Learning Semantic-Aware Dynamics for Video Prediction

Xinzhu Bei
UCLA Vision Lab
xzbei@cs.ucla.edu

Yanchao Yang
Stanford University
yanchaoy@cs.stanford.edu

Stefano Soatto
UCLA Vision Lab
soatto@cs.ucla.edu

Abstract

We propose an architecture and training scheme to predict video frames by explicitly modeling dis-occlusions and capturing the evolution of semantically consistent regions in the video. The scene layout (semantic map) and motion (optical flow) are decomposed into layers, which are predicted and fused with their context to generate future layouts and motions. The appearance of the scene is warped from past frames using the predicted motion in co-visible regions; dis-occluded regions are synthesized with content-aware inpainting utilizing the predicted scene layout. The result is a predictive model that explicitly represents objects and learns their class-specific motion, which we evaluate on video prediction benchmarks.

1. Introduction

Anticipating the future is critical for autonomous agents to operate intelligently in the environment, such as for navigation, manipulation, and other forms of physical interaction. We hypothesize that decomposing the scene into independent entities, each with its own attributes, is beneficial to prediction. For example, in Fig. 1, different objects have different geometry and motion, which induces distinctive temporal changes in the video.

We propose a video prediction architecture that explicitly models the different dynamics of semantically consistent regions (Fig. 2). The model, described in detail in Sec. 3.1, decomposes the video into regions, corresponding to different semantic classes in the scene, and learns class-specific characteristics while ensuring that their re-composition can predict the image, along with class labels and flow fields.

Unlike warping the past using globally predicted flow fields [20, 19, 27, 9], in our semantic-aware dynamic model (SADM), local regions are represented by binary semantic masks, whose evolution is simpler and easier to learn than the motion of the entire video frames (see Fig. 1). Each of the regions is predicted and then fused with its content to generate future semantic maps and flow fields. The prediction in co-visible regions of future frames is warped from the past, with dis-occlusion detection mediated by the predicted semantic maps. Furthermore, the *dis-occluded* regions are filled-in by a generative model or conditional renderer, trained with not only the warped images, but also the predicted semantic maps, enabling more structured and semantically-aware synthesis. Modeling dis-occlusions explicitly spares the model the effort otherwise needed to learn this complex phenomenon.

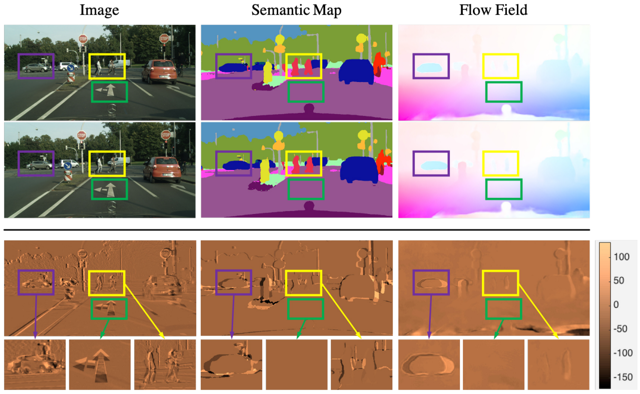


Figure 1. Different representations (video frame, semantic map, flow field) have dynamics with different complexity. Also, different classes have different dynamics within a given representation. Top: a sequence of video frames (left), semantic maps (middle), and flow fields (right). Bottom: dynamics or changes visualized in terms of their difference. The dynamics in video frames is much more complex than that in semantic maps and flow fields.

We incorporate semantic segmentation (scene layout), optical flow (scene motion) and synthesis (scene appearance) into a complete generative model for videos, which facilitates semantically and geometrically consistent prediction of complete video frames. SADM achieves state-of-the-art performance in video prediction benchmarks such as [7, 11, 10].

2. Related Work

Video generation methods produce image sequences either from noise [41] or other input including pose [5] and text [26]. SVG-LP [8] proposes to sample noise from

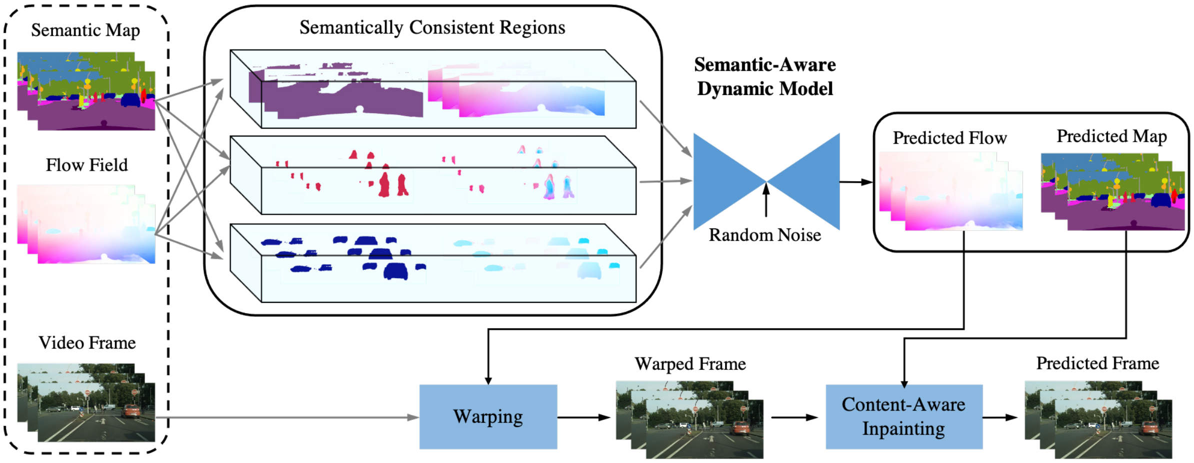


Figure 2. Our video prediction architecture with learned semantic-aware dynamics. It first decomposes the scene into semantically consistent regions to facilitate the modeling of class-specific characteristics. Each region is predicted and fused to generate the future scene layout (semantic map) and motion (flow field) using the proposed semantic-aware dynamic model. Content-aware video inpainting for dis-occlusions is performed after warping to generate the future video frames.

