

Uncertainty Estimates and Multi-Hypotheses Networks for Optical Flow

Eddy Ilg*, Özgün Çiçek*, Silvio Galli*, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox

University of Freiburg, Germany

{ilg,cicek,gallisso,kleinaa,makansio,fh,brox}@cs.uni-freiburg.de

Abstract. Optical flow estimation can be formulated as an end-to-end supervised learning problem, which yields estimates with a superior accuracy-runtime tradeoff compared to alternative methodology. In this paper, we make such networks estimate their local uncertainty about the correctness of their prediction, which is vital information when building decisions on top of the estimations. For the first time we compare several strategies and techniques to estimate uncertainty in a large-scale computer vision task like optical flow estimation. Moreover, we introduce a new network architecture and loss function that enforce complementary hypotheses and provide uncertainty estimates efficiently with a single forward pass and without the need for sampling or ensembles. We demonstrate the quality of the uncertainty estimates, which is clearly above previous confidence measures on optical flow and allows for interactive frame rates.

Keywords: Convolutional Neural Networks, Optical Flow Estimation, Uncertainty Estimation

1 Introduction

Recent research has shown that deep networks typically outperform handcrafted approaches in computer vision in terms of accuracy and speed. Optical flow estimation is one example: FlowNet [8,14] yields high accuracy optical flow at interactive frame rates, which is relevant for many applications in the automotive domain or for activity understanding.

A valid critique of learning-based approaches is their black-box nature: since all parts of the problem are learned from data, there is no strict understanding on how the problem is solved by the network. Although FlowNet 2.0 [14] was shown to generalize well across various datasets, there is no guarantee that it will also work in different scenarios that contain unknown challenges. In real-world scenarios, such as control of an autonomously driving car, an erroneous decision can be fatal; thus it is not possible to deploy such a system without information about how reliable the underlying estimates are. We should expect an additional

* equal contribution

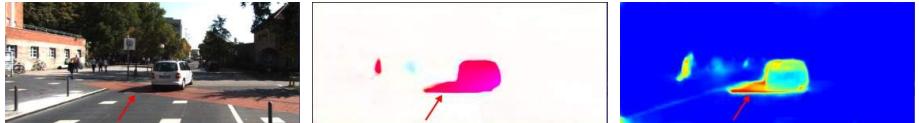


Fig. 1: Joint estimation of optical flow and its uncertainty. **Left:** Image from a KITTI 2015 sequence. **Middle:** Estimated optical flow. **Right:** The estimated uncertainty (visualized as heatmap) marks the optical flow in the shadow of the car as unreliable (pointed by the red arrow), contrary to the car itself, which is estimated with higher certainty. Marked as most reliable is the optical flow for the static background.

estimate of the network’s own uncertainty, such that the network can highlight hard cases where it cannot reliably estimate the optical flow or where it must decide among multiple probable hypotheses; see Figure 1. However, deep networks in computer vision typically yield only their single preferred prediction rather than the parameters of a distribution.

The first contribution of this paper is an answer to the open question which of the many approaches for uncertainty estimation, most of which have been applied only to small problems so far, are most efficient for high-resolution encoder-decoder regression networks. We provide a comprehensive study of empirical ensembles, predictive models, and predictive ensembles. The first category comprises frequentist methods, the second one relies on the estimation of a parametric output distribution, and the third one combines the properties of the previous two. We implemented these approaches for FlowNet using the common MC dropout technique [9], the less common Bootstrapped Ensembles [21] and snapshot ensembles [13]. We find that in general all these approaches yield surprisingly good uncertainty estimates, where the best performance is achieved with uncertainty estimates derived from Bootstrapped Ensembles of predictive networks.

While such ensembles are a good way to obtain uncertainty estimates, they must run multiple networks to create sufficiently many samples. This drawback increases the computational load and memory footprint at training and test time linearly with the number of samples, such that these approaches are not applicable in real-time.

As a second contribution, we present a multi-headed network architecture that yields multiple hypotheses in a single network without the need of sampling. We use a loss that only penalizes the best hypothesis. This pushes the network to make multiple different predictions in case of doubt. We train a second network to optimally combine the hypotheses and to estimate the final uncertainty. This network yields the same good uncertainty estimates as Bootstrapped Ensembles, but allows for interactive frame rates. Thus, in this paper, we address all three important aspects for deployment of optical flow estimation in automotive systems: high accuracy inherited from the base network, a measure of reliability, and a fast runtime.

2 Related Work

Confidence measures for optical flow. While there is a large number of optical flow estimation methods, only few of them provide uncertainty estimates.

Post-hoc methods apply post-processing to already estimated flow fields. Kondermann et al. [18] used a learned linear subspace of typical displacement neighborhoods to test the reliability of a model. In their follow-up work [19], they proposed a hypothesis testing method based on probabilistic motion models learned from ground-truth data. Aodha et al. [1] trained a binary classifier to predict whether the endpoint error of each pixel is bigger or smaller than a certain threshold and used the predicted classifier’s probability as an uncertainty measure. All post-hoc methods ignore information given by the model structure.

Model-inherent methods, in contrast, produce their uncertainty estimates using the internal estimation model, i.e., energy minimization models. Bruhn and Weickert [4] used the inverse of the energy functional as a measure of the deviation from the model assumptions. Kybic and Nieuwenhuus [20] performed bootstrap sampling on the data term of an energy-based method in order to obtain meaningful statistics of the flow prediction. The most recent work by Wannenwetsch et al. [32] derived a probabilistic approximation of the posterior of the flow field from the energy functional and computed flow mean and covariance via Bayesian optimization. Ummenhofer et al. [31] presented a depth estimation CNN that internally uses a predictor for the deviation of the estimated optical flow from the ground-truth. This yields a confidence map for the intermediate optical flow that is used internally within the network. However, this approach treats flow and confidence separately and there was no evaluation for the reliability of the confidence measure.

Uncertainty estimation with CNNs. Bayesian neural networks (BNNs) have been shown to obtain well-calibrated uncertainty estimates while maintaining the properties of standard neural networks [26,24]. Early work [26] mostly used Markov Chain Monte Carlo (MCMC) methods to sample networks from the distribution of the weights, where some, for instance Hamiltonian Monte Carlo, can make use of the gradient information provided by the backpropagation algorithm. More recent methods generalize traditional gradient based MCMC methods to the stochastic mini-batch setting, where only noisy estimates of the true gradient are available [7,33]. However, even these recent MCMC methods do not scale well to high-dimensional spaces, and since contemporary encoder-decoder networks like FlowNet have millions of weights, they do not apply in this setting.

Instead of sampling, variational inference methods try to approximate the distribution of the weights by a more tractable distribution [10,3]. Even though they usually scale much better with the number of datapoints and the number of weights than their MCMC counterparts, they have been applied only to much smaller networks [12,3] than in the present paper.

Gal and Ghahramani [9] sampled the weights by using dropout after each layer and estimated the *epistemic* uncertainty of neural networks. In a follow-up work by Kendall and Gal [17], this idea was applied to vision tasks, and the *aleatoric* uncertainty (which explains the noise in the observations) and

the epistemic uncertainty (which explains model uncertainty) were studied in a joint framework. We show in this paper, that the dropout strategy used in all previous computer vision applications [17,28] is not the best one per-se, and other strategies yield better results.

In contrast to Bayesian approaches, such as MCMC sampling, bootstrapping is a frequentist method that is easy to implement and scales nicely to high-dimensional spaces, since it only requires point estimates of the weights. The idea is to train M neural networks independently on M different bootstrapped subsets of the training data and to treat them as independent samples from the weight distribution. While bootstrapping does not ensure diversity of the models and in the worst case could lead to M identical models, Lakshminarayanan et al. [21] argued that ensemble model averaging can be seen as dropout averaging. They trained individual networks with random initialization and random data shuffling, where each network predicts a mean and a variance. During test time, they combined the individual model predictions to account for the epistemic uncertainty of the network. We also consider so-called *snapshot ensembles* [13] in our experiments. These are obtained rather efficiently via Stochastic Gradient Descent with warm Restarts (SGDR) [23].

Multi-hypotheses estimation. The loss function for the proposed multi-hypotheses network is related to Guzman-Rivera et al. [11], who proposed a similar loss function for SSVMs. Lee et al. [22] applied the loss to network ensembles and Chen & Koltun [6] to a single CNN.

3 Uncertainty Estimation with Deep Networks

Assume we have a dataset $\mathcal{D} = \{(\mathbf{x}_0, \mathbf{y}_0^{\text{gt}}), \dots, (\mathbf{x}_N, \mathbf{y}_N^{\text{gt}})\}$, which is generated by sampling from a joint distribution $p(\mathbf{x}, \mathbf{y})$. In CNNs, it is assumed that there is a unique mapping from \mathbf{x} to \mathbf{y} by a function $f_{\mathbf{w}}(\mathbf{x})$, which is parametrized by weights \mathbf{w} that are optimized according to a given loss function on \mathcal{D} .

