ACCURATE MULTISCALE SELECTIVE FUSION OF CT AND VIDEO IMAGES FOR REAL-TIME ENDOSCOPIC CAMERA 3D TRACKING IN ROBOTIC SURGERY

Xiongbiao Luo

Department of Computer Science, Xiamen University, Xiamen 36105, China

ABSTRACT

Robotic surgery requires endoscope 3D tracking to navigate the endoscope in the body. This paper proposes an accurate multiscale selective fusion framework to register 2D endoscopic video images to 3D pre-operative CT data for endoscope 3D tracking. Current video-based 3D tracking depends on the performance of the 2D-3D fusion procedure that suffers from inaccurate similarity and image uncertainties. To boost video-based 3D tracking, we develop multiscale selective similarity characterization to enhance the 2D-3D fusion procedure. Such fusion not only uses image pyramids in multiple scales to represent endoscopic images but also selects specific structure information from these multiscale images to compute the similarity. We validated our method on clinical data. Our method can reduce the current tracking error from 8.9 to 5.4 mm without using any external trackers, while it provides surgeons with robust real-time surgical 3D tracking.

***Index Terms*—** Endoscope Tracking, Surgical Navigation, 2D-3D Registration, Image Similarity, Robotic Surgery

1. INTRODUCTION

Robotic endoscopy commonly uses surgical endoscopes to visually inspect the interiors of the hollow organs (e.g., sinus, colon, stomach, and bladder). These endoscopes are usually a flexible (or rigid) and thin tube integrated with video cameras at its distal tip, providing surgeons with two-dimensional (2D) video sequences observed on medical monitors in the operating room. The most important function of the endoscope is to guide physicians for disease diagnosis and treatment on site, e.g., exploring, removing, or repairing suspicious tissues inside the body. An unavoidable challenge facing physicians is to determine the endoscope position relative to suspicious or precancerous tissues given online 2D endoscopic video images. Fortunately, endoscopic 3D camera tracking and navigation systems have been proposed to assist physicians by temporally locating the endoscope in a reference coordinate system such as pre-operative computed tomography (CT) or magnetic resonance (MR) image coordinate system [1].

This work was supported partly by the National Natural Science Foundation of China under Grant 61971367 and the Natural Science Foundation of Fujian Province of China under Grant 2020J01004.

Current 3D endoscope tracking techniques use either external devices or video-based methods to navigate the endoscope. The former – particularly electromagnetic tracking [2, 3] – limits itself to inherent system errors and patient movements. The latter employs 2D-3D registration to fuse preoperative and intraoperative images to determine the endoscope location [4, 5]. While Shen et al. [6] recognized the bifurcations to locate the bronchoscope, more recently they introduce a context-aware depth estimation method for bronchoscope 3D tracking and navigation [7]. Compared to external tracking devices, video-based endoscope tracking has several advantages, e.g., pure software-driven methods, cost-effective, simple and easy settings in the operating room, without additional calibration for hardware devices, and robust to tissue deformation. Unfortunately, video-based 3D tracking methods suffer from image uncertainties or artifacts (e.g., blurring, illumination variations, mucus or bubbles), which easily collapse the 2D-3D fusion procedure to fail endoscope tracking since current similarity functions remain challenging to precisely characterize these image artifacts during optimization.

This work aims to address the problem of image artifacts and boost the 2D-3D fusion procedure to enhance video-based 3D endoscopic camera tracking. The contribution of this paper is twofold. We develop a multiscale selective fusion framework with an accurate similarity function to precisely describe image difference in 2D-3D registration methods for surgical tracking and and navigation. Such a multiscale selective similarity function is robust and effective to image deformation and artifacts. On the other hand, we also contribute to image registration since our multiscale selective similarity function is applicable to multimodal image registration or segmentation, e.g., mediastinal or abdominal lymph node detection. Moreover, we apply our proposed similarity function to perform 2D-3D fusion for endoscope 3D tracking. The multiscale selective fusion driven 3D tracking method outperforms current available video-based approaches for robotic and navigated endoscopic procedures.

