# StereoDiffusion: Temporally Consistent Stereo Depth Estimation with Diffusion Models

Haozheng Xu[1], Chi Xu[1], and Stamatia Giannarou[1]

The Hamlyn Centre for Robotic Surgery, Department of Surgery and Cancer, Imperial College London, London, SW7 2AZ, UK
{haozheng.xu19,chi.xu20,stamatia.giannarou}@imperial.ac.uk

**Abstract.** In Minimally Invasive Surgery (MIS), temporally consistent depth estimation is necessary for accurate intraoperative surgical navigation and robotic control. Despite the plethora of stereo depth estimation methods, estimating temporally consistent disparity is still challenging due to scene and camera dynamics. The aim of this paper is to introduce the StereoDiffusion framework for temporally consistent disparity estimation. For the first time, a latent diffusion model is incorporated into stereo depth estimation. Advancing existing depth estimation methods based on diffusion models, StereoDiffusion uses prior knowledge to refine disparity. Prior knowledge is generated using optical flow to warp the disparity map of the previous frame and predict a reprojected disparity map in the current frame to be refined. For efficient inference, fewer denoising steps and an efficient denoising scheduler have been used. Extensive validation on MIS stereo datasets and comparison to state-of-the-art (SOTA) methods show that StereoDiffusion achieves the best performance and provides temporally consistent disparity estimation with high-fidelity details, despite having been trained on natural scenes only.

**Keywords:** Deep Learning · Depth Estimation · Diffusion Model.

## 1 Introduction

In MIS, the depth of surgical scenes is essential information for applications including augmented reality, 3D tissue reconstruction, and surgical navigation. Temporal consistency in depth estimation is required to enable reliable guidance of surgical robots. Depth estimation using stereoscopic cameras has been a popular approach in surgical vision. This involves matching pixels between two views from a calibrated stereo camera and estimating the disparity between them to measure depth. Recently, deep learning methods have been proposed for stereo depth estimation [8][7][17][16]. They rely on specialized architectures and task-specific loss functions, such as cost volumes, feature warps, or photometric reprojection losses. Although these methods have achieved great performance, estimating temporally consistent depth from video sequences is still challenging due to issues including the presence of dynamic objects, scene deformations,

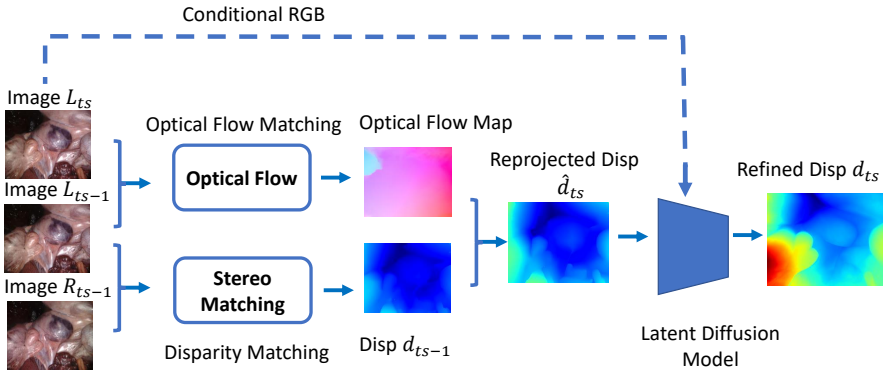


**Fig. 1.** Outline of the StereoDiffusion framework.

and camera motion which make feature matching difficult. The problem is further compounded in scenes with uniform textures, changing illumination, and occlusions, conditions particularly prevalent in MIS.

