Forecasting of depth and ego-motion with transformers and self-supervision

Houssem Eddine Boulahbal, Adrian Voicila, Andrew I. Comport

Forecasting of depth and ego-motion with transformers and self-supervision

Houssem eddine BOULAHBAL
Renault Software Factory and
CNRS-I3S,
Côte d'Azur University
Houssem-eddine.Boulahbal@renault.com

Adrian VOICILA
Renault Software Factory
Adrian.Voicila@renault.com

Andrew I. COMPORT
CNRS-I3S,
Côte d'Azur University
Andrew.Comport@cnrs.fr

***Abstract*—This paper addresses the problem of end-to-end self-supervised forecasting of depth and ego motion. Given a sequence of raw images, the aim is to forecast both the geometry and ego-motion using a self supervised photometric loss. The architecture is designed using both convolution and transformer modules. This leverages the benefits of both modules: Inductive bias of CNN, and the multi-head attention of transformers, thus enabling a rich spatio-temporal representation that enables accurate depth forecasting. Prior work attempts to solve this problem using multi-modal input/output with supervised ground-truth data which is not practical since a large annotated dataset is required. Alternatively to prior methods, this paper forecasts depth and ego motion using only self-supervised raw images as input. The approach performs significantly well on the KITTI dataset benchmark with several performance criteria being even comparable to prior non-forecasting self-supervised monocular depth inference methods.**

I. Introduction

Forecasting the future is crucial for intelligent decision making. It is a remarkable ability of human beings to effortlessly forecast what will happen next, based on the current context and prior knowledge of the scene. Forecasting sequences in real-world settings, particularly from raw sensor measurements, is a complex problem due to the the exponential time-space space dimensionality, the probabilistic nature of the future and the complex dynamics of the scene. Whilst much effort from the research community has been devoted to video forecasting [18], [44], [50], [64] and semantic forecasting [3], [24], [53], [57], depth and ego-motion forecasting have not received the same interest despite their importance. The geometry of the scene is essential for applications such as planning the trajectory of an agent.

Anticipating is therefore important for autonomous driving auto-pilots or human/robot interaction as it is critical for the agent to quickly respond to changes in the external environment.

The first work that studied depth forecasting was carried out by Mahjourian *et al.* [42], the aim of that paper was to use the forecasted depth to render the next RGB image frame. They supervised the depth loss using ground-truth LiDAR scans and the warping was done using ground-truth poses. [47] used additional modalities for input, namely, a multi-modal RGB, depth, semantic and optical flow and forecasted the same future modalities. The supervision was carried out using the aforementioned ground-truth labels. [29] developed a probabilistic approach for forecasting using only input images and generated a diverse and plausible multi-modal future including

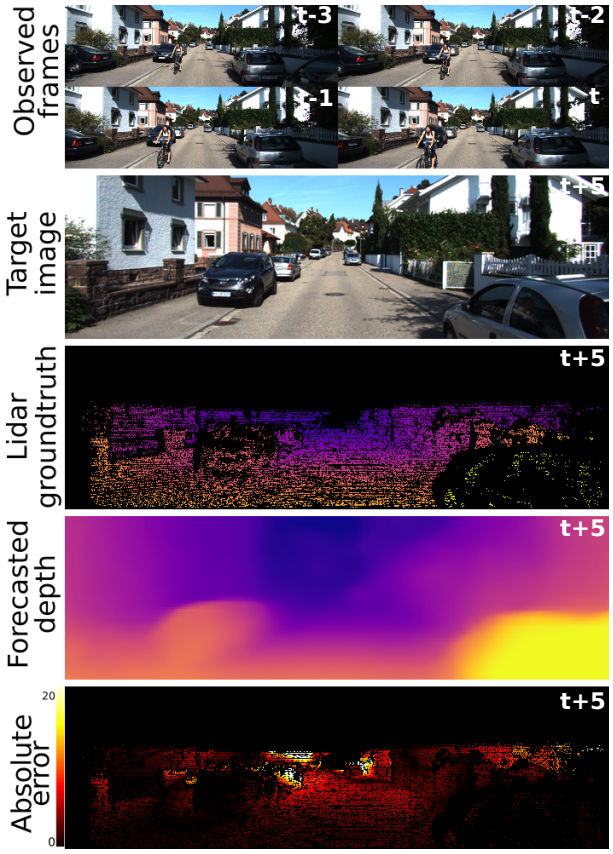

Fig. 1. The proposed method is trained using only raw sensor self-supervision, it is able to forecast an accurate geometry of the scene. The network only accesses the frames $\mathbf{I}_{t-3:t}$ and forecast the depth $\mathbf{D}_{t+n}$ and the pose ${}^{t}\mathbf{T}_{t+n}$. Not only the network forecasts accurately the ego-motion (the depth of the static object is accurate) but also handles the dynamic objects.

This paper is a preprint (Accepted in ICPR 2022).