learned priors; MoCoGAN [38] samples latent variables from the motion and content spaces separately to improve temporal consistency. Similarly, TGAN [34] employs a temporal generator and an image generator to model temporal correlations; [13] models the dynamics in the latent space with attribute controls. Given that the visual scene is highly structured, [40, 5, 53] propose to generate a sequence of poses, which are transformed into images for human action sequences; [57] generates videos of a single object by first generating a sequence of conditions using a 3D morphable model, while [12] controls the video generation using sparse trajectories specified by the user. VGAN [41] trains video generators with explicit separation of the foreground and background, assuming static background. Seg2vid [27] resorts to warping using flows generated by the semantic mask, hoping to preserve the scene structure implicitly. We also employ semantic maps in the generation of future flows but with a semantic-aware dynamic model. [54] decomposes images into objects utilizing contextual information separation [55] and synthesizes motion of single objects through perturbations in the object-centric latent space.

Video prediction models are typically approximations of conditional generative models [14, 1, 47, 41, 34, 8, 38, 30, 50]. The quality of predictions is typically evaluated by image quality and temporal consistency. Given the high complexity and dimensionality of the signal to be predicted, the process usually requires explicit modeling or constraints [27, 9]. PredNet [23] proposes a predictive model with coding-based regularization. ContextVP [4] uses a context-aware module with parallel LSTMs. SDC-Net [30] applies flow guided spatially-displaced convolutions, while [15] predicts with dynamic filters that depend on the inputs. DDPAE [14] and [47] map the observed images to a low-dimensional space, so temporal correlations are easier to learn. TPK [42] predicts future poses to guide appearance changes. To address the loss of realism, [29, 20, 25, 19] explicitly model the flows, and DVF [22] uses flow to synthesize future frames. Similar to MCNet [39] and [52], DPG [9] proposes motion-specific propagation and motion-agnostic generation with confidence-based occlusion maps. [24] predicts future semantic maps, and [16] jointly predicts the future semantic maps and flow fields. We use a semantic-aware model such that the predicted maps can exploit class-specific motion priors. Our method generates both future optical flows and semantic maps before rendering future images. In [48], moving object segmentation masks are used, but restricted to 2D affine motions, with two categories: moving and static.

Image inpainting [28, 51, 17], image synthesis [3, 31, 44], and video-to-video synthesis [43] are also related to our handling of dis-occlusions.

3. Method

Notation and goal. Let $x^t \in \mathbb{R}^{H\times W\times 3}$ be a video frame at time t, with $f^t \in \mathbb{R}^{H\times W\times 2}$ and $m^t \in \{1, 2, ..., C\}^{H\times W}$ be the corresponding optical flow field and semantic map respectively. Here C is the number of semantic classes in the semantic map. Given past observations $\{x^t, f^t, m^t\}_{t=1}^{T}$ up to time T, our goal is to predict K frames into the future, i.e., $\{x^t\}_{t=T+1}^{T+K}$. The predictions should match the statistics, quality and content of past frames of the same scene, and exhibit variations that are consistent with the motion of objects within. As illustrated in Fig. 2, our approach falls into the direction of prediction by propagation, where video

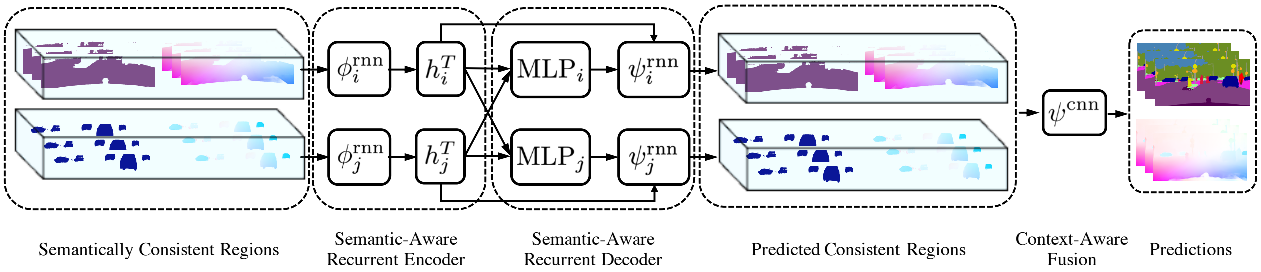


Figure 3. The architecture of our semantic-aware dynamic model (SADM) for learning class-specific dynamics of the scene layout and motion. Input semantic maps and flow fields are parsed and processed by the semantic-aware recurrent encoders ϕ^{rnn} and decoders ψ^{rnn}, with context incorporated into the prediction through a multi-layer perceptron. The predictions of semantically consistent regions are combined by the fusion network ψ^{cnn} to generate the prediction on the whole image domain, which further improves contextual compatibility in the predicted semantic maps and flow fields. The illustration is for two classes, but can be easily extended to more classes.

prediction for the co-visible part[1] of the scene can be accomplished by warping, i.e., propagating pixels via the corresponding flow field.

