Warp to the Future: Joint Forecasting of Features and Feature Motion

Josip Šarić[1] Marin Oršić[1] Tonći Antunović[2] Sacha Vražić[2] Siniša Šegvić[1]
[1]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
[2]Rimac Automobili, Sveta Nedelja, Croatia

Abstract

We address anticipation of scene development by forecasting semantic segmentation of future frames. Several previous works approach this problem by F2F (feature-to-feature) forecasting where future features are regressed from observed features. Different from previous work, we consider a novel F2M (feature-to-motion) formulation, which performs the forecast by warping observed features according to regressed feature flow. This formulation models a causal relationship between the past and the future, and regularizes inference by reducing dimensionality of the forecasting target. However, emergence of future scenery which was not visible in observed frames can not be explained by warping. We propose to address this issue by complementing F2M forecasting with the classic F2F approach. We realize this idea as a multi-head F2MF model built atop shared features. Experiments show that the F2M head prevails in static parts of the scene while the F2F head kicks-in to fill-in the novel regions. The proposed F2MF model operates in synergy with correlation features and outperforms all previous approaches both in short-term and mid-term forecast on the Cityscapes dataset.

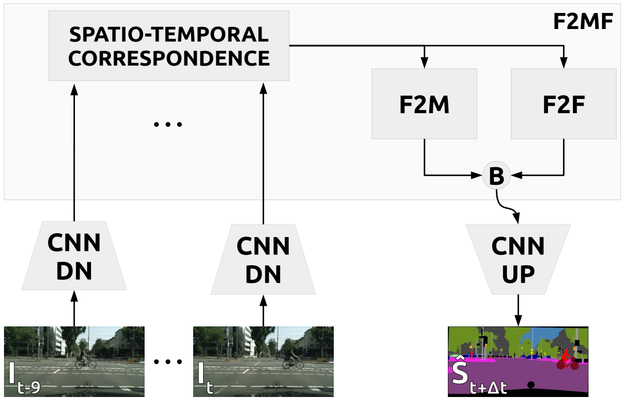

Figure 1. Overview of the proposed F2MF forecasting approach. Observed RGB images I_τ, $\tau \in \{t-9, t-6, t-3, t\}$, are processed into low-resolution features $\mathbf{X}_\tau$ by a pre-trained recognition module (CNN DN). The features are enriched with their spatio-temporal correlation coefficients and forwarded to F2M and F2F modules which specialize for forecasting previously observed and novel scenery. Forecasted future features $\hat{\mathbf{X}}_{t+\Delta t}$ are a blend (B) of F2M and F2F outputs. Dense predictions $\hat{\mathbf{S}}_{t+\Delta t}$ are finally recovered through a pre-trained upsampling module (CNN UP).

1. Introduction

Anticipated future [1, 29, 30] is invaluable input to many decision making systems. For example, in autonomous driving, future pedestrian location could enable potentially life-saving decisions. Models for forecasting future events can often be trained on unlabeled videos, which are an inexhaustible source of training data. Some recent work [20, 35, 26] addresses forecasting future RGB frames given the past frames. However, this difficult task is not required in many interesting applications. For instance, in the autonomous driving context, we are more concerned about future semantics [37] than about future appearance. Hence, semantic forecasting [18] represents an interesting alternative with clear potential to improve accuracy and speed.

Several approaches have been proposed for future anticipation on the semantic level. Direct semantic forecasting maps past predictions into future ones [18, 14, 2, 21, 4, 38]. Unfortunately, this approach risks propagating single-frame prediction errors into the forecast. Additionally, successful forecasting requires establishing at least implicit correspondence across the past frames, which is not easily achieved at the level of final predictions. Finally, this approach can not be realized in a task agnostic manner.

Flow-based forecasting operates on dense image motion vectors [34]. It receives reconstructed optical flow from the past few frames and targets the optical flow between the future frame and the last observed frame. Future predictions can be recovered by warping past predictions with the forecasted flow. However, this approach requires pre-computed optical flow, which implies separate training and decreases inference speed. Additionally, purely geometric forecasting can not take advantage of semantic information and generate ad-hoc content in disoccluded pixels.

Feature-level forecasting receives intermediate features from the past frames, and targets their future counterparts. This approach has been successfully applied for semantic segmentation [37, 28, 5], instance segmentation [17, 7, 32] and action recognition [36]. In comparison with the previous two approaches, feature-level forecasting stands a better chance to avoid propagating single-frame prediction errors, since features are not constrained to commit to particular predictions. Additionally, deep convolutional representations are typically subsampled w.r.t. input which allows for efficient implementations in terms of memory footprint and computational speed. There is also a promising potential for end-to-end training and task-agnostic operation.

Most previous feature-level approaches express forecasting as a pure recognition task [17, 28, 32]. This does not appear satisfactory since it ignores the geometric nature of future anticipation and makes it difficult for the model to disentangle individual factors of variation. The forecasting problem involves several geometric degrees of freedom such as camera motion, depth, and individual object motion. We hypothesize that learning and inference would be easier if some of these factors were explicitly present in the model. Feature-level forecasting can also be expressed as a pure 3D reconstruction task [37] given reconstructed ego-motion and depth. However, 3D interpretation may introduce undesired noise , while perhaps not being necessary for achieving optimal performance. This especially concerns the process of ”imagining” unobserved scenery. Hence, we prefer to formulate the forecast as 2D motion of previously observed structure plus 2D generation of novel scenery.

