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

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

1. Results on Cityscapes test

We submit our best model trained on Cityscapes trainval for online evaluation on the Cityscapes benchmark [2]. Table 1 shows the results. We achieve considerably better test results (1.7 pp mIoU MO at mid-term) which indicates absence of bias towards the validation set.

			Short-term		Mid-term	
model	train	eval	All	MO	All	MO
F2MF-DN121	train	val	69.6	67.7	57.9	54.6
F2MF-DN121	train+val	test	70.2	68.7	59.1	56.3

Table 1. Forecasting on Cityscapes test outperforms the validation mIoU accuracy presented in the main paper.

2. Per-class results

Table 2 shows per-class accuracies (IoU) of four F2MF models and two oracles. The two sections are dedicated to models based on ResNet-18 and DenseNet-121. Each section first presents the oracle and then compares it to short-term and mid-term forecasts. The last eight classes in the table represent moving objects. We do not show F2M and F2F forecasts since they are almost always worse than F2MF.

We observe that all forecasting models achieve the lowest accuracy on the class pole. Forecasting poles is hard since their elongation is perpendicular to the motion: even a small displacement can miss an entire object. Thus, our F2MF models often opt to entirely omit some poles in order to avoid double punishment (mIoU counts both false positives and false negatives). This can be confirmed by comparing incidence of pole pixels in oracle prediction (1.14%), with the corresponding statistics in short-term (1.00%) and mid-term (0.69%) forecasts. Among the moving object classes, persons cause the largest performance deterioration: 14.7 mIoU pp for short-term and 31.8 mIoU pp for mid-term period. This indicates that our F2MF models find person motion much less predictable than the motion of vehicles. People assume different motion styles and cause considerable (dis-)occlusion especially since they often move in groups. Their vertical elongation leads to similar problems as in the case of poles. These facts make forecasting of future person locations and poses quite challenging.

We also observe a somewhat unexpected finding. A short-term forecast of the ResNet-based model for the class `truck` outperforms the oracle for 5pp mIoU (cf. Table 2, section RN-18). This result should be taken with a grain of salt, since there are only 120 truck instances in Cityscapes

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mean
Oracle RN-18	**97.5**	**81.6**	**90.7**	50.1	**53.4**	**56.1**	**60.3**	**70.8**	**90.9**	**60.9**	**92.9**	**75.9**	**53.0**	**93.2**	67.4	**84.4**	**72.0**	**54.5**	**71.7**	**72.5**
F2MF w/o d.a. short-term	96.3	74.9	87.8	**50.4**	50.6	40.0	53.0	59.9	87.7	56.6	89.3	61.2	45.5	87.6	**72.1**	78.1	65.8	52.3	62.1	66.9
F2MF w/o d.a. mid-term	94.1	64.5	81.8	46.8	45.2	19.1	35.5	40.9	80.9	50.6	82.8	44.1	27.9	76.6	67.3	68.5	49.0	38.7	46.6	55.9
Oracle DN-121	**97.8**	**82.9**	**91.8**	**60.1**	**59.4**	**59.8**	**65.0**	**74.2**	**91.4**	**62.0**	**93.5**	**78.2**	**58.4**	**94.2**	**80.8**	**85.0**	**68.9**	**61.6**	**73.8**	**75.8**
F2MF w/ d.a. short-term	96.7	76.5	89.0	57.8	56.5	44.2	57.5	63.9	88.5	59.0	90.4	64.7	49.8	88.8	77.5	81.3	63.2	50.5	65.2	69.6
F2MF w/ d.a. mid-term	94.6	66.4	83.0	50.6	49.9	19.2	38.4	42.9	81.9	51.5	83.6	45.9	30.5	78.4	71.1	73.1	47.6	41.0	48.8	57.9

Table 2. Per-class results (IoU) on Cityscapes val for models based on ResNet-18 and DenseNet-121. Only DenseNet-based F2MF models are trained with data augmentation in order to show the full spectrum of achievable performance.