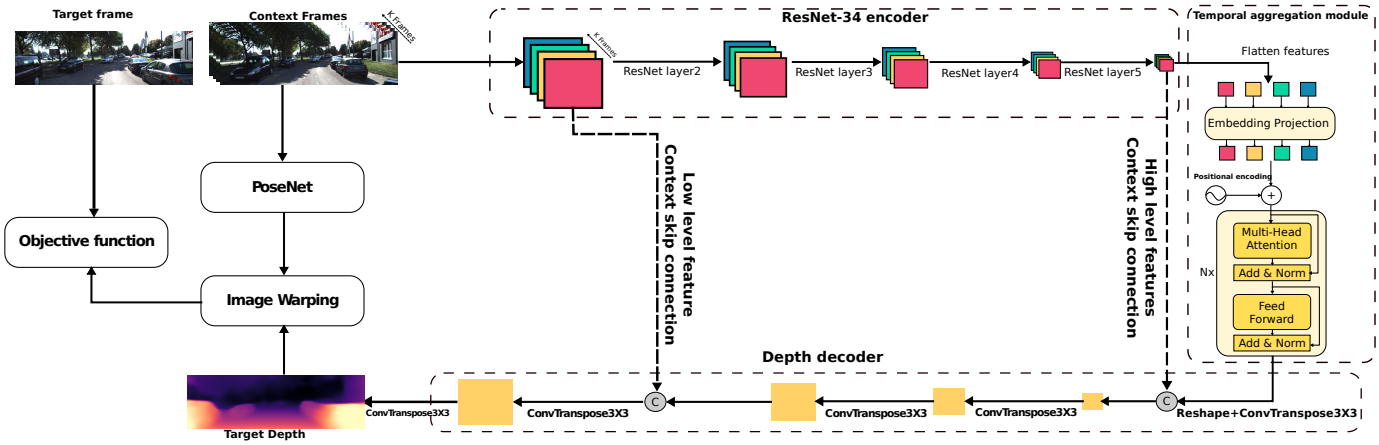

Fig. 2. Illustration of the proposed architecture. Two sub-networks are used for training: The PoseNetwork as in network [21], [32] is used to forecast the ego-motion. The depth network combines both CNN and transformers. The Resnet34 [25] encoder extracts the spatial features for each context frame. The embedding projection module projects these features into $R^{k \times d_{model}}$ where $k = 4$ is the context frames. $N = 3$ transformer encoders are used to fuse the spatial temporal to obtain a rich spatio-temporal features. The output of the transformer module encodes the motion of the scene. The decoder uses simple transposed convolution. In order to recover the context, skip connections are pooled from the encoder. Only the last frame features are pooled for the context. The decoder outputs a disparity map that will be used along with the pose network to warp the source images onto the target.

depth, semantics and optical flow. However, it was supervised through ground-truth labels and the final loss was a weighted sum of future segmentation, depth and optical flow losses similar to [47]. While these methods enable forecasting the depth, they suffer from two shortcomings: [29], [42], [47] require the ground-truth labels for supervision during training and testing and [47] uses a multi-modal input for inference that requires either ground-truth labels or a separate network.

The work presented in this paper addresses the problem of depth and ego-motion forecasting using only monocular images with self-supervision. Monocular depth and ego-motion inference has been successful for self-supervised training [1], [9], [10], [21], [22], [31], [51], [52], [60], [66]. The basic idea is to jointly learn depth and ego-motion supervised by a photometric reconstruction loss. In this paper, it is demonstrated that it is possible to extend this self-supervised training to sequence forecasting. An accurate forecasting requires a knowledge of the ego-motion, semantics, and the motion of dynamic objects. Powered by the advances of transformers [7], [14], [15], [40], [58], and using only sensor input, the network learns a rich spatio-temporal representation that encodes the semantics, the ego-motion and the dynamic objects. Therefore avoiding the need for extra labels for training and testing. The results on the KITTI benchmark [20] show that the proposed method is able to forecast the depth accurately and outperform even non-forecasting methods [17], [38], [65], [69].

II. Related work

Previous works [21], [62], [69] have used the term "prediction" for methods that infer depth from a single image irrespective of past and future events. This could easily lead to confusion since the term "prediction" could equally be employed to refer to the prediction of future events. Therefore, the term "monocular depth inference" will be used to replace this ambiguous term. As the goal of this paper is related to time-series future prediction, the term "forecasting" will be used to refer to methods [29], [42], [47] that use observations of the past and present to predict a future depth of the scene.

A. Ego-motion and monocular depth self-supervised learning

Monocular depth inference is an ill-posed problem as an infinite number of 3D scenes can be projected onto the same 2D scene. The earliest deep learning methods [16], [17] utilized LiDAR data as ground truth for supervision. For inferring pose, [32] proposed to use a simple regression network, based on a pretrained classifier [56], supervised by an Euclidean loss. The work on spatial transformers [30] offered an end-to-end differentiable warping function, thus, enabling several works [4], [10], [21], [23], [34], [51], [52], [60], [62], [68], [69] to formulate the problem as self-supervised training of depth and pose. To account for the ill-posed nature of this task, several works have addressed different challenges. [4], [10], [52], [60] proposed to enforce depth scale and structure consistency. [23], [54] proposed more robust image reconstruction losses to handle occlusion due to moving objects. [10], [34], [51], [66] proposed multi-modal training to better supervise the network. In this paper, it will be shown that it is not only possible to infer the current depth but also forecast future depths accurately.

B. Sequence forecasting

Anticipation of the future state of a sequence is a fundamental part of the intelligent decision making process. The forecasted sequence could be a RGB video sequence [2], [18], [35], [44], [50], [64], depth image sequence [42], [63], semantic segmentation sequence [3], [12], [24], [24], [29], [41], [53], [57] or even a multi-modal sequence [29], [47]. Early deep learning models for RGB video future forecasting leveraged several techniques including: Recurrent models [64],

variational autoencoder VAE [2], generative adversarial network [44], autoregressive model [50] and normalizing flows [35]. These techniques have inspired subsequent sequence forecasting methods. Despite the importance of geometry for developing better decision making, depth forecasting is still in early development. [42] used supervised forecasted depth along with supervised future pose to warp the current image and generate the future image. Instead of using images as input, [63] used LiDAR scans and forecast a sparse depth up to 3.0s in the future on the KITTI benchmark [20]. [47] used a multi-modal input/output and forecast the depth among other modalities. [29] handled the diverse future generation by utilizing a variational model to forecast a multi-modality output. The use of multi-modalities requires additional labels or pretrained networks. This makes the training more complicated. Instead, the work presented in this paper leverages only raw images and forecasts in a self-supervised manner.

