Multi-View Stereo by Temporal Nonparametric Fusion

Aalto University

Yuxin Hou
Aalto University
yuxin.hou@aalto.fi

Juho Kannala
Aalto University
juho.kannala@aalto.fi

Arno Solin
Aalto University
arno.solin@aalto.fi

Introduction

- Novel idea for depth estimation from image-pose **sequences**
- Existing methods are frame-by-frame, leading to **flickery** results
- Leverage **multi-view information** without extra costs
- A **pose-kernel prior** to encode similarity of the camera poses
- Encourages similar poses to have **resembling latent spaces**
- Suitable both for **batch** estimation and **online** estimation
- Can be combined with a post-processing stage

Pose-kernel Gaussian process prior

- Define a distance measure between two camera poses P_i and P_j

$$D[P_i, P_j] = \sqrt{\|\mathbf{t_i} - \mathbf{t_j}\|^2 + \frac{2}{3}\mathrm{tr}(\mathbf{I} - \mathbf{R}_i^\top \mathbf{R_j})},$$

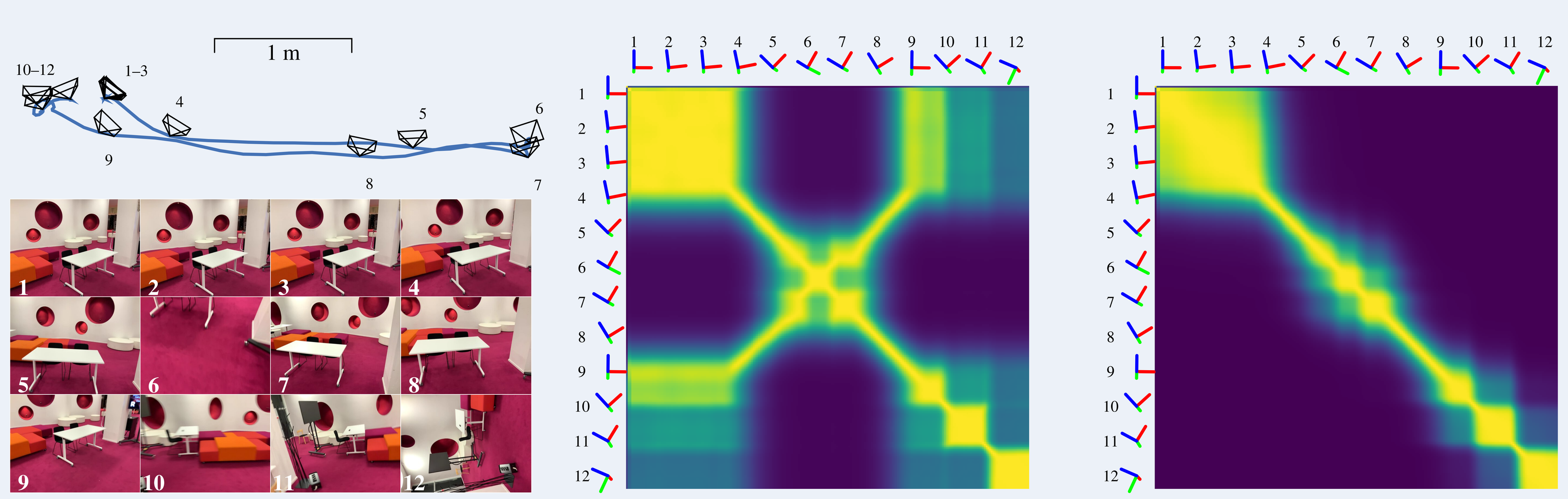
  where the poses are $P = \{\mathbf{t}, \mathbf{R}\}$, residing in $\mathbb{R}^3 \times \mathrm{SO}(3)$
- Use the **Matérn class** as covariance function (kernel) structure

$$\kappa(P, P') = \gamma^2 \left(1 + \frac{\sqrt{3}\, D[P, P']}{\ell}\right) \exp\left(-\frac{\sqrt{3}\, D[P, P']}{\ell}\right)$$

  to enable the latent space to behave in a continuous and smooth fashion
- State inference problem as a GP regression model

$$z_j(t) \sim \mathrm{GP}(0, \kappa(P[t], P[t'])),$$
$$y_{j,i} = z_j(t_i) + \varepsilon_{j,i}, \quad \varepsilon_{j,i} \sim \mathrm{N}(0, \sigma^2)$$

  assign independent GP priors to z_i, and consider the encoder outputs y_i to be noise-corrupted latent code



