CT-Video Matching for Retrograde Intrarenal Surgery Based on Depth Prediction and Style Transfer

Honglin Lei, Yanqi Pan, Tao Yu, Zuoming Fu, Chongan Zhang, Xinsen Zhang, Peng Wang, Jiquan Liu *
Xuesong Ye and Huilong Duan

Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering &Instrument Science, Zhejiang University, Hangzhou 310027, China; 21715095@zju.edu.cn (H.L.); 21715095@zju.edu.cn (Y.P.); yutao11615043@zju.edu.cn (T.Y.); fzm21315045@zju.edu.cn (Z.F.); kevin_07@zju.edu.cn (C.Z.); zxs1997@zju.edu.cn (X.Z.); 0016902@zju.edu.cn (P.W.); yexuesong@zju.edu.cn (X.Y.); duanhl@zju.edu.cn (H.D.)
*
Correspondence: liujq@zju.edu.cn

Abstract: Retrograde intrarenal surgery (RIRS) is a minimally invasive endoscopic procedure for the treatment of kidney stones. Traditionally, RIRS is usually performed by reconstructing a 3D model of the kidney from preoperative CT images in order to locate the kidney stones; then, the surgeon finds and removes the stones with experience in endoscopic video. However, due to the many branches within the kidney, it can be difficult to relocate each lesion and to ensure that all branches are searched, which may result in the misdiagnosis of some kidney stones. To avoid this situation, we propose a convolutional neural network (CNN)-based method for matching preoperative CT images and intraoperative videos for the navigation of ureteroscopic procedures. First, a pair of synthetic images and depth maps reflecting preoperative information are obtained from a 3D model of the kidney. Then, a style transfer network is introduced to transfer the ureteroscopic images to the synthetic images, which can generate the associated depth maps. Finally, the fusion and matching of depth maps of preoperative images and intraoperative video images are realized based on semantic features. Compared with the traditional CT-video matching method, our method achieved a five times improvement in time performance and a 26% improvement in the top 10 accuracy.

Keywords: ureteroscopy; CNN; CT-video; style transfer

1. Introduction

Image-guided endoscopic navigation has been a hot topic of research in surgical navigation, which can provide visual aids to clinicians for interventional surgery. Retrograde intrarenal surgery (RIRS) is one of the procedures for the image-guided treatment of kidney stones, performed through a ureteroscopy. RIRS has become one of the important methods of treating kidney stones, especially for the removal of large calculi. In traditional RIRS, the surgeons usually use preoperative CT to reconstruct a 3D model of the kidney by preoperative CT to determine the location of the kidney stones and understand the structure of the kidney, and then finds and removes the stones through intraoperative endoscopic video images based on experience. Due to the kidney having many internal branches, it can be difficult to locate stones and guarantee that all branches have been searched for during surgery, which may cause misdiagnosis. It is therefore important to assist the surgeon in RIRS by locating the ureteroscope through image navigation. At present, image fusion of preoperative CT images and intra-operative video images has become a popular solution.

There are mainly two methods to realize the fusion of preoperative CT and intra-operative video, one is the 2D–2D registration method [1–3], and the other is the 3D–3D registration method [4–7]. However, both methods have their limitations in RIRS. Traditional 2D–2D registration methods, which were based on global pixel points of grayscale images, were time-consuming. While 3D–3D registration methods were used for matching
or registration after a point cloud reconstruction based on video images. The accuracy of point cloud reconstruction is affected heavily by kidney stones, water-filled, bubbles, flocculent, and other impurities in RIRS, and is not suitable for the RIRS scene.

2. Materials and Methods

Due to the limitations of traditional methods depicted above, a new method of preoperative CT and intraoperative video information fusion was proposed to solve the CT-video matching problem in ureteroscopic surgical guidance in this paper. Inspired by the application of deep learning and depth maps in the field of endoscopic guidance [8,9], we proposed to use depth maps as the intermediate connecting medium to match preoperative CT with intraoperative video. Our proposed method could be specified through the following steps.

