Dense Semantic Forecasting in Video by Joint Regression of Features and Feature Motion

Josip Šarić, Sacha Vražić, and Siniša Šegvić, Member, IEEE

Abstract—Dense semantic forecasting anticipates future events in the video by inferring pixel-level semantics of an unobserved future image. We present a novel approach that is applicable to various single-frame architectures and tasks. Our approach consists of two modules. The feature-to-motion (F2M) module forecasts a dense deformation field that warps past features into their future positions. The feature-to-feature (F2F) module regresses the future features directly and is, therefore, able to account for emergent scenery. The compound F2MF model decouples the effects of motion from the effects of novelty in a task-agnostic manner. We aim to apply F2MF forecasting to the most subsampled and the most abstract representation of the desired single-frame model. Our design takes advantage of deformable convolutions and spatial correlation coefficients across neighboring time instants. We perform experiments on three dense prediction tasks: semantic segmentation, instance-level segmentation, and panoptic segmentation. The results reveal state-of-the-art forecasting accuracy across three dense prediction tasks.

Index Terms—Computer vision, deep learning, dense semantic forecasting, future prediction.

I. INTRODUCTION

VISUAL perception is an important challenge toward the development of autonomous robots and vehicles. A recent discovery of deep learning methods triggered great progress in single-image dense prediction tasks, such as instance segmentation [1] or panoptic segmentation [2]. However, the reasoning of an intelligent agent must not be limited to the present moment in time since the consequences of our current actions and the goals of our missions occur in the future. Consequently, the anticipation of future events could be an important ingredient toward making our present systems better and more intelligent.

Early visual forecasting methods relied on handcrafted models of scene dynamics [3]. However, this approach is prone to systematic errors due to insufficient modeling accuracy. For instance, popular camera models [4] are unable to express arbitrary lens distortions [5]. Likewise, object-level forecasting by geometric filtering [6] is vulnerable to propagation of detection and association errors. Consequently, recent work attempts to implicitly capture the laws of scene dynamics through deep learning in video [7]. This can be carried out by forecasting either future RGB frames [8], [9] or the corresponding semantic content [10]–[12].

Semantic forecasting is an appealing approach since decision-making systems mainly care about the content of future scenes rather than about their appearance [13]. Empirical evidence indicates that direct semantic forecasting performs much better than inferring from a forecast RGB frame while requiring significantly less computational effort [11]. Intuition suggests similar conclusions. It seems appropriate to first figure out what is going to happen at an abstract level before moving on to RGB pixels and the associated details of lighting, reflectance, and surface normals. Indeed, it appears that semantic forecasting should be a prerequisite for RGB forecasting rather than vice versa.

There are multiple approaches for expressing semantic forecasting. Semantics-to-semantics (S2S) forecasting maps observed semantic maps into their future counterparts [11].
Motion-to-motion (M2M) forecasting receives the optical flow between observed images and produces the optical flow between the future image and the last observed image [14]. Finally, feature-to-feature (F2F) forecasting operates on abstract convolutional features [12]. There are several advantages of F2F forecasting with respect to the remaining two approaches. In comparison with S2S, it offers better spatiotemporal correspondence, more expressive features, and task-agnosticism. In comparison with M2M, it offers semantic reasoning and in-painting the novel scenery.

This article builds on previous feature-level forecasting approaches [12], [15] and enriches them with the following contributions. We propose feature-to-motion (F2M) forecasting that expresses future features by warping their current counterparts throughout a dense deformation field. This formulation expresses feature-level forecasting as a causal relationship between the past and the future. However, F2M forecasting is unable to predict the emergence of novel scenery. Hence, we finally propose F2MF forecasting (F2M-and-feature) by blending F2M and F2F with densely regressed weight factors, as illustrated in Fig. 1. This allows our method to foresee the emergence of unobserved scenery and exploit that information for decoupling variation due to novelty from variation due to motion. We account for the geometric nature of our task by leveraging spatiotemporal correlation coefficients and deformable convolutions. We show that dense semantic forecasting can be expressed in terms of the most compressed representation of the single-frame model, and such an organization can lead to competitive accuracy. The resulting generalization performance surpasses the state of the art by a wide margin. Our method is very well suited for real-time implementation due to low computational overhead with respect to single-frame prediction [16]. The method can be adapted for other dense prediction tasks, which we explain next.

In addition to contributions from previous conference accounts [16], [17], here, we extend F2MF forecasting with two new single-frame architectures [1], [2] and show that single-level forecasting can preserve instance-related information. Thus, this is the first forecasting method that can be applied to three dense prediction tasks: semantic segmentation, instance-level segmentation, and panoptic segmentation. We note that transition toward instance-aware tasks is not trivial since the corresponding metrics (AP and PQ) aggregate instance-level instead of pixel-level recognition quality. Hence, correct recognition of small objects becomes much more important than in semantic segmentation. This article also presents the computational advantages of our method by comparing its complexity with other dense semantic forecasting approaches. Finally, it demonstrates that our method is able to generalize over different cameras, resolutions, and frame rates, and produce longer term forecasts in an autoregressive manner.

II. RELATED WORK

Semantic forecasting anticipates semantic contents of an unobserved future image. Conceptually, this task can be factorized into RGB forecasting [8] and semantic prediction in the forecast image. However, we prefer to carry it out as a single processing step [11] due to the advantages stated in the introduction. In particular, we consider forecasting three dense prediction tasks: semantic segmentation [18], instance segmentation [1], and panoptic segmentation [2]. Our method warps features from observed images into their future positions, which makes it related to optical flow [19]. Our work is most related to previous dense semantic forecasting approaches F2F [12] and M2M [14].

A. Optical Flow

Optical flow is a dense 2-D-motion field between neighboring image frames \(I_t\) and \(I_{t+1}\). A future image \(I_{t+1}\) can be approximated either by forward warping \(I_t\) with the forward flow \(f_{t+1}^{fw} = \text{flow}(I_t, I_{t+1})\) or backward warping it with the backward flow \(f_{t+1}^{bw} = \text{flow}(I_{t+1}, I_t)\) [20]

\[
I_{t+1} \approx \text{warp}_{fw}(I_t, f_{t+1}^{fw}) \approx \text{warp}_{bw}(I_{t+1}, I_t). \tag{1}
\]

The approximate equality reminds us that bijective mapping cannot be established due to occlusions and disocclusions.