2. MULTISCALE SELECTIVE SIMILARITY

This section defines the multiscale selective similarity function for 2D-3D fusion or registration. Basically, the computation of the multiscale selective similarity consists of three

steps: (1) Gaussian pyramids for multiscale images, (2) specific shape selection from images, and (3) mathematical formulation of the intensity similarity. We then establish a multiscale selective similarity driven fusion framework with deterministic optimization for robust robotic endoscope 3D tracking, precisely locating the endoscope during surgery.

Multiscale representation is widely introduced for image processing. Several multiscale methods are used in the literature, e.g., Gaussian pyramid, Laplacian pyramid, and wavelet methods [8]. We employ Gaussian pyramids to process for endoscopic video images since they are simply implemented and keep color information without distortion.

The Gaussian pyramid is a collection of successively lowpass-filtered images. Let $\mathbf{I}_k(x, y)$ (which equals $\mathbf{G}_k^0(x, y)$) with pixel (x, y) be the k-th endoscopic video image. The Gaussian pyramid of $\mathbf{I}_k(x, y)$ is recursively constructed by:

$$\mathbf{G}_k^{i+1}(x, y) = \sum_{u=-2}^{u=2} \sum_{v=-2}^{v=2} \alpha(u, v)\mathbf{G}_k^i(2x + u, 2y + v), \quad (1)$$

where $\alpha(u, v)$ is the Gaussian kernel with the separable and symmetric properties to weight each pixel, pyramid level $i = 0, 1, \cdots, L - 1$, and total L levels are obtained.

2.1. Specific Shape Selection

The human visual system is sensitive to specific shape or structural information on video images, e.g., fold and bifurcation or hole shapes. This implies that the similarity computation can discriminate the specific region or shape information from insensitive information on endoscopic images. Therefore, we first extract the shape regions.

For an input endoscopic video image with the size of width W and height H, we separate it into $A \cdot B$ regions (Fig. 1(a)). A region or patch $R_{a,b}$ ($a \in A$, $b \in B$) with $w \cdot h$ pixels is represented by the following formulation

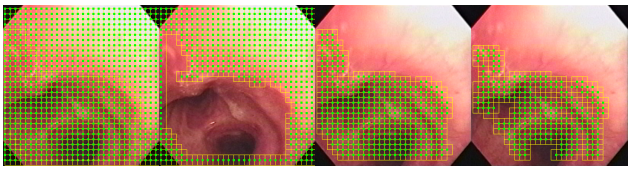
$$R_{a,b} = (o_x, o_y), w = W/A, h = H/B, \quad (2)$$

where (o_x, o_y) is the center of region $R_{a,b}$. To detect shape on the region, variance $\beta_{a,b}$ is calculated by

$$\beta_{a,b}^2 = \frac{1}{M} \sum_{R_{a,b}} \left(R_{a,b}(x, y) - \bar{R}_{a,b}\right)^2, \quad (3)$$

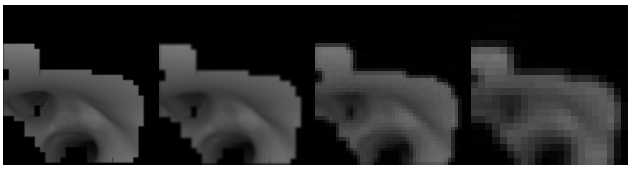
where M and $\bar{R}_{a,b}$ mean the number of pixels and the average intensity in the region $R_{a,b}$. The contrast $\chi_{a,b}$ for shape detection depends on the function $\delta(R_{a,b}(x, y))$ related to color saturation and lightness at a pixel location (x, y).

The regions without specific shape information (Fig. 1(b)) were discarded by $\chi_{a,b} \geq \eta$ (a predefined factor) since they are insensitive to the human visual system. The remained regions with the shape information (Fig. 1(c)) were sorted in descending order on the basis of value $\beta_{a,b}$. Finally, we select N shape regions $\mathcal{R} = \{R_n\}_{n=1}^N$ from the sorted regions for the similarity characterization. Fig. 2 illustrates the selected shape regions of the multiscale images.



(a) Separated (b) Discarded (c) Remained (d) Selected

Fig. 1: Specific shape selection: A *yellow* square denotes a region and *green* dots mean the centers of all the regions.