3.1. Semantic-Aware Dynamic Model

We aim to explicitly model the semantic aware dynamics of both the semantic maps (scene layout) and flow fields (scene motion). In general, the proposed semantic aware dynamic model takes as input the flow fields and semantic maps up to time T, i.e., $\{(m^t, f^t)\}_{t=1}^T$ and outputs the K future flow fields and semantic maps $\{(m^t, f^t)\}_{t=T+1}^{T+K}$, as shown in Fig. 3. The components therein are elaborated below.

Semantic-aware recurrent encoder. Let $m_c^t = \mathbb{1}(m^t = c)$ be the binary mask that indicates the region of semantic class c at time t. Similarly, $f_c^t = f^t \cdot m_c^t$ is the masked flow field showing only the motion of the pixels that are classified as c. The semantic aware recurrent encoder ϕ_c^{rnn} will be operating recursively to produce a hidden representation of the past $\{(m_c^t, f_c^t)\}_{t=1}^T$ while enforcing temporal continuity of the representation:

$$h_c^t = \phi_c^{\text{rnn}}([m_c^t, f_c^t], h_c^{t-1}) \tag{1}$$

with h_c^T the hidden representation that summarizes the past regions and flow fields of the pixels within class c, up to time T. For now, we instantiate C such semantic aware recurrent encoders, $\{\phi_c^{\text{rnn}}\}_{c=1}^C$, which together generate the hidden representation $H^T = \{h_c^T\}_{c=1}^C$ that summarizes the past semantic maps and flow fields, covering all semantic classes.

Note that the collection H^T explicitly represents independent objects. While this may appear inefficient, in reality the model reduces the number of parameters needed, since the individual objects are simpler to represent. We also carry out an ablation study (in the supplementary) on different C's by merging some of the semantically similar classes, showing the accuracy-efficiency trade-offs. Moreover, we can easily parallelize the computation using the grouped convolution operator proposed in [18]. Next, we describe the procedure to predict the future semantic maps and flow fields.

Semantic aware recurrent decoder. Given the hidden representation of the past, $H^T = \{h_c^T\}_{c=1}^C$, the semantic aware recurrent decoder produces K future semantic maps and flow fields $\{(m^t, f^t)\}_{t=T+1}^{T+K}$. We first describe a deterministic decoding procedure, for simplicity, which can then be easily adapted to a stochastic one to account for the randomness of the future.

Again, we consider decoders that learn the dynamics and predict the future in a semantic aware manner. Let ψ_c^{rnn} be the recurrent decoder for semantic class c, which generates the prediction for $\{(m_c^t, f_c^t)\}_{t=T+1}^{T+K}$ by recursively executing the following procedures:

$$h_c^t, e_c^t = \psi_c^{\text{rnn}}(h_c^{t-1}, \text{MLP}_c(H^T)), t \geq T+1 \tag{2}$$

$$\tilde{m}_c^t = \psi_{c,m}^{\text{rnn}}(e_c^t); \quad \tilde{f}_c^t = \psi_{c,f}^{\text{rnn}}(e_c^t) \tag{3}$$

Here we abuse the notation ψ_c^{rnn} to refer to the recurrent unit that updates the latent representation h_c^t, while generating a common embedding e_c^t, which is then decoded into the predicted semantic mask $\tilde{m}_c^t$ and flow fields $\tilde{f}_c^t$, respectively through separate decoding heads $\psi_{c,m}^{\text{rnn}}$ and $\psi_{c,f}^{\text{rnn}}$. This separate decoding design aligns with the practice that improves the decoding efficiency in multi-task learning. Note, we also apply a multi-layer perceptron MLP_c (due to its efficiency) on the collection of the hidden representations for all classes $H^T = \{h_c^T\}_{c=1}^C$, to ensure that the semantic aware decoder has access to the context provided by other classes within the scene (Fig. 3).

The decoders for each class $\{\psi_{c,m}^{\text{rnn}}, \psi_{c,f}^{\text{rnn}}\}_{c=1}^C$ can also be running in parallel, so that we have the semantic aware predictions for each class, i.e., $\{(\tilde{m}_c^t, \tilde{f}_c^t)\}$ with $t \in \{T+1, ..., T+K\}$ and $c \in \{1, ..., C\}$. Next, we apply late fusion to get predictions that can be directly compared to the ground-truth semantic maps and flow fields

[1] image regions that are observed/visible across multiple frames

$\{(m^t, f^t)\}_{t=T+1}^{T+K}$, and to further improve the contextual compatibility between different classes.

Context-aware late fusion. Given $\{(\tilde{m}_c^t, \tilde{f}_c^t)\}_{t=T+1}^{T+K}$ for each $c \in \{1, ..., C\}$, we apply a three-layer ConvNet to first fuse the binary semantic maps:

$$\tilde{m}^t = \psi^{\text{cnn}}(\text{concat}(\{\tilde{m}_c^t\}_{c=1}^{C})) \tag{4}$$

where the dimension of $\tilde{m}^t$ is $H \times W \times C$. We use softmax as the last layer for ψ^{cnn}, such that each slice of $\tilde{m}^t$ indexed by the last dimension, i.e., $\tilde{m}^t(c) = \tilde{m}^t[:, :, c]$ (in Python style), is still a scalar field indicating the probability of each pixel belonging to class c. And the fused flow field f^t is obtained as following:

$$\tilde{f}^t(i,j) = \sum_c \tilde{m}^t(i,j,c) \cdot \tilde{f}_c^t(i,j); \quad i \leq H, j \leq W \tag{5}$$

which is a linear combination of the flow vectors predicted by each semantic aware recurrent decoder, whose visibility comes from the fused semantic map $\tilde{m}^t$.