Recent optical flow methods leverage deep convolutional models [21], [22] due to the capability to guess motion where the correspondence is absent or ambiguous. These approaches exploit explicit 2-D correlation across the local neighborhood [21] and local embeddings that act as a correspondence metric [23]. Our method is especially related to self-supervised approaches [19] since they learn to reconstruct optical flow without any ground-truth information.

B. RGB Forecasting

RGB forecasting is also known as video prediction [8]. The task is to predict one or more future video frames given a few recently observed frames of the same scene [7]. This is especially interesting due to the opportunity for self-supervised representation learning on practically unlimited data.

Mathieu et al. [8] express RGB forecasting as image generation with a multiscale adversarial network. Vukotić et al. [24] embed the desired temporal offset in the latent representation of the observed scenery. Reda et al. [25] warp observed frames by applying a regressed kernel at the location determined by the forecast flow. Some works generate video from still images by warping the image with forecast flow [26], [27]. Our approach forecasts multiple flows that warp multiple previous features toward a single future feature tensor. This allows resolving some disocclusions by warping from a suitable past image.

Some works decompose RGB forecasting into reconstruction from the past and in-painting the novel scenery [9], [26], [28]. Their forecast includes the future warp and the respective disocclusion map. The past frame is first warped with forecast flow, and then, the disocclusions are filled by in-painting. This setup is conceptually similar to the proposed F2MF approach; however, there are two important differences as follows. First, we exploit spatiotemporal correlation features and deformable convolutions. Second, we perform the forecast on heavily subsampled (16 × or 32 ×) abstract features instead of pixels. This allows our method to operate on Mpixel images.
and achieve state-of-the-art accuracy in semantic forecasting while incurring only a modest computational overhead with respect to single-frame prediction.

C. Dense Semantic Prediction

Several computer vision tasks address scene understanding at the pixel level. Semantic segmentation [18] assigns each pixel to a suitable semantic class. This includes stuff classes, such as road or vegetation, and object classes, such as person or bicycle. Instance segmentation [1] detects instances of object classes and associates them with the respective image regions. Panoptic segmentation [2] subsumes the previous two tasks by assigning pixels both the semantic class and the instance index. Today, all these tasks are solved by fully convolutional models for dense prediction.

Several successful architectures for dense semantic prediction have an asymmetric hourglass-shaped structure consisting of the recognition backbone and a lean upsampling data path [2], [29], [30]. The recognition backbone usually corresponds to a fully convolutional portion of a model designed for image classification. The role of this component is to convert the input image into a subsampled latent representation. Most authors exploit knowledge transfer from ImageNet since image-wide supervision requires much less effort than dense semantic supervision. The upsampling data path recovers dense semantic predictions by blending semantics of deep features with the location accuracy of their shallow counterparts. The blending is usually implemented by means of skip connections from the backbone toward the upsampling path. Empirical studies show advantages of asymmetric designs that complement deep and thick recognition with shallow and thin upsampling [31]. This suggests that recognition requires much more capacity than guessing the borders when rough semantics is known.

D. Forecasting at the Level of Semantic Predictions (S2S)

Luc et al. [11] were the first to propose direct semantic forecasting. Their S2S model maps past semantic segmentation into the future semantic segmentation. Bhattacharyya et al. [32] try to account for multimodal future with variational inference based on the MC dropout while conditioning the forecast on measurements from the vehicle odometer. Nabavi et al. [33] formulate the forecasting in a recurrent fashion with shared parameters between each two frames. Chen and Han [34] improve their work by leveraging deformable convolutions and enforcing temporal consistency between neighboring feature tensors. Their attention-based blending is related to forward warping based on pairwise correlation features [17]. However, the forecasting accuracy of these approaches is considerably lower than in our ResNet-18 experiments despite greater forecasting capacity and better single-frame performance. This suggests that ease of correspondence and avoiding error propagation may be important for successful forecasting. Concurrent work [35] forecasts panoptic segmentation by separately regressing things and stuff from previous predictions.

E. Flow-Based Forecasting (M2M)

Direct semantic forecasting requires a lot of training data due to the necessity to learn all motion patterns one by one. This can be improved by allowing the forecasting model to access geometric features that reflect 2-D motion in the image plane [36]. Further development of that idea brings us to warping the last dense prediction according to forecast optical flow. A prominent instance of this approach can be succinctly described as M2M forecasting since it receives optical flows from three observed frames and produces the future optical flow. The corresponding implementation based on convolutional LSTM had achieved state-of-the-art semantic forecasting accuracy prior to our work [14]. This approach is related to our F2M module, which also forecasts by warping with the regressed flow. However, our F2M module operates on abstract convolutional features and does not require external computational resources and additional supervision for training and evaluating the flow model. Our approach discourages error propagation due to end-to-end training and implies very efficient inference due to subsampled resolution and feature sharing between motion reconstruction and dense recognition. In addition, we take into account features from all past four frames instead of relying only on the last prediction. This allows our F2M module to detect complex disocclusion patterns and simply copy from the past where possible. Furthermore, our module has access to raw semantic features that are complementary to flow patterns [37] and often strongly correlated with future motion (consider, e.g., cars versus pedestrians). Finally, we complement our F2M module with pure recognition-based F2F forecasting that outperforms F2M on previously unobserved scenery.

F. Feature-Level Forecasting (F2F)

F2F forecasting maps past features to their future counterparts. The first F2F approach operated on image-wide features from a fully connected AlexNet layer [38]. Luc et al. [12] propose dense F2F forecasting by regressing all features along the FPN-style [29] upsampling path. Further work [15], [39] improves by using a convolutional LSTM module at each level of the feature pyramid and proposing interlevel connections for context sharing. However, forecasting at fine resolution is computationally expensive [40]. Hence, we propose single-level forecasting of the coarsest features [16], [41]. Such an approach is advantageous due to small interframe displacements, rich contextual information, and small computational footprint [16]. It also has an intuitive appeal of prioritizing the big picture before moving on to the fine details. The most recent work [42] follows and extends our idea of single-level feature forecasting. They introduce a variational autoencoder, which provides a compressed representation of a multilevel feature pyramid. After the forecasting, this representation is decoded into tensors of the FPN style decoder.