Denoising Diffusion Probabilistic Models (DDPMs) [5] have illustrated the capability of generating high-fidelity images with remarkable details by learning to iteratively reverse images degraded with Gaussian noise to capture rich knowledge about the visual information. A lot of attention has been given in transferring the diffusion model into classical computer vision tasks [12][6][2]. In the Marigold architecture [6], the diffusion model has been applied for monocular depth estimation. The diffusion model predicts relative scene depth (based on random noise and conditional RGB images), and a global scale is then applied to transform the relative depth to actual depth values. Although this work can generate detailed relative depth maps, it is quite sensitive to the type of the scene and struggles to recover actual depth in scenarios such as surgical scenes due to the lack of training data. Therefore, our aim is to overcome the above limitations and integrate diffusion models into the stereo depth estimation pipeline.In this paper, we introduce the StereoDiffusion framework for stereo depth estimation. For the first time, a latent diffusion model is used for disparity refinement. Advancing existing depth estimation methods based on diffusion models, in our work instead of treating the diffusion model as a regression model, we use prior knowledge to refine disparity. To enable real-time disparity inference, fewer denoising steps and an efficient denoising scheduler have been used. The prior knowledge fed to the diffusion model is generated using optical flow to warp the disparity map of the previous frame and predict a reprojected disparity map in the current frame which requires refinement due to flaws or imperfections caused by the scene dynamics. Performance evaluation on MIS data and comparison to SOTA methods verifies the robustness, temporal consistency and generalisability of StereoDiffusion, despite having been trained on natural scenes only.

## 2 Methods

The StereoDiffusion framework proposed in this work for temporally consistent disparity estimation is composed of the following two main parts, (1) Disparity and optical flow estimation given a sequence of stereo images (2) Disparity refinement using a latent diffusion model.

### 2.1 Disparity and Optical Flow Estimation

The first part of our framework focuses on the estimation of disparity and optical flow as shown in Fig. 1. The aim is to warp the disparity estimated in the previous frame with the optical flow between the previous and the current frames to generate the disparity map at the current frame. Given the left and right camera images $L_{ts-1}$ and $R_{ts-1}$, respectively at time step $ts-1$, the disparity module predicts the disparity map $\mathbf{d}_{ts-1}$. Given the image pair $L_{ts-1}$ and $L_{ts}$, the optical flow module predicts the optical flow map from $L_{ts-1}$ to $L_{ts}$. Then $\mathbf{d}_{ts-1}$ is warped with the optical flow to generate the reprojected disparity $\hat{\mathbf{d}}_{ts}$. This process is iteratively conducted frame by frame. In this work, we use the off-the-shelf pre-trained RAFT models [8][15] for both disparity and optical flow estimation. However, any other disparity and optical flow estimation model can be used.

### 2.2 Disparity Refinement with Latent Diffusion Model

The reprojected disparity $\hat{\mathbf{d}}_{ts}$ contains flaws or imperfections caused by the scene dynamics and errors in the optical flow. Hence, the second part of our framework focuses on refining the disparity $\hat{\mathbf{d}}_{ts}$ using a latent diffusion model to improve both its temporal consistency and accuracy.

For this purpose, we approach the disparity estimation as a conditional denoising diffusion generation task. In this work, the Stable Diffusion V2 [10] is used as our latent diffusion model because of its memory and time efficiency. Contrary to diffusion models operating directly on data [12], latent diffusion models perform diffusion in a compressed latent space for computational efficiency and high-resolution image generation. This latent space is established via a variational autoencoder (VAE) which extracts a latent representation which compresses the original data.

The above latent diffusion model is fine-tuned with the conditional distribution $D(\mathbf{z}^{(\mathbf{d})}) \mid \mathbf{z}^{(\mathbf{x})})$ over disparity $\mathbf{d} \in \mathbb{R}^{W \times H}$, where the condition $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$ is an RGB image which in our work is the current left image $L_{ts}$. The latent representation of a disparity map and of a conditioning image is estimated as $\mathbf{z}^{(\mathbf{d})}$ and $\mathbf{z}^{(\mathbf{x})}$, respectively. By applying the same VAE to both the disparity map and the conditioning image, we ensure that their latent representations are well-aligned for further processing. Fig. 2 illustrates the training processes.

The denoising pipeline consists of the *forward* and *reverse* processes. In the *forward* process, Gaussian noise is incrementally added to the latent representation of the disparity map at levels $t \in \{1, ..., T\}$ to generate noisy samples $\mathbf{z}_t^{(\mathbf{d})}$,

