On the Uncertain Single-View Depths in Endoscopies

Javier Rodríguez-Puigvert, David Recasens, Javier Civera, Rubén Martínez-Cantín

Abstract—Estimating depth from endoscopic images is a pre-requisite for a wide set of AI-assisted technologies, namely accurate localization, measurement of tumors, or identification of non-inspected areas. As the domain specificity of colonoscopies— a deformable low-texture environment with fluids, poor lighting conditions and abrupt sensor motions—pose challenges to multi-view approaches, single-view depth learning stands out as a promising line of research. In this paper, we explore for the first time Bayesian deep networks for single-view depth estimation in colonoscopies. Their uncertainty quantification offers great potential for such a critical application area. Our specific contribution is two-fold: 1) an exhaustive analysis of Bayesian deep networks for depth estimation in three different datasets, highlighting challenges and conclusions regarding synthetic-to-real domain changes and supervised vs. self-supervised methods; and 2) a novel teacher-student approach to deep depth learning that takes into account the teacher uncertainty.

Index Terms—Single-view depth, Bayesian deep networks, depth from monocular endoscopies

I. INTRODUCTION

MONOCULAR depth perception inside the human body is a cornerstone to enable automated assistance tools in medical procedures (e.g., virtual augmentations and annotations, accurate measurements or 3D registration of tools and interest regions) and, in the long run, the full automation of certain procedures and medical robotics. Monocular cameras stand out as very convenient sensors, as they are minimally invasive for in-vivo patients. However, estimating depth from single or multiple views remains challenging. Multi-view approaches show a high degree of accuracy and robustness in natural images [1], [2], but assume certain rigidity, texture and illumination conditions that are not generally fulfilled in in-body images. From a geometric perspective, single-view depth is an ill-posed problem since infinite 3D scenes can explain a single 2D view [3]. In this last case, deep neural networks [4] have shown excellent results in last years [5]–[7]. However, very often, the accuracy and validity of their predictions are taken for granted, which could result in false assumptions that we cannot rely on in such a critical environment as the human body. Uncertainty quantification is a must-have for interpretable and safe artificial intelligence systems.

This work was supported by EndoMapper GA 863146 (EU-H2020), RTI2018-096903-B-I00, BES-2016-078426, PGC2018-096367-B-I00 (FEDER / Spanish Government), DGA-T45 17RFSE (Aragón Government). The authors are with I3A, Universidad de Zaragoza, Spain. {jrp, recasens, jcivera, rmcantin}@unizar.es

Bayesian learning has shown to be a successful tool for uncertainty quantification in adaptive perception and control systems. For high-dimensional deep neural networks, however, Bayesian inference is intractable and even most approximate inference methods are infeasible. In practice, scalable approaches such as deep ensembles [8] and Monte Carlo dropout [9] have shown to produce well-calibrated uncertainties in many computer vision tasks [10].

In this work, we present for the first time the application of Bayesian deep networks for depth prediction into endoscopic images. In Fig. 1 we illustrate the predicted depth and uncertainties of one of our Bayesian models in a real colonoscopy image. Dark/bright colors stands for near/far depths and blue/red stands for low/high uncertainties. Notice the higher depth uncertainties in darker and farther areas and in specular reflections.

Fig. 1: Depth prediction and its uncertainty (epistemic and aleatoric) for a single colonoscopy image. Dark/bright colors stands for near/far depths and blue/red stands for low/high uncertainties. Notice the higher depth uncertainties in darker and farther areas and in specular reflections.
its strengths and limitations in the medical area. We show how
domain changes (synthetic colonoscopy–real colonoscopy–real
laparoscopy) impact depth and uncertainty estimations, as well
as the generalization that uncertainty offers in such models.
Very importantly, we demonstrate that Bayesian models that
perform well in synthetic data can be transferred adequately to
similar domains. Finally, we propose a novel teacher-student
architecture that achieves state-of-the-art results compared to
previous approaches.

II. PRELIMINARIES AND RELATED WORK

A. Bayesian Deep Learning

Bayesian deep learning extends deep learning to the proba-
bilistic setting by using Bayesian learning rules and inference.
The key advantage of Bayesian deep learning is that we can
compute the posterior distribution over the network parameters
and the predictive posterior distribution of any input. Having a
full predictive posterior enables more robust results by explic-
itly representing the uncertainty of every prediction. Bayesian
deep learning also allows to reason about the learning process
and the model capabilities by tracking the source of the
uncertainty. It is important to differentiate the possible types of
uncertainty that we can find in the predictive posterior
distribution. First, the lack of information of the input data
(aleatoric uncertainty) and what the model does not know
about the estimation (epistemic uncertainty).

