Video Generation from Single Semantic Label Map

Junting Pan[1,4], Chengyu Wang[1], Xu Jia[2], Jing Shao[1], Lu Sheng[3,4]*, Junjie Yan[1], and Xiaogang Wang[1,4]
Sensetime Research[1], Huawei Noah's Ark Lab[2], College of Software, Beihang University[3],
CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong[4]

Abstract

This paper proposes the novel task of video generation conditioned on a SINGLE semantic label map, which provides a good balance between flexibility and quality in the generation process. Different from typical end-to-end approaches, which model both scene content and dynamics in a single step, we propose to decompose this difficult task into two sub-problems. As current image generation methods do better than video generation in terms of detail, we synthesize high quality content by only generating the first frame. Then we animate the scene based on its semantic meaning to obtain temporally coherent video, giving us excellent results overall. We employ a cVAE for predicting optical flow as a beneficial intermediate step to generate a video sequence conditioned on the initial single frame. A semantic label map is integrated into the flow prediction module to achieve major improvements in the image-to-video generation process. Extensive experiments on the Cityscapes dataset show that our method outperforms all competing methods. The source code will be released on https://github.com/junting/seg2vid.

1. Introduction

A typical visual scene is composed of foreground objects and the background. In a dynamic scene, motion of the background is determined by camera movement which is independent of the motion of foreground objects. Scene understanding, which include both understanding how foreground objects and background look and how they change, is essential to advancing the development of computer vision. Scene understanding, besides using recognition models, can be accomplished by generative methods[34]. In this work we focus on using generative models to understand our visual world.

There has been much progress in image generation to address static scene modeling. Researchers have proposed methods to generate images from only noise [10] or from pre-defined conditions such as attribute, text and pose [41, 20]. In recent works, people also pay attention to image generation conditioned on semantic information with either paired [12] or unpaired data [42]. The conditional image generation methods provide a way to manipulate existing images and have potential value as a data augmentation strategy to assist other computer vision tasks. While image generation tasks only model static scenes, for video prediction, it is essential to also investigate the temporal dynamics. Models are trained to predict raw pixels of the future frame by learning from historical motion patterns. There is another line of work on video synthesis without any history frames.

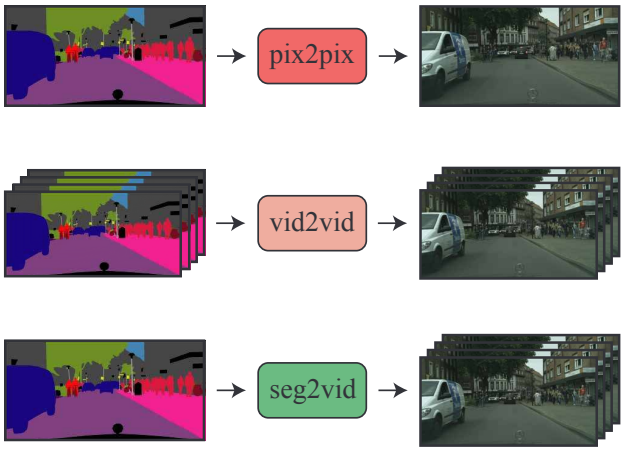

Figure 1: Comparison with existing generation tasks. From top: Image-to-image translation, video-to-video, and our image-to-video synthesis. Our method only takes *one* semantic label map as input and synthesizes a sequence of photo-realistic video frames.

Similar to research on image generation, some work investigated unconditional video generation. That is, directly generating video clips from noise by using generative adversarial networks to learn a mapping between spatial-temporal latent space and video clips [31, 25]. Another group of researchers worked on video-to-video translation [37], where a sequence of frames are generated according to a sequence of aligned semantic representations.

*Lu Sheng is the corresponding author.

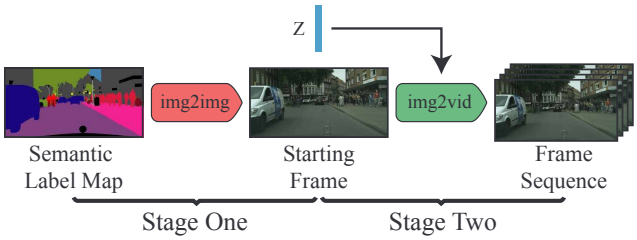

Figure 2: Overview of our two step generation network. In the first stage, we generate the starting frame by mapping from a semantic label map. In the second stage, we use our flow prediction network to transform the initial frame to a video sequence.