For optical flow, we denote the trained network as a mapping from the input images $\mathbf{x} = (\mathbf{I}_1, \mathbf{I}_2)$ to the output optical flow $\mathbf{y} = (\mathbf{u}, \mathbf{v})$ as $\mathbf{y} = f_{\mathbf{w}}(\mathbf{I}_1, \mathbf{I}_2)$, where \mathbf{u}, \mathbf{v} are the x- and y-components of the optical flow. The FlowNet by Dosovitskiy et al. [8] minimizes the per-pixel endpoint error

$$\text{EPE} = \sqrt{(u - u^{\text{gt}})^2 + (v - v^{\text{gt}})^2} , \quad (1)$$

where the pixel coordinates are omitted for brevity. This network, as depicted in Figure 2a, is fully deterministic and yields only the network's preferred output $\mathbf{y} = f_{\mathbf{w}}(\mathbf{x})$. Depending on the loss function, this typically corresponds to the mean of the distribution $p(\mathbf{y}|\mathbf{x}, \mathcal{D})$. In this paper, we investigate three major approaches to estimate also the variance σ^2 . These are based on the empirical variance of the distribution of an ensemble, a parametric model of the distribution, and a combination of both. The variance in all these approaches serves as an estimate of the uncertainty.

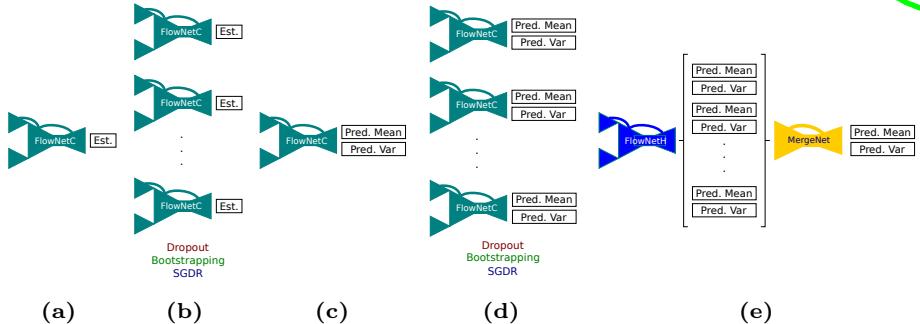


Fig. 2: Overview of the networks considered in this paper. (a) FlowNetC trained with EPE. (b) Same network as (a), where an ensemble is built using dropout, bootstrapping or SGDR. (c) FlowNetC trained with -log-likelihood to predict mean and variance. (d) Same network as (c), where an ensemble is built using dropout, bootstrapping or SGDR. (e) FlowNetH trained to predict multiple hypotheses with variances, which are merged to a single distributional output. Only (a) exists in this form for optical flow.

3.1 Empirical Uncertainty Estimation

A straightforward approach to get variance estimates is to train M different models independently, such that the mean and the variance of the distribution $p(\mathbf{y}|\mathbf{x}, \mathcal{D})$ can be approximated with the empirical mean and variance of the individual model's predictions. Let $f_{\mathbf{w}_i}(\mathbf{x})$ denote model i of an ensemble of M models with outputs $\mathbf{u}_{\mathbf{w}_i}$ and $\mathbf{v}_{\mathbf{w}_i}$. We can compute the empirical mean and variance for the \mathbf{u} -component by:

$$\mu_{\mathbf{u}} = \frac{1}{M} \sum_{i=1}^M \mathbf{u}_{\mathbf{w}_i}(\mathbf{x}) \quad (2)$$

$$\sigma_{\mathbf{u}}^2 = \frac{1}{M} \sum_{i=1}^M (\mathbf{u}_{\mathbf{w}_i}(\mathbf{x}) - \mu_{\mathbf{u}})^2 \quad (3)$$

and accordingly for the \mathbf{v} -component of the optical flow. Such an ensemble of M networks, as depicted in Figure 2b, can be built in multiple ways. The most common way is via Monte Carlo Dropout [9]. Using dropout also at test time, it is possible to randomly sample from network weights M times to build an ensemble. Alternatively, ensembles of individual networks can be trained with random weight initialization, data shuffling, and bootstrapping as proposed by Lakshminarayanan et al. [21]. A more efficient way of building an ensemble is to use M pre-converged snapshots of a single network trained with the SGDR [23] learning scheme, as proposed by Huang et al. [13]. We investigate these three ways of building ensembles for flow estimation and refer to them as Dropout, Bootstrapped Ensembles and SGDR Ensembles, respectively.

3.2 Predictive Uncertainty Estimation

Alternatively, we can train a network to output the parameters θ of a parametric model of the distribution $p(\mathbf{y}|\mathbf{x}, \mathcal{D})$ as introduced by Nix and Weigend [27]. In the literature, Gaussian distributions (where θ parameterizes the distribution's mean and the variance) are most common, but any type of parametric distribution is possible. Such networks can be optimized by maximizing their log-likelihood:

$$\log p(\mathcal{D} | \mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \log p(\mathbf{y}_i | \theta(\mathbf{x}_i, \mathbf{w})) \quad (4)$$

w.r.t. \mathbf{w} . The predictive distribution for an input \mathbf{x} is then defined as:

$$p(\mathbf{y} | \mathbf{x}, \mathbf{w}) \equiv p(\mathbf{y} | \theta(\mathbf{x}, \mathbf{w})). \quad (5)$$

While negative log-likelihood of a Gaussian corresponds to L_2 loss, FlowNet is trained with an EPE loss, which has more robustness to outliers. Thus, we model the predictive distribution by a Laplacian, which corresponds to an L_1 loss. The univariate Laplace distribution has two parameters a and b and is defined as:

$$\mathcal{L}(u|a, b) = \frac{1}{2b} e^{-\frac{|u-a|}{b}}. \quad (6)$$

As Wannewetsch et al. [32], we model the u and v components of the optical flow to be independent. The approximation yields:

$$\mathcal{L}(u, v | a_u, a_v, b_u, b_v) \approx \mathcal{L}(u | a_u, b_u) \cdot \mathcal{L}(v | a_v, b_v). \quad (7)$$

We obtain a probabilistic version of FlowNet with outputs a_u, a_v, b_u, b_v by minimizing the negative log-likelihood of Eq. 7:

$$-\log(\mathcal{L}(u | a_u, b_u) \cdot \mathcal{L}(v | a_v, b_v)) = \frac{|u - a_u|}{b_u} + \log b_u + \frac{|v - a_v|}{b_v} + \log b_v. \quad (8)$$

As an uncertainty estimate we use the variance of the predictive distribution, which is $\sigma^2 = 2b^2$ in this case. This case corresponds to a single FlowNetC predicting flow and uncertainty as illustrated in Figure 2c.

3.3 Bayesian Uncertainty Estimation

From a Bayesian perspective, to obtain an estimate of model uncertainty, rather than choosing a point estimate for \mathbf{w} , we would marginalize over all possible values:

$$p(\mathbf{y} | \mathbf{x}, \mathcal{D}) = \int p(\mathbf{y} | \mathbf{x}, \mathbf{w}) p(\mathbf{w} | \mathcal{D}) d\mathbf{w} \quad (9)$$

$$= \int p(\mathbf{y} | \theta(\mathbf{x}, \mathbf{w})) p(\mathbf{w} | \mathcal{D}) d\mathbf{w}. \quad (10)$$

This integral cannot be computed in closed form, but by sampling M networks $\mathbf{w}_i \sim p(\mathbf{w} | \mathcal{D})$ from the posterior distribution and using a Monte-Carlo approximation [26], we can approximate its mean and variance as:

$$p(\mathbf{y} | \mathbf{x}, \mathcal{D}) \approx \sum_{i=1}^M p(\mathbf{y} | \theta(\mathbf{x}, \mathbf{w}_i)). \quad (11)$$

Since every parametric distribution has a mean and a variance, also the distributions predicted by each individual network with weights \mathbf{w}_i yield a mean μ_i and a variance σ_i^2 . The mean and variance of the mixture distribution in Eq. 11 can then be computed by the law of total variance for the \mathbf{u} -component (as well as for the \mathbf{v} -component) as:

$$\mu_u = \frac{1}{M} \sum_{i=1}^M \mu_{u,i} \quad (12)$$

$$\sigma_u^2 = \frac{1}{M} \sum_{i=1}^M \left((\mu_{u,i} - \mu_u)^2 + \sigma_{u,i}^2 \right). \quad (13)$$

This again can be implemented as ensembles obtained by predictive variants of dropout [9], bootstrapping [21] or SGDR [13], where the ideas from Section 3.1 and Section 3.2 are combined as shown in Figure 2d.

4 Predicting Multiple Hypotheses within a Single Network

The methods presented in the Sections 3.1 and 3.3 require multiple forward passes to obtain multiple samples with the drawback of a much increased computational cost at runtime. In this section, we propose a loss function to make multiple predictions within a single network. We call these predictions *hypotheses*. For the predicted hypotheses, we encourage multimodality by the design of the loss function [11,22,6]. This makes the predictions more diverse and leads to capturing more different solutions, but does not allow for merging by simply computing the mean as for the ensembles presented in the last section. Therefore, we also propose to use a second network that merges the hypotheses to a single prediction and variance, as depicted in Figure 2e.

