A disparity refinement framework for learning-based stereo matching methods in cross-domain setting for laparoscopic images

Zixin Yang, Richard Simon, Cristian Linte

Zixin Yang, Richard Simon, Cristian Linte, "A disparity refinement framework for learning-based stereo matching methods in cross-domain setting for laparoscopic images," Proc. SPIE 12466, Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling, 1246607 (3 April 2023); doi: 10.1117/12.2654804

SPIE.

Event: SPIE Medical Imaging, 2023, San Diego, California, United States

A Disparity Refinement Framework for Learning-based Stereo Matching Methods in Cross-domain Setting for Laparoscopic Images

Zixin Yang[a], Richard Simon[b], and Cristian Linte[a,b]

[a]Center for Imaging Science, Rochester Institute of Technology Rochester, NY 14623, USA
[b]Department of Biomedical Engineering, Rochester Institute of Technology Rochester, NY 14623, USA

ABSTRACT

Stereo matching methods that enable depth estimation are crucial for visualization enhancement applications in computer-assisted surgery (CAS). Learning-based stereo matching methods are promising to predict accurate results for applications involving video images. However, they require a large amount of training data, and their performance may be degraded due to domain shifts. Maintaining robustness and improving performance of learning-based methods are still open problems. To overcome the limitations of learning-based methods, we propose a disparity refinement framework consisting of a local disparity refinement method and a global disparity refinement method to improve the results of learning-based stereo matching methods in a cross-domain setting. Those learning-based stereo matching methods are pre-trained on a large public dataset of natural images and are tested on a dataset of laparoscopic images. Results from the SERV-CT dataset showed that our proposed framework can effectively refine disparity maps on an unseen dataset even when they are corrupted by noise, and without compromising correct prediction, provided the network can generalize well on unseen datasets. As such, our proposed disparity refinement framework has the potential to work with learning-based methods to achieve robust and accurate disparity prediction. Yet, as a large laparoscopic dataset for training learning-based methods does not exist and the generalization ability of networks remains to be improved, it will be beneficial to incorporate the proposed disparity refinement framework into existing networks for more accurate and robust depth estimation.

Keywords: Stereo Matching, Disparity Refinement, Endoscopy, Variational Model, Cross-domain Generalization, Optical Flow.

1. INTRODUCTION

Depth estimation plays a key role in surgical navigation[1] and visualization enhancement applications.[2] Stereo endoscopy is commonly used to enable depth estimation.[3] Stereo correspondences represented by disparity maps can be estimated via stereo matching techniques[4] to provide depth measurements with known intrinsic and extrinsic camera calibration.

In the era of deep learning, learning-based stereo matching methods are reported to achieve high performance on several public benchmarks and outperform traditional methods.[5,6] However, their achievements are based on several prerequisites: 1. A large amount of data is available for training. 2. The training dataset and the testing dataset are identically distributed. Those prerequisites are not generally satisfied in the surgical data science arena. Obtaining a large dataset of specific surgical scenes to train learning-based methods is impractical, usually with millions of parameters, especially for fully-supervised learning methods that require accurate ground truth. Given various texture and surgery settings, there is also no guarantee that the distribution difference between the training dataset and the testing dataset is negligible. As a large laparoscopic dataset with accurate ground

Further author information: For correspondence send to Zixin Yang
Zixin Yang: E-mail: yy8898@rit.edu
Richard Simon: E-mail: rasbme@rit.edu
Cristian A. Linte: E-mail: calbme@rit.edu

Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling, edited by
Cristian A. Linte, Jeffrey H. Siewerdsen, Proc. of SPIE Vol. 12466, 1246607
© 2023 SPIE · 1605-7422 · doi: 10.1117/12.2654804

truth does not exist, several methods[3,7] use cross-domain datasets[5,6] for training, which may not be sufficiently robust, as the distribution change or domain shift can jeopardize performance.[8]

To overcome the limitations of learning-based methods, we propose a framework for refinement of disparity maps of laparoscopic images predicted from learning-based methods trained on a large stereo dataset of natural images. The proposed disparity refinement framework consists of local and global refinement methods (LDR and GDR).

Our contributions are summarized as follows: 1) We present a disparity refinement framework based on traditional methods for learning-based stereo matching methods consisting of LDR and GDR methods in a cross-domain setting. 2) We present an LDR method to measure the confidence of disparity maps and refine disparity values in low confidence regions. The method is designed to refine errors concentrated in small regions and provide a more robust initialization for the subsequent global disparity refinement method. 3) We present a GDR method using our illumination invariant multi-resolution variational model to refine various artifacts, especially for errors concentrated in large regions.