Training loss for $\phi_c^{\text{rnn}}, \text{MLP}_c, \psi_c^{\text{rnn}}, \psi_{c,m}^{\text{rnn}}, \psi_{c,f}^{\text{rnn}}, \psi^{\text{cnn}}$. With the ground-truth $\{(m^t, f^t)\}_{t=T+1}^{T+K}$, the training loss for flow fields is the $L1$ loss:

$$\mathcal{L}_f = \sum_{t=T+1}^{T+K} \|\tilde{f}^t - f^t\|_1 \tag{6}$$

which penalizes the discrepancy between the predicted flow and the ones computed from the ground-truth images. For the semantic maps we apply the cross entropy loss:

$$\mathcal{L}_m = \sum_{t=T+1}^{T+K} (1 + \alpha\, \mathcal{G} * \nabla m^t) \cdot \mathcal{H}(m^t, \tilde{m}^t) \tag{7}$$

Here $\mathcal{H}$ represents the cross-entropy, which is weighted by whether the pixel is near the boundaries between different classes or not. Note ∇ is the gradient operator, and we binarize its response to 0 and 1 to discount the artifacts caused by naming different classes with different integers. The binarized boundary map is then smoothed by a Gaussian kernel $\mathcal{G}$ to expand the weights to nearby pixels, making the boundaries thicker. The variance of the Gaussian, which determines the spatial extent of the boundaries is set to 9.0 and fixed. With this weighting scheme, the network will focus more on the pixels near the semantic boundaries, thus better preserves the shape of each semantic segment in the prediction. The relative importance between boundary and non-boundary pixels is controlled by the scalar α, which is set to 5.0 for all experiments.

So far, we have described the proposed semantic aware dynamic model in its deterministic mode. However, extending it to account for the stochasticity of the future is straightforward. For this purpose, we instantiate C semantic aware recurrent encoders θ_c^{rnn}, which operate in a similar way as the encoders for the past:

$$z_c^t = \theta_c^{\text{rnn}}([m_c^t, f_c^t], z_c^{t-1}); \quad z_c^t = [u_c^t, v_c^t], \ t \geq T+1 \tag{8}$$

The goal of the recurrent encoder θ_c^{rnn} is to generate a random variable z_c^t, represented by its mean and variance $[u_c^t, v_c^t]$ through reparameterization, whose initial value is set to $z_c^T = [h_c^T, I]$.[2] At the end of the recursion, we would like $z_c^{T+K} = [u_c^{T+K}, v_c^{T+K}]$ to be a zero-mean unit-variance Gaussian. Then, $Z^K = \{(u_c^{T+K}, v_c^{T+K})\}_{c=1}^{C}$ will be added to H^T in Eq. (2), also through reparameterization, for decoding the future with randomness. To learn θ_c^{rnn}'s, we add a KL-divergence term to the loss:

$$\mathcal{L}_{\text{kl}} = \mathbb{KL}(\mathcal{N}(u^{T+K}, v^{T+K}), \mathcal{N}(\mathbf{0}, \mathbf{I})) \tag{9}$$

where $\mathcal{N}$ represents the normal distribution. We summarize the training loss for the stochastic semantic aware dynamic model in the following:

$$\mathcal{L}_{\text{dynamic}} = \mathcal{L}_f + \mathcal{L}_m + \beta \mathcal{L}_{kl} \tag{10}$$

with β the weight on the KL-divergence term. As in VAEs, $\{\theta_c^{\text{rnn}}\}_{c=1}^{C}$ are used only during the training for the stochastic decoder, and will not be used during testing since the random noise can be directly sampled from the prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.2. Warping with Semantic Informed Dis-occlusion

We warp the past video frames to provide an anchor point for future synthesis using the predicted future semantic masks and flows fields $\{(\tilde{m}^t, \tilde{f}^t)\}_{t=T+1}^{T+K}$ from the semantic-aware dynamic model. To ease the warping and comply with the literature, here we mark the predicted flow $\tilde{f}^t$ as the backward flow, i.e., from $t+1$ to t. The warping can be simply performed via bilinear interpolation:

$$\hat{x}^{t+1}(p) = x^t(p + \tilde{f}^{t+1}(p)), p \notin \Omega_d \tag{11}$$

The key is to estimate the dis-occluded area Ω_d, which invalidates the assumption that a pixel in frame x^{t+1} is propagated from the previous frame x^t.

Note, [9] proposes to use pixel occupancy for dis-occlusion detection, however, miss-detection happens due to errors in flow prediction on the object boundaries where dis-occlusion resides (see Fig. 4). Given that semantic masks are easier to predict than flows, particularly, with our semantic-aware dynamic model, we propose a semantic consistency criteria for dis-occlusion estimation, i.e., $p \in \Omega_d$, if $\tilde{m}^{t+1}(p) \neq m^t(p)$. The above semantic consistency criteria can still correctly detect dis-occlusions even if the flow is wrong as shown in Fig. 4. In our experiments, we use both the pixel occupancy and the proposed semantic

[2] this ensures that the generation of the random variable is conditioned on the past.