(1) We first reconstructed a 3D model of the kidney based on CT images.
(2) We used a virtual camera to simulate the movement path of the real ureteroscope to generate pairs of images. Each image pair consists of one simulated image (SI) taken by a virtual camera and its corresponding depth map (DM). We call these pairwise images datasets (simulated images and depth maps) as SI-DM. Based on the SI-DM dataset, we trained a model to predict the depth of simulated images.
(3) We trained a model to transfer the style of endoscopic images (EI) into the style of simulated images mentioned in (2). Then we could indirectly obtain endoscopic images’ depth maps.
(4) Finally, by extracting features of the depth maps from SI and EI and calculating their similarity, we realized CT-video matching to avoid kidney stone misdiagnosis based on the depth maps from (2) and the depth maps from (3).

In this method, both the depth prediction network and the style transfer network are crucial. The style transfer network can transfer one modal data to another modal data, which means that if we know one modal data’s depth maps, another modal data’s depth maps can be predicted indirectly. This approach has been shown to be effective when applied to natural images [10]. Moreover, to avoid using traditional feature extraction methods to compute feature descriptors for matching, we use convolutional neural networks (CNNs) instead to extract deep semantic features of the depth map, which can improve the time performance.

The main contributions of our paper are as follows:

• We established a corresponding mapping relationship based on the depth map between the white-light ureteroscopic image and the virtual endoscopic image. In other words, we achieved depth prediction based on a single white-light endoscopic image.
• We extract abstract semantic features of the depth map from ureteroscopic images and simulated images captured by the virtual camera for CT-video matching. This approach achieves effective matching and significantly reduces the computational time consumption.
• The results show that our method achieved a 26% improvement in top 10 matching accuracy with a five times improvement in time performance.

Figure 1 shows the flow chart of our approach to navigate the ureteroscope by matching preoperative CT images and intraoperative video frames. In detail, first, a 3D model of the kidney was reconstructed by CT images. Then we generated an SI-DM dataset based on the 3D kidney and virtual reality to train the depth prediction model. Meanwhile, we introduced a style transfer network to transform ureteroscopic images into simulated images, whose style was like the SI dataset. Once we obtained both ureteroscopic images’ depth maps and SI dataset’s depth maps, we could extract their depth semantic features, respectively. Finally, based on the depth maps’ features, we realized a CT-video matching task to make sure each branch of the kidney was examined.
while, we introduced a style transfer network to transform ureteroscopic images into simulated images, whose style was like the SI dataset. Once we obtained both ureteroscopic images’ depth maps and SI dataset’s depth maps, we could extract their depth semantic features, respectively. Finally, based on the depth maps’ features, we realized a CT-video matching task to make sure each branch of the kidney was examined.

Figure 1. Flowchart of our proposed method.

2.1. Depth Prediction

A depth map was used as the connecting medium to realize the matching of CT and video. A depth map describes the spatial geometry information, with each pixel value representing the spatial distance of that pixel from the camera, and it can be applied in virtual reality, 3D reconstruction, and other fields. There have been a lot of related studies on natural images [11–17] that have studied how to obtain depth information from monocular images, that is, monocular depth prediction. In monocular depth prediction, it can be regarded as a regression problem, just like image segmentation. In this paper, we used the SI-DM dataset obtained from CT to train the depth map prediction model, mainly because we cannot obtain the real ureteroscopic images’ depth maps. This step is a fundamental part of solving the depth map of ureteroscope video images through style transfer. For virtual endoscopic images, a depth map can be recovered from CT image series directly. As to real endoscopic images, they are transferred to the virtual endoscopic domain at first, and then the depth maps are generated using the depth prediction model trained from virtual endoscopic images.

This paper uses an encoder–decoder network structure which was first proposed by Alhashim et al. [17].