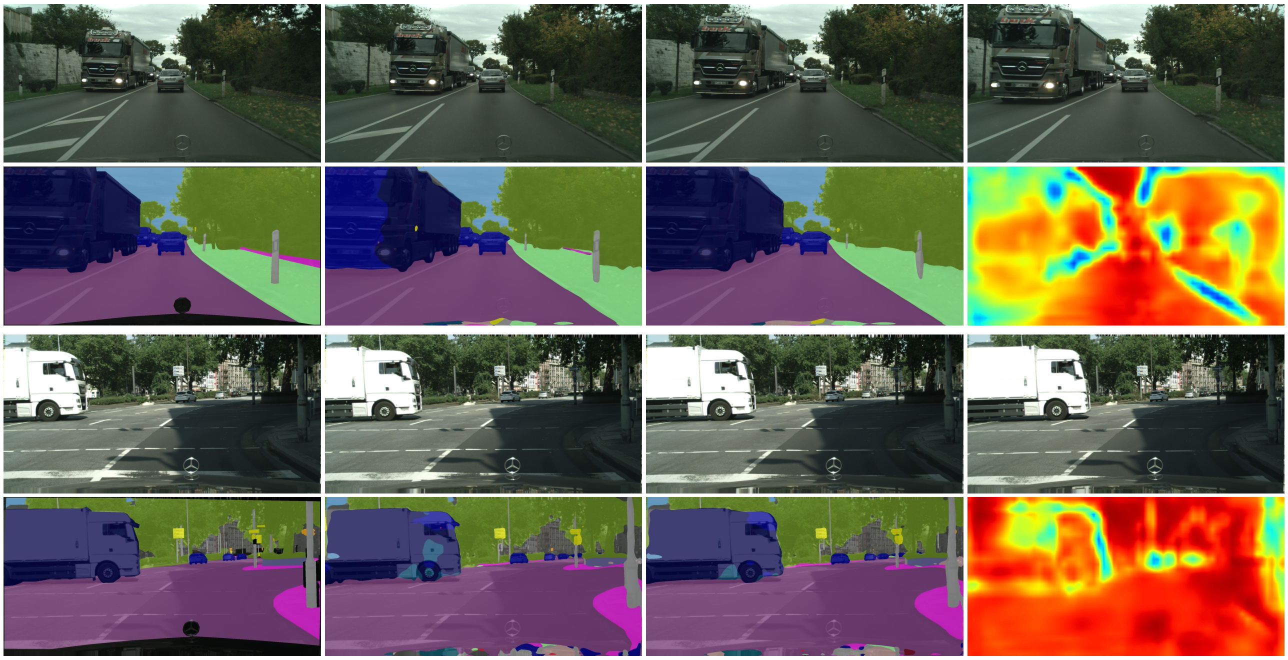

Figure 1. A clip in which F2MF-RN18 short-term forecast beats the oracle. Top row shows the four observed frames. Bottom row shows ground truth, oracle, our forecast and dense w^{F2M} weights. The forecast has an opportunity to observe the entire truck at a reasonable scale.

val. Nevertheless, a closer look reveals that forecasting presents real advantages in particular clips, as illustrated in Figure 1. There are two ways how forecast may beat the oracle. First, if a very large vehicle moves towards the camera (which is the usual case in road-driving), then the vehicle is often more recognizable at middle distance than in the close-up case. A forecasting model has a chance to recognize the object when it is far from the camera, and then to simply propagate that information to the forecast. The oracle does not have that luxury and has to deal only with the single image where the object may be larger than the receptive field. Second, F2M forecasting has an opportunity to convert a near miss to a hit by ensembling warped representations from different frames.

Note that the oracle beats our forecasts at all classes in the case of the more powerful single-frame model (cf. Table 2, section DN-121). However, previously described effects seem still to be present, since forecasted trucks and buses lose less IoU performance with respect to the oracle than all classes with moving objects.

3. Failure cases

Figure 2 shows some failure cases of our best model. The first row shows the last observed image while the subsequent three rows show the future image overlayed with ground truth, oracle prediction, and F2MF-DN121 forecast. The last row visualizes the F2M weighting factor w^{F2M} which reveals whether the particular pixel has been principally forecasted by F2M (red) or F2F (blue). The columns correspond to three short-term (columns 1-3) and three mid-term (columns 4-6) forecasts which we pick among the clips with the largest cross-entropy loss on Cityscapes val.

The first clip shows a previously unobserved cyclist entering the future frame from the left. Our forecasting model is unable to reason correctly here, since no visual evidence about the cyclist was present in the observed frames. In the next two clips, forecasting errors are caused by poor single-frame performance. A significant part of the road is misinterpreted for sidewalk (column 2), while some terrain is classified as vegetation (column 3) both in the oracle prediction and the forecast. We see that F2MF model is unable to recover from consistent mis-prediction by the single-frame model.

The first two mid-term examples (columns 4 and 5) show poor forecast of thin traffic signs, which is likely due to coarse resolution of our F2MF setup. We note a correct F2F preference at novel scenery in the top-left corner of the future image in column 4 (please note that the ego-car is turning left). We also note the incorrect forecast of the tram motion in column 5, where the tram does not re-appear at the other side of the minivan. The last clip features prominent, independent and articulated motion due to nearby pedestrians crossing the street. The forecasting model disregards the F2M branch (blue color in row 5) while the F2F branch seems overwhelmed by the sheer complexity of the scene dynamics and consequently produces blobby predictions.