Aleatoric uncertainty is the encoded random variation be-
tween different inputs, which cannot be reduced by additional
training data. In regression tasks, aleatoric uncertainty is
commonly assumed to be homoscedastic, that is, the predictive
uncertainty is independent of the input data. Heteroscedastic
uncertainty requires a model that is able to predict the
variability of uncertainty for different inputs. In this work,
we use a common backbone to predict depth and aleatoric
uncertainty estimates from two heads of the neural network.
Epistemic uncertainty represents the lack of knowledge about
the best prediction model. In contrast to aleatoric uncertainty,
using more data during training might reduce the epistemic
uncertainty if the model is able to incorporate the new data
and distill the information. This is specially relevant in large
capacity models such as deep neural networks. Epistemic
uncertainty can be used for detecting out-of-distribution data
and it is related to the models’ ability to generalize.

In the case of Bayesian deep learning, the epistemic uncer-
tainty is estimated directly from the posterior model parame-
ters by using the Bayes rule: $p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$, based on
a prior distribution $p(\theta)$ over the network parameters $\theta$ and
its likelihood model $p(D|\theta)$ over the dataset $D$. Approximate
Bayesian inference is required due to the intractability of the
normalizing constant $p(D)$, which requires the computation of
a complex, high-dimensional integral. In this work, we focus
on deep ensembles as approximate Bayesian inference due to
their success approximating the epistemic uncertainty.

A deep ensemble is a set of deep neural networks with
randomization [8]. Such randomization coming in this case
from both the initial weight values and the selection order of
data points (or minibatches) during training. Thus, each
network from the ensemble can be trained in parallel. In
the original work of Lakshminarayanan et al. [8] they use a
network with two output channels, one for the predicted mean
$\mu(x)$ and other for the variance $\sigma^2(x)$, as discussed before.
Those channels can be combined in a single loss based on a
Gaussian likelihood, more concretely, they update the negative
log likelihood

$$\mathcal{L}_{GL}(x, y) = \frac{||y - \mu(x)||_2}{\sigma^2(x)} + \log \sigma^2(x)$$

being $x$ a data sample and $y$ its label. However, they
acknowledge that the maximum likelihood approach might
result in overfitting and that the Gaussian distribution might be
too restrictive. We address the overfitting problem by setting a
prior on the weights and minimizing a negative log a posteriori
distribution

$$\mathcal{L}_{MAP} = \mathcal{L}_{GL} + \mathcal{L}_{prior}$$

where we assume a Gaussian prior on the model weights which
results in: $\mathcal{L}_{prior} = ||\theta||^2$. We also consider a likelihood with
heavier tails, such as the Laplace distribution, which has shown
to be more effective for depth estimation [12], [13]

$$\mathcal{L}_{LL}(x, y) = \frac{||y - \mu(x)||_1}{\sigma(x)} + \log \sigma(x)$$

Deep ensembles, as other sample approximations to neural
networks such as Monte Carlo dropout [9], allow to com-
pute more accurate predictions using Bayesian model average
through computing the predictive posterior by marginalization
of the model weights $\theta$ trained on dataset $D$. That is $p(y|x) = \int p(y|x, \theta)p(\theta|D)d\theta$. Although strictly speaking, the deep
ensemble samples are not generated from the full posterior
distribution $p(\theta|D)$. The fact that they start from random
locations guarantees that they capture the multimodality of
the predictive posterior distribution [10]. Furthermore, by
optimizing each sample on the MAP loss we guarantee that
each sample has high probability, which is fundamental for
deep models where the weight space $\theta$ is huge, but the number of
samples $M$ must remain small for computational tractability.

B. Single-View Depth Learning

Deep learning has demonstrated a remarkable performace
for single-view depth prediction. The first works in this area
[14], [15] used ground truth depth to learn supervisedly.
Later research contributed mainly by proposing architectural
innovations [16]–[19]. All these methods rely on accurate
ground truth labels at training, which is not trivial in many
application domains. Self-supervision without depth labels
was achieved by enforcing multi-view photometric consistency
during training [20]–[23]. Other works, e.g. Li et al. [24],
trained depth networks with the multi-view reconstructions
computed by Structure from Motion (SFM) pipelines, mainly
COLMAP [25], [26].

In the medical domain, supervised learning was
addressed by Visentini et al. [27] using an autoencoder and in
Shen et al. [28] with a generative adversarial network (GAN),
both using ground truth from phantom models. Other works
also based on GANs were trained with synthetic models [29]–[32]; and Cheng et al. [33] added a temporal consistency loss. GANs have shown an extraordinary performance on many vision tasks [34]–[36]. However, this architecture is in disuse for depth learning since GANs produce depth maps that prioritize realism over accuracy. Self-supervised learning is a natural choice for real medical endoscopes to overcome the lack of depth labels on the target domain [37]–[39]. Although depth or stereo sensors are not common for in-vivo procedures, several works are trained with real stereo endoscopies [40]–[42]. Others train in phantoms [43] and with synthetic [44], [45], facing the risk of not generalizing in the target domain (in this paper we study the limits of such generalization).