Vora et al. [43] formulate feature-level forecasting as reprojection of reconstructed features to the forecast future ego-location. However, such an approach underperforms in presence of (dis-)occlusions and large changes of perspective. In addition, it makes it difficult to account for the independent
Fig. 2. Details of F2MF forecasting. F2M and F2F modules receive a processed concatenation of features from observed frames and their spatiotemporal correlation coefficients. The F2M module regresses future feature flow that warps past features into their future locations. The F2F module forecasts the future features directly. The compound forecast $\hat{X}_{t+\Delta t}$ is a weighted blend (B) of F2M and F2F forecasts.

motion of moving objects. The large empirical advantage of our method suggests that optimal forecasting performance requires a careful balance between reconstruction and recognition, as well as that explicit 3-D reasoning may bring minor benefits.

G. Semantic Forecasting in Video

Dense semantic forecasting has been used to increase the level of supervision in partially labeled video [44]. This could be especially useful when only few [45] or weak [46] labels are available. Semantic forecasting can improve inference in the video by alleviating disturbances due to noisy input [3]. Learned object dynamics could especially improve few-shot approaches that rely on inference-time optimization [47].

III. DENSE SEMANTIC FORECASTING WITH F2MF

We present a novel forecasting approach that combines standard feature-level forecasting (F2F) with its regularized variant (F2M) that warps past representations into the future. Fig. 2 illustrates the structure of the proposed joint model. On the input, we receive $T = 4$ past feature tensors $X_{t-9}$, $X_{t-6}$, $X_{t-3}$, and $X_t$ ($X_{t-9,3}$ for short) extracted with the front end of the desired single-frame model for dense semantic prediction. Our F2MF model maps the past features into the future feature tensor $\hat{X}_{t+\Delta t}$. This tensor is transformed to semantic predictions $\hat{S}_{t+\Delta t}$ throughout the back end of the chosen single-frame model as shown in Fig. 1. Note that $\Delta t$ denotes the temporal offset between the last observed frame and our forecast.

F2MF forecasting proceeds as follows. Input features are concatenated across the semantic dimension and fused with a single convolutional layer in order to reduce the number of channels. In parallel, the correlation module computes pairwise correlation coefficients across a local neighborhood. The fused representation is concatenated with the correlation features. The result is further processed through six convolutional layers in order to recover the shared representation for F2M and F2F forecasting. The weight module (block W in Fig. 2) outputs five feature maps: $w_{t-9}^{F2M}$, $w_{t-6}^{F2M}$, $w_{t-3}^{F2M}$, $w_t^{F2M}$, and $w_t^{F2F}$. The four maps $\{w_t^{F2M}\}, t \in \{t-9, t-6, t-3, t\}$, represent contributions of the four past feature tensors in the F2M forecast. The fifth map $w_t^{F2F}$ represents the contribution of the F2F head in the compound forecast. All convolutional layers are implemented as BN-ReLU-dconv, where dconv stands for deformable convolution [48].

A. F2M Module

The F2M module assumes that the future can be explained as a geometrical transformation of the observed past. Therefore, it outputs a dense field of motion vectors $\hat{f}_{t+\Delta t}$ for each of the $T = 4$ input feature tensors $X_t$, $t \in \{t-9, t-6, t-3, t\}$. The future tensors $\hat{X}_{t+\Delta t}^{(r)}$ are estimated by backward warping [20] the corresponding input features $X_t$ with the regressed flow $\hat{f}_{t+\Delta t}$. The resulting warped tensors are subsequently blended with regressed weights that we activate with per-pixel softmax. Thus, the F2M forecast is a weighted sum of warped features from the observed images

$$\hat{X}_{t+\Delta t}^{F2M} = \text{warp}_{\text{bw}}(X_t, \hat{f}_{t+\Delta t})$$

$$\alpha = \text{softmax}(\{a_t^{F2M}\}_{t \in \{t-9, t-3\}})$$

This allows the F2M module to choose the most suitable previous image for forecasting a particular region. Such opportunity is particularly beneficial in some occlusion patterns as we illustrate in Fig. 10.

However, the assumption that the future can be reconstructed from the past is only partially true since the future often brings unpredictable novelty. In addition, sometimes, the correspondence will be hard to determine due to large ego-motion and changes in perspective. Consequently, some parts of the scene are going to be particularly hard to forecast by simple warping from the past. Accurate forecasting of such regions will require imagination ability as we describe next.

B. F2F Module

The F2F module directly regresses future features $\hat{X}_{t+\Delta t}^{F2F}$ from the shared representation. It consists of a single BN-ReLU-dconv unit, where the number of output feature maps is the same as in a single past feature tensor. As it is not bound to reconstruction from the past, it has a chance to in-paint the features in novel regions. This is similar to some previous [12] and concurrent work [15], [41]; however, there are three important differences. First, we aim at single-level F2F forecasting. Second, we use deformable convolutions. Third, our F2F module has access to spatiotemporal correlation features that relieve the need to learn correspondence from scratch. Our experiments show a clear advantage of these novelties. The contribution of correlation features suggests that correspondence is not easily learned on existing datasets.

C. Compound F2MF Forecasting

We hypothesize that F2M and F2F forecasts might be complementary, and their combination might lead to improved
accuracy. Consequently, our F2MF module formulates future features as a weighted average of the F2M and F2F forecasts (as before, $\tau \in \{t - 9 : t : 3\})$
\[
\hat{X}_{t+\Delta t}^{F2MF} = \beta^{F2F} \cdot \hat{X}_{t+\Delta t}^{F2F} + \sum_{\tau} \beta^{F2M}_{\tau} \cdot \hat{X}_{t+\Delta t}^{(\tau)}
\]
\[
\beta = [\beta^{F2F} || \beta^{F2M}] = softmax([W^{F2F} || W^{F2M}]). \tag{5}
\]
This formulation encourages specialization of the F2F module for the novel parts of the scene. It also relaxes the penalty of the F2M module in such regions and allows it to focus on parts where correspondence can be established. Nevertheless, this kind of separation might weaken the learning signals within the two individual modules. Consequently, we propose to train F2MF forecasting with three loss terms. The main loss $L_{F2MF}$ involves the F2MF forecast (5). The two auxiliary losses $L_{F2F}$ and $L_{F2M}$ affect the corresponding outputs $\hat{X}_{t+\Delta t}^{F2F}$ and $\hat{X}_{t+\Delta t}^{F2M}$. All three losses are formulated as mean squared L2 distance between the forecast and actual features computed with the single-frame model in the future frame. Consequently, the model can be trained with self-supervision on unlabeled video.

D. Correlation Module

Our correlation module determines spatiotemporal correspondence between neighboring frames. On the input, it receives convolutional features $X_{t-9:3}$ in the form of a $T \times C \times H \times W$ tensor. On the output, it produces spatiotemporal correlation coefficients across a $d \times d$ neighborhood for each of the $T - 1$ pairs of neighboring frames.

