Temporal Performance Prediction for Deep Convolutional Long Short-Term Memory Networks

Laura Fieback[1][0009−0003−3456−6766], Bidya Binayam Dash[1][0000−0002−8043−6669], Jakob Spiegelberg[1][0000−0002−6550−0087], and Hanno Gottschalk[2][0000−0003−2167−2028]

[1] Volkswagen AG, Berliner Ring 2, 38440 Wolfsburg, Germany
{laura.fieback,bidya.binayam.dash,jakob.spiegelberg}@volkswagen.de
[2] Mathematical Modeling of Industrial Life Cycles, Institute of Mathematics, TU Berlin, Berlin, Germany
gottschalk@math.tu-berlin.de

Abstract. Quantifying predictive uncertainty of deep semantic segmentation networks is essential in safety-critical tasks. In applications like autonomous driving, where video data is available, convolutional long short-term memory networks are capable of not only providing semantic segmentations but also predicting the segmentations of the next timesteps. These models use cell states to broadcast information from previous data by taking a time series of inputs to predict one or even further steps into the future. We present a temporal postprocessing method which estimates the prediction performance of convolutional long short-term memory networks by either predicting the intersection over union of predicted and ground truth segments or classifying between intersection over union being equal to zero or greater than zero. To this end, we create temporal cell state-based input metrics per segment and investigate different models for the estimation of the predictive quality based on these metrics. We further study the influence of the number of considered cell states for the proposed metrics.

Keywords: Uncertainty quantification · Video frame prediction · Semantic segmentation.

1 Introduction

Retrieving information from images is an important task for scene understanding. Semantic image segmentation is a common approach to gain knowledge about image content by assigning each pixel a label from a predefined label space using neural networks. In safety-critical applications like autonomous driving [11] or medical diagnostics [28], information about the reliability of a prediction is indispensable for decision making. While most approaches to uncertainty quantification focus on a single frame only, temporal information is often available as

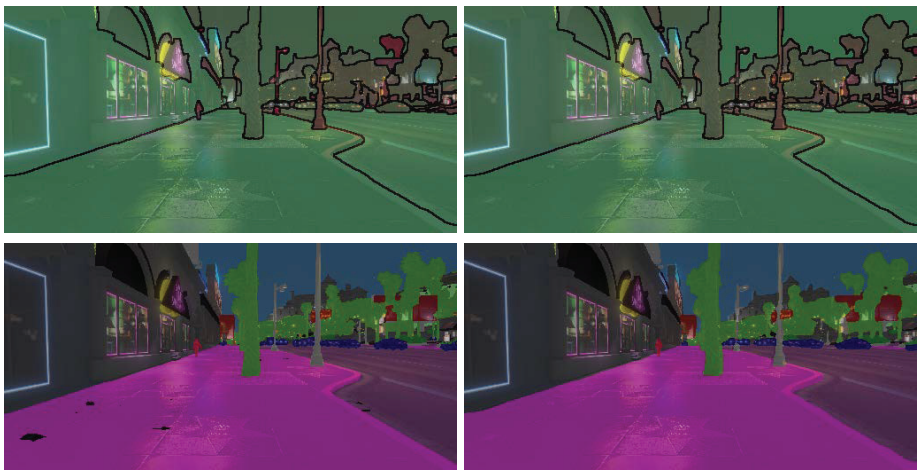


Fig. 1. Visualization of the meta regression task. Ground truth semantic segmentation (bottom left), predicted semantic segmentation via ConvLSTM (bottom right), true IoU_{adj} of prediction and ground truth per segment, where green colors represent high values of IoU_{adj} and red colors represent low values (top left), predicted IoU_{adj} via meta regression (top right).

in the case of video data. To leverage on this, we build on the meta classification and regression approach from [23] and [17]. The method introduced in [23] provides a postprocessing framework to predict the performance of a segmentation network based on its softmax output, i.e., to predict the intersection over union IoU (also known as Jaccard index [13]) per segment from metrics derived from its aggregated softmax outputs (meta regression) or classifying between $IoU = 0$ and $IoU > 0$ (meta classification). Fig. 1 provides a visualization of the meta regression task. Note that this approach can be equipped with any pixel-wise uncertainty measure. In [17], the approach of [23] is extended to time series metrics using a light-weight tracking algorithm. In this work, we investigate temporal metrics retrieved from convolutional long short-term memory networks (ConvLSTMs). Long short-term memory networks (LSTMs) [9] take time series as inputs to make predictions for future timesteps. Thus, the metrics presented in this work express uncertainties in single frames by taking account of temporal information from LSTM outputs. Moreover, we use the light-weight tracking algorithm from [17] to investigate the power of LSTM meta models. This is the first work that conducts meta classification and regression by considering LSTM-based temporal metrics and meta models. Note that our procedure requires a semantic segmentation LSTM network and a video stream of input data.

