applied sciences

MDPI

Article

Dense Semantic Forecasting with Multi-Level Feature Warping

Iva Sović, Josip Šarić and Siniša Šegvić *

Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia
* Correspondence: sinisa.segvic@fer.hr

Abstract: Anticipation of per-pixel semantics in a future unobserved frame is also known as dense semantic forecasting. State-of-the-art methods are based on single-level regression of a subsampled abstract representation of a recognition model. However, single-level regression cannot account for skip connections from the backbone to the upsampling path. We propose to address this shortcoming by warping shallow features from observed images with upsampled feature flow. Our goal is not straightforward, since warping with coarse feature flow introduces noise into the forecasted features. We therefore base our work on single-frame models that are more resistant to the noise in skip connections. To achieve this, we propose a training procedure that enables recognition models to operate reasonably well with or without skip connections. Validation experiments reveal interesting insights into the influence of particular skip connections on recognition accuracy. Our forecasting method delivers 70.2% mIoU 0.18 s into the future and 58.5% mIoU 0.54 s into the future. These experiments show 0.6 mIoU points of improved accuracy with respect to the baseline and reveal promising directions for future work.

Keywords: dense semantic forecasting; dense prediction; semantic segmentation; feature forecasting; future prediction; deep learning; computer vision

Citation: Sović, I.; Šarić, J.; Šegvić, S. Dense Semantic Forecasting with Multi-Level Feature Warping. *Appl. Sci.* **2023**, *13*, 400. https://doi.org/10.3390/app13010400

Academic Editor: Athanasios Nikolaidis

Received: 15 November 2022
Revised: 20 December 2022
Accepted: 20 December 2022
Published: 28 December 2022

1. Introduction

Anticipation of future events is a critical ingredient of intelligent behavior [1,2]. Timely reactions of autonomous robotic systems may avoid material damage or human injury [3,4]. Visual perception is especially prominent in this context due to its suitability for environments designed for humans. We focus on dense semantic forecasting from observed RGB frames [5–7]. State-of-the-art approaches exploit conciseness and rich semantic content of subsampled abstract features [8,9] through single-level feature-to-feature regression. However, such an approach is unable to deliver forecasts with fine-grained details and is especially ineffective for recovering thin vertical objects such as poles, tree trunks, or humans.

Most modern dense prediction models consist of a backbone and the upsampling path. The backbone extracts a set of subsampled features from the input image. The upsampling path gradually increases the resolution of the extracted features in order to achieve spatially precise dense predictions. A prominent approach to improve fine-grained spatial details in single-frame dense prediction relies on skip connections between the backbone and the upsampling path. This approach has been known as feature pyramid networks [10] and ladder-style upsampling [11]. Thus, some dense semantic forecasting approaches independently regress all four intermediate representations that are used along the upsampling path [5,12]. However, application of the forecasting module to the fine-resolution skip connections introduces large computational cost. Furthermore, convolutional forecasting at high resolution poses additional challenges. First, the magnitude of the object displacements increases with the resolution, which makes the convolutional regression very hard to learn. Second, fine-grained forecasting makes little sense in previously unobserved pixels of the scene. Such pixels typically arise due to disocclusion. For instance, this can happen when the ego-camera turns around a large obstacle or when another moving object

uncovers the background scenery. In such instances, even a human operator could forecast only very rough positions of the most common elements of the scene, such as the road, sidewalk, or buildings. Fine-grained semantic information may be quite reliably forecasted in pixels projected from the previously observed parts of the scene. We note that such intermediate representations need not be generated from scratch since they already exist in the forecasting pipeline. If the model succeeds in recognizing the relative motion between the ego camera and the corresponding part of the scene, the forecasting could be carried out through simple copy-pasting of the observed intermediate representations. This discussion suggests that dense semantic forecasting involves two distinct classes of pixels, and that these two classes may call for different forecasting approaches.

This paper considers the problem of reconciling single-level forecasting [9,13] with feature pyramid networks [5,10]. We propose to forecast skip connections by warping intermediate features from the observed images with the feature flow. We follow the best practices from the related work by using a single forecasting module on the coarsest feature resolution. We reuse the forecasted feature flow and upsample it to the corresponding resolution. We promote tolerance of noisy or pruned skip connections by proposing a novel training procedure, which we denote as MixSkip. Furthermore, we fine-tune the upsampling path of the forecasting system in order to give our model a chance to adapt to the forecasted skip connections. For example the model can suppress the skip connections in novel parts of the scene. Experiments show that MixSkip models work well with pruned skip connections, and that it delivers competitive forecasting performance on the Cityscapes dataset.

2. Related Work

Deep learning caused a quantum leap in the performance of visual recognition systems. The progress was first demonstrated on the core computer vision task of image classification [14,15], but the momentum quickly transferred to other tasks as well. For example, fully convolutional networks [16] successfully adapted image classification architectures for dense prediction of semantic classes at the pixel level. Further work [11,17–19] utilized the classification architectures as the backbones of the semantic segmentation models and considered various options to recover the spatial details lost due to the downsampling. SegNet [18] recovers exact positions of pooled features in the backbone by memorizing corresponding indices and gradually recovers the spatial resolution through upsampling and convolutional layers. DeepLab [19] drops some of the pooling layers from the backbone to avoid downsampling and dilates the corresponding convolutional layers to increase the receptive field. PSPNet [17] introduces a module based on pyramid pooling of a convolutional feature map, which introduces a global context into the local features. U-Net [20] architecture is specialized for medical image segmentation and introduces skip connections between the downsampling and the upsampling path. Skip-connections enable the model to mix semantically poor and spatially accurate features from the backbone with semantically rich features from the upsampling path. LadderDenseNet [11] proved that the lean upsampling path with skip connections is a more accurate and efficient solution for recovering the lost spatial details in large-resolution images. Skip-connections have emerged as an indispensable part of the state-of-the-art semantic segmentation architectures [21–23]. This is particularly the case in models aiming at real-time inference speed [24,25]. Our semantic segmentation models follow the SwiftNet baseline [25] architecture with spatial pyramid pooling and skip connections in the upsampling path. We demonstrate that models trained with the proposed MixSkip training procedure are able to perform well with and without skip connections. This suggests that such models are more resistant to noise in skip connections. This is particularly important for forecasting applications because of the noise introduced due to the future uncertainty.

