The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Gaussian Process Priors for View-Aware Inference

Yuxin Hou[1,*], **Ari Heljakka**[1,2,*], **Arno Solin**[1]
[1]Aalto University, Espoo, Finland,
[2]GenMind Ltd., Finland
{yuxin.hou, ari.heljakka, arno.solin}@aalto.fi

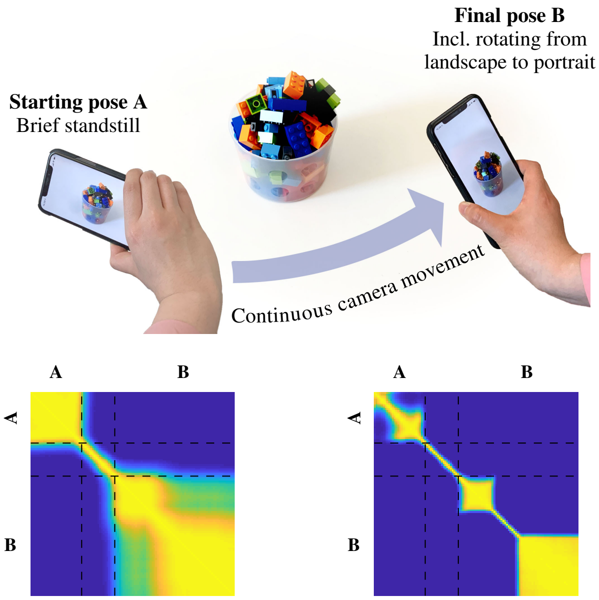
Abstract

While frame-independent predictions with deep neural networks have become the prominent solutions to many computer vision tasks, the potential benefits of utilizing correlations between frames have received less attention. Even though probabilistic machine learning provides the ability to encode correlation as prior knowledge for inference, there is a tangible gap between the theory and practice of applying probabilistic methods to modern vision problems. For this, we derive a principled framework to combine information coupling between camera poses (translation and orientation) with deep models. We proposed a novel view kernel that generalizes the standard periodic kernel in $\mathrm{SO}(3)$. We show how this soft-prior knowledge can aid several pose-related vision tasks like novel view synthesis and predict arbitrary points in the latent space of generative models, pointing towards a range of new applications for inter-frame reasoning.

Introduction

Gaussian processes (GPs, Rasmussen and Williams 2006) provide a flexible probabilistic framework for combining *a priori* knowledge with forecasting, noise removal, and explaining data. Their strengths are in many ways complementary to those of deep neural networks which perform best in applications where large training data sets are available and the test points reside close to the training samples. The tremendous success of deep neural networks in solving many fundamental computer vision tasks has largely dictated the research in the past years, but recent interest in prediction under incomplete inputs has motivated combining the extreme flexibility and expressive power of current computer vision models with structured constraints encoded by GP priors. Application areas include uncertainty quantification (see discussion in Blundell et al. 2015; Kendall and Gal 2017), auxiliary data fusion, and prediction under scarce data. These are instrumental for delivering practical methods and robustifying inference.

In this paper, we aim to fill a tangible gap between the theory and practice of applying probabilistic methods to certain computer vision tasks. We propose a tailored Gaussian process prior for encoding knowledge of camera poses into probabilistic models. In GPs, prior assumptions are encoded by a covariance function. As illustrated in Fig. 1, we aim to encode the notion of *similarity* of camera views given the known camera movement.



(a) Translation cov

(b) Orientation cov

Figure 1: We propose a GP prior for encoding known six degrees-of-freedom camera poses into probabilistic models. In region A, the phone starts from standstill with minor rotation (high overall covariance in (a)). Between A and B, it moves to the right while rotating (low overall covariance in (a) and (b)). In B, the phone firstly stands still (high overall covariance), then rotates from portrait to landscape (low covariance in (b), higher in (a)), and is finally still again.

In practice, the camera movement estimation is typically fused with motion information from inertial sensors. New consumer hardware in smartphones and cars typically have these capabilities built-in—Apple iPhones/iPads run ARKit and Android devices Google ARCore, both exposing real-time six degrees-of-freedom camera pose data. This readily available motion information could be utilized as priors for improving standard visual regression and classification tasks.

[*] Equal contribution.

However, typical computer vision methods operating on a stream of images consider the frames independently and merely post-process the outputs by, *e.g.*, linear interpolation or temporal low-pass filtering.

This paper is *bridging*: We emphasize the principled link between computer vision and non-parametric inference for encoding probabilistic information between camera poses, advocating for the use of more principled strategies for inter-frame reasoning in computer vision. Our contributions in this paper are: *(i)* We propose a novel view covariance function for encoding 3D camera orientation which extends the theory of GP models towards vision applications. *(ii)* We push the boundaries of GP applications in computer vision. For the first time, we use a GP model on an autoencoder to predict learnt shapes in arbitrary angles. *(iii)* We also introduce an approach to non-linear latent space interpolation in generative image models, using our view kernel.

Background

Gaussian processes (GPs) provide a probabilistic plug-and-play framework for specifying prior knowledge inside models. As a general-purpose machine learning paradigm they are instrumental in applications for discovering structure in signals (Duvenaud 2014), regression tasks (Bui et al. 2016), data-efficient reinforcement learning (Deisenroth and Rasmussen 2011), and probabilistic numerics (Hennig, Osborne, and Girolami 2015). In theory, their applicability is only limited by the availability of prior knowlege that can be encoded.

We focus on GP models that admit the form of a *Gaussian process prior* $f(\mathbf{x}) \sim \mathrm{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ and *likelihood* $\mathbf{y} \,|\, \mathbf{f} \sim \prod_{i=1}^{n} p(y_i \,|\, f(\mathbf{x}_i))$, where the data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ are input–output pairs, $\mu(\mathbf{x})$ the mean, and $\kappa(\mathbf{x}, \mathbf{x}')$ the covariance function of the GP prior. This family covers many standard modelling problems, including regression and classification tasks.

GPs are typically associated with two issues hindering their wider use: (*i*) prohibitive cubic scaling in the number of training samples n and (*ii*) the need for approximative inference when dealing with non-Gaussian likelihoods. Recent research has delivered methods to overcome these limitations by methods such as basis function projection (Lázaro-Gredilla et al. 2010; Hensman, Durrande, and Solin 2018), matrix structure exploiting (Wilson and Nickisch 2015; Wang et al. 2019), stochastic inference (Hensman, Fusi, and Lawrence 2013; Krauth et al. 2017), and temporal models (Särkkä, Solin, and Hartikainen 2013; Solin, Hensman, and Turner 2018). The availability of GPU-accelerated software libraries such as GPflow (Matthews et al. 2017) and GPyTorch (Gardner et al. 2018) have recently made GP models more applicable as building blocks for larger models. Therefore, the traditional limitations are now less severe, allowing GPs to provide exciting opportunities for computer vision applications.