Self-supervised approaches in the medical domain, and we extend Poggi et al. by including the self-supervised monocular loss. Poggi et al. [12] are the first ones to propose a heteroscedastic model for the aleatoric uncertainty, and aleatoric variance \( \hat{\sigma}_a, m \) of a deep model with per pixel depth labels \( d, \sigma_d \) from stereo, \( d_t \) truth depth, \( \sigma_d_t \) corresponds to \( \sigma_d^2 \). For these three supervised learning alternatives, we model our prediction with a Gaussian distribution \( N(\hat{d}, \hat{\sigma}_d^2) \), \( \hat{d} \) results averaging all ensembles depth output and \( \hat{\sigma}_d^2 \) results of joining aleatoric and epistemic uncertainty.

\[
\hat{d} = \frac{1}{M} \sum_{m=0}^{M} \hat{d}_m, \quad \hat{\sigma}_d^2 = \frac{1}{M} \sum_{m=0}^{M} \hat{\sigma}_d^2, \quad \hat{\sigma}_e^2 = \frac{1}{M} \sum_{m=0}^{M} (\hat{d} - \hat{d}_m)^2.
\]

where \( \lfloor \cdot \rfloor \) is the sampling operator and \( j \in \Omega \) refers to the 2-dimensional pixel coordinates in the image domain \( \Omega \).

The per-pixel depth labels \( d \) vary in our endoscopic sequences. Therefore, our depth annotations \( d \) encompasses three different labels: when the depth values are perfect ground truth depth, \( d \) corresponds to \( d^{GT} \); if the depth labels come from stereo, \( d \) corresponds to \( d^{ST} \); and if the depth labels are from a 3D up-to-scale reconstruction using SfM [1], \( d \) corresponds to \( d^{SM} \). For these three supervised learning alternatives, we model our prediction with a Gaussian distribution per pixel \( N(\hat{d}, \hat{\sigma}_d^2) \), that we predict using a deep ensemble. Our ensemble is composed by \( M \) networks, each of them trained separately starting from a different random seed for its weights. The output of the \( m^{th} \) ensemble follows a Gaussian distribution \( N(\hat{d}_m, \hat{\sigma}_a, m^2) \), of which \( \hat{d}_m \) is the pixel-wise depth prediction and \( \hat{\sigma}_a, m^2 \) describes the aleatoric uncertainty (see Fig. 3). We obtain the mean depth of the ensemble \( \hat{d} \) and its epistemic uncertainty \( \hat{\sigma}_e^2 \) using the mean and variance of \( \hat{d}_m \) of all its networks. The total uncertainty \( \hat{\sigma}_d^2 \) combines the aleatoric and epistemic uncertainty which results from the law of total variance.

\[
\hat{d} = \frac{1}{M} \sum_{m=0}^{M} \hat{d}_m, \quad \hat{\sigma}_d^2 = \frac{1}{M} \sum_{m=0}^{M} \hat{\sigma}_e^2, \quad \hat{\sigma}_e^2 = \frac{1}{M} \sum_{m=0}^{M} (\hat{d} - \hat{d}_m)^2.
\]

\[\text{IV. SELF-SUPERVISED LEARNING USING DEEP ENSEMBLES}\]

Self-supervised approaches for depth prediction aim at learning without depth labels, i.e. our dataset is composed of input images \( I \).
only by $N$ RGB images $D = \{\mathcal{I}_1, \ldots, \mathcal{I}_N\}$. See Fig. 4 for an illustration. As before, our network architecture is an encoder-decoder with skip connections [7]. Differently from supervised approaches, however, our self-supervision uses two networks: the first one learning depth and a photometric uncertainty parameter $\hat{u}$, and the second one learning to predict relative camera motion between two frames. The relative motion is needed by the photometric consistency, which acts as supervisory signal, as explained in the next paragraphs. In this case, we use a pseudo-Laplacian likelihood for the loss function. That is, for an instance $m$ of a deep ensemble, and per image $\mathcal{I}$:

$$\mathcal{L}_{LL,m} = \frac{1}{w_h} \sum_{j \in \Omega_I} \left( \frac{F_p[j]}{u_m[j]} + \log \hat{u}_m[j] \right)$$  \hspace{1cm} (2)

This pseudo-likelihood is based on a photometric residual $F_p$ and the uncertainty parameter $\hat{u}_m$. As the reprojection estimation is an ill-posed problem, the prior loss incorporates an edge-aware smoothness term $F_s$ which regularizes the problem (see Godard et al. [23] for more details about $F_s$). Thus, for a network $m$, the loss $\mathcal{L}_m$ becomes the sum of the pseudo-likelihood $\mathcal{L}_{LL,m}$ plus the prior term $\mathcal{L}_{prior} = ||\theta||^2 + \lambda_u F_s[j]$, where $\lambda_u$ calibrates the effect of the smoothness in terms of the reprojection uncertainty. The photometric residual $F_p[j]$ of pixel $j$ in a target image $\mathcal{I}_t$ is the minimum between the warped images $\mathcal{I}_{t'\rightarrow t}$ from the previous and posterior images $\mathcal{I}_{t'}$ to the target one $\mathcal{I}_t$ of the sum of the photometric reprojection error and Structural Similarity Index Measure (SSIM) [52]