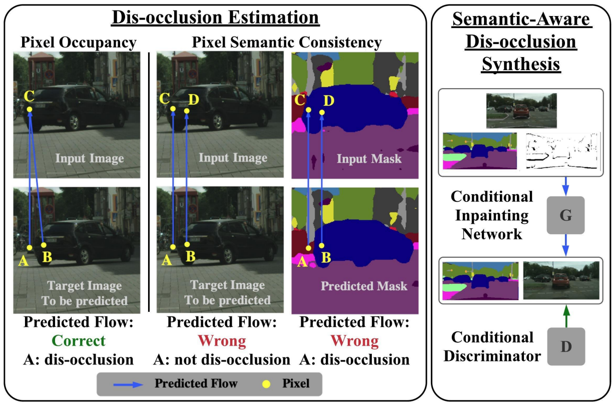


Figure 4. **Left**: Two criteria for dis-occlusion detection. *Pixel Occupancy*: pixels A, B in the target image domain are mapped onto pixel C in the input image domain, which is occupied by more than one pixel when the predicted (backward) flow (blue arrows) is correct; in this case, pixel A, as the cause of over-occupancy, can be detected as dis-occlusion. *Pixel Semantic Consistency*: if the predicted flow is incorrect, pixel occupancy fails in detecting A as dis-occlusion; however, given that semantic mask can be accurately predicted, A will be mapped to pixel C with inconsistent semantic labels, thus can be correctly detected as dis-occlusion. **Right**: Semantic-aware dis-occlusion synthesis, where both the generator and the discriminator take in predicted semantic mask, and the generator is dis-occlusion aware.

consistency criterion for dis-occlusion detection given their complementarity.

After warping, we end up with the future frames warped from the past and the corresponding dis-occlusion masks, i.e., $\{(\hat{x}^t, \Omega_d^t)\}_{t=T+1}^{T+K}$. Note $\hat{x}^t$ is only valid (up to noise) in the complement of Ω_d^t, which will be extrapolated as we describe next.

3.3. Semantic-Aware Dis-occlusion Synthesis

Using the warped frames $\{\hat{x}^t\}_{t=T+1}^{T+K}$ as the anchor, we employ a conditional inpainting network to further complete the dis-occluded parts and improve the quality of the synthesized images via adversarial training. The conditional inpainting network φ takes as input the anchor frame $\hat{x}^t$, and tries to complete the missing region indicated by Ω_d^t based on the predicted semantic map $\tilde{m}^t$ in a content-aware manner:

$$\tilde{x}^t = \varphi(\hat{x}^t, \Omega_d^t, \tilde{m}^t) \tag{12}$$

Note, image details and their temporal consistency can be improved by semantic maps informing the scene content as shown in [43], which only focuses on translating known semantic maps to images. Given the ability to model the dynamics of the semantic map and its prediction, our conditional inpainting network can be informed about the scene content, thus able to generate better synthesis (see Fig. 5).

To help the training of the content-aware conditional inpainting, we also employ two discriminators D_v, D_x for the video clip and frame respectively, with D_v focuses on the temporal continuity and D_x focuses on the image quality. So the training loss for the content-aware inpainting network is:

$$\begin{aligned}\mathcal{L}_\varphi = &\sum_{t=T+1}^{T+K} (1-\Omega_d^t)\cdot\|\tilde{x}^t - \hat{x}^t\|_1 + \lambda\mathcal{L}_{per}(\tilde{x}^t, x^t)\\ &+ \gamma\sum_{t=T+1}^{T+K} D_x(\tilde{x}^t, \tilde{m}^t) + \eta D_v(\{\tilde{x}^t\}, \{\tilde{m}^t\})\end{aligned} \tag{13}$$

where the first term measures the discrepancy between the completed image and the warped image in the co-visible area; the loss $\mathcal{L}_{per}$ measures the perceptual similarity between the generated images and the real images [27]. D_x and D_v measure the plausibility of images/videos conditioned on the semantic content. The final predictions of our model are $\{(\tilde{m}^t, \tilde{f}^t, \tilde{x}^t)\}_{t=T+1}^{T+K}$, i.e., predicted semantic maps, flow fields and video frames with the semantic-aware dynamics model as their driving force.

4. Experiments

Datasets. We evaluate our method on multiple *prediction* tasks, e.g., video frames and semantic maps, using three commonly used datasets, Cityscapes [7], KITTI Flow [11] and KITTI Raw [10]. Cityscapes [7] contains driving sequences recorded in 50 different cities. We use the training split for training our semantic-aware dynamic model and the validation set for evaluation. The training and evaluation subsets contain 2975 and 500 videos, respectively. Pixel-wise annotations for semantic segmentation are only available every 20 frames for the Cityscapes dataset. KITTI Raw [10] contains 156 long sequences. Following [48], we use 4 of them for testing, and the rest for training. KITTI Flow [11] is designed for benchmarking optical flow algorithms, and is more challenging than KITTI Raw [10]. It consists of 200 training videos and 200 test videos. Following [9], we downsample the videos to 128 × 424 and then center-crop to 128 × 256, yielding 4000 clips for both training and testing. Since per-frame dense semantic maps and optical flow annotations are not available, we leverage the off-the-shelf semantic segmentation network DeepLabV3 [6] to extrapolate annotations for 20 classes, and compute the optical flow using the PWC-Net [36].

Implementation and training. We adapt the grouped Conv-LSTM network [49] for the semantic-aware recurrent encoders and decoders to perform temporal aggregation of semantic maps and flow fields. The inpainting network is a modified U-Net [32], conditioned on predicted anchor frames and semantic maps. We also replace the convolutional layers in the encoder with the partial convolution proposed in [21], which masks dis-occluded area in the feature

Method	**MS-SSIM** (×1e-2) ↑ t+1	 t+5	**LPIPS** (×1e-2) ↓ t+1	 t+5
PredNet [23]	84.03	75.21	25.99	36.03
MCNET [39]	89.69	70.58	18.88	37.34
Voxel Flow [22]	83.85	71.11	17.37	28.79
Vid2vid [43]	88.16	75.13	10.58	20.14
Seg2vid [27]	88.32	61.63	9.69	25.99
FVS [48]	89.10	75.68	8.50	16.50
SADM	**95.99**	**83.51**	**7.67**	**14.93**

Table 1. Quantitative comparison on the Cityscapes dataset.

space at different resolutions. The video and image discriminators are similar to those in CycleGAN [58], except that the video discriminator has ordinary 2D convolutions replaced with 3D convolutions.