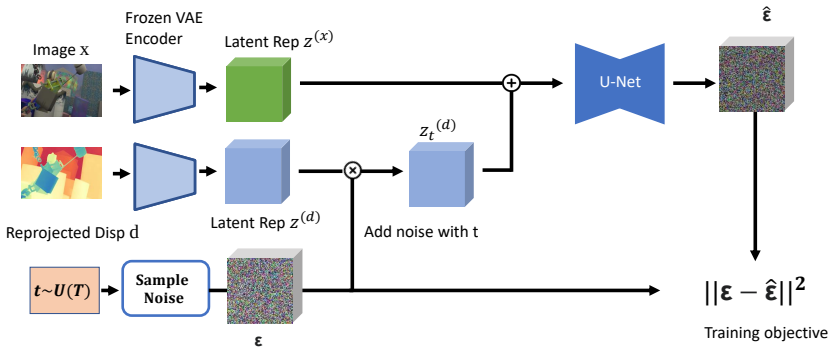


**Fig. 2.** Training Phase of the Latent Diffusion Model.

as:

$$\mathbf{z}_t^{(\mathbf{d})} = \sqrt{\overline{\alpha}_t}\mathbf{z}_0^{(\mathbf{d})} + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{1}$$

where, $\epsilon \sim \mathcal{N}(0, I)$ and $\overline{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$ where $\beta$ has fixed value from the set $\{\beta_1, ..., \beta_T\}$. In the *reverse* process, the conditional denoising model $\epsilon_\theta(\cdot)$, with parameters $\theta$, iteratively removes noise from $\mathbf{z}_t^{(\mathbf{d})}$ to recover $\mathbf{z}_{t-1}^{(\mathbf{d})}$.

The denoising model $\epsilon_\theta(\cdot)$ in this work is the U-Net. It is trained by selecting a pair $(\mathbf{d}, \mathbf{x})$ from the training dataset, applying to disparity $\mathbf{d}$ noise of a random level $t$, computing the noise estimate $\hat{\epsilon} = \epsilon_\theta(\mathbf{d}_t, \mathbf{x}, t)$, and learning to minimize the denoising diffusion objective function $\mathcal{L}$, defined as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{d}_0, \boldsymbol{\epsilon}\sim\mathcal{N}(0,I), t\sim\mathcal{U}(T)} \left\|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\right\|_2^2. \tag{2}$$

During inference, the refined disparity $\hat{\mathbf{d}}$ (corresponding to $\mathbf{d}_0$) is reconstructed from the reprojected disparity map, by iteratively applying the trained denoiser $\epsilon_\theta(\mathbf{z}_t^{(\mathbf{d})}, \mathbf{x}, t)$.

**Disparity Normalization:** To guarantee the generalizability of the diffusion model to different types of scenes, we normalize every reprojected disparity map prior to feeding it to the diffusion model. We use min-max normalization and adjust the disparity values to the range $[-1, 1]$, as:

$$\mathbf{d}_{norm} = \left(\frac{\mathbf{d} - \mathbf{d}_{min}}{\mathbf{d}_{max} - \mathbf{d}_{min}} - 0.5\right) \times 2, \tag{3}$$

where, $d$ is the original disparity map and $d_{norm}$ is the normalized one. This disparity range adjustment fits the conventional data range of the pretrained VAE for better training. Moreover, this normalization makes the model focus on the relative disparity values rather than predict disparity values directly. This helps the diffusion model refine the reprojected disparity map based on its disparity data distribution and can be transferred into other scenes, ensuring generalizability across different scenarios. In addition, the min-max normalisation is superior

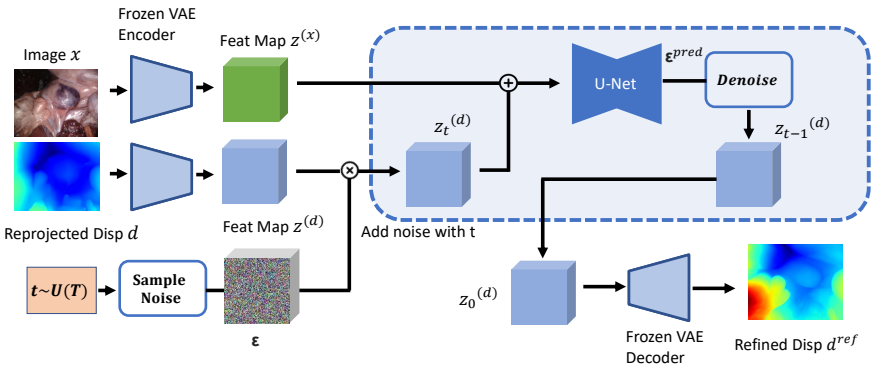


**Fig. 3.** Inference Phase of the Latent Diffusion Model.

than normalizing the disparity by the image width as we found that the latter makes the final prediction unstable and with noisy patches during inference.