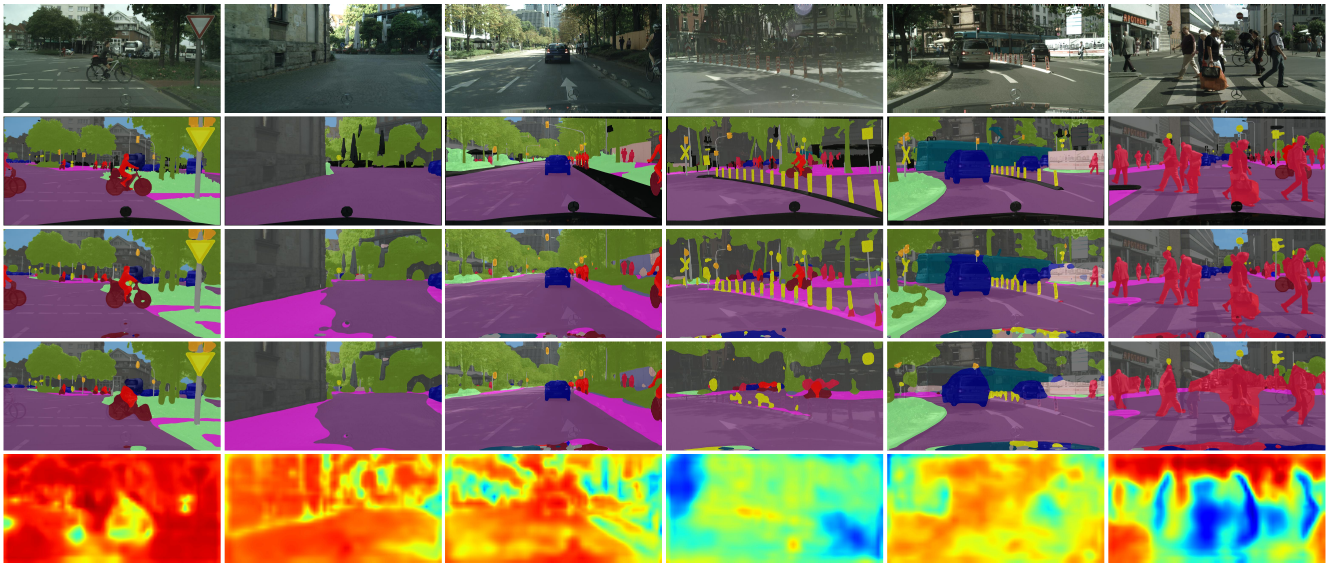

Figure 2. Six failure cases of our best performing model. The rows contain i) the last observed image, ii) semantic segmentation ground-truth, iii) prediction by our oracle, iv) F2MF-DN121 forecast, and v) the heat map of w^{F2M} where red denotes F2M preference. Rows ii), iii) and iv) are overlaid with the unobserved future image. Each column corresponds to a different clip from Cityscapes val. We show three examples of short-term forecast (columns 1-3) and three examples of mid-term forecast (columns 4-6).

4. Visualization of the feature flow

Figure 3 shows feature flows predicted by F2M models with forward and backward warping. The columns correspond to two Cityscapes val clips. Rows 1 and 4 show the last observed image (I_t) and the unobserved future image ($\mathrm{I}_{\mathrm{t}+3}$). Row 2 shows the forward feature flow $\hat{\mathbf{f}}_t^{t+3}$ predicted by our F2M-RN18-FW model which achieves 64.6 mIoU. Row 3 shows the backward feature flow $\hat{\mathbf{f}}_{t+3}^{t}$ predicted by our F2M-RN18-BW model which achieves 64.8 mIoU (cf. Table 4 in the article).

We encode flow with the standard color-code [1] where cyan means left, yellow — down, red — right, and blue — up, while the saturation is proportional to the magnitude. We observe i) that forward flow aligns with object locations in the observed image, ii) that backward flow aligns with the future object locations, and iii) that corresponding motion vectors are opposite for the two F2M variants. This is in concordance with equation (1) and the discussion from section 3.4 in the paper.

In the first column, we can see that the backward flow aligns with the future cyclist location. On the other hand, the forward flow is better aligned with the cyclist in the observed image. The same pattern occurs at the distant moving pedestrian in the left part of the image. Motion of the distant car in the image center has been detected only by the backward F2M model. The complementary nature of the two feature flows is clearly visible in the second column as well. Additionally, there we note qualitatively correct flow on stationary parts of the scene, which occurs due to ego-motion of the camera.