Though our model can be trained end-to-end, we split the training into two stages to manage on a single work-station: 1) Training the semantic-aware dynamic model with loss (10); 2) Training the inpainting network to fill-in the disoccluded area with loss (13). The training is performed on 2 GeForce GTX 1080 Ti GPUs with batch size equals to 6, and each video clip in the batch contains 10 frames. Learning rates for both stages start from 0.001 and decay by 0.8 every 20 epochs. The weights of each term in the losses are detailed in the supplementary. The training of the semantic aware dynamic model needs 40 hours to converge, and the training of the inpainting network takes about 20 hours.

Model complexity and inference. The semantic-aware dynamic model for predicting semantic maps and flow fields contains 8.6M parameters. The inpainting network contains 5.18M parameters. Inference can be performed on a single GeForce GTX 1080 GPU with 8GB memory. The inference runs at 19 frames per second.

4.1. Quantitative Results

Video prediction. Table 1 and 2 report the multi-frame video prediction performance evaluated in terms of Multi-scale Structural Similarity Index Measure (MS-SSIM) [46] and LPIPS [56], on the Cityscapes and the KITTI Raw datasets respectively. Higher MS-SSIM scores and lower LPIPS distances suggest better performance. Specifically, on longer horizon prediction ($t+5$), our model improves Seg2vid [27], which also employs semantic segmentation, by 35.50% (MS-SSIM) and 20.85% (LPIPS). Moreover, our model outperforms FVS [48], which infers 2D affine transformations of moving objects, by 10.35% (MS-SSIM) and 9.39% (LPIPS) on the $t+5$ predictions.

Method	**MS-SSIM** (×1e-2) ↑ t+1	 t+3	 t+5	**LPIPS** (×1e-2) ↓ t+1	 t+3	 t+5
PredNet [23]	56.26	51.47	47.56	55.35	58.66	62.95
MCNet [39]	75.35	63.52	55.48	24.05	31.71	37.39
Voxel Flow [22]	53.93	46.99	42.62	32.47	37.43	41.59
FVS [48]	79.28	67.65	60.77	18.48	24.61	**30.49**
SADM	**83.06**	**72.44**	**64.72**	**14.41**	**24.58**	31.16

Table 2. Quantitative comparison on the KITTI Raw dataset.

Method	**PSNR**↑	**SSIM** (×1e-2)↑	**LPIPS** (×1e-2)↓
Repeat [9]	16.5	48.9	19.0
PredNet [23]	17.0	52.7	26.3
SVP-LP [8]	18.5	56.4	20.2
MCNet [39]	18.9	58.7	23.7
MoCoGAN [38]	19.2	57.2	18.6
DVF [22]	22.1	68.3	16.3
CtrlGen [12]	21.8	67.8	17.9
DPG [9]	*22.3*	*69.6*	*11.4*
SADM	**24.47**	**71.1**	**10.9**

Table 3. Quantitative comparison in next-frame prediction on the KITTI Flow dataset.

Method	t+1	t+5	t+9
Repeat	67.1	52.1	38.3
S2S-dil [24]	-	59.4	47.8
PSPNet [35]	71.3	60	-
Jin [16]	66.1	-	-
Terwilliger [37]	73.2	67.1	51.5
Bayes-WD-SL [2]	**74.1**	65.1	51.2
F2MF-DN121 [33]	-	69.6	57.9
SADM	73.8	**70.3**	**60.1**

Table 4. Quantitative results of semantic map prediction on the Cityscapes dataset measured by the mIoU score.

Following DPG [9], we report the next-frame prediction results on the KITTI Flow dataset in Table 3. For a fair comparison, we also include the commonly used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [45] as the evaluation metrics, in addition to Learned Perceptual Image Patch Similarity (LPIPS) [56]. Our model improves DPG [9], the previous state-of-the-art method evaluated on this dataset, by 9.7% in terms of PSNR, 2.2% in terms of SSIM, and 4.4% in terms of LPIPS.

Semantic segmentation mask prediction. We evaluate our model's performance on semantic mask prediction using the Cityscapes dataset, with the standard mean Intersection-over-Union score (mIoU) as the evaluation metric. Following [24], the scores are computed with respect to the ground-truth segmentation of the 20th frame in each sequence. Table 4 shows the semantic mask prediction performance on multiple prediction lengths. Our method performs the best among the other methods, especially when the prediction horizon gets longer.