Semantic forecasting deals with predicting the future at the semantic level while observing only frames from the past. Dense semantic forecasting predicts future outputs of dense prediction algorithms such as: semantic segmentation [5], instance segmenta-

tion [26], panoptic segmentation [7], etc. Early work focused on predicting the future semantics by observing semantic predictions from the past. Such an approach is often called semantics-to-semantics (S2S) mapping. The first such work considered forecasting semantic segmentation with a dilated convolutional model that receives past per-class semantic maps [5]. This approach was extended with Bayesian variational inference in order to allow multi-modal forecasts [27]. Large inter-frame motion has been alleviated by means of an encoder–decoder architecture with Conv-LSTM skip connections [28]. This is related to our approach, since we also propose a dense semantic forecasting model with skip connections; however, we propose to transform skip connections through forecasted displacements. Additionally, our method is more efficient as heavy computation is done on a single most-condensed representation.

Many state-of-the-art methods follow the feature-to-feature (F2F) forecasting approach [8,9,12]. The first such approach was applied to instance segmentation based on the Mask R-CNN model above the feature pyramid network [26]. Each level of the feature pyramid is forecasted separately with a sequence of dilated convolutions. APANet [12] improves their approach by introducing connections between the pyramid levels and by forecasting skip connections through ConvLSTM modules. Further work leverages a VAE encoder in order to transform the feature pyramid into a concise representation [9]. The future feature pyramid is obtained by applying the VAE decoder to the forecasted representation. Our approach also combines feature pyramid networks with single-level F2F forecasting. However, we forecast skip connections by warping past features with forecasted feature flow. In comparison, our approach is much simpler to train since it does not require separate training of a VAE model.

DeformF2F [13] uses a single-frame semantic segmentation model without skip connections, which enables single-level F2F forecasting. F2MF [8] extends that idea with the feature-to-motion head and the correlation module to capture spatio–temporal feature relations. The F2M head outputs feature flow, which is used for warping past features into the future. Our approach also uses the F2MF forecasting module but adapts it to the single-frame model with skip connections. We estimate future skip connections by warping their past correspondences with the upsampled feature flow. This is an efficient solution for skip-connection forecasting, but it introduces noise because of the sub-optimal feature flow. The noise could be fatal for semantic segmentation accuracy, so we train our single-frame model to be resistant with the proposed MixSkip procedure.

3. Method

We propose an efficient model for semantic segmentation of future frames based on feature forecasting. Different from previous approaches [12,26,29], our feature-forecasting module operates only on coarse features and efficiently forecasts skip connections at all levels by warping from the past with the upsampled flow. We identify two important methodological contributions that support successful implementation of such a system. First, we propose a novel training scheme that enables semantic segmentation models to perform well in regimes with and without skip connections. This procedure prevents catastrophic performance deterioration of regular ladder-style models in the absence of skip connections. Second, we present an extension to a competitive feature forecasting module [29], which attains compatibility with ladder-style dense prediction architectures.

The rest of this section is organized as follows. We first describe the properties of the ladder-style models suitable for noisy multi-level feature forecasting. Then, we briefly recap the single-frame dense prediction model used in this work. We follow with the presentation of the proposed training procedure that delivers the desired properties. Finally, we describe the proposed extension for multi-level feature forecasting.

3.1. Dense Prediction with Optional Skip Connections

Our goal is to train a dense prediction model that is resistant to possible noise present in the skip connections. This property is important for later application in a future-prediction

scenario, since feature forecasting introduces additional noise. We achieve this by combining a specific ladder-style architecture with a custom training procedure. Our implementation grants usage of the skip connections in two ways and ensures comparable performances for any desired configuration. Skip connections are constructed as detachable links, meaning each one can be turned on or off, enabling the model to be trained at the same time in a single-level and a multi-level way. Turning the skip connections off before they are added to the upsampled features ensures the robustness of the model to the lack of information from the skip connections. Implementation-wise, this flexibility requires a ladder-style model that combines the skip connections and upsampled features through element-wise addition.

We adopt a general single-scale SwiftNet [25] model, as it utilizes a ladder-style architecture that satisfies the before-mentioned requirements. Furthermore, this architecture achieves a great accuracy/efficiency ratio, which is important for our computationally intensive forecasting experiments. Figure 1a illustrates the SwiftNet architecture. It consists of a backbone (downsampling path, yellow), spatial pyramid pooling (green), and a decoder (upsampling path, blue). The two paths are linked in a ladder-style manner. As usual, the backbone corresponds to the ImageNet pre-trained [30] classification model without the final fully connected layer. We experiment with ResNet-18 [31] and DenseNet-121 [32]. The upsampling path consists of three upsampling modules, each increasing the spatial resolution by a factor of 2. Figure 1b gives a closer look at a single upsampling module. The module receives two inputs: features extracted by the backbone (a skip connection) and smaller-resolution features from the upsampling path. The upsampling path features are first bilinearly interpolated ($2\times\uparrow$) to match the spatial resolution of a skip connection. The skip connection features are processed with a single BN-ReLU-CONV unit in order to match the channel dimension with the upsampling path. The features are finally fused via element-wise addition and processed with another BN-ReLU-CONV unit.