C. Vision transformers

The introduction of the Transformers in 2017 [58] revolutionized natural language processing resulting in remarkable results [7], [14], [48]. The year 2020 [8], [15] marked one of the earliest pure vision transformer networks. As opposed to recurrent networks that process sequence elements recursively and can only attend to short-term context, transformers can attend to complete sequences thereby learning long and short relationships. The multi-head attention could be considered as fully connected graph of the sequence's features. It demonstrated its success by outperforming convolution based networks on several benchmarks including classification [15], [37], [67], detection [8], [36], [37] and segmentation [11], [40]. The has led to a paradigm shift [39]. transformers are slowly wining "The Hardware Lottery" [28]. However, training vision transformers is complicated as these modules are not memory efficient for images and needs a large dataset pretraining. [8] has demonstrated that it is possible to combine convolution and transformers to learn a good representation without requiring large pertaining. This paper proposes to leverage a hybrid CNN and transformer network as in [8] that is designed to forecast the geometry of the scene. The proposed network is simple and yet efficient. It outperforms even prior monocular depth inference methods [17], [38], [65], [69] that access the image of the depth frame.

III. The method

A. The problem formulation

Let $\mathbf{I}_t \in \mathbb{R}^{w \times h \times c}$ be the t-th frame in a video sequence $\mathbf{I} = \{\mathbf{I}_{t-k:t+n}\}$. The frames $\mathbf{I}_c = \{\mathbf{I}_{\mathbf{t-k:t}}\}$ are the context of $\mathbf{I}_t$ and $\mathbf{I}_f = \{\mathbf{I}_{t+1:t+n}\}$ is the future of $\mathbf{I}_t$. The goal of the future depth and ego-motion forecasting is to predict the future geometry of the scene $\mathbf{D}_{t+n}$ and the ego-motion ${}^{t+n}\mathbf{T}_t$ corresponding to $\mathbf{I}_{t+n}$ given only the context frames $\mathbf{I}_c$:

$$(\widehat{\mathbf{D}}_{t+n}, {}^{t+n}\widehat{\mathbf{T}}_t) = f(\mathbf{I}_c; \boldsymbol{\theta}) \tag{1}$$

where f is a neural network with parameters $\boldsymbol{\theta}$.

In self-supervised learning depth inference, the problem is formulated as novel view synthesis by warping the source frame $\mathbf{I}_s$ into the target frames $\mathbf{I}_{tar}$ using the depth and the ${}^{s}\mathbf{T}_{tar} \in \mathbb{SE}[3]$ pose target to source pose. The warping is defined as:

$$\widehat{\mathbf{p}}_s = \pi({}^{s}\mathbf{T}_{tar} H(\pi^{-1}(\mathbf{p}_{tar}, D(\mathbf{p}_{tar})))) \tag{2}$$

where π is the inverse camera projection defined as:

$$\pi^{-1}\big(\mathbf{p}, D(\mathbf{p})\big) = D(\mathbf{p})\Big(\frac{x - c_x}{f_x}, \frac{y - c_y}{f_y}, 1\Big)^\top \tag{3}$$

(f_x, f_y, c_x, c_y) are the camera intrinsic parameters. $\mathbf{H}$ is the homogeneous coordinates transformation and ${}^{s}\mathbf{T}_{tar} \in \mathbb{R}^{3\times4}$ is the ego-motion. Similarly, a pixel $\mathbf{p}$ is projected to a 3D point $\mathbf{P}$ given its depth $D(\mathbf{p})$ using the operator π defined as:

$$\pi(\mathbf{p}) = (f_x \frac{X}{Z} + c_x, f_y \frac{Y}{Z} + c_y) \tag{4}$$

using the spatial transformers [30], The reverse warping defined here is fully differentiable and uses a bilinear interpolation. Similarly, the self-supervised depth and ego-motion forecasting is defined here by considering the source frames $\mathbf{I}_c$ and the target $\mathbf{I}_{t+n}$ with the corresponding poses ${}^{t_i}\mathbf{T}_{t+n} : \forall i \in \{t-k, ..., t\}$. The frame $\widehat{\mathbf{I}}_{t+n}$ is obtained by reverse warping the context frames.

Reconstructing the frame $\mathbf{I}_{t+n}$ using the depth and the pose from only the context by a warping could be formulated as a maximum likelihood problem:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\mathbf{I}_{t+n}|\mathbf{I}_c; \boldsymbol{\theta}) \equiv \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_m P_{model}(\mathbf{I}^m_{t+n}|\mathbf{I}^m_c) \tag{5}$$

m is the number of samples. If P_{model} is assumed to follow a Laplacian distribution $P_{model}(\mathbf{I}_{t+n}|\mathbf{I}_c) \sim Lap(\mathbf{I}_{t+n}; \boldsymbol{\mu} = \widehat{\mathbf{I}}_{t+n}; \boldsymbol{\beta} = \sigma^2\mathbf{I})$. $\widehat{\mathbf{I}}_{t+n}$ is the warped image. Maximizing the Eq. 5 is equivalent to minimizing an L_1 error of $\widehat{\mathbf{I}}_{t+n}$ and the known image frame $\mathbf{I}_{t+n}$. Similarly, if the distribution is assumed to follow a Gaussian distribution, the maximization is equivalent to minimizing an L_2 error.

B. The architecture

The architecture of the network is depicted in Fig. 2. The forecasting network is composed of two sub-networks: a pose net to forecast the future ${}^{t_i}\mathbf{T}_{t+n}$ and a depth network that forecast $\mathbf{D}_{t+n}$ (see Eq. 1) Similar to [32], the pose-net is composed of a classification network [25] as a feature extractor followed by a simple pose decoder as in [21]. The pose network forecasts 6 parameters using the axis-angle representation. The depth network leverages a hybrid CNN and Transformer network as in [8] that is designed to forecast the geometry of the scene. This network benefits from both modules. The convolution module is used to extract the spatial features of the frames as it is memory efficient, easy to train and does not require large pretraining. The transformer module is used for better temporal feature aggregation. The multi-head attention could be considered as a fully-connected graph of the features of each frame. Therefore the information is correlated across all the frames rather than incrementally, one

step at a time, as in LSTM [27]. The architecture consists of three modules: an encoder, temporal aggregation module and a decoder.

