



Stereo Matching for Wireless Capsule Endoscopy Using Direct Attenuation Model

Min-Gyu Park^(✉), Ju Hong Yoon, and Youngbae Hwang

Intelligent Image Processing Research Center, Korea Electronics Technology Institute
(KETI), Gwangju, South Korea
{mpark, jhyoon, ybhwang}@keti.re.kr

Abstract. We propose a robust approach to estimate depth maps designed for stereo camera-based wireless capsule endoscopy. Since there is no external light source except ones attached to the capsule, we employ the direct attenuation model to estimate a depth map up to a scale factor. Afterward, we estimate the scale factor by using sparse feature correspondences. Finally, the estimated depth map is used to guide stereo matching to recover the detailed structure of the captured scene. We experimentally verify the proposed method with various images captured by stereo-type endoscopic capsules in the gastrointestinal tract.

1 Introduction

The wireless capsule endoscope (WCE) is a powerful device to acquire images of the gastrointestinal (GI) tract for screening, diagnostic, and therapeutic endoscopic procedures [1]. Especially, the WCE captures the images of the small intestine where current wired endoscopic devices cannot reach. In this paper, we introduce a method to recover the 3D structure from stereo images captured by a stereo-type WCE, shown in Fig. 1.

To perceive depth from endoscopic images, many researchers have brought various computer vision techniques such as stereo matching [4], shape-from-shading (SfS) [2, 13], shape-from-focus (SfF) [11], and shape-from-motion (SfM) [3]. Ciuti et al. [2] adopted the SfS technique because the position of light sources are known and shading is an important cue in the endoscopic images. Visentini et al. [13] fused the SfS cue and image feature correspondences to estimate accurate dense disparity maps. Takeshita et al. [11] introduced an endoscopic device that estimates depth by using the SfF technique, which utilizes multiple images captured with different focus settings at the same camera position. Fan et al. [3] established sparse feature correspondences between consequent images, and then, they calculated camera poses and the 3D structure of the scene by using the SfM technique. They generated 3D meshes through Delaunay triangulation by using triangulated feature points.

Stereo matching is also a well-known technique to estimate a depth map from images, which can be divided into active and passive [9] approaches. We refer



Fig. 1. Stereo-type wireless endoscopic capsule, wireless receiver, and captured images in the stomach and the small bowel, from the left.

structured light-based stereo matching [10] to the active approach which projects a visible or IR pattern to the scene to leverage correspondence searching between the images. However, the active approach is not suitable for wireless endoscopy mainly because of the limited resources, e.g., battery capacity and the size of a capsule. Therefore, previous studies [4] focused on minimally invasive surgery rather than WCE-based GI examination. For the same reason, most commercial wireless endoscopic capsules typically adopt conventional passive image sensors.

To the best of our knowledge, commercially available WCE products are not capable of estimating depth information. This is the first attempt to estimate the geometric structure of the scene inside the GI tract captured by a WCE. To achieve this goal, we designed a stereo-type WCE as shown in Fig. 1 without enlarging the diameter of the capsule. This sensor can capture about 0.12 million images for the entire GI tract as described in Fig. 1 ranging from the stomach to the large bowel. Having captured stereo images in one hand, we estimate a fully dense depth map by using the direct attenuation model. Since there is no external light source except ones attached to the capsule, farther objects look darker than nearer one in the captured image. Therefore, we consider the attenuation trend of the light to estimate depth maps assuming that the medium inside the GI tract is homogeneous. We firstly employ the direct attenuation model to compute an up-to-scale depth map, and then, solve the scale ambiguity by using sparse feature correspondences. Afterward, we utilize the rescaled depth map to guide a popularly used algorithm, i.e., semi-global matching (SGM) [6]. The detailed description of the proposed method is given in the following section.

2 Proposed Method

2.1 Capsule Specification

Our wireless endoscopic capsule consists of two cameras, four led lights, a wireless transmitter, and the battery. Two cameras are displaced about 4 mm, the viewing angle is 170° , and the resolution of captured images is 320×320 . The capsule captures three pairs of images per second. In total, it captures more than 115,000 images for eight hours in the GI tract. Four led lights are attached around the cameras as shown in Fig. 1. The lights are synchronized with the cameras to

minimize the battery usage. Captured images are transmitted to the receiver because the capsule does not have an internal storage. The length of the capsule is 24 mm and the diameter is 11 mm.