$$F_p[j] = \min((1 - \alpha)||\mathcal{I}_t[j] - \mathcal{I}_{t'\rightarrow t}[j]||_1 + \frac{\alpha}{2}(1 - \text{SSIM}(\mathcal{I}_t, \mathcal{I}_{t'\rightarrow t}, j)))$$  \hspace{1cm} (3)

being $\alpha \in [0, 1]$ the relative weight of the addends; and $\mathcal{I}_t[j]$ and $\mathcal{I}_{t'\rightarrow t}[j] = \mathcal{I}_{t'}[j']$ the color values of pixel $j$ of the target image $\mathcal{I}_t$ and of the warped image $\mathcal{I}_{t'\rightarrow t}$. In order to obtain this latter term we warp every pixel $j$ from the target image domain $\Omega_t$ to that of the source image $\Omega_{t'}$ using

$$j' = \pi(\mathbf{R} t_{i',i}; \pi^{-1}(j, \hat{d}_t[j]) + t_{i',i})$$  \hspace{1cm} (4)

where $\mathbf{R} t_{i',i} \in \text{SO}(3)$ and $t_{i',i} \in \mathbb{R}^3$ are the rotation and translation from $\Omega_t$ to $\Omega_{t'}$, and $\pi$ and $\pi^{-1}$ the projection and back-projection functions (3D point to pixel, and vice versa).

As the true scale is unobservable by the multi-view models in self-supervised losses, it is convenient to predict pixel-wise disparity $\hat{d}_t[j]$ (distance in pixels between a point $j$ in a target image $\mathcal{I}_t$ and the same point but in a source image $\mathcal{I}_{t'}$). This disparity can be converted to scaled depth $d_t[j] = 1/(\hat{d}_t[j] + b)$.

Finally, we can obtain the average prediction of the ensemble by model averaging as in the supervised case

$$d = \frac{1}{M} \sum_{m=1}^{M} \hat{d}_m, \quad \hat{u}^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{u}_m^2, \quad \hat{\sigma}^2 = \frac{1}{M} \sum_{m=1}^{M} (\hat{d} - \hat{d}_m)^2$$

Note that, in this case we cannot obtain aleatoric uncertainty for the depth estimate, because the uncertainty parameter estimated by the network $\hat{u}^2$ is associated with the photometric residual [12]. Therefore, we cannot obtain the total uncertainty.

**V. SELF-SUPERVISED LEARNING USING UNCERTAIN TEACHER-STUDENT**

In the endoscopic field, there are accurate depth labels only from RGB-D endoscopes (which are scarce) or synthetic data (that is affected by domain change). Teacher-student approaches [12] train a student network using as supervisory labels the predictions of another network which acts as the teacher. The teacher is trained beforehand, optionally with supervised data, in a similar domain to the target one. As key advantages, teacher-student architectures avoid photometric losses and pose regression networks, which might be unstable and have low accuracy, and the uncertainty can be learned in depth units. A relevant limitation in our case is that the teacher labels are noisy and affected by the bias of the training domain.

In supervised learning, the labels $d$ used in training are generally assumed to come from a perfect oracle. However, in our novel teacher-student architecture (illustrated in Fig. 5), the teacher is a deep ensemble, and hence the label comes
from the predictive posterior distribution of the teacher \(d \sim \mathcal{N}(\hat{d}, \sigma_T^2)\), where \(\sigma_T^2\) is the total teacher variance (aleatoric and epistemic). Thus, the likelihood distribution must incorporate both the teacher and student distributions, which is used in the training loss for the student. As before, the loss is based on a Laplacian likelihood.

\[
\mathcal{L}_{LL,m} = \frac{1}{w \cdot h} \sum_{j \in \Omega_i} \left( \frac{||\hat{d}_T[j] - \hat{d}[j]||_1}{\sigma_m[j]} + \log \hat{\sigma}_m[j] \right)
\]

where the variance for each pixel \([j]\) is the sum of the teacher predictive variance, both aleatoric and epistemic, and the aleatoric variance predicted by the student \(\hat{\sigma}_m^2 = \sigma_T^2 + \sigma_a^2\).

VI. EXPERIMENTAL RESULTS

Datasets. We present results in three different datasets. The first one is the synthetic data (thus with ground truth) generated by Rau et al. [32], containing 16,016 RGB images rendered from a 3D model of the colon in 15 different texture and illumination conditions. The second one, the dataset recorded within the EndoMapper project\(^1\) containing real monocular colonoscopies. Finally, we also present results on the Hamlyn Centre stereo dataset [53]–[56], containing images of 22 laparoscopic procedures of in-vivo and ex-vivo animal organs. Both real datasets present challenges such as tissue motion, surgical tools and strong specular reflections, aggravated in the EndoMapper sequences by more abrupt motions and liquid presence.