In this paper, the main contributions relate to the GP prior, where the *a priori* assumptions are encoded by the covariance function (kernel) $\kappa(\cdot, \cdot)$. Without loss of generality, we constrain our interest to models with $\mu(\mathbf{x}) = 0$. Some SLAM methods exploit GP priors in $\mathrm{SE}(3)$ for continuous trajectory estimation (Anderson and Barfoot 2015). For computer vision and graphics applications, recent work in kernel design has focused more on encoding the ignorance rather than the knowledge about orientation. Invariant kernels (see, *e.g.*, Haasdonk and Burkhardt 2007) can robustify deep convolutional models against rotation, while translation insensitive kernels (Dutordoir et al. 2020) can account for problems with patch similarity across images. We, however, aim to encode explicit prior knowledge about inter-image camera poses—view similarity—by crafting a view kernel that accounts for camera translation and orientation. Song et al. (2009) proposed an inner product kernel between rotations, which can be regarded as a linear model in the Hilbert space, while we span a multi-dimensional periodic model in that space. This line of research also connects to distance measures between rigid bodies (Mazzotti, Sancisi, and Parenti-Castelli 2016).

Perhaps due to the two limitations mentioned earlier, GPs have not been extensively used in computer vision applications. Sufficient and necessary conditions for Gaussian kernels on metric spaces are derived in Jayasumana et al. (2013), with the focus on theoretical ground-work. GP priors for rigid motions applied to object tracking is extensively studied in Lang and Hirche (2017); Lang, Kleinsteuber, and Hirche (2018), which we also compare against. There has also been previous work in combining variational autoencoders with GP priors in vision (Eleftheriadis et al. 2016; Casale et al. 2018) and GP based latent variable models for multi-view and view-invariant facial expression recognition (Eleftheriadis, Rudovic, and Pantic 2015a,b). In Casale et al. (2018), GPs are applied to face image modelling, where the GP accounts for the pose, and in Urtasun, Fleet, and Fua (2006) used them for 3D people tracking.

From an application point of view, leveraging information from consecutive views lies at the heart of many subfields in computer vision. Video analysis, multi-view methods, optical flow, visual tracking, and motion estimation and correction all directly build on the object or camera movement cues in consecutive image frames. View priors can also help in semantic processing of video (Everingham, Sivic, and Zisserman 2006) or depth estimation (Hou, Kannala, and Solin 2019; Hou et al. 2021). However, in many 'one-shot' applications in visual regression and classification, the frames of the image sequence are treated as independent, and typically processed with linear interpolation or low-pass filtering.

Camera Pose Priors

In geometric computer vision (*e.g.*, Hartley and Zisserman 2003), the standard description of a camera projection model is characterized by *extrinsic* and *intrinsic* camera parameters. The extrinsic parameters denote the coordinate system transformations from world coordinates to camera coordinates, while the intrinsic parameters map the camera coordinates to image coordinates. In the standard *pinhole camera* model, this corresponds to

$$\begin{pmatrix} u & v & 1 \end{pmatrix}^{\mathsf{T}} \propto \mathbf{K} \begin{pmatrix} \mathbf{R}^{\mathsf{T}} & -\mathbf{R}^{\mathsf{T}}\mathbf{p} \end{pmatrix} \begin{pmatrix} x & y & z & 1 \end{pmatrix}^{\mathsf{T}}, \quad (1)$$

where (u, v) are the image (pixel) coordinates, $(x, y, z) \in \mathbb{R}^3$ are the world coordinates, $\mathbf{K}$ is the intrinsic matrix and the $\mathbf{p} \in \mathbb{R}^3$ and $\mathbf{R}$ describe the position of the camera centre and the orientation in world coordinates respectively. From Eq. (1), given a set of fixed world coordinates and a known motion between frames the driver for changes in pixel values (u, v) is the camera pose $P = \{\mathbf{p}, \mathbf{R}\}$.

Kernels in $\mathrm{SE}(3)$

In the mathematical sense, the three-dimensional camera poses belong to the special Euclidean group, $\mathrm{SE}(3)$, whose elements are called rigid motions or Euclidean motions. They comprise arbitrary combinations of translations and rotations, but not reflections. This group contains transformations represented as a translation followed by a rotation: $\mathrm{SO}(3)\times\mathrm{T}(3)$, where the former denotes the special orthogonal rotation group and the latter the group of translations. A camera pose $P = \{\mathbf{p}, \mathbf{R}\}$ is an element of this group. We consider the orientation and translation contributions entering the prior separately: $\kappa_{\mathrm{pose}}(P, P') = \kappa_{\mathrm{trans.}}(\mathbf{p}, \mathbf{p}')\, \kappa_{\mathrm{view}}(\mathbf{R}, \mathbf{R}')$, since in the general case separability imposes a less informative prior. As the translation vectors reside in $\mathbb{R}^3$, we may directly write the translation kernel as any suitable covariance function (see, *e.g.*, Rasmussen and Williams 2006; Duvenaud 2014). An apparent first choice is the so-called squared exponential (RBF, exponentiated quadratic) covariance function:

$$\kappa(\mathbf{p}, \mathbf{p}') = \sigma^2 \exp\left(-\frac{\|\mathbf{p} - \mathbf{p}'\|^2}{2\ell^2}\right), \qquad (2)$$

where σ^2 denotes a magnitude and $\ell > 0$ is a characteristic lengthscale hyperparameter. This particular choice of covariance function encodes continuity, smoothness, and translation invariance in $\mathbf{p}$. An example realization of the translation covariance matrix is visualized in Fig. 1a.