In this work, we study video generation with a setting similar to the video-to-video work [37] except that it is only conditioned on a single frame's semantic label map. Compared to previous works on video generation, our setting not only provides control over the generation process but also allows high variability in the results. Conditioning the generation on semantic label map helps avoid producing undesirable results (*e.g.* a car driving on the pavement) which often occurs in unconditional generation. Furthermore, we can generate cars moving at different speeds or in different directions, which is not possible in the video-to-video setting. One intuitive idea to address this new task would be to train an end-to-end conditional generative model. However, it is not easy to apply such a model to datasets composed of diverse objects and background, *i.e.* different objects in different scenes have different motions. In reality, training a single end-to-end model to simultaneously model both appearance and motion of these objects and scenes is very hard. Therefore, as illustrated in Fig. 2, we take a divide-and-conquer strategy, designed to model appearance and motion in a progressive manner.

In the first stage, we aim to transform a semantic label map to a frame such that the appearance of scene is synthesized, which falls into the category of image-to-image translation. During translation process, the model only focuses on producing an image of good quality with reasonable content.

In the next stage, future motion of the scene is predicted based on the generated frame. Specifically, a conditional VAE is employed to model uncertainty of future motion. Different from existing video prediction tasks where motion information can be estimated from historical frames, in our setting, we only have one semantic label map and one generated frame available. We argue that it is important for the model to leverage the semantic meaning of the first frame when predicting motion. For example, buildings and pedestrians have very distinctive motion. We take both the semantic label map and the generated frame as input and feed them into a motion prediction model. Empirical results demonstrate that with semantic representation as input, the model can learn better motion for dynamic objects than without that, specially for complex scenes with multiple classes of objects. We model motion with optical flow. Once flows are predicted, they are directly applied to warp the first frame to synthesize future frames. Finally, a post-processing network is added to rectify imperfection caused during the warping operation. Inspired by[21], we further improve the performance of flow prediction and future frame generation using bidirectional flows and geometric consistency. Experimental results demonstrate the effectiveness of the proposed method in video generation.

Our contributions are the following.

1. We introduce the novel task of conditioning video generation on a single semantic label map, allowing a good balance between flexibility and quality compared to existing video generation approaches.

2. The difficult task is divided into two sub-problems, *i.e.*, image generation followed by image-to-sequence generation, such that each stage can specialize on one problem.

3. We make full use of the semantic categorical prior in motion prediction when only one starting frame is available. It helps predict more accurate optical flow, thereby producing better future frames.

2. Related Work

Image generation Many work exists regarding image generation which generally can be classified into two categories, unconditional generation and conditional generation. In unconditional generation, some work extends GANs [10] or VAE [16] to map from noise to real data distribution. Auto-regressive architectures model the image on a per-pixel basis [32, 22]. In the second category, conditional models generate images given either class category, textual descriptions, scene graphs or images [20, 2, 41, 15, 26]. Especially for image translation task, researchers study how to generate a meaningful image from a semantic representation such as semantic label maps (paired and unpaired) ([12, 42, 38, 3, 26]). However, in image generation tasks, photo-realism of the scene is modeled without considering their motion information.

Video Generation Similar to Image generation, video generation can also be divided into two categories: conditional and unconditional. For the former category, VideoGAN [34] explicitly disentangles a scene's foreground from background under the assumption that the background is stationary. The model is limited to only simple cases and cannot handle scenes with a moving background due to camera movement. TGAN [25] first generates a sequence of latent variables and then synthesize a sequence of frames