2.1.1. Datasets

In this paper, the SI-DM dataset was like the RGB-D dataset [18–20]. It was generated from preoperative CT image sequences. The whole process is depicted in Figure 2. (1) We segmented CT images according to grey-scale thresholding and extracted interested anatomical sites; (2) We used the Marching Cubes (MC) [21] algorithm to reconstruct the interested anatomical sites for the kidney 3D model; (3) We extracted the center path of the kidney 3D model and used a virtual camera to simulate the movement path of a real ureteroscope; (4) We obtained pairwise SI-DM dataset by using the 3D object rendering imaging principle.
segmented CT images according to grey-scale thresholding and extracted interested anatomical sites; (2) We used the Marching Cubes (MC) [21] algorithm to reconstruct the interested anatomical sites for the kidney 3D model; (3) We extracted the center path of the kidney 3D model and used a virtual camera to simulate the movement path of a real ureteroscope; (4) We obtained pairwise SI-DM dataset by using the 3D object rendering imaging principle.

![Figure 2](image1.png)

**Figure 2.** Dataset acquisition process. (a) shows CT images and the segmentation of key anatomical sites; (b) represents the reconstructed 3D model of the kidney and its center path; (c) denotes a pair of SI-DM images.

We collected CT images for 3D model reconstructions from two patients at Shanghai Changhai Hospital. A Philips CT scanner was used with 1.5 mm scan thickness and 512 * 512 image resolution. In generating the synthetic data, we considered the points on the center path as the virtual camera’s location and rotated the virtual camera at different angles at each interval point in order to get more multi-angle virtual endoscopic images in the virtual endoscope. Here, the virtual camera can perform the functions of an RGB-D camera and can collect the depth map corresponding to each frame of the virtual image. As shown in Table 1, we generated 29,608 SI-DM images, of which 21,429 were training sets and 8179 were test sets.

Table 1. Dataset composition of deep prediction model.

| Total Datasets | Training Sets | Test Sets |
|----------------|---------------|-----------|
| 29,608         | 21,429        | 8179      |

2.1.2. Loss Function

In order to obtain better model precision, we combine different loss functions to conduct experiments. For in depth prediction tasks, researchers usually apply the L1 loss function and the L2 loss function. However, training with a sole reconstruction loss function can cause the model to tend to generate an average value, which can lead to ambiguous outputs. In the work of Alhashim. et al. [17], structural similarity (SSIM) error as a loss function was also mentioned. As mentioned in that work, using structural similarity as a loss function is a good loss term for depth estimating CNNs. We also tried this loss in this article. The formulas of the loss functions were as follows: y represents the ground truth of the depth value and \( \hat{y} \) represents the prediction of the depth value:

\[
L_{1\text{loss}} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (1)
\]

\[
L_{2\text{loss}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (2)
\]
\[ L_{SSIMloss} = \frac{1 - SSIM(y, \hat{y})}{2} \]  

We quantitatively evaluated the performance of the model. The quantitative evaluation generally uses threshold accuracy, root mean square error (RMSE), root mean square log error RMSE (log), average log10 error, and relative error (REL) to judge model performance. The formula of evaluation functions are as follows:

\[ \delta = \max \left( \frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i} \right) < \text{threshold} \]  

\[ L_{rmse} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \]  

\[ L_{rmse(log)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\log(y_i) - \log(\hat{y}_i))^2} \]  

\[ L_{log10} = \frac{1}{n} \sum_{i=1}^{n} |\log_{10}(y_i) - \log_{10}(\hat{y}_i)| \]  

\[ L_{rel} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \]  

Usually, in formula (4), the value of the threshold is set as 1.25, 1.25², or 1.25³.

2.1.3. Implementation Details

All implementation and training were done in PyTorch [22], using densenet-169 [23] pre-trained on ImageNet [24] to initialize the encoder parameters, and the decoder used skip connection and up-sample structure. The Adma [25] optimizer with parameter values of \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \) was used. The learning rate was set to 0.001, the training epoch was set to 50, the batch size was set to four, and the training input image size was 480 × 640.