Since a ground-truth is available only for the single true solution, the question arises of how to train a network to predict multiple hypotheses and how to ensure that each hypothesis comprises meaningful information. To this end, we use a loss that punishes only the best among the network output hypotheses $\mathbf{y}_1, \dots, \mathbf{y}_M$ [11]. Let the loss between a predicted flow vector $\mathbf{y}(i, j)$ and its ground-truth $\mathbf{y}^{gt}(i, j)$ at pixel i, j be defined by a loss function l . We minimize:

$$L_{hyp} = \sum_{i,j} l(\mathbf{y}_{\text{best_idx}(i,j)}, \mathbf{y}^{gt}(i, j)) + \Delta(i, j), \quad (14)$$

where $\text{best_idx}(i, j)$ selects the best hypothesis per pixel according to the ground-truth:

$$\text{best_idx}(i, j) = \operatorname{argmin}_k [\text{EPE}(\mathbf{y}_k(i, j), \mathbf{y}^{\text{gt}}(i, j))]. \quad (15)$$

$\Delta = \Delta_u + \Delta_v$ encourages similar solutions to be from the same hypothesis k via one-sided differences, e.g. for the \mathbf{u} component:

$$\begin{aligned} \Delta_u(i, j) = & \sum_{k: i > 1; j} |y_{k,u}(i, j) - y_{k,u}(i-1, j)| + \\ & \sum_{k: i; j > 1} |y_{k,u}(i, j) - y_{k,u}(i, j-1)| \end{aligned} \quad (16)$$

For l , we either use the endpoint error from Eq. 1 or the negative log-likelihood from Eq. 8. In the latter case, each hypothesis is combined with an uncertainty estimation and l also operates on a variance σ . Equations 15 and 16 remain unaffected. For the best index selection we stick to the EPE since it is the main optimization goal.

To minimize L_{hyp} , the network must make a prediction close to the ground-truth in at least one of the hypotheses. In locations where multiple solutions exist and the network cannot decide for one of them, the network will predict several different likely solutions to increase the chance that the true solution is among these predictions. Consequently, the network will favor making diverse hypotheses in cases of uncertainty. In Tables 3 and 4 of the supplemental material we provide visualizations of such hypotheses.

In principle, L_{hyp} could collapse to use only one of the hypotheses' outputs. In this case the other hypotheses would have very high error and would never be selected for back-propagation. However, due to the variability in the data and the stochasticity in training, such collapse is very unlikely. We never observed that one of the hypotheses was not used by the network, and for the oracle merging we observed that all hypotheses contribute more or less equally. We show this diversity in our experiments.

5 Experiments

To evaluate the different strategies for uncertainty estimation while keeping the computational cost tractable, we chose as a base model the FlowNetC architecture from Dosovitsky et al. [8] with improved training settings by Ilg et al. [14] and by us. A single FlowNetC shows a larger endpoint error (EPE) than the full, stacked FlowNet 2.0 [14], but trains much faster. Note that this work aims for uncertainty estimation and not for improving the optical flow over the base model. The use of ensembles may lead to minor improvements of the optical flow estimates due to the averaging effect, but these improvements are not of major concern here. In the end, we will also show results for a large stacked network to demonstrate that the uncertainty estimation as such is not limited to small, simple networks.

5.1 Training Details

In contrast to Ilg et al. [14], we use Batch Normalization [15] and a continuously dropping cosine learning rate schedule [23]. This yields shorter training times and improves the results a little; see Table 1. We train on FlyingChairs [8] and start with a learning rate of $2e - 4$. For all networks, we fix a training budget of 600k iterations per network, with an exception for SGDR, where we also evaluate performing some pre-cycles. For SGDR Ensembles, we perform restarts every 75k iterations. We fix the T_{mult} to 1, so that each annealing takes the same number of iterations. We experiment with different variants of building ensembles using snapshots at the end of each annealing. We always take the latest M snapshots when building an ensemble. For dropout experiments, we use a dropout ratio of 0.2 as suggested by Kendall et al. [17]. For Bootstrapped Ensembles, we train M FlowNetC in parallel with bootstrapping, such that each network sees different 67% of the training data. For the final version of our method, we perform an additional training of 250k iterations on FlyingThings3D [25] per network, starting with a learning rate of $2e - 5$ that is decaying with cosine annealing. We use the Caffe [16] framework for network training and evaluate all runtimes on an Nvidia GTX 1080Ti. We will make the source code and the final models publicly available.

For the ensembles, we must choose the size M of the ensemble. The sampling error for the mean and the variance decreases with increasing M . However, since networks for optical flow estimation are quite large, we are limited in the tractable sample size and restrict it to $M = 8$. We also use $M = 8$ for FlowNetH.

For SGDR there is an additional pre-cycle parameter: snapshots in the beginning have usually not yet converged and the number of pre-cycles is the number of snapshots we discard before building the ensemble. In the supplemental material we show that the later the snapshots are taken, the better the results are in terms of EPE and AUSE. We use 8 pre-cycles in the following experiments.

5.2 Evaluation Metrics

Sparsification Plots. To assess the quality of the uncertainty measures, we use so-called sparsification plots, which are commonly used for this purpose [1, 32, 19, 20]. Such plots reveal on how much the estimated uncertainty coincides with the true errors. If the estimated variance is a good representation of the model uncertainty, and the pixels with the highest variance are removed gradually, the error should monotonically decrease. Such a plot of our method is shown in Figure 3. The best possible ranking of uncertainties is ranking by the true error to the ground-truth. We refer to this curve as *Oracle Sparsification*. Figure 3 reveals that our uncertainty estimate is very close to this oracle.

| | Iter. | EPE |
|---------------|-------|------|
| FlowNetC [14] | 600k | 3.77 |
| FlowNetC [14] | 1.2m | 3.58 |
| FlowNetC ours | 600k | 3.40 |

Table 1: Optical flow quality on Sintel train clean with the original FlowNetC [14] and our implementation.

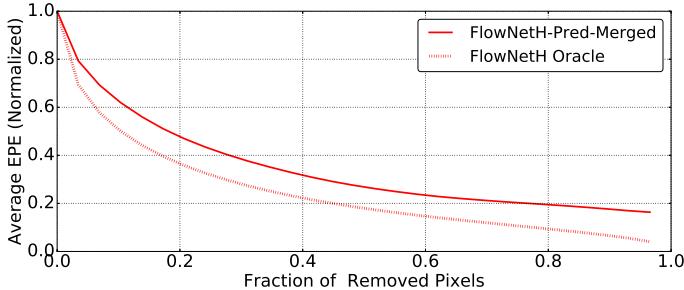


Fig. 3: Sparsification plot of FlowNetH-Pred-Merged for the Sintel train clean dataset. The plot shows the average endpoint error (AEPE) for each fraction of pixels having the highest uncertainties removed. The oracle sparsification shows the lower bound by removing each fraction of pixels ranked by the ground-truth endpoint error. Removing 20 percent of the pixels results in halving the average endpoint error.

| | empirical (Emp) | | | | predictive (Pred) | | | | Runtime |
|-----------------------------|-----------------|------|------------|--------------|-------------------|------|------------|--------------|---------|
| | AUSE | EPE | Oracle EPE | Var. | AUSE | EPE | Oracle EPE | Var. | |
| FlowNetC | - | 3.40 | - | - | 0.133 | 3.62 | - | - | 38ms |
| Dropout | 0.212 | 3.67 | 2.56 | 5.05 | 0.158 | 3.99 | 2.96 | 3.80 | 320ms |
| SGDREnsemble | 0.191 | 3.25 | 2.56 | 3.50 | 0.134 | 3.40 | 2.87 | 1.52 | 304ms |
| BootstrappedEnsemble | 0.209 | 3.41 | 2.17 | 9.52 | 0.127 | 3.46 | 2.49 | 6.15 | 304ms |
| BootstrappedEnsemble-Merged | | | | | 0.102 | 3.20 | 2.49 | 6.15 | 332ms |
| FlowNetH-Merged | - | 3.50 | 1.73 | 83.32 | 0.095 | 3.36 | 1.89 | 52.85 | 60ms |

Table 2: Comparison of flow and uncertainty predictions of all proposed methods with $M = 8$ on the Sintel train clean dataset. Oracle-EPE is the EPE of the pixel-wise best selection from the samples or hypotheses determined by the ground-truth. Var. is the average empirical variance over the 8 samples or hypotheses. Predictive versions (Pred) generally outperform empirical versions (Emp). Including a merging network increases the performance. FlowNetH-Pred-Merged performs best for predicting uncertainties and has a comparatively low runtime.

Sparsification Error. For each approach the oracle is different, hence a comparison among approaches using a single sparsification plot is not possible. To this end, we introduce a measure, which we call *Sparsification Error*. It is defined as the difference between the sparsification and its oracle. Since this measure normalizes the oracle out, a fair comparison of different methods is possible. In Figure 4a, we show sparsification errors for all methods we present in this paper. To quantify the sparsification error with a single number, we use the Area Under the Sparsification Error curve (*AUSE*).

Oracle EPE. For each ensemble, we also compute the hypothetical endpoint error by considering the pixel-wise best selection from each member (decided by the ground-truth). We report this error together with the empirical variances among the members in Table 2.

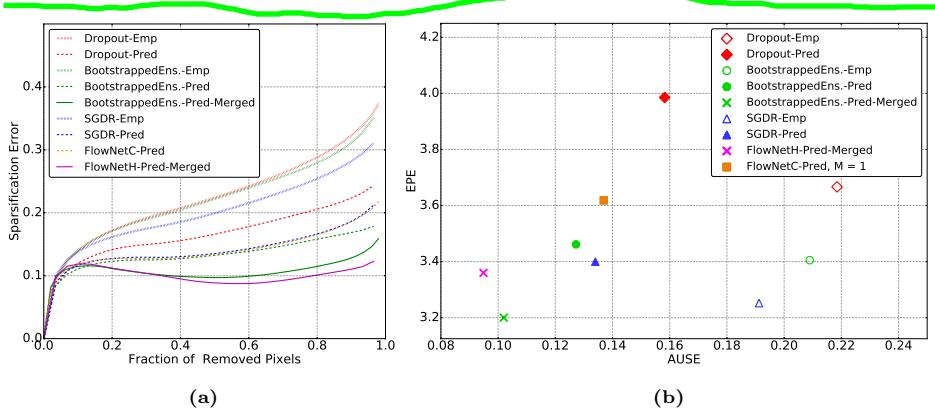


Fig. 4: (a) Sparsification error on the Sintel train clean dataset. The sparsification error (smaller is better) is the proposed measure for comparing the uncertainty estimates among different methods. FlowNetH-Pred-Merged and BootstrappedEnsemble-Pred-Merged perform best in almost all sections of the plot. (b) Scatter plot of AEPE vs. AUSE for the tested approaches visualizing some content of Table 2.