In our experiments, we predict the performance of a ConvLSTM network [26] trained on the VIsual PERception (VIPER) dataset [21]. This network takes a

time series of semantic segmentations as input to predict the segmentation for the next timestep. We achieve meta classification accuracy of 96.15%(±0.17%) and Area Under Receiver Operating Characteristic ($AUROC$) of 95.04%(±0.22%). The best meta classification results using time series temporal metrics are obtained by our proposed LSTM meta model. For meta regression, we obtain R^2 values of 74.31%(±0.33%).

The remainder of this work is organized as follows. An overview over related work in the field of uncertainty quantification and object tracking is provided in section 2. In section 3, we introduce the temporal metrics for time-dynamic uncertainty quantification followed by the light-weight tracking algorithm in section 4. In section 5 we describe the meta classification and regression method for time-dynamic performance prediction. Finally, we present our numerical results in section 6.

2 Related Work

2.1 Uncertainty Quantification

Modern neural networks tend to be overconfident in their predictions [8, 19]. Temperature scaling [8] and Dirichlet calibration [15] are scaling methods to calibrate the model's confidence estimates. Another common approach to quantify model uncertainty are Bayesian models [18]. Different methods have been established to conduct Bayesian inference via variational approximations like [4] and [5]. In [11], the sampling procedure is simulated based on temporal information in video data. Besides, Monte Carlo dropout [7] is widely used to approximate Bayesian neural networks. In [16], deep ensembles are proposed to quantify predictive uncertainty based on the variance of the ensemble prediction. Other approaches like [22] and [10] propose to model predictive uncertainty based on gradients. In [23], a meta learning approach for semantic segmentation networks is introduced for false positive detection (meta classification) and performance prediction in terms of IoU (meta regression). In [25] and [17], this work is extended by adding resolution dependent uncertainty and temporal metrics, respectively. In [6], performance metrics for video object segmentation and tracking are introduced.

2.2 Object Tracking

Most works in the field of object tracking refer to the task of multi-object tracking, that is, tracking multiple objects in videos by means of bounding boxes [3,20]. Tracking-by-detection [1] is a common approach for this task, which separates objects from the background. The approaches in [27] and [2] are based on segmentation and perform tracking using fully-convolutional Siamese networks and particle filters, respectively. Video panoptic segmentation [14] combines the task of semantic segmentation and object tracking at the same time. Recent works in this field [12, 14] propose end-to-end architectures to fulfill both

tasks simultaneously. In [17], a tracking algorithm is introduced which builds up on a semantic segmentation and matches segments of the same class based on their overlap in consecutive video frames.

3 Segment-wise Dispersion and Temporal Metrics

We build input metrics for the meta classification and regression task based on the output of our ConvLSTM video frame prediction model. The aim of our model is to predict the semantic segmentation of the next timestep given a video sequence of previous segmentations. Semantic segmentation can be viewed as a pixel-wise classification task, where each pixel z of an input image x is classified as a label $y \in C = \{y_1, \ldots, y_c\}$ with c possible output labels. The network's softmax output $f_z(y|x,w)$ can be interpreted as a probability distribution over the output labels $y \in C = \{y_1, \ldots, y_c\}$ given the input image x and the network weights w. The predicted class for a pixel z is then given by the largest softmax value, i.e.,

$$\hat{y}_z(x,w) = \underset{y \in C}{\operatorname{argmax}} f_z(y|x,w). \tag{1}$$

The degree of randomness in a network's softmax output can be quantified using dispersion measures. Thus, we build metrics for the meta classification and regression task based on uncertainty heatmaps representing pixel-wise dispersion measure as proposed in [25]. We consider the entropy

$$E_z(x,w) = -\frac{1}{\log(c)} \sum_{y \in C} f_z(y|x,w) \log f_z(y|x,w), \tag{2}$$

the variation ratio

$$V_z(x,w) = 1 - \max_{y \in C} f_z(y|x,w), \tag{3}$$

as well as the probability margin

$$M_z(x,w) = 1 - \max_{y \in C} f_z(y|x,w) + \max_{y \in C \setminus \hat{y}_z} f_z(y|x,w). \tag{4}$$

Note that, for better comparison, these quantities have been normalized to the interval $[0,1]$.

Let $\hat{S}_x = \{\hat{y}_z(x,w) \,|\, z \in x\}$ denote the predicted semantic segmentation for an image x and $\hat{\mathcal{K}}_x$ the set of all predicted segments k in x, i.e., the set of all connected components of pixels z' with the same predicted class c', that is, $\hat{y}_{z'} = c'$ for all pixels z'. The segment-wise dispersion metrics based on the pixel-wise uncertainty heatmaps introduced above are defined as

$$\bar{D} = \frac{1}{S} \sum_{z \in k} D_z(x,w), \tag{5}$$

where $D_z \in \{E_z, V_z, M_z\}$ and $S = |\{z \in k\}|$ denotes the segment size, i.e., the number of pixels contained in k. As proposed in [23], we define segment-wise