2.2. Style Transfer

In Section 2.1, we have been able to predict depth maps from simulated virtual endoscopic monocular images (SI-DM dataset). However, our purpose was to obtain depth maps from real ureteroscopic images. Since video images and virtual images belong to two domains, migration of different modality data is a problem of domain adaptation [26]. In order to achieve this, we introduced a style transfer neural network [27]. The style in the paper refers to the different data distribution domains of these two kinds of images, such as color, texture obtained with the same content. Style transfer refers to retaining the image content and transferring the image to the target style from the source style.

In this paper, we use the image style transfer method to transfer the real endoscopic images to the virtual endoscopic images because the depth image matching under the same data domain is more suitable than the cross-domain matching from the perspective of data distribution and image alignment, so that the depth prediction model trained in Section 2.1 was effective for the real endoscopic images. There are many methods to realize style transfer [28–31]. In [32], CycleGAN was adopted and unpaired data was used for training. The structure of the CycleGAN is shown in Figure 3. A represents the real endoscopic image domain (EI), while B represents the simulated images domain (SI). It consists of two discriminators and two generators. The input image Input_A generates Fake_B through Generator A to B, and Fake_B generates Rec_A through Generator B to A. After two transformations, Rec_A belongs to the A domain, and the model is optimized by comparing the similarity between the Input_A and the Rec_A, and Input_B is also processed in the same way.
The same data domain is more suitable than the cross-domain matching from the perspective of data distribution and image alignment, so the depth prediction model trained in Section 2.1 was effective for the real endoscopic images. There are many methods to realize style transfer [28–31]. In [32], CycleGAN was adopted and unpaired data was used for training. The structure of the CycleGAN is shown in Figure 3. A represents the real endoscopic image domain (EI), while B represents the simulated images domain (SI). It consists of two discriminators and two generators. The input image Input_A generates Fake_B through Generator A to B, and Fake_B generates Rec_A through Generator B to A. After two transformations, Rec_A belongs to the A domain, and the model is optimized by comparing the similarity between the Input_A and the Rec_A, and Input_B is also processed in the same way.

**Figure 3.** A style migration model of CycleGAN was used to transform white light images into virtual endoscopic images.

### 2.2.1. Datasets

The advantage of the CycleGAN adopted in this paper is that it does not need a pair of source-target domain matching images, which provides the convenience of our dataset acquisition. We only need to obtain the images from the two domains separately, without considering their relationship, and even the number of image datasets in the two domains do not need to be consistent (however, in order to guarantee model training more successfully, the gap of the two domains’ datasets should not be too large). The source domain dataset we used was derived from two videos of clinical ureteroscopic lithotomy in Shanghai Changhai Hospital. In order to collect high-quality endoscopic images, we got rid of invalid frames from the surgical videos, as shown in Figure 4. At the same time, the target domain dataset used in this paper was the simulated virtual endoscopic images (SI) obtained in Section 2.1. As shown in Table 2, we used 1747 source domain images (EI) and in order to ensure that the number of datasets in the two domains would not be too different, 2767 target domain images were randomly selected from the SI to train style transfer model.

| Total Datasets | Source Domain Images | Target Domain Images |
|----------------|----------------------|---------------------|
| 4514           | 1747                 | 2767                |
Table 2. Examples of typical ureteroscope video images. The first row shows problematic or uninformative frames due to being blurry, kidney stones, or flocculent, and the second row displays informative images without any impurity.

![Ureteroscope video images](image)

**Figure 4.** Examples of typical ureteroscope video images. The first row shows problematic or uninformative frames due to being blurry, kidney stones, or flocculent, and the second row displays informative images without any impurity.

2.2.2. Implementation Details

This paper conducted experiments based on CycleGAN. The generator used the ResNet [33] architecture, the discriminator used PatchGAN [31], and the cycle loss function was the L1 loss function. The model training platform was the same as Section 2.1. The initial learning rate was 0.0002, the Adam optimizer with parameter values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used, the batch size was set to 4, and the training epoch was 250.