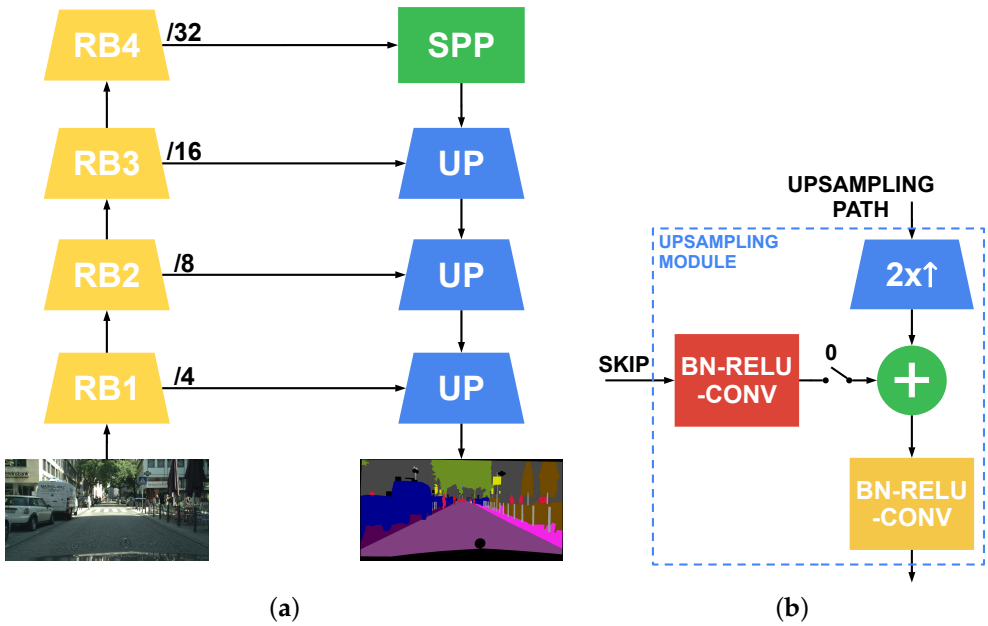

Figure 1. (**a**) SwiftNet model for semantic segmentation based on ladder-style architecture. (**b**) Closer look at the SwiftNet upsampling module which fuses skip connections with the main features from the upsampling path via element-wise addition.

Prior to the addition, we visualize a switch that can turn off the skip connection depending on the desired configuration. The switch is managed by the training procedure, which we describe next.

3.2. MixSkip Training Procedure

The proposed training procedure, which we call MixSkip, optimizes the model in such a way that it minimizes the loss function with and without skip connections. A single training step is divided in a couple of steps. First, we acquire a batch of labeled images and split it into two equal parts. Second, we turn all skip connections on, run a forward pass, and compute the loss for the first part of the batch. Third, we turn all skip connections off, run a forward pass, and compute the loss for the second part of the batch. The total loss, accordingly, consists of two parts:

$$\mathcal{L} = \mathcal{L}_{\text{skip}} + \mathcal{L}_{\text{no\_skip}} \tag{1}$$

As we optimize the total loss $\mathcal{L}$, the gradients of both components are averaged. Except for the skip connections, part of the graph is optimized only by the $\mathcal{L}_{\text{skip}}$ component. Figure 2 illustrates a single training step of the proposed MixSkip procedure using PyTorch-like code.

```
skip_batch = batch[:bs//2]
no_skip_batch = batch[bs//2:]

model.set_upsample_skip(True)
loss_skip = model.loss(skip_batch)

model.set_upsample_skip(False)
loss_no_skip = model.loss(no_skip_batch)

loss = (loss_skip + loss_no_skip) / 2
loss.backward()
optimizer.step()
```

Figure 2. Code illustration of our MixSkip training procedure.

3.3. F2MF-Based Multi-Level Feature Forecasting

Dense semantic forecasting estimates future semantic predictions from observed frames. Recently, a method based on feature-to-feature (F2F) forecasting achieved a great accuracy/efficiency ratio [29]. In order to be efficient, this method requires a single-frame recognition model with no skip connections. However, this requirement comes at the cost of reduced single-frame recognition accuracy, which affects the forecasting performance. We propose to mitigate this issue by training the recognition model with the MixSkip procedure and combining it with F2MF module adapted for multi-level feature forecasting.

The F2MF module forecasts future features by combining the predictions from two heads. The F2F head forecasts the features directly. The F2M head forecasts by separately warping each past feature representation into the future according to the predicted flow. The multiple forecasts are blended according to the densely predicted weights. The original work only considers single-level applications of the F2MF module to custom dense recognition architectures. Our extension enables application of the F2MF module to ladder-style architectures. In particular, we reuse the flow and weight predictions from the F2M module to warp the representations coming from skip connections. This minimally affects the forecasting efficiency since most of the computation is still done on the most-condensed representation of the recognition model. However, the predicted feature flow and blending weights must be upsampled prior to warping in order to match the resolution of the corresponding skip connections. This is suboptimal because the upsampled flow is coarse and noisy. Consequently, we train our recognition model to be more resistant to noise in skip connections. Additionally, we fine-tune the upsampling path in order to adapt to the forecasted features.

