

# Deep Laparoscopic Stereo Matching with Transformers

\*Xuelian Cheng<sup>1</sup>, \*Yiran Zhong<sup>2,3</sup>, Mehrtash Harandi<sup>1,4</sup>, Tom Drummond<sup>5</sup>,  
Zhiyong Wang<sup>6</sup>, and Zongyuan Ge<sup>1,7,8</sup>(✉)

<sup>1</sup> Faculty of Engineering, Monash University, Melbourne, Australia

<sup>2</sup> SenseTime Research, Shanghai, China

<sup>3</sup> Shanghai AI Laboratory, Shanghai, China

<sup>4</sup> Data61, CSIRO, Australia

<sup>5</sup> University of Melbourne, Melbourne, Australia

<sup>6</sup> The University of Sydney, Sydney, Australia

<sup>7</sup> eResearch Centre, Monash University, Melbourne, Australia

<sup>8</sup> Monash-AirDoc Research Centre, Melbourne, Australia

zongyuan.ge@monash.edu, <https://mmai.group>

**Abstract.** The self-attention mechanism, successfully employed with the transformer structure is shown promise in many computer vision tasks including image recognition, and object detection. Despite the surge, the use of the transformer for the problem of stereo matching remains relatively unexplored. In this paper, we comprehensively investigate the use of the transformer for the problem of stereo matching, especially for laparoscopic videos, and propose a new hybrid deep stereo matching framework (HybridStereoNet) that combines the best of the CNN and the transformer in a unified design. To be specific, we investigate several ways to introduce transformers to volumetric stereo matching pipelines by analyzing the loss landscape of the designs and in-domain/cross-domain accuracy. Our analysis suggests that employing transformers for feature representation learning, while using CNNs for cost aggregation will lead to faster convergence, higher accuracy and better generalization than other options. Our extensive experiments on SceneFlow, SCARED2019 and dVPN datasets demonstrate the superior performance of our HybridStereoNet.

**Keywords:** Stereo Matching · Transformer · Laparoscopic video

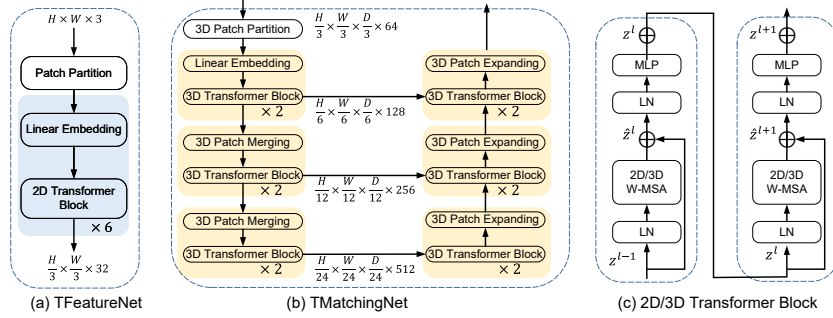
## 1 Introduction

3D information and stereo vision are important for robotic-assisted minimally invasive surgeries (MIS) [2, 17]. Given the success of modern deep learning systems [4, 5, 34, 36, 37] on natural stereo pairs, a promising next challenge is surgical stereo vision, *e.g.*, laparoscopic and endoscopic images. It has received

---

\* Indicates equal contribution





**Fig. 2.** The architecture of our transformer-based TFeatureNet and TMatchNet.

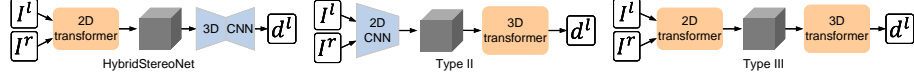
designed transformer-based structure and compare the accuracy, generalization ability, and the loss landscapes to analyze the behavior of the transformer for the stereo matching task. The insights obtained from those analyses have led us to propose a new hybrid architecture for laparoscopic stereo videos that achieves better performance than both convolution-based methods and pure transformer-based methods. Specifically, we find that the transformers tend to find flatter local minima while the CNNs can find a lower one but sharper. By using transformers as the feature extractor, and CNNs for cost aggregation, the network can find a local minima both lower and flatter than pure CNN-based methods, which leads to better generalization ability. Thanks to the global field of view of transformers, our HybridStereoNet has better textureless handling capability as well.