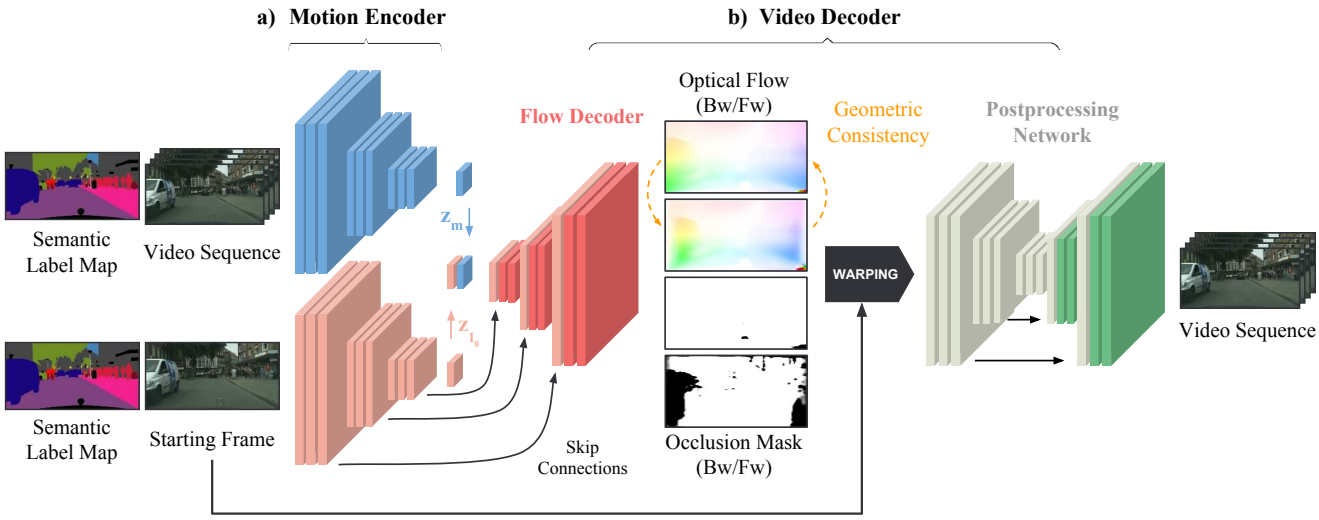

Figure 3: Overall architecture of the proposed image-to-video generation network. It consists two components: a) Motion Encoder and b) Video Decoder. For any pair of bidirectional flow predictions, consistency check is computed only in non occluded areas.

based on those latent variables. MoCoGAN [31] also tries to map a sequence of random vectors to a sequence of frames. However, their framework decomposes video into content subspace and motion subspace, making video generation process more controllable. For conditional video generation, it is still at its early stage. One recent work is vid2vid [37] in which authors aim at transforming a sequence of semantic representation, e.g. semantic label map and sketch map, to a sequence of video frames. Our work falls into the category of conditional video generation, but unlike vid2vid, our method only requires a single semantic label map as input which enables more freedom over the generation process.

Video prediction Some work model future motion in a deterministic manner. In [23, 29, 33], future prediction is carried out in a latent space, and the representation of future frames is projected back to image domain. These models are directly trained to optimize a reconstruction loss, such as Mean Squared Error (MSE), between the predicted frames and ground truth frames. However, they are prone to converging to blurry results as they compute an average of all possible future outcomes for the same starting frame. In [19, 13, 8], future motion is predicted using either optical flow or filter, where estimation and then corresponding spatial transformation is applied to history frames to produce future frames. The result is sharp but lacks diversity. A group of researchers [39, 36, 7, 1] introduced conditional variational autoencoders for video prediction to model uncertainty in future motion allowing the results to be both sharp and diverse. Similar to our work, Walker et al. [35] and Li et al. [18] attempt to predict multiple future frames from a static image. In the training phase, they take the ground truth optical flow, either human annotated or computed, as supervision to predict such flow, and transform the given frame to future frames. Contrary to Walker et al. [35] and Li et al. [18], we learn optical flow in an unsupervised manner, *i.e.*, without taking any pre-computed flow as supervision.

3. Semantic Label Map to Video Generation

Generating a video sequence $V = \{I_0, I_1, ..., I_T\}$ from a single semantic label map S allows more flexibility compared to translating multiple label maps to a video, but is also more challenging. In this work we propose to divide such a difficult task into two relatively easy sub-problems and address each one separately, i.e., i) *Image-to-Image* (I2I): an image generation model based on conditional GANs [38] that maps a given semantic label map S to the starting frame $\hat{I}_0$ of a sequence, and ii) *Image-to-Video* (I2V): an image-sequence generation network that produces a sequence of frames $\hat{V} = \{\hat{I}_0, \hat{I}_1, ..., \hat{I}_T\}$ based on the generated starting frame $\hat{I}_0$ and a latent variable z. In each stage we have a model specializing on the corresponding task such that the overall performance is good.

3.1. Image-to-Image (I2I)

Image-to-image translation aims at learning the mapping of an image in the source domain to its corresponding image in the target domain. Among the existing methods [12, 42, 38, 3], we adopt the state-of-the-art image translation model pix2pixHD [38] to generate an image from a semantic label map. It includes a coarse-to-fine architecture to progressively produce high quality images with fine

details while keeping global consistency. Note that the translation stage is not restricted to this method and other image translation approaches can substitute pix2pixHD.

3.2. Image-to-Video (I2V)

In this section, we present how to use cVAE for image sequence generation conditioned on an initial frame obtained from Sec. 3.1. It is composed of two sub-modules, i.e., flow prediction and video frame generation from flow. Fig. 3 shows the network structure and the components of the proposed Image-to-Video model.