(a) Camera pose track and frames (b) Pose-kernel in batch mode (c) Pose-kernel in chain mode (online)

Fig. 1: Illustrative example of our pose-kernel.

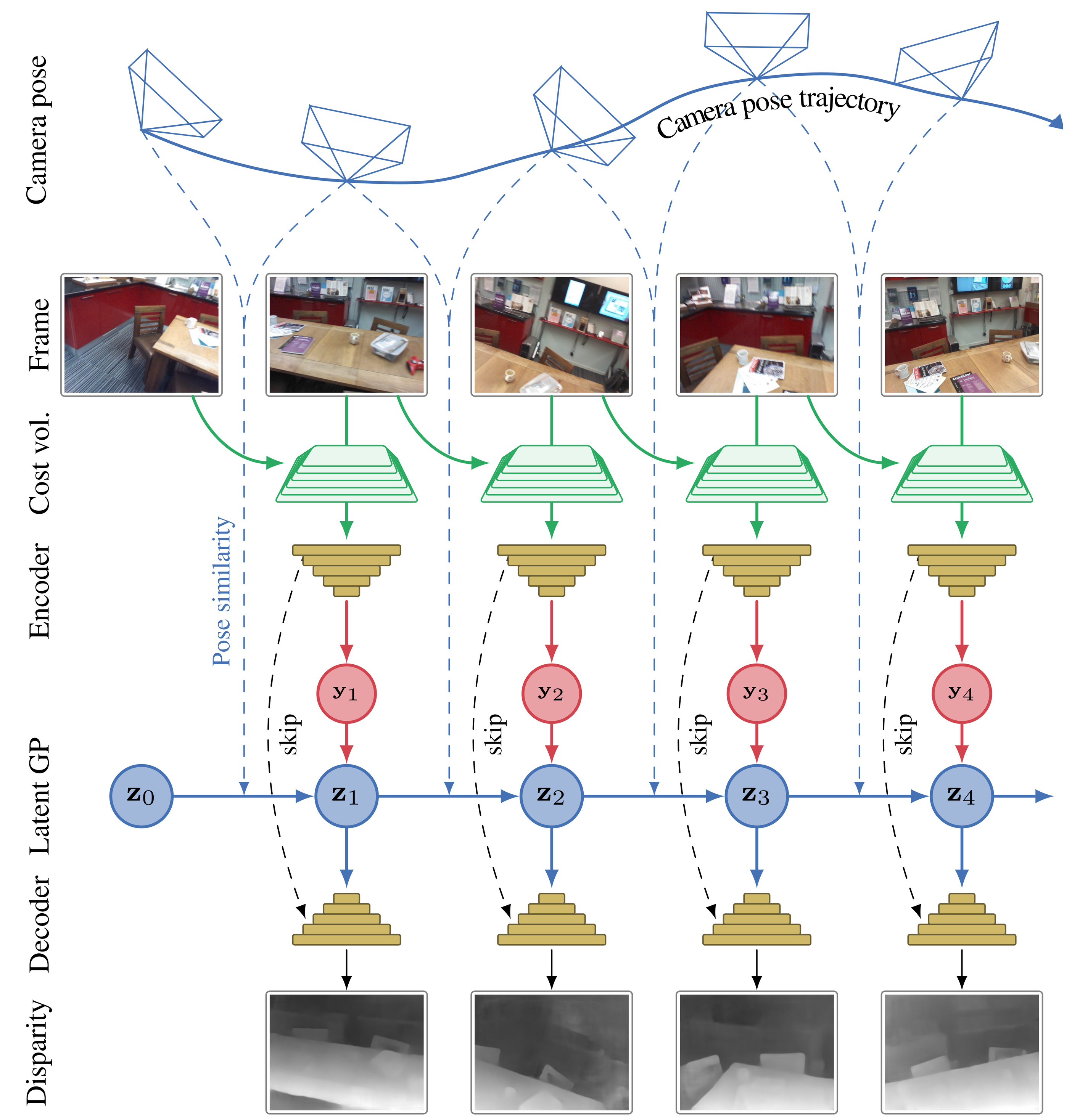
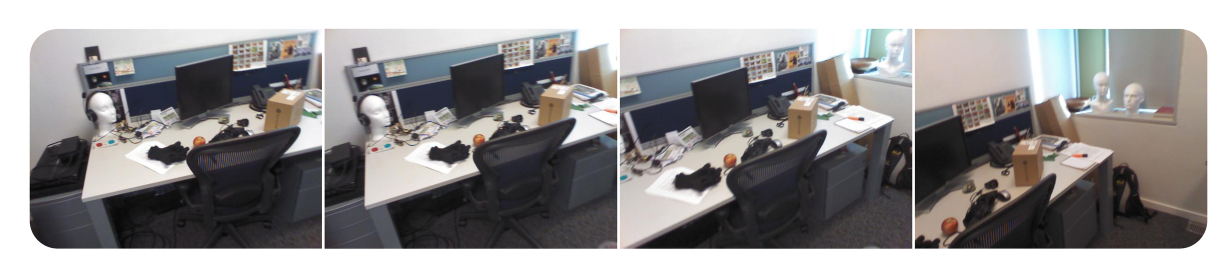
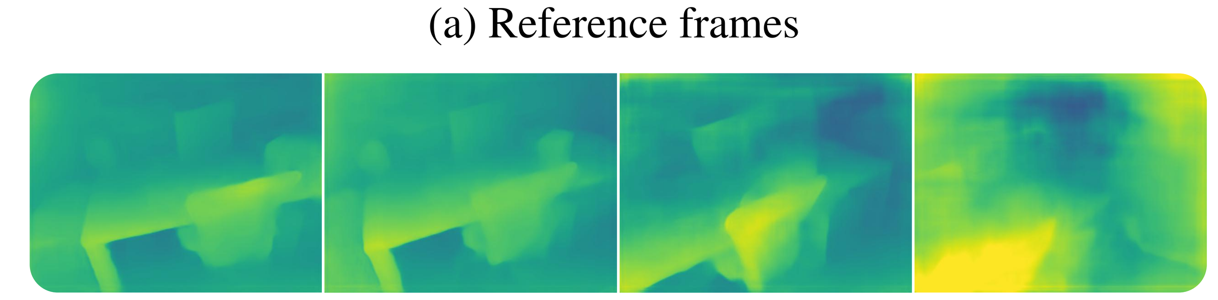