## 2 Method

In this section, we first illustrate our hybrid stereo matching architecture and then provide a detailed analysis of transformers in stereo matching networks. Our design is inspired by LEAStereo which is an state-of-the-art model for the natural stereo matching task. For the sake of discussion, we denote the feature net in LEAStereo [5] as CFeatureNet, and the matching net as CMatchNet.

### 2.1 The HybridStereoNet

Our proposed HybridStereoNet architecture is shown in Fig 1. We adapt the standard volumetric stereo matching pipeline that consists of a feature net to extract features from input stereo images, a 4D feature volume that is constructed by concatenating features from stereo image pairs through epipolar lines [32], a matching net that regularizes the 4D volume to generate a 3D cost volume, and a projection layer to project the 3D cost volume to a 2D disparity map. In this



**Fig. 3.** The overall pipeline of variant networks. We follow the feature extraction – 4D feature volume construction – dense matching pipeline for deep stereo matching. The variants change the FeatureNet and Matching Net with either transformer or CNNs. The gray box represents 4D cost volume.

pipeline, only the feature net and the matching net contain trainable parameters. We replace both networks with transformer-based structures to make them suitable for laparoscopic stereo images.

We show our transformer-based feature net (TFeatureNet) in Fig. 2 (a) and matching net (TMatchNet) in Fig. 2 (b). For a fair comparison with the convolutional structure of the LEAStereo [5], we use the same number of layers  $L$  for our TFeatureNet and TMatchNet as in the LEAStereo, *i.e.*,  $L^F = 6$  for the TFeatureNet and  $L^M = 12$  for the TMatchNet. The 4D feature volume is also built in  $1/3$  resolution.

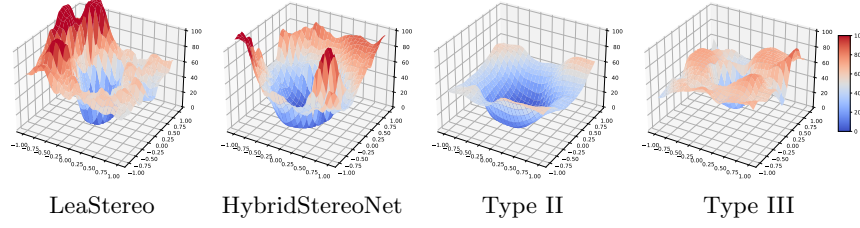
**TFeatureNet** is a Siamese network with shared weights to extract features from input stereo pairs of size  $H \times W$ . We adapt the same patching technique as in ViT [7] and split the image into  $N$  non-overlapping patches before feeding them to a vision transformer. We set the patch size to  $3 \times 3$  to obtain  $\frac{H}{3} \times \frac{W}{3}$  tokens and thus ensure the cost volume built in  $1/3$  resolution. A linear embedding layer is applied on the raw-valued features and project them to a  $C$  dimensional space. We set the embedding feature channel to 32. We empirically select the swin transformer block [15] as our 2D transformer block.

**TMatchNet** is a U-shaped encoder-decoder network with 3D transformers. The encoder is a three-stage down-sampling architecture with stride of 2. In contrast, the decoder is a three-stage up-sampling architecture as shown in Fig 2 (b). Similar to the TFeatureNet, we use a 3D Patch Partitioning layer to split the 4D volume  $H' \times W' \times D' \times C$  into  $N$  non-overlapping 3D patches and feed them into a 3D transformer. Unlike previous 3D transformers [16] that keep the dimension  $D$  unchanged when down-sampling the spatial dimensions  $H, W$ , we change  $D$  accordingly with the stride to enforce the correct geometrical constraints.

We compare the functionality of transformers in feature extraction and cost aggregation. As we will show soon, we find that transformers are good for feature representation learning while convolutions are good for cost aggregation. Therefore, in our HybridStereoNet, we use our TFeatureNet as the feature extractor and CMatchNet for the cost aggregation. We provide a detailed comparison and analysis in the following section.

## 2.2 Analyzing transformer in laparoscopic stereo

To integrate the transformer in the volumetric stereo pipeline, we have three options as shown in Fig 3. Type I (HybridStereoNet): TFeatureNet with CMatchNet; Type II: CFeatureNet with TMatchNet; and Type III: TFeatureNet with

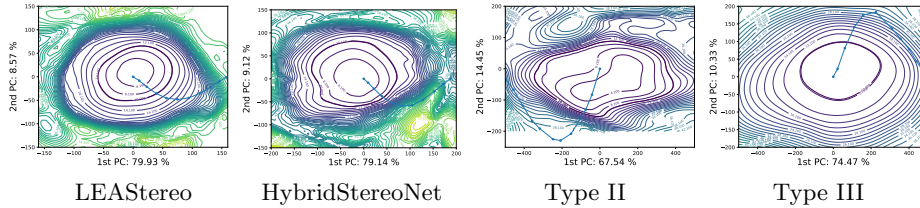


**Fig. 4.** Loss Landscape Visualization on SCARED2019 dataset. 3D surfaces of the gradient variance from HybridStereoNet and its variants on SceneFlow. The two axes mean two random directions with filter-wise normalization. The height of the surface indicates the value of the gradient variance.