2. METHODOLOGY

We first use the LDR to detect low confidence regions on the predicted disparity map by assuming that outliers in the disparity map strongly violate the smoothness and photometric consistency assumptions. After the detection, the disparity values of low confidence regions are interpolated from the surrounding high confidence regions. Subsequently, we use the GDR, which introduces a multi-resolution variational model to further refine the disparity from the previous step.

2.1 Local Disparity Refinement

Several assumptions are made to estimate the confidence of the disparity. Firstly, we assume that the tissue surface is relatively smooth. The pixel $\mathbf{x}$ should have high confidence in a smoothness confidence map $C_s(\mathbf{x})$ if the disparity value $u(\mathbf{x})$ is consistent with its surrounding pixels $\overline{u}_w(\mathbf{x})$:

$$C_s(\mathbf{x}) = 1 - \alpha_s \cdot \left| \frac{u(\mathbf{x}) - \overline{u}_w(\mathbf{x})}{\overline{u}_w(\mathbf{x})} \right|, \tag{1}$$

where $\overline{u}_w(\mathbf{x})$ is the mean disparity value of the local window with the size w, and α_s is the hyper-parameter.

Secondly, we assume that outliers in the disparity map tend to violate the photo-consistency assumption strongly. Intensities of outliers $I_s(\mathbf{x})$ in the source (left) image and their matched points $I_t(\mathbf{x} + u(\mathbf{x}))$ in the target (right) image would have a large difference. Due to illumination differences, the intensity values of corresponding images may not be the same. However, it is reasonable that outliers would strongly violate this assumption. Therefore, the photo-consistency confidence $C_p(\mathbf{x})$ is defined as:

$$C_p(\mathbf{x}) = 1 - \alpha_p \cdot \left| \frac{I_s(\mathbf{x}) - I_t(\mathbf{x} + u(\mathbf{x}))}{I_s(\mathbf{x})} \right|, \tag{2}$$

where α_p is the hyper-parameter. Hence, pixels with incorrect disparity values have low confidence values in $C_p(\mathbf{x})$.

Thirdly, we assume that in specular highlights and border occlusions,[9] predicted disparities would tend to be unreliable. In specular highlights, pixel intensities are saturated and uniform. Border occlusions result from the fact that the right camera misses some of the leftmost portions of the field of view of the left camera.

We set confidence values in these regions as zeros by introducing the specular highlight mask $M_s(\mathbf{x})$ and the boundary occlusions mask $M_b(\mathbf{x})$:

$$M_s(\mathbf{x}) = \begin{cases} 1 & \text{if } S(x) > th_s \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$M_b(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} + D(\mathbf{x}) \text{ exists in the right image} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where $S(x) \in [0, 1]$ is the value in HSV color space.

The final confidence map $C_f(\mathbf{x})$ is defined as the product between the above confidence maps and masks:

$$C_f(\mathbf{x}) = M_v(\mathbf{x}) \cdot M_s(\mathbf{x}) \cdot C_p(\mathbf{x}) \cdot C_s(\mathbf{x}). \tag{5}$$

Pixels are selected as outliers if their confidence values are below the threshold value of th_f. Next, we search for inliers along with eight directions for each outlier, similar to SGM.[10] Then, for each outlier, its disparity is replaced with the median value of the eight inliers.

2.2 Global Disparity Refinement

We formulate the stereo matching as a variational problem. Disparity values $u(x)$ between the source I_s and target I_t images are predicted by the minimization of an energy function composed of a data term E_{data}, and a regularization term E_S:

$$\min_u \left[\lambda \, E_{data}(u, I_t, I_s) + E_S(u)\right], \tag{6}$$

where λ denotes the weight between E_{data} and E_S. E_{data} measures the similarity of pixels in I_s and I_t using:

$$E_{data}(\mathbf{u}) = \int_\Omega \left|\mathbf{D}(P(x+u, I_t)) - \mathbf{D}(P(x, I_s))\right|^2 dx. \tag{7}$$

Here, Ω denotes the image domain, x presents the pixel location, $P(x, I)$ is the patch that contains the local intensities, and $\mathbf{D}$ is a novel illumination invariant descriptor:

$$P(x) = \begin{bmatrix} I(x_4) & I(x_3) & I(x_2) \\ I(x_5) & I(x) & I(x_1) \\ I(x_6) & I(x_7) & I(x_8) \end{bmatrix}. \tag{8}$$

$$\mathbf{D}(P(x_0, I)) = \frac{\mathbf{A}(P(x_0, I))}{\|\mathbf{A}(P(x_0, I))\|}. \tag{9}$$

$$\mathbf{A}(P(x_0, I)) = \begin{bmatrix} |I(x_0) - I(x_1)| \\ |I(x_0) - I(x_2)| \\ \vdots \\ |I(x_0) - I(x_8)| \end{bmatrix}. \tag{10}$$

.