(a) Level 1 (b) Level 2 (c) Level 3 (d) Level 4

Fig. 2: Selective specific shape regions on multiscale images

2.2. Similarity Formulation

After selecting these shape regions $\mathcal{R}$, we propose a function to mathematically measure the similarity. Inspired by the work in [9], we define a cost function $\mathcal{F}(R_n)$ as

$$\mathcal{F}(R_n^r, R_n^d) = \underbrace{\mathcal{L}^\kappa(R_n^r, R_n^d)}_{Luminance} \cdot \underbrace{\mathcal{C}^\lambda(R_n^r, R_n^d)}_{Contrast} \cdot \underbrace{\mathcal{S}^\mu(R_n^r, R_n^d)}_{Structure}, \quad (4)$$

where parameters κ, λ, and μ weigh the importance of three elements of luminance $\mathcal{L}(\,,\,)$, contrast $\mathcal{C}(\,,\,)$, and structure $\mathcal{S}(\,,\,)$. For selected regions R_n^r and R_n^d (where r and d denote the regions from the reference and distorted images, respectively), $\mathcal{S}(R_n^r, R_n^d)$, $\mathcal{L}(R_n^r, R_n^d)$ and $\mathcal{C}(R_n^r, R_n^d)$ are

$$\mathcal{S}(R_n^r, R_n^d) = \frac{\sigma_{r,d} + C_1}{\sigma_r \sigma_d + C_1}, \quad (5)$$

$$\mathcal{L}(R_n^r, R_n^d) = \frac{2\xi_r \xi_d + C_2}{\xi_r^2 + \xi_d^2 + C_2}, \quad (6)$$

$$\mathcal{C}(R_n^r, R_n^d) = \frac{2\sigma_r \sigma_d + C_3}{\sigma_r^2 + \sigma_d^2 + C_3}, \quad (7)$$

where ξ_r and ξ_d are the average intensities in regions R_n^r and R_n^d that were detected from reference and distorted images, σ_r and σ_d are the variance, $\sigma_{r,d}$ is the correlation, and three constants, C_1, C_2, and C_3 are used to avoid the unstable similarity computation. The function $\mathcal{F}(R_n^r, R_n^d)$ with three elements were proved to be very effective for describing the similarity under image uncertainties [9]. Based on the multiscale endoscopic images generated by the Gaussian pyramid and the similarity function $\mathcal{F}(R_n^r, R_n^d)$, a multiscale similarity characterization function can be defined as

$$\hat{\mathcal{F}}(R_n^r, R_n^d) = \mathcal{X} \cdot \prod_{i=0}^{L-1} \mathcal{C}^{\lambda_i}(R_n^r, R_n^d) \cdot \mathcal{S}^{\mu_i}(R_n^r, R_n^d), \quad (8)$$

$$\mathcal{X} = \mathcal{L}_*^{\varphi_{L-1}}(R_n^r, R_n^d), \tag{9}$$

where φ_{L-1} is the weight at level $L-1$ and $\mathcal{L}_*(R_n^r, R_n^d)$ is

$$\mathcal{L}_*(R_n^r, R_n^d) = \arg \max_{i=0,\cdots,L-1} \mathcal{L}_i(R_n^r, R_n^d). \tag{10}$$

We simply determine $\varphi_i = \lambda_i = \mu_i$ and $\sum_{i=0}^{L-1} \mu_i = 1$.

The multiscale similarity characterization function is formulated (Eq. 8) as it involves image luminance, contrast, and structure information from specific shape regions that were detected and selected from these multiscale images. Therefore, such a similarity function is robust and insensitive to image artifices or uncertainties to enhance the fusion procedure. We apply this similarity to 2D-3D fusion to boost video-based 3D endoscope tracking described in the following.

3. APPLICATION TO 3D ENDOSCOPE TRACKING

Video-based 3D endoscopic camera tracking requires to fuse 2D endoscopic video images with preoperative 3D CT data, i.e., to compute the spatial transformation from the endoscopic camera to the CT image coordinate system.