This paper expresses feature-level forecasting by disentangling variation caused by motion from variation due to novelty. Our contributions are as follows. First, we improve feature-based forecasting by enriching features with their spatio-temporal correlation coefficients across the local neighbourhood. This promotes generalization across semantic classes and simplifies establishing temporal correspondence. Second, we model variation due to motion by warping observed features with regressed feature flow. We denote this procedure as F2M (feature-to-motion) forecasting in order to emphasize its relation towards the classic F2F (feature-to-feature) approach [17]. Third, we leverage the complementary nature of F2F and F2M approaches by blending their forecasts according to densely regressed weight factors, as illustrated in Figure 1. The proposed F2MF forecasting model outperforms the classic F2F approach by improving the accuracy in previously observed regions, and encouraging the F2F module to focus on ”imagining” the novel scenery. F2M forecast can be implemented either with forward or backward warping [33]. The two approaches achieve equally good performance in our experimental setup. Experiments on the Cityscapes dataset show clear advantage of F2MF forecasting over the classic F2F approach, both at short-term and mid-term period.

2. Related work

Optical flow. Optical flow reconstructs dense 2D-motion between neighbouring image frames I_t and I_{t+1}. The flow can be defined either in the forward or in the backward direction. The future image I_{t+1} can be approximated either by forward warping [33] previous image I_t with the forward flow $\mathbf{f}_t^{t+1}$= $\mathrm{flow}(\mathrm{I}_t, \mathrm{I}_{t+1})$, or by backward warping I_t with the backward flow $\mathbf{f}_{t+1}^{t}$= $\mathrm{flow}(\mathrm{I}_{t+1}, \mathrm{I}_t)$:

$$\mathrm{I}_{t+1} \approx \mathrm{warp\_fw}(\mathrm{I}_t, \mathbf{f}_t^{t+1}) \approx \mathrm{warp\_bw}(\mathrm{I}_t, \mathbf{f}_{t+1}^{t}) \qquad (1)$$

Approximate equality in (1) reminds us that a bijective mapping between two successive images often can not be established due to (dis-)occlusions and changes of perspective.

Recent optical flow research leverages deep convolutional models [8, 31] due to end-to-end trained correspondence and capability to guess motion in (dis-)occluded regions where correspondences are absent. These models are based on local embeddings which act as a correspondence metric, and explicit 2D motion recovery within the correlation layer [8]. Note that correct flow-based forecasting requires optical flow estimation between the past and the future frame which is yet to be observed. Consequently, straightforward extrapolation of past optical flow is bound to achieve suboptimal accuracy even for short-term forecasting, especially at articulated objects such as pedestrians.

Temporal alignment. Semantic forecasting is related to temporal alignment of observed images. Features from a segmented key-frame can be warped towards the current frame in order to speed up semantic segmentation in video [41]. Groundtruth labels can be warped to surrounding unlabeled frames in order to enlarge the training dataset [42]. Current predictions can be improved by enforcing temporal consistency with respect to past frames [10, 27].

Direct semantic forecasting. Luc et al. [18] were the first to forecast future semantic segmentation. Their S2S model follows the direct forecasting approach by taking past segmentations on the input, and producing the future segmentation on the output. Bhattacharyya et al. [2] point out the multimodal nature of the future and try to account for it with dropout-based variational inference. Nabavi et al. [21] formulate the forecasting in a recurrent fashion, with shared parameters between each two frames. Their work has been improved by enforcing temporal consistency between neighbouring feature tensors and leveraging deformable convolutions [4]. This results in attention-based blending, which is related to our forward warping based on pairwise correlation features. However, the forecasting accuracy of these approaches is considerably lower than in our ResNet-18 experiments despite considerable forecasting capacity and better single-frame performance. This suggests that ease of correspondence and avoiding error propagation may be important for successful forecasting.

Flow-based forecasting. Direct semantic forecasting requires a lot of training data due to necessity to learn all motion patterns one by one. This has been improved by allowing the forecasting model to access geometric features which reflect 2D motion in the image plane [14]. Further development of that idea brings us to flow-based forecasting which warps the last dense prediction according to forecasted optical flow [34] as illustrated in (1). This approach has achieved state-of-the-art short-term forecasting accuracy prior to our work. Their convolutional LSTM model receives backward optical flows from three observed frames, and produces the backward optical flow for the future frame. Such formulation is related to our F2M module which also forecasts by warping with regressed flow. However, our F2M module operates on abstract convolutional features, and requires neither external components nor additional supervision. We achieve that by joint training of our compound deep model with feature regression loss. This implies very efficient inference due to subsampled resolution and discourages error propagation due to end-to-end training. Additionally, 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. Further, our module has access to raw semantic features which are complementary to flow patterns [9], and often strongly correlated with future motion (consider for example cars vs pedestrians). Finally, we complement our F2M module with pure recognition-based F2F forecasting which outperforms F2M on previously unobserved scenery. Optical flow has also been used for generating multi-modal future video from single-frame input [16, 24]. Our F2M method takes an opposite approach: we also forecast multiple flows, however our flows connect a single future frame with several past frames. Multi-modal forecasting would be an interesting extension of our present work.

Feature-level forecasting. These approaches map past features to their future counterparts, which is also known as F2F (feature-to-feature) forecasting. The first F2F approach [36] operated on image-wide features from a fully connected AlexNet layer. Later work addressed dense forecasting by regressing features along all levels of the FPN-style upsampling path [17]. However, forecasting at fine resolution is computationally expensive [7]. Hence, some later work reverted to forecasting on the coarse feature level [5]. State-of-the-art mid-term accuracy has been achieved by leveraging deformable convolutions in the F2F module, fine-tuning of the upsampling path with cross-entropy, and a single-frame model without skip-connections [28]. Forecasting at coarse resolution is advantageous due to small inter-frame displacements, rich contextual information and small computational footprint, although some information for recovering small objects may be lost in the process.

Our work improves on [28] as follows. First, we show that forecasting accuracy can be improved by forecasting normalized SPP features. Second, we model explicit correspondence across neighbouring frames by recovering spatio-temporal correlation between embedded convolutional features. Such geometric insight further improves the forecasting accuracy. Third, we introduce F2M forecasting which operates by warping previous features with regressed feature flow. We show that F2M and F2F approaches complement each other in a multi-head F2MF model with shared features. F2F proves better in novel parts of the scene where the model has to imagine what will happen, while F2M prevails on previously observed scenery.