Figure 3 shows our dense-semantic-forecasting system. The inference pipeline is as follows. First, we extract features from each of the four past images independently. The

feature extraction involves the backbone (yellow) and the SPP module (green). As usual, the F2MF module (red) maps past SPP features into the future counterparts. Furthermore, we bring the predicted feature flow and blending weights to the three warp modules (purple). Each warp module additionally takes the past features from the corresponding skip connection. First, it upsamples the feature flow and the blending weights. Note that the flow vectors also have to be scaled with the upsampling factor. Second, it warps and blends the skip connections features into the future representation. Finally, all of the forecasted features (SPP and skip connections) are interpreted by the upsampling path in order to recover the future semantic predictions.

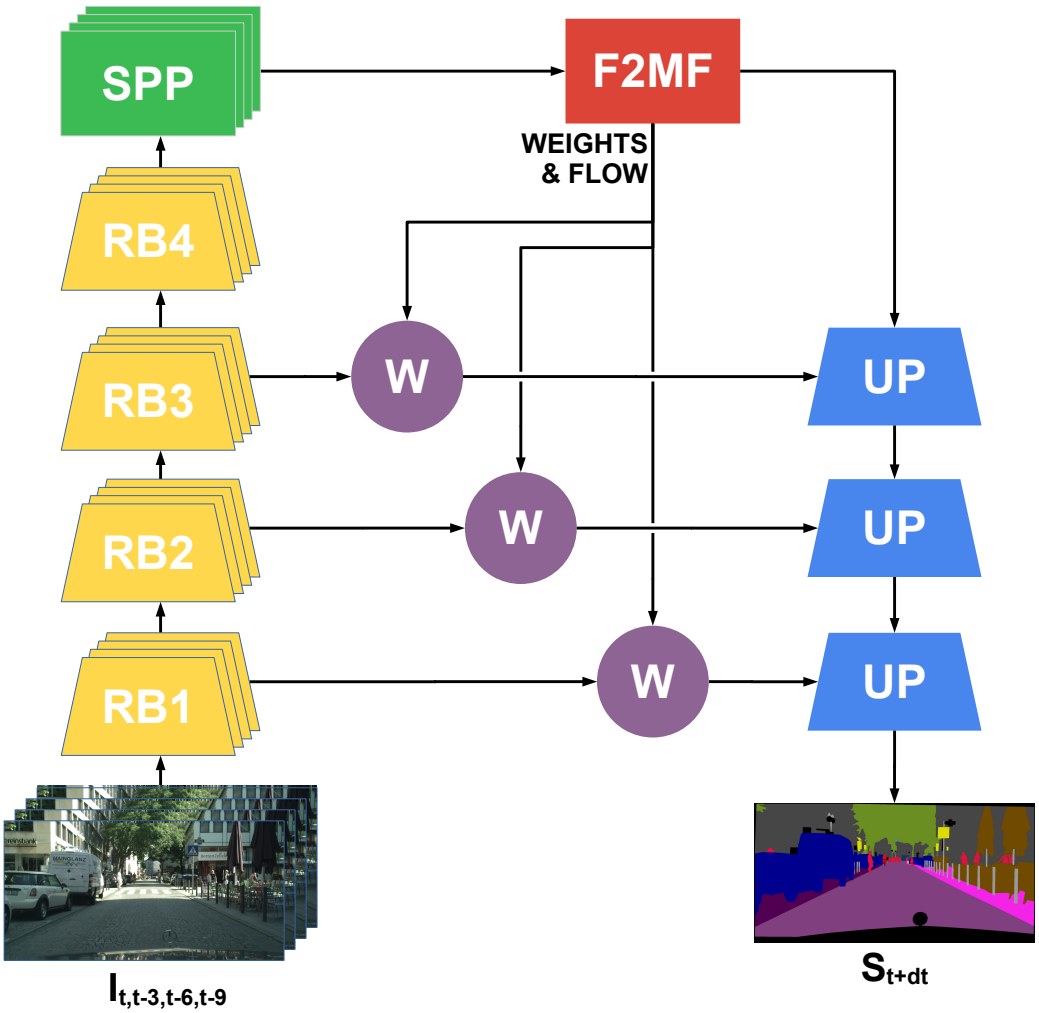

Figure 3. Dense semantic forecasting based on F2F regression and multi-level F2M warping.

4. Experiments

We conduct our experiments on the Cityscapes dataset [33], which contains 2975 training, 1525 testing, and 500 validation video snippets. Each snippet contains 30 images, with the 20th image being annotated. Each image has a resolution of 1024 × 2048 pixels. We estimate our models' performances using mean Intersection over Union for all classes (mIoU) and separately for eight classes that represent moving objects (mIoU-MO). We evaluate our single-frame semantic segmentation model as well as the forecasting variant. For completeness, here we briefly recap the mIoU metric for semantic segmentation. As shown in Equatioon (2), mIoU measures the ratio of intersection and union between predictions and ground truth averaged over all classes. If there is a perfect match between the predictions and the ground truth, this ratio is equal to 1. We follow a common practice in

the community and aggregate predictions across all images of the corresponding subset and then compute the metric.

$$\text{mIoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\text{PRED}_c \cap \text{GT}_c}{\text{PRED}_c \cup \text{GT}_c} \tag{2}$$

We conduct our experiments on a single NVIDIA A4000 GPU with 16 GB of RAM. Training of the semantic segmentation models with the MixSkip procedure takes around 2 days on our hardware. During the first stage of the forecasting training, we load cached features from the SSD drive, which speeds up the training. This stage takes around 6 h. The fine-tuning stage takes around 8 h because in each training step we have to extract features for all frames in the batch.