Let $\mathbf{Q}_k$ be the transformation including 3D camera position and orientation at frame k. To estimate the endoscopic camera pose with position and orientation information in spatial transformation $\mathbf{Q}_k$, the 2D-3D registration or fusion procedure is formulated as an optimization process w.r.t. the similarity function to compute the difference between two images

$$\tilde{\mathbf{Q}}_k = \arg\max_{\mathbf{Q}_k} \mathcal{F}\left(\mathbf{I}_k, \mathbf{I}_v(\mathbf{Q}_k)\right), \tag{11}$$

where $\mathbf{I}_k$ denotes the endoscopic video image at frame k and $\mathbf{I}_v(\mathbf{Q}_k)$ indicates the 2D virtual rendering image that was generated from the camera pose parameters $\mathbf{Q}_k$ using volume rendering techniques. According to the proposed multiscale selective similarity characterization function (Eq. 8), the objective function $\mathcal{F}(\mathbf{I}_k, \mathbf{I}_v(\mathbf{Q}_k))$ can be rewritten

$$\mathcal{F}\left(\mathbf{I}_k, \mathbf{I}_v(\mathbf{Q}_k)\right) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_*^{\varphi_{L-1}}(R_n^k, R_n^v) \cdot \mathcal{Y}, \tag{12}$$

$$\mathcal{Y} = \prod_{i=0}^{L-1} \mathcal{C}^{\lambda_i}(R_n^k, R_n^v) \cdot \mathcal{S}^{\mu_i}(R_n^k, R_n^v), \tag{13}$$

where N is the number of specific shape regions R_n^k and R_n^v that were extracted and selected from video image $\mathbf{I}_k$ at frame k and 2D virtual rendering image $\mathbf{I}_v(\mathbf{Q}_k)$, respectively. Before running the optimizer, we need to initialize transformation $\mathbf{Q}_k$. It is possible to initialize $\mathbf{Q}_k = \tilde{\mathbf{Q}}_{k-1}$, i.e., directly use previous pose estimate $\tilde{\mathbf{Q}}_{k-1}$. However, such an initialization loses temporal motion $\Delta\mathbf{Q}_k$ between two continuous images $k-1$ and k. To compensate this loss, we use history estimates $\tilde{\mathbf{Q}}_{k-3}$, $\tilde{\mathbf{Q}}_{k-2}$, and $\tilde{\mathbf{Q}}_{k-1}$ to initialize $\mathbf{Q}_k$

$$\mathbf{Q}_k = \frac{\tilde{\mathbf{Q}}_{k-1}(\tilde{\mathbf{Q}}_{k-2})^\tau + \tilde{\mathbf{Q}}_{k-2}(\tilde{\mathbf{Q}}_{k-3})^\tau}{2} \tilde{\mathbf{Q}}_{k-1}, \tag{14}$$

where symbol τ means to inverse the transformation.

4. EXPERIMENTAL SETTINGS

We evaluate our method on five cases of clinical datasets with videos and their CT volumes. We manually generated ground truth for these datasets and compute the tracking accuracy of different methods. In the specific shape selection, parameters and thresholds are $A = B = 30$, $\eta = 0.9$, and N equals the number of the 30% sorted regions. We compare three methods: (1) Merritt et al. [5]: using the similarity measure of normalized sum squared difference, (2) Shen et al. [7]: a context-aware depth estimation based tracking method, (3) our method, as discussed in Sections 2 and 3. For the multiscale similarity, coefficients φ_i, λ_i, μ_i are determined by normalizing the tracking accuracy at each scale itself. We also performed the single scale selective similarity function (Eq. 4) for video-based 3D tracking and predefine these coefficients as $\varphi = \lambda = \mu = 1$ for the singel-scale case.

5. RESULTS AND DISCUSSION

Table 1 summarizes the tracking accuracy from five cases of clinical experiments. The proposed method significantly improved the tracking accuracy from 8.9 to 5.4 mm, while the orientation error was reduced from 39.3° to 20.9°. Fig. 3 visually compares the tracking results that were used to generate 2D virtual rendering images whether they resemble endoscopic video images. Note that these 2D virtual rendering images were generated by the navigated or tracked endoscope position and orientation using volume rendering techniques. Fig. 4 shows examples of comparing the tracking accuracy of using different methods tested on Case 3. While Fig. 5 illustrates the various single scale based similarity function used for the 2D-3D fusion or registration procedure, Fig. 6 further compares the results of the fusion of using different signle scales. Additionally, we used graphic processing unit techniques to accelerate our method up to 32 frames per second.