1) Encoder: : ResNet [25] is one of the most used foundation models [5]. It has demonstrated its success as a task agnostic feature extractor for nearly all vision tasks. In this work, ResNet34 is used as feature extractor. It is pretrained on ImageNet [13] for better convergence. Each context frame is fed-forward and a pyramid of features is extracted. These features encode the spatial relationship between each scene separately. Thus, at the output of this module, a pyramid of spatial features for each frame is constructed. These features will be correlated temporally using the Temporal aggregation module TAM.

2) Temporal aggregation module: Since its introduction, transformer have demonstrated their performance outperforming their LSTM/RNN counterparts in various sequence learning benchmarks [7], [14], [49], [59]. forecasting accurate depth requires knowledge of the static objects, accurate ego-motion and knowledge of the motion of the dynamic objects. The last layer of the encoder is assumed to encode higher abstraction features (*e. g.* recognizing objects). Therefore, correlating temporally these features allows the extraction of the motion features of the scene. The TAM consists of two sub-modules:

- **Embedding projection:** The dimensions after flattening the feature output of the last layer of the encoder is not memory efficient for the transformers. The embedding projection maps these features as: $\mathbb{R}^{K\times C\times H\times W} \longrightarrow \mathbb{R}^{K\times d_{enc}}$.
- **Transformer encoder:** After projecting the features using the embedding layer, a Transformer encoder with N layer, m multi-head attention and d_{enc} is used. It correlates the spatial features of the sequence producing a spatio-temporal fused features.

3) Depth decoder: After the spatio-temporal fusion, the decoder takes these spatio-temporal features along with the context features as input and decodes them to produce a disparity map. As depicted in Fig. 2, the context of the scene is obtained by pooling the features of last frame in the encoder. Two levels are pooled and concatenated. ${}^{dec}f_{t+n} = [{}^{enc}f_t\ ,\ TAM({}^{enc}f_{t-4:t})]$. The high level features (skip connection before the TAM) enable learning the motion while the low level features (skip connection at the start of ResNet) recover the finer details lost by the down-sampling. Therefore, the decoder maps (context + motion $\longrightarrow$ depth).

Each level of the decoder consists of a simple sequential layer of: transposed convolution with kernel of 3×3 with similar channels as the encoder, batch normalization and Relu activation in that order. The forecasting head consists of a convolution with a kernel of 1×1 and a Sigmoid activation. The output of this activation, σ, is re-scaled to obtain the depth $D = \frac{1}{a\sigma+b}$, where a and b are chosen to constrain D between 0.1 and 100 units, similar to [21]. For training, each level has a forecasting head but only the last head is used for inference.

The batch is omitted

C. Objective functions

As formulated in Sec. III-A, learning the parameters $\widehat{\boldsymbol{\theta}}$ involves maximizing the maximum likelihood of P_{model}. Similar to the prior work, along with the L_1 error other additional regularization terms are used and defined as:

- **Photometric loss:** Following [21], [52], [69] The photometric loss seeks to reconstruct the target image by warping the source images using the forecast pose and depth. An L_1 loss is defined as follows:

$$\mathcal{L}_{rec}(\mathbf{I}_{t+n}, \widehat{\mathbf{I}}_{t+n}) = \sum_{\mathbf{p}} |\mathbf{I}_{t+n}(\mathbf{p}) - \widehat{\mathbf{I}}_{t+n}(\mathbf{p})| \tag{6}$$

  where $\widehat{\mathbf{I}}_{t+n}(\mathbf{p})$ is the reverse warped target image obtained by Eq. 2. This simple L_1 is regularized using SSIM [61] that has a similar objective to reconstruct the image. The final photometric loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{pe}}(\mathbf{I}_{t+n}, \widehat{\mathbf{I}}_{t+n}) = \sum_{\mathbf{p}} \big[(1-\alpha)\ \text{SSIM}[\mathbf{I}_{t+n}(\mathbf{p}) - \widehat{\mathbf{I}}_{t+n}(\mathbf{p})] \\ + \alpha|\mathbf{I}_{t+n}(\mathbf{p}) - \widehat{\mathbf{I}}_{t+n}(\mathbf{p})|\big]\end{aligned} \tag{7}$$

- **Depth smoothness:** An edge-aware gradient smoothness constraint is used to regularize the photometric loss. The disparity map is constrained to be locally smooth through the use an image-edge weighted L_1 penalty, as discontinuities often occur at image gradients. This regularization is defined as [26]:

$$\begin{aligned}\mathcal{L}_s(D_{t+n}) = \sum_{p} \big[|\partial_x D_{t+n}(\mathbf{p})|e^{-|\partial_x \mathbf{I}_{t+n}(\mathbf{p})|} + \\ |\partial_y D_{t+n}(\mathbf{p})|e^{-|\partial_y \mathbf{I}_{t+n}(\mathbf{p})|}\big]\end{aligned} \tag{8}$$

Training with these loss functions is subject to major challenges: gradient locality, occlusion and out of view-objects. Gradient locality is a result of bilinear interpolation [30], [69]. The supervision is derived from the fours neighbors of $I(\mathbf{p}_s)$ which could degrade training if that region is low-textured. Following [19], [21], [22], an explicit multi-scale approach is used to allow the gradient to be derived from larger spatial regions. A forecasting head is used at each level to obtain each level's disparity map during training. Eq. 2 assumes global ego-motion to calculate the disparity. Supervising directly using this objective is inaccurate when this assumption is violated (*e. g.* the camera is static or a dynamic object moves with same velocity as the camera). According to [21] this problem can manifest itself as 'holes' of infinite depth. This could be mitigated by masking the pixels that do not change the appearance from one frame to the next. A commonly used solution [21], [69] is to learn a mask μ that weigh the contribution of each pixel, while [69] uses an additional branch to learn this mask. This paper uses the auto-masking defined in [21] to learn a binary mask μ as follows:

$$\mu(\mathbf{I}_{t+n}, \widehat{\mathbf{I}}_{t+n}, \mathbf{I}_t) = \mathcal{L}_{pe}(\mathbf{I}_{t+n}, \widehat{\mathbf{I}}_{t+n}) < \mathcal{L}_{pe}(\mathbf{I}_{t+n}, \mathbf{I}_t) \tag{9}$$