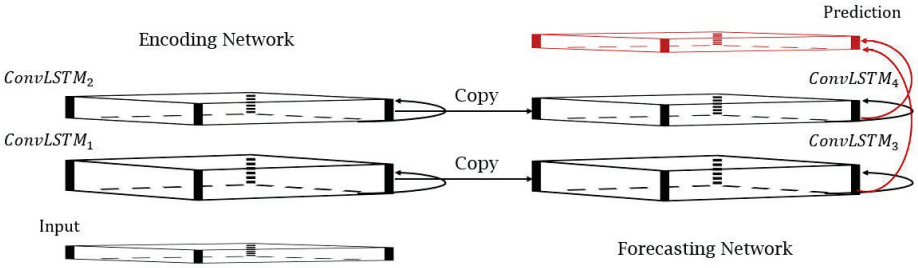


Fig. 2. Depiction of a ConvLSTM block with shared hidden states and cell states between the encoding and forecasting network (from [26]). Here, both networks consist of two ConvLSTM cells, respectively. $ConvLSTM_1$ and $ConvLSTM_3$ share the same states as well as $ConvLSTM_2$ and $ConvLSTM_4$.

inner dispersion metrics and boundary dispersion metrics, since we typically observe high values of D_z for boundary pixels. To this end, let $k_{in} \subset k$ denote the set of all inner pixels of segment k, where a pixel $z \in k$ is called an inner pixel of k if all eight neighboring pixels are an element of k, and let $k_{bd} = k \setminus k_{in}$ denote the set of boundary pixels of segment k. We obtain further segment-wise dispersion metrics by averaging the pixel-wise uncertainty heatmaps over all inner pixels and boundary pixels by analogy with equation (5) yielding the inner and boundary dispersion metrics $\bar{D}_{in}$ and $\bar{D}_{bd}$, respectively, as well as S_{in} and S_{bd}. Based on these metrics, we obtain the respective relative metrics $\tilde{S} = S/S_{bd}$, $\tilde{S}_{in} = S_{in}/S_{bd}$, $\tilde{D} = \bar{D}\tilde{S}$ and $\tilde{D}_{in} = \bar{D}_{in}\tilde{S}_{in}$ with $D \in \{E, V, M\}$. Our set of metrics further contains the geometric center

$$\bar{k} = \left(\bar{k}_1, \bar{k}_1\right) = \frac{1}{S} \sum_{z \in k} \left(z_1, z_1\right), \tag{6}$$

where z_1 and z_2 are the vertical and horizontal coordinates of pixel z as well as the mean class probabilities for each class $y \in C = \{y_1, \ldots, y_c\}$,

$$P\left(y|k\right) = \frac{1}{S} \sum_{z \in k} f_z\left(y|x, w\right). \tag{7}$$

This results in the following set of metrics (see [17])

$$\begin{aligned} U = &\{\bar{D}, \bar{D}_{in}, \bar{D}_{bd}, \tilde{D}, \tilde{D}_{in} \mid D \in \{E, V, M\}\} \cup \{\bar{k}\} \\ &\cup \{S, S_{in}, S_{bd}, \tilde{S}, \tilde{S}_{in}\} \cup \{P\left(y|k\right) | y = y_1, \ldots, y_c\}. \end{aligned} \tag{8}$$

We use these metrics as a baseline in our tests and define additional metrics based on the cell states of our ConvLSTM video frame prediction model. Our model consists of $l = 10$ ConvLSTM blocks (see Fig. 2) using ten previous semantic segmentations $x_{t-i},\ i = 1, \ldots, 10$, of a video to predict the semantic segmentation of the next video frame $\hat{x}_t$. Note that every ConvLSTM block itself consists

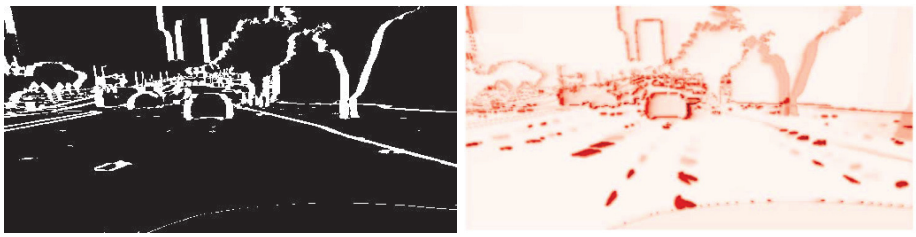


Fig. 3. Prediction error between ground truth and predicted semantic segmentation mask via ConvLSTM (left), where black areas correspond to correctly predicted pixels and white areas to misclassified pixels, and pixel-wise temporal cell state-based heatmap C_z^9 (right).