The module first embeds features from all time instants into a space with enhanced metric properties by a shared $3 \times 3 \times C'$ convolution ($C' = 128$). We hypothesize that this mapping might recover distinguishing information, which is not needed for single-frame inference. Subsequently, we construct our metric embedding by normalizing $C'$-dimensional features to unit norm so that cosine similarity becomes dot product. This metric embedding by normalizing $C'$-dimensional features to

This approach is also advantageous since low spatial resolution encourages the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture. We propose to forecast succinct abstract features in order to encourage the forecasting module to focus on the big picture.

F. Forecasting Semantic Segmentation

In the case of semantic segmentation, we achieve our best results with custom single-frame models without any skip connections. The single-frame penalty of this design decision is around three-percentage point (pp) mean intersection over union (mIoU) on Cityscapes val. Despite this handicap, our F2MF forecasting addresses 512-D features produced by the SPP module, as well as that the semantic forecast is formed by the upsampling path of the single-frame model.

G. Forecasting Instance Segmentation

In the case of instance segmentation, we choose a third-party implementation of a Mask R-CNN without skip connections [1]. The single-frame penalty of this design decision
is around 1-pp AP COCO and 2.7-pp AP50 on Cityscapes val. Despite this handicap, our model performs favorably with respect to the state of the art. Fig. 4 shows that our F2MF forecasting addresses 1024-D features produced by the second-to-last residual block of the ResNet-50 backbone. Subsequently, the RPN head extracts object candidates from the forecast features. Instance segmentations are obtained by processing each object candidate with the last residual block and the two Mask R-CNN heads. Future semantics are formed by resizing the inferred instance segmentations to input resolution.

H. Forecasting Panoptic Segmentation

In the case of panoptic segmentation, we train a custom Panoptic DeepLab model [2] with only one skip connection. The single-frame penalty of this design decision is around 2.3-pp PQ on Cityscapes val. Despite this handicap, our model performs favorably with respect to the state of the art. Fig. 4 shows that our F2MF forecasting addresses 1024-D features produced by the second-to-last residual block of the ResNet-50 backbone. Subsequently, the forecast features are processed by the last residual block and the two upsampling paths. Future semantics are finally formed by postprocessing object centers and per-pixel semantic information.

IV. EXPERIMENTS

We train our method on the Cityscapes train [49] and evaluate on the Cityscapes val test and the CamVid test. We consider short- and mid-term forecasting experiments that target $\Delta t = 3$ and $\Delta t = 9$ time steps into the future (180 and 540 ms, respectively) [11]. We estimate the forecasting performance by computing the usual dense prediction metrics in the future frame. We use mIoU for semantic segmentation, COCO average precision (AP) for instance segmentation, and panoptic, segmentation, and recognition qualities (PQ, SQ, and RQ) for panoptic segmentation.

Our training procedure involves several separate steps. First, the backbone of a single-frame model is pretrained on ImageNet. Second, we train the single-frame model on labeled images in the task-specific setup. Finally, we train the forecasting module on unlabeled images to map the past features to their future counterparts with MSE loss. We extract all training features (both the past and the future) by applying the single-frame model to the corresponding video frames. Note that the forecasting module does not require any annotations in spite of being trained in a supervised fashion.

We train our model for 160 epochs with an ADAM [50] optimizer and early stopping. We set the initial learning rate to $4 \cdot 10^{-4}$ and reduce it to $10^{-7}$ through cosine annealing. We implement our model in Pytorch. Training with cached features takes 12 h on a single GTX1080Ti. Training with online data augmentation requires two days and three GPUs. During training, we normalize inputs and outputs of our F2MF model to dataset-wide zero mean and unit variance. During inference, we normalize the input and denormalize the output. This facilitates the training process and improves generalization. We set the weights of all three components of the F2MF loss to 1. Our data augmentation policy includes sliding the input tuple across the video clip and horizontal flipping. We refrain from using any auxiliary information, such as vehicle odometry or depth maps.

A. Semantic Segmentation Forecasting on Cityscapes

We experiment with two single-frame models of different capacities. Both of them have an encoder–decoder architecture [51] without skip connections. The encoder corresponds to standard ImageNet classification models. The decoder consists of a spatial pyramid pooling and three upsampling modules, as shown in Fig. 3. The two models differ in backbone architectures and the width of the decoder. The weaker model is based on a ResNet-18 and uses 128 feature maps along the upsampling path. It achieves 72.5% mIoU on Cityscapes val. The stronger model uses a DenseNet-121 backbone, 512 feature maps in the SPP, 256 maps in the first upsampling module, and 128 maps in the last two upsampling modules. It achieves 75.8% mIoU on Cityscapes val.

Table I shows the forecasting accuracy for semantic segmentation on Cityscapes val. The top section shows the
TABLE I
EVALUATION OF OUR F2MF MODEL FOR SEMANTIC SEGMENTATION FORECASTING ON CITYSCAPES VAL. ALL DENOTES ALL CLASSES, MO—MOVING OBJECTS, D.A.—DATA AUGMENTATION, AND †—TEST SET