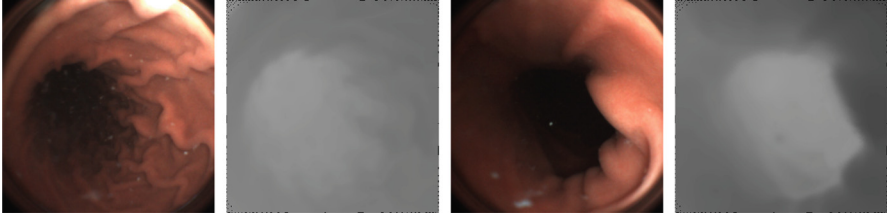


Fig. 2. Sample input images and the depth maps computed by Eq. (5). Here, bright pixels indicates they are farther than dark ones.

2.2 Depth Estimation with the Direct Attenuation Model

Since the captured image has poor visibility, the image for each pixel \mathbf{p} can be modeled as [5]

$$I(\mathbf{p}) = J(\mathbf{p})t(\mathbf{p}) + A(1 - t(\mathbf{p})), \quad (1)$$

where J is the scene radiance, I is the observed intensity, t is the transmission map, and A is the atmospheric light. Since there is no source of natural illumination such as sunlight, A can be dropped from Eq. (1). Then, t can be defined as

$$t(\mathbf{p}) = I(\mathbf{p})/J(\mathbf{p}). \quad (2)$$

The transmission map also can be defined by Bouguer’s exponential law of attenuation [8],

$$t(\mathbf{p}) = \exp(-\beta(\mathbf{p})d(\mathbf{p})), \quad (3)$$

where an attenuation coefficient $\beta(\mathbf{p})$ is typically represented by sum of absorption and scattering coefficients, $\beta(\mathbf{p}) = \beta_{\text{absorption}}(\mathbf{p}) + \beta_{\text{scatter}}(\mathbf{p})$. By combining Eqs. (2) and (3), the depth of a pixel \mathbf{p} can be estimated as

$$d(\mathbf{p}) = \frac{\ln(J(\mathbf{p})) - \ln(I(\mathbf{p}))}{\beta(\mathbf{p})} \approx \frac{\ln(\bar{I}) - \ln(I(\mathbf{p}))}{\beta}. \quad (4)$$

To simplify Eq.(4), we approximate two terms $J(\mathbf{p})$ and $\beta(\mathbf{p})$ by considering characteristics of the GI tract. First, assuming that the GI tract is filled with a homogeneous matter such as water, the attenuation coefficient $\beta(\mathbf{p})$ is approximated as a constant value for all pixels, $\beta \approx \beta(\mathbf{p})$. Second, we also approximate the scene radiance as the mean of all pixel values as $J(\mathbf{p}) \approx \bar{I}$ based on the assumption that most pixels have a similar color in a local GI region. Based on the second assumption, we easily obtain the depth map up to a scale factor β ,

$$d_{\beta}(\mathbf{p}) = \beta d(\mathbf{p}) = \ln(\bar{I}) - \ln(I(\mathbf{p})). \quad (5)$$

Here, the depth map $d_\beta(\mathbf{p})$ indicates a depth map up to scale factor β . In the following section, we estimate β . Beforehand, we apply a noise removal filter to smooth $d_\beta(\mathbf{p})$ by using a well-known bilateral filter [12].