2.3. Semantic Feature Matching

After acquiring depth maps of the two domains' images, this section explains how to realize the matching method between real endoscopic images and corresponding SI images. In order to improve time performance, we extracted semantic features of the depth maps to match endoscopic images with SI images. We used an auto-encoder [34] network to extract high-level semantic features of each image. We encoded the $240 \times 320$ image into the feature vector of 3840 dimensions.

We used vector Euclidean Distance to calculate the similarity of semantic features, which was a common but effective method. In the matching process, we built a semantic feature database by using the preoperative depth map dataset, and searched for the best match in the database by using the intraoperative depth map features to realize matching between semantic features. Then, CT-video registration was re-implemented based on the mapping relationship between semantic features and preoperative and intraoperative anatomical positions. The process of semantic feature matching is shown in Figure 5. The valid frames of the intraoperative video were successively inputted to style transfer network, depth prediction, and semantic feature extraction. Similarity analysis was conducted between the calculated features and the semantic feature database constructed from the semantic features of the preoperative depth map dataset, and the top-10 images with the best matches were output after sorting. When SI-DM images were acquired in Section 2.1.1, the virtual camera’s position and angle corresponding to each simulated image were saved. After obtaining the matched top-10 simulated images, the positions of the frames could be displayed in the 3D model, which could reflect whether the match was correct according to the position of the branch. When one or more of the matched top-10 images matched the correct anatomical location, the match was considered a successful case.
3. Results

This section verifies the effectiveness of the proposed methods. First, we examined the results of our depth prediction model. As shown in Table 3, we compared the effects of different loss functions on the experiment. We found that the difference in loss functions had little impact on the performance of the model. The loss functions and error functions are shown in 2.1.2, with $\delta_1$, $\delta_2$, and $\delta_3$ set as 1.25, 1.25$^2$, and 1.25$^3$, respectively. The accuracy of the model could achieve 94.1% in depth prediction on the SI-DM dataset using the loss function of L2 + SSIM. Figure 6 shows some examples of SI-DM’s test dataset. We could find that the predicted depth map was generally consistent with the ground truth.

![Figure 5. The process of semantic feature matching.](image)

**Figure 5.** The process of semantic feature matching.

![Figure 6. Examples of model depth map prediction results.](image)

**Figure 6.** Examples of model depth map prediction results. The first row shows virtual endoscopic images (SI), the second row shows the predicted depth map, and the last row displays the ground truth of the depth map.
Table 3. Comparison of model depth prediction results under different loss functions.

| Loss Function | Error↓ | Threshold Accuracy↑ |
|---------------|--------|---------------------|
|               | RMSE   | RMSE (log) | log10 | Rel | δ1  | δ2  | δ3  |
| L1            | 0.088  | 0.223      | 0.197 | 0.256 | 0.730 | 0.870 | 0.933 |
| L2            | 0.095  | 0.235      | 0.208 | 0.269 | 0.701 | 0.856 | 0.928 |
| SSIM          | 0.102  | 0.258      | 0.227 | 0.304 | 0.675 | 0.844 | 0.917 |
| L1 + SSIM     | 0.098  | 0.239      | 0.210 | 0.271 | 0.689 | 0.869 | 0.936 |
| L2 + SSIM     | 0.091  | 0.223      | 0.198 | 0.256 | 0.721 | 0.881 | 0.941 |

Then, we evaluated the effectiveness of style transfer by analyzing the feature distributions of the source domain images (EI), the target domain images (SI), and the images generated by the CycleGAN. We input each of these three types of images into the discriminator of the target domain (Discriminator B in Figure 3, which determines whether an image belongs to the target domain). Each image was denoted by a $1 \times 1024$ feature vector. We used t-SNE (t-distributed stochastic neighbor embedding) to visualize these features in the same 2D coordinate space (see Figure 7a). We could see that the images generated by the CycleGAN were closer to those in the target domain, but due to the discriminator having a discriminatory effect, the feature distribution of both has a certain boundary. Similarly, these images were also inputted to ResNet-50 for feature vectoring and visualization. The results are shown in Figure 7b; the distribution of the images generated by the CycleGAN and target domain images are similar with some overlapping areas, which meant that the conventional CNN classifier could not distinguish between the two types of images correctly. This proved that we could achieve the migration between domain distributions through style transfer. The examples of style transfer results are shown in Figure 8.