Conditional VAE - Compared to future prediction from *multiple* frames, where the future motion can be estimated based on past sequence, motion predicted from *one* single frame can be more diverse. We employ the conditional VAE (cVAE) model [39] as the backbone to capture multiple possible future motions conditioned on a static image. The proposed cVAE is composed of an encoder and a decoder. The encoder $Q(z|V, I_0)$ learns to map a starting frame I_0 and the subsequent frames $V = \{I_1, ..., I_T\}$ into a latent variable z that carries information about motion distribution conditioned on the first frame I_0. To achieve such mapping, the latent variable z is composed of two parts, one projecting from the whole sequence including both I_0 and V, and the other from only the initial frame I_0. The decoder $P(V|z, I_0)$ then reconstructs the sequence and outputs $\hat{V}$ based on a sampled z and I_0. During training, the encoder $Q(z|V, I_0)$ learns to match the standard normal distribution, $N(0, I)$. When running inference, the cVAE will generate a video sequence from a given starting frame I_0 and a latent variable z sampled from $N(0, I)$ without the need of the motion encoder.

Flow Prediction - We first use an image encoder to transform the starting frame into a latent vector z_{I_0} as a part of the latent variable z. The whole sequence is sent to another sequence encoder to compute z_m, which makes up the other part of z for uncertainty modeling. z_{I_0} and z_m are concatenated as one vector z which is fed to a decoder to compute future optical flow. For motion generation, we predict bidirectional flows, *i.e.* both forward flow from the initial frame to future frames and backward flow from future frames to the initial frame. Computing cycle flow allows us to perform forward-backward consistency checks. For regions which appear in both frames (A and B), correspondence between two frames can be captured both from A to B and from B to A. We compute an occlusion mask to omit regions which are either occluded or missing in the generated frame so that the consistency check is only conducted on non-occluded regions. Putting all this together, the resulting output of the cVAE is the optical flow as well as the occlusion mask for both forward and backward directions, defined as:

$$W^f, W^b, O^f, O^b = \mathcal{F}(I_0), \tag{1}$$

Where $\mathcal{F}$ is the flow prediction module that is composed of the motion encoder and the flow decoder as shown in Fig 3. $W^f = \{\mathbf{w}_1^f, ..., \mathbf{w}_T^f\}$, where $\mathbf{w}_t^f = (u^f, v^f)$ is the forward optical flow from I_0 to I_t and $W^b = \{\mathbf{w}_1^b, ..., \mathbf{w}_T^b\}$, with $\mathbf{w}_t^b = (u^b, v^b)$ is the backward optical flow. $O^f = \{o_1^f, ..., o_T^f\}$ and $O^b = \{o_1^b, ..., o_T^b\}$ are the multi-frame forward-backward occlusion maps. We define a pixel value in the occlusion map to be zero when there is no correspondence between frames. All optical flows and occlusion maps are jointly predicted by our image-to-flow module. Note that both bidirectional and occlusion maps are learned without any pre-computed flow as supervision.

Video frame Generation - With the predicted optical flow, we can directly produce future frames by warping the initial frame. However, the generated frames obtained solely by warping has inherent flaws, as some parts of the objects may not be visible in one frame but appears in another. To fill in the holes caused by either occlusion or objects entering or leaving the scene, we propose to add a post-processing network after frame warping. It takes a warped frame and its corresponding occlusion mask O^b as the input, and generates the refined frame. The final output of our model is defined as follows:

$$\hat{I}_t(\mathbf{x}) = \mathcal{P}(o_t^b(\mathbf{x}) \cdot I_0(\mathbf{x} + \mathbf{w}_t^b(\mathbf{x}))), \tag{2}$$

where $\mathcal{P}$ is the post-processing network and $\mathbf{x}$ denotes the coordinates of a position in the frame.

Loss Function - Our loss function contains both per-pixel reconstruction and uncertainty modeling. For the per-pixel reconstruction, we compute losses in both the forward and backward direction, formulated as

$$\begin{aligned}\mathcal{L}_r(W^f, W^b, V) = \sum_t^T \sum_{\mathbf{x}} o_t^f(\mathbf{x}) |I_0(\mathbf{x}) - I_t(\mathbf{x} + \mathbf{w}_t^f(\mathbf{x}))|_1 \\ + o_t^b(\mathbf{x}) |I_t(\mathbf{x}) - I_0(\mathbf{x} + \mathbf{w}_t^b(\mathbf{x}))|_1,\end{aligned} \tag{3}$$

where T is the length of the generated sequence. We only compute reconstruction in non-occluded regions to avoid learning incorrect deformations. Neighboring pixels usually belong to the same object, thus they tend to have similar displacement. Therefore, similar to previous work [40, 30] we also add a smoothness constraint to encourage flow in a local neighborhood to be similar.