| Model               | Accuracy (mIoU) | Short term: Δt=3 All | Mid term: Δt=9 All | Oracle (ours) | F2F 4× [12] | ConvLSTM F2F 4× [15] | ApaNet [39] | F2F-Corr (ours) | F2MF (ours) |
|---------------------|-----------------|----------------------|--------------------|---------------|-------------|----------------------|------------|----------------|-------------|
| Oracle-DN121        | 75.8            | 75.2                 | 75.8               | 36.3          | 36.1        | 36.3                 | 36.1       | 36.3          | 36.1        |
| Oracle-RN18         | 72.5            | 71.5                 | 72.5               | 37.3          | 65.8        | 37.3                 | 65.8       | 37.3          | 65.8        |
| Copy last (DN121)   | 53.7            | 48.7                 | 39.1               | 9.6           | 22.9        | 2.2                  | 8.1        | 2.2           | 8.1         |
| 3Dcnn-F2F [41]      | 56.0            | /                    | /                  | 36.3          | 36.1        | 36.3                 | 36.1       | 36.3          | 36.1        |
| DinLSTM-S2S [11]    | 50.4            | 55.3                 | 47.8               | 37.3          | 65.8        | 37.3                 | 65.8       | 37.3          | 65.8        |
| Mask-F2F [12]       | 60.1            | 51.2                 | 41.2               | 22.1          | 44.3        | 37.3                 | 65.8       | 37.3          | 65.8        |
| FeatProj3D [43]     | 61.5            | 54.4                 | /                  | 23.2          | 46.1        | 12.9                 | 29.2       | 12.9          | 29.2        |
| Bayesian S2S [32]   | 65.1            | /                    | /                  | 21.2          | 43.3        | 9.4                  | 19.2       | 9.4           | 19.2        |
| LSTM AM S2S [34]    | 65.8            | /                    | /                  | 23.6          | 47.2        | 11.5                 | 24.2       | 11.5          | 24.2        |
| LSTM M2M [14]       | 67.1            | 65.1                 | 51.5               | 70.2          | 68.7        | 59.1                 | 56.3       | 70.2          | 68.7        |
| F2MF-RN18 w/o d.a.  | 66.9            | 65.6                 | 55.9               | 52.4          | 52.4        | 52.4                 | 52.4       | 52.4          | 52.4        |
| F2MF-DN121 w/o d.a. | 68.7            | 66.8                 | 56.8               | 53.1          | 53.1        | 53.1                 | 53.1       | 53.1          | 53.1        |
| F2MF-DN121 w/ d.a.  | 69.6            | 67.7                 | 57.9               | 54.6          | 54.6        | 54.6                 | 54.6       | 54.6          | 54.6        |
| F2MF-DN121 w/ d.a. †| 70.2            | 68.7                 | 59.1               | 56.3          | 56.3        | 56.3                 | 56.3       | 56.3          | 56.3        |

TABLE II
INSTANCE SEGMENTATION FORECASTING ON CITYSCAPES VAL

| Model               | Short term: Δt=3 AP | Mid term: Δt=9 AP | Oracle (ours) | COPY LAST |
|---------------------|----------------------|--------------------|---------------|----------|
| AP                  |                      |                    |               |          |
| P0                  |                      |                    |               |          |
| R0                  |                      |                    |               |          |
| PQ                  |                      |                    |               |          |
| SQ                  |                      |                    |               |          |
| RQ                  |                      |                    |               |          |
| Oracle (ours)       | 36.3                 | 36.1               | 36.3          | 36.1     |
| Copy last           | 9.6                  | 22.9               | 2.2           | 8.1      |
| F2F 4× [12]         | 19.4                 | 39.9               | 7.7           | 19.4     |
| ConvLSTM F2F 4× [15]| 22.1                 | 44.3               | 11.2          | 25.6     |
| ApaNet [39]         | 23.2                 | 46.1               | 12.9          | 29.2     |
| F2F-Corr (ours)     | 21.2                 | 43.3               | 9.4           | 19.2     |
| F2MF (ours)         | 23.6                 | 47.2               | 11.5          | 24.2     |

TABLE III
PANOPTIC SEGMENTATION FORECASTING ON CITYSCAPES VAL

| Model               | Short term: Δt=3 PQ | Mid term: Δt=9 PQ | Oracle (ours) | COPY LAST |
|---------------------|----------------------|--------------------|---------------|----------|
| PQ                  |                      |                    |               |          |
| SQ                  |                      |                    |               |          |
| RQ                  |                      |                    |               |          |
| Oracle (ours)       | 57.5                 | 79.7               | 70.8          |          |
| Copy last           | 32.3                 | 70.9               | 68.1          | 30.2     |
| F2F-Corr (ours)     | 43.3                 | 74.4               | 55.5          | 36.0     |
| F2MF (ours)         | 47.3                 | 75.1               | 60.6          | 43.3     |

B. Instance Segmentation Forecasting on Cityscapes

Table II presents the accuracy of our forecasting approach on the instance segmentation task and compares it to the related work on Cityscapes val. In the case of semantic segmentation, we use single-level forecasting in order to reduce the memory footprint and improve the speed of training. We demonstrate that our approach is a plug-in solution for any dense-prediction model without skip connections, by experimenting on a public Mask R-CNN model [52] that we do not customize (cf. Fig. 4). The chosen model (R50-C4) attaches the RPN prediction head to 16× subsampled features produced by the penultimate residual block of ResNet-50. We compare our approach with previous forecasting approaches [12], [15], [39] that use Mask R-CNN with FPN upsampling [29]. The two previous approaches require forecasting all four levels of the feature pyramid. The table shows that our approach achieves the best short-term accuracy in spite of considerably weaker oracle and significantly lower computational complexity. The F2F-Corr baseline ablates the F2M module from our design. The resulting performance reveals a significant contribution of F2M forecasting in the case of instance segmentation.

C. Panoptic Segmentation Forecasting on Cityscapes

Table III presents our performance on the panoptic segmentation task. Our single-frame model is a custom Panoptic Deeplab [2] with ResNet-50 backbone and a single skip connection from the backbone to the decoder/upsampling path (cf. Fig. 5). As in the case of instance segmentation, we target the features at 16× subsampled resolution from the penultimate residual block. To the best of our knowledge, this is the first attempt at forecasting panoptic segmentation. The table clearly shows that our F2MF model outperforms the copy-last baseline by a large margin. The performance of the F2F-Corr baseline reveals a significant contribution of F2M forecasting in the case of panoptic segmentation.

Tables I–III show that F2MF forecasting can be applied to three different dense semantic prediction tasks and, therefore, confirm that our method indeed is task-agnostic.

D. Qualitative Results

Fig. 6 visualizes our short- and mid-term forecasts, and compares them with our oracle. The columns show the last observed frame, the oracle prediction overlaid on top of the future frame, our F2MF forecast overlaid on top of the future frame, and the F2F heatmap $\beta_{F2F} = 1 - \beta_{F2M}$. The red regions correspond to pixels forecast by warping with the F2M flow, while the blue regions correspond to pixels forecast by the F2F module. The blue pixels usually
correspond to unoccluded scenery, which has to be imagined by the model, since it was not visible in any of the observed frames. We observe that the forecast F2M heatmaps are quite accurate in all images, which suggests that F2MF is versatile and task agnostic.

The top two rows of Fig. 6 correspond to semantic segmentation. Most pixels from row 1 of this group are predicted by F2M forecasting because there is little motion in the scene. On the other hand, there is a large blue region in the bottom left of the F2M heatmap in row 2. The blue region was disoccluded by the car passing by the camera. As it was never observed before, the F2M module does not stand a chance, so the F2F module has to in-paint novel content. We observe that the prediction is sound although it misses some of the people behind. We also provide a video presentation of our mid-term forecasting performance on the Frankfurt video.  

