

# World-Consistent Video-to-Video Synthesis

Arun Mallya\*, Ting-Chun Wang\*, Karan Sapra, and Ming-Yu Liu

NVIDIA

{amallya,tingchunw,ksapra,mingyul}@nvidia.com  
<https://nvlabs.github.io/wc-vid2vid>

**Abstract.** Video-to-video synthesis (vid2vid) aims for converting high-level semantic inputs to photorealistic videos. While existing vid2vid methods can achieve short-term temporal consistency, they fail to ensure the long-term one. This is because they lack knowledge of the 3D world being rendered and generate each frame only based on the past few frames. To address the limitation, we introduce a novel vid2vid framework that efficiently and effectively utilizes all past generated frames during rendering. This is achieved by *condensing* the 3D world rendered so far into a physically-grounded estimate of the current frame, which we call the *guidance image*. We further propose a novel neural network architecture to take advantage of the information stored in the guidance images. Extensive experimental results on several challenging datasets verify the effectiveness of our approach in achieving *world consistency*—the output video is consistent within the entire rendered 3D world.

**Keywords:** neural rendering, video synthesis, GAN

## 1 Introduction

Video-to-video synthesis [80] concerns generating a sequence of photorealistic images given a sequence of semantic representations extracted from a source 3D world. For example, the representations can be the semantic segmentation masks rendered by a graphics engine while driving a car in a virtual city [80]. The representations can also be the pose maps extracted from a source video of a person dancing, and the application is to create a video of a different person performing the same dance [8]. From the creation of a new class of digital artworks to applications in computer graphics, the video-to-video synthesis task has many exciting practical use-cases. A key requirement of any such video-to-video synthesis model is the ability to generate images that are not only individually photorealistic, but also temporally smooth. Moreover, the generated images have to follow the geometric and semantic structure of the source 3D world.

While we have observed steady improvement in photorealism and short-term temporal stability in the generation results, we argue that one crucial aspect of the problem has been largely overlooked, which is the *long-term temporal consistency* problem. As a specific example, when visiting the same location in the

---

\* Equal contribution

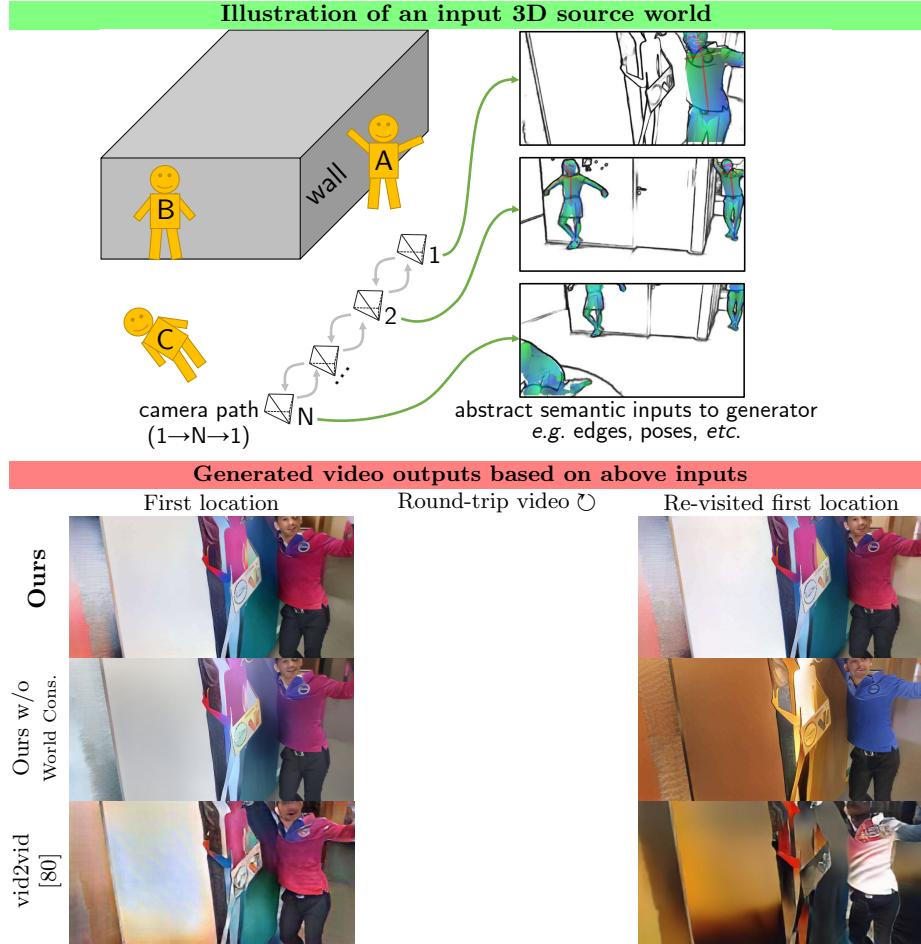


Fig. 1: Imagine moving around in a world such as the one abstracted at the top. As you move from locations  $1 \rightarrow N \rightarrow 1$ , you would expect the appearance of previously seen walls and people to remain unchanged. However, current video-to-video synthesis methods such as vid2vid [80] or our improved architecture combining vid2vid with SPADE [61] cannot produce such world-consistent videos (third and second rows). Only our method is able to produce videos consistent over viewpoints by adding a mechanism for world consistency (first row).