μ is set to only include the loss when the photometric loss of the warped image $\widehat{\mathbf{I}}_{t+n}$ is lower than the original unwarped image $\mathbf{I}_t$. The final objective function is defined as:

$$\mathcal{L} = \sum_{l} [\,\mu\,\mathcal{L}_p + \alpha_d L_s\,] \tag{10}$$

Method	Forecasting	Resolution	Supervision	Abs Rel	Sq Rel	RMSE log	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al.* [17]	-	576 x 271	D	0.203	1.548	0.282	6.307	0.702	0.898	0.967
Liu *et al.* [38]	-	640 x 192	D	0.201	1.584	0.273	6.471	0.680	0.898	0.967
SfMLearner [69]	-	416 x 128	SS	0.198	1.836	0.275	6.565	0.718	0.901	0.960
Yang *et al.* [65]	-	416 x 128	SS	0.182	1.481	0.267	6.501	0.725	0.906	0.963
Vid2Depth [43]	-	416 x 128	SS	0.159	1.231	0.243	5.912	0.784	0.923	0.970
Monodepth2 [21]	-	640 x 192	SS	0.115	0.882	0.190	4.701	0.879	0.961	0.982
Wang *et al.* [60]	-	640 x 192	SS	0.109	0.779	0.186	4.641	0.883	0.962	0.982
LiDAR Train set mean	-	1240 x 374	-	0.361	4.826	0.377	8.102	0.638	0.804	0.894
ForecastMonodepth2	0.5sec	640 x 192	SS	0.201	**1.588**	0.275	**6.166**	0.702	0.897	0.960
Ours	0.5sec	640 x 192	SS	**0.178**	1.645	**0.257**	6.196	**0.761**	**0.914**	**0.964**
Copy last LiDAR scan	1sec	1240 x 374	-	0.698	10.502	15.901	7.626	0.294	0.323	0.335
ForecastMonodepth2	1sec	640 x 192	SS	0.231	**1.696**	0.303	6.685	0.617	0.869	0.954
Ours	1sec	640 x 192	SS	**0.208**	1.894	**0.291**	**6.617**	**0.701**	**0.882**	**0.949**

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON OF ON THE KITTI BENCHMARK WITH EIGEN SPLIT [20] FOR DISTANCES UP TO 80M. IN THE *Supervision* COLUMN, D REFERS TO DEPTH SUPERVISION USING LIDAR GROUNDTRUTH AND (SS) SELF-SUPERVISION. AT TEST-TIME, ALL MONOCULAR METHODS (M) SCALE THE DEPTHS WITH MEDIAN GROUND-TRUTH LIDAR.

Method	forecasting	Seq.09	Seq.10
Mean Odom	-	0.032 ± 0.026	0.028 ± 0.023
ORB-SLAM [45]	-	0.014 ± 0.008	0.012 ± 0.011
SfMLearner [69]	-	0.021 ± 0.017	0.020 ± 0.015
Monodepth2 [21]	-	0.017 ± 0.008	0.015 ± 0.010
Wang *et al.* [60]	-	0.014 ± 0.008	0.014 ± 0.010
Ours	0.5s	0.020 ± 0.011	0.018 ± 0.011

TABLE II

ATE ERROR OF THE PROPOSED METHOD AND THE PRIOR NON-FORECASTING METHODS ON KITTI [20]. THE PROPOSED METHOD IS COMPARABLE TO THESE METHODS EVEN IF IT ONLY ACCESSES PAST FRAMES.

where l is the scale level of the forecasted depth.

IV. EXPERIMENTS

A. Setting

1) KITTI benchmark [20]: : Following the prior work [17], [21], [38], [43], [60], [65], [69], the Eigen *et al.* [17] split is used with Zhou *et al.* [69] pre-processing to remove static frames. Frames without sufficient context are excluded from the training and testing. This split has become the defacto benchmark for training and evaluating depth that is used by nearly all depth methods.

2) Baselines: : As discussed above, previous work on depth forecasting has been supervised using LiDAR scans, and has used a multimodal network that provides depth. Their evaluation is not performed on the Eigen split, nor does it use the defacto self-supervised metrics. In order to fairly evaluate the proposed method, a self-supervised monocular formulation will be used to compare performance with the KITTI Eigen split benchmark. Comparisons will be made with three approaches: prior work on self-supervised depth inference [17], [21], [38], [60], [65], [69]; copy of the last observed LiDAR frame as done in [47]; and ForecastMonodepth2, a modified version of [21] that is adapted for forecasting pose/depth.

3) Evaluation metrics: : For evaluation, the metrics of previous works [17] are used for the depth. During the evaluation, the depth is capped to 80m. To resolve the scale ambiguity, the forecasted depth map is multiplied by median scaling where $s = \frac{median(D_{gt})}{median(D_{pred})}$. For the pose evaluation, the Absolute Trajectory Error (ATE) defined in [55] is evaluated on KITTI odometery benchmark [20] sequences 09 and 10.