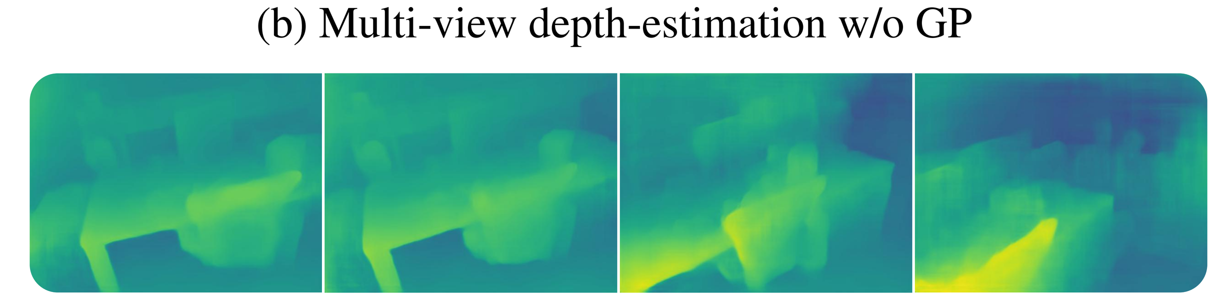
Fig. 2: Illustrative sketch of our MVS approach.



(a) Reference frames



(b) Multi-view depth-estimation w/o GP



(c) Multi-view depth-estimation with GP

Fig. 3: Introducing information sharing in the latent space makes results more stable and edges sharper.

Batch estimation

- Solve independent GP regression tasks with one matrix inversion

$$\mathbb{E}[\mathbf{Z} \mid \{(P_i, \mathbf{y}_i)\}_{i=1}^N] = \mathbf{C}\,(\mathbf{C} + \sigma^2\,\mathbf{I})^{-1}\,\mathbf{Y},$$
$$\mathbb{V}[\mathbf{Z} \mid \{(P_i, \mathbf{y}_i)\}_{i=1}^N] = \mathrm{diag}(\mathbf{C} - \mathbf{C}\,(\mathbf{C} + \sigma^2\,\mathbf{I})^{-1}\,\mathbf{C})$$

  where $\mathbf{C}_{i,j} = \kappa(P_i, P_j)$ and $\mathbf{Y} = (\mathbf{y}_1\ \mathbf{y}_2\ \ldots\ \mathbf{y}_N)^\top$ are outputs from the encoder

Online estimation

- Solve GP inference in **state-space form**

$$\mathbf{\Phi}_i = \exp\left[\begin{pmatrix} 0 & 1 \\ -3/\ell^2 & -2\sqrt{3}/\ell \end{pmatrix} \Delta P_i\right],$$

  where $\Delta P_i = D[P_i, P_{i-1}]$ is the pose-distance

$$\mathbf{z}_i \mid \mathbf{y}_{1:i-1} \sim \mathrm{N}(\bar{\boldsymbol{\mu}}_i, \bar{\mathbf{\Sigma}}_i), \qquad \bar{\boldsymbol{\mu}}_i = \mathbf{\Phi}_i\,\boldsymbol{\mu}_{i-1}, \quad \bar{\mathbf{\Sigma}}_i = \mathbf{\Phi}_i\,\mathbf{\Sigma}_{i-1}\,\mathbf{\Phi}_i^\top + \mathbf{Q}_i,$$

  where $\mathbf{Q}_i = \mathbf{\Sigma}_0 - \mathbf{\Phi}_i\,\mathbf{\Sigma}_0\,\mathbf{\Phi}_i^\top$ The posterior mean and covariance is then given by:

$$\boldsymbol{\mu}_i = \bar{\boldsymbol{\mu}}_i + \mathbf{k}_i\,(\mathbf{y}_i^\top - \mathbf{h}^\top \bar{\boldsymbol{\mu}}_i) \quad \text{and} \quad \mathbf{\Sigma}_i = \bar{\mathbf{\Sigma}}_i - \mathbf{k}_i\,\mathbf{h}^\top \bar{\mathbf{\Sigma}}_i$$

Experiments

- Trained with mixed data set of **SUN3D**, **RGBD**, **MVS**, and **Scenes11**
- **Jointly train the GP hyperparameters** with mini-sequences of length three
- Robust to neighbour frame selection
- Better 3D reconstruction results demonstrate **temporal consistency**
- A **real-time iOS app** to demonstrate the efficiency

Conclusion

- We show that our method enables the model to leverage multi-view information but keeps the model structure simple and time-efficient
- We show that our pose-kernel can measure the 'closeness' between frames and the GP prior improves the accuracy
- Using a confidence measure to penalize wrong predictions from propagating further might improve the method

Presented at ICCV 2019, Seoul, Korea. Contact details: yuxin.hou@aalto.fi