$$\mathcal{L}_{fs}(W^f, W^b) = |\nabla W^f|_1 + |\nabla W^b|_1 \tag{4}$$

We compute forward-backward consistency loss for non-occluded regions:

$$\begin{aligned}\mathcal{L}_{fc}(W^f, W^b) = \sum_t^T \sum_{\mathbf{x}} o_t^f(\mathbf{x}) |\mathbf{w}_t^f(\mathbf{x}) - \mathbf{w}_t^b(\mathbf{x} + \mathbf{w}_t^f(\mathbf{x}))|_1 \\ + o_t^b(\mathbf{x}) |\mathbf{w}_t^b(\mathbf{x}) - \mathbf{w}_t^f(\mathbf{x} + \mathbf{w}_t^b(\mathbf{x}))|_1,\end{aligned} \tag{5}$$

To train the in-painting network, we applied an $L1$ loss together with a perceptual loss [14] that has been shown to be useful for image generation. Therefore, our data loss can be formulated as a weighed sum of the above terms.

$$\begin{aligned}\mathcal{L}_{data}(\hat{V}, V) = \lambda_r \mathcal{L}_r + \lambda_{fs} \mathcal{L}_{fs} + \lambda_{fc} \mathcal{L}_{fc} \\ + \mathcal{L}_{l1}(\hat{V}, V) + \mathcal{L}_{l1}(\phi(\hat{V}), \phi(V)) \\ + \lambda_p |1 - O^b|_1 + \lambda_p |1 - O^f|_1,\end{aligned} \tag{6}$$

where ϕ is VGG-19 [27] from where we extract and collect features from the first 16 layers. We add a penalty on the occlusion maps for $\lambda_p = 0.1$ to avoid the trivial solution where all pixels become occluded (we define the value in a position of O^b to be 0 when the pixels is becoming occluded in the next frame). The weights are set to be: $\lambda_r = \lambda_{fs} = \lambda_{fc} = \lambda_{l1} = 1$ and $\beta = 0.1$. To model the motion uncertainty we incorporate the KL-divergence loss such that $Q(z|X)$ matches $N(0, I)$. The training loss for the cVAE is a data loss combined with a KL-divergence loss.

$$\mathcal{L}_{cVAE}(\hat{V}, V) = \mathcal{L}_{data} + \beta \mathcal{D}_{kl}(p_\phi(z|V)||p(z)). \tag{7}$$

3.3. Flow prediction with semantic label maps

Different from video prediction conditioned on multiple frames, generating a video from a static frame has no access to historical motion information. To infer future motion of a object in a static frame, the model needs to understand the semantic category of that object and its interaction with other objects and background. For example, the car will stop when the traffic light is red and move on when is green. To promote future motion estimation for the whole frame, we incorporate semantic label map which describes semantic information of the whole scene into the flow prediction module discussed in previous sub-section.

We explore two ways of integrating the semantic label map for flow prediction. In the first method, we expand a semantic label map into several heatmaps which is filled with ones on positions correspond to a semantic category and zeros elsewhere. These heatmaps are concatenated with the generated starting frame and fed to the cVAE model for future frame synthesis. In the other method, we further divide the heatmaps into two sets, *i.e.*, foreground heatmaps and background heatmaps, as shown in Fig. 4. Each set of heatmaps is fed to a separate sequence encoder to get a latent vector z_{FG} and z_{BG}. They are then concatenated with z_{I_0} becoming the input to the flow decoder. In Section 4, experimental results demonstrate that integrating semantic label map helps computing more accurate flow and accordingly improve the video generation performance.

4. Experiments

In this section we present the dataset and describe the details about the implementation. We evaluate our method against several baseline methods with both qualitative and quantitative metrics. We also perform ablation studies to confirm the effectiveness of using semantic label maps for video generation.

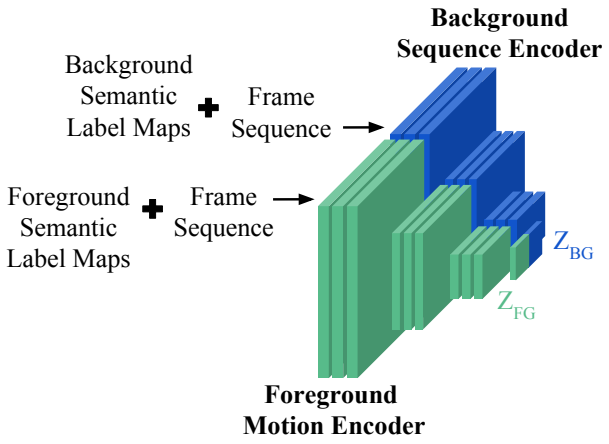

Figure 4: Semantic sequence encoder. Each sequence encoder only focuses on learning either foreground or background motion.