![Figure 7](image-url)  
Figure 7. Feature distribution results. (a) represents the feature distribution results of the images input into the discriminator (D) network, and (b) displays the feature distribution results of the images input into the ResNet-50 network. Since D has the discriminant effect, while ResNet-50 does not have, so b_real and b_fake are closer in the feature distribution.

In addition, in order to further prove the effectiveness of the style transfer, we directly input the ureteroscopic images into the depth prediction model, and the results of the depth maps are shown in Figure 9b. For comparison, we input the same ureteroscopic images into the depth prediction model after the style transfer and got the depth maps as shown in Figure 9d. As shown in Figure 9, we can see that, by comparison, the depth...
map without style transfer had poor performance, which showed that style transfer could greatly improve the results of depth prediction.

Figure 7. Feature distribution results. (a) represents the feature distribution results of the images input into the discriminator (D) network, and (b) displays the feature distribution results of the images input into the ResNet-50 network. Since D has the discriminant effect, while ResNet-50 does not have, so b_real and b_fake are closer in the feature distribution.

Figure 8. Examples of style transfer results. The first row represents the ureteroscopy images, and the second row represents the results of the style transfer.

Figure 9. Examples of depth prediction of ureteroscopic video images by style transfer. (a) shows ureteroscopic images, (b) shows the results of direct depth prediction, (c) shows style transfer, and (d) shows depth prediction after style transfer.

Finally, we evaluate the performance of the matching method. In order to highlight the advantages of our method, we also used the traditional 2D–2D matching method on the same data. The traditional method was based on the direct similarity calculation of EI and SI. The similarity calculation method was based on the similarity calculation of image
pixel points, and the commonly used method was the SSIM calculation. We compared our methods (see Figure 10c) with the traditional 2D–2D methods (see Figure 10b), and our methods showed better performance with more accurate matching positions. For further analysis, we used 1039 ureteroscopy images (EI) and the feature database, which included 2077 preoperative depth maps’ semantic features to do the matching experiment. We analyzed the matching accuracy and matching time consumption for Top-1, Top-5, and Top-10, and the experimental results are shown in Table 4. Accuracy is the ratio of correct matches to the total number of matches, and the matching elapsed time is calculated by dividing the time to complete all images by the number of images. It could be seen that although our method had no advantage in terms of top-1, our methods were 16 and 26% more accurate in terms of top-5 and top-10 respectively, with a 5 five times improvement in time performance, which proved that our method works better than the traditional methods.

Figure 9. Examples of depth prediction of ureteroscopic video images by style transfer. (a) shows ureteroscopic images, (b) shows the results of direct depth prediction, (c) shows style transfer, and (d) shows depth prediction after style transfer.

Finally, we evaluate the performance of the matching method. In order to highlight the advantages of our method, we also used the traditional 2D–2D matching method on the same data. The traditional method was based on the direct similarity calculation of EI and SI. The similarity calculation method was based on the similarity calculation of image pixel points, and the commonly used method was the SSIM calculation. We compared our methods (see Figure 10c) with the traditional 2D–2D methods (see Figure 10b), and our methods showed better performance with more accurate matching positions. For further analysis, we used 1039 ureteroscopy images (EI) and the feature database, which included 2077 preoperative depth maps’ semantic features to do the matching experiment. We analyzed the matching accuracy and matching time consumption for Top-1, Top-5, and Top-10, and the experimental results are shown in Table 4. Accuracy is the ratio of correct matches to the total number of matches, and the matching elapsed time is calculated by dividing the time to complete all images by the number of images. It could be seen that although our method had no advantage in terms of top-1, our methods were 16 and 26% more accurate in terms of top-5 and top-10 respectively, with a 5 five times improvement in time performance, which proved that our method works better than the traditional methods.