### 2.3 Inference of StereoDiffusion Model

The inference pipeline for the disparity refinement is illustrated in Fig. 3. Similar to the training phase, the disparities are normalised following Eq. (3). Given the reprojected disparity map and the conditional RGB image, the frozen VAE first converts the two inputs into their latent representations. Then, Gaussian noise is added to the latent representation of the reprojected disparity map. The denoiser is then applied iteratively to estimate the refined disparity map in the latent space. To recover the disparity from the latent space, the frozen VAE decoder $\mathcal{D}$ is used to reconstruct the refined disparity $\hat{d} = \mathcal{D}(\mathbf{z^{(d)}})$. Since the VAE used to establish the latent space of the diffusion model, is designed for 3-channel data input, we replicate the disparity map into the 3 channels at the encoder and decoder of the VAE to satisfy the data format. Hence, the refined disparity map is estimated as the average of the three channels recovered by the VAE decoder. To accelerate inference, we apply the DDIM [14] scheduler to perform non-Markovian sampling with re-spaced steps. Also, we use only 10 denoising steps which as shown in our results, it achieves high accuracy.

### 2.4 Implementation

The Stable Diffusion v2 [10] is used with pre-trained model weight setup with a v-objective [11]. The text conditioning module is disabled to save memory usage. We use the DDPM noise scheduler [5] with 1000 diffusion steps for training and the DDIM scheduler [14] with 10 sampling steps for fast inference. We train with a batch size of 4 on 2 Nvidia RTX A5000 GPUs with accumulative steps set equal to 8. Half-precision training is used to save memory usage and the Adam optimizer with a learning rate of $3 \cdot 10^{-5}$.

## 3 Experiments and Results

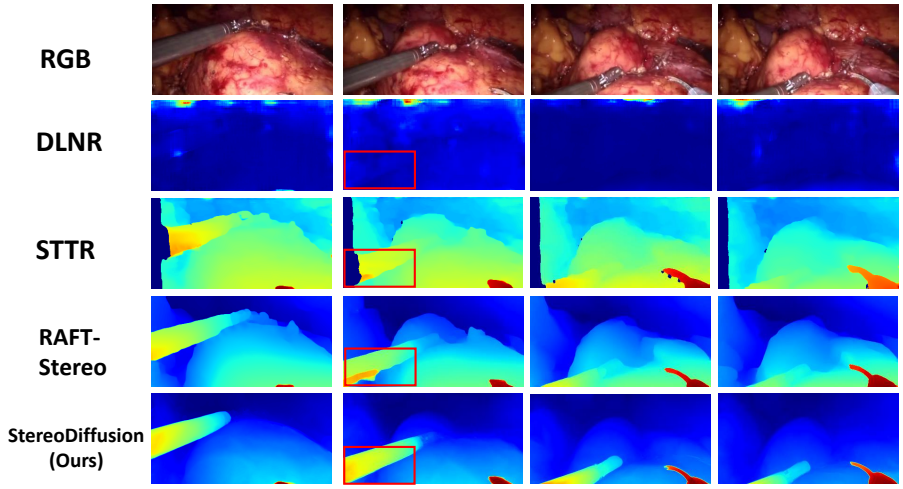


**Fig. 4.** Temporal Consistency Comparison on the Hamlyn Dataset.

**Table 1.** Validation on the SCARED dataset. The best results are highlighted in bold.

| Method | SCARED 2019 Test | | | | | | SCARED Small | | |
|---|---|---|---|---|---|---|---|---|---|
| | Testdata 1 | | | Testdata 2 | | | Keyframes | | |
| | EPE ($px$) | D3 (%) | MAE ($mm$) | EPE | D3 | MAE | EPE | 3 px | MAE |
| DLNR [17] | 4.14 | 33.91 | 3.98 | 5.32 | 36.37 | 4.68 | 1.45 | 4.12 | 1.32 |
| STTR [7] | 6.52 | 39.96 | 4.14 | 8.23 | 40.13 | 5.91 | 6.03 | 9.52 | 11.31 |
| RAFT-Stereo [8] | 3.89 | 36.91 | 3.74 | 4.22 | 38.52 | 4.28 | 1.16 | 4.59 | 1.01 |
| StereoDiffusion | **3.22** | **25.22** | **3.13** | **3.67** | **27.34** | **3.67** | **1.05** | **3.55** | **1.00** |