View Orientation Kernels

Since translations can be considered directly, our main interest is formulating a proper orientation covariance function in $\mathrm{SO}(3)$. Here, the first choice could be to leverage the standard periodic kernel, which can be derived following MacKay (1998): Given a valid covariance function $\kappa(\mathbf{u}, \mathbf{u}')$, we can introduce a non-linear mapping $\mathbf{x} \mapsto \mathbf{u}(\mathbf{x})$, through which to define a new covariance function $\kappa'(\mathbf{x}, \mathbf{x}') \triangleq \kappa(\mathbf{u}(\mathbf{x}), \mathbf{u}(\mathbf{x}'))$. The standard periodic kernel (*cf.*, Rasmussen and Williams 2006) is usually derived by the mapping $\theta \mapsto \mathbf{u}$ that warps θ to the unit circle: $\mathbf{u}(\theta) = (\cos(\theta), \sin(\theta))$. Combining this with the covariance function in Eq. (2) gives

$$\kappa(\theta, \theta') = \exp\left(-\frac{2\sin^2((\theta - \theta')/2)}{\ell^2}\right), \qquad (3)$$

which can be used for imposing a periodic prior over inputs $\theta \in \mathbb{R}$. We aim to extend this 1D standard periodic kernel to 3D rotations (see also Hamsici and Martinez 2008).

Euler angle formalism Assuming Euler angles $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ to be fully separable, we can extend Eq. (3) to 3D rotations directly. This would correspond to a separable view kernel (see Fig. 2d for the corresponding distance function):

$$\kappa(\boldsymbol{\theta}, \boldsymbol{\theta}') = \prod_{j=\{1,2,3\}} \exp\left(-\frac{2\sin^2((\theta_j - \theta'_j)/2)}{\ell_j^2}\right). \qquad (4)$$

This, however, can suffer from issues related to Euler angles like possibly singular representations and gimbal lock (loss of one degree of freedom, see, *e.g.*, (Diebel 2006; Featherstone 2014)), and should thus be avoided as an internal representation of orientation.

Quaternion formalism Instead of Euler angles, common representations for orientation are given in terms of rotation matrices or quaternions. The set of unit quaternions, $\mathbf{q} = (q_{\mathrm{w}}, q_{\mathrm{x}}, q_{\mathrm{y}}, q_{\mathrm{z}})$, s.t. $\|\mathbf{q}\| \equiv 1$, forms the 3D rotation group $\mathrm{SO}(3)$ covering the S^3 sphere. In order to seek a similar, but higher-dimensional, form of Eq. (3), the quaternion representation can directly be used as a mapping. This would make sense, as the derivation of the standard periodic covariance function can be viewed as a mapping onto the complex plane and quaternions represent a 4D extension of complex numbers. So we may define the distance between quaternions $\mathbf{q}_1$ and $\mathbf{q}_2$ as the norm of their difference:

$$d_{\mathrm{quat}}(\mathbf{q}_1, \mathbf{q}_2) = 2\|\mathbf{q}_1 - \mathbf{q}_2\|. \qquad (5)$$

The quaternion model has previously been discussed by Lang and Hirche (2017) and Lang, Kleinsteuber, and Hirche (2018). However, the resulting covariance function is not well-behaved in all orientations—due to non-uniqueness of quaternions—as can be seen from Fig. 2b, where full-turn (2π) correlations are close to zero.

Rotation matrix formalism The peculiarities with the previous formulations, as visualized in Fig. 2, acted as a motivation to seek a more principled generalization of the periodic covariance function with rotation matrices. Since there is no direct way to use a rotation matrix as a mapping to extend Eq. (3), we consider the geodesic (arc) distance. Considering the eigendecomposition of $\mathbf{R}$ that define the rotation axis and angle (see supplement), we have the geodesic distance defined by rotation matrices $\mathbf{R}$:

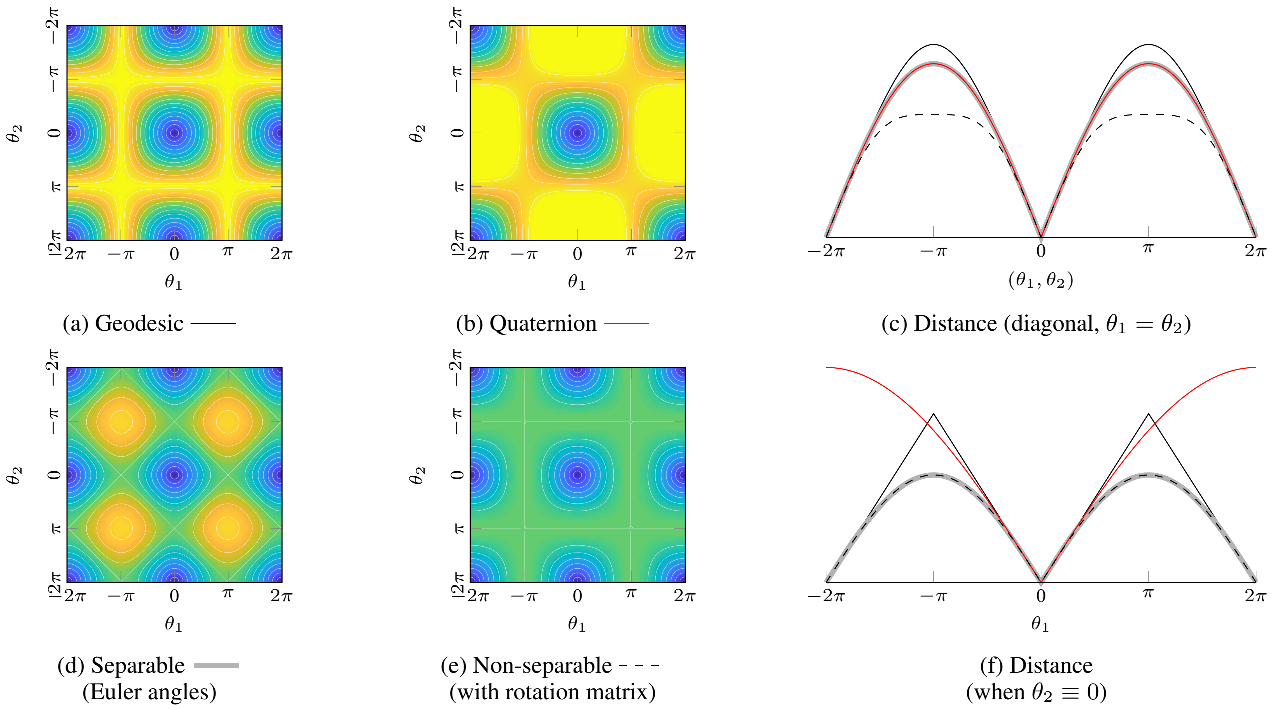
$$d_{\mathrm{g}}(\mathbf{R}, \mathbf{R}') = \arccos\left(\frac{1}{2}(\mathrm{tr}(\mathbf{R}^\mathsf{T}\mathbf{R}') - 1)\right). \qquad (6)$$