2.3 Resolving the Scale Ambiguity of d_β

To resolve the scale ambiguity of $d_\beta(\mathbf{p})$, we compute β from by using sparse feature correspondences. First, we detect and match corner points. Then, we compute the depth of \mathbf{p} , $d_s(\mathbf{p})$,

$$d_s(\mathbf{p}) = \frac{fB}{|p_x^L - p_x^R|}, \quad (6)$$

where p_x^L and p_x^R are the positions of matched points from the left and right images along the x-axis, f is the focal length of the left camera, B is the baseline between two cameras. Since each corner point has corresponding $d_\beta(\mathbf{p})$, β can be computed by

$$\beta = d_\beta(\mathbf{p})/d_s(\mathbf{p}). \quad (7)$$

Assuming that β is constant for all pixels, we find an optimal β , β^* , that maximizes the number of inlier points whose error is smaller than a threshold value, τ_c .

$$\begin{aligned} \beta^* &= \arg \max_{\beta \in \mathcal{B}} \sum_{\mathbf{p} \in \mathcal{S}} T(\mathbf{p}, \beta, \tau_c), \\ T(\mathbf{p}, \beta, \tau_c) &= \begin{cases} 1 & \text{if } |d_s(\mathbf{p}) - d_\beta(\mathbf{p})/\beta| \leq \tau_c \\ 0 & \text{otherwise.} \end{cases}, \end{aligned} \quad (8)$$

where \mathcal{B} is the set of β values computed from all feature correspondences and \mathcal{S} is the set of correspondences' positions in the image coordinate. The function T gives 1, if the discrepancy between $d_s(\mathbf{p})$ and rescaled $d_\beta(\mathbf{p})$ is small, and 0, otherwise. Therefore, the estimated β^* minimizes the gap between d_s and d_β/β . We thus rescale $d_\beta(\mathbf{p})$ and compute its corresponding disparity map as

$$\bar{d}_\beta(\mathbf{p}) = \frac{d_\beta(\mathbf{p})}{\beta^*}, \quad \bar{D}_\beta(\mathbf{p}) = \frac{fB}{\bar{d}_\beta(\mathbf{p})}. \quad (9)$$

We utilize the rescaled disparity map $\bar{D}_\beta(\mathbf{p})$ to leverage stereo matching.

2.4 Robust Stereo Matching Using a Guidance Depth Map

We slightly modify the SGM algorithm [6] to compute the disparity map $D(\mathbf{p})$ which minimizes the following energy function,

$$\begin{aligned} E(d) &= \sum_{\mathbf{p}} (\phi(\mathbf{p}, D(\mathbf{p})) + \psi(\mathbf{p}, D(\mathbf{p}))) \\ &+ \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_1 T[|D(\mathbf{p}) - D(\mathbf{q})| = 1] + \sum_{\mathbf{q} \in N_{\mathbf{p}}} P_2 T[|D(\mathbf{p}) - D(\mathbf{q})| > 1]. \end{aligned} \quad (10)$$

In the first term, the function $\phi(\cdot, \cdot)$ is the pixel-wise matching cost, computed by using Census-based hamming distance and absolute difference of intensities (AD-CENSUS). The function $\psi(\cdot, \cdot)$ is also the pixel-wise matching cost computed by using $\bar{D}_\beta(\mathbf{p})$,

$$\psi(\mathbf{p}, D(\mathbf{p})) = \begin{cases} |\bar{D}_\beta(\mathbf{p}) - D(\mathbf{p})| & \text{if } |\bar{D}_\beta(\mathbf{p}) - D(\mathbf{p})| \leq \tau_{\text{err}} \\ c & \text{otherwise} \end{cases}. \quad (11)$$

The second term gives the penalty P_1 for the pixels having small disparity differences with the neighboring pixels $\mathbf{q} \in N_{\mathbf{p}}$, i.e., $T[|D(\mathbf{p}) - D(\mathbf{q})| = 1]$ gives 1 when the difference of disparity values is 1. Similarly, the third term gives the large penalty P_2 such that $P_2 > P_1$ for the pixels having disparity differences greater than 1 with the neighboring pixels. We minimize Eq. 10 by using the SGM method [6]. As a post-processing step, we apply the weighted median filter [7]. Finally, we obtain the depth map from the disparity map by $d(\mathbf{p}) = fB/D(\mathbf{p})$.