### 3.1 Datasets for taining and testing

In our framework, only the diffusion model is trainable while all the other components are pretrained. We train the diffusion model in our framework on the Sceneflow [9] and Middlebury [4] datasets. The Sceneflow dataset consists of three subsets namely, FlyingThings3D, Driving and Monkaa. We use the final-pass set of Sceneflow which contains 36k images of dynamic scenes with motion blur and defocus blur. The diffusion model is then finetuned on the 23 Middlebury training images to enable it to generate images with more realistic details.

For testing, three different datasets have been used namely, the SCARED, STIR and Hamlyn datasets. The SCARED dataset [1] includes 7 training and 2 test subsets. We use the 2 test datasets which contain 8 video clips in total. The videos have a resolution of 1280x1024. The STIR dataset [13] includes 566 stereo video clips from both in vivo and ex vivo settings, with more than 3,000 sparse points in the whole dataset, visible in the infrared spectrum. The centre point

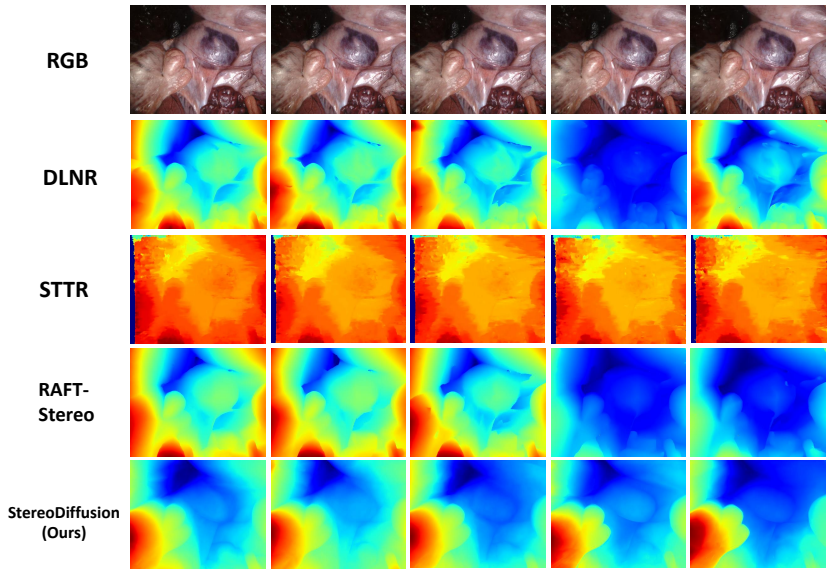


**Fig. 5.** Temporal Consistency Comparison on the SCARED dataset.

of the labeled bounding boxes on each pair of stereo images is used to estimate ground truth disparity for our evaluation. From the Hamlyn dataset [3], we used the partial nephrectomy video for qualitative validation as there is no ground truth.

### 3.2 Performance evaluation study

We compare the performance of StereoDiffusion with two state-of-the-art disparity models namely, RAFT-stereo [8], STTR [7] and DLNR [17] on the test datasets. The performance is evaluated in terms of the average End-Point-Error (EPE) of disparity, the percent of pixels with EPE greater than 3 pixels (D3) and the Mean Absolute Error (MAE) of depth on SCARED. And we use EPE, D3 and Intersection Over Union (IOU) to evaluate disparity results on STIR to estimate the accuracy in matching bounding boxes. Temporal consistency is validated only qualitatively due to the lack of ground truth optical flow. All the models have been trained on the natural scene datasets only and tested on the medical scene data, without finetuning on the medical datasets.

Tables 1 and 2, show the quantitative evaluation of our model compared with the SOTA stereo depth estimation models. As it can be seen, StereoDiffusion outperforms the other models in both benchmarks, especially in terms of the D3 metric, which verifies that it predicts more accurate disparity maps, with less pixels having disparity error over 3 pixels. Validation only on the keyframes of the SCARED dataset with accurate ground truth shows that StereoDiffusion not only has superior performance but it also achieves $1mm$ accuracy.