To derive the 3D counterpart of the standard periodic kernel, a Taylor expansion (see supplement) for the geodesic distance around the origin gives a mapping $d_{\mathrm{g}}(\mathbf{R}, \mathbf{R}') \approx \sqrt{\mathrm{tr}(\mathbf{I} - \mathbf{R}^\mathsf{T}\mathbf{R}')}$ (visualized in Fig. 2e) that we use for the non-separable covariance function:

$$\kappa_{\mathrm{view}}(\mathbf{R}, \mathbf{R}') = \exp\left(-\frac{\mathrm{tr}(\mathbf{I} - \mathbf{R}^\mathsf{T}\mathbf{R}')}{2\ell^2}\right). \qquad (7)$$

This proposed 3D kernel Eq. (7) gives the standard periodic kernel as a special case where there is only rotation around one of the axes (see Fig. 2f). Moreover, the proposed Eq. (7) may be generalized to $\kappa_{\mathrm{view}}(\mathbf{R}, \mathbf{R}') = \exp(-\frac{1}{2}\,\mathrm{tr}(\boldsymbol{\Lambda} - \mathbf{R}^\mathsf{T}\boldsymbol{\Lambda}\mathbf{R}'))$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\ell_{\mathrm{x}}^{-2}, \ell_{\mathrm{y}}^{-2}, \ell_{\mathrm{z}}^{-2})$, which can account for different characteristic scaling per axis flexibly. (NB: The ℓs are coupled and its interpretation is not as straightforward as scaling for the respective axes)

To summarize, we propose the non-separable orientation covariance function $\kappa_{\mathrm{view}}(\cdot, \cdot)$ that preserves a symmetric correlation structure around origin (like the geodesic model), does not suffer from the degeneracy of Euler angles, and generalizes the gold-standard (one-dimensional) periodic kernel to high-dimensional rotations.



(a) Geodesic ——

(b) Quaternion ——

(c) Distance (diagonal, $\theta_1 = \theta_2$)

(d) Separable ▬ (Euler angles)

(e) Non-separable - - - (with rotation matrix)

(f) Distance (when $\theta_2 \equiv 0$)

Figure 2: Characterization of differences between different orientation distance measures. Left: Distance matrices between two degrees-of-freedom rotations. (a) shows the geodesic distance (see Eq. (6)), (b) the quaternion norm distance (see Eq. (5)), (d) the separable periodic distance (Eq. (4)), and (e) the non-separable orientation distance (Eq. (7)). Right: Distance evaluations along the diagonal and when $\theta_2 \equiv 0$, showing that in 1D (d) and (e) coincide, while (e) is symmetric in 2D/3D.

Application Experiments

In the experiments, we show examples of real-world applications of the view kernel in probabilistic view synthesis. In the first experiment, we extend the GP variational autoencoder model with our view kernel for a view synthesis task. The second experiment is concerned with latent space interpolation for human face modelling, showcasing the general applicability of the kernel. Further examples and comparisons are included in the supplement.

View Synthesis with a GP Prior VAE

We consider the task of using a variational autoencoder (VAE) to predict how objects look in orientations that are not in the training set. We first describe how the problem was previously addressed by Casale et al. (2018) with the Gaussian Process Prior Variational Autoencoder (GPPVAE), explain a major limitation in this approach, and then overcome this limitation with our kernel. GPPVAE is a *fully probabilistic* model that captures correlations in both object identities and views by leveraging covariance structure in latent space. The kernel defines a prior for latent code $\mathbf{z}$. Given an object ID and view angle, the encoder and GP posterior predict the posterior $\mathbf{z}$. Intuitively, the prediction is based on the relation between training samples.

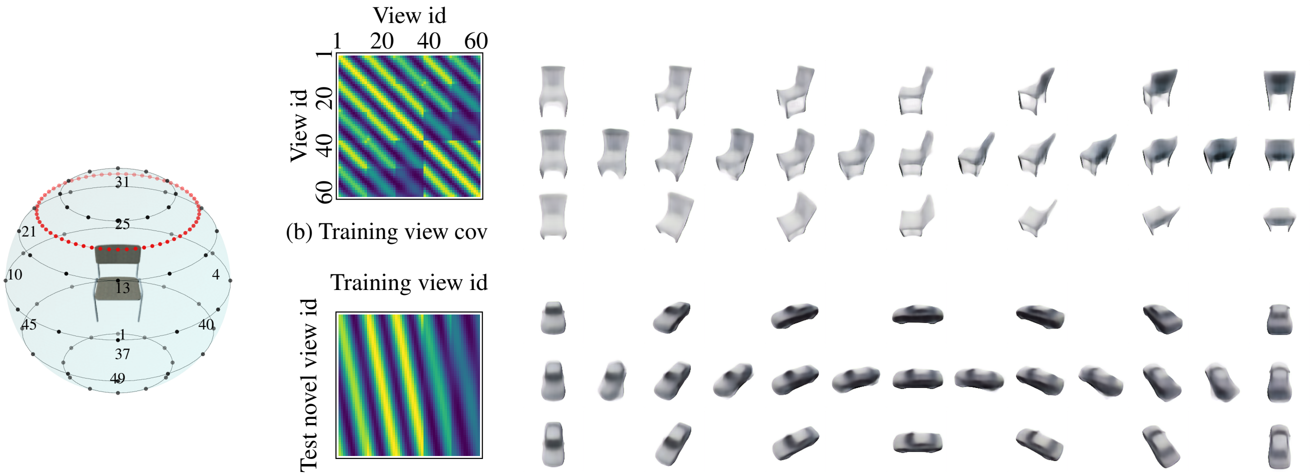
Given training images $\mathcal{Y}$, training object feature vectors $\mathcal{X}$, and training views $\mathcal{P}$, the predictive posterior for an image $\mathbf{y}_\star$ for an object with features $\mathbf{x}_\star$ seen from a view $P_\star$ is given (see detailed presentation in Casale et al. 2018) by

$$p(\mathbf{y}_\star \mid \mathbf{x}_\star, \mathcal{Y}, \mathcal{X}, \mathcal{P}) \approx \int \underbrace{p(\mathbf{y}_\star \mid \mathbf{z}_\star)}_{\text{decode prediction}} \underbrace{p(\mathbf{z}_\star \mid \mathbf{x}_\star, P_\star, \mathcal{Z}, \mathcal{X}, \mathcal{P})}_{\text{GP predictive posterior}} \underbrace{q(\mathcal{Z} \mid \mathcal{Y})}_{\text{encode training data}} \,\mathrm{d}\mathbf{z}_\star \,\mathrm{d}\mathcal{Z}, \quad (8)$$

where $\mathbf{z}_\star$ are the predicted latent representations and $\mathcal{Z}$ are latent representations of training images. Given fixed views and objects, the task of GPPVAE is to predict images $\mathbf{y}_\star$ for an object in the view $P_\star$ that remained unobserved.