The goal of this work is to address the accuracy and robustness of 2D-3D registration that fuses multimodal medical information for 3D endoscope tracking and navigation. The similarity metric plays an essential role in image registration. To this end, we propose a framework that calculate the sim-

Table 1: Comparison of 3D endoscopic camera tracking errors of using different methods. The units of (position, orientation) are millimeter and degree, respectively.

Cases	Merritt et al. [5]	Shen et al. [7]	Ours
1	21.4, 30.9	10.9, 36.8	6.50, 25.6
2	5.30, 24.4	12.4, 60.9	3.80, 16.6
3	16.4, 120.2	9.70, 57.5	2.60, 13.8
4	21.9, 122.3	5.80, 19.7	5.10, 18.4
5	17.0, 53.8	5.60, 21.4	9.20, 30.3
Average	16.4, 70.3	8.9, 39.3	5.40, 20.9

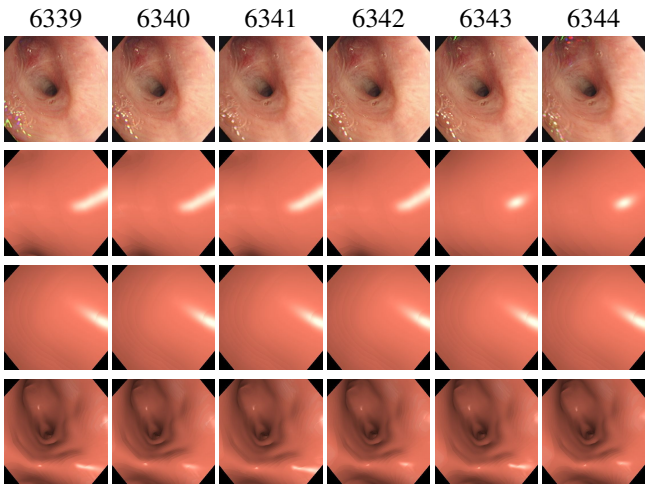


Fig. 3: Visual comparison of the tracking results including endoscope positions and orientations used to generate virtual images whether resemble video ones. The first row shows the frame numbers of a small sequence of continuous video images and the second row their corresponding video images with the observed artifacts (i.e., bubbles and reflection). The virtual images in Rows 3∼5 correspond to the tracking results from Merritt et al. [5], Shen et al. [7], and our method, respectively. Our approach works better than other methods.

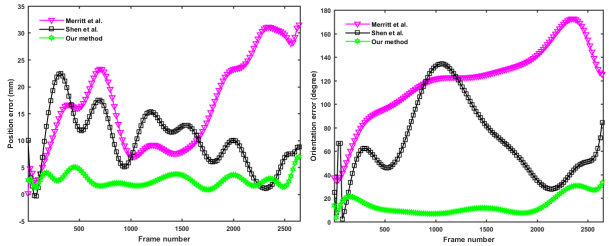


Fig. 4: 3D endoscope tracking position and orientation accuracy of using Merritt et al. [5], Shen et al. [7], and our method

ilarity between target and reference images on the basis of specific structural information. This framework follows the human visual system that is usually sensitive to typical structural information on images, while it also takes luminance and contrast changes into consideration (Eq. 12). Moreover, multiscale representation is powerful to describe image information. These properties of our proposed strategy are robust to image uncertainties (e.g., bubbles in Fig. 3). Hence, we contribute the improvement of video-based endoscope tracking to the multiscale selective similarity.

Although our method works well, it still involves limitations. Our method possibly gets trapped in image artifacts such as uninformative images failing to detect the specific region detection. On the other hand, we used a deterministic optimization method to resolve the registration equation. Such an optimizer depends critically on its initial guess. It

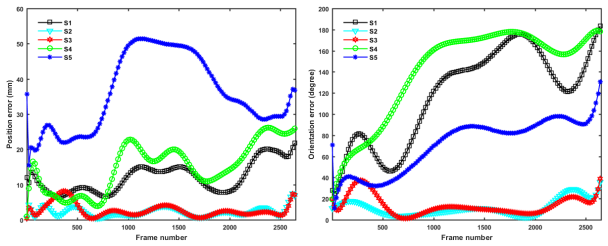


Fig. 5: Illustration of the fusion accuracy of using the five single-scale selective similarity functions (S1∼S5) (Case 3)