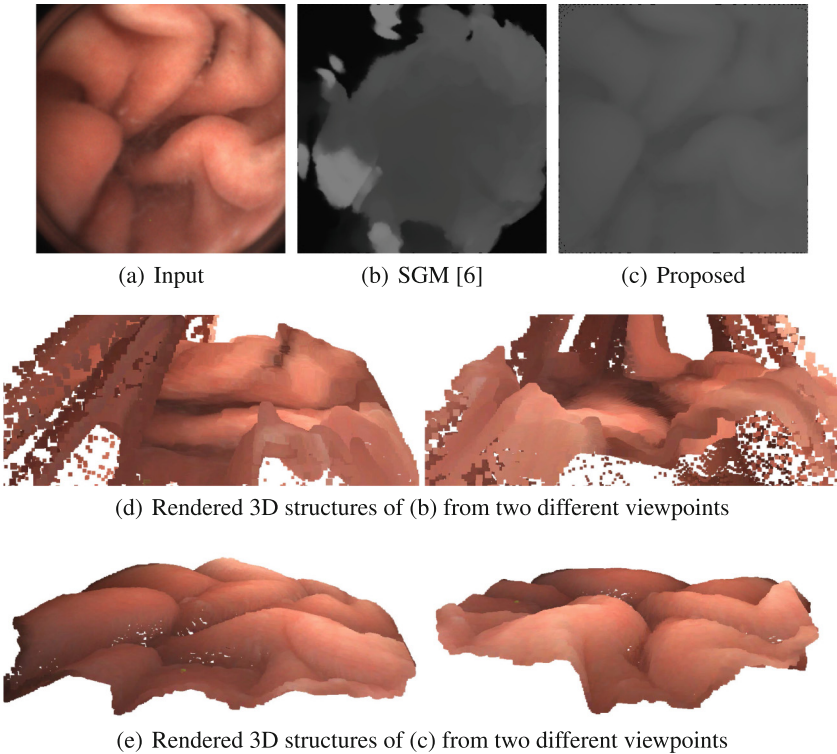


Fig. 3. Comparison of disparity maps and reconstructed 3D structures.

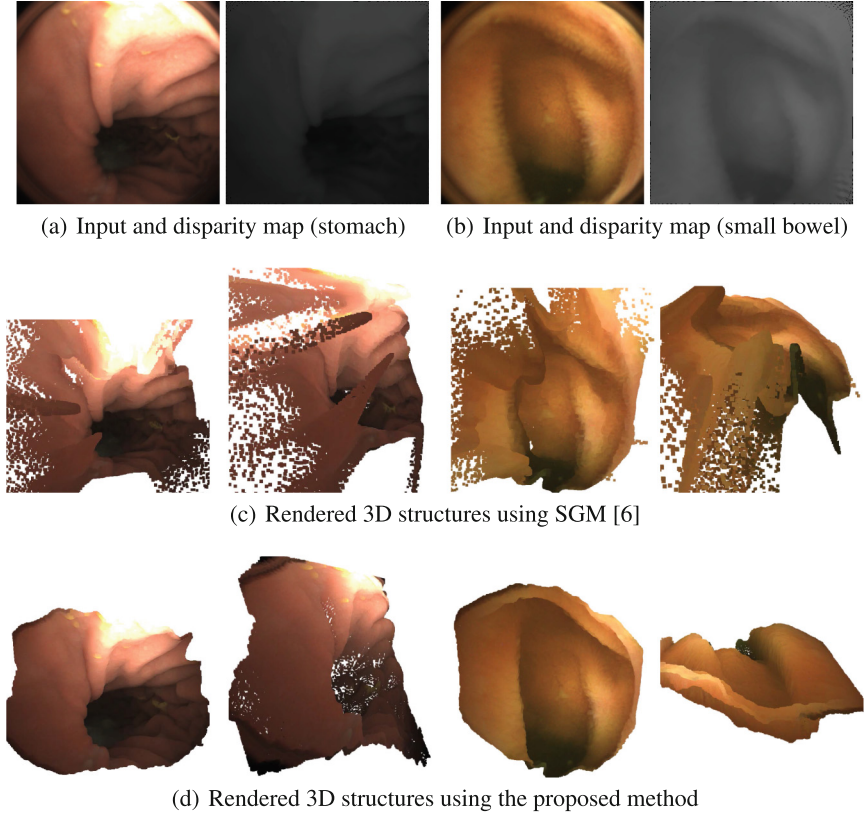


Fig. 4. Comparison reconstructed 3D structures.