However, though Casale et al. (2018) present the task as ‘out-of-sample’ prediction, their approach of brute learning the covariance does not support arbitrary 3D angles. Rather, it is defined based on the assumption that all query views in the test set have already been observed for at least one object in the training set. When that assumption does not hold, only a fixed number of 3D rotations are available. In GPPVAE, all experiments only consider rotations in one dimension, modelled with the 1D standard periodic kernel or the fully-learned kernel. The 1D standard periodic kernel cannot handle 3D rotations and the fully-learned kernel can only capture the correlations within fixed training views. In contrast, our proposed kernel that extends the 1D standard periodic kernel to $\mathrm{SO}(3)$ can work with arbitrary 3D angles.

To showcase our kernel with 3D rotations, we carried out an experiment with ShapeNet (Chang et al. 2015) 3D chair models at 128×128 resolution. We use 1660 different chairs in total. For each object, we ren-



(a) View poses (b) Training view cov (c) Cross-cov (d) Results of novel views synthesis

Figure 3: ShapeNet experiments with a GPPVAE. (a) Visualization of the 60 view angles (black dots) in the training data. (b) The cov matrix for 60 training views. (c) The cross-cov matrix for the test novel views (red dots in (a)) and training views (black dots). (d) We experimented with both chairs and cars. Our proposed kernel allows to predict arbitrary views which are not presented in training data. For each category, the first (elevation 30°) and the third row (elevation 60°) show predictions for angles found in the training set, while the whole second row shows predictions for angles not found in the training set (tight red dots in (a)).

der images from 60 fixed views, considering both azimuth angles $(0°, 30°, 60°, \ldots, 330°)$ and elevation angles $(0°, 30°, 60°, -30°, -60°)$. The camera view angles are shown in Fig. 3a. We randomly selected 80% images for training (81,312 images), 10% for validation (10,164 images) and 10% for testing (10,164 images). Following original GPPVAE, we compute the view covariance based only on orientation angles (cameras at fixed radius from the object centre; translation seen as function of orientation). For the object covariance, we use a linear kernel between learned object features. The resulting composite kernel $\kappa(\mathbf{x}, \mathbf{R}; \mathbf{x}', \mathbf{R}')$ expresses the covariance between two chair images in terms of the relative view orientation between orientations $\mathbf{R}$ and $\mathbf{R}'$ and object feature vectors $\mathbf{x}$ and $\mathbf{x}'$:

$$\kappa(\mathbf{x}, \mathbf{R}; \mathbf{x}', \mathbf{R}') = \underbrace{\mathbf{x}^\mathsf{T}\mathbf{x}'}_{\text{object}} \underbrace{\exp\big(-\frac{1}{2}\mathrm{tr}(\mathbf{\Lambda} - \mathbf{R}^\mathsf{T}\mathbf{\Lambda}\mathbf{R}')\big)}_{\text{view}}, \quad (9)$$

where $\mathbf{\Lambda} = \mathrm{diag}(\ell_\mathrm{x}^{-2}, \ell_\mathrm{y}^{-2}, \ell_\mathrm{z}^{-2})$ and we learn the lengthscale hyperparameters $\ell_\mathrm{x}, \ell_\mathrm{y}, \ell_\mathrm{z}$ as part of the training. Due to rich variability in chair shapes, we consider a higher rank $(M = 128)$ than the original setup for the object covariance (see supplement for details). We first experiment on same task as GPPVAE (in-sample evaluation). For the proposed view kernel, the MSE is 0.025 ± 0.012, which still has slightly better performance than the fully-learned view-covariance matrix as in Casale et al. (2018) (0.026 ± 0.012). This also shows that encoding the information through a view kernel (with only hyperparameters to learn), rather than through brute free-form optimization, is sensible.

Fig. 3d demonstrates the capability of our kernel for novel view predictions conditioned on an object ID, with truly 'out-of-sample' views (novel viewpoints in red in Fig. 3a). The closest views within the training set are also visualized, which demonstrates that our model has learned to disentangle view and content by the aid of the view prior. The qualitative results on ShapeNet cars also show the generalizability.

We evaluate MSEs for the novel view prediction for each kernel, using the trained lengthscale and magnitude hyperparameters from the view kernel (the parameters have the same interpretation across kernels). The practical degeneracy of the separable kernels (based on Euler angles) and quaternion kernels can make training unstable. For our non-separable view kernel we get an MSE of 0.036. Given the hyperparameters trained with the non-separable model, the separable model performs almost equally well. The quaternion distance kernel fails at this task (MSE 0.058).

Robust Interpolation for Face Reconstruction

As a second example of inter-frame reasoning, we consider view-aware GP interpolation in the latent space of a Generative Adversarial Network (GAN, Goodfellow et al. 2014) for face generation. A GAN incorporates a generator network that acts as a feature extractor, allowing an image to be represented by a low-dimensional latent code. By utilizing the pose information of the view-aware kernel, we can do GP regression in the latent space. The data comprises short video sequences of faces of four volunteers captured by an Apple iPhone XS. We used a custom app for capturing the video stream (1440×1920 at 60 Hz) interleaved with camera poses from the Apple ARKit API.

In absence of a built-in encoder, as in case of most GANs, we use an optimization setup to find out the best latent code for an image j (similarly to Abdal, Qin, and Wonka 2019). The traditional approach has been to learn these codes from i.i.d. training data, and under the assumption that we essen-