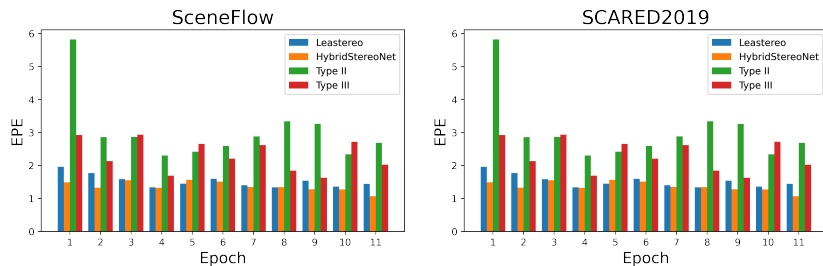
TMatchNet. In this section, we investigate these options in terms of loss landscapes, projected learning trajectories, in-domain/cross-domain accuracy, and texture-less handling capabilities in laparoscopic stereo.

**Loss landscape.** We visualize the loss landscape on the SCARED2019 dataset for types described above along with LEAStereo using the random direction approach [12] in Fig 4. The plotting details can be found in the supplemental material. The visualization suggests that Type II and Type III tend to find a flatter loss landscape but with a higher local minima. Compared with LEAStereo, the landscape of the HybridStereoNet shares a similar local minima but is flatter, which leads to better cross-domain performance [3, 11].

**Projected learning trajectory.** We also compare the convergence curve of each type with the projected learning trajectory [12]. Fig. 5 shows the learning trajectories along the contours of loss surfaces. Let  $\theta_i$  denote model parameters at epoch  $i$ . The final parameters of the model after  $n$  epochs of training are shown by  $\theta_n$ . Given  $n$  training epochs, we can apply PCA [24] to the matrix  $M = [\theta_0 - \theta_n; \dots; \theta_{n-1} - \theta_n]$ , and then select the two most explanatory directions. This enables us to visualize the optimizer trajectory (blue dots in Fig. 5) and loss surfaces along PCA directions. On each axis, we measure the amount of variation in the descent path captured by that PCA direction. Note that the loss landscape dynamically changes during training and we only present the “local” landscape around the final solution.



**Fig. 5.** Projected learning trajectories use normalized PCA directions for our variants. Please zoom in for better visualization.



**Fig. 6. Accuracy comparison for our proposed variants.** (a) In-domain results on SceneFlow dataset. (b) Cross domain results on SCARED2019 dataset.

**Table 1.** Memory footprint for the proposed variant networks.

Method	LeaStereo	Type II	Type III	HybridStereoNet
Params [M]	1.81	9.54	9.62	1.89
Flops [M]	3713.53	4144.85	811.68	380.01
Runtime [s]	0.30	0.48	0.50	0.32

As we can see from the Fig. 5, the Type II variant misses the local minima; Type III directly heads to a local minima after circling around the loss landscape in its first few epochs. It prohibits the network from finding a lower local minima. The LEAStereo and our HybridStereoNet both find some lower local minimas but the HybridStereoNet uses a sharper descending pathway on the loss landscape and therefore leads to a lower local minima. We use the SGD optimizer and the same training setting with LEAStereo [5] for side-by-side comparison.

**Accuracy.** In Fig 6, we compare the in-domain and cross-domain accuracy of all types for the first several epochs. All of these variants are trained on a large nature scene synthetic stereo dataset called SceneFlow [18] with over 30k stereo pairs. For in-domain performance, we plot the validation end point error (EPE) on the SceneFlow dataset in Fig 6 (a). The HybridStereoNet consistently achieves low error rates than all the other variants. For the cross-domain performance, we directly test the trained models on the SCARED2019 laparoscopic stereo dataset and plot the EPE in Fig 6 (b). Again, the HybridStereoNet achieves lower cross-domain error rates in most epochs.