$I(x_i) \in [1, 2, ..., 8]$ denotes locations relative to the central pixel x_0. $\mathbf{D}$ is a 8 component vector calculated using Eq. 9 and 10. Our descriptor $\mathbf{D}$ is a simpler form of a descriptor proposed in.[11] It represents the normalized image gradient about the central pixel of patch $P(x_0)$ which was shown to be invariant to linear illumination changes:

$$\mathbf{D}(P(x+u, I)) = \mathbf{D}(aP(x, I) + b). \tag{11}$$

To preserve discontinuities at sharp object transitions and avoid staircasing artifacts in the in the calculated disparity map, we use a Huber function as a regularization term:

$$E_S = \int_\Omega |\nabla u|_\epsilon dx, \tag{12}$$

where

$$|r|_\epsilon = \begin{cases} \frac{r^2}{2\epsilon} & 0 \leq |r| \leq \epsilon, \\ |r| - \frac{\epsilon}{2} & \epsilon < |r|. \end{cases} . \tag{13}$$

and ϵ is a small positive constant.

Finally, Eq. 6 takes the following form:

$$\min_u \left[\lambda \int_\Omega \left| \mathbf{D}(P(x+u, I_t)) - \mathbf{D}(P(x, I_s)) \right|^2 dx + \int_\Omega |\nabla u(x)|_\epsilon dx \right], \tag{14}$$

which is solved by applying a primal-dual minimization scheme proposed in.[12] We use a coarse-to-fine warping framework to tackle large displacements. A scale factor of 0.5 is used to construct an image pyramid of n levels. At each level, we perform m warping iterations of optimizing energy functional Eq. 14. In each level, the warping iteration is initialized with the current disparity field u, and a target image is warped towards the source image using the current disparity map.

3. RESULTS

We use several state-of-the-art learning-based methods, including PSMnet,[13] AAnet,[14] LEAStereo,[15] and STTR,[7] to generate raw disparity maps of images from the SERV-CT[3] dataset. All the above learning methods are executed by training their public models on the SceneFlow dataset.[5] Root mean square disparity error (RMSE Disparity) and root mean square depth error (RMSE Depth) are used to evaluate errors between estimated and ground truth results. We examine the refinement performance of our proposed LDR and GDR and compare them with the closet work SDR[16] to ours. Parameters of methods used in experiments are shown in Table 1.

Table 1. Summary of parameters used in our methods.

Parameter	Function	Value
α_s	Hyper-parameter in Eq. 1	20
α_p	Hyper-parameter in Eq. 2	2
th_f	Threshold to select outliers from the final confidence map	0.5
λ	Weight between data term and regularization term in Eq. 6	0.5
ϵ	Hyper-parameter in Huber norm regularizer Eq. 7	0.1
m	Warping iterations at each image pyramid to solve Eq. 14	50
n	Levels of image pyramid to solve Eq. 14	4

Table 2. Evaluation results on SERV-CT dataset. The statistical significance between the errors before refinement and after refinement is identified by $*(p < 0.05)$.