4.1. Datasets and Evaluation Metrics

Datasets We have conducted experiments on the Cityscapes dataset while we have provided qualitative results on the many other datasets. **Cityscapes** [6] consists of urban scene videos recorded from a car driving on the street. It contains 2,975 training, 500 validation and 1,525 test video sequences, each containing 30 frames. The ground truth semantic segmentation mask is only available for the 20th frame of every video. We use DeepLabV3[5] to compute semantic segmentation maps for all frames, which are used for training and testing. We train the model using all videos from the training set, and test it on the validation set. **UCF101** [28] The dataset contains $13,220$ videos of 101 action classes. **KTH Action dataset** [17] consists of 600 videos of people performing one of the six actions(walking, jogging, running, boxing, handwaving, hand-clapping). **KITTI** [9] similar to Cityscpes was recorded from a car traversing streets.

Evaluation Metrics We provide both quantitative and qualitative evaluation results in this section. For qualitative evaluation, we conducted a human subjective study to evaluate our method as well as the baseline methods. We randomly generated 100 video sequences for each method, pairing each generated video with the result of another randomly chosen method. The participants are asked to choose from each pair the most realistic looking video. We calculate the human preference score after each pair of videos was evaluated by 10 participants.

The Fréchet Inception Distance (FID) [11] measures the similarity between two sets of images. It was shown to correlate well with human judgment of visual quality and

	MoCoGAN	FG	vid2vid	Ours
FID	8.77	3.69	4.86	**3.52**

Table 1: Comparison of video generation methods where the input is a single semantic label map.

is most often used to evaluate the quality of samples from GANs. FID is calculated by computing the Fréchet distance between two feature representations of the Inception network. Similar to [37], we use the video inception network [4] to extract spatio-temporal feature representations.

4.2. Implementation details

Our method takes a single semantic label map S and predict $T = 8$ frames in a single step. We resize all frames to 128×128 and extract the semantic segmentation maps with DeepLabV3 [5] for training. We do not use any flow map as ground truth for training. In the cVAE, the motion encoder is built upon stacks of 2D convolutional layers intercepted with max pooling layers. The latent vector z has dimension 1024, 896 for foreground motion and 128 for background motion. For the flow encoder, we use three blocks each consisting of 3D convolutional layers intercepted with bilinear upsampling layer that progressively recovers the input resolution in both spatial and temporal dimensions. For the postprocessing network, we adopt the U-Net architecture from [24].

4.3. Ablation Studies

We conduct extensive experiments on the Cityscapes dataset to analyze the contribution of the semantic label map and optical flow for motion prediction. We have shown that optical flow is reliable motion representation to convey motion between frames and preserver better visual quality. Fig. 9 shows that the model without optical flow produces blurry frames. In contrast, our flow based solution preserves better details even on fast moving objects and produces fewer artifacts.

We also compare frame sequences generated by the model without semantic label map and two ways of integrating that. As shown in Fig. 10, the model integrating semantic label map is able to capture both foreground object motion and background motion, whereas the one without that fails to estimate the independent foreground object motion. By further separating semantic label maps into background and foreground, it can capture more details in structure marked by the red rectangles. As expected, semantic information plays an important role in generating object motion when predicting from a single frame. We show further improvements by separating semantic classes into two groups based on background and foreground.

Figure 5: Comparison between different approaches of video prediction from a static image. Top left: ground truth. Top right: FG. Bottom left: MoCoGAN. Bottom right: img2vid (ours). Our method preserve the the visual quality while other method rapidly degrades.

	MoCoGAN	FG	Ours
FID	7.06	2.86	**1.80**

Table 2: Comparison of video prediction methods that take a single starting frame as input.

4.4. Baselines

We compare our network with five state-of-the-art baseline methods trained on the Cityscapes dataset.

MoCoGAN [31] is an unconditional video generation model. Here, we also compared the conditional setting of MoCoGAN, given the initial frame x_0 as input.

FlowGrounded (FG) [18] is a video prediction model from a static image. We compare our image-to-video stage with this method on both video generation and video prediction tasks.

Vid2Vid [37], the goal of vid2vid is to map a sequence of semantic represenation to a sequence of video frames, where future motion is approximately given in the semantic segmentation sequence. We evaluate vid2vid to see whether our method is comparable to this "upper bound".

4.5. Results

Quantitative Results In Table 1 we report the results on the Cityscapes dataset. In terms of performance, the lower the FID, the better the model. In Table 1, we show that our method has the lowest FID compared to all competing methods. Notice that the results here are slightly different from what is reported by Wang et al. [38] because we only evaluate 8-frame sequences with a resolution of 1024×512 due to GPU memory limitations. We generated a total of 500 short sequences on the validation set. We also provide results for video prediction when only the starting frame is given. As shown in Table 2, our method outperforms all other state-of-the-art approaches in video prediction from a

Figure 6: Comparisons with other competing baselines. Notice that vid2vid uses a sequence of semantic label maps while other methods only take **one** as input. Please zoom in for best view.