Our work is also related to [37] who formulate feature-level forecasting as reprojection of reconstructed features to the forecasted future ego-location. However, such purely geometric approach is clearly suboptimal in presence of (dis-)occlusions and changes of perspective. Additionally, it is difficult to account for independent motion of moving objects. Our method outperforms [37] by a wide margin, which suggests that optimal forecasting performance requires a careful balance between reconstruction and recognition while explicit 3D reasoning may not be necessary.

3. Semantic forecasting with feature flow

We propose a feature-level forecasting approach which complements recognition-based inference with causal geometric insight as illustrated in Figure 2. The proposed

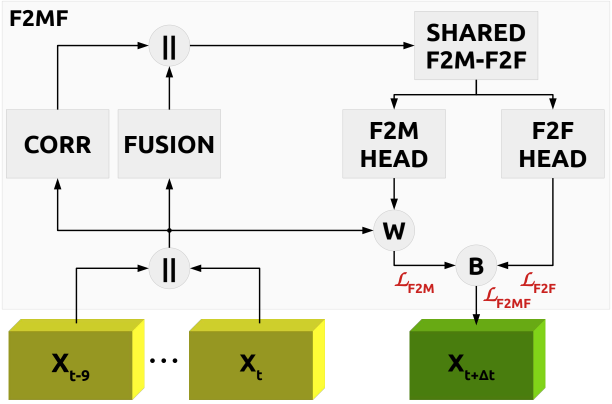

Figure 2. Details of the proposed F2MF forecasting approach. F2M and F2F heads receive a processed concatenation (||) of features $\mathbf{X}_{\mathtt{t-9:t:3}}$ from observed frames, and their spatio-temporal correlation coefficients. The F2M head regresses future feature flow which warps (W) past features into their future locations. The F2F head forecasts the future features directly. The compound forecast $\mathbf{X}_{t+\Delta t}$ is a blend (B) of F2M and F2F forecasts.

F2MF model receives convolutional features $\mathbf{X}_{t-9}$, $\mathbf{X}_{t-6}$, $\mathbf{X}_{t-3}$, $\mathbf{X}_t$ ($\mathbf{X}_{\mathtt{t-9:t:3}}$ for short) extracted by a pre-trained convolutional backbone (CNN DN). On output, the F2MF module forecasts the corresponding future features $\hat{\mathbf{X}}_{t+\Delta t}$, which are subsequently converted to dense predictions $\hat{\mathbf{S}}_{t+\Delta t}$ by a pre-trained upsampling module (CNN UP).

3.1. Single-frame model

Figure 3 shows our single-frame dense prediction model. The downsampling path (CNN-DN) contains an ImageNet-pretrained backbone and a pyramid pooling module [39]. The resulting features $\mathbf{X}_t$ are $32\times$ subsampled with respect to the input resolution. The upsampling path (CNN-UP) has three trained upsampling modules [15, 23] and a 1×1 dense classifier, and concludes with $4\times$ bilinear upsampling.

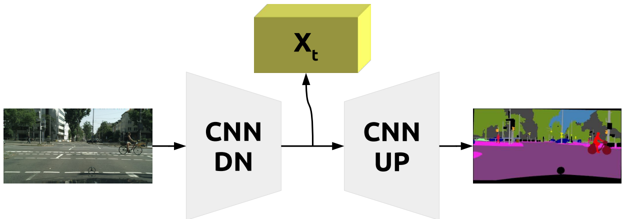

Figure 3. Our single-frame model is a SwiftNet [23] without skip-connections. The downsampling path (CNN-DN) converts the input image I_t to a condensed representation $\mathbf{X}_t$. The upsampling path (CNN-UP) produces dense semantic output $\hat{\mathbf{S}}_t$.

3.2. Spatio-temporal correlation features

Our correlation module determines spatio-temporal correspondence between neighbouring frames. On input, it receives a T×C×H×W tensor with convolutional features $\mathbf{X}_{\mathtt{t-9:t:3}}$. In all experiments we have T=4 (time instants), H=32, and W=64. We first embed features from all time instants into a space with enhanced metric properties by a shared 3×3×C' convolution (C'=128). This mapping can recover distinguishing information which is not needed for single-frame inference. Subsequently, we construct our metric embedding $\mathbf{F}$ by normalizing C'-dimensional features to unit norm so that cosine similarity become dot product. Finally, we produce d^2 correspondence maps between features at time τ and their counterparts at $\tau-3$ within $d\times d$ neighborhood, for each $\tau\in\{t-6,t-3,t\}$. The correlation tensor $\mathbf{C}^\tau$ at location $\mathbf{q}$ and feature map $ud+v$ is a dot product of a metric feature at time τ and location $\mathbf{q}\in\mathcal{D}(\mathbf{F})$, with its counterpart at time $\tau-3$ offset by (u,v) [8, 13]:

$$\mathbf{C}^{\tau}_{ud+v,\mathbf{q}} = \mathbf{F}^{\tau\top}_{\mathbf{q}}\mathbf{F}^{\tau-3}_{\mathbf{q}+[u-\frac{d}{2},v-\frac{d}{2}]}, \text{where } u,v\in[0\,..\,d). \quad (2)$$

3.3. F2F forecasting

Our feature-to-feature module receives processed input features and directly regresses the future features $\mathbf{X}_{t+\Delta t}$. This is similar to previous work [17, 5, 28], however there is one important difference. Our F2F module has access to spatio-temporal correlation features which relieve the need to learn correspondence from scratch. Our experiments show clear advantage of these features which suggests that correspondence is not easily learned on existing datasets.