\(^1\)http://endomapper.eu/

Metrics. We report the metrics that are standard in the field. For depth errors [5]: Absolute Relative difference: \(1/w \cdot h \sum_{j \in \Omega_i} |d(j) - \hat{d}(j)| / d(j)\), Square Relative difference: \(1/w \cdot h \sum_{j \in \Omega_i} (d(j) - \hat{d}(j))^2 / d(j)^2\), Root Mean Square Error: \((1/w \cdot h \sum_{j \in \Omega_i} (d(j) - \hat{d}(j))^2)^{1/2}\), RMSE log: \((1/w \cdot h \sum_{j \in \Omega_i} (\log d(j) - \log \hat{d}(j))^2)^{1/2}\) and \(\delta < 1.25^i\) with \(i \in \{1, 2, 3\}\): \(\delta < 1.25^1 = 1/w \cdot h \sum_{j \in \Omega_i} \max(d(j), \hat{d}(j)) < 1.25^1\). Regarding uncertainty, we report the Area Under the Sparsification Error curve (AUSE) as proposed in Gustafsson et al. [10] and the Area Under the Calibration Error (AUCE) [57]. The reader is referred to the sources for details. The AUSE indicates the quality of the uncertainty estimation by comparing the model (specifically the ordering of the estimated predictive uncertainty), and the oracle (the ordering of the true prediction error in terms of RMSE). This metrics shows the extent so which the estimated uncertainty serves to sort the prediction errors, hence being a relative metric. The AUCE is a measure of the uncertainty calibration. To compute it, we define intervals of confidence level \(p \in [0, 1]\) using the cumulative density function of our predicted distribution. The calibration will be perfect when the ratio of the prediction intervals which covers the target \(\hat{p}\) is identical to the confidence level \(p\). The AUCE is specifically the area between the absolute error curve and a perfect calibration \(|p - \hat{p}|\) for a predefined number of confidence intervals.

A. Synthetic colon dataset

Goals. In this experiment, we have two objectives. First, evaluating three different training alternatives for depth learning: supervision with ground truth depth, supervision with SfM and photometric self-supervision. And secondly, we aim to clarify if SfM can be used for benchmarking depth learning methods in real data, where ground truth is not available.

Experimental details. In all SfM-related experiments we use COLMAP [26], a state-of-the-art SfM framework that estimates a quasi-dense reconstruction from a set of images.
In Fig. [6a] we show a 3D reconstruction using the synthetic colon images. Note that COLMAP, as every SfM framework, might not be able to register all images, being the main reason the deficient texture. For a fair evaluation of supervised and self-supervised approaches, we only use the subset of images that COLMAP used for the reconstruction. In this experiment, this amounts to 6,550 images for training and 720 images for testing. From empirical analysis, we observed that training more than 18 networks per ensemble does not improve the performance significantly, so we use this number in our experiments. We initialize randomly the weights of each network in the ensemble using a Gaussian distribution $\mathcal{N}(0, 10^{-2})$ (also in the experiments of next subsections).

1) Ground truth depth supervision: Following the approach in Section III we train each ensemble with perfect depth labels $d^{GT}$ using the loss defined in Eq. [1].

2) SfM depth supervision: We train each network using the SfM depth $d^{SfM}$ as supervisory signal. Although the SfM reconstruction is rather accurate, it may be affected by spurious data points and holes with no depth values.

3) Photometric self-supervision: The networks are trained directly from monocular images without the need of depth labels, as explained in Section [4]. As the aleatoric and epistemic uncertainties cannot be added in this case, our results only contain the epistemic one.

Results. Table [I] shows the metrics for the depth error and its uncertainty, evaluated with the ground truth labels $d^{GT}$. Table [II] shows these metrics evaluated using the SfM depth $d^{SfM}$.

In Table [I] we see that the AUSE metric is quite different between the two tables, validating the use of SfM to benchmark single-view methods in the absence of ground truth. Finally, note that the metrics for SfM labels are significantly smaller in Table [II] compared to Table [I] and even smaller than those in the last row in Table [I].

In the absence of ground truth, the metrics for SfM labels suggests that this model is overfitting, and hence such numbers should not be trusted. Note finally that the AUSE metric is quite different between the two tables, and should not be trusted either in following experiments.

B. EndoMapper dataset

Goals. Our objective in these experiments is to evaluate the performance of Bayesian depth networks in real data. Specifically, we first evaluate the performance of the models trained with ground truth depth $d^{GT}$ and SfM depth $d^{SfM}$ in the synthetic data of the previous section. We also show results from the self-supervised approach trained on the real images. Finally, we demonstrate that our novel teacher-student loss can combine the strengths of both data sources, synthetic and real, and outperforms all the other models.