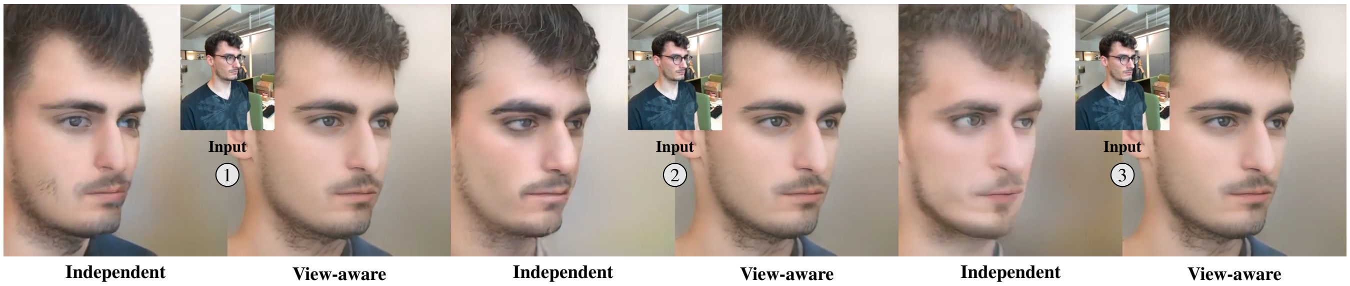


Figure 4: View-aware manipulations in the latent space of StyleGAN (Karras, Laine, and Aila 2019). Example of denoising of the GAN reconstructions for consecutive frames (note the noisy independent reconstructions) in a longer video, where every frame is treated as a noisy observation. See the supplement for video examples.

tially have only a single 'observation' of each entity that the image represents. We now relax this assumption and consider the more general case where we postulate, for each input image frame, the existence of a hidden 'correct' latent code $\mathbf{f}_j \in \mathbb{R}^d$ that encodes both the time-invariant aspect (face identity) and the time-dependent aspect (pose of the face), and then re-interpret each latent code produced by an encoder or optimizer as a noisy 'observation' $\mathbf{y}_j \in \mathbb{R}^d$ of the correct code. Consider the case of images that depict a face with fixed identity. We cast this as a GP regression problem in which each latent dimension, $i = 1, 2, \ldots, d$ is independent. The likelihood is $y_{j,i} = f_i(P_j) + \varepsilon_{j,i}$, $\varepsilon_{j,i} \sim \mathrm{N}(0, \sigma_\mathrm{n}^2)$, for frames $j = 1, 2, \ldots, n$. The GP prior is over the camera poses P_j: $f_i(P) \sim \mathrm{GP}(0, \kappa_\mathrm{view}(P, P'))$. Solving these independent GP regression problems only requires inverting one $n \times n$ covariance matrix, which makes inference fast. We use two or more images of a sequence to predict the expected latent code, $\mathrm{E}[\mathbf{f}(P_j) \,|\, \mathcal{D}]$, for any image in the sequence, without necessarily ever running that image through the encoder. We can apply these predictions in several ways, here focusing separately on noise reduction (leveraging all available image frames) and view synthesis (leveraging as few as two frames).

We demonstrate this approach in the 18×512 latent space of StyleGAN (Karras, Laine, and Aila 2019) based on four image sequences, each depicting a specific face identity (see Fig. 5 and the supplement). We find the 'observed' latent codes using an optimizer, leveraging VGG16 feature projections (Simonyan and Zisserman 2015; Puzer (GitHub user) 2019). Separately for each face identity, our method infers the 'correct' latent codes for each pose. The GAN generator then decodes those back to 1024×1024 image space. The values for the three hyperparameters were chosen to $\sigma^2 = 0.1$, $\ell = 1.098$, and $\sigma_\mathrm{n}^2 = 0.0001$ (pre-trained on an independent task w.r.t. marginal likelihood). Even if the GAN encoding produced stable results, the considerable slowness of finding the latent codes by optimization (in range of minutes per single image) motivates the present approach, as we now need to encode only a small subset of frames and match the camera movement by GP prediction.

Reconstruction mode	Mean±std	Median	LPIPS-Δ
1-by-1 GAN proj. (all f.)	0.33±0.10	0.36	0.154
Sep. kernel (all f.)	0.41±0.12	0.42	0.026
Quat kernel (all f.)	0.41±0.12	0.43	**0.021**
View kernel (all frames)	**0.39±0.13**	**0.41**	0.031
Lin. interp. (first–last)	0.45±0.07	0.46	0.020
Sep. kernel (f–l)	0.44±0.07	0.46	0.024
Quat kernel (f–l)	0.45±0.10	**0.44**	**0.012**
View kernel (f–l)	**0.42±0.08**	**0.44**	0.020

Table 1: LPIPS similarities between ground-truth and frames generated with different methods, center-cropped, using camera runs on 4 face identities ($N = 1570$). Smaller is better.

Noise reduction Given a sequence of images of the same object, we can use the encoder (optimizer) to find the corresponding latent codes. As we decode the codes back to individual images, they are mutually inconsistent (no temporal consistency). The issue may not be clear when visually examining single frames, but it is plain when the frames are combined into a video (see the supplement for video examples). We 'denoise' the sequence of latent codes with GP regression, and decode the new sequential images as video, making it smoother and reducing artifacts. Fig. 4 shows three consecutive input frames from a video and their respective independent GAN reconstructions. Partly due to the tilted angle, the quality and preservation of identity in face reconstructions for independent frames varies. GP regression with our view-aware prior makes the motion smooth and preserves the identity better throughout the video. The smoothness can be measured using the mean difference of the learned perceptual image path similarity metric (LPIPS, Zhang et al. 2018) between consequtive frames, considerably smaller for the GP interpolation using all frames (the LPIPS-Δ in Table 1).

View synthesis Next, we take only a subset of the frames—the extreme case with only a single start and a single end frame (see Fig. 5)—and interpolate the rest of the frames in the latent space by predicting the latent codes, $\mathrm{E}[\mathbf{f}(P_\star) \,|\, \mathcal{D}]$, for unseen views $P_\star$, following the *correlation structure of the original camera movement*. In Fig. 5, we compare to independent frame-by-frame reconstructions. For certain input head poses, the quality is gapped by suboptimal StyleGAN projections (leading to some variation in face alignment). As a baseline, we also linearly interpolate between the first and last frame, which (for apparent reasons) fails to capture the