4.1. Semantic Segmentation

4.1.1. Training Details

We train our single-frame semantic segmentation model according to the proposed MixSkip procedure. Note that we initialize our backbone with ImageNet [30] weights. We train our model for 500 epochs, with batch size 14 and cross-entropy loss. We set the learning rate to 4×10^{-4} and decay it with cosine annealing to the minimal value of 10^{-6}. Our data augmentation involves random scaling, horizontal flipping, and cropping with size set to 768×768 pixels.

Table 1 compares the regular semantic segmentation training with the proposed MixSkip procedure. We evaluate two models based on ResNet-18 and DenseNet-121 backbones in two regimes: with and without skip connections. The results show that both procedures achieve comparable results when evaluating the models with skip connections turned on. However, models trained with the regular procedure reveal serious performance deterioration when evaluated without the skip connections. In particular, the model based on ResNet-18 loses 31 mIoU points if trained with the regular procedure and 2.3 mIoU points if trained with MixSkip. The difference is even bigger with the model based on DenseNet-121. Regular training leads to almost 60 mIoU points lost, while MixSkip reduces that drop to 2.8 mIoU points. We also notice that the stronger DenseNet-121 backbone improves by 2.4 points over the weaker ResNet-18 backbone. These results suggest that the models trained with the MixSkip procedure are more noise resistant and more appropriate for the forecasting application.

Table 1. Evaluation (mIoU) of semantic segmentation models with and without skip connections on Cityscapes val. Models are trained with regular or mix-skip procedure.

Model	w/o Skips	w/ Skips
RN-18	44.1	75.0
RN-18-MixSkip	72.6	74.9
DN-121	18.0	77.4
DN-121-MixSkip	75.5	78.3

4.1.2. Evaluation of Different Skip Connection Configurations

Table 2 evaluates the two models trained with the MixSkip procedure with each possible skip-connection configuration. We denote each skip connection according to the corresponding output stride. For example, "/16" denotes the coarsest skip connection, with feature resolution 16× smaller compared to that of the input image. Each row of the table evaluates the same model but with different skip-connection configurations: ✓—denotes that the corresponding skip connection is used by the model, while ✗—denotes that it is set to zero. The results show that the coarsest skip connection brings small improvements or even reduces the accuracy compared to the model without skips. The biggest contribution comes from the other two feature resolutions. Nevertheless, both models work best when

all skip connections are turned on. This suggests that the synergy of skips is better than selecting only those that contribute the most.

Table 2. Validation of per-skip contributions on semantic segmentation accuracy (mIoU) on Cityscapes val.

Configuration			Model	
/16	/8	/4	RN-18-MixSkip	DN-121-MixSkip
✗	✗	✗	72.6	75.5
✗	✗	✓	73.9	77.0
✗	✓	✗	74.3	77.4
✗	✓	✓	74.8	78.0
✓	✗	✗	72.5	75.7
✓	✗	✓	74.0	77.3
✓	✓	✗	74.2	77.5
✓	✓	✓	**74.9**	**78.3**

4.1.3. Validation of Number of Epochs in MixSkip Training Setup

We hypothesize that the MixSkip procedure needs more training steps to converge due to the fact that the model is trained with two configurations in parallel. Consequently, each model configuration in a single training step sees only half of the batch. Thus, we train our models for 500 epochs, which is 2× longer than usual. Table 3 shows the results for DenseNet-121- and ResNet-18-based models evaluated with and without skip connections. The results show that longer training consistently improves accuracy for both configurations. Longer training improves the ResNet model for 0.2 mIoU points when evaluated with skip connections and 0.5 mIoU points when evaluated without skip connections. The difference is bigger for the DenseNet model: 0.5 mIoU points with skip connections on and 0.8 mIoU points without the skip connections. We believe this is due to the larger capacity of the DenseNet model.

Table 3. Influence of number of training epochs on different models trained with the MixSkip procedure. Models are evaluated with and without skip connections on Cityscapes val.

Model	#Epochs	w/o Skips	w/ Skips
RN-18-MixSkip	250	72.1	74.7
RN-18-MixSkip	500	72.6	74.9
DN-121-MixSkip	250	74.7	77.8
DN-121-MixSkip	500	75.5	78.3

4.2. Semantic Segmentation Forecasting

4.2.1. Training Details

The training pipeline for our dense semantic forecasting system involves a couple of steps. First, we independently train the semantic segmentation model according to the MixSkip procedure. We use that model to extract the SPP features used for the forecasting training. Second, we train the F2MF module as proposed in [29]. Finally, we fine-tune our upsampling path with cross-entropy loss. We train the forecasting model for 240 epochs. We set auxiliary loss to 1.7 and 0.8 for F2M and F2F forecasting, respectively. We set the learning rate to 4×10^{-4} and decay it with cosine annealing to the minimal value of 1×10^{-7}. We fine-tune our model for 10 epochs. During fine-tuning, we set the learning rate to 4×10^{-5} and decay it with cosine annealing to the minimal value of 1×10^{-7}. All skip connections are turned on during fine-tuning.