Experimental details. Notice that the viewpoint might change abruptly during a real colonoscopy and the images might be saturated or blurry. There might also appear a considerable amount of liquid in some frames. For this reasons, we pre-process the dataset removing images with partial or total visibility issues. We use 6,912 images out of the 14,400 images in the complete endoscopic procedure. Similarly to the

| Approach          | ABErel | SBErel | RMSE [mm] | RMSErel | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | AUCE ↓ | AUSE ↓ |
|-------------------|--------|--------|-----------|---------|----------------|------------------|-----------------|-------|--------|
| GT labels         | 0.103  | 0.382  | 4.129     | 0.119   | 0.985          | 0.997            | 0.999           | 0.264 | 0.300  |
| SfM labels        | 0.172  | 2.567  | 7.410     | 0.270   | 0.853          | 0.939            | 0.963           | 0.055 | 0.127  |
| Self-supervised   | 0.179  | 1.774  | 7.600     | 0.244   | 0.792          | 0.939            | 0.973           | 0.166 | 0.111  |

| Approach          | ABErel | SBErel | RMSE [mm] | RMSErel | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ | AUCE ↓ | AUSE ↓ |
|-------------------|--------|--------|-----------|---------|----------------|------------------|-----------------|-------|--------|
| GT labels         | 0.136  | 2.113  | 10.643    | 0.323   | 0.893          | 0.938            | 0.954           | 0.227 | 0.699  |
| SfM labels        | 0.091  | 1.353  | 8.166     | 0.218   | 0.916          | 0.953            | 0.970           | 0.043 | 0.045  |
| Self-supervised   | 0.165  | 2.262  | 10.357    | 0.288   | 0.800          | 0.932            | 0.960           | 0.164 | 0.331  |
| Baseline (GT)     | 0.128  | 2.598  | 11.362    | 0.312   | 0.880          | 0.934            | 0.951           | -     | -      |
As explained in section 5. As a baseline, we also train the teacher network in the forward pass, using the same input image as the student, and the student has the best metrics, supporting our initial hypothesis that including the teacher’s uncertainty in the loss is beneficial.

The first row shows the performance of the teacher trained with GT labels in synthetic data, the second row the one of the teacher-student methods. Table IV shows the metrics for the teacher-student methods. The first row shows the performance of the teacher trained with GT labels in synthetic data, the second row the one of the teacher-student methods. For the student network.

Table IV shows the metrics for the teacher-student methods. The first row shows the performance of the teacher trained with GT labels in synthetic data, the second row the one of the teacher-student methods. For the student network.

### TABLE V: Depth and uncertainty metrics for the Hamlyn dataset.

| Test seq. | Train data | Approach            | Per-image scaling | Abg_{log} ↓ | S_{log} ↓ | RMSE↓ | RMSE_{log} ↓ | δ < 1.25↑ | δ < 1.25^2↑ | δ < 1.25^3↑ | AUCE↓ | AUCE↓ |
|-----------|------------|---------------------|-------------------|-------------|-----------|-------|-------------|-----------|-------------|-------------|-------|-------|
| 1         | Synthetic  | GT labels           | ✓                 | 0.186       | 5.123     | 20.834| 0.265       | 0.654     | 0.918       | 0.981       | 0.150 | 0.307 |
|           | Synthetic  | SIM labels          | ✓                 | 0.192       | 4.719     | 19.969| 0.241       | 0.645     | 0.938       | 0.994       | 0.489 | 0.212 |
|           | Hamlyn A   | Stereo labels       | ✓                 | 0.149       | 1.295     | 7.322 | 0.240       | 0.796     | 0.962       | 1.000       | 0.474 | 0.365 |
|           | Hamlyn A   | Self-supervised     | ✓                 | 0.116       | 1.857     | 12.555| 0.154       | 0.869     | 0.992       | 0.998       | 0.481 | 0.402 |
| 4         | Synthetic  | GT labels           | ✓                 | 0.186       | 2.135     | 9.453 | 0.250       | 0.674     | 0.915       | 0.988       | 0.071 | 0.279 |
|           | Synthetic  | SIM labels          | ✓                 | 0.192       | 4.719     | 19.969| 0.241       | 0.645     | 0.938       | 0.994       | 0.489 | 0.212 |
|           | Hamlyn B   | Stereo labels       | ✓                 | 0.064       | 0.227     | 3.195 | 0.072       | 1.000     | 1.000       | 1.000       | 0.358 | 0.188 |
|           | Hamlyn B   | Self-supervised     | ✓                 | 0.065       | 0.296     | 3.732 | 0.085       | 0.977     | 1.000       | 1.000       | 0.478 | 0.441 |
| 19        | Synthetic  | GT labels           | ✓                 | 0.335       | 12.273    | 25.409| 0.366       | 0.532     | 0.777       | 0.907       | 0.218 | 0.079 |
|           | Synthetic  | SIM labels          | ✓                 | 0.365       | 14.131    | 27.473| 0.389       | 0.847     | 0.949       | 0.996       | 0.493 | 0.073 |
|           | Hamlyn C   | Stereo labels       | ✓                 | 0.074       | 1.003     | 9.038 | 0.121       | 0.955     | 0.992       | 0.996       | 0.128 | 0.127 |
|           | Hamlyn C   | Self-supervised     | ✓                 | 0.211       | 10.871    | 23.962| 0.276       | 0.737     | 0.942       | 0.974       | 0.468 | 0.157 |
| 20        | Synthetic  | GT labels           | ✓                 | 0.220       | 4.409     | 14.804| 0.282       | 0.805     | 0.888       | 0.971       | 0.147 | 0.191 |
|           | Hamlyn D   | Stereo labels       | ✓                 | 0.096       | 0.814     | 6.720 | 0.130       | 0.900     | 1.000       | 1.000       | 0.140 | 0.189 |
|           | Hamlyn D   | Self-supervised     | ✓                 | 0.139       | 4.905     | 15.093| 0.195       | 0.869     | 0.982       | 0.988       | 0.483 | 0.077 |