5.3 Comparison among Uncertainties from CNNs

Nomenclature. When a single network is trained against the endpoint error, we refer to this single network and the resulting ensemble as empirical (abbreviated as *Emp*; Figures 2a and 2b), while when the single network is trained against the negative log-likelihood, we refer to the single network and the ensemble as predictive (*Pred*; Figures 2c and 2d). When multiple samples or solutions are merged with a network, we add *Merged* to the name. E.g. FlowNetH-Pred-Merged refers to a FlowNetH that predicts multiple hypotheses and merges them with a network, using the loss for a predictive distribution for both, hypotheses and merging, respectively (Figure 2e). Table 2 and Figures 4a, 4b show results for all models evaluated in this paper.

Empirical Uncertainty Estimation. The results show that uncertainty estimation with empirical ensembles is good, but worse than the other methods presented in this paper. However, in comparison to predictive counterparts, empirical ensembles tend to yield slightly better EPES, as will be discussed in the following.

Predictive Uncertainty Estimation. The estimated uncertainty is better with predictive models than with the empirical ones. Even a single FlowNetC with predictive uncertainty yields much better uncertainty estimates than any empirical ensemble in terms of AUSE. This is because when training against a predictive loss function, the network has the possibility to explain outliers with the uncertainty. This is known as *loss attenuation* [17]. While the EPE loss tries to enforce correct solutions also for outliers, the log-likelihood loss attenuates them. The experiments confirm this effect and show that it is advantageous to let a network estimate its own uncertainty.

Predictive Ensembles. Comparing ensembles of predictive networks to a single predictive network shows that a single network is already very close to the predictive ensembles and that the benefit of an ensemble is limited. We attribute this also to loss attenuation: different ensemble members appear to attenuate outliers in a similar manner and induce less diversity, as can be seen by the variance among the members of the ensemble (column ‘Var.’ in Table 2).

When comparing empirical to predictive ensembles, we can draw the following conclusions: a.) empirical estimation provides more diversity within the ensemble (variance column in Table 2), b.) empirical estimation provides lower EPEs and Oracle EPEs, c.) all empirical setups provide worse uncertainty estimates than predictive setups.

Ensemble Types. We see that the commonly used dropout [9] technique performs worst in terms of EPE and AUSE, although the differences between the predictive ensemble types are not very large. SGDR Ensembles provide better uncertainties, yet the variance among the samples is the smallest. This is likely because later ensemble members are derived from previous snapshots of the same model. Furthermore, because of the 8 pre-cycles, SGDR experiments ran the largest number of training iterations, which could be an explanation why they provide a slightly better EPE than other ensembles. Bootstrapped Ensembles provide the highest sample variance and the lowest AUSE among the predictive ensembles.

FlowNetH and Uncertainty Estimation with Merging Networks. Besides FlowNetH we also investigated putting a merging network on top of the predictive Bootstrapped Ensembles. Results show that the multi-hypotheses network (FlowNetH-Pred-Merged) is on-par with BootstrappedEnsemble-Pred-Merged in terms of AUSE and EPE. However, including the runtime, FlowNetH-Pred-Merged yields the best trade-off; see Table 2. Only FlowNetC and FlowNetH-Pred-Merged allow a deployment at interactive frame rates. Table 2 also shows that FlowNetH has a much higher sample variance and the lowest oracle EPE. This indicates that it internally has very diverse and potentially useful hypotheses that could be exploited better in the future. For some visual examples, we refer to Tables 3 and 4 in the supplemental material.

5.4 Comparison to Energy-Based Uncertainty Estimation

We compare the favored approach from the previous section (FlowNetH-Pred-Merged) to ProbFlow [32], which is an energy minimization approach and currently the state-of-the-art for estimating the uncertainty of optical flow. Figure 5 shows the sparsification plots for the Sintel train final. ProbFlow has almost the same oracle as FlowNetH-Pred-Merged, i.e. the flow field from ProbFlow can equally benefit from sparsification, but the actual sparsification error due to its estimated uncertainty is higher. This shows that FlowNetH-Pred-Merged has superior uncertainty estimates. In Table 3 we show that this also holds for the KITTI dataset. FlowNetH outperforms ProbFlow also in terms of EPE in this case. This shows that the superior uncertainty estimates are not due to a weaker optical flow model, i.e. from obvious mistakes that are easy to predict.

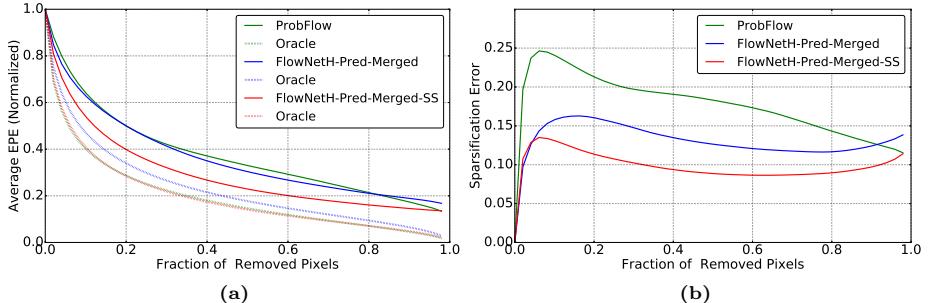


Fig. 5: Plots of the sparsification curves with their respective oracles (a) and of the sparsification errors (b) for ProbFlow, FlowNetH-Pred-Merged and FlowNetH-Pred-Merged-SS (version with 2 refinement networks stacked on top) on the Sintel train final dataset. KITTI versions are similar and provided in the supplemental material.

| | Sintel Clean | | Sintel Final | | KITTI | | runtime |
|-------------------------------|--------------|------|--------------|------|--------------|------|--------------------|
| | AUSE | EPE | AUSE | EPE | AUSE | EPE | |
| ProbFlow [32] | 0.162 | 1.87 | 0.173 | 3.34 | 0.466 | 8.95 | 38.1s [†] |
| FlowNetH-Pred-Merged-FT-KITTI | - | - | - | - | 0.086 | 3.12 | 60ms |
| FlowNetH-Pred-Merged | 0.117 | 2.58 | 0.128 | 3.78 | 0.151 | 7.84 | 60ms |
| FlowNetH-Pred-Merged-S | 0.091 | 2.29 | 0.098 | 3.51 | 0.102 | 6.86 | 86ms |
| FlowNetH-Pred-Merged-SS | 0.089 | 2.19 | 0.096 | 3.40 | 0.091 | 6.50 | 99ms |

Table 3: Comparison of FlowNetH to the state-of-the-art uncertainty estimation method ProbFlow [32] on the Sintel train clean, Sintel train final and our KITTI 2012+2015 validation split datasets. The '-FT-KITTI' version is trained on FlyingChairs [8] first and then on FlyingThings3D [25], as described in Sec. 5.1 and subsequently fine-tuned on our KITTI 2012+2015 training split. FlowNetH-Pred-Merged, -S and -SS are all trained with the FlowNet2 [14] schedule described in supplemental material Fig. 6. Our method outperforms ProbFlow in AUSE by a large margin and also in terms of EPE for the KITTI dataset. [†]runtime taken from [32], please see the supplemental material for details on the computation of the ProbFlow outputs.

Table 3 further shows that the uncertainty estimation is not limited to simple encoder-decoder networks, but can also be applied successfully to state-of-the-art stacked networks [14]. To this end, we follow Ilg et al. [14] and stack refinement networks on top of FlowNetH-Pred-Merged. Different from [14], each refinement network yields the residual of the flow field and the uncertainty, as recently proposed by [29]. We refer to the network with the 1st refinement network as FlowNetH-Pred-Merged-S and with the second refinement network as FlowNetH-Pred-Merged-SS, since each refinement network is a FlowNetS [14].

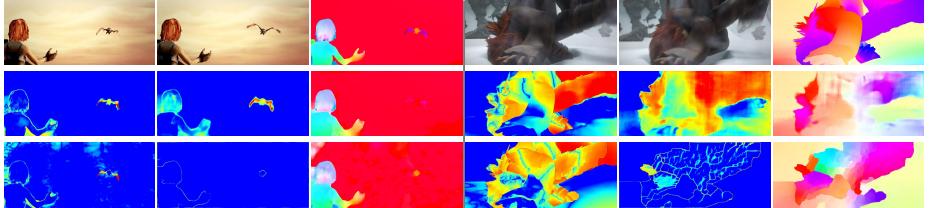


Fig. 6: Comparison between FlowNetH-Pred-Merged and ProbFlow [32]. The first row shows the image pair followed by its ground-truth flow for two different scenes from the Sintel final dataset. The second row shows FlowNetH-Pred-Merged results: entropy from a Laplace distribution with ground-truth error (we refer to this as *Oracle Entropy* to represent the optimal uncertainty as explained in the supplemental material), predicted entropy and predicted flow. Similar to the second row, the third row shows the results for ProbFlow. Although both methods fail at estimating the motion of the dragon on the left scene and the motion of the arm and the leg in the right scene, our method is better at predicting the uncertainties in these regions.