	Occlusions included		Occlusions not included	
Method	RMSE Disparity (pixel)	RMSE Depth (mm)	RMSE Disparity (pixel)	RMSE Depth (mm)
PSMnet[13]	38.91 ± 17.11	25.16 ± 8.77	33.07 ± 16.30	23.01 ± 8.88
PSMnet[13] + LDR	31.70 ± 17.57	22.55 ± 8.76	29.04 ± 16.63	21.33 ± 9.14
PSMnet[13] + LDR + GDR	∗ 8.17 ± 10.38	∗ 7.89 ± 7.85	∗ 6.59 ± 9.99	∗ 6.48 ± 7.69
PSMnet[13] + SDR[16]	32.04 ± 21.03	22.30 ± 8.07	28.00 ± 19.78	20.79 ± 8.16
AAnet[14]	11.71 ± 7.21	13.33 ± 5.90	9.35 ± 6.20	11.85 ± 5.74
AAnet[14] + LDR	9.39 ± 7.13	10.04 ± 6.72	7.53 ± 6.29	8.39 ± 6.32
AAnet[14] + LDR + GDR	∗ 4.15 ± 2.08	∗ 4.86 ± 2.70	∗ 2.76 ± 1.43	∗ 3.47 ± 2.40
AAnet[14] + SDR[16]	9.22 ± 4.92	11.36 ± 4.87	6.98 ± 3.80	9.51 ± 4.60
LEAStereo[15]	7.79 ± 5.58	12.21 ± 7.39	6.27 ± 5.10	10.27 ± 6.69
LEAStereo[15] + LDR	6.06 ± 4.66	9.66 ± 6.54	4.49 ± 3.89	7.05 ± 5.57
LEAStereo[15] + LDR + GDR	∗ 3.96 ± 1.79	∗ 4.45 ± 2.03	∗ 2.58 ± 1.31	∗ 3.06 ± 1.73
LEAStereo[15] + SDR[16]	5.05 ± 2.83	7.24 ± 3.68	4.17 ± 2.38	6.42 ± 3.56
STTR[7]	17.22 ± 6.38	27.06 ± 5.16	4.27 ± 3.47	5.34 ± 4.00
STTR[7] + LDR	13.23 ± 6.25	∗ 22.44 ± 5.77	3.34 ± 3.13	4.20 ± 3.81
STTR[7] + LDR + GDR	∗ 4.86 ± 3.04	∗ 5.97 ± 3.57	2.95 ± 1.68	3.36 ± 1.70
STTR[7] + SDR[16]	10.71 ± 4.84	16.38 ± 6.15	3.64 ± 3.49	4.83 ± 3.63

Quantitative results are presented in Table 2. Decreases in errors are observed after each refinement stage, especially in the GDR stage. When including the occluded region in the evaluations, raw disparity maps estimated from LEAStereo[15] have the lowest 2D and 3D errors, with an RMSE of 7.79 ± 5.58 pixel (12.21 ± 7.39 mm). The errors are minimized to 6.06 ± 4.66 and 9.66 ± 6.54 after LDR stage, and 3.96 ± 1.79 and 4.45 ± 2.03 after GDR stage. All results from all networks have higher accuracy after excluding occluded regions, and STTR[7] has the lowest error with 4.27 ± 3.47 pixel (5.34 ± 4.00 mm). The results of STTR[7] can be further improved to 3.34 ± 3.13 pixel (4.20 ± 3.81 mm) at the LDR stage, and 2.95 ± 1.68 pixel and 3.36 ± 1.70 mm at the GDR stage. Our method can also refine disparity maps predicted by PSMnet with significant errors. Excluding occluded regions, errors of PSMnet are refined from 31.70 ± 17.57 pixels to 6.59 ± 9.99 pixels.