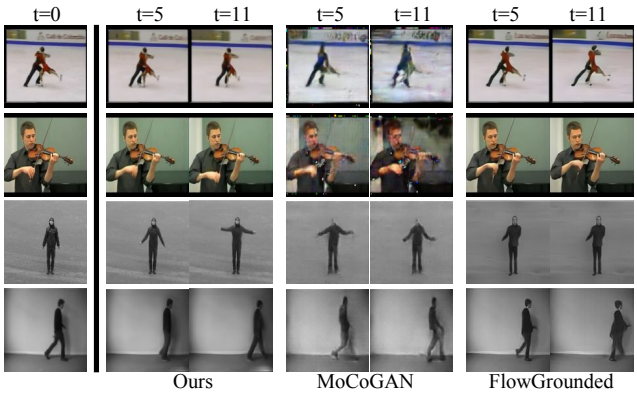

Figure 7: Comparisons with other competing baselines on UCF-101 dataset and KTH human dataset. Please zoom in to see the details.

static image.

Qualitative Results Fig. 6 compares our generation results with other approaches. MoCoGAN has limited capability in modeling video sequences (both motion and appearance). FG fails to synthesize the details of the scene,*e.g.* windows of the background building are completely missing,

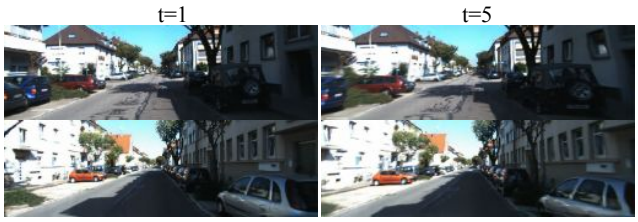

Figure 8: Samples of KITTI generated from model trained on the cityscapes dataset.

increasing blurriness. Our method maintains the semantic structure of the scene for the duration of the sequence and contains finer details than the previous two methods. The proposed method makes reasonable estimates of the objects' future motion and produces temporally coherent video sequence. Compared to the ground truth sequence, our model can generate semantically correct samples but with different properties, *e.g.*, a white car in the ground truth sequence appears as a silver car in our result. For vid2vid, where the input is a sequence of semantic label maps, shows realistic images with great details, but limited on preserving the temporal consistency across frames, *e.g.* the silver car in $t = 3$

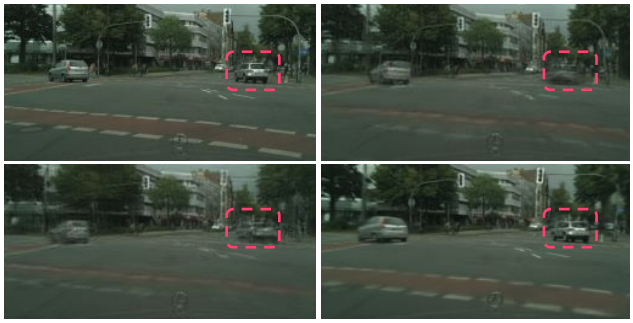

Figure 9: Ablation studies of our method. Top left: GT. Top right: w/o segmentation label map and flow. Bottom left:w/o flow. Bottom right: our full model. Our method preserve better the visual quality.

Human Preference Score	
seg2vid(ours) / MoCoGAN	**1.0** / 0.0
seg2vid(ours) / FG	**0.78** / 0.22
seg2vid(ours) /vid2vid	0.37 / **0.63**

Table 3: User study on video generation methods.

Human Preference Score	
seg2vid(ours) / MoCoGAN	**1.0** / 0.0
seg2vid(ours) / FG	**0.82** / 0.18

Table 4: User study on video prediction methods.

has turned into black in $t = 7$, while our methods keeps the same color. To further show the effectiveness of our method on predicting general motions, we provide visual results on UCF-101 dataset and KTH action dataset that mainly consist on people performing actions. As shown in Fig. 7, our method preserves well the body structure and synthesizing complex non-linear motions such as people skiing, playing violin and walking. We trained the model on Cityscapes and tested on samples from KITTI to show the method's generalization ability, shown in Fig. 8.

The user study illustrated in Table. 3 also shown that our method is the most favored except vid2vid. Additionally to the results of synthesized data, we also reported results for video prediction task. As shown in Fig. 5 our method can predict well background motion and simultaneously captured the movement of the car on the left side. The details and structure of the scene is well preserved with our approach while other methods suffer severe deformation. Table 4 shows that participants find our method to be more realistic.