### TABLE III: Depth and uncertainty metrics in the EndoMapper dataset.

| Approach       | Train data | Abg_{log} ↓ | S_{log} ↓ | RMSE↓ | RMSE_{log} ↓ | δ < 1.25↑ | δ < 1.25^2↑ | δ < 1.25^3↑ | AUCE↓ | AUCE↓ |
|----------------|------------|-------------|-----------|-------|-------------|-----------|-------------|-------------|-------|-------|
| GT labels      | Synthetic  | 0.253       | 1.418     | 6.055 | 0.340       | 0.612     | 0.857       | 0.943       | 0.086 | 0.145 |
| SfM labels     | Synthetic  | 0.241       | 1.271     | 5.700 | 0.321       | 0.632     | 0.874       | 0.952       | 0.169 | 0.146 |
| Self-supervised| Endomapper | 0.343       | 1.376     | 4.603 | 0.440       | 0.449     | 0.716       | 0.872       | 0.329 | 0.265 |

### TABLE IV: Depth and uncertainty metrics in the EndoMapper dataset.

1) Ground truth depth and SfM supervision with domain change: We evaluate the models trained supervisedly in synthetic colonoscopy data in real colonoscopy images.

2) Photometric self-supervision: We train 18 ensembles using the self-supervised approach in Section 2. This setting is challenging due to reflections, fluids, and deformations, all of them aspects that are not modeled in the photometric reprojection model of self-supervised losses. As aleatoric and epistemic uncertainties do not refer to the same units and cannot be added, we only report the epistemic uncertainty that corresponds directly to depth prediction.

3) Teacher-student self-supervision: We use the deep ensemble trained in the synthetic dataset with GT labels to teach depth and uncertainty labels to the network referred to as student. The teacher predicts depth and uncertainty labels in the forward pass, using the same input image as the student, as explained in section 5. As a baseline, we also train the teacher-student architecture of [12].

Results. Table III shows the depth and uncertainty metrics for models trained with supervision on synthetic data, and self-supervised models trained in real colonoscopy images. The models supervised by synthetic data present better performance than the self-supervised ones. Self-supervised approaches indeed struggle in our application due to specularities and saturated images. Regarding uncertainty, learning from GT labels leads to a better correlation between depth errors and uncertainties (see Fig. 8). Note that the aleatoric uncertainty captures light reflection and depth discontinuities in supervised learning. On the other hand, the epistemic uncertainty stands out the deeper areas. From these results, we can conclude that the domain change from synthetic to real colon images is not significant. Models trained on synthetic data generalize to real images and outperform models trained with self-supervision on the target domain, due to the challenges mentioned in the previous paragraphs. In the table, we observe that the model trained with SfM depth outperforms the one trained with GT depth. However, the analysis from the previous section shows that such numbers should not be trusted, as the model trained with SfM depth might be overfitting.

Table IV shows the metrics for the teacher-student methods. The first row shows the performance of the teacher trained with GT labels in synthetic data, the second row the one of the teacher-student method from [12], and the third row the one of our uncertain teacher-student method. Our uncertain teacher-student has the best metrics, supporting our initial hypothesis that including the teacher’s uncertainty in the loss is beneficial for the student network.

C. Hamlyn dataset

Goals. In this experiments, we evaluate supervised and self-supervised approaches in the Hamlyn dataset. In particular, we use the stereo depth labels d_{ST} and rectified images from Recasens et al. [39]. We compare learning methods trained in the target domain (Hamlyn) against those trained with domain change (in synthetic colonoscopies).

Experimental details. We divided the Hamlyn dataset into four splits, each of them composed of videos with similar resolution and camera intrinsics. For each test sequence, the model was trained in the rest of the sequences of the same split. The splits, called Hamlyn A, B, C and D, have 15,624,
Fig. 8: Qualitative depth and uncertainty examples of domain change (supervised Learning, supervised Learning SfM), self supervised Learning and uncertain teacher-student in real colonoscopy images. Note that the aleatoric uncertainty in the self-supervised case is related to the photometric residual and not the depth error, which explains the differences with respect to other aleatoric uncertainties. Dark/bright colors stands for near/far depths and low/high errors; and blue/red stands for low/high uncertainties.