The uncertainty estimation is not negatively influenced by the stacking, despite the improving flow fields. This shows again that the uncertainty estimation works reliably notwithstanding if the predicted optical flow is good or bad.

Figure 6 shows a qualitative comparison to ProbFlow. Clearly, the uncertainty estimate of FlowNet-Pred-Merged also performs well outside motion boundaries and covers many other causes for brittle optical flow estimates. More results on challenging real-world data are shown in the supplemental video which can also be found on <https://youtu.be/HvyovWSo8uE>.

6 Conclusion

We presented and evaluated several methods to estimate the uncertainty of deep regression networks for optical flow estimation. We showed that SGDR and Bootstrapped Ensembles perform better than the commonly used dropout technique. Furthermore, we found that a single network can estimate its own uncertainty surprisingly well and that this estimate outperforms every empirical ensemble. We believe that these results will apply to many other computer vision tasks, too. Moreover, we presented a multi-hypotheses network that shows very good performance and is faster than sampling-based approaches and ensembles. The fact that networks can estimate their own uncertainty reliably and in real-time is of high practical relevance. Humans tend to trust an engineered method much more than a trained network, of which nobody knows exactly how it solves the task. However, if networks say when they are confident and when they are not, we can trust them a bit more than we do today.

Acknowledgements

We gratefully acknowledge funding by the German Research Foundation (SPP 1527 grants BR 3815/8-1 and HU 1900/3-1, CRC-1140 KIDGEM Z02) and by the Horizon 2020 program of the EU via the ERC Starting Grant 716721 and the project Trimb0t2020.

References

1. Aodha, O.M., Humayun, A., Pollefeys, M., Brostow, G.J.: Learning a confidence measure for optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(5), 1107–1120 (May 2013). <https://doi.org/10.1109/TPAMI.2012.171>
2. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: *IEEE Int. Conference on Computer Vision (ICCV)* (2015)
3. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*. pp. 1613–1622
4. Bruhn, A., Weickert, J.: A Confidence Measure for Variational Optic flow Methods, pp. 283–298. Springer Netherlands, Dordrecht (2006), https://doi.org/10.1007/1-4020-3858-8_15
5. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: *European Conference on Computer Vision (ECCV)* (2012)
6. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: *IEEE Int. Conference on Computer Vision (ICCV)* (2017)
7. Chen, T., Fox, E., Guestrin, C.: Stochastic gradient Hamiltonian Monte Carlo. In: *Proceedings of the 31th International Conference on Machine Learning, (ICML'14)* (2014)
8. Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., v.d. Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: *IEEE Int. Conference on Computer Vision (ICCV)* (2015)
9. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *Int. Conference on Machine Learning (ICML)* (2016)
10. Graves, A.: Practical variational inference for neural networks. In: *In Advances in Neural Information Processing Systems (NIPS) 2011*. p. 23482356 (2011)
11. Guzman-Rivera, A., Batra, D., Kohli, P.: Multiple choice learning: Learning to produce multiple structured outputs. In: *Int. Conference on Neural Information Processing Systems (NIPS)* (2012)
12. Hernández-Lobato, J., Adams, R.: Probabilistic backpropagation for scalable learning of Bayesian neural networks. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)* (2015)
13. Huang, G., Li, Y., Pleiss, G.: Snapshot ensembles: Train 1, get M for free. In: *Int. Conference on Learning Representations (ICLR)* (2017)
14. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)

15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR, Lille, France (07–09 Jul 2015), <http://proceedings.mlr.press/v37/ioffe15.html>
16. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proc. ACMMM. pp. 675–678 (2014)
17. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Int. Conference on Neural Information Processing Systems (NIPS) (2017)
18. Kondermann, C., Kondermann, D., Jähne, B., Garbe, C.: An Adaptive Confidence Measure for Optical Flows Based on Linear Subspace Projections, pp. 132–141. Springer Berlin Heidelberg, Berlin, Heidelberg (2007), https://doi.org/10.1007/978-3-540-74936-3_14
19. Kondermann, C., Mester, R., Garbe, C.: A Statistical Confidence Measure for Optical Flows, pp. 290–301. Springer Berlin Heidelberg, Berlin, Heidelberg (2008), https://doi.org/10.1007/978-3-540-88690-7_22
20. Kybic, J., Nieuwenhuis, C.: Bootstrap optical flow confidence and uncertainty measure. Computer Vision and Image Understanding **115**(10), 1449 – 1462 (2011). <https://doi.org/https://doi.org/10.1016/j.cviu.2011.06.008>, <http://www.sciencedirect.com/science/article/pii/S1077314211001536>
21. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS workshop (2016)
22. Lee, S., Purushwalkam, S., Cogswell, M., Ranjan, V., Crandall, D., Batra, D.: Stochastic multiple choice learning for training diverse deep ensembles. In: Int. Conference on Neural Information Processing Systems (NIPS) (2016)
23. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. In: Int. Conference on Learning Representations (ICLR) (2017)
24. MacKay, D.J.C.: A practical bayesian framework for backpropagation networks. Neural Computation **4**(3), 448–472 (May 1992)
25. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4040–4048 (June 2016). <https://doi.org/10.1109/CVPR.2016.438>
26. Neal, R.: Bayesian learning for neural networks. PhD thesis, University of Toronto (1996)
27. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Neural Networks, 1994. IEEE World Congress on Computational Intelligence. vol. 1, pp. 55–60 vol.1 (June 1994). <https://doi.org/10.1109/ICNN.1994.374138>
28. Novotny, D., Larlus, D., Vedaldi, A.: Learning 3D object categories by looking around them. In: IEEE Int. Conference on Computer Vision (ICCV) (2017)
29. Pang, J., Sun, W., Ren, J.S.J., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: IEEE Int. Conference on Computer Vision (ICCV) Workshop (2017)
30. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: EpicFlow: Edge-Preserving Interpolation of Correspondences for Optical Flow. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

31. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), <http://lmb.informatik.uni-freiburg.de//Publications/2017/UZUMIDB17>
32. Wannenwetsch, A.S., Keuper, M., Roth, S.: Probflow: Joint optical flow and uncertainty estimation. In: IEEE Int. Conference on Computer Vision (ICCV) (Oct 2017)
33. Welling, M., Teh, Y.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML'11) (2011)

Supplementary Material

1 Video

Please see the supplementary video for qualitative results on a number of diverse real-world video sequences and a comparison to ProbFlow [32]. The video is also available on <https://youtu.be/HvyovWSo8uE>.

2 Color Coding

For optical flow visualization we use the color coding of Butler et al. [5]. The color coding scheme is illustrated in Figure 1. Hue represents the direction of the displacement vector, while the intensity of the color represents its magnitude. White color corresponds to no motion. Because the range of motions is very different in different image sequences, we scale the flow fields before visualization: independently for each image pair shown in figures, and independently for each video fragment in the supplementary video. Scaling is always the same for all methods being compared.

For uncertainty visualizations we show the predicted entropy, which we compute as:

$$H = \log(2b_x e) + \log(2b_y e), \quad (1)$$

where b_x and b_y are estimated scale parameters from our Laplace distribution model for x and y dimensions and e is Euler's number. To assess the quality of our uncertainty estimations, we compare our estimated entropies against the limiting cases, where b_x and b_y correspond to exactly the estimation errors $|u_{pred} - u_{gt}|$ and $|v_{pred} - v_{gt}|$. We visualize this as the *Oracle Entropy* in all cases where ground-truth is present. For ProbFlow [32], the underlying distribution is Gaussian and therefore we use the entropy of a Gaussian distribution as:

$$H = 0.5 * \log(2e\sigma_x^2\pi) + 0.5 * \log(2e\sigma_y^2\pi), \quad (2)$$

and set σ_x and σ_y to $|u_{pred} - u_{gt}|$ and $|v_{pred} - v_{gt}|$, respectively. To compare to this oracle entropy we normalize to the same range, but when comparing our method to ProbFlow, we allow to normalize to different ranges to show the most interesting aspects of the entropy.

3 Sparsification Plots

Sparsification is a way to assess the quality of uncertainty estimates for optical flow. Already popular in literature [4,19,20,1], it works by progressively discarding percentages of the pixels the model is most uncertain about and verifying

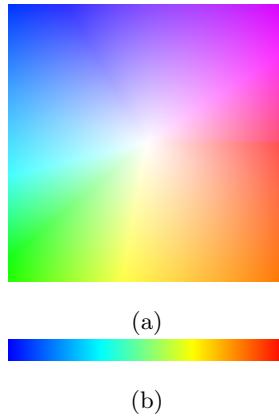


Fig. 1: (a) Flow field color coding used in this paper. The displacement of every pixel in this illustration is the vector from the center of the square to this pixel. The central pixel does not move. The value is scaled differently for different images to best visualize the most interesting range. (b) The color coding used for displaying the entropy maps, from the lowest value (blue), to the highest (red).

whether this corresponds to a proportional decrease in the remaining average endpoint error. To make the results of different experiments comparable, the errors are normalized to 1.

Image-wise sparsification. The method, including the normalization, is typically applied to images individually and the sparsification plots of all images are then averaged. In the main paper we also follow this procedure. However, this approach weights images where the uncertainty estimation is easy equally to images where the uncertainty estimation is hard. Also, due to the normalization, pixels with very large endpoint error from one image can be treated equally to pixels with very small endpoint error from another image.