5. Conclusion

In this work, we introduced the new video generation task conditioned only on a single semantic label map, and proposed a novel method for this task. Instead of learning the generation end-to-end, which is very challenging, we

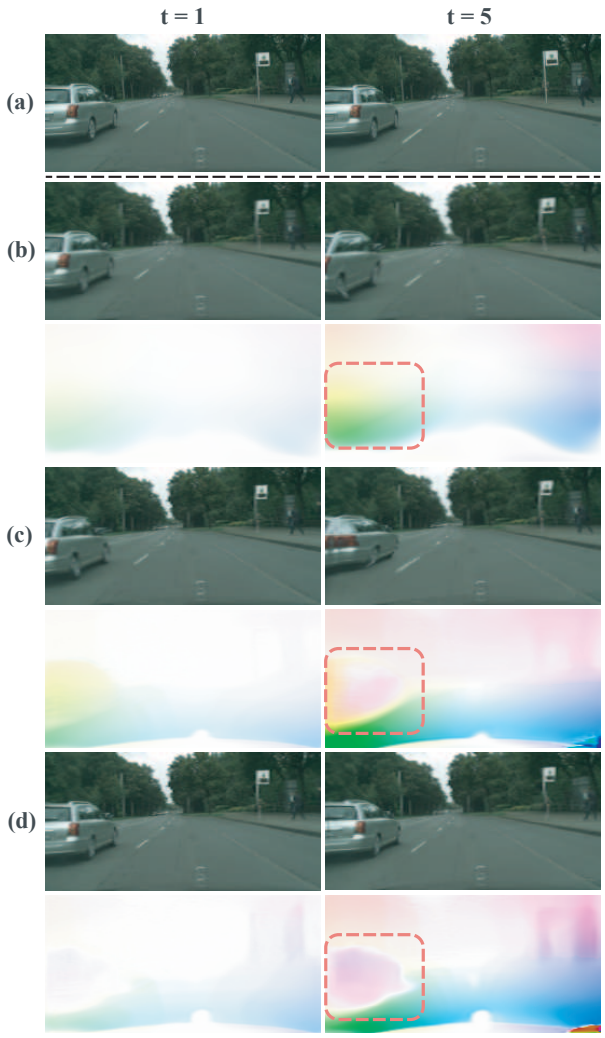

Figure 10: We compare three different variants of using semantic label map for flow and frame prediction. (a) ground truth, (b) w/o semantic label maps, (c) with semantic label maps, (d) with separate semantic label maps for background and foreground objects.

employed a divide and conquer strategy to model appearance and motion in a progressive manner to obtain quality results. We demonstrated that introducing semantic information brings large improvement when predicting motion from static content. The impressive performance compared to other baselines indicate the effectiveness of the proposed method for video generation.

Acknowledgements. This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14208417, CUHK14239816. We also want to thank Yucong Zhou for his technical support.

References

[1] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017. 3

[2] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. *arXiv preprint arXiv:1804.07739*, 2018. 2

[3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017. 2, 3

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 6

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 5, 6

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[7] E. Denton and R. Fergus. Stochastic video generation with a learned prior. *arXiv preprint arXiv:1802.07687*, 2018. 3

[8] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016. 3

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 1, 2, 3

[13] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016. 3

[14] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[15] J. Johnson, A. Gupta, and L. Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018. 2

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[17] I. Laptev, B. Caputo, et al. Recognizing human actions: a local svm approach. In *null*, pages 32–36. IEEE, 2004. 5

[18] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Flow-grounded spatial-temporal video prediction from still images. *arXiv preprint arXiv:1807.09755*, 2018. 3, 6

[19] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. *arXiv preprint arXiv:1701.01821*, 2, 2017. 3

[20] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 1, 2

[21] S. Meister, J. Hur, and S. Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017. 2

[22] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. 2

[23] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 3

[24] O. Ronneberger, P. Fischer, and T. 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. 6

[25] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 5, 2017. 1, 2

[26] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5, 2017. 2

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[28] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[29] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015. 3

[30] W. Trobin, T. Pock, D. Cremers, and H. Bischof. An unbiased second-order prior for high-accuracy motion estimation. In *Joint Pattern Recognition Symposium*, pages 396–405. Springer, 2008. 4

[31] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*, 2017. 1, 3, 6

[32] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle. Neural autoregressive distribution estimation. *The Journal of Machine Learning Research*, 17(1):7184–7220, 2016. 2

[33] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 3

[34] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016. 1, 2

[35] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016. 3

[36] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3352–3361. IEEE, 2017. 3

[37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 1, 2, 3, 6

[38] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. 2, 3, 6

[39] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016. 3, 4

[40] C. Zhang, Z. Li, R. Cai, H. Chao, and Y. Rui. As-rigid-as-possible stereo under second order smoothness priors. In *European Conference on Computer Vision*, pages 112–126. Springer, 2014. 4

[41] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint*, 2017. 1, 2

[42] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 1, 2, 3