The middle two rows illustrate instance forecasting on Cityscapes val. We observe that our model accurately predicts the future leg stance for some of the pedestrians in row 3. The F2M heatmap reveals that most human pixels are forecast by the F2F module except for the heads that exhibit more regular motion than the bodies. Furthermore, our model correctly predicts the future position of the moving taxi in row 4. The corresponding F2M heatmap reveals that the F2F module in-paints the features in the disoccluded area. Of course, we are unable to forecast the objects behind the taxi.

The bottom two rows address panoptic forecasting on Cityscapes val. Notice that pixels corresponding to object classes are colored in different shades of the original class color. We observe that the forecast segmentation appears sharp and accurate. Similarly, as in semantic segmentation forecasting, we observe that the F2F module is in charge of novel regions. This is best seen in the last row, where the car on the right leaves the scene and dis-occludes a large part of the background. The model correctly forecasts that the pixels behind the car correspond to buildings, sidewalks, and roads.

E. Ablation and Validation Study

Table IV shows validation experiments that quantify contributions from our F2F, F2M, and correlation modules. We perform this study on semantic segmentation forecasting with our single-frame model based on ResNet-18. Note that

---

1https://www.youtube.com/watch?v=Wgo_5xnvn8M
we do not use data augmentation in this experiment. This enables caching the features on the SSD drive and, therefore, allows significantly faster training. Every model configuration presented in the table is separately trained.

First, we compare F2F and F2M forecasting individually. In experiments without the correlation module, F2F outperforms F2M for 0.6 mIoU points. In experiments with the correlation module, F2F outperforms F2M for 0.7 mIoU points at the short-term forecast, and they perform equally in the mid-term forecast. These results are plausible because the F2F approach is more expressive. While F2M has to explain the future by warping past features, F2F is unconstrained and can imagine anything. Nevertheless, the compound F2MF model always outperforms any of the two single-head approaches. In a setup without the correlation module, F2MF outperforms F2M by 1.0–1.2 mIoU points and the individual F2F by 0.4–0.6 mIoU points. In a setup with the correlation module, F2MF outperforms the F2M by 1.3–1.4 mIoU points and the F2F by 0.6–1.4 mIoU points. These results hint that F2M and F2F are indeed complementary.

The correlation module improves the performance in all setups. It contributes 0.8 and 2.3 mIoU points in short- and mid-term F2M forecasts, respectively. Likewise, the correlation module improves the F2F performance for 0.9 and 1.7 mIoU points and boosts the compound F2MF model for 1.1–2.5 mIoU points at short term and mid-term, respectively.

We further investigate the complementary nature of the F2F and F2M approaches by comparing the accuracy of independent single-head models in stratified groups of pixels. Previous results showed that the F2F model performs slightly better in general. However, we know that predicting the future in novel regions is particularly hard for the F2M model. Therefore, we hypothesize that F2M might perform better in the previously observed scenery. We divide pixels into bins according to the \( \beta \) values as predicted by the compound F2MF model and then show the forecasting accuracy of the two independently trained F2F and F2M models in Fig. 7. On the left \( y \)-axis, we show per-bin forecasting accuracy. On the right \( y \)-axis, we show the relative share of pixels in that particular bin. We show the F2M weights \( \beta_{F2M} \) on the \( x \)-axis for each bin. The figure separately considers the short- (left) and mid-term (right) forecastings. We observe that the pixel histogram is skewed toward higher \( \beta_{F2M} \) values. This suggests that the F2MF model delegates the majority of the pixels to the F2M module. We also observe that this effect is less pronounced in the mid-term forecast. This makes sense because we observe more novelty there. The plot also shows that the F2M model performs better in pixels with higher \( \beta_{F2M} \) values, which justifies the delegating decision from the compound model. These plots support our hypothesis that the two approaches are complementary.

Table V validates the number of input frames by exploring F2MF accuracy for different look-back windows. Models that observe only the most recent frame perform poorly as they are unable to estimate the scene dynamics with a single reference point. Note that these models do not have the correlation module because at least two frames are needed to compute spatiotemporal coefficients. Models with two to five input frames perform comparably well, while the model with four input frames performs the best. This confirms the suitability of the default forecasting setup [11] on Cityscapes. Please note that mid-term experiments with five input frames are not feasible for \( \Delta t = 3 \) due to the limited length of Cityscapes clips.

We have also attempted to forecast three and nine steps into the future by observing input frames with stride 2 \((t \in \{ t − 6, t − 4, t − 2, t \})\). Our F2MF-RN18 model achieved 67.3 mIoU in the short-term forecast and 55.6 mIoU in the mid-term forecast. Thus, forecasting with input stride 2 improved for 0.4 points with respect to the standard setup at short term while underperforming for 0.3 points at mid-term. This suggests that mid-term forecasting profits from a larger look-back window, which can also be observed in Table V.

Table VI validates the contribution of the proposed weight module, which outputs five dense feature maps used to blend the four F2M forecasts and the F2F forecast. We compare the proposed design with two baselines. The first baseline simply averages all of the five feature forecasts. The second baseline predicts five scalars that represent the weights for each of the five future feature predictions. This baseline uses tensor-wide instead of the original per-pixel blending. The table shows that tensor-wide weights outperform simple averaging by 1.4 mIoU points in the short-term forecast and 0.7–1.3 mIoU points in

---

**TABLE IV**

ABLATION OF CORRELATION, F2F, AND F2M ON CITYSCAPES VAL. STANDALONE F2F AND F2M MODELS ARE TRAINED INDEPENDENTLY

| Configuration (F2MF-RN18) | F2F | F2M | Correlation | Short-term mIoU | Mid-term mIoU |
|--------------------------|-----|-----|-------------|----------------|--------------|
| F2F                      | ✓   | ✓   | ✓           | 64.8           | 52.2         |
| F2M                      | ✓   | ✓   | ✓           | 65.4           | 52.8         |
| Correlation              | ✓   | ✓   | ✓           | 65.8           | 53.4         |
|                          | ✓   | ✓   |             | 65.6           | 54.5         |
|                          | ✓   | ✓   |             | 66.3           | 54.5         |
|                          | ✓   | ✓   |             | 66.9           | 55.9         |