4) Implementation details: : PyTorch [46] is used for all models. The networks are trained for 20 epochs, with a batch size of 8. The Adam optimizer [33] is used with a learning rate of $lr = 10^{-4}$ and $(\beta_1, \beta_2) = (0.9, 0.999)$. As training proceeds, the learning rate is decayed at epoch 15 to 10^{-5}. The SSIM weight is set to $\alpha = 0.15$ and the smoothing regularization weight to $\alpha_d = 0.001$. $l = 4$ scales are used for each output of the decoder. At each scale, the depth is upscaled to the target image size. $d_{model} = 2048$, $m = 16$ and $N = 3$ for the TAM projection. The input images are resized to 192×640. Two data augmentations were performed: horizontal flips with probability $p = 0.5$ and color jitter with $p = 1$. $k = 4$ frames are used for context sequence and $n = 5$ is used for short term forecasting and $n = 10$ for mid-term forecasting as in [47] which corresponds to forecasting 0.5s and 1.0s into the future. The ForecastedMonodepth2 is the same as [21] with a modified input. The context images are concatenated and used as input for both depth and pose networks.

B. Depth forecasting results

Table I shows the results of the proposed method on the KITTI benchmark [20]. As specified in Sec. IV-A2, the method is compared to three approaches: prior work on depth inference; copying last frame; and adapting monodepth2 [21] for future forecasting. The proposed method outperforms the forecasting baselines for both short and mid-term forecasting especially for short range forecating. The results are even comparable to non-forecasting methods [17], [38], [65], [69] that have access to $\mathbf{I}_{t+n}$. The gap between state-of-the-art depth inference and the proposed forecasting method is reasonable due to the uncertainty of the future, the unobservability of certain events such as a new object entering the scene and the complexity of natural videos that requires modeling correlations across space-time with much higher input dimensions.

Fig. 1 shows an example of depth forecasting on Eigen test split. Several observation can be made:

- The network handles correctly the out-of-view object (*i. e.* the bicycle on the scene).

Method	Abs Rel	Sq Rel	RMSE log	RMSE
Ours	0.178	1.645	0.257	6.196
Without TAM	0.205	1.745	0.296	6.565
Shared pose/depth features	0.208	1.745	0.282	6.529
Single scale	0.208	1.950	0.283	6.595
Disable auto-masking	0.193	1.774	0.273	6.374

TABLE III
ABLATION STUDY OF THE DIFFERENT COMPONENT OF THE PROPOSED METHOD. THE EVALUATION WAS DONE ON KITTI BENCHMARK USING EIGEN SPLIT [17]

- The network learned the correct ego-motion: The position of the static objects is accurate.

These results suggest that the network is able to learn a rich spatio-temporal representation that enable learning the motion, geometry and the semantics of the scene. Thus, extending the self-supervision depth inference to perform future forecasting with comparable results. A further analysis is done to evaluate and validate the choices of the network in Sec. IV-D.

C. Ego-Motion forecasting results

Table II shows the results of the proposed network on the KITTI odometry benchmark [20]. Similar to depth, the assessment is made by comparing with non-forecasting prior works. To avoid data leakage, the network is trained from scratch on the sequences 00-08 of the KITTI odometry benchmark. The network takes only the context images $\mathbf{I}_c$ and forecasts ${}^{t}\mathbf{T}_{t+n}$. The ATE results in Table II show the proposed network achieves a competitive result relative to other non-forecasting approaches. All the methods are trained in the monocular setting, and therefore scaled at test time using ground-truth. These results suggest that using the proposed architecture along with the self-supervised loss function successfully learns the future joint depth and ego-motion.

D. Ablation study

To further analyse the network, several ablations are made. Table III depicts a comparison of the proposed model with several variants. The evaluation is done for short-term forecasting $n = 5$ using $k = 4$ context frames.

1) Effect of the Temporal Aggregation Module: In order to evaluate the contribution of the multi-head attention, a variant of the proposed method is designed by replacing the TAM module by a simple concatenation of the last layer features. From Table III, the improvement induced by the TAM module is significant across all metrics. These results suggest that the performance obtained by the proposed method is achieved through the TAM module. Since the TAM aggregates the temporal information across all frames using a learned attention, the temporal features are better correlated and the final representation successfully encodes the spatio-temporal relationship between the images.

2) Effect of sharing the encoder of depth and ego-motion: Since both pose and depth networks encode the future motion and geometry of the scene, it is expected that sharing the encoders of these networks yield better results. However, as reported by Table III, the degradation is significant. Even though these tasks are collaborative, sharing the encoder will result in a set of parameters $\hat{\boldsymbol{\theta}}$ that are neither the best local optima for the depth nor for the pose. By alleviating this restriction and separating the encoders, the network learns better local optima for both pose and depth.

3) The benefit of using multiple scales: In order to evaluate the multi-scale extension, a variant of the proposed method that uses only one scale is trained. As illustrated in the Table III The network benefits from the multi-scale. The reverse warping uses bi-linear interpolation. As mentioned earlier, each depth point depends only on the four neighboring warped points. By using a multi-scale depth at training-time the gradient is derived from a larger spatial region directly at each scale.

4) Effect of auto-masking: Table III compares the proposed method with a variant without using the auto-masking defined is Sec. III-C. The results show that using auto-masking improves all four evaluation criteria. This demonstrates that using auto-masking, for pixels that do not change appearance, reject these outliers that inhibit the optimization. This leads to better accuracy of the forecasted depth.

E. Limitations and perspective

Even though the proposed forcasting method yields good results, there exists a gap with respect to non-forecasting methods. Several limitations contribute to this:

- A common assumption across SOTA methods is that the environment is deterministic and that there is only one possible future. However, this is not accurate since there are multiple plausible futures. Given the stochastic nature of the forecasting proposed here, the network will tend to forecast the mean of all the possible outcomes [2].
- The network does not forecast the correct boundaries of the objects. This is due to the formulation as a maximum likelihood problem with a Laplacien distribution assumption and the deterministic nature of the architecture. As a result, the boundaries of the dynamic objects are smoothed.
- Due to the problem formulation, the scale of the forecasted depth is ambiguous.
- The model fails to account for the motion of distant dynamic objects due to lack of parallax. and fine objects are ignored by the network.