4.2.2. Multi-Level Feature Forecasting for Future Semantic Segmentation

Table 4 compares our forecasting model with the results from the literature. As in the literature, we report results for short-term (0.18 s) and mid-term (0.54 s) forecasting. We

also show our upper- and lower-bound forecasting performance. The upper bound (oracle) is attained by applying our single-frame DN-121-MixSkip model in the future frame. It is called the oracle because this model sees the future frame, which is otherwise unavailable in the forecasting setup. Thus, it represents the upper-bound performance because we try to imitate its predictions through forecasting. Our oracle with skip connections achieves 78.3% mIoU. The lower bound is set with the copy-last (DN-121) model. This baseline takes the semantic segmentation of the last input frame and copies it to the output [5]. Thus, it assumes a static scene that did not change between the last observed and the future moment. This baseline achieves 52.6% and 38.6% mIoU for short-term and mid-term forecasting, respectively. Our forecasting model with the DenseNet-121 backbone beats this simple baseline by a large margin and achieves competitive results. In particular, we achieve 70.2% and 58.5% mIoU accuracy for short-term and mid-term, respectively. The state-of-the-art [9] achieves 0.9 mIoU points better accuracy in short-term and 1.8 mIoU points in mid-term forecasting. We notice that our method improves over the F2MF baseline by 0.6 mIoU points in short-term and mid-term forecasting.

Table 4. Evaluation of short-term and mid-term forecasting accuracy (mIoU) on Cityscapes val. *All* denotes all classes; *MO* denotes moving objects.

Accuracy (mIoU)	**Short-Term**		**Mid-Term**	
	All	**MO**	**All**	**MO**
Oracle (DN-121-MixSkip)	78.3	78.8	78.3	78.8
Copy last (DN-121-MixSkip)	52.6	48.5	38.6	30.0
3Dconv-F2F [34]	57.0	/	40.8	/
Dil10-S2S [5]	59.4	55.3	47.8	40.8
LSTM S2S [28]	60.1	/	/	/
Mask-F2F [26]	/	61.2	/	41.2
FeatReproj3D [35]	61.5	/	45.4	/
Bayesian S2S [27]	65.1	/	51.2	/
DeformF2F [13]	65.5	63.8	53.6	49.9
LSTM AM S2S [36]	65.8	/	51.3	/
LSTM M2M [6]	67.1	65.1	51.5	46.3
ApaNet [12]	/	64.9	/	51.4
F2MF [8]	69.6	67.7	57.9	54.6
Panoptic Forecasting [7]	67.6	60.8	58.1	52.1
LSTM-VAE-ML-F2F [9]	71.1	69.2	60.3	56.7
ML-F2MF-DN-121 (ours)	70.2	69.0	58.5	55.9

4.2.3. Significance of Skip Connections for Feature Forecasting

Table 5 investigates the influence of skip connections for both the single-frame (oracle) and the forecasting model. The top section refers to the ResNet-based models and the bottom section to the DenseNet-based models. For the segmentation model, we observe 2.3pp drop in accuracy for RN18 and 2.8pp for DN-121 when we do not use skip connections. Although the forecasting model (denoted as ML-F2MF) does not achieve comparable improvement, we still observe a consistent gain in accuracy when we use skip connections. For RN-18, we achieve 68.3% mIoU for short-term and 55.9% mIoU for mid-term, which is 0.6% and 0.5% better, respectively, compared to the setting in which our model does not use skip connections. For DN-121, this improvement is 0.4% and 0.2% for short-term and mid-term, respectively.

Table 5. Evaluation (mIoU) of our ML-F2MF forecasting model with single-frame segmentation models based on ResNet-18 (top section) and DenseNet-121 (bottom section) backbones. We evaluate models with and without skip-connection forecasting.

Model	Short-Term		Mid-Term	
	All	MO	All	MO
Oracle-RN-18 w/o skips	72.6	71.9	72.6	71.9
Oracle-RN-18 w/ skips	74.9	73.9	74.9	73.9
ML-F2MF-RN-18 w/o skips	67.7	66.5	55.4	51.2
ML-F2MF-RN-18 w/ skips	68.3	67.4	55.9	51.8
Oracle-DN-121 w/o skips	75.5	76.4	75.5	76.4
Oracle-DN-121 w/ skips	78.3	78.8	78.3	78.8
ML-F2MF-DN-121 w/o skips	69.8	68.7	58.3	55.7
ML-F2MF-DN-121 w/ skips	70.2	69.0	58.5	55.9

4.2.4. Fine-Tuning of the Upsampling Path

Table 6 investigates the influence of fine-tuning. During fine-tuning, we optimize parameters of the upsampling path of the segmentation model. This allows the upsampling path to adapt to the distribution shift of the features due to the forecasting. All values presented in the table are obtained by using the skip connections. Note that models denoted with 'w/o ft.' are not fine-tuned, but they use skip connections. The first section of the table evaluates models based on ResNet-18. Fine-tuning brings no improvement to the short-term forecast, while it improves the mid-term forecast by 0.2 mIoU points. The second section shows the results for the model based on DenseNet-121. For DenseNet-121, we observe 0.6pp improvement in both short-term and mid-term mIoU accuracy.

Table 6. Validation of upsampling path fine-tuning for semantic segmentation forecasting.