---

**TABLE V**

VALIDATION OF THE NUMBER OF INPUT FRAMES ON CITYSCAPES

| Input frames (F2MF-RN18) | Short-term mIoU | Mid-term mIoU |
|--------------------------|-----------------|---------------|
|                          | All MO          | All MO        |
| \{t\}                    | 57.9            | 45.7          |
| \{t − 3, t\}             | 66.4            | 54.9          |
| \{t − 6, t − 3, t\}      | 66.9            | 55.1          |
| \{t − 9, t − 6, t − 3, t\} | 66.9           | 55.9          |
| \{t − 12, t − 9, t − 6, t − 3, t\} | 66.7          | /             |

---

**TABLE VI**

VALIDATION OF THE NUMBER OF FUTURE FEATURE PREDICTIONS

| Future feature predictions | Short-term mIoU | Mid-term mIoU |
|----------------------------|-----------------|---------------|
|                            | All MO          | All MO        |
| Simple Averaging           | 64.8            | 52.2          |
| Tensor-Wide Weights        | 65.6            | 53.4          |

---

**Fig. 7.** Histograms of F2F and F2M accuracy and overall pixel incidence over \( \beta_{F2M} \) bins as inferred by the compound F2MF model on Cityscapes val for short- (left) and mid-term (right) forecastings.
TABLE VI

| Blending method (F2MF-RN18) | Short-term mIoU | Mid-term mIoU |
|----------------------------|-----------------|---------------|
|                            | All             | MO            | All             | MO             |
| mean weights               | 65.3            | 63.7          | 54.6            | 50.4           |
| image-level weights        | 66.7            | 65.1          | 55.3            | 51.7           |
| per-pixel weights          | 66.9            | 65.6          | 55.9            | 52.4           |

Fig. 8. Dependence of the semantic segmentation accuracy of our autoregressive models on different forecasting offsets on Frankfurt videos from Cityscapes val. FT denotes autoregressive fine-tuning.

the mid-term forecast. The proposed per-pixel weights further improve the performance for 0.2–0.5 mIoU points in the short-term forecast and 0.6–0.7 mIoU points in the mid-term forecast.

F. Autoregressive Long-Term Forecasting

We investigate the autoregressive application of our short-term model for long-term forecasting. Each subsequent invocation of our forecasting approach takes the precedent forecast as the most recent input. This enables forecasting an arbitrary number of time steps into the future. We evaluate the performance only on Frankfurt scenes from Cityscapes val because video clips from other cities have only very short videos. We fine-tune our best short-term model for autoregressive forecast by accumulating the loss in timesteps $t+3$, $t+6$, and $t+9$ and backpropagating gradients through time. Fig. 8 shows the mIoU accuracy for different forecasting times both for the regular (F2MF-AR) and the fine-tuned model (F2MF-AR-FT).

G. Cross-Dataset Generalization

We investigate whether our feature-based forecasting generalizes to unseen data beyond the validation subset, as proposed in [11]. We aim to test our models on the CamVid dataset [53], which differs from Cityscapes in terms of image resolution, the video frame rate, and the set of semantic labels. No adaptation is needed regarding the image resolution because our model is fully convolutional. We sample CamVid videos each five timesteps, which is approximately equal to three timesteps in Cityscapes. We evaluate the mIoU accuracy according to a mapping from 19 Cityscapes classes to 11 CamVid classes.

Table VII shows the accuracy of our single-step and autoregressive models and also the autoregressive approach from [11]. The columns show the oracle accuracy (mIoU), the forecast accuracy (mIoU), and the relative performance drop w.r.t. to the oracle model. We observe that our approach achieves the highest accuracy and also the lowest relative performance drop.

H. Interpreting the Model Decisions

This section further investigates the difference between F2MF and F2F forecastings for semantic segmentation. Fig. 10 considers two particular output pixels and visualizes pixels with top 1‰ gradients of log-max-softmax w.r.t. to the input frames (red dots). Columns 1–4 show the four input frames. Columns 5 and 6 show our forecast and the ground truth. Pairs of rows correspond to two output pixels that are designated with a green square in the predictions and a yellow square in the observed frames. Rows 1 and 2 consider a pixel located at the rear bicycle wheel in the future frame. We observe that most of the gradients are located at the corresponding spot in the most recent frame for both models. However, the gradients of the F2MF model are less spread and more focused toward the corresponding input pixel. Rows 3 and 4 consider a pixel, which is occluded by the passing bicyclist in the most recent frame, while being disoccluded in the future frame. We observe that the F2F model relies on the context by looking around the pixel location in the most recent frame. On the other hand, the F2MF model realizes that this part of the scene had been visible in the most distant frame. Thus, our F2MF module makes the forecast just by copying the representation from the most distant feature tensor. This shows that our F2MF model is able to learn to handle complex occlusion patterns.

I. Computational Complexity

We analyze the computational complexity of our methods and compare it with the corresponding state of the art. We assume that the system is able to cache the necessary past activations and, therefore, consider the computational effort, which is required to make the forecast after acquiring the
current RGB frame. We report the count of multiply-and-accumulate operations (MAC) as measured with \texttt{thop} [54].

Table VIII compares our F2MF-based semantic segmentation forecasting with the runner-up method [14] from Table I. We observe that our forecasting incurs 12× less complexity. The advantage remains considerable (4×) even if we consider a comparison with single-frame semantics included. The difference mostly stems from the fact that our forecasting model operates on 1/32 resolution and does not require high-resolution optical flow. The approach [14] has to evaluate PSP-Net and FlowNet2-C at 0.5 MPx and forecast flow features on 1/8 resolution. Our estimate assumes that their forecast requires evaluation of a single ConvLSTM cell with 74 feature maps both in the hidden state and the input [14].

Table IX compares our F2MF-based instance-level forecasting with the multilevel F2F approach [12]. The table shows the cost of single-frame inference and the cost of forecasting at each subsampling level. Our single-frame inference is a bit inefficient and requires around 1.6× more compute than FPN-based Mask R-CNN. Nevertheless, our total forecasting cost is 18× smaller than multilevel F2F [12]. This is because we only forecast features that are 16× subsampled w.r.t. to the input resolution. On the other hand, multilevel approaches [12], [15] additionally have to forecast expensive fine-resolution features at 1/8× and 1/4× resolutions. We expect that [15] would suffer from the same problem although their source code is not available.