of an encoding network and a forecasting network, where both networks consist of the same number of convolutional LSTM cells with shared hidden states and cell states (see Fig. 2). The shared hidden states and cell states between both networks are the same states, which are broadcasted to the next ConvLSTM block. In our model, the last convolutional LSTM cell of the forecasting network of each ConvLSTM block outputs states of the same height and width as the model's prediction with 64 features. Thus, for every ConvLSTM block, we focus on the cell state of the last convolutional LSTM cell and define the mean cell state $\bar{C}^i,\ i = 1, \ldots, 10$, of block i as the mean over the 64 features. Based on this, we build temporal heatmaps from the stability of the mean cell state $\bar{C}^i$ over i ConvLSTM blocks. To this end, we define the stability of cell state j for an image x, a pixel z and network weights w as

$$C_z^j\,(x,w) = |\bar{C}_z^1\,(x,w) - \bar{C}_z^{j+1}\,(x,w)\,|,\ j = 1, \ldots, 9. \tag{9}$$

Fig. 3 (right) shows a temporal heatmap obtained from C_z^9, that is, the stability of cell state $j = 9$. As for the uncertainty heatmaps introduced above, we define segment-wise temporal metrics based on the temporal heatmaps as

$$\bar{T} = \frac{1}{S} \sum_{z \in k} T_z\,(x,w)\,, \tag{10}$$

with $T_z \in \{C_z^j,\ j = 1, \ldots, 9\}$. With the notation above, we define our proposed set of metrics for $m = 1, \ldots, 9$ as

$$V_m = U \cup CS_m, \tag{11}$$

where

$$CS_m = \{\bar{T}, \bar{T}_{in}, \bar{T}_{bd}, \tilde{T}, \tilde{T}_{in} \mid T \in \{C^j,\ j = 1, \ldots, m\}\}. \tag{12}$$

Note that all of these metrics can be calculated from our model output without any knowledge of the ground truth.

4 Tracking Algorithm

For the investigation of LSTM meta models, we apply the tracking algorithm proposed in [17]. This algorithm builds on a video sequence of semantic segmentations and performs tracking based on the overlap of segments of the same class in consecutive frames. It does not require additional training. Within this procedure, every segment is assigned a tracking id. To this end, let $\{x_1, \ldots, x_T\}$ denote a sequence of T consecutive semantic segmentations. The overlap of a segment k with segment j is defined as

$$O_{j,k} = \frac{|\{z \in k\} \cap \{z \in j\}|}{|\{z \in j\}|}. \tag{13}$$

The algorithm is applied sequentially to each segmentation $x_t,\ t = 1, \ldots, T$, where for each frame, the segments are prioritized based on their segment size. In detail, the algorithm consists of five steps starting with the largest segment $k^{S_{max}} \in \hat{\mathcal{K}}_{x_t}$ in each step. Once a segment $k \in \hat{\mathcal{K}}_{x_t}$ has been matched with a segment from a previous frame, it is ignored in the following steps. Matched segments receive the same tracking id. To this end, we denote a matched segment k in x_t as k_t.

Step 1 matches segments of the same class in x_t which are close to each other, i.e., with a distance less than a constant c_{near}, and thus, are regarded as one segment.

Step 2 matches segments based on their geometric center. If a segment k exists in two consecutive frames, i.e., $k \in \hat{\mathcal{K}}_{x_{t-1}} \cap \hat{\mathcal{K}}_{x_{t-2}}$, segment k_{t-1} is shifted by $\left(\bar{k}_{t-1} - \bar{k}_{t-2}\right)$ and segments $j \in \hat{\mathcal{K}}_{x_t}$ are matched with the shifted segment $\hat{k}_t$, if the overlap $O_{j,\hat{k}_t}$ is higher than a constant c_{over} or if the distance between the geometric centers $\bar{j}$ and $\bar{\hat{k}}_t$ is smaller than a constant c_{dist}. If segment $k \in \hat{\mathcal{K}}_{x_{t-1}}$ does not exist in two consecutive frames, i.e., $k \notin \hat{\mathcal{K}}_{x_{t-2}}$, segments $j \in \hat{\mathcal{K}}_{x_t}$ are matched based on the distance of the geometric centers $\bar{j}$ and $\bar{k}_{t-1}$.

Step 3 matches segments in consecutive frames based on their overlap, i.e., segments $k \in \hat{\mathcal{K}}_{x_{t-1}}$ and $j \in \hat{\mathcal{K}}_{x_t}$ are matched if $O_{j,k} \geq c_{over}$.

Step 4 accounts for flashing predicted segments due to occlusions or false predictions. It aims at matching segments that are more than one frame apart in temporal direction. To this end, a linear regression model is used to predict the geometric center of segment k in x_t if k was matched in at least two of the last lr segmentations $x_{t-lr}, \ldots, x_{t-1}$. Segments $j \in \hat{\mathcal{K}}_{x_t}$ are matched if the distance between the predicted geometric center $\hat{\bar{k}}_t$ and $\bar{j}$ is less than a constant c_{lin}.

Step 5 assigns a new id to all segments $j \in \hat{\mathcal{K}}_{x_t}$, that have not yet been matched.