**Memory footprint.** We provide details of four variants regarding the running time, and memory footprint in Table 1, which were tested on a Quadro GV100 with the input size  $504 \times 840$ . We keep relatively similar number of learnable parameters to make a fair comparison.

### 3 Experiments

#### 3.1 Datasets

We evaluate our HybridStereoNet on two public laparoscopic stereo datasets: the SCARED2019 dataset [1] and the dVPN dataset [31].

**Table 2.** The mean absolute depth error for the SCARED2019 *Test-Original* set (unit: mm). Each test set containing 5 keyframes, denoted as  $kf_n, n \in [1, 5]$ . Note that our method and STTR are not fine-tuned on the target dataset. The lower the better.

Method	Test Set 1						Test Set 2					
	$kf_1$	$kf_2$	$kf_3$	$kf_4$	$kf_5$	Avg.	$kf_1$	$kf_2$	$kf_3$	$kf_4$	$kf_5$	Avg.
Lalith Sharan [1]	30.63	46.51	45.79	38.99	53.23	43.03	35.46	50.09	25.24	62.37	70.45	48.72
Xiaohong Li [1]	34.42	20.66	17.84	27.92	13.00	22.77	24.58	16.80	29.92	11.37	19.93	20.52
Huoling Luo [1]	29.68	16.36	13.71	22.42	15.43	19.52	20.83	11.27	35.74	8.26	14.97	18.21
Zhu Zhanshi [1]	14.64	7.77	7.03	7.36	11.22	9.60	14.41	12.55	16.30	27.87	34.86	21.20
Wenyao Xia [1]	<b>5.70</b>	7.18	6.98	8.66	5.13	6.73	13.80	6.85	13.10	5.70	7.73	9.44
Trevor Zeffiro [1]	7.91	2.97	<b>1.71</b>	2.52	2.91	3.60	5.39	1.67	4.34	3.18	2.79	3.47
Congcong Wang [1]	6.30	2.15	3.41	3.86	4.80	4.10	6.57	2.56	6.72	4.34	1.19	4.28
J.C. Rosenthal [1]	8.25	3.36	2.21	2.03	1.33	3.44	8.26	2.29	7.04	2.22	0.42	4.05
D.P. 1 [1]	7.73	2.07	1.94	2.63	<b>0.62</b>	3.00	4.85	<b>0.65</b>	<b>1.62</b>	<b>0.77</b>	0.41	<b>1.67</b>
D.P. 2 [1]	7.41	<b>2.03</b>	1.92	2.75	0.65	2.95	4.78	1.19	3.34	1.82	<b>0.36</b>	2.30
STTR [13]	9.24	4.42	2.67	<b>2.03</b>	2.36	4.14	7.42	7.40	3.95	7.83	2.93	5.91
HybridStereoNet	7.96	2.31	2.23	3.03	1.01	3.31	<b>4.57</b>	1.39	3.06	2.21	0.52	2.35

**SCARED2019** is released during the Endovis challenge at MICCAI 2019, including 7 training subsets and 2 test subsets captured by a da Vinci Xi surgical robot. The original dataset only provides the raw video data, the depth data of each key frame and corresponding camera intrinsic parameters. We perform additional dataset curation to make it suitable for stereo matching. After curation, the SCARED2019 contains 17206 stereo pairs for training and 5907 pairs for testing. We use the official code to assess the mean absolute depth error on all the subsequent frames provided by [1], named *Test-Original*.

However, as pointed by STTR [13], the depth of following frames are interpolated by forwarding kinematics information of the point cloud. This would lead to the synchronization issues and kinematics offsets, resulting in inaccurate depth values for subsequent frames. Following STTR [13], we further collect the first frame of each video and build our *Test-19* set, which subset consists of 19 images of resolution  $1080 \times 1024$  with the maximum disparity of 263 pixels. The left and right 100 pixels were cropped due to invalidity after rectification. We further provide a complete *one-key evaluation toolbox* for disparity evaluation. **dVPN** is provided by Hamlyn Centre Laparoscopic, with 34320 pairs of rectified stereo images for training and 14382 pairs for testing. There is no ground truth depth for these frames. To compare the performance of our model with other methods, we use the image warping accuracy [32] as our evaluation metrics, *i.e.*, Structural Similarity Index Measure (SSIM) [29], and Peak-Signal-to-Noise Ratio (PSNR) [9]. Note that for a fair comparison, we exclude self-supervised methods in our comparison as they directly optimize the disparity with image warping losses [32].