*Please view with Acrobat Reader. Click any middle column image to play video.*

virtual city, an existing vid2vid method [79,80] could generate an image that is very different from the one it generated when the car first visited the location, despite using the same semantic inputs. Existing vid2vid methods rely on optical flow warping and generate an image conditioned on the past few generated images. While such operations can ensure short-term temporal stability, they

cannot guarantee long-term temporal consistency. Existing vid2vid models have no knowledge of what they have rendered in the past. Even for a short round-trip in a virtual room, these methods fail to preserve the appearances of the wall and the person in the generated video, as illustrated in Fig. 1.

In this paper, we attempt to address the long-term temporal consistency problem, by bolstering vid2vid models with memories of the past frames. By combining ideas from scene flow [74] and conditional image synthesis models [61], we propose a novel architecture that explicitly enforces consistency in the entire generated sequence. We perform extensive experiments on several benchmark datasets, with comparisons to the state-of-the-art methods. Both quantitative and visual results verify that our approach achieves significantly better image quality and long-term temporal stability. On the application side, we also show that our approach can be used to generate videos consistent across multiple viewpoints, enabling simultaneous multi-agent world creation and exploration.

## 2 Related work

**Semantic Image Synthesis** [11,51,61,62,81] refers to the problem of converting a single input semantic representation to an output photorealistic image. Built on top of the generative adversarial networks (GAN) [24] framework, existing methods [51,61,81] propose various novel network architectures to advance state-of-the-art. Our work is built on the SPADE architecture proposed by Park *et al.* [61] but focuses on the temporal stability issue in video synthesis.

**Conditional GANs** synthesize data conditioned on user input. This stands in contrast to unconditional GANs that synthesize data solely based on random variable inputs [24,26,38,39]. Based on the input type, there exist label-conditional GANs [6,57,59,90], text-conditional GANs [63,87,91], image-conditional GANs [3,5,14,31,33,43,49,50,65,70,96], scene-graph conditional GANs [34], and layout-conditional GANs [92]. Our method is a video-conditional GAN, where we generate a video conditioned on an input video. We address the long-term temporal stability issue that the state-of-the-art overlooks [8,79,80].

**Video synthesis** exists in many forms, including 1) unconditional video synthesis [64,73,76], which converts random variable inputs to video clips, 2) future video prediction [17,19,27,30,36,42,44,47,53,54,60,69,75,77,78,88], which generates future video frames based on the observed ones, and 3) video-to-video synthesis [8,12,22,79,80,95], which converts an input semantic video to a real video. Our work belongs to the last category. Our method treats the input video as one from a self-consistent world so that when the agent returns to a spot that it has previously visited, the newly generated frames should be consistent with the past generated frames. While a few works have focused on improving the temporal consistency of an input video [4,41,89], our method does not treat consistency as a post-processing step, but rather as a core part of the video generation process.

**Novel-view synthesis** aims to synthesize images at unseen viewpoints given some viewpoints of the scene. Most of the existing works require images at

multiple reference viewpoints as input [13,20,21,28,35,56,94]. While some works can synthesize novel views based on a single image [68,82,85], the synthesized views are usually close to the reference views. Our work differs from these works in the sense that our input is different – instead of using a set of RGB images, our network takes in a sequence of semantic maps. If we directly treat all past synthesized frames as reference views, it makes the memory requirement grow linearly with respect to the video length. If we only use the latest frames, the system cannot handle long-term consistency as shown in Fig. 1. Instead, we propose a novel framework to keep track of the synthesis history in this work.

The closest related works are those on neural rendering [2,55,67,71], which can re-render a scene from arbitrary viewpoints after training on a set of given viewpoints. However, note that these methods still require RGB images from different viewpoints as input, making it unsuitable for applications such as those to game engines. On the other hand, our method can directly generate RGB images using semantic inputs, so rendering a virtual world becomes more effortless. Moreover, they need to train a separate model (or part of the model) for each scene, while we only need one model per dataset, or domain.

### 3 World-consistent video-to-video synthesis

**Background.** Recent image-to-image translation methods perform extremely well when turning semantic images to realistic outputs. To produce videos instead of images, simply doing it frame-by-frame will usually result in severe flickering artifacts [80]. To resolve this, vid2vid [80] proposes to take both the semantic inputs and  $L$  previously generated frames as input to the network (*e.g.*  $L = 3$ ). The network then generates three outputs – a hallucinated frame, a flow map, and a (soft) mask. The flow map is used to warp the previous frame and linearly combined with the hallucinated frame using the soft mask. Ideally, the network should reuse the content in the warped frame as much as possible, and only use the disoccluded parts from the hallucinated frame.