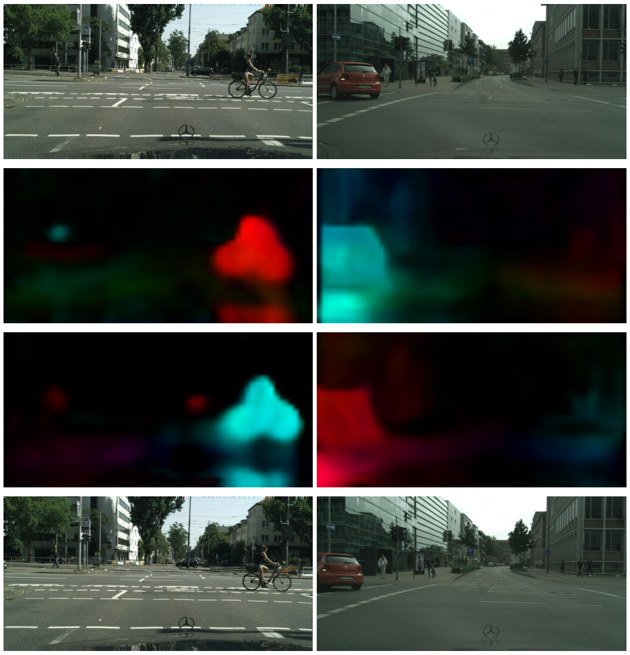

Figure 3. Visualization of the feature flows forecasted by two F2M models on Cityscapes val. Row 1 shows the last observed image. Rows 2-3 show forward and backward feature flows as forecasted by corresponding independent F2M-RN18 models. Row 4 shows the unobserved future image. Outlines of the moving objects indicate that the feature flows are qualitatively correct.

	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mean
F2M-RN18 short-term	98.6	79.1	95.6	61.0	54.3	52.1	61.1	73.1	94.8	66.7	96.8	62.9	50.9	95.5	83.4	85.8	78.9	66.0	78.4	75.5
F2F-RN18 short-term	98.5	78.7	95.6	62.0	56.1	51.6	60.4	71.9	94.5	66.7	96.8	61.0	49.1	95.5	85.0	86.9	81.4	64.1	76.7	75.2
F2M-RN18 mid-term	96.8	69.9	89.2	50.5	48.0	31.8	45.4	48.8	86.6	54.7	88.6	37.4	32.8	76.7	64.7	74.5	52.4	46.3	56.4	60.7
F2F-RN18 mid-term	96.6	67.9	88.6	51.8	47.7	29.9	43.5	47.2	86.0	53.1	88.1	35.8	32.1	75.5	64.7	74.1	43.3	45.9	54.2	59.3

Table 4. Independent F2M model outperforms independent F2F model at pixels which are assigned to the F2M head ($w^{F2M} > .7$) by the F2MF model. We show per-class IoU accuracy on Cityscapes val of our short-term (top) and mid-term (bottom) forecasts.

Forecasted flows from Figure 3 should not be compared to predictions made by dedicated optical flow models. First, our feature flows are predicted by recognizing and extrapolating past events, without observing the future image. Second, our models are not trained with optical flow groundtruth. Our forecasts are trained to reconstruct intermediate feature representations at 32× subsampled resolution. This goal is loose and does not require as accurate flow prediction as in the reconstruction of the RGB frames. Still, the recovered motion is quite good and allows independent F2M models to outperform F2F in many image pixels as will be shown next.

5. Validation of the correlation embedding

Table 3 validates the choice of feature embedding which we apply prior to the spatio-temporal correlation inference. Row 1 shows the baseline mIoU accuracy (no embedding). In this case, the correlation is established across unit feature vectors from the pyramid pooling module. Rows 2-4 correspond to embeddings with one 1×1, one 3×3 and two 3×3 convolutional layers. Single 3×3 convolution outperforms the baseline for 0.5 (short-term) and 1.1 (mid-term) pp mIoU. Single 1×1 convolution is only slightly better than the baseline. This suggests that the success of single-layer 3×3 embedding is due to relative spatial information. The two-layer 3×3 is better than no embedding and single-layer 1×1, but overall worse than single-layer 3×3, which indicates overfitting.

	Short-term		Mid-term	
Embedding	All	MO	All	MO
None	66.5	65.0	54.8	50.9
Conv 1×1	66.6	65.4	54.9	51.2
Conv 3×3	**66.9**	**65.6**	**55.9**	**52.4**
2× Conv 3×3	**66.9**	65.2	55.3	51.4

Table 3. Validation of different metric embeddings for the correlation module presented in the main paper. We show F2MF-RN18 forecasting performance (mIoU accuracy) on Cityscapes val.

6. Per-class accuracy in F2M pixels

Table 4 shows per class results in Cityscapes val pixels which are assigned to the F2M head with probability $w^{F2M} > .7$. These pixels account for 49.9% image content in short-term forecast and 39.8% at mid-term. We observe that F2M outperforms F2F for 0.3pp (short-term) and 1.4pp (mid-term). Note that these metrics can not be obtained from the content of Fig. 6 in the main paper although the two experiments are related. Note that the compound F2MF model would still come out as the winner even under these terms, despite having neglectably larger capacity than independent models.

References

[1] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. 3

[2] 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. 1