3.4. F2M forecasting

Our F2M module provides a regularized variant of F2F forecasting. It assumes that there is a causal relationship between the past and the future, which can be explained by 2D warping. It receives processed input features and outputs a dense displacement field for warping each of the four feature tensors into its future counterpart $\hat{\mathbf{X}}^{(\tau)}_{t+3}$, $\tau\in\{t-9,t-6,t-3,t\}$. We finally blend the four forecasts with trained per-pixel weight vectors which we activate with softmax. Consequently, the forecast can utilize the observed frame with the best view onto a disoccluded part of the scene. We demonstrate this in Fig 7.

F2M with backward warp. F2M forecast can be constructed either by backward warping with $\hat{\mathbf{f}}^{\tau}_{t+\Delta t}$ or by forward warping with $\hat{\mathbf{f}}^{t+\Delta t}_{\tau}$, as shown in (1). In the backward case, the F2M model forecasts feature flows at time $t+\Delta t$:

$$\hat{\mathbf{f}}^{\tau}_{t+\Delta t} = \mathrm{F2M}^{(\tau)}_{\mathrm{bw}}(\mathbf{X}_{\mathtt{t-9:t:3}})\;, \tau\in\{t-9,\dots,t\}\;. \quad (3)$$

Future feature tensors are subsequently obtained by backward warping each of the four previous feature vectors:

$$\hat{\mathbf{X}}^{(\tau)}_{t+\Delta t} = \mathrm{warp\_bw}(\mathbf{X}_\tau, \hat{\mathbf{f}}^{\tau}_{t+\Delta t})\;. \quad (4)$$

Backward warp obtains future activations by interpolating at non-integer locations of the forecasted backward flow:

$$\hat{\mathbf{X}}^{(\tau)}_{t+\Delta t}[\mathbf{q}] = \mathrm{bilinear\_interp}(\mathbf{X}_\tau, \mathbf{q}+\hat{\mathbf{f}}^{\tau}_{t+\Delta t}[\mathbf{q}])\;. \quad (5)$$

F2M with forward warp. This F2M variant forecasts forward feature flow at times $\tau\in\{t-9,t-6,t-3,t\}$:

$$\hat{\mathbf{f}}^{t+\Delta t}_{\tau} = \mathrm{F2M}^{(\tau)}_{\mathrm{fw}}(\mathbf{X}_{\mathtt{t-9:t:3}}) \quad (6)$$

Future feature tensors are obtained by forward warping each of the four previous feature vectors:

$$\hat{\mathbf{X}}^{(\tau)}_{t+\Delta t} = \mathrm{warp\_fw}(\mathbf{X}_\tau, \hat{\mathbf{f}}^{t+\Delta t}_{\tau}) \quad (7)$$

This produces future activations by splatting [33] observed features at non-integer locations provided by the forecasted forward flow. Unfortunately, a splatting implementation for GPU hardware [22] became available only after the time of our experiments. We have therefore devised a simple although inefficient implementation based on `matmul`:

$$\hat{\mathbf{X}}^{(\tau)}_{t+\Delta t}[\mathbf{q}] = \frac{1}{N_{\mathbf{q}}}\sum_{\mathbf{p}\in\mathcal{D}(\mathbf{X})} k(\mathbf{p}+\hat{\mathbf{f}}^{t+\Delta t}_{\tau}[\mathbf{p}],\mathbf{q})\cdot\mathbf{X}_t[\mathbf{p}]. \quad (8)$$

In the above equations, k represents the RBF kernel, while $N_{\mathbf{q}}$ is a normalizing factor which ensures that the norm of the forecasted features remains within the usual range:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) , \tag{9}$$

$$N_{\mathbf{q}} = \sum_{\mathbf{p} \in \mathcal{D}(\mathbf{X})} k(\mathbf{p} + \hat{\mathbf{f}}_{\tau}^{t+\Delta t}[\mathbf{p}], \mathbf{q}) . \tag{10}$$

Expression (8) is computationally much more intensive than (5), but it is nevertheless feasible due to small resolution.

The presented two formulations of feature flow are quite different. The forward flow (6) is aligned with the *observed* features, while the corresponding backward flow (3) aligns with the *forecasted* features. Consider a pixel at some moving object in the last observed image. Its forward flow is inferred by looking (convolutionally speaking) at the present object location. On the other hand, the backward flow has to look at the future object location. Hence, backward flow requires larger receptive field in order to operate correctly.

Backward F2M addresses *effects* of the motion: it makes decisions by considering all possible observed activations which may ”come” into the particular location of the future tensor. Consequently, it stands a good chance to correctly resolve contention due to occlusion, provided its receptive field is large enough. On the other hand, forward F2M addresses *causes* of the motion: it makes decisions by considering observed motion of feature activations. Hence, forward F2M is able to model a probabilistic distribution over feasible displacements, which may make it an interesting choice for longer-term forecasting of multi-modal future.

3.5. Compound F2MF model

The compound F2MF model blends F2M and F2F outputs with dense softmax activated weights w^{F2F} and w_{τ}^{F2M}:

$$\hat{\mathbf{X}}_{t+\Delta t}^{\mathrm{F2MF}} = w^{\mathrm{F2F}} \cdot \hat{\mathbf{X}}_{t+\Delta t}^{\mathrm{F2F}} + \sum_{\tau} w_{\tau}^{\mathrm{F2M}} \cdot \hat{\mathbf{X}}_{t+\Delta t}^{(\tau)} \tag{11}$$

Note that the F2MF model reuses softmax preactivations of w_{τ}^{F2M} which are regressed by F2M. There is 1 convolutional layer in the fusion module, 6 layers in the shared module, and 1 layer in F2F and F2M heads. All layers are BN-ReLU-dconv where dconv stands for deformable convolution [40]. We use two auxiliary losses $\mathcal{L}_{\mathrm{F2M}}$ and $\mathcal{L}_{\mathrm{F2F}}$, as well as the compound loss $\mathcal{L}_{\mathrm{F2MF}}$, as shown in Fig. 2. All losses have equal contribution.