Dataset-wise sparsification. Alternatively, one can perform the sparsification on a whole dataset. In this variant, the sparsification is performed first (by ranking across the whole dataset) and normalization is performed last. With this approach, the effect of the outliers is better visible in the sparsification curves, which show larger slopes with respect to the previous version.

In Figures 2a, and 2b we present the figures from the main paper again with the dataset-wise sparsification. In Figure 2a we observe that the FlowNetH-Pred-Merged and BootstrappedEnsemble-Pred-Merged perform slightly worse than other ensembles for very high uncertainties, when sparsified on the whole dataset. As also observed in the main paper the best performing model in terms of AUSE is FlowNetH-Pred-Merged.

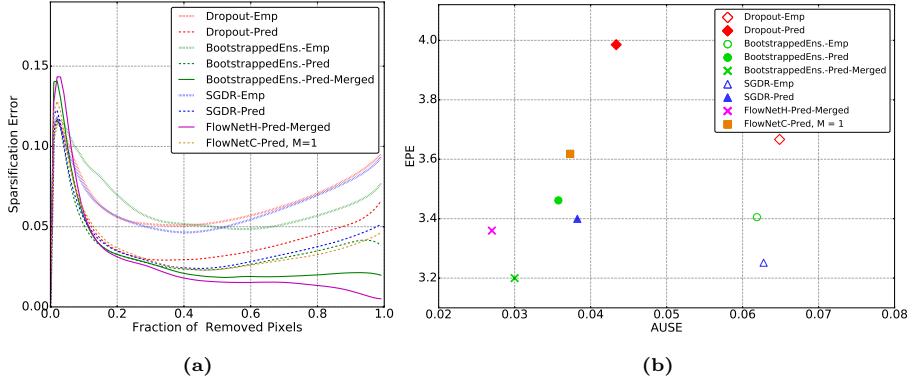


Fig. 2: NOTE: In this version we normalize to dataset-wise instead of image-wise. (a) Sparsification error plots on Sintel train clean dataset. Number of ensemble members are fixed to $M = 8$ and number of pre-cycles for SGDR to 8. We observe that in this case FlowNetH-Pred-Merged and the BootstrappedEnsemble-Pred-Merged perform slightly worse for very high uncertainties, while still showing the best performance for remaining uncertainties. (b) Scatter plot of EPE vs. AUSE for proposed ensemble types. For SGDR, we take the last M available snapshots. The behavior of the different models is not drastically different from the one visible in the per-image sparsification scatter plots in Figure 5 from the main paper. The best performing model in terms of AUSE is FlowNetH-Pred-Merged.

4 Effect of Pre-Cycles for SGDR Ensembles

For SGDR ensembles not only the ensemble size M , but also the models discarded from earlier cycles matter (pre-cycles). Therefore, we have further experimented with pre-cycle counts from 0 to 8 (with a constant ensemble size of $M = 8$). The scatter plots of EPE vs. AUSE can be seen in Figure 3. Figure 3a shows the plot where image-wise normalization is used for sparsification, while Figure 3b shows the plot for dataset-wise normalization. From both plots we can see that the later the models are taken, the lower the EPE gets without a significant change in the AUSE. When compared to other ensemble types, SGDR ensembles are derived from earlier snapshots and the ones with more pre-cycles are trained for more iterations in total. This might be the reason why they show a lower EPE. However, it also means they can converge more and we actually observe the lowest variance for SGDR ensembles among all the models (see Var. column of Table 2 in the main paper).

5 Evaluation on KITTI and Comparison to ProbFlow

We perform the final evaluation of FlowNetH also on the KITTI datasets. We therefore mix KITTI2012 and KITTI2015 and split into 75%/25% training and

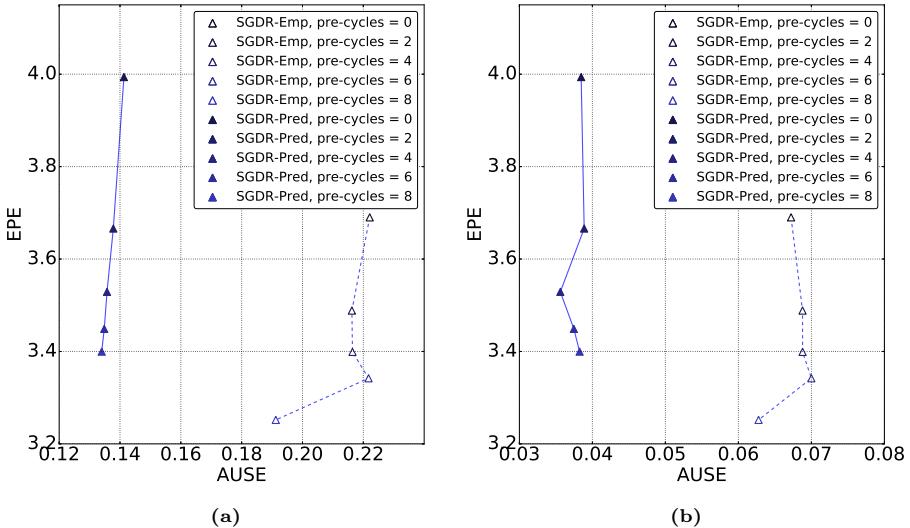


Fig. 3: Scatter plot of EPE vs. AUSE showing the effect of different number of pre-cycles for SGDR ensembles with ensemble size $M = 8$. (a) shows the plot for the image-wise sparsification and (b) shows the plot for the dataset-wise sparsification, as explained in Section 3. It can be seen that a larger number of pre-cycles always positively affects the EPE without penalizing the AUSE score.

test data. In Figure 4 and Table 3 from the main paper, we show the performance of our method compared to ProbFlow [32]. As can be seen from Table 3 in the main paper, fine-tuning significantly reduces the endpoint error, as well as AUSE for FlowNetH-Pred-Merged. This concludes that the quality of the uncertainty estimation of FlowNetH-Pred-Merged is outperforming ProbFlow independent of the flow accuracy.

5.1 Details on ProbFlow results

In order to reproduce ProbFlow results, the ProbFlowFields algorithm contained in the official software package was used. In particular flow initializations were obtained from FlowFields matches [2] and subsequently interpolated with EpicFlow [30]. For FlowFields on KITTI, we have found the best parameter combination to be $r = 5$, $r2 = 4$, $\epsilon = 5$, $e = 4$ and $s = 50$ (the search was conducted around the values suggested in the paper, and full parameter optimization was not performed), and on Sintel we employed $r = 8$, $r2 = 6$, $\epsilon = 5$, $e = 4$ and $s = 50$ as reported by the authors of ProbFlow. In contrast to [32] we evaluate on the complete Sintel train set instead of the selected subset from [32]. This explains the difference in the resulting EPE comparing to the original paper.

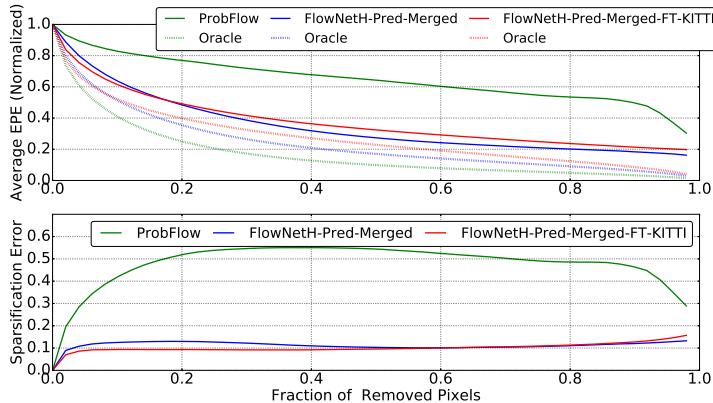


Fig. 4: Sparsification and sparsification error plots for ProbFlow [32], FlowNetH-Pred-Merged (further trained on the FlyingThings3D dataset for 250k iterations) and FlowNetH-Pred-Merged-FT-KITTI (subsequently fine-tuned on our joint KITTI2012 and 2015 training split). One can observe that fine-tuning does not change the sparsification error drastically, while EPE reduces significantly (see Table 3 in the main paper). We see that FlowNetH outperforms ProbFlow both in terms of EPE and AUSE. Note that although the average EPE for ProbFlow is higher, due to the effect of normalization of the oracle sparsification curve, it appears to be the lowest.

6 Training Details for Stacked Networks

The work of FlowNet 2.0 [14] presented how several refinement networks could be stacked on top of a FlowNetC [14] to obtain improved flow fields. We also build such a stack for our FlowNetH-Merged. As illustrated in Figure 5, we stack two refinement networks on top. Different from [14], each refinement network yields the residual of the flow field and the uncertainty, as recently proposed by [29]. Particularly, we find that the residual connections yield much lower convergence times. While we found Batch Normalization and the cosine annealing schedule to be useful for uncertainty estimation networks, we found that refinement networks perform better without Batch Normalization. Therefore, we use the schedule proposed by FlowNet 2.0 [14] scaled down to half the number of iterations and no Batch Normalization for the refinement networks. An overview of the training steps is provided in Figure 6.

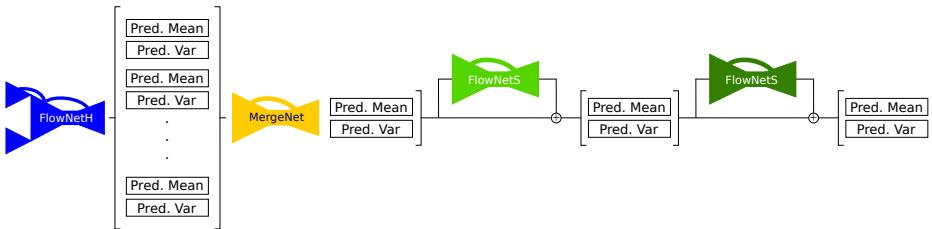


Fig. 5: Illustration of our full uncertainty estimation stack. The first two networks are the FlowNetH and the merging network as described in the paper. The two stacked FlowNetS architecture networks estimate residual refinements.