To address these limitations future work will investigate constraining depth to be conditional on the input image structure as in [6]. Concerning scale, one solution could be to infer directly from the scene as in [4], [10], [60].

V. CONCLUSION

This paper proposed an approach for forecasting future depth and ego motion using only raw images as input. This problem is addressed as end-to-end self-supervised forecasting of the future depth and ego motion. Results showed significant performances on several KITTI dataset benchmarks [20]. The performance criteria are even comparable with non-forecasting self-supervised monocular depth inference methods [17], [38],

[65], [69]. The proposed architecture demonstrates the effectiveness of combining the inductive bias of the CNN as a spatial feature extractor and the multi-head attention of transformers for temporal aggregation. The proposed method learns a spatio-temporal representation that captures the context and the motion of the scene.

REFERENCES

[1] Yasin Almalioglu, Muhamad Risqi U. Saputra, Pedro P.B. De Gusmao, Andrew Markham, and Niki Trigoni. GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In *Proceedings - IEEE International Conference on Robotics and Automation*, volume 2019-May, pages 5474–5480, 2019.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy Campbell, and Sergey Levine. Stochastic variational video prediction. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[3] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[4] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32:35–45, 2019.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Houssem-eddine Boulahbal, Adrian Voicila, and Andrew Comport. Are conditional gans explicitly conditional? *arXiv preprint arXiv:2106.15011*, 2021.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020-December, 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[9] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29:730–738, 2016.

[10] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:7062–7071, 2019.

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021.

[12] Hsu Kuang Chiu, Ehsan Adeli, and Juan Carlos Niebles. Segmenting the Future. *IEEE Robotics and Automation Letters*, 5(3):4202–4209, 2020.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186, 2019.

This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-011011931). The authors would like to acknowledge the Association Nationale Recherche Technologie (ANRT) for CIFRE funding (n°2019/1649).

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 2650–2658, 2015.

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 3, pages 2366–2374, 2014.

[18] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016.

[19] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016.

[20] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[21] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October(1):3827–3837, 2019.

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6602–6611, 2017.

[23] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:8976–8985, 2019.

[24] Colin Graber, Grace Tsai, Michael Firman, Gabriel Brostow, and Alexander Schwing. Panoptic segmentation forecasting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 2279–2288, 2021.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Philipp Heise, Sebastian Klose, Brian Jensen, and Alois Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2360–2367, 2013.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[28] Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

[29] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic Future Prediction for Video Scene Understanding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12361 LNCS, pages 767–785, 2020.

[30] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[31] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 2015-Janua:2017–2025, 2015.

[32] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-dof camera relocalization. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:2938–2946, 2015.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Marvin Klingner, Jan Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12365 LNCS:582–600, 2020.

[35] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. *arXiv preprint arXiv:1903.01434*, 2019.

[36] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2021.

[37] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021.

[38] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2016.

[39] Ruiyang Liu, Yinghui Li, Dun Liang, Linmi Tao, Shimin Hu, and Hai-Tao Zheng. Are we ready for a new paradigm shift? a survey on visual deep mlp. *arXiv preprint arXiv:2111.04060*, 2021.

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[41] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann Lecun. Predicting Deeper into the Future of Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 648–657, 2017.

[42] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Geometry-based next frame prediction from monocular video. In *IEEE Intelligent Vehicles Symposium, Proceedings*, pages 1700–1707, 2017.

[43] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[44] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[45] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[47] Xiaojuan Qi, Zhengzhe Liu, Qifeng Chen, and Jiaya Jia. 3D motion decomposition for RGBD future dynamic scene synthesis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 7665–7674, 2019.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[50] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. In *VISIGRAPP 2021 - Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 101–112, 2021.

[51] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 12232–12241, 2019.

[52] Vitor Guizilini Rares, Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2020.

[53] Josip Saric, Marin Orsic, Tonci Antunovic, Sacha Vrazic, and Sinisa Segvic. Warp to the Future: Joint Forecasting of Features and Feature Motion. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10645–10654, 2020.

[54] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-Metric Loss for Self-supervised Learning of Depth and Egomotion. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12364 LNCS:572–588, 2020.

[55] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[57] Adam M Terwilliger, Garrick Brazil, and Xiaoming Liu. Recurrent flow-guided semantic forecasting. In *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pages 1703–1712, 2019.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[59] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021.

[60] Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, and Huchuan Lu. Can Scale-Consistent Monocular Depth Be Learned in a Self-Supervised Scale-Invariant Manner? *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12727–12736, 2021.

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[62] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021.

[63] Xinshuo Weng, Jianren Wang, Sergey Levine, Kris Kitani, and Nicholas Rhinehart. Inverting the pose forecasting pipeline with spf2: Sequential pointcloud forecasting for sequential pose forecasting. *arXiv preprint arXiv:2003.08376*, 2020.

[64] Gabriel Kreiman William Lotter and David Cox. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016.

[65] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Thirty-Second AAAI conference on artificial intelligence*, 2018.

[66] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.

[67] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[68] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian M. Reid. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[69] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6612–6621, 2017.