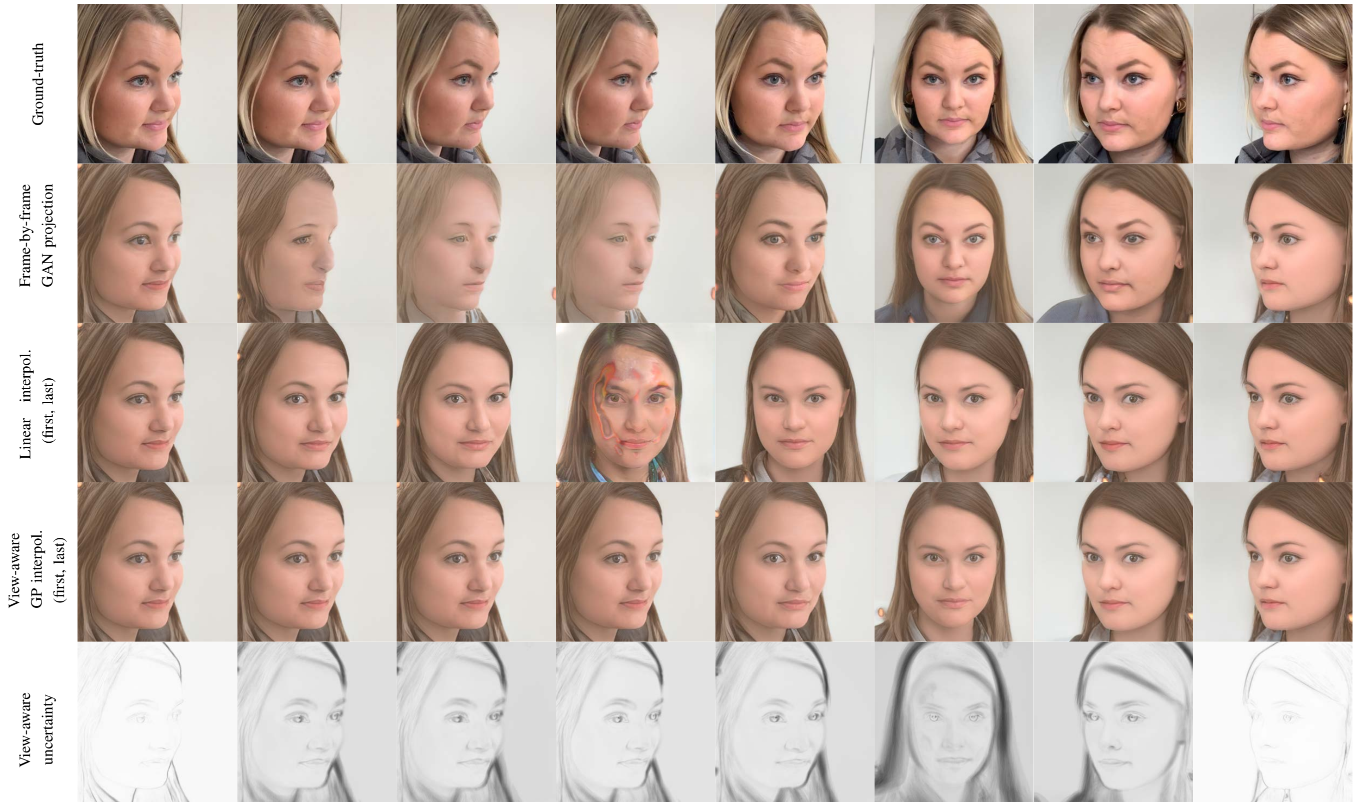

Ground-truth
Frame-by-frame GAN projection
Linear interpol. (first, last)
View-aware GP interpol. (first, last)
View-aware uncertainty

Figure 5: View-aware GP interpolation between two input frames: Row #1: Frames separated by equal time intervals from a camera run. Row #2: Independently GAN reconstructions. Row #3: Linear interpolation of intermediate frames in GAN latent space between first–last frame (note lost azimuth angle). Row #4: Interpolation in GAN latent space between first and last frame by our view-aware GP prior. Row #5: Per-pixel GP posterior uncertainty visualized in the form of marginal standard deviation.

varying camera motion, with mismatches in the head angle. The GP solution with our view prior smoothly matches the view orientation while maintaining the face features. Also, we visualize the frame-wise marginal uncertainty (posterior variance $\mathrm{V}[\mathbf{f}(P_j) \mid \mathcal{D}]$) of the GP predictions as a standard deviation map in image space. We create the maps by drawing 100 samples from the posterior process and calculating the standard deviation over faces. The uncertainty is small in the beginning/end (where the inputs are) and highest towards the part where the linear interpolation has the largest error—showing the practical uncertainty quantification capabilities of the model. We also measure the differences to ground-truth images (LPIPS in Table 1). One expects the direct StyleGAN projection that uses all frames to yield the minimum LPIPS, but it has poor temporal consistency (LPIPS-Δ). The separable and quaternion kernels have it *vice versa*: Their high consistency (low LPIPS-Δ) is irrelevant as it is due to losing the original diversity (increasing direct LPIPS, visuals in the supplement). The start and end frames were selected for reasonable symmetry to fairly compare to linear interpolation. Still, the GP interpolation is clearly superior to the linear case. As expected, although GP interpolation with all frames reduces jitter (see supplementary video), it has less frame-by-frame similarity to the originals than direct projection.

Discussion and Conclusion

We have presented a new GP covariance function to encode *a priori* knowledge about camera movement into computer vision tasks, advocating more principled approaches for inter-frame reasoning in computer vision. We consider this view kernel an important building block for applying Gaussian process priors to many computer vision models. The covariance function itself is simple, yet elegant, and circumvents possible problems related to degeneracy and gimbal lock related to the alternative approaches. The model directly generalizes the standard periodic covariance function to high-dimensional rotations, filling a tangible gap in the existing GP tool set.

To underline the practical importance of our work, we considered real-world applications for the proposed model. Our quantitative experiments showed that the view prior can encode authentic movement and provide a soft-prior for view synthesis. We also showed how the model can be of direct practical value by acting as a camera-motion-aware interpolator. Combining probabilistic models with computer vision tasks come with a promise of better data efficiency (not everything needs to be learned from data, as demonstrated in the comparison and uncertainty quantification.

Code and material related to this paper is available at https://aaltoml.github.io/view-aware-inference.

Acknowledgments

We thank Jaakko Lehtinen and Janne Hellsten (NVIDIA) for the StyleGAN latent space projection script and advice on its usage. Authors acknowledge funding from GenMind Ltd. and Academy of Finland (grant numbers 324345 and 308640). We acknowledge the computational resources provided by the Aalto Science-IT project.

Ethical Impact

Following the breakthroughs of deep neural networks in recent years, broader societal concerns have increasingly shifted from maximizing the accuracy under controlled conditions to aspects such as robustness and explainability. In real-world applications, machine learning systems are expected to generalize despite limited amount of training data, yield principled quantification of uncertainty, and allow for human interpretation of the inference process.