5 IoU Prediction

For the task of semantic segmentation, a common measure for predictive quality is the IoU. In our experiments, we use a slight modification proposed in [23],

the IoU_{adj}, which is less prone to fragmented objects. We perform segment-wise meta classification, i.e., classifying between $IoU_{adj} = 0$ and $IoU_{adj} > 0$ as well as segment-wise meta regression, that is, predicting the performance of our ConvLSTM semantic segmentation for each segment in terms of IoU_{adj} by means of the metrics defined in section 3. Note that all of these metrics can be calculated from the ConvLSTM's output without any knowledge of the ground truth. An illustration of the meta regression task is given in Fig. 1. We analyze the information gain induced by the temporal metrics for both single frame metrics and time series metrics as proposed in [17]. Those time series metrics are based on the tracking algorithm introduced in section 4. For each segment $k_t \in \hat{\mathcal{K}}_{x_t}$, we obtain single-frame based metrics $V_m^k = V_{m,t}^k$ according to section 3 as well as their history $V_{m,t-1}^k, \ldots, V_{m,t-T}^k$ due to tracking of segment k over T previous frames. In our experiments, we investigate the influence of metric histories for up to $T = 10$ timesteps. In [17], different models for the meta tasks were investigated. We choose the best performing models, i.e., the linear model (LR), the shallow neural network (NN) as well as the gradient boosting model (GB) for our experiments (for implementation details, see [17]). In addition, we investigate the performance of a shallow LSTM neural network (in the following referred to as LSTM) with 50 neurons only for both meta tasks. The number of LSTM cells depends on the respective number of considered timesteps T of the time series metrics.

6 Numerical Results

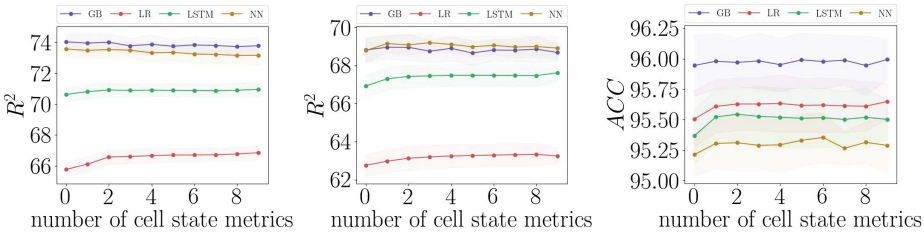
In this section, we investigate the properties of the temporal metrics defined in section 3. We further investigate the influence of time series metrics as described in the previous section and consider different models for meta classification and regression. To this end, we train a ConvLSTM network with ten ConvLSTM blocks (see Fig. 2), each of them built by five convolutional LSTM cells. We train our model on the synthetic VIPER dataset [21]. The dataset consists of more than $250,000$ frames all annotated with ground truth semantic labels with a resolution of 1920×1080 pixels per frame. Since the ground truth annotation has very fine labels, we apply the smoothing algorithm proposed in [24] to generate a coarse ground truth by blurring each class using a normalized box filter. Moreover, we resize the images to 256×512 pixels for computational reasons. The VIPER dataset contains 32 different classes with 23 proposed training ids. Out of these, we further cluster highly underrepresented classes to a misc class which results in a total of 17 training classes. We train our ConvLSTM model on 19 training folders which contain $30,168$ images in total and 8 validation folders yielding a total of $7,021$ images. In our experiments, we compare two different models from our training procedure: The "strong model" (S) which was trained for 18 epochs yielding a mean IoU ($mIoU$) of 82.82%, as well as the "weak model" (W) which obtained an $mIoU$ of 79.45% after 4 epochs of training. We implement the tracking algorithm from section 4 with parameters $c_{near} = 10$, $c_{over} = 0.35$, $c_{dist} = 100$ and $c_{lin} = 50$.

Table 1. Results for meta classification and regression based on temporal metrics for different meta models and the entropy baseline for both the weak (W) and the strong (S) model. The superscript denotes the number of cell state metrics, where the best performance and in particular the given values are reached. The best results are highlighted.

Meta Classification $IoU_{adj} = 0, > 0$					
Entropy Baseline (W): $ACC = 93.40\%(\pm 0.20\%)$ $AUROC = 81.63\%(\pm 0.78\%)$					
Entropy Baseline (S): $ACC = 95.27\%(\pm 0.20\%)$ $AUROC = 80.45\%(\pm 0.71\%)$					
		GB	LR	LSTM	NN
ACC	W	**94.72%(±0.22%)^{7}**	94.39%(±0.16%)^{1}	94.01%(±0.16%)^{6}	93.72%(±0.22%)^{2}
	S	**95.99%(±0.17%)^{9}**	95.65%(±0.15%)^{9}	95.54%(±0.22%)^{2}	95.35%(±0.21%)^{6}
$AUROC$	W	**94.54%(±0.44%)^{0}**	93.69%(±0.47%)^{2}	93.28%(±0.53%)^{0}	92.85%(±0.59%)^{0}
	S	**93.87%(±0.43%)^{2}**	92.57%(±0.42%)^{9}	92.25%(±0.44%)^{9}	91.87%(±0.45%)^{9}
Meta Regression IoU_{adj}					
Entropy Baseline (W): $\sigma = 0.227(\pm 0.002)$ $R^2 = 42.80\%(\pm 0.70\%)$					
Entropy Baseline (S): $\sigma = 0.225(\pm 0.003)$ $R^2 = 38.58\%(\pm 0.81\%)$					
		GB	LR	LSTM	NN
σ	W	**0.154%(±0.002%)^{8}**	0.175%(±0.002%)^{0}	0.162%(±0.001%)^{0}	0.155%(±0.002%)^{8}
	S	0.161%(±0.001%)^{9}	0.175%(±0.002%)^{0}	0.165%(±0.001%)^{0}	**0.160%(±0.002%)^{9}**
R^2	W	**74.04%(±0.52%)^{0}**	66.85%(±0.43%)^{9}	70.96%(±0.47%)^{9}	73.57%(±0.46%)^{0}
	S	68.95%(±0.61%)^{1}	63.33%(±0.59%)^{8}	67.61%(±0.43%)^{9}	**69.19%(±0.47%)^{3}**