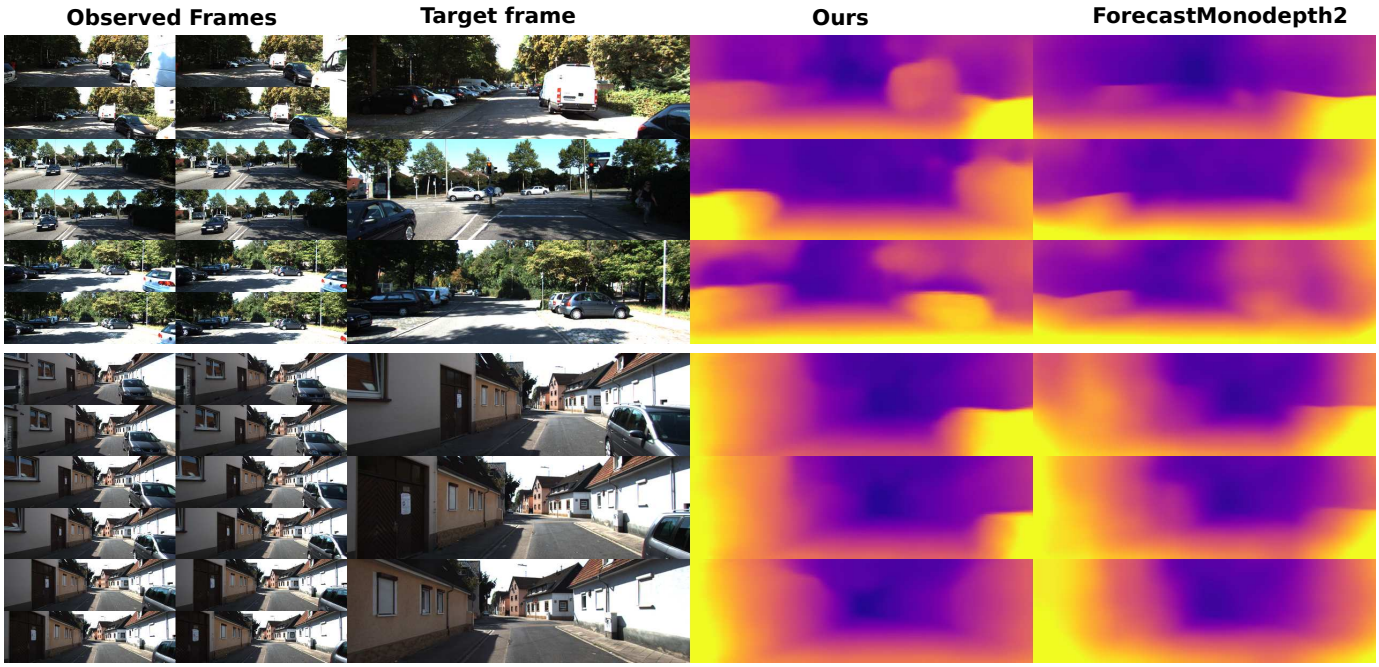

Fig. 3. Qualitative results of the comparison of the proposed method with the ForecastMonodepth2 baseline. This comparison shows that the proposed method performs better than the baseline, especially for nearby dynamic objects. This observation is further validated in Table IV. In addition, the baseline method is showing a lack of detection of moving objects, which leads to a degradation of the forecasted depth. The proposed method is able to detect moving objects, thus accurately forecasting the depth of the scene.

Range	Method	Forecasting	Abs Rel	Sq Rel	RMSE log	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
[00m 10m]	Monodepth2 [21]	-	0.066	0.264	0.106	1.076	0.959	0.987	0.994
	ForecastMonodepth2	0.5s	0.138	**0.586**	0.178	1.697	0.847	0.957	0.985
	Ours	0.5s	**0.112**	0.595	**0.155**	**1.573**	**0.893**	**0.964**	**0.986**
[10m 30m]	Monodepth2 [21]	-	0.119	0.858	0.192	3.706	0.876	0.956	0.978
	ForecastMonodepth2	0.5s	0.192	1.673	0.258	5.169	0.725	0.906	0.963
	Ours	0.5s	**0.167**	**1.453**	**0.241**	**4.803**	**0.782**	**0.921**	**0.965**
[30m 80m]	Monodepth2 [21]	-	0.188	3.094	11.115	0.288	0.709	0.897	0.950
	ForecastMonodepth2	0.5s	**0.213**	**3.526**	**11.940**	**0.292**	**0.631**	**0.874**	**0.953**
	Ours	0.5s	0.224	4.052	12.638	0.312	0.622	0.862	0.941

TABLE IV

QUANTITATIVE PERFORMANCE COMPARISON ON THE KITTI BENCHMARK WITH EIGEN SPLIT [20] FOR MULTIPLE DISTANCES RANGE. FOR ABS REL, SQ REL, RMSE AND RMSE LOG LOWER IS BETTER, AND FOR $\delta < 1.25$, $\delta < 1.25^2$ AND $\delta < 1.25^3$ HIGHER IS BETTER. THREE RANGES ARE CONSIDERED: SHORT RANGE [0 10M] WHICH REPRESENTS 37.95%, MEDIUM-RANGE [10 30]WHICH REPRESENTS 50.74% AND LONG-RANGE[30 80] WHICH REPRESENTS 11.30%. THE RESULTS SHOWS THAT THE PROPOSED METHOD IS ABLE TO FORECAST GOOD DEPTH AND OUTPERFORM THE BASELINE AT SHORT AND MEDIUM FORECASTING RANGE.

VI. MORE QUALITATIVE RESULTS

In order to further analyse the depth forecasting results, an assessment based on the ground-truth LiDAR distance is done. Table IV shows the comparison of the non-forecasting method Monodepth2 [21], ForecastMonodepth2 and the proposed methods.

The results suggest that the proposed method outperform the adaptation of Monodepth2 for short-range with a improvement of the Abs Rel of -16.7% and medium-range with a improvement of the Abs Rel of -8.8%. These regions are the most significant regions of the forecasting as it has enough parallax for the ego-motion and dynamic object motion. Besides, this region assess several challenges including out-of-view objects and occlusion. For the long-range forecasting, the results shows that the two methods performer badly due to the lack of parallax in these region and down-sampling that ignore small objects. Moreover, this region has a high likelihood of new-object entering the scene which the forecasting is unable to handle by definition. The reported performances and the qualitative results suggest that the two forecasting networks only fit the road and completely ignore any other object. These results are shown qualitatively in Fig. 3.