Probabilistic methods provide natural solutions to these requirements. Yet, current Bayesian deep learning approaches fall short of ways to encode *interpretable* priors into models, in which non-parametric priors such as Gaussian processes can help. These tools are widely used in, for instance, finance, navigation, and medical tasks, while computer vision applications have seen less benefit. Our work offers a principled building block that extends the gold standard Gaussian process tooling to allow utilization of Gaussian process priors across a range of computer vision tasks, of which we showcase just a few representative examples. We hope this work inspires computer vision practitioners of a variety of different subdomains to increasingly integrate probabilistic methods in their work, as well as motivate the researchers in probabilistic methods to explore models in computer vision applications.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *International Conference on Computer Vision (ICCV)*, 4432–4441.

Anderson, S.; and Barfoot, T. D. 2015. Full STEAM ahead: Exactly sparse gaussian process regression for batch continuous-time trajectory estimation on SE(3). In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 157–164.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 1613–1622. PMLR.

Bui, T.; Hernández-Lobato, D.; Hernandez-Lobato, J.; Li, Y.; and Turner, R. 2016. Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, 1472–1481. PMLR.

Casale, F. P.; Dalca, A.; Saglietti, L.; Listgarten, J.; and Fusi, N. 2018. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 10369–10380. Curran Associates, Inc.

Chang, A. X.; Funkhouser, T. A.; Guibas, L. J.; Hanrahan, P.; Huang, Q.-X.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012* .

Deisenroth, M. P.; and Rasmussen, C. E. 2011. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 465–472. Omnipress.

Diebel, J. 2006. Representing attitude: Euler angles, unit quaternions, and rotation vectors. *Matrix* 58(15-16): 1–35.

Dutordoir, V.; van der Wilk, M.; Artemev, A.; Tomczak, M.; and Hensman, J. 2020. Bayesian Image Classification with Deep Convolutional Gaussian Processes. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, 1529–1539.

Duvenaud, D. 2014. *Automatic Model Construction with Gaussian Processes*. Ph.D. thesis, Computational and Biological Learning Laboratory, University of Cambridge, Cambridge, UK.

Eleftheriadis, S.; Rudovic, O.; Deisenroth, M. P.; and Pantic, M. 2016. Variational Gaussian process auto-encoder for ordinal prediction of facial action units. In *Asian Conference on Computer Vision (ACCV)*, 154–170. Springer.

Eleftheriadis, S.; Rudovic, O.; and Pantic, M. 2015a. Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing* 24(1): 189–204.

Eleftheriadis, S.; Rudovic, O.; and Pantic, M. 2015b. Multi-conditional latent variable model for joint facial action unit detection. In *IEEE International Conference on Computer Vision (ICCV)*, 3792–3800.

Everingham, M.; Sivic, J.; and Zisserman, A. 2006. "Hello! My name is... Buffy"–Automatic naming of characters in TV video. In *British Machine Vision Conference (BMVC)*.

Featherstone, R. 2014. *Rigid Body Dynamics Algorithms*. New York: Springer.

Gardner, J.; Pleiss, G.; Weinberger, K. Q.; Bindel, D.; and Wilson, A. G. 2018. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 7576–7586. Curran Associates, Inc.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2672–2680. Curran Associates, Inc.

Haasdonk, B.; and Burkhardt, H. 2007. Invariant kernel functions for pattern analysis and machine learning. *Machine Learning* 68(1): 35–61.

Hamsici, O. C.; and Martinez, A. M. 2008. Rotation invariant kernels and their application to shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(11): 1985–1999.

Hartley, R.; and Zisserman, A. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Hennig, P.; Osborne, M. A.; and Girolami, M. 2015. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 471(2179).

Hensman, J.; Durrande, N.; and Solin, A. 2018. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research (JMLR)* 18(151): 1–52.

Hensman, J.; Fusi, N.; and Lawrence, N. D. 2013. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence (UAI)*, 282–290. AUAI Press.

Hou, Y.; Janjua, M. K.; Kannala, J.; and Solin, A. 2021. Movement-induced Priors for Deep Stereo. In *International Conference on Pattern Recognition (ICPR)*.

Hou, Y.; Kannala, J.; and Solin, A. 2019. Multi-view stereo by temporal nonparametric fusion. In *IEEE International Conference on Computer Vision (ICCV)*, 2651–2660.

Jayasumana, S.; Hartley, R.; Salzmann, M.; Li, H.; and Harandi, M. 2013. Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 73–80.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NIPS)*, 5574–5584.

Krauth, K.; Bonilla, E. V.; Cutajar, K.; and Filippone, M. 2017. AutoGP: Exploring the capabilities and limitations of Gaussian process models. In *Uncertainty in Artificial Intelligence (UAI)*. AUAI Press.

Lang, M.; and Hirche, S. 2017. Computationally efficient rigid-body Gaussian process for motion dynamics. *IEEE Robotics and Automation Letters* 2(3): 1601–1608.

Lang, M.; Kleinsteuber, M.; and Hirche, S. 2018. Gaussian process for 6-DoF rigid motions. *Autonomous Robots* 42(6): 1151–1167.

Lázaro-Gredilla, M.; Quiñonero-Candela, J.; Rasmussen, C. E.; and Figueiras-Vidal, A. R. 2010. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research (JMLR)* 11: 1865–1881.

MacKay, D. J. 1998. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences* 168: 133–166.

Matthews, A. G. d. G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrá, P.; Ghahramani, Z.; and Hensman, J. 2017. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research (JMLR)* 18(1): 1299–1304.

Mazzotti, C.; Sancisi, N.; and Parenti-Castelli, V. 2016. A measure of the distance between two rigid-body poses based on the use of platonic solids. In *ROMANSY 21-Robot Design, Dynamics and Control*, 81–89. Springer.

Puzer (GitHub user). 2019. StyleGAN Encoder – Converts real images to latent space. https://github.com/Puzer/stylegan-encoder. GitHub; accessed 2019 Feb 19.

Rasmussen, C. E.; and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. The MIT Press.

Särkkä, S.; Solin, A.; and Hartikainen, J. 2013. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing. *IEEE Signal Processing Magazine* 30(4): 51–61.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.

Solin, A.; Hensman, J.; and Turner, R. E. 2018. Infinite-horizon Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3486–3495. Curran Associates, Inc.

Song, L.; Huang, J.; Smola, A.; and Fukumizu, K. 2009. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 961–968.

Urtasun, R.; Fleet, D. J.; and Fua, P. 2006. 3D people tracking with Gaussian process dynamical models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 238–245.

Wang, K. A.; Pleiss, G.; Gardner, J. R.; Tyree, S.; Weinberger, K. Q.; and Wilson, A. G. 2019. Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems (NeurIPS)*, 14622–14632. Curran Associates, Inc.

Wilson, A. G.; and Nickisch, H. 2015. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, 1775–1784.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.