For the meta tasks, we use 5 validation folders, not yet used during the training procedure of the ConvLSTM model, which sum up to $3,464$ images. This results in a total of $46,587,336$ segments for the weak model (not yet matched over time) of which $110,739$ have non-empty interior. Out of these, $7,649$ segments have $IoU_{adj} = 0$. For the strong model, we obtain $42,295,440$ segments, $113,286$ with non-empty interior of which $5,622$ segments have $IoU_{adj} = 0$. The corresponding naive classification baseline discussed in [23] and [17] yields an accuracy of 93.09% for the weak model and 95.04% for the strong model. This baseline is obtained by random guessing, i.e., randomly assigning a probability to each segment and thresholding on it. The classification accuracy is the number of correct predictions divided by the total number of predictions made. The corresponding $AUROC$ value is 50%. This baseline is clearly outperformed. To this end note that, the stronger the ConvLSTM model, the higher the naive accuracy. We improve the naive accuracy by further $1.63pp$ for the weak model and $0.95pp$ for the strong model.

In all our experiments, we average our results over ten randomly sampled train/val/test (70%/10%/20%) splits using a total of $38,000$ segments in each split. In tables, the corresponding standard deviations are given in brackets, whereas, in figures, they are given by shades. All meta models considered yield an inference time for all $38,000$ segments together of less than one second. We measure the classification performance of our method in terms of classification accuracy (ACC) and Area Under Receiver Operating Characteristic ($AUROC$), which is obtained by varying the decision threshold between $IoU_{adj} = 0$ and $IoU_{adj} > 0$. For meta regression, we state the results in terms of the regression standard error σ and the R^2 value.



(a) Weak Model R^2 (b) Strong Model R^2 (c) Strong Model ACC

Fig. 4. A selection of results for meta classification in terms of ACC and meta regression in terms of R^2 as functions of the number of considered cell state metrics. Meta regression via the weak model (a), meta regression via the strong model (b), meta classification via the strong model (c).

6.1 Evaluation of Temporal Metrics

First, we investigate the influence of single-frame temporal metrics $V_m = V_{m,t}$ by considering the stability of cell states over $m \in \{1, \ldots, 9\}$ ConvLSTM blocks. Table 1 shows the best results for different meta models. The superscript denotes the number of considered cell state metrics, where the best performance and in particular the given values are reached. Note that the superscript being equal to 0 refers to the metric set U_t without any cell state metrics. For the weak model, we achieve test $AUROC$ values of up to 94.54%(±0.44%) and classification accuracies of up to 94.72%(±0.22%). For the strong model, a test accuracy of 95.99%(±0.17%) is reached and $AUROC$ value up to 93.87%(±0.43%). As in [17], GB performs best for meta classification. With regard to meta regression, we obtain R^2 values up to 74.04%(±0.52%) for the weak model and 69.19%(±0.47%) for the strong model. As a baseline, we consider the approach from [23], i.e., the metric set U_t without any cell state metrics. In almost every experiment, best results are obtained when considering temporal metrics. In those cases where the best results are obtained without temporal metrics, we observe vanishing differences between the respective performance metrics for temporal metrics (e.g., see R^2 values for GB and NN in Fig. 4(a)). In [23], the results are compared with the entropy as a single-metric baseline and with the naive baseline introduced above. For the entropy baseline (see Table 1), we use single-frame gradient boosting as suggested in [17]. Both baselines are clearly outperformed. In contrast to the results in [17], the GB meta regression model does not outperform the neural network in all settings, even though it yields the best results in most of the experiments.

Fig. 4 shows the influence of temporal metrics with respect to R^2 value and classification accuracy. For the linear meta regression model based on the weak ConvLSTM (Fig. 4(a)), we obtain R^2 values up to 66.85%(±0.43%) when taking account of all $m = 9$ temporal metrics, whereas the baseline metrics U_t (0 considered cell state metrics) only achieve averaged R^2 values of 65.77%(±0.45%).