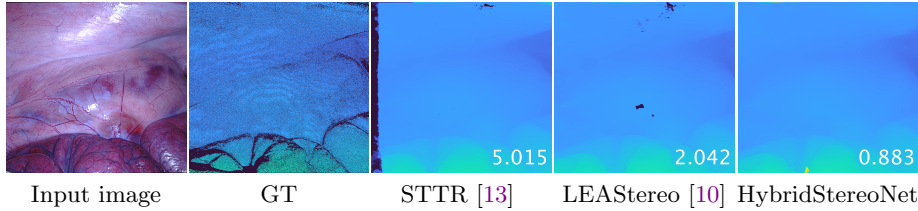
### 3.2 Implementation

We implemented all the architectures in Pytorch. A random crop with size  $336 \times 336$  is the only data argumentation technique used in this work. We use the SGD optimizer with momentum 0.9, cosine learning rate that decays from



**Table 3.** Quantitative results on the *Test-19* set (evaluated on all pixels). We compare our method with various state-of-the-art methods, by bad pixel ratio disparity errors.

Methods	EPE [px] ↓	RMSE [px] ↓	bad 2.0 [%] ↓	bad 3.0 [%] ↓	bad 5.0 [%] ↓
STTR [13]	6.1869	20.4903	8.4266	8.0428	7.5234
LEAStereo [5]	1.5224	<b>4.1135</b>	4.5251	3.6580	2.1338
HybridStereoNet	<b>1.4096</b>	4.1336	<b>4.1859</b>	<b>3.4061</b>	<b>2.0125</b>

**Fig. 7.** Qualitative results with bad 3.0 value on the *Test-19* set. Our model predicts dense fine-grained details even for occlusion areas.

0.025 to 0.001, and weight decay 0.0003. Our pretrained models on SceneFlow are conducted on two Quadro GV100 GPUs. Due to the limitation of public laparoscopic data and the ground truth, we train the proposed variant models on a synthetic dataset, SceneFlow [18], which has per-pixel ground truth disparities. It contains 35,454 training and 4,370 testing rectified image pairs with a typical resolution of  $540 \times 960$ . We use the “finalpass” version as it is more realistic.

### 3.3 Results

**SCARED2019.** We summarized the evaluation results on *Test-Original* in Table 2, including methods reported in the challenge summary paper [1]. We also provided unsupervised methods from [10] in the supplementary material. Note that our model never seen the training set. As shown in Table 3, our results show an improvement compared with the state-of-the-art Pure CNN method LEAStereo and a transformer-based method STTR [13] on our reorganized *Test-19* set. Please refer to supplementary material for more results on non-occluded areas.

**dVPN.** As shown in Table 4, our results are better than other competitors. Noting that DSSR [17] opts for the same structure with STTR [13]. Several unsupervised methods, *e.g.*, SADepth [10], are not included in this table as they are trained with reconstruction losses which will lead to a high value of the evaluated SSIM metric. However, for the sake of completeness, we provide their results in supplementary material.

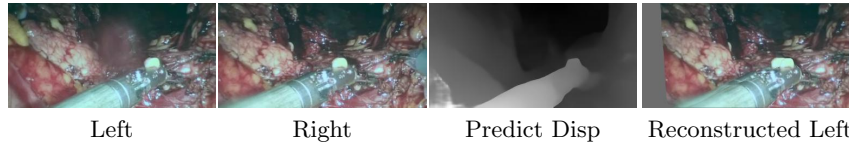
## 4 Conclusion

In this paper, we extensively investigated the effect of transformers for laparoscopic stereo matching in terms of loss landscapes, projected learning trajectories, in-domain/cross-domain accuracy, and proposed a new hybrid stereo



**Table 4.** Evaluation on the *dVPN* test set ( $\uparrow$  means higher is better). We directly report results of ELAS and SPS from [10]. Results of E-DSSR and DSSR are from [17].

Method	Training	Mean SSIM $\uparrow$	Mean PSNR $\uparrow$
ELAS [8]	No training	47.3	-
SPS [30]	No training	54.7	-
E-DSSR [17]	No training	$41.97 \pm 7.32$	$13.09 \pm 2.14$
DSSR [17]	No training	$42.41 \pm 7.12$	$12.85 \pm 2.03$
LEAStereo [5]	No training	55.67	15.25
HybridStereoNet	No training	<b>56.98</b>	<b>15.45</b>



**Fig. 8.** Quantitative results on the *dVPN* dataset. The invalid areas on the left side of the reconstructed image are the occluded areas.

matching framework. We empirically found that for laparoscopic stereo matching, using transformers to learn feature presentations and CNNs to aggregate matching costs can lead to faster convergence, higher accuracy and better generalization. Our proposed HybridStereoNet surpasses state-of-the-art methods on SCARED2019 and *dVPN* datasets.