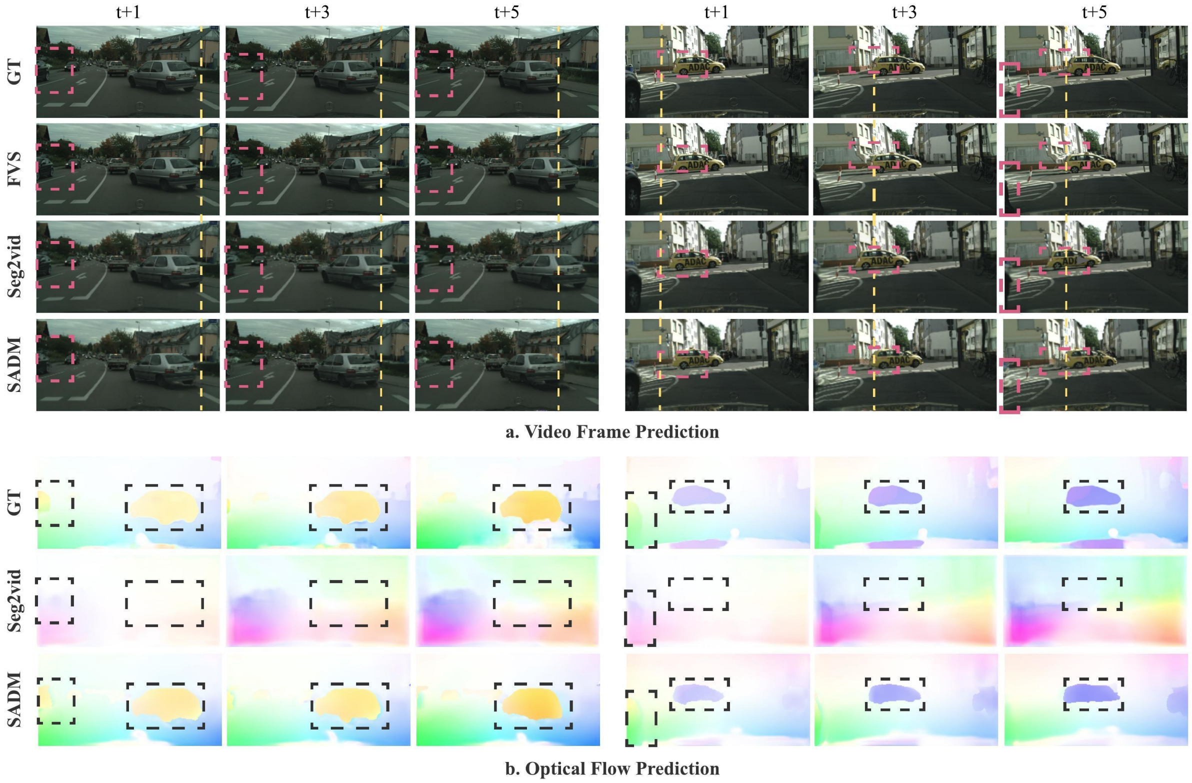


Figure 5. Visual comparison on the Cityscapes dataset. Both the predicted video frames (a) and flow fields (b) are presented. Left: the flow predicted by our network clearly shows the silver car moves to the left and the camera moves forward, while the flow predicted from Seg2vid [27] is dominated by "major" camera motion exhibited in the dataset, i.e., zooming-in caused by the movement of the running car. FVS [48] wrongly predicts the motion of the silver car, resulting in incorrect car locations at $t + 3$ and $t + 5$. Note that, at $t + 5$, the black car on the left should move outside the image domain, which is only captured by our model. Again, the "ghost effect" presents near the objects' boundaries in the predictions from the other two methods. Right: without conditioning on semantic segmentation masks, the dis-occluded area of the building is incorrectly inpainted by the inpainting network as part of the moving car, causing distortions in the results from FVS.

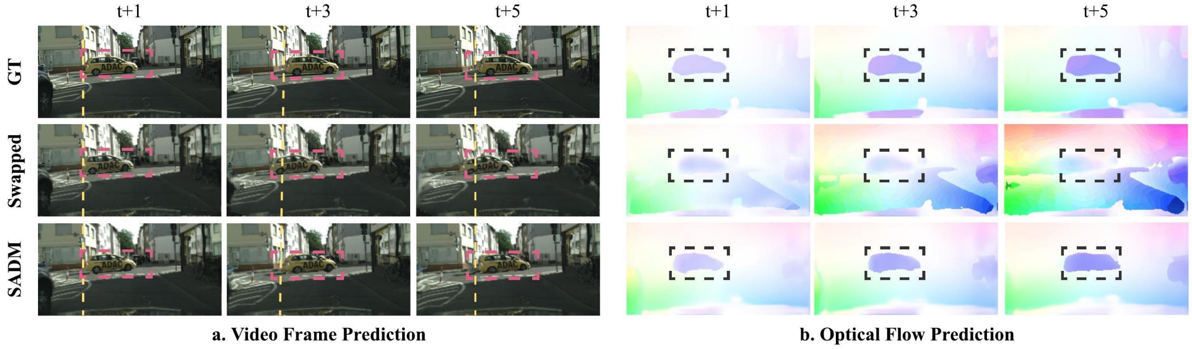


Figure 6. Ablation study: our model predicts video frames and flow fields by swapping semantic-aware encoders for the classes of "car" and "road", verifying that semantic-aware dynamics is learned with the proposed model.

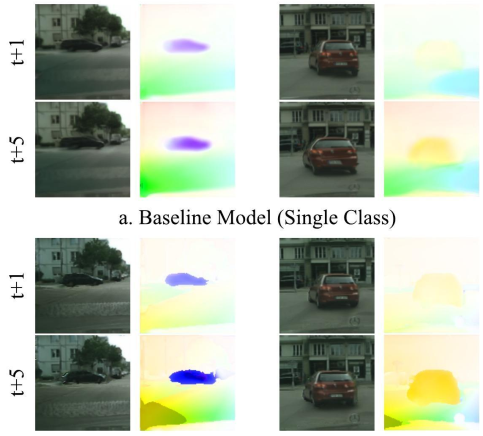
4.2. Qualitative Results

In Fig. 5, we compare to FVS [48] and Seg2vid [27], two most recent methods that employ semantic segmentation or moving object segmentation to facilitate video prediction. The motivation of Seg2vid [27] is that the high-level semantics of the scene will result in more accurate predictions. However, without an explicit modeling, such as SADM, predicted flow fields from [27] still suffer from over-smoothing. Moreover, given that most of the videos in Cityscapes are captured by a camera moving forward with a car, there is a strong tendency in the model from Seg2vid to produce flow fields showing zooming-in motion. On the other hand, FVS [48] separates the whole scene into moving and non-moving segments by moving object detection. Although 2D affine transformations are predicted per frame to approximate the object motion, complex motion, including deformation and 3D rotation, may not be captured by a single 2D affine transformation. Even for non-moving rigid objects whose motion is induced by the camera motion, e.g., parked cars and buildings, their projected 2D flow fields still depend on the 3D geometries and thus are not 2D affine. With semantic-aware dynamics and inpainting, our model can generate high-quality video frames with more accurate motions.