3 Experimental Results

Computing depth maps from endoscopic images is difficult, mainly because of difficulties caused by the characteristics of the GI tract and the limited resources of the capsule. The difficulties are summarized as (1) inaccurate matching due to lens distortion, (2) low resolution image, (3) image noise caused by the lack of illumination.

Note that the proposed method without the direct attenuation model and its cost function terms in Eqs. (10) and (11) is identical to the popularly used stereo matching algorithm, SGM [6]. Therefore, we qualitatively compare results of the proposed method with the conventional SGM to demonstrate the advantages of the proposed method under the aforementioned difficulties. To achieve fair comparison and to explicitly demonstrate the advantages of the proposed method, we used the same similarity measure and parameters for the SGM and the proposed method.

In the first experiment, we used the images of a stomach and a small bowel captured by our stereo-type WCE. In the second experiment, we used the large bowel Phantom model¹ not only to capture endoscopic images but also to compare the actual size of an object with the estimated size.

Qualitative Evaluation: Most importantly, the proposed method acquires fully dense depth maps whereas the conventional approach fails at non-overlapping regions because pixels of the left image in non-overlapping regions do not have corresponding pixels of the right image. Moreover, since we used a large field of view cameras, the proportion of non-overlapping regions is about 30~50% depending on the distance of the captured scene from the camera. Computed disparity values in non-overlapping regions are noisy as shown in Fig. 3(b) where noisy disparity values are exceptionally brighter than others in the disparity map although they should show similar disparity as seen in the scene structure of Fig. 3(a). The noisy disparity values become more conspicuous when they are represented in the 3D space as shown in Fig. 3(d), and those noisy disparity values seem to float the space so that they obstruct to see the underlying 3D structure. Differently, the proposed method accurately recovers the 3D structure of the scene as shown in Fig. 3(e) because the proposed cost function with the direct attenuation model well suppresses uncertainties caused by radial distortion and low light noise. In addition, the depth map based on the direct attenuation model effectively enforces the proposed cost function to reconstruct the depth in non-overlapping regions as shown in Fig. 3(c).

As discussed in the introduction, the main advantage of the WCE is that it can capture not only stomach images but also images of small bowel where typical wired endoscopic devices cannot reach. Similar to the results demonstrated in Fig. 3, the proposed method reconstruct 3D structures of the local stomach and small bowel regions more robustly than the SGM as shown in Figs. 4(c) and (d), and effectively estimates dense depth maps in non-overlapping regions as shown in Figs. 4(a) and (b).

Application for Diagnosis: We also show an application of the proposed method for diagnosis. Using estimated depth information, we estimate the size of an object of interest by clicking two points from the image as shown in Figs. 5(a) and (c). In this experiment, we used the large bowel Phantom model and two different types of polyps whose size is known. As shown in Figs. 5, the estimated size is quite similar to the actual size in which the error was at most 0.5 mm. This procedure has long relied on the experienced doctor or endoscopist.

Parameter Settings and Running Time: We used 7×9 window to compute census-based matching cost computation, and set P1 and P2 to 11 and 19, respectively. The average running time of the proposed method was about 10ms, implemented on a modern GPU, GTX Titan Xp.

¹ https://www.buyamag.com/digestive_system_models.php.