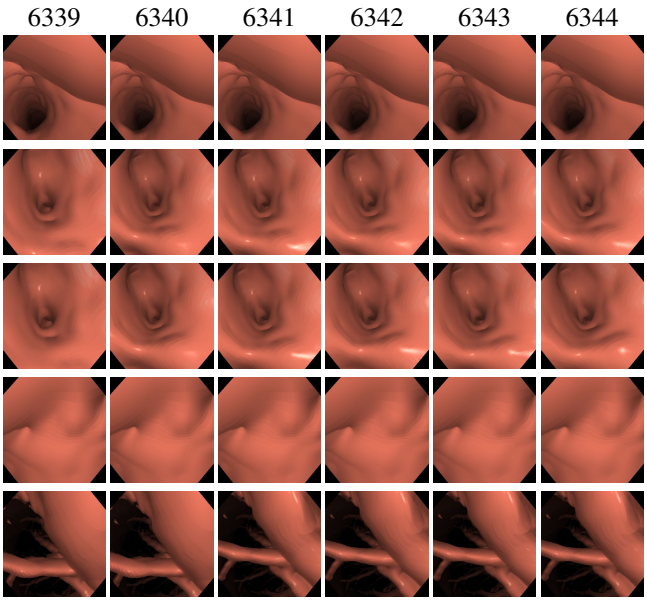


Fig. 6: Comparison of fusion in different scales: Row 1 shows the frame numbers of real images (Fig. 3) and Rows 2∼6 different single scale-based tracking results from scales 1 to 5.

commonly uses the estimated endoscope pose at the previous frame as the initialization of the current frame. Unfortunately, we cannot guarantee that the previous pose estimate is correct. This also fails to 3D tracking. In addition, tissue deformation such as respiratory motion is also a challenge to CT-video fusion based 3D endoscopic camera tracking and navigation.

6. CONCLUSION

This work proposes an accurate multiscale selective similarity characterization to boost 2D-3D registration for endoscope 3D tracking in robotic surgery. With its application to video-based endoscope 3D tracking, the experimental results demonstrate that our proposed strategy outperforms other methods. In particular, the 3D tracking accuracy was improved from 8.9 to 5.4 mm. Future work includes removal of endoscopic uninformative images, improvement of initialization in optimization, and tissue deformation modeling.

7. REFERENCES

[1] X. Luo, K. Mori, and T. Peters, "Advanced endoscopic navigation: Surgical big data, methodology, and applications," *Annual Review of Biomedical Engineering*, vol. 20, pp. 221–251, 2018.

[2] H. Sorger, E. Hofstad, T. Amundsen, T. Lango, and H. Leira, "A novel platform for electromagnetic navigated ultrasound bronchoscopy (ebus)," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 8, pp. 1431–1443, 2016.

[3] H. Stephen, H. Jaeger, R. Burke, B. O'Sullivan, J. Keane, F. Trauzettel, B. Marques, S. Cotin, B. Brian, H. Leira, E. Hofstad, O. Solberg, T. Lango, and P. Cantillon-Murphy, "An open electromagnetic tracking framework applied to targeted liver tumour ablation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, pp. 1475–1484, 2019.

[4] X. Luo, M. Feuerstein, D. Deguchi, T. Kitasaka, H. Takabatake, and K. Mori, "Development and comparison of new hybrid motion tracking for bronchoscopic navigation," *Medical Image Analysis*, vol. 16, no. 3, pp. 577–596, 2012.

[5] S. A. Merritt, R. Khare, and W. Higgins, "Interactive ct-video registration for the continuous guidance of bronchoscopy," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1376–1396, 2013.

[6] M. Shen, B. Giannarou, P. Shah, and G.-Z. Yang, "BRANCH: Bifurcation recognition for airway navigation based on structural characteristics," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2017, pp. 182–189.

[7] M. Shen, Y. Gu, N. Liu, and G.-Z. Yang, "Context-aware depth and pose estimation for bronchoscopic navigation," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 732–739, 2019.

[8] R. Szeliski, *Computer Vision: Algorithms and Applications*, Springer-Verlag London, 2011.

[9] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: new challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2017.