4. Experiments

We perform experiments on finely annotated subset of the Cityscapes dataset with 2975 train, 500 validation, and 1525 test images. Each labeled image corresponds to the 20-th frame of a 1.8 second long video clip (30 frames) [6]. We use pre-trained single-frame models based on DenseNet-121 [12] or ResNet-18 [11]. Our forecasts target normalized features from the most condensed representation of the single-frame model (cf. 3.1). We train for 160 epochs with L2 loss, early stopping, batch size 12, and ADAM with cosine annealing (lr$_{\max}$=5e-4, lr$_{\min}$=1e-7).

We evaluate F2MF forecasting on the semantic segmentation task in short-term (Δt=3 frames ahead, 180 ms) and mid-term (Δt=9 frames, 540 ms) experiments. We report the accuracy for all classes (mIoU All), and 8 classes with moving objects (mIoU MO) [18, 17]. In some experiments we augment the training data by horizontal flipping and random sliding of the training tuple across the video clip. Most experiments use backward warping due to better efficiency.

4.1. Comparison with previous state of the art

Table 1 compares our F2MF model with previous work on Cityscapes val. The first section presents the usual upper bound (oracle) and the usual baseline (copy last segmentation) [18]. The second section shows results from the literature where LSTM M2M [34] and DeformF2F [28] achieve best short-term (67.1 mIoU) and mid-term (53.6 mIoU) performance, respectively. The last section presents our DenseNet-based F2MF model trained without and with data augmentation. Our best model achieves state-of-the-art both in short-term and mid-term forecasting while outperforming the two runner-ups by 2.5 and 4.3 mIoU points.

4.2. Qualitative results

Figures 4 and 5 show our short-term and mid-term semantic segmentation forecasts on six clips from Cityscapes val. The first three rows show the last observed image, and

	Short term: Δt=3		Mid term: Δt=9	
Accuracy (mIoU)	All	MO	All	MO
Oracle	75.8	75.2	75.8	75.2
Copy last segmentation	53.3	48.7	39.1	29.7
3Dconv-F2F [5]	57.0	/	40.8	/
Dil10-S2S [18]	59.4	55.3	47.8	40.8
LSTM S2S [21]	60.1	/	/	/
Mask-F2F [17]	/	61.2	/	41.2
FeatReproj3D [37]	61.5	/	45.4	/
Bayesian S2S [2]	65.1	/	51.2	/
DeformF2F [28]	65.5	63.8	53.6	49.9
LSTM AM S2S [4]	65.8	/	51.3	/
LSTM M2M [34]	67.1	65.1	51.5	46.3
F2MF-DN121 w/o d.a.	68.7	66.8	56.8	53.1
F2MF-DN121 w/ d.a.	**69.6**	**67.7**	**57.9**	**54.6**

Table 1. Evaluation of our DenseNet-121-based F2MF model for semantic segmentation forecasting on Cityscapes val. *All* denotes all classes, *MO* — moving objects, and *d.a.* — data augmentation.

the future image overlayed with the oracle prediction and our F2MF forecast. The last row visualizes $w^{\mathrm{F2M}} = 1 - w^{\mathrm{F2F}} = \sum_{\tau} w_{\tau}^{\mathrm{F2M}}$ which reveals whether the particular pixel is forecasted by F2M (red) or F2F (blue). We observe that our forecasts incur some loss of details (cf. classes pole and person), but are otherwise quite accurate. The F2M head is preferred in static regions where establishing correspondence is relatively easy (cf. red w^{F2M} in columns 3, 4 in Fig. 4, and columns 2, 4 in Fig. 5). The F2F head contributes to dynamic scenery and assumes full responsibility in previously unobserved pixels (cf. blue w^{F2M} in column 2 in Fig. 4, and columns 1, 3, 6 in Fig. 5). Contribution of the F2F head is best visible in column 1 of Fig. 5. A car on the right leaves the scene while disoccluding a large part of previously unobserved background. Our model assigns disoccluded pixels to the F2F head which correctly fills-in road, sidewalk and building pixels. This suggests that F2F and F2M complement each other.

4.3. Influence of the single-frame model

Table 2 explores influence of single-frame performance to the forecasting accuracy. We consider two backbones with very lean representations in the last convolutional layer. The model based on DenseNet-121 has a more accurate backbone and wider pyramid pooling (C=512 vs. C=128). These advantages result in 3.3 pp mIoU higher single-frame performance as shown in columns 2-3 (oracle). However, some of this performance gain does not transfer onto the forecasting tasks. The advantage drops to 1.8 pp mIoU at short-term and to 0.9 pp mIoU at mid-term. Thus, we use the model based on ResNet-18 in all further experiments in order to speed up the processing.

	Oracle		Short-term		Mid-term	
Accuracy (mIoU)	All	MO	All	MO	All	MO
F2MF-RN18	72.5	71.5	66.9	65.6	55.9	52.4
F2MF-DN121	75.8	75.2	68.7	66.8	56.8	53.1

Table 2. Influence of the single-frame semantic segmentation model to the forecasting performance on Cityscapes val. We do not use data augmentation in order to speed up the training.

4.4. Ablation and validation experiments

Table 3 evaluates the contribution of correlation features and the F2M head. We first compare independent F2F and F2M approaches (row 1 vs row 2, and row 4 vs row 5). F2F is somewhat better overall (up to 1 pp mIoU), except in mid-term forecast with correlation features where the two approaches perform equally well. Subsequently we explore the contribution of correlation features (rows 1, 2, 3 vs. rows 4, 5, 6). We note a consistent performance improvement, 0.8-1.1 pp mIoU at short-term, and 1.7-3.1 pp mIoU at mid-term. The compound F2MF model profits more than the independent F2F model. Finally, we observe that the compound model outperforms independent models even though its capacity is only marginally larger (most of F2F and F2M features are shared). Hence, the improvement is likely due to stronger learning signal. F2MF outperforms F2F for 0.4-1.1 pp mIoU (without correlation), and 0.6-1.6 pp mIoU (with correlation). Overall, correlation features and F2M

F2MF-RN18 Configuration			Short-term mIoU		Mid-term mIoU	
F2F	F2M	Correlation	All	MO	All	MO
	✓		64.8	63.4	52.2	47.6
✓			65.4	64.0	52.8	48.6
✓	✓		65.8	64.7	53.4	49.7
	✓	✓	65.6	64.4	54.5	50.7
✓		✓	66.3	64.9	54.5	50.8
✓	✓	✓	**66.9**	**65.6**	**55.9**	**52.4**

Table 3. Ablation of correlation, F2F, and F2M on Cityscapes val. Standalone F2F and F2M models are trained independently.