Table 2. Results for meta classification and regression based on time series temporal metrics for different meta models and the GB baseline from [17] for both the weak (W) and the strong (S) model. The superscript denotes the number of frames, where the best performance and in particular the given values are reached. The best results are highlighted.

Meta Classification $IoU_{adj} = 0, > 0$					
Baseline [17] (W): $ACC = 94.93\%(\pm 0.32\%)$ $AUROC = 94.99\%(\pm 0.35\%)$					
Baseline [17] (S): $ACC = 96.03\%(\pm 0.18\%)$ $AUROC = 94.12\%(\pm 0.43\%)$					
		GB	LR	LSTM	NN
ACC	W	$94.95\%(\pm 0.24\%)^{9}$	$94.64\%(\pm 0.24\%)^{8}$	$\mathbf{95.25\%(\pm 0.22\%)^{1}}$	$94.09\%(\pm 0.23\%)^{6}$
	S	$\mathbf{96.15\%(\pm 0.17\%)^{1}}$	$95.88\%(\pm 0.23\%)^{1}$	$\mathbf{96.15\%(\pm 0.17\%)^{9}}$	$95.54\%(\pm 0.30\%)^{1}$
$AUROC$	W	$95.00\%(\pm 0.28\%)^{1}$	$94.24\%(\pm 0.34\%)^{1}$	$\mathbf{95.04\%(\pm 0.22\%)^{1}}$	$93.32\%(\pm 0.47\%)^{1}$
	S	$\mathbf{94.23\%(\pm 0.42\%)^{9}}$	$92.85\%(\pm 0.39\%)^{1}$	$93.65\%(\pm 0.46\%)^{1}$	$91.92\%(\pm 0.64\%)^{0}$
Meta Regression IoU_{adj}					
Baseline [17] (W): $\sigma = 0.153(\pm 0.002)$ $R^2 = 74.00\%(\pm 0.65\%)$					
Baseline [17] (S): $\sigma = 0.161(\pm 0.001)$ $R^2 = 68.27\%(\pm 0.53\%)$					
		GB	LR	LSTM	NN
σ	W	$\mathbf{0.154\%(\pm 0.001\%)^{6}}$	$0.168\%(\pm 0.001\%)^{6}$	$0.157\%(\pm 0.003\%)^{6}$	$0.157\%(\pm 0.003\%)^{7}$
	S	$\mathbf{0.162\%(\pm 0.002\%)^{8}}$	$0.173\%(\pm 0.002\%)^{8}$	$\mathbf{0.162\%(\pm 0.002\%)^{8}}$	$0.163\%(\pm 0.002\%)^{8}$
R^2	W	$\mathbf{74.31\%(\pm 0.33\%)^{0}}$	$69.15\%(\pm 0.46\%)^{3}$	$73.58\%(\pm 0.74\%)^{3}$	$73.54\%(\pm 0.39\%)^{0}$
	S	$68.97\%(\pm 0.81\%)^{1}$	$64.44\%(\pm 0.51\%)^{1}$	$\mathbf{69.00\%(\pm 0.98\%)^{6}}$	$68.53\%(\pm 1.04\%)^{6}$

For the stronger ConvLSTM model (Fig. 4(b)), the best results are obtained for 8 cell state metrics, that is, $R^2 = 63.33\%(\pm 0.59\%)$, whereas the baseline metrics only obtain R^2 values up to $62.76\%(\pm 0.58\%)$. These results are in line with the findings in [23] and [17], that is, stronger segmentation models yield worse meta performance with respect to R^2. Moreover, the analysis of time series metrics in [17] showed a performance gain for linear models, whereas the stronger gradient boosting models do not benefit as much from time series metrics. We observe the same effects with regard to temporal metrics. Finally, with regard to meta classification based on the strong model (Fig. 4(c)), we observe that all models benefit from the temporal metrics, while the linear model outperforms the shallow LSTM and neural network by $0.15pp$ and $0.26pp$, respectively. Note that even though the linear model is only slightly better than the shallow network, this result is not in line with the findings of [23] and [17], where the neural networks outperformed the linear models in all experiments.

6.2 Evaluation of Time Series Temporal Metrics

Next, we investigate time series metrics $\{V_{m,t}, V_{m,t-1}, \ldots, V_{m,t-T}\}$ with $m = 9$ and a length of up to $T = 10$ previous timesteps, yielding 11 different sets of metrics. The results are summarized in Table 2. Since the gradient boosting model performs best in [17] as well as in most of our experiments, we consider the gradient boosting model equipped with the metric set $\{U_t, U_{t-1}, \ldots, U_{t-10}\}$ as the baseline model. This baseline is outperformed for both meta tasks and both the strong and the weak model. For the weak model, we achieve classification accuracy up to $95.25\%(\pm 0.22\%)$ with our proposed LSTM meta model

considering 1 cell state metric. For meta regression, we obtain R^2 values up to 74.31%(±0.33%) for the gradient boosting model. For the strong model, we achieve best results for the classification task by means of the gradient boosting model, while our proposed LSTM meta model outperforms the gradient boosting model in the regression task yielding R^2 values of 69.00%(±0.98%) with 6 considered cell state metrics.