While the above framework reduces flickering between neighboring frames, it still struggles to ensure long-term consistency. This is because it only keeps track of the past  $L$  frames, and cannot memorize everything in the past. Consider the scenario in Fig. 1, where an object moves out of and back in the field-of-view. In this case, we would want to make sure its appearance is similar during the revisit, but that cannot be handled by existing frameworks like vid2vid [80].

In light of this, we propose a new framework to handle *world-consistency*. It is a superset of *temporal consistency*, which only ensures consistency between frames in a video. A world-consistent video should not only be temporally stable, but also be consistent across the entire 3D world the user is viewing. This not only makes the output look more realistic, but also enables applications such as the multi-player scenario where different players can view the same scene from different viewpoints. We achieve this by using a novel *guidance image* conditional scheme, which is detailed below.

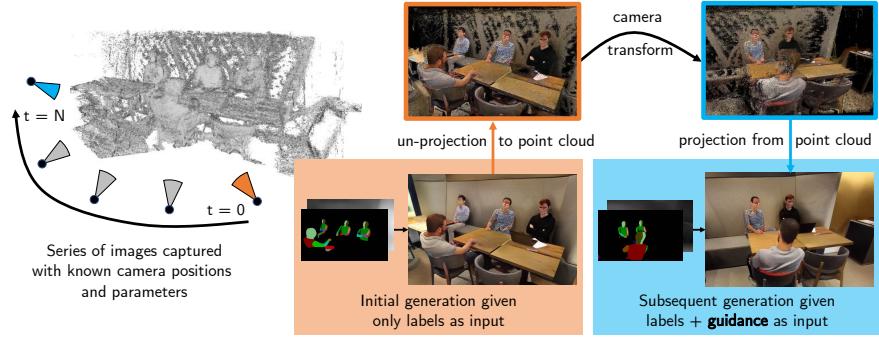


Fig. 2: Overview of *guidance image* generation for training. Consider a scene in which a camera(s) with known parameters and positions travels over time  $t = 0, \dots, N$ . At  $t = 0$ , the scene is textureless and an output image is generated for this viewpoint. The output image is then back-projected to the scene and a *guidance image* for a subsequent camera position is generated by projecting the partially textured point cloud. Using this guidance image, the generative method can produce an output that is consistent across views and smooth over time. Note that the guidance image can be noisy, misaligned, and have holes, and the generation method should be robust to such inputs.

**Guidance images and their generation.** The lack of knowledge about the world structure being generated limits the ability of vid2vid to generate view-consistent outputs. As shown in Fig. 5 and Sec. 4, the color and structure of the objects generated by vid2vid [80] tend to drift over time. We believe that in order to produce realistic outputs that are consistent over time and viewpoint change, an ideal method must be aware of the 3D structure of the world.

To achieve this, we introduce the concept of “*guidance images*”, which are physically-grounded estimates of what the next output frame should look like, based on how the world has been generated so far. As alluded to in their name, the role of these “*guidance images*” is to guide the generative model to produce colors and textures that respect previous outputs. Prior works including vid2vid [80] rely on optical flows to warp the previous frame for producing an estimate of the next frame. Our guidance image differs from this warped frame in two aspects. First, instead of using optical flow, the guidance image should be generated by using the motion field, or scene flow, which describes the true motion of each 3D point in the world<sup>1</sup>. Second, the guidance image should aggregate information from *all* past viewpoints (and thus frames), instead of only the direct previous frames as in vid2vid. This makes sure that the generated frame is consistent with the entire history.

While estimating motion fields without an RGB-D sensor [23] or a rendering engine [18] is not easy, we can obtain motion fields for the static parts of the world

<sup>1</sup> As an example, consider a textureless sphere rotating under constant illumination. In this case, the optical flow would be zero, but the motion field would be nonzero.

by reconstructing part of the 3D world using structure from motion (SfM) [52, 72]. This enables us to generate guidance images as shown in Fig. 2 for training our video-to-video synthesis method using datasets captured by regular cameras. Once we have the 3D point cloud of the world, the video synthesis process can be thought of as a camera moving through the world and texturing every new 3D point it sees. Consider a camera moving through space and time as shown in the left part of Fig. 2. Suppose we generate an output image at  $t = 0$ . This image can be back-projected to the 3D point cloud and colors can be assigned to the points, so as to create a persistent representation of the world. At a later time step,  $t = N$ , we can obtain the projection of the 3D point cloud to the camera and create a guidance image leveraging estimated motion fields. Our method can then generate an output frame based on the guidance image.

Although we generate guidance images using the projection of 3D point clouds, it can also be generated by any other method that gives a reasonable estimate. This makes the concept powerful, as we can use different sources to generate guidance images