## References

1. 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)
2. Cartucho, J., Tukra, S., Li, Y., S. Elson, D., Giannarou, S.: Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. *CMBBE: Imaging & Visualization* **9**(4), 331–338 (2021)
3. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R.: Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment* **2019**(12), 124018 (2019)
4. Cheng, X., Zhong, Y., Dai, Y., Ji, P., Li, H.: Noise-aware unsupervised deep lidar-stereo fusion. In: *CVPR* (2019)
5. Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., Drummond, T., Ge, Z.: Hierarchical neural architecture search for deep stereo matching. In: *NeurIPS*. vol. 33 (2020)
6. Chong, N., Si, Y., Zhao, W., Zhang, Q., Yin, B., Zhao, Y.: Virtual reality application for laparoscope in clinical surgery based on siamese network and census transformation. In: *MICAD*. pp. 59–70. Springer (2021)
7. 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)

8. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: ACCV. pp. 25–38. Springer (2010)
9. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
10. 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: MICCAI. pp. 227–237. Springer (2021)
11. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. ICLR (2017)
12. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. *NeurIPS* **31** (2018)
13. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: ICCV. pp. 6197–6206 (October 2021)
14. Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. arXiv preprint arXiv:2109.07547 (2021)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
16. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
17. 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: MICCAI. pp. 415–425. Springer (2021)
18. 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: CVPR. pp. 4040–4048 (2016)
19. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
20. Nicolau, S., Soler, L., Mutter, D., Marescaux, J.: Augmented reality in laparoscopic surgical oncology. *Surgical oncology* **20**(3), 189–201 (2011)
21. Overley, S.C., Cho, S.K., Mehta, A.I., Arnold, P.M.: Navigation and robotics in spinal surgery: where are we now? *Neurosurgery* **80**(3S), S86–S99 (2017)
22. Qin, Z., Sun, W., Deng, H., Li, D., Wei, Y., Lv, B., Yan, J., Kong, L., Zhong, Y.: cosformer: Rethinking softmax in attention. In: ICLR (2022)
23. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition. pp. 31–42. Springer (2014)
24. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* **10**(5), 1299–1319 (1998)
25. Sun, W., Qin, Z., Deng, H., Wang, J., Zhang, Y., Zhang, K., Barnes, N., Birchfield, S., Kong, L., Zhong, Y.: Vicinity vision transformer. In: arxiv. p. 2206.10552 (2022)
26. Wang, J., Zhong, Y., Dai, Y., Birchfield, S., Zhang, K., Smolyanskiy, N., Li, H.: Deep two-view structure-from-motion revisited. In: CVPR. pp. 8953–8962 (June 2021)
27. Wang, J., Zhong, Y., Dai, Y., Zhang, K., Ji, P., Li, H.: Displacement-invariant matching cost learning for accurate optical flow estimation. In: NeurIPS (2020)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: ICCV. pp. 568–578 (2021)

29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *TIP* **13**(4), 600–612 (2004)
30. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: *ECCV*. pp. 756–771. Springer (2014)
31. 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)
32. Zhong, Y., Dai, Y., Li, H.: Self-supervised learning for stereo matching with self-improving ability (2017)
33. Zhong, Y., Dai, Y., Li, H.: 3d geometry-aware semantic labeling of outdoor street scenes. In: *ICPR* (2018)
34. Zhong, Y., Dai, Y., Li, H.: Stereo computation for a single mixture image. In: *ECCV* (September 2018)
35. Zhong, Y., Ji, P., Wang, J., Dai, Y., Li, H.: Unsupervised deep epipolar flow for stationary or dynamic scenes. In: *CVPR* (2019)
36. Zhong, Y., Li, H., Dai, Y.: Open-world stereo video matching with deep rnn. In: *ECCV* (2018)
37. Zhong, Y., Loop, C.T., Byeon, W., Birchfield, S., Dai, Y., Zhang, K., Kamenev, A., Breuel, T.M., Li, H., Kautz, J.: Displacement-invariant cost computation for stereo matching. In: *IJCV* (March 2022)