

# Online 3D reconstruction and dense tracking in endoscopic videos - Supplementary material

Michel Hayoz<sup>1</sup>, Christopher Hahne<sup>1</sup>, Thomas Kurmann<sup>3</sup>, Max Allan<sup>3</sup>, Guido Beldi<sup>2</sup> and Daniel Candinas<sup>2</sup> and Pablo Márquez-Neila<sup>1</sup>, and Raphael Sznitman<sup>1</sup>

<sup>1</sup> ARTORG Center, University of Bern, Switzerland

<sup>2</sup> Dept. of Visceral Surgery and Medicine, Inselspital, Switzerland

<sup>3</sup> Applied Research, Intuitive Surgical, USA

michel.hayoz@unibe.ch

## 0.1 Details on initialization of the control points

We obtain a collection of screen-space offsets  $\{\delta \mathbf{u}_p\}_{p=1}^P$  from optical flow between the frame image and synthetic image. Each offset  $\delta \mathbf{u}_p \in \mathbb{R}^2$  corresponding to pixel  $\mathbf{u}_p$  is projected back to the 3D scene as

$$\delta \mathbf{x}_p = \pi_{3D}(\mathbf{u}_p + \delta \mathbf{u}_p; \mathbf{d}_t, \mathbf{P}_t) - \pi_{3D}(\mathbf{u}_p; \mathbf{d}_t, \mathbf{P}_t) \quad (1)$$

to obtain a scene-space offset at location  $\mathbf{x}_p = \pi_{3D}(\mathbf{u}_p; \mathbf{d}_t, \mathbf{P}_t)$ . The offsets  $\delta \mu_k$  of the control points are then initialized to minimize the difference between the offsets modeled by the translation field  $\Delta^\mu$  and the offsets computed with optical flow,

$$\underset{\{\delta \mu_k\}_{k=0}^{K_t}}{\operatorname{argmin}} \sum_{p=1}^P \|\Delta^\mu(\mathbf{x}_p) - \delta \mathbf{x}_p\|_2^2, \quad (2)$$

where the sum is taken over the  $P$  pixels of the image. This optimization is quadratic with respect to the parameters  $\{\delta \mu_k\}_k$  and has an efficient closed-form solution via least squares.

## 0.2 Point tracking in EndoNerf and EndoSurf

Any point in space  $\mathbf{x}$  at time  $t$  can be mapped to the canonical space as  $\bar{\mathbf{x}} = \mathbf{x} + \Delta(\mathbf{x}, t)$ , where  $\Delta(\cdot)$  represents a translation vector field parameterized by a multi-layer perception. Since  $\Delta(\cdot)$  cannot be inverted, tracking needs to be performed in multiple steps. First, we render depth  $\hat{\mathbf{d}}_0$  at  $t = 0$  and reproject the point to be tracked into 3D world space using  $\mathbf{x}_{\text{ref}} = \pi_{3D}(\mathbf{u}_{\text{ref}}, \hat{\mathbf{d}}_0, \mathbf{P}_0)$  and compute the canonical reference point  $\bar{\mathbf{x}}_{\text{ref}}$ . Similarly, we render the depth and compute the collection of visible surface points  $\{\mathbf{x}_{p,t} = \pi_{3D}(\mathbf{u}_p, \hat{\mathbf{d}}_t, \mathbf{P}_t)\}_{p=1}^P$  and their respective canonical points  $\{\bar{\mathbf{x}}_{p,t}\}_{p=1}^P$  for all subsequent frames. Finally, we establish correspondences by identifying the closest point in the canonical space:

$$p^* = \underset{\{p\}_{p=1}^P}{\operatorname{argmin}} \|\bar{\mathbf{x}}_{\text{ref}} - \bar{\mathbf{x}}_p\|_2 \quad (3)$$

This process results in the tracked 3D point  $\mathbf{x}_{p^*}$  and 2D point  $\mathbf{u}_{p^*}$ .

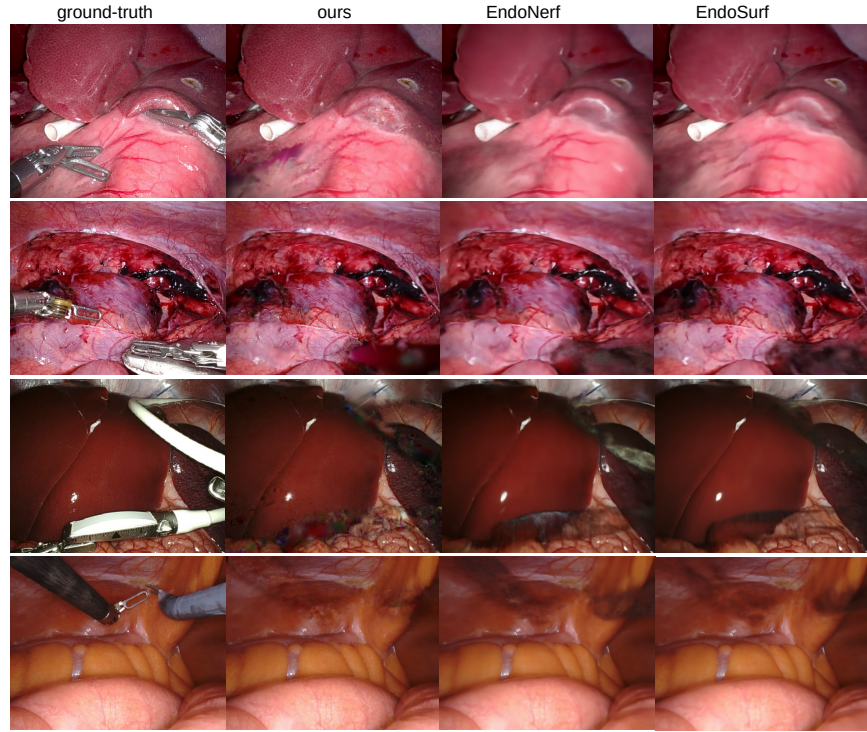


Fig. 1: Examples of input RGB images and rendered images from our method and offline 3D reconstruction methods. From top to bottom: P3\_2  $t = 45$ , H1\_1  $t = 45$ , P2\_0  $t = 100$ , and H3\_1  $t = 34$