,y)\right| e^{-\left\|\partial_x I^{l}(x,y)\right\|} + \left|\partial_y d^{l}(x,y)\right| e^{-\left\|\partial_y I^{l}(x,y)\right\|} \tag{8}$$

The left-right disparity consistency loss is given by:

$$L_{lr} = \frac{1}{hw} \sum_{x,y} |d^{l}(x,y) - d^{r}((x + d^{l}(x,y)), y)| \tag{9}$$

Losses for an image pair. Finally, we combine all loses from both left and right images as:

$$L = \lambda_H(L_H^l + L_H^r) + \lambda_F(L_F^l + L_F^r) + \lambda_{ap}(L_{ap}^l + L_{ap}^r) + \lambda_{ds}(L_{ds}^l + L_{ds}^r) + \lambda_{lr}L_{lr} \tag{10}$$

where λ_H, λ_F, λ_{ap}, λ_{ds} and λ_{lr} are parameters controlling the relative weights of different loss terms.

2.3 Implementation details.

Input video clips are rescaled to a resolution of 192×384 with a frame-length of 4. We empirically adopt Gaussian kernel size $k = 21$, spectral mask radius $r = 115$, $batchsize = 4$, $\lambda_H = 0.5$, $\lambda_F = 0.5$, $\lambda_{ap} = 0.85$, $\lambda_{ds} = 0.1$ and $\lambda_{lr} = 1.0$. The network is implemented in Pytorch framework and trained on a workstation with two GeForce RTX 3090 graphics cards. We trained the network with the AdamW optimizer [18] with a weight decay of 0.05 and learning rate of 1e-4 for approximately 21 epochs. A cosine learning rate scheduler [19] is adopted with 1 epoch of linear warm-up and initial learning rate of 5e-7.

3 Experiments and Results

To demonstrate the performance of the proposed method, we designed and conducted comprehensive experiments on two publicly available datasets. The first was the dVPN dataset, collected from da Vinci partial nephrectomy [9] and the second was the stereo correspondence and reconstruction of endoscopic data (SCARED) [20].

Results on the dVPN dataset. The dVPN dataset contains 34320 pairs of rectified training images and 14383 pairs of testing images. We adopted the same evaluation metrics as proposed in [9], which was the mean and standard deviation of Structural Similarity Index (SSI) between pairs of input and reconstructed image of both left images and right images. The experimental results when compared with SOTA unsupervised depth estimation methods are shown in Table 1-left. SFDDE achieved the best performance with a mean SSI of 86.4 $\pm$ 3.5%.

Table 1. Results of validation study on both datasets.

Dataset	dVPN		SCARED	
	Test set		Test set 1	Test set 2
Methods	mean SSI(%)	std. SSI(%)	mADE (mm)	mADE(mm)
Ye et al. [9]	60.4	6.6	-	-
Godard et al. [8]	54.9	8.7	23.56	21.62
Godard et al. [21]	71.2	7.5	21.92	15.25
Allan et al. [20]	-	-	20.94	17.22
Huang et al. [4]	79.6	4.9	17.42	11.23
SFDDE (Ours)	**86.4**	3.5	**9.61**	**6.39**

Table 2. Ablation study on different components of SFDDE.

Components			Performance	
VFE	ViT	Frequency losses	mean SSI(%)	std. SSI(%)
✓			83.9	4.0
✓	✓		84.4	3.9
✓	✓	✓	**86.4**	3.5

Results on the SCARED dataset. The SCARED dataset is from MICCAI2019 Endovis challenge, which consists of 7 training datasets and 2 test datasets of rectified stereo videos and ground truth depth labels. Following evaluation metric in [20], we used mean absolute depth error (mADE) to compare with other unsupervised methods. As shown in Table 1-right, SFDDE demonstrated lowest mADE of $9.61mm$ on test set 1 and lowest mADE of $6.39mm$ on test set 2.

Ablation study. We conducted an ablation study on dVPN dataset with the same study protocol as presented above. As shown in Table 2, without using ViT module and the proposed high-frequency losses, the model obtained a mean SSI of 83.9 ± 4.0%. By adding the ViT module, the model generated a better mean SSI of 84.4 ± 3.9%. By adding both ViT module and the high-frequency losses, our method achieved a further improved mean SSI of 86.4 ± 3.5%.

4 Conclusions

In this paper, we proposed a novel frequency domain depth estimation framework for accurate depth estimation with fine-grained details and temporal consistency from stereoscopic surgical videos. Unlike previous works, our method took full advantage of the complementary information of visual and temporal features learned from the stereoscopic video sequences. SFDDE can learn high-level representations that encode both visual features and temporal dependencies in an end-to-end architecture to improve the depth estimation accuracy. Results from experiments conducted on two publicly available datasets demonstrated the superior performance of SFDDE over other state-of-the-art methods.

References

1. Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2010) 275–282
2. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al.: Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Medical image analysis **17**(8) (2013) 974–996
3. Xia, W., Chen, E.C., Pautler, S., Peters, T.M.: A robust edge-preserving stereo matching method for laparoscopic images. IEEE Transactions on Medical Imaging (2022)
4. Huang, B., Zheng, J.Q., Nguyen, A., Tuch, D., Vyas, K., Giannarou, S., Elson, D.S.: Self-supervised generative adversarial network for depth estimation in laparoscopic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2021) 227–237
5. Yang, Z., Simon, R., Li, Y., Linte, C.A.: Dense depth estimation from stereo endoscopy videos using unsupervised optical flow methods. In: Annual Conference on Medical Image Understanding and Analysis, Springer (2021) 337–349
6. Long, Y., Li, Z., Yee, C.H., Ng, C.F., Taylor, R.H., Unberath, M., Dou, Q.: E-dssr: efficient dynamic surgical scene reconstruction with transformer-based stereoscopic depth perception. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2021) 415–425
7. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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) 4040–4048
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 270–279
9. Ye, M., Johns, E., Handa, A., Zhang, L., Pratt, P., Yang, G.Z.: Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. arXiv preprint arXiv:1705.08260 (2017)
10. Tukra, S., Giannarou, S.: Randomly connected neural networks for self-supervised monocular depth estimation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization (2021) 1–10
11. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Chen, C., Chen, X., Cheng, H.: On the over-smoothing problem of cnn based disparity estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 8997–9005
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

15. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9) (2015) 1904–1916
16. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 5410–5418
17. Heideman, M.T., Johnson, D.H., Burrus, C.S.: Gauss and the history of the fast fourier transform. Archive for history of exact sciences (1985) 265–277
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
19. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
20. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
21. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 3828–3838