7 Conclusion and Outlook

In this paper, we extended the approach from [23] and [17] for deep ConvLSTM networks. We introduced temporal metrics based on the cell states broadcasted through LSTM cells as additional inputs for meta classification and regression. In our experiments, we studied the influence of different numbers of considered cell state metrics for four meta models, i.e., linear models, gradient boosting, shallow neural networks as well as shallow LSTM models. Moreover, we investigated the influence of LSTM meta models for time series metrics proposed in [17]. In all experiments, our approach slightly improved the state of the art results [23] and [17]. More precisely, we achieve classification accuracy of 96.15%(±0.17%) and $AUROC$ of 95.04% (±0.22%) using our proposed LSTM meta model with temporal metrics. For meta regression, we obtain R^2 values of 74.31%(±0.33%). We plan to develop further LSTM-based metrics for uncertainty quantification applied to the task of predicting several steps into the future.

Disclaimer The results, opinions and conclusions expressed in this publication are not necessarily those of Volkswagen Aktiengesellschaft.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 983–990 (2009). https://doi.org/10.1109/CVPR.2009.5206737
2. Belagiannis, V., Schubert, F., Navab, N., Ilic, S.: Segmentation based particle filtering for real-time 2d object tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 842–855. Springer Berlin Heidelberg (2012)
3. Bergmann, P., Meinhardt, T., Leal-Taixé, L.: Tracking without bells and whistles. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 941–951 (2019). https://doi.org/10.1109/ICCV.2019.00103
4. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1613–1622. PMLR (2015), https://proceedings.mlr.press/v37/blundell15.html
5. Duvenaud, D., Maclaurin, D., Adams, R.: Early stopping as nonparametric variational inference. In: Gretton, A., Robert, C.C. (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. Proceedings

of Machine Learning Research, vol. 51, pp. 1070–1077. PMLR (2016), https://proceedings.mlr.press/v51/duvenaud16.html
6. Erdem, C.E., Sankur, B., Tekalp, A.M.: Performance measures for video object segmentation and tracking **13**(7), 937–951 (2004). https://doi.org/10.1109/TIP.2004.828427
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR (2016), https://proceedings.mlr.press/v48/gal16.html
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017), https://proceedings.mlr.press/v70/guo17a.html
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
10. Hornauer, J., Belagiannis, V.: Gradient-based uncertainty for monocular depth estimation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 613–630. Springer Nature Switzerland (2022)
11. Huang, P.Y., Hsu, W.T., Chiu, C.Y., Wu, T.F., Sun, M.: Efficient uncertainty estimation for semantic segmentation in videos. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 536–552. Springer International Publishing (2018)
12. Hurtado, J.V., Mohan, R., Burgard, W., Valada, A.: Mopt: Multi-object panoptic tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Scalability in Autonomous Driving (2020)
13. Jaccard, P.: The distribution of the flora in the alpine zone.1 **11**, 37–50 (1912)
14. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video panoptic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9856–9865 (2020). https://doi.org/10.1109/CVPR42600.2020.00988
15. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.d., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/8ca01ea920679a0fe3728441494041b9-Paper.pdf
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf
17. Maag, K., Rottmann, M., Gottschalk, H.: Time-dynamic estimates of the reliability of deep semantic segmentation networks. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 502–509 (2020). https://doi.org/10.1109/ICTAI50040.2020.00084
18. MacKay, D.J.C.: A practical bayesian framework for backpropagation networks **4**(3), 448–472 (1992). https://doi.org/10.1162/neco.1992.4.3.448

19. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 15682–15694. Curran Associates, Inc (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/8420d359404024567b5aefda1231af24-Paper.pdf
20. Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 145–161. Springer International Publishing (2020)
21. Richter, S.R., Hayder, Z., Koltun, V.: Playing for benchmarks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2232–2241 (2017). https://doi.org/10.1109/ICCV.2017.243
22. Riedlinger, T., Rottmann, M., Schubert, M., Gottschalk, H.: Gradient-based quantification of epistemic uncertainty for deep object detectors. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3910–3920 (2023). https://doi.org/10.1109/WACV56688.2023.00391
23. Rottmann, M., Colling, P., Paul Hack, T., Chan, R., Hüger, F., Schlicht, P., Gottschalk, H.: Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (2020). https://doi.org/10.1109/IJCNN48605.2020.9206659
24. Rottmann, M., Reese, M.: Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3213–3222 (2023). https://doi.org/10.1109/WACV56688.2023.00323
25. Rottmann, M., Schubert, M.: Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1361–1369 (2019). https://doi.org/10.1109/CVPRW.2019.00176
26. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. pp. 802–810. NIPS'15, MIT Press (2015)
27. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1328–1338 (2019). https://doi.org/10.1109/CVPR.2019.00142
28. Wickstrøm, K., Kampffmeyer, M., Jenssen, R.: Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation. In: 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6 (2018). https://doi.org/10.1109/MLSP.2018.8516998