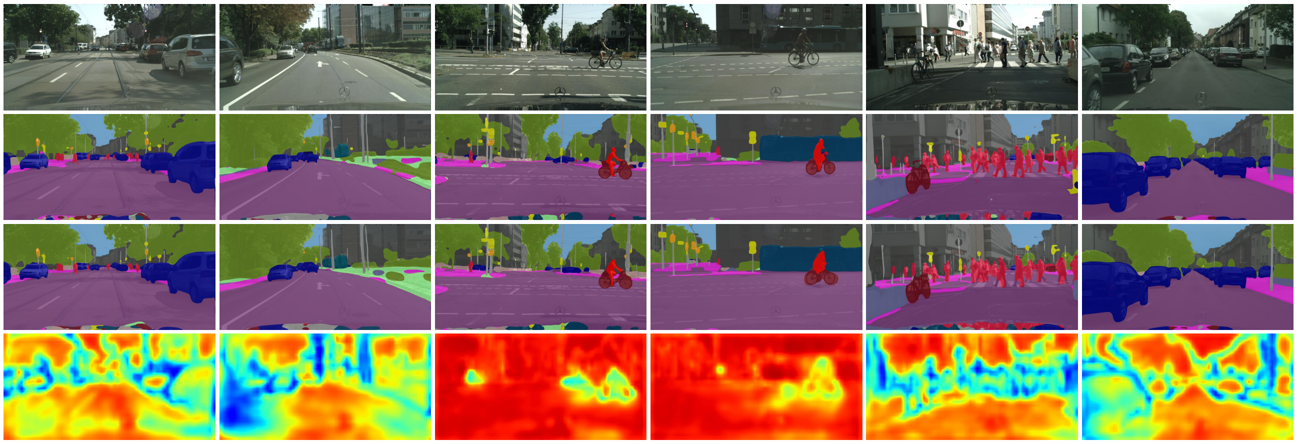

Figure 4. Short-term accuracy of our best model. The rows contain i) the last observed image, ii) prediction by our oracle, iii) our forecast, and iv) heat map of w^{F2M} where red denotes F2MF preference of F2M forecast. Rows ii) and iii) are overlaid with the future image.

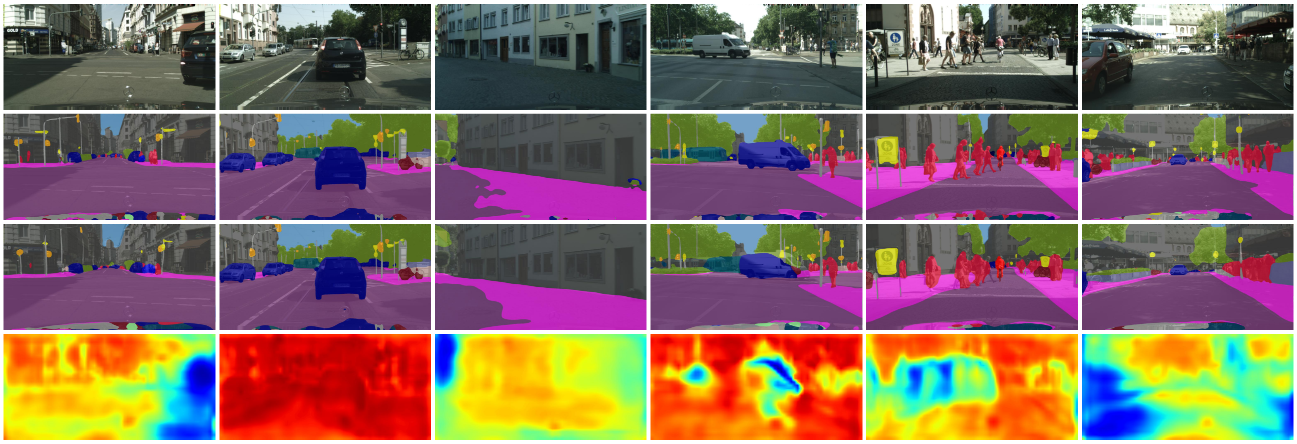

Figure 5. Mid-term accuracy of our best model. The rows contain i) the last observed image, ii) prediction by our oracle, iii) our forecast, and iv) heat map of w^{F2M} where red denotes F2MF preference of F2M forecast. Rows ii) and iii) are overlaid with the future image.

head bring significant improvement upon the F2F baseline: 1.5 pp mIoU at short-term and over 3 pp mIoU at mid-term.

Table 4 compares backward and forward formulations of independent F2M forecasting as presented in Section 3.4. Forward warping uses the RBF kernel with $\sigma^2 = 0.125$. The first section shows that, interestingly, the two approaches achieve very similar results in the standard setup. Hence, we use the backward formulation in all other experiments as a more efficient option. The second section considers a variant of the F2M model which has only three convolutional layers (instead of eight), and uses regular instead of deformable convolutions. These experiments show clear advantage of forward warping in case of limited receptive field, and support our hypothesis that F2M with backward warp requires a larger receptive field.