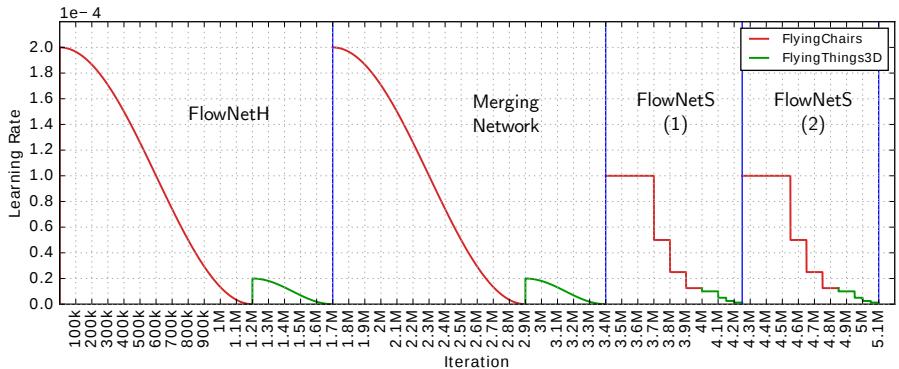


Fig. 6: The learning rate curve of our full stack. FlowNetH and Merging network are trained with Batch Normalization and cosine annealing, while refinement networks are trained without Batch Normalization and with the schedules proposed for FlowNet 2.0 [14] scaled to half the number of iterations. As in [14] the networks are trained step by step, i.e. indicated by each blue line in the figure, the bottom networks are fixed and not trained any more.

7 Qualitative Evaluation

We provide qualitative results on real world videos, Sintel train clean, KITTI2012 and KITTI2015 datasets for FlowNetH-Pred-Merged and ProbFlow in Figure 7, Figure 8 and Figure 9.

At last, we show the outputs for all ensemble members for a simple and a difficult case in Tables 1,2 and Tables 3,4. We note that comparing to the other ensembles, hypothesis from FlowNetH generate the most diverse results.

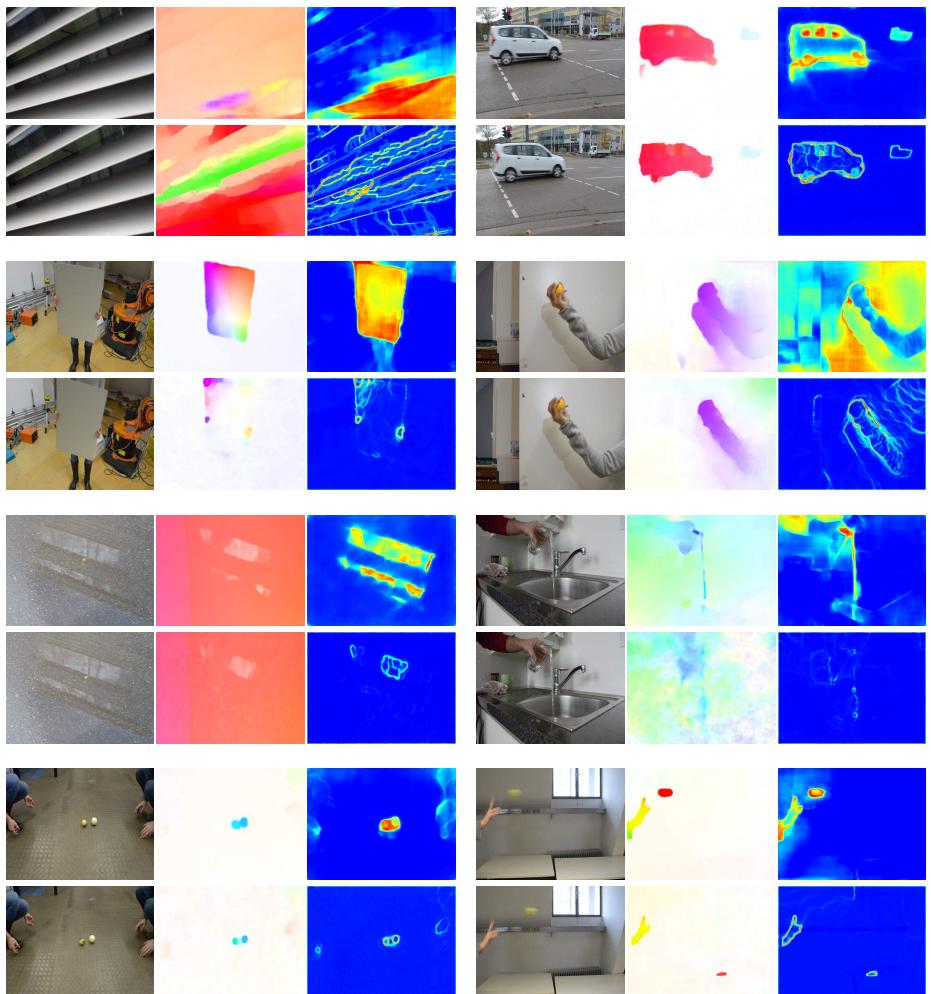


Fig. 7: Examples from real world data. Examples are arranged in a coarse 4x2 grid, where in each we follow the convention: **first column:** original image pair, **second column:** flow predicted by FlowNetH-Pred-Merged and flow predicted by ProbFlow, **third column:** predicted entropy by FlowNetH-Pred-Merged and predicted entropy by ProbFlow. For the full videos of the real world dataset and further comments please see the supplementary video which can also be found on <https://youtu.be/HvyovWS08uE>.

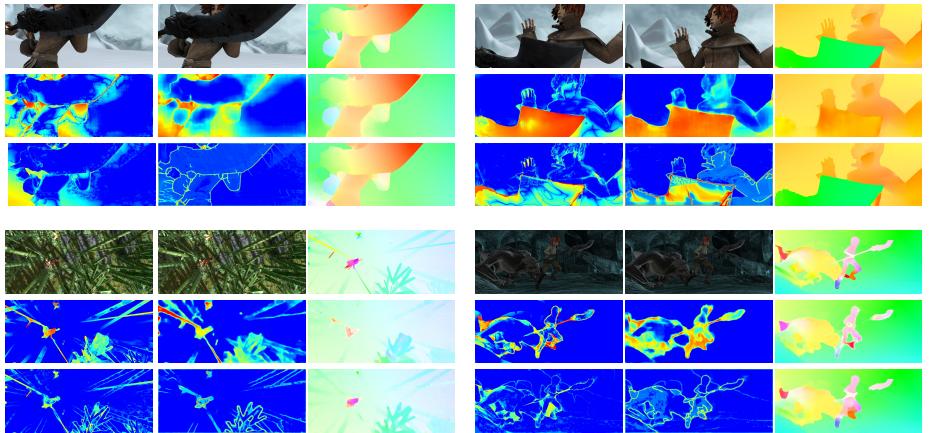


Fig. 8: Four examples from the Sintel train clean dataset for qualitative comparison between FlowNetH-Pred-Merged and ProbFlow. For each example: **first row** shows the original image sequence followed by its ground truth flow field. **Second row** shows FlowNetH-Pred-Merged results: oracle entropy (representing the optimal uncertainty), predicted entropy and predicted flow. Similar to the second row, the **third row** shows the results for ProbFlow. While our method is predicting uncertainties on large areas, ProbFlow shows uncertainties mainly only on the motion or image boundaries and sometimes shows overconfidence in the regions where its prediction is wrong. This is visible e.g. in the upper left example for the lower left corner, where the estimation is wrong, but the uncertainty is low.

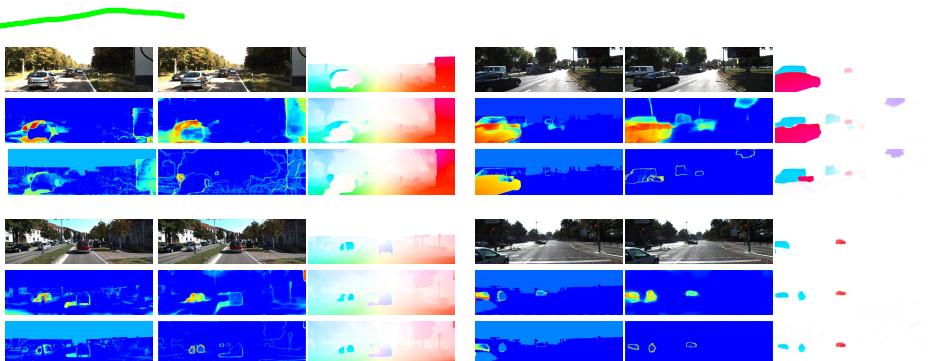


Fig. 9: Examples from KITTI2012 and KITTI2015 datasets for qualitative comparison between FlowNetH-Pred-Merged and ProbFlow. For each example: **first row** shows the original image sequence followed by its ground truth flow field (bilinearly interpolated from sparse ground truth). **Second row** shows FlowNetH-Pred-Merged results: oracle entropy (representing the optimal uncertainty), predicted entropy and predicted flow. Similar to second row, **third row** shows the results for ProbFlow. Remark: In the sky the groundtruth provided in the datasets are invalid due to data acquisition.