Fig. 9: Depth predictions, aleatoric, epistemic and total uncertainties, and depth errors in a Hamlyn image using four different approaches. Note that the aleatoric uncertainty in the self-supervised case is related to the photometric residual and not the depth error, which explains the differences with respect to other aleatoric uncertainties. Dark/bright colors stands for near/far depths and low/high errors; and blue/red stands for low/high uncertainties.

Results We show qualitative and quantitative results in Fig. 9 and Table V. All the approaches, except for stereo supervision, use the per-image scaling defined previously. Even those trained with synthetic ground truth need it here, since learning in a different domain induces a depth bias that will be reflected in the metrics.

We can observe that models trained on the synthetic colon are less accurate than those with stereo supervision in the
target domain: predictions are less smooth and erroneous in textures not present at training, such as veins or organs intersections. In addition, such models are biased to the tubular shape of the colon and prone to predict deeper depths in certain areas. However, they still predict correctly which zones are closer and further. The aleatoric uncertainty concentrates in veins, tools and specularities for models supervised by GT or SfM in the synthetic domain, since these elements did not appear in the training data. The epistemic uncertainty captures the same aspects in a smoothed fashion, i.e., regions with small veins are covered with a diffuse epistemic uncertainty area. The epistemic uncertainty is generally higher in deeper zones, while the aleatoric one is not depth-dependent. The epistemic uncertainty is in general higher than the aleatoric one. This is relevant, as the epistemic uncertainty can be reduced either improving the training procedures or adding more data (the aleatoric one cannot be reduced). Finally, observe in the figure that the total uncertainty has a high correlation with the predicted depth error: pixels with higher depth errors have, in general, higher uncertainty, which indicates qualitatively that uncertainty predictions are trustable.

Regarding the domain change, both approaches—supervised with synthetic and SfM labels—perform similarly in terms of depth metrics but SfM supervision shows worse uncertainty metrics. This suggests that Bayesian losses allow networks to learn from noisy labels, absorbing the irregularities—which remains reflected in the AUCE, meaning the network is less confident in general—, and to achieve high-quality depth predictions. However, the domain change is noticeable here, and models trained in colonoscopy synthetic data show a worse generalization in Hamlyn than in real colonoscopies. This leads to supervised networks trained in synthetic colonoscopy data performing worse than the ones self-supervised in the target domain, in contrast with the results in real colonoscopies in the previous section. As observed both qualitative and quantitative, learning with depth labels from the target domain outperforms photometric self-supervision with the exception of test sequence 1, for which the stereo baseline is not big enough to capture the parallax effectively.

**VII. Conclusion**

Any system that requires depth predictions from color images can benefit from uncertainty estimates, in order to obtain robust, explainable and dependable assistance and decisions. In this paper, we have explored for the first time supervised and self-supervised approaches for depth and uncertainty estimations from single-views of endoscopies. From our experimental results, we can extract several conclusions. Firstly, when ground truth depth labels are available, using them as supervisory labels outperforms self-supervised learning and results in better calibrated models.

Secondly, approaches based on photometric self-supervision and on SfM supervision coexist in the literature and there is a lack of analysis and results showing which type is more convenient. In our study we validate the use of SfM reconstructions to rank single-view depth learning approaches and show up to what extent such evaluations should be trusted.

Thirdly, our experiments show that models trained in synthetic colonoscopy data are able to generalize to real colonoscopy images. We also show that this does not occur when the domain gap is too large, as it happens between synthetic colonoscopies and real laparoscopies.

Finally, we propose a novel teacher-student architecture that incorporates the teacher uncertainty in the loss, and we demonstrate that it improves the depth prediction accuracy over previous teacher-student architectures.

**References**

[1] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[2] R. Mur-Artal and J. D. Tardós, “Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras,” *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[3] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003.

[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *arXiv preprint arXiv:1406.2283*, 2014.

[6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2013.

[7] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” *Advances in Neural Information Processing Systems*, vol. 2017-December, no. Nips, pp. 6404–6414, 2017.

[9] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[10] F. K. Gustafsson, M. Danelljan, and T. B. Schön, “Evaluating scalable bayesian deep learning methods for robust computer vision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 318–319.

[11] E. Brachmann, M. Humenberger, C. Rother, and T. Sattler, “On the limits of pseudo ground truth in visual camera re-localisation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6218–6228.

[12] M. Poggio, F. Alloitt, F. Tosi, and S. Mattoccia, “On the uncertainty of self-supervised monocular depth estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3227–3237.

[13] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?”, *Advances in Neural Information Processing Systems*, vol. 2017-December, no. Nips, pp. 5575–5585, 2017.

[14] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. of the 28th Conference on Neural Information Processing Systems*, 2014.

[15] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *IEEE international conference on computer vision*, 2015.

[16] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *Fourth international conference on 3D vision*, 2016.

[17] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep @in-proceedings@eigen2015predicting, title=Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, author=Eigen, David and Fergus, Rob, booktitle=Proceedings of the IEEE international conference on computer vision, pages=2650–2658, year=2015 ordinal regression network for monocular depth estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