Model	Short-Term		Mid-Term	
	All	MO	All	MO
ML-F2MF-RN-18 w/o ft.	68.3	67.4	55.7	51.4
ML-F2MF-RN-18 w/ ft.	68.3	67.4	55.9	51.8
ML-F2MF-DN-121 w/o ft.	69.6	68.7	57.9	55.1
ML-F2MF-DN-121 w/ ft.	70.2	69.0	58.5	55.9

4.2.5. Data Augmentation in Feature-Forecasting Training

We investigate the influence of data augmentation on the forecasting performance. Our data augmentation includes random-use horizontal flipping and sliding of time frames. The latter is possible since we do not need ground-truth labels for our basic forecasting training. Thus, we can randomly pick our start frame and every following frame accordingly. Table 7 shows consistent improvement of mIoU across all models and for both short-term and mid-term forecasting. The forecasting model obviously benefits from a slight increase in data diversity. With RN-18, we achieve 0.7pp improvement in accuracy for short-term and 0.6pp for mid-term forecasting. Our best model, based on DenseNet-121, benefits 0.6 mIoU points in short-term and 1.1 mIoU points in mid-term forecasting due to fine-tuning.

Table 7. Ablation of data augmentation during forecasting training.

Model	Short-Term		Mid-Term	
	All	**MO**	**All**	**MO**
RN-18 w/o d.a.	67.6	66.7	55.3	50.7
RN-18 w/ d.a.	68.3	67.4	55.9	51.8
DN-121 w/o d.a.	69.6	68.4	57.4	54.0
DN-121 w/ d.a.	70.2	69.0	58.5	55.9

4.2.6. Evaluation with Oracle Skip Connections

We investigate if our forecasting performance could benefit from better forecasting of skip connections exclusively. In particular, we evaluate our forecasting model with skip connections originating from one of the oracles. The first oracle corresponds to the model that uses oracle optical flow for warping the most recent features into the future. We recover optical flow between the last observed frame and the future frame using RAFT [37]. Note that the difference between the two frames is three timesteps in the short-term forecast, and nine timesteps in mid-term forecast. The second oracle uses skip connections computed by applying the single-frame model in the future frame. Table 8 shows the results of the forecasting model and two oracles. The optical flow oracle achieves 2.2 mIoU points better performance in short-term forecasting than our ML-F2MF. However, in mid-term forecasting, oracle flow aggravates the performance by 0.6 mIoU points. This is not surprising since mid-term forecasting implies large displacements due to the nine-timestep difference between the future and the most recent frame. We believe that this is unnatural for the optical flow model. Thus, it struggles to recover accurate flow. The last row in the table shows the performance of our model when we use oracle skip connections instead of the ones we get with forecasting. The forecasting model lags behind the oracle by 4.7 mIoU points in short-term and 5.7 mIoU points in mid-term forecasting. The results suggest that there is space for improvement in the skip-connection forecasting.

Table 8. Evaluation of the forecasting model with oracle skip connections.

Model	Short-Term		Mid-Term	
	All	**MO**	**All**	**MO**
ML-F2MF-DN-121	70.2	69.0	58.5	55.9
w/ oracle optical flow	72.4	70.8	57.9	55.4
w/ oracle skip conn.	74.9	73.8	64.2	61.0

4.2.7. Qualitative Results

Figure 4 shows qualitative results of our short-term and mid-term forecast. The rows show: the last observed frame, the future frame (unobserved), ground truth, and our forecast. We overlay the ground truth and our forecast over the future frame for the sake of more informative visualization. We also measure the ratio of correctly classified pixels for each of the examples individually to get a better sense of model accuracy. The results are as follows (col. 1–4): 93.3%, 90.0%, 95.0%, and 94.1%. The results show that our forecast is mainly accurate. In some cases, we are even able to recover some fascinating details. For example, in the second example of the mid-term forecast, our model accurately forecasts the pose of the pedestrians legs, which is extremely hard.

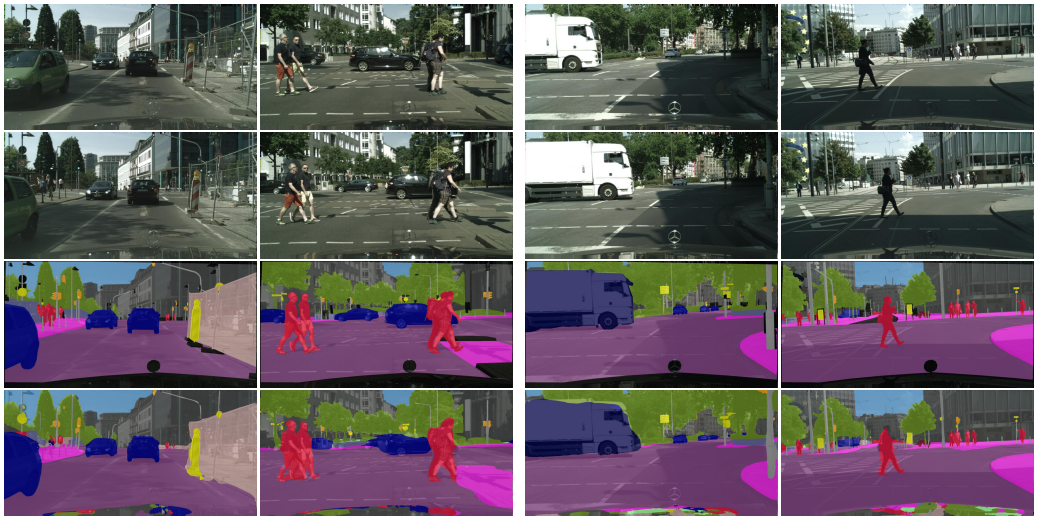

Figure 4. Two qualitative examples of short-term and two of mid-term forecasting on Cityscapes val. The rows show: the last observed frame, future frame (unobserved), ground truth, and our forecast.