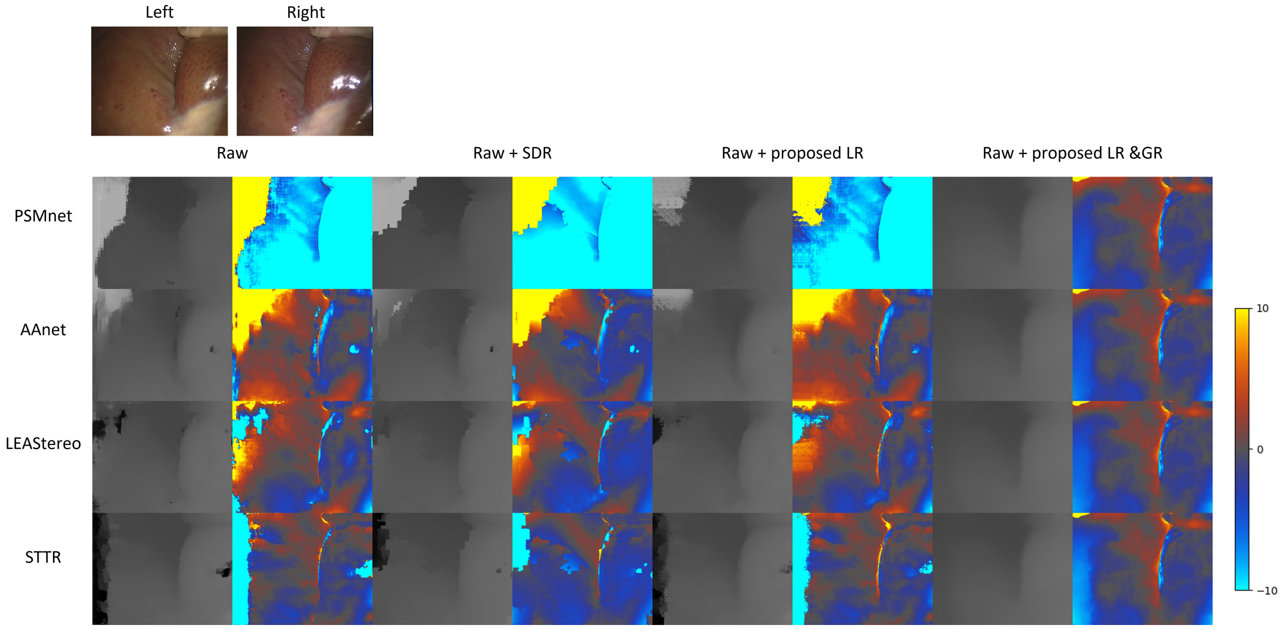

Figure 1. Disparity maps of learning-based methods (PSMnet,[13] AAnet,[14] LEAStereo,[15] STTR[7]) and their refined results by SDR,[16] our proposed LDR, and LDR + GDR, with disparity error maps compared with ground truth. Stereo endoscopic images and ground truth disparity map are shown on the top.

We visualize predicted disparity maps and their refined results by our methods in Fig. 1. We observe that raw disparity maps generated from the learning-based methods show significant disparity errors, especially in regions containing imaging artifacts such as specular highlights, occlusions, low texture and illumination differences. These error regions (the noise-corrupted regions) are characterized by spikes, strikes, and holes and are not continuous with their surrounding areas. These imaging artifacts are common in endoscopic images but not in the natural images that they are trained on. Small error regions can be refined effectively via the proposed LDR, and SDR.[16] However, they are not able to refine large error-corrupted areas, such as errors around boundary occlusions. The proposed GDR could further improve the results refined by LDR, and furthermore, various imaging artifacts still existing can also be effectively refined via the GDR.

4. CONCLUSION

We have presented a robust and accurate disparity refinement framework that integrates the successful use of local and global disparity refinement methods for learning-based stereo matching methods in a cross-domain setting. Applications that require visualization augmentation, such as surgical navigation and 3D organ visualization, rely on stereo matching. Learning-based approaches produce encouraging outcomes, however, they are constrained when used with endoscopic images for several reasons: they require a lot of training data, which is not readily available for endoscopic medical images and their performance suffers, due to the domain gap and large number of artifacts. Our proposed framework removes these barriers and demonstrates that by combining learning-based and traditional methods, we can yield robust and accurate results. Our proposed method provides assistance with applications in need of visualization enhancement that employ learning-based techniques that rely on laparoscopic images featuring small training datasets, limited ground truth data available, and various artifacts for accurate depth estimation.

REFERENCES

[1] B. Lin, Y. Sun, X. Qian, *et al.*, "Video-based 3d reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey," *The International Journal of Medical Robotics and Computer Assisted Surgery* **12**(2), 158–178 (2016).

[2] R. Modrzejewski, T. Collins, B. Seeliger, *et al.*, "An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions," *International journal of computer assisted radiology and surgery* **14**(7), 1237–1245 (2019).

[3] P. E. Edwards, D. Psychogyios, S. Speidel, *et al.*, "Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction," *Medical image analysis* **76**, 102302 (2022).

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision* **47**(1), 7–42 (2002).

[5] N. Mayer, E. Ilg, P. Hausser, *et al.*, "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*, 4040–4048 (2016).

[6] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3061–3070 (2015).

[7] Z. Li, X. Liu, N. Drenkow, *et al.*, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6197–6206 (2021).

[8] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing* **312**, 135–153 (2018).

[9] S. Huq, A. Koschan, and M. Abidi, "Occlusion filling in stereo: Theory and experiments," *Computer Vision and Image Understanding* **117**(6), 688–704 (2013).

[10] H. Hirschmuller, "Stereo vision based mapping and immediate virtual walkthroughs," (2003).

[11] D.-H. Trinh and C. Daul, "On illumination-invariant variational optical flow for weakly textured scenes," *Computer Vision and Image Understanding* **179**, 1–18 (2019).

[12] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of mathematical imaging and vision* **40**(1), 120–145 (2011).

[13] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5410–5418 (2018).

[14] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1959–1968 (2020).

[15] X. Cheng, Y. Zhong, M. Harandi, *et al.*, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems* **33**, 22158–22169 (2020).

[16] T. Yan, Y. Gan, Z. Xia, *et al.*, "Segment-based disparity refinement with occlusion handling for stereo matching," *IEEE Transactions on Image Processing* **28**(8), 3885–3897 (2019).