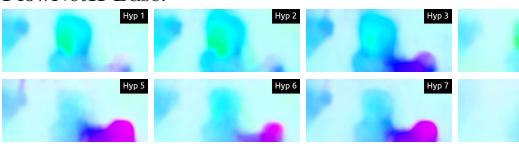
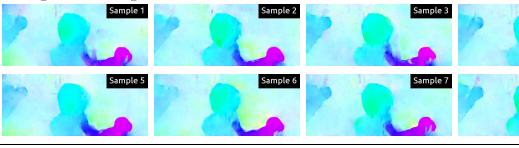
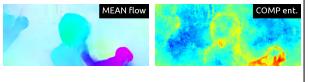
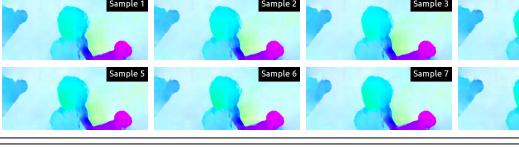
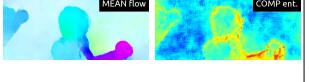
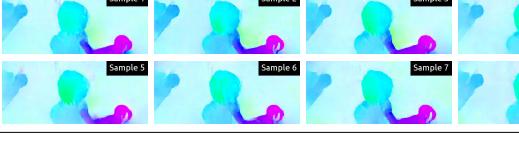
| | | |
|----------------------------|--|--|
| Data: |  |  |
| FlowNetC Emp: | |  |
| FlowNetH Base: |  |  |
| Dropout Emp: |  |  |
| SGDR Emp: |  |  |
| Bootstrapped Ensemble Emp: |  |  |

Table 1: In this table we show the outputs of empirical experiments with all presented methods for an easy Sintel example as well as the averaged flows and computed entropies. Because the example is easy, the networks are certain and not much variety is visible in the outputs.

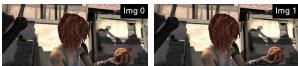
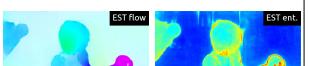
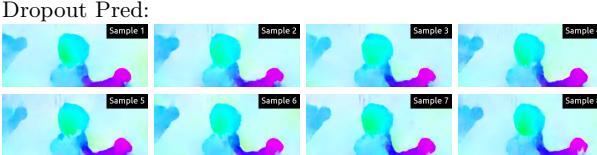
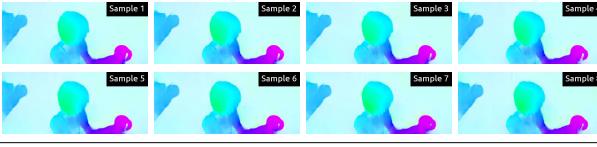
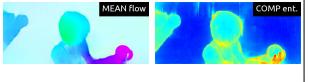
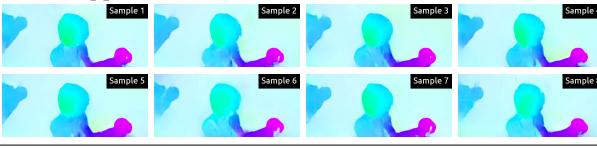
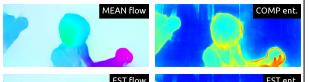
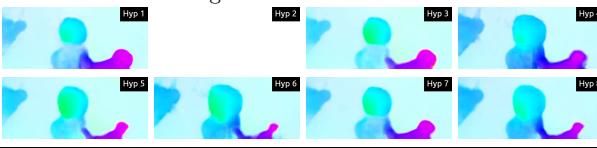
| | | |
|----------------------------|--|--|
| Data: |  |  |
| FlowNetC Pred: | |  |
| Dropout Pred: |  |  |
| SGDR Pred: |  |  |
| BootstrappedEnsemble Pred: |  |  |
| FlowNetH Pred-Merged: |  |  |

Table 2: In this table we show the outputs of predictive experiments with all presented methods for an easy Sintel example as well as the averaged flows and computed entropies. For Bootstrapped-Ensemble-Pred-Merged and FlowNetH-Pred-Merged we show also the estimated flow and estimated entropy as the output of the merging network on top.

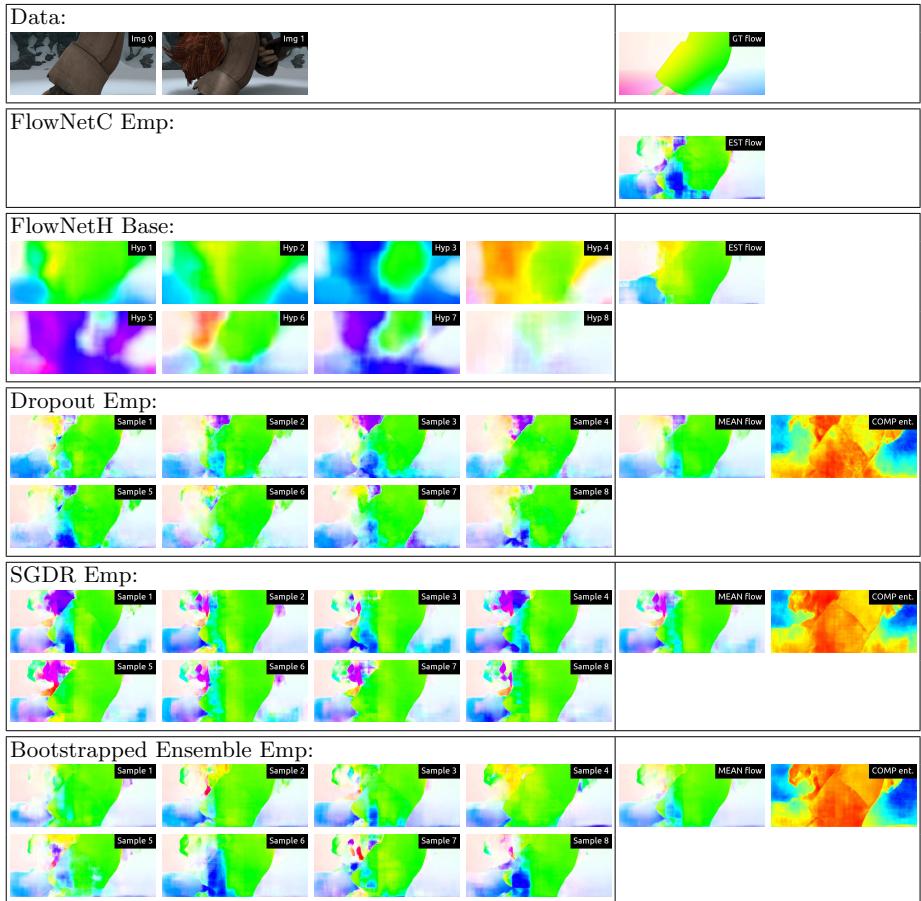


Table 3: In this table we show the outputs of empirical experiments with all presented methods for a hard Sintel example as well as the averaged flows and computed entropies. Some variety of each method is visible, while FlowNetH provides a different kind of output with much more variety.

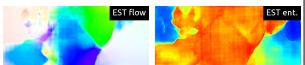
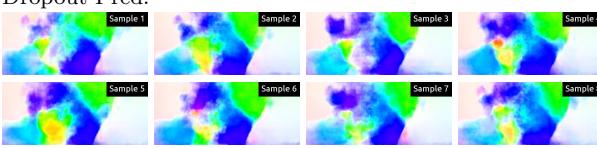
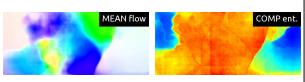
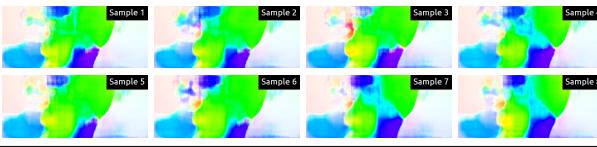
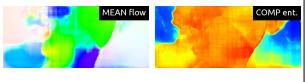
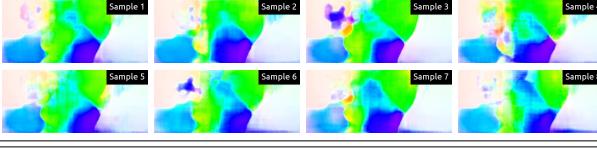
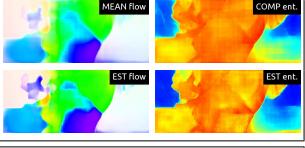
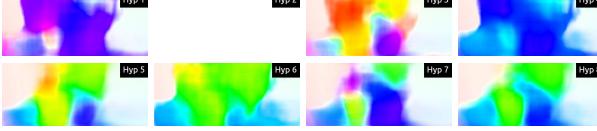
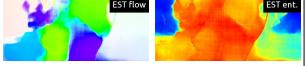
| | | |
|----------------------------|--|--|
| Data: |  |  |
| FlowNetC Pred: | |  |
| Dropout Pred: |  |  |
| SGDR Pred: |  |  |
| BootstrappedEnsemble Pred: |  |  |
| FlowNetH Pred-Merged: |  |  |

Table 4: In this table we show the outputs of predictive experiments with all presented methods for a hard Sintel example, as well as the averaged flows and computed entropies. For BootstrappedEnsemble-Pred-Merged and FlowNetH-Pred-Merged we show also the estimated flow and estimated entropy as the output of the merging network on top. The hypothesis estimated by FlowNetH-Pred-Merged is the most diverse one. In the second hypothesis, the motion predicted is very small and could be corresponding to the background.