Figure 10. Visual comparison of the traditional 2D–2D method and our method. (a) represents ureteroscopic images, (b) represents the matching results of the traditional method, and (c) represents the matching results of our method. Our method based on the depth map showed a better performance.

Table 4. A quantitative comparison of CT-video matching accuracy and time performance by the different methods.

|          | 2D-2D       | Our Approach |
|----------|-------------|--------------|
| Top1     | 42.92%      | 42.31%       |
| Top5     | 47.06%      | 63.26%       |
| Top10    | 50.43%      | 76.54%       |
| time (per picture) | 6.23 s | 1.26 s |

4. Discussion
This paper aims to solve the difficulty of RIRS caused by complex lumen in ureteroscopy. Nevertheless, our method has some limitations. The premise that our method can be effective is that the depth information varies from anatomical structure to anatomical structure.
However, in actual clinical images, the depth information distribution of different anatomical locations may be similar, in which case, our method will have an increased matching failure rate and obtain false-negative results, i.e., there is a matching ambiguity problem (as shown in Figure 11).

At the same time, the amount of our data needs to be increased, although the thousands of virtual endoscopy frames generated from our collected CT data are sufficient for training pre-training-based depth estimation and style transfer models. The evaluation results show that our method has significantly surpassed traditional matching methods, and we believe that the proposed matching method can bring more inspiration for the exploration of artificial intelligence methods in the field of nephrology surgery. We will collect more different data for further experiments in the future.

5. Conclusions

A CT-video matching method based on a depth map was proposed for the ureteroscopy scene. We applied the method to clinical data and compared it with the traditional 2D–2D registration method. The results show that our method outperforms the traditional method in terms of accuracy and time performance, with a 26% improvement in accuracy and a five times improvement in speed for the top 10. In addition, even though the time performance is improved by a factor of five, our time performance of 1.26 s per image is still far from meeting the clinical requirements. In fact, the depth map-based matching method in this paper is not limited to ureteroscopy application scenarios but can also be considered for other endoscopy scenarios. We believe that that with the continuous exploration of deep learning technology, future research work can optimize the matching method based on this paper to achieve better matching accuracy and speed trade-off.

Author Contributions: Methodology, Y.P. and T.Y.; software, H.L. and Y.P.; validation, Z.F., C.Z. and X.Z.; investigation, Y.P.; resources, P.W., J.L., X.Y. and H.D.; data curation, P.W.; writing—original draft preparation, H.L. and Y.P.; writing—review and editing, J.L., P.W. and T.Y.; visualization, H.L.; project administration, J.L. All authors have read and agreed to the published version of the manuscript.
**Funding:** This research was funded by the National Natural Science Foundation of China, Grant/Award Numbers: 31771072, 81827804 and the National Key Research and Development Program of China, Grant/Award Numbers: 2017YFB1302803.

**Institutional Review Board Statement:** Ethical review and approval were waived for this study, due to the retrospective design of the study and the fact that all data used were from existing and anonymized clinical datasets.

**Informed Consent Statement:** Patient consent was waived due to the retrospective nature of this study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Acknowledgments:** The authors would like to thank Xiaofeng Gao, Zeyu Wang, and Ling Li for their help in preparing the data in this paper.

**Conflicts of Interest:** No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication. I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

**References**

1. Merritt, S.A.; Khare, R.; Bascom, R.; Higgins, W.E. Interactive CT-video registration for the continuous guidance of bronchoscopy. *IEEE Trans. Med. Imaging* **2013**, *32*, 1376–1396. [CrossRef] [PubMed]

2. Wan, Y.; Wu, Q.; He, X. Dense Feature Correspondence for Video-Based Endoscope Three-Dimensional Motion Tracking. In Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2014), Valencia, Spain, 1–4 June 2014; pp. 49–52.