Table X shows unoptimized execution profiles of our forecasting models on a GTX1080Ti GPU. We measure the average time for three inference stages: feature extraction in four input frames, feature forecasting, and semantic formation. Sections correspond to three dense prediction tasks: semantic segmentation (top), instance segmentation (middle), and panoptic segmentation (bottom). We observe that forecasting is much faster than feature extraction and single-frame inference. F2MF-RN18 allows real-time inference since online execution environments allow to cache the past features and perform feature extraction in only one frame—the most recent one.

V. CONCLUSION

Anticipation of future semantics is a prerequisite for intelligent planning of current actions. Recent work addresses this problem by implicitly capturing laws of scene dynamics throughout deep learning in the video. However, the existing
approaches are unable to distinguish disoccluded and emerging scenery from previously observed parts of the scene. This is suboptimal since the former requires pure recognition, while the latter can be explained by warping. Different from all previous approaches, our method predicts the emergence of unobserved scenery and exploits that information for disentangling variation caused by novelty from variation due to motion.

Our method performs dense semantic forecasting on the feature level. Different than previous such approaches, we regularize the forecasting process by expressing it as a causal relationship between the past and the future. The proposed F2M forecasting generalizes better than the classic F2F approach at many (but not all) image locations. We achieve the best of both worlds by blending F2M and F2F predictions with densely regressed weight factors. We empirically confirm that low F2M weights occur at unobserved scenery. The resulting F2MF approach surpasses the state of the art in semantic segmentation forecasting on the Cityscapes dataset by a wide margin.

We complement convolutional features with their respective correlation coefficients organized within a cost volume over a small set of discrete displacements. Our forecasting models use deformable convolutions in order to account for the geometric nature of F2F forecasting. These two improvements bring a clear advantage in all three feature-level approaches: F2F, F2M, and F2MF. To the best of our knowledge, this is the first account of panoptic forecasting in the scientific literature.

Unlike previous methods, our single-frame model for semantic segmentation does not use skip connections along the upsampling path. Consequently, we are able to forecast condensed abstract features at the far end of the downsampling path with a single F2MF module. This greatly improves the inference speed and favors the forecasting accuracy due to the coarse resolution and high semantic content of the involved features. We were unable to outperform this approach with multilevel F2F forecasting in spite of significantly better single-frame accuracy.

We also propose an adaptation of the proposed F2MF method for two additional dense prediction tasks: instance segmentation and panoptic segmentation. These experiments use third-party single-frame models and, therefore, show that our method can be successfully used as a drop-in solution for converting any kind of dense prediction model into its competitive forecasting counterpart. To the best of our knowledge, this is the first account of panoptic forecasting in the scientific literature.

The proposed method offers many exciting directions for future work. In particular, our method does not address the multimodal future, which is a key to long-term forecasting and worst case reasoning. Other suitable extensions include overcoming obstacles toward end-to-end training, extension to RGB forecasting, and enforcing temporally consistent predictions in neighboring video frames.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. ICCV, 2017, pp. 2961–2969.
[2] B. Cheng et al., “Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12475–12485.
[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-time single camera SLAM,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 1052–1067, Jun. 2007.
[4] Z. Zhang, “Camera calibration: A personal retrospective,” Mach. Vis. Appl., vol. 27, no. 7, pp. 963–965, Oct. 2016.
[5] I. Krešo and S. Šegvić, “Improving the egomotion estimation by correcting the calibration bias,” in Proc. 10th Int. Conf. Comput. Vis. Theory Appl., X. Zhou et al., pp. 347–356, Sep. 2015.
[6] S. Reuter, B.-T. Vo, B.-N. Vo, and K. Dietmayer, “The labeled multi-Bernoulli filter,” IEEE Trans. Signal Process., vol. 62, no. 12, pp. 3246–3260, Jun. 2014.
[7] S. Oprea et al., “A review on deep learning techniques for video prediction,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Dec. 15, 2020, doi: 10.1109/TPAMI.2020.3045007.
[8] M. Mathieu, C.ouprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” in Proc. ICLR, 2016, pp. 1–14.
[9] H. Gao, H. Xu, Q.-Z. Cai, R. Wang, F. Yu, and T. Darrell, “Disentangling propagation and generation for video prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9006–9015.
[10] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 961–970.
[11] P. Luc, N. Neverova, Couprie, J. Verbeek, and Y. LeCun, “Predicting deeper into the future of semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 648–657.
[12] P. Luc, C.ouprie, Y. Lecun, and J. Verbeek, “Predicting future instance segmentation by forecasting convolutional features,” in Proc. ECCV, 2018, pp. 584–599.
[13] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, “Ego-centric vision-based future vehicle localization for intelligent driving assistance systems,” in Proc. Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 9711–9717.
[14] A. Terwilliger, G. Brazil, and X. Liu, “Recurrent flow-guided semantic forecasting,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2019, pp. 1703–1712.
[15] J. Sun et al., “Predicting future instance segmentation with contextual pyramid ConvLSTMs,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2043–2051.
[16] J. Šaric, M. Oršić, T. Antunović, S. Vražić, and S. Šegvić, “Single level feature-to-feature forecasting with deformable convolutions,” in Proc. German Conf. Pattern Recognit. Cham, Switzerland: Springer, 2019, pp. 189–202.
[17] J. Šaric, M. Oršić, T. Antunović, S. Vražić, and S. Šegvić, “Warp to the future: Joint forecasting of features and feature motion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10648–10657.
[18] B. Zhou et al., “Semantic understanding of scenes through the ADE20K dataset,” Int. J. Comput. Vis., vol. 127, no. 3, pp. 302–321, Mar. 2019.
[19] P. Liu, M. Lyu, I. King, and J. Xu, “SelFlow: Self-supervised learning of optical flow,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4571–4580.
[20] R. Szeliski, Computer Vision: Algorithms and Applications. London, U.K.: Springer, 2010.
[21] A. Dosovitskiy et al., “Flownet: Learning optical flow with convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2758–2766.
[22] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PW-CNet: CNNs for optical flow using pyramid, warping, and cost volume,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8934–8943.
[23] J. Zhontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 2287–2318, Oct. 2016.
[24] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. van Gemert, “One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network,” in Proc. Int. Conf. Image Anal. Process. Cham, Switzerland: Springer, 2017, pp. 140–151.
[25] F. A. Reda et al., “SDC-Net: Video prediction using spatially-displaced convolution,” in Proc. ECCV, 2018, pp. 718–733.
[26] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, “Flow-grounded spatial-temporal video prediction from still images,” in Proc. ECCV, 2018, pp. 600–615.
[27] J. Pan et al., “Video generation from single semantic label map,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3733–3742.