4.3. Ablation Study

To demonstrate the effectiveness of decomposing the video into semantically consistent regions for video prediction, we train a baseline model ("single class" in Fig. 7) with the same network architecture as SADM, and with a naive concatenation of semantic masks and flow fields as the input to the baseline model. The encoder and decoder of this baseline model share similar structures as those in SADM, besides that ordinary convolutions are used. Note that an ordinary convolution layer has more parameters than a grouped convolutional layer ($O(K^2)$ v.s. $O(K)$). As shown in Fig. 7, without explicitly modeling the class-wise dynamics, the baseline model trained to predict flow fields or video frames has difficulties estimating the motion near the boundaries between different semantic regions (the black car in the top left). There is a heavy over-smoothing near the boundary of the car, which is problematic for the consecutive warping procedure, since the flow there will warp pixels on the car to the background or vice versa, generating the "ghost effect". With the proposed semantic aware dynamic model, the motion of either the car or the background can be accurately estimated since the influence on motion estimation from occlusions is automatically handled through the decomposition. Similarly, in the bottom left of Fig. 7, the warped image using the flow predicted by the baseline model shows far more artifacts on the red car.

To show that the learned dynamics from our model is indeed semantic-aware, we test the model with semantic labels intentionally swapped in the input. For example, we input the "car" segments to the semantic-aware encoder that learns the dynamics of the "road" class, and vice versa. As expected (Fig. 6), the predicted motion of the "car" segments using the encoder for the "road" class (middle row) now looks like the one of the "road" (bottom row). Similarly, the predicted motion of the "road" segments using the encoder for the "car" class (middle row) now looks more like the one from the "car" (bottom row). This shows exactly that semantic-aware dynamics is captured by the proposed model.



a. Baseline Model (Single Class)

b. SADM

Figure 7. A baseline model without explicit modeling of the semantic-aware dynamics (single class) shows less accurate motion prediction than SADM and has more artifacts in the predicted video frames.

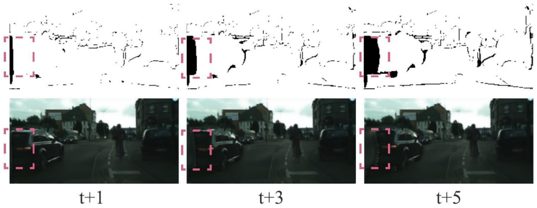


Figure 8. Complicated dis-occluded regions cause difficulties for the inpainting network, even if the estimated occlusions are accurate (top row).

5. Discussion

We have tested the hypothesis that representing object-level motion in a video can be beneficial for prediction. To that end, we have proposed a model that captures occlusions explicitly, and represents class-specific motion. While such high-level modeling is beneficial to prediction, there are failure cases. Specifically, hallucinating the dis-occluded regions can lead to failure when the background is complex (Fig. 8). As the time horizon grows, the prediction becomes increasingly unrealistic, as with other video prediction models, but the explicit modeling of objects and class-specific motion yields improvements over generic models. Also, we are constrained by classes for which we have training data, which limits generalization. So, our work is only a first step to incorporate dynamical models that are informed by the semantics of objects in the scene, which we expect will ultimately facilitate intelligent interaction with physical scenes by autonomous agents.

Acknowledgment

Research supported by ARL W911NF-20-1-0158 and ONR N00014-17-1-2072.

References

[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

[2] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. *arXiv preprint arXiv:1810.00746*, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 753–769, 2018.

[5] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018.

[9] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9006–9015, 2019.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

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

[12] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018.

[13] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018.

[14] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526, 2018.

[15] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[16] 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.

[17] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[19] 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.

[20] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1744–1752, 2017.

[21] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[22] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017.

[23] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[24] 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.

[25] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2203–2212, 2017.

[26] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1434, 2017.

[27] 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.

[28] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmío. Video inpainting under constrained camera motion. *IEEE Transactions on Image Processing*, 16(2):545–553, 2007.

[29] Silvia L Pintea, Jan C van Gemert, and Arnold WM Smeulders. Déja vu. In *European Conference on Computer Vision*, pages 172–187. Springer, 2014.

[30] 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.

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Josip S. et al. Warp to the future. In *CVPR*, 2020.

[34] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2839, 2017.

[35] Seyed shahabeddin Nabavi, Mrigank Rochan, and Yang Wang. Future semantic segmentation with convolutional lstm. In *BMVC*, page 137, 2018.

[36] 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.

[37] 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.

[38] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[39] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

[40] 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.

[41] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[42] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3332–3341, 2017.

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

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

[46] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[47] Nevan Wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. *arXiv preprint arXiv:1806.04768*, 2018.

[48] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5539–5548, 2020.

[49] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[50] Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, and Xiaokang Yang. Structure preserving video prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1460–1469, 2018.

[51] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. *arXiv preprint arXiv:1905.02884*, 2019.

[52] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in neural information processing systems*, pages 91–99, 2016.

[53] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[54] Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to manipulate individual objects in an image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2020.

[55] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019.

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of

deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[57] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 387–403, 2018.

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.