3. Luo, X.; Takabatake, H.; Natori, H.; Mori, K. Robust Real-Time Image-Guided Endoscopy: A New Discriminative Structural Similarity Measure for Video to Volume Registration. In Proceedings of the International Conference on Information Processing in Computer-Assisted Interventions (IPCAI 2013), Heidelberg, Germany, 26 June 2013; pp. 91–100.

4. Billings, S.D.; Sinha, A.; Reiter, A.; Leonard, S.; Ishii, M.; Hager, G.D.; Taylor, R.H. Anatomically Constrained Video-CT Registration via the V-IMLOP Algorithm. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2016), Athens, Greece, 17–21 June 2016; pp. 133–141.

5. Mirotta, D.J.; Wang, H.; Taylor, R.H.; Ishii, M.; Gallia, G.L.; Hager, G.D. A system for video-based navigation for endoscopic endonasal skull base surgery. *IEEE Trans. Med. Imaging* **2011**, *31*, 963–976. [CrossRef] [PubMed]

6. Leonard, S.; Reiter, A.; Sinha, A.; Ishii, M.; Taylor, R.H.; Hager, G.D. Image-Based Navigation for Functional Endoscopic Sinus Surgery Using Structure from Motion. In Proceedings of the Medical Imaging 2016: Image Processing, San Diego, CA, USA, 1–3 March 2016; p. 97840V.

7. Leonard, S.; Sinha, A.; Reiter, A.; Ishii, M.; Gallia, G.L.; Taylor, R.H.; Hager, G.D. Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data. *IEEE Trans. Med. Imaging* **2018**, *37*, 2185–2195. [CrossRef] [PubMed]

8. Visentini-Scarzanella, M.; Sugitani, T.; Kaneko, T.; Koto, S. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. *Int. J. Comput. Assist. Radial. Surg.* **2017**, *12*, 1089–1099. [CrossRef] [PubMed]

9. Luo, X.; Zeng, H.-Q.; Du, Y.-P.; Cheng, X. Towards Multiple Instance Learning and Hermann Weyl’s Discrepancy for Robust Image-Guided Bronchoscopic Intervention. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019), Shenzhen, China, 13–17 October 2019; pp. 403–411.

10. Atapour-Abarghouei, A.; Breckon, T.P. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2800–2810.

11. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374.

12. Liu, F.; Shen, C.; Lin, G.; Reid, I. earning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2024–2039. [CrossRef] [PubMed]

13. Cao, Y.; Wu, Z.; Shen, C. Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3174–3182. [CrossRef]

14. Li, B.; Shen, C.; Dai, Y.; Hengel, A.V.D.; He, M. Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFs. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 July 2015; pp. 1119–1127.
15. Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; Yuille, A.L. Towards Unified Depth and Semantic Prediction from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2800–2809.

16. Zhang, Z.; Schwing, A.G.; Fidler, S.; Urtasun, R. Monocular Object Instance Segmentation and Depth Ordering with Cnns. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2614–2622.

17. Alhashim, I.; Wonka, P. High quality monocular depth estimation via transfer learning. *arXiv* 2018, arXiv:1812.11941.

18. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.

19. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Learning 3D scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* 2009, 31, 824–840. [CrossRef] [PubMed]

20. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* 2013, 32, 1231–1237. [CrossRef]

21. Lorensen, W.E.; Cline, H.E. Marching cubes: A high resolution 3D surface construction algorithm. *ACM Siggraph Comput. Graph.* 1987, 21, 163–169. [CrossRef]

22. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic Differentiation in PyTorch. In Proceedings of the 2017 Neural Information Processing Systems, Long Bench, CA, USA, 4–9 December 2017.

23. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.

24. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the 2016 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016. [CrossRef]

25. Xiang, Y.; Wang, N.; Liu, J.; Hou, X. Demystifying Neural Style Transfer. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 19–25 August 2017; pp. 2230–2236. [CrossRef] [PubMed]