5. Conclusions

We have presented a novel feature-to-feature forecasting system for semantic segmentation of future frames. Our system is based on efficient multi-level feature forecasting. We base our work on a feature-to-motion module that forecasts the future flow and uses it to warp past representations into the future. Our method applies this module on the coarsest features to preserve efficiency, and reuses the predicted feature flow to warp the skip connections. The warped skip-connection features are noisy because of the coarse feature flow. To account for that, we train our recognition model with a novel training procedure called MixSkip. MixSkip allows us to train a single ladder-style model that performs comparably to specialized models that train with and without skip connections. Consequently, MixSkip models become more resistant to the noise in skip connections. Our forecasting system achieves competitive accuracy in semantic segmentation forecasting on the Cityscapes dataset. Additionally, we provide extensive validation experiments that reveal the influence of particular skip connections on semantic segmentation accuracy. Experiments also suggest that better skip-connection forecasting may significantly improve the dense semantic forecasting performance. Thus, one direction for future work involves research on efficient methods for skip-connection forecasting. Other possible directions include multi-modal semantic forecasting and applications to different recognition tasks such as instance segmentation or panoptic segmentation.

Author Contributions: Conceptualization, S.Š.; methodology, I.S., J.Š. and S.Š.; software, I.S.; validation, I.S.; investigation, I.S. and J.Š.; writing—original draft preparation, I.S.; writing—review and editing, J.Š. and S.Š.; visualization, J.Š. and I.S.; supervision, S.Š.; project administration, S.Š.; funding acquisition, S.Š. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by Rimac Technologies, Croatian Science Foundation under grant IP-2020-02-5851 ADEPT, and the European Regional Development Fund under grant KK.01.1.1.01.0009 DATACROSS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; Savarese, S. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
2. Vondrick, C.; Pirsiavash, H.; Torralba, A. Anticipating Visual Representations from Unlabeled Video. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 98–106.
3. Yao, Y.; Xu, M.; Choi, C.; Crandall, D.J.; Atkins, E.M.; Dariush, B. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 9711–9717.
4. Hu, A.; Cotter, F.; Mohan, N.; Gurau, C.; Kendall, A. Probabilistic Future Prediction for Video Scene Understanding. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020.
5. Luc, P.; Neverova, N.; Couprie, C.; Verbeek, J.; LeCun, Y. Predicting deeper into the future of semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 648–657.
6. Terwilliger, A.; Brazil, G.; Liu, X. Recurrent Flow-Guided Semantic Forecasting. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1703–1712.
7. Graber, C.; Tsai, G.; Firman, M.; Brostow, G.; Schwing, A.G. Panoptic Segmentation Forecasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12517–12526.
8. Saric, J.; Orsic, M.; Antunovic, T.; Vrazic, S.; Segvic, S. Warp to the Future: Joint Forecasting of Features and Feature Motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
9. Lin, Z.; Sun, J.; Hu, J.F.; Yu, Q.; Lai, J.H.; Zheng, W.S. Predictive Feature Learning for Future Segmentation Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 7365–7374.
10. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
11. Kreso, I.; Segvic, S.; Krapac, J. Ladder-Style DenseNets for Semantic Segmentation of Large Natural Images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017.
12. Hu, J.F.; Sun, J.; Lin, Z.; Lai, J.H.; Zeng, W.; Zheng, W.S. APANet: Auto-Path Aggregation for Future Instance Segmentation Prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3386–3403. [CrossRef] [PubMed]
13. Šarić, J.; Oršić, M.; Antunović, T.; Vražić, S.; Šegvić, S. Single level feature-to-feature forecasting with deformable convolutions. In Proceedings of the German Conference on Pattern Recognition, Dortmund, Germany, 10–13 September 2019; Springer: Cham, Switzerland, 2019; pp. 189–202.
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
15. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
17. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
18. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
19. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
21. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
22. Cheng, B.; Collins, M.D.; Zhu, Y.; Liu, T.; Huang, T.S.; Adam, H.; Chen, L.C. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12475–12485.
23. Krešo, I.; Krapac, J.; Šegvić, S. Efficient ladder-style densenets for semantic segmentation of large images. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4951–4961. [CrossRef]

24. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
25. Orsic, M.; Kreso, I.; Bevandic, P.; Segvic, S. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
26. Luc, P.; Couprie, C.; Lecun, Y.; Verbeek, J. Predicting Future Instance Segmentation by Forecasting Convolutional Features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 584–599.
27. Bhattacharyya, A.; Fritz, M.; Schiele, B. Bayesian Prediction of Future Street Scenes using Synthetic Likelihoods. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
28. Nabavi, S.S.; Rochan, M.; Wang, Y. Future Semantic Segmentation with Convolutional LSTM. *arXiv* **2018**, arXiv:1807.07946.
29. Šarić, J.; Vražić, S.; Šegvić, S. Dense semantic forecasting in video by joint regression of features and feature motion. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–13. [CrossRef] [PubMed]
30. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
32. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
33. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
34. Chiu, H.k.; Adeli, E.; Niebles, J.C. Segmenting the future. *IEEE Robot. Autom. Lett.* **2020**, *5*, 4202–4209. [CrossRef]
35. Vora, S.; Mahjourian, R.; Pirk, S.; Angelova, A. Future Semantic Segmentation Using 3D Structure. In Proceedings of the ECCV 3D Reconstruction meets Semantics Workshop, Munich, Germany, 8–14 September 2018.
36. Chen, X.; Han, Y. Multi-Timescale Context Encoding for Scene Parsing Prediction. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1624–1629.
37. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 402–419.