	Short-term		Mid-term	
Accuracy (mIoU)	All	MO	All	MO
F2M-BW	64.8	63.4	52.2	47.6
F2M-FW	64.6	63.2	52.2	47.3
F2M-BW (limited r.f.)	60.4	58.1	45.4	37.8
F2M-FW (limited r.f.)	61.2	59.1	47.6	41.1

Table 4. Comparison of backward and forward warping in terms of independent F2M forecasting accuracy on Cityscapes val. Forward F2M has an edge in experiments with small receptive field.

4.5. F2M vs F2F performance across pixel groups

Table 3 shows that, overall, independent F2F outperforms independent F2M. However, we know that F2M performs very poorly in novel pixels, and hence hypothesize that F2M may outperform F2F in previously observed regions. We therefore stratify pixels with respect to F2M weights w^{F2M} (as predicted by the F2MF model), and test our hypothesis by comparing the forecasting accuracy across ten pixel groups as shown in Fig. 6.

The x-axis shows F2M weights, the left y-axis shows the accuracy (bar plot) while the right y-axis shows pixel incidence (red line). We omit the pixel group with w^{F2M}=0.05 since very few pixels belong there. The pixel incidence curve shows that F2MF believes F2M in majority of pixels. This is correct behaviour, because independent F2M outperforms independent F2F in the right parts of the two plots (w^{F2M} ≥0.75). However, the F2F model prevails in hard pixels (left parts of the two plots, w^{F2M} ≤0.45).

Note that here, as well as in subsection 4.4, we consider independently trained F2M and F2F models in order to avoid interference of the compound training. This analysis corroborates experiments from Section 4.2 which show that F2MF model succeeds to output low w^{F2M} at novel pixels and high w^{F2M} at static scenery. These weights can be seen as a proxy for how easy the F2M forecast is at a particular pixel. Therefore, these results also confirm our hypothesis that the two approaches complement each other.

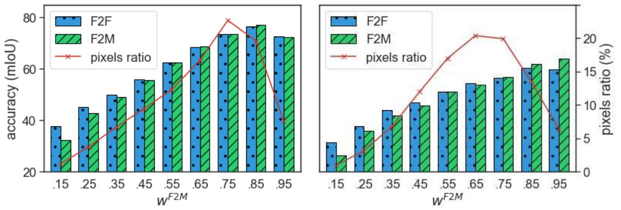

Figure 6. Stratified comparison of independent F2F and F2M forecast on Cityscapes val at short-term (left) and mid-term (right). We present the accuracy of the two models (bar plot, mIoU) and the F2MF pixel incidence (red line, %) across w^{F2M} bins.

4.6. Importance of feature normalization

Table 5 explores the influence of feature normalization with training set mean and variance to the forecasting accuracy. This facilitates the optimization process by making all feature maps equally important. Note that this also requires denormalization of the forecasted features before the upsampling path. The normalization improves the accuracy by 1.4 and 3.1 pp mIoU at short-term and mid-term period.

	Short-term		Mid-term	
Accuracy (mIoU)	All	MO	All	MO
F2MF w/ norm.	66.9	65.6	55.9	52.4
F2MF w/o norm.	65.5	64.1	52.8	48.1

Table 5. Influence of feature normalization to F2MF accuracy on Cityscapes val. We normalize w.r.t. training mean and variance.

4.7. Visual interpretation of model decisions

Fig. 7 provides further insight into difference between F2MF and F2F forecasting by visually comparing their gradients w.r.t. the input frames [19]. The columns show four observed frames, and the future frame overlayed with semantic forecast and ground truth. We focus at the pixel designated with the green square and explain the corresponding model decision (max-log-softmax) by showing locations of top 0.1% largest gradients w.r.t. the input pixels (red dots). First, we consider a pixel on the bicycle wheel (rows 1-2). F2F gradients spread in an irregular manner over the whole bicycle and the background (row 1), while F2MF gradients concentrate around the wheel position in the last frame (row 2). Second, we consider a background pixel which is disoccluded by the cyclist motion. The F2F model tries to reconstruct the forecast from the context by looking around the cyclist in the last frame. On the other hand, the F2MF model succeeds to detect that this part of the scene has actually been observed in the most distant past frame. Thus, it performs the forecast by simply copying the corresponding representation into the future.

5. Conclusion

We have presented a novel feature-level forecasting approach which regularizes the inference by modeling a causal relationship between the past and the future. The proposed F2M (feature-to-motion) forecasting generalizes better than the classic F2F (feature-to-feature) approach in many (but not all) image locations. We achieve the best of both worlds by blending F2M and F2F predictions with densely regressed weight factors. The resulting F2MF model surpasses the state-of-the-art in semantic segmentation forecasting on the Cityscapes dataset by a wide margin.

To the best of our knowledge, this is the first account of using correlation features for semantic forecasting. Our experiments show that these features bring clear advantage in all three feature-level approaches: F2F, F2M, and F2MF. We have considered two F2M variants with respect to warp direction. F2M with forward warping performs better in setups with small receptive field and allows probabilistic modeling of motion uncertainty. However, F2M with backward warping generalizes equally well in our regular setup.

Despite encouraging results, real-world applications will require lots of future work. In particular, our method does not address multi-modal future, which is a key to long-term forecasting and worst-case reasoning. Other suitable extensions include overcoming obstacles towards end-to-end training, applications to other tasks and RGB forecasting, and enforcement of temporal consistency.

Acknowledgment

This work has been funded by Rimac Automobili and supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 DATACROSS. Parts of our experimental implementation benefited from [3, 25].