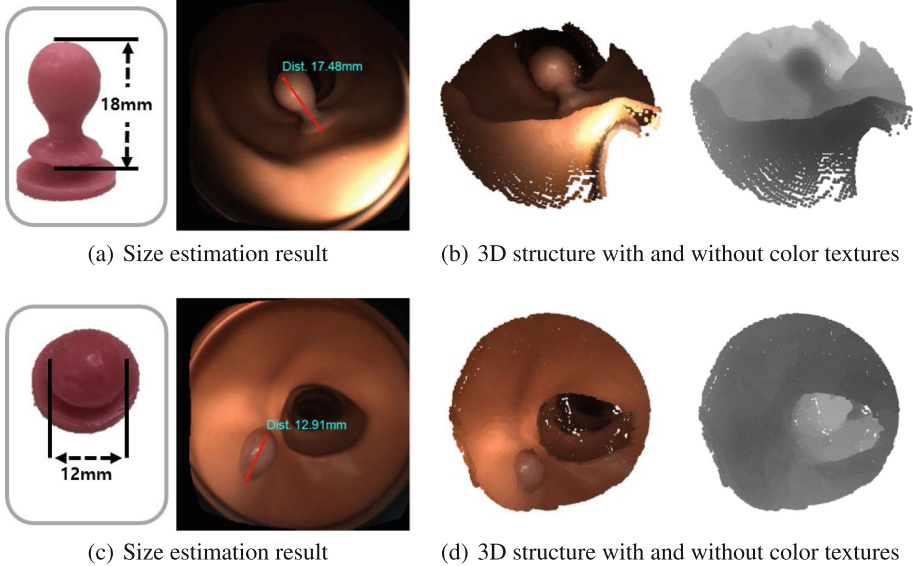


Fig. 5. Size estimation results with a large bowel Phantom model

4 Conclusion

We have proposed a stereo matching algorithm designed for a stereo-type wireless capsule endoscopy. We obtained an up-to-scale depth map by using the direct attenuation model because of the light source around the capsule in the completely dark environment. Thereafter, we employed the up-to-scale depth map to guide conventional stereo matching algorithms after resolving the scale ambiguity. Through the experiments, we observed that the proposed method can estimate depth maps accurately and robustly in the GI tract.

Acknowledgment. This work was supported by ‘The Cross-Ministry Giga KOREA Project’ grant funded by the Korea government(MSIT) (No. K18P0200, Development of 4D reconstruction and dynamic deformable action model based hyper-realistic service technology) and a gift from Intramedic.

References

1. Ciuti, G., Menciassi, A., Dario, P.: Capsule endoscopy: from current achievements to open challenges. *IEEE Rev. Biomed. Eng.* **4**, 59–72 (2011)
2. Ciuti, G., Visentini-Scarzanella, M., Dore, A., Menciassi, A., Dario, P., Yang, G.Z.: Intra-operative monocular 3d reconstruction for image-guided navigation in active locomotion capsule endoscopy. In: 2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), pp. 768–774, June 2012

3. Fan, Y., Meng, M.Q.H., Li, B.: 3d reconstruction of wireless capsule endoscopy images. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 5149–5152, August 2010
4. Furukawa, R., Sanomura, Y., Tanaka, S., Yoshida, S., Sagawa, R., Visentini-Scarzanella, M., Kawasaki, H.: 3d endoscope system using doe projector. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2091–2094, August 2016
5. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(12), 2341–2353 (2011). <https://doi.org/10.1109/TPAMI.2010.168>
6. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(2), 328–341 (2008)
7. Ma, Z., He, K., Wei, Y., Sun, J., Wu, E.: Constant time weighted median filtering for stereo matching and beyond. In: IEEE International Conference on Computer Vision (ICCV) (2013)
8. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. *Int. J. Comput. Vis.* **48**(3), 233–254 (2002). <https://doi.org/10.1023/A:1016328200723>
9. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* **47**, 7–42 (2002)
10. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2003, pp. 195–202. IEEE Computer Society, Washington, DC, USA (2003). <http://dl.acm.org/citation.cfm?id=1965841.1965865>
11. Takeshita, T., Kim, M., Nakajima, Y.: 3-d shape measurement endoscope using a single-lens system. *Int. J. Comput. Assist. Radiol. Surg.* **8**(3), 451–459 (2013). <https://doi.org/10.1007/s11548-012-0794-2>
12. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: IEEE International Conference on Computer Vision (ICCV) (1998)
13. Visentini-Scarzanella, M., Stoyanov, D.: Stereo and shape-from-shading cue fusion for dense 3d reconstruction in endoscopic surgery. In: 3rd Joint Workshop on New Technologies for Computer/Robot Assisted Surgery (CRAS) (2013). https://drive.google.com/open?id=0B0x0v_kN6YuManhfYXVtSjJDYnc