**Table 2.** Validation on the STIR dataset. Best results are highlighted in bold.

| Method | Disparity Evaluation | | Bounding Box Evaluation |
|---|---|---|---|
| | EPE ($px$) ↓ | D3 (%) ↓ | IOU ↑ |
| DLNR [17] | 2.98 | 22.54 | 0.823 |
| STTR [7] | 3.48 | 24.03 | 0.818 |
| RAFT-Stereo [8] | 2.61 | 23.12 | 0.832 |
| StereoDiffusion | **2.15** | **15.84** | **0.855** |

Fig. 4 and Fig. 5 show the disparity maps extracted by the compared models on a short sequence of consecutive frames from the Hamlyn and the SCARED datasets, respectively. The RGB refer to the left RGB images to save space.

In the disparity map generated by DLNR in Fig. 4 it is not possible to distinguish different scene structures due to the outliers observed at the top of the map. STTR generates disparity by processing the images in patches. Hence, it misses fine details as can be noticed in Fig. 4 along the edges of the surgical tools and the border between the tissue and the background. Also, it generates uneven disparity with holes and gaps. RAFT and DLNR can better preserve details on the different structures present in the surgical scene compared to STTR. However, the temporal consistency of all these models is poor as the range of the disparity values varies significantly across the image sequence, creating flickering results. StereoDiffusion generates more precise and sharper disparity maps while the distribution of the predicted disparity values remains consistent. This verifies the ability of our method to generate temporally consistent disparity maps. In Fig. 4, the light reflection creates a bright line close to the edge of the shaft of the left-hand needle driver. RAFT treats it as an independent object and estimates along this line disparities which are significantly different compared to those of the main body of the needle driver. StereoDiffusion manages to generate smooth disparities along the shaft of the surgical tool. Video results of the above sequences have been provided in the Supplemental material. Despite not training on surgical scenes, StereoDiffusion consistently outperformes SOTA disparity estimation models by generating sharper and temporally consistent disparity maps, verifying its generalizability.

## 4 Conclusion

In this paper, we have proposed the StereoDiffusion framework which uses a latent diffusion model with prior knowledge for disparity refinement with improved temporal consistency. Compared with previous work which utilizes diffusion model for direct depth estimation, we combine both stereo depth estimation and diffusion model to provide a more generalizable and robust framework. The experimental results verify that although the model has been trained on natural scene data only, it can still estimate temporally consistent disparity of surgical scenes without any finetuning. Our future work will focus on transferring our proposed framework to other computer vision tasks such as image segmentation.

**Acknowledgments.** This work was supported by the Royal Society [URF\R\201014].

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Allan, M., McLeod, A.J., Wang, C.C., et. al, J.R.: Stereo correspondence and reconstruction of endoscopic data challenge. CoRR **abs/2101.01133** (2021)
2. Amit, T., Nachmani, E., Shaharabany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. CoRR **abs/2112.00390** (2021), https://arxiv.org/abs/2112.00390
3. Hamlyn Centre Laparoscopic / Endoscopic Video Datasets: Hamlyn Centre Laparoscopic / Endoscopic Video Datasets. https://hamlyn.doc.ic.ac.uk/vision/
4. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. 2007 IEEE Conference on Computer Vision and Pattern Recognition pp. 1–8 (2007)
5. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
6. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation (2023)
7. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6197–6206 (October 2021)
8. Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. arXiv preprint arXiv:2109.07547 (2021)
9. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. CVPR pp. 4040–4048 (2016)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
11. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
12. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
13. Schmidt, A., Mohareri, O., DiMaio, S., Salcudean, S.: Stir: Surgical tattoos in infrared (2023). https://doi.org/10.21227/w8g4-g548
14. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=St1giarCHLP
15. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II. p. 402–419. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58536-5_24, https://doi.org/10.1007/978-3-030-58536-5_24
16. Tukra, S., Xu, H., Xu, C., Giannarou, S.: Generalizable stereo depth estimation with masked image modelling. Healthcare Technology Letters (12 2023). https://doi.org/10.1049/htl2.12067
17. Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1327–1336 (2023)