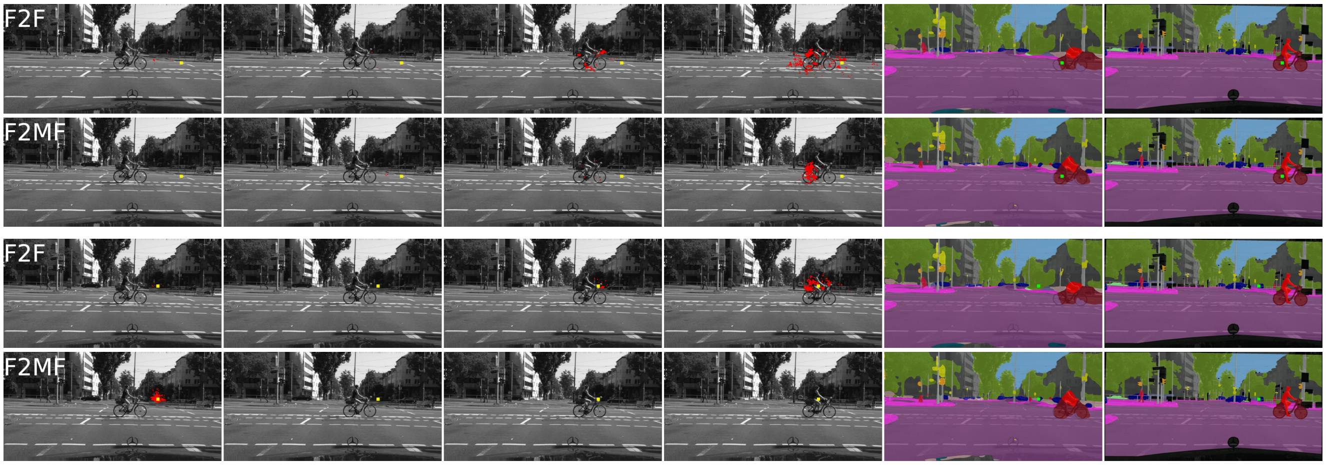

Figure 7. Interpretation of F2F (rows 1,3) and F2MF (rows 2,4) decisions in two pixels denoted with green squares. We consider a pixel on the bicycle (rows 1-2) and a pixel on disoccluded background (rows 3-4). The columns show the four input frames, the forecasted semantic map, and the groundtruth with overlayed future frame. Red dots show top gradients of the green pixel max-log-softmax w.r.t. input.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 1

[2] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. In *International Conference on Learning Representations*, 2019. 1, 2, 5

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 8

[4] Xin Chen and Yahong Han. Multi-timescale context encoding for scene parsing prediction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1624–1629. IEEE, 2019. 1, 2, 5

[5] Hsu-Kuang Chiu, Ehsan Adeli, and Juan Carlos Niebles. Segmenting the future. *CoRR*, abs/1904.10666, 2019. 2, 3, 4, 5

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[7] Camille Couprie, Pauline Luc, and Jakob Verbeek. Joint Future Semantic and Instance Segmentation Prediction. In *ECCV Workshop on Anticipating Human Behavior*, pages 154–168, 2018. 2, 3

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2, 4

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1933–1941, 2016. 3

[10] Raghudeep Gadde, Varun Jampani, and Peter V. Gehler. Semantic video cnns through representation warping. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4463–4472, 2017. 2

[11] 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. 5

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269, 2017. 5

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[14] Xiaojie Jin, Huaxin Xiao, Xiaohui Shen, Jimei Yang, Zhe Lin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Predicting scene parsing and motion dynamics in the future. In *Advances in Neural Information Processing Systems*, pages 6915–6924, 2017. 1, 3

[15] Ivan Kreso, Josip Krapac, and Sinisa Segvic. Ladder-style densenets for semantic segmentation of large natural images. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 238–245. IEEE Computer Society, 2017. 4

[16] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 600–615, 2018. 3

[17] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 584–599, 2018. 2, 3, 4, 5

[18] 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*, pages 648–657, 2017. 1, 2, 5

[19] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*, pages 4898–4906, 2016. 8

[20] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 1

[21] Seyed Shahabeddin Nabavi, Mrigank Rochan, and Yang Wang. Future semantic segmentation with convolutional lstm. *BMVC*, 2018. 1, 2, 5

[22] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. *CoRR*, abs/2003.05534, 2020. 4

[23] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12607–12616, 2019. 4

[24] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2019. 3

[25] Clement Pinard. Pytorch-correlation-extension. https://github.com/ClementPinard/Pytorch-Correlation-extension, 2020. 8

[26] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–733, 2018. 1

[27] Timo Saemann, Karl Amende, Stefan Milz, and Horst-Michael Gross. Leverage temporal consistency for robust semantic video segmentation. In *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, 2019. 2

[28] Josip Šarić, Marin Oršić, Tonći Antunović, Sacha Vražić, and Siniša Šegvić. Single level feature-to-feature forecasting with deformable convolutions. In *German Conference on Pattern Recognition*, pages 189–202. Springer, 2019. 2, 3, 4, 5

[29] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1501–1510, 2017. 1

[30] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 273–283, 2019. 1

[31] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2

[32] Jiangxin Sun, Jiafeng Xie, Jianfang Hu, Zihang Lin, Jianhuang Lai, Wenjun Zeng, and Wei-Shi Zheng. Predicting future instance segmentation with contextual pyramid convlstms. In *ACM MM*, pages 2043–2051, 2019. 2

[33] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 2, 4

[34] Adam Terwilliger, Garrick Brazil, and Xiaoming Liu. Recurrent flow-guided semantic forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1703–1712. IEEE, 2019. 1, 3, 5

[35] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3560–3569. JMLR. org, 2017. 1

[36] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *arXiv preprint arXiv:1504.08023*, 2, 2015. 2, 3

[37] Suhani Vora, Reza Mahjourian, Soeren Pirk, and Anelia Angelova. Future semantic segmentation using 3d structure. In *ECCV 3D Reconstruction meets Semantics Workshop*, 2018. 1, 2, 3, 5

[38] Yu Yao, Mingze Xu, Chiho Choi, David J. Crandall, Ella M. Atkins, and Behzad Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *ICRA*, 2019. 1

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[40] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *arXiv preprint arXiv:1811.11168*, 2018. 5

[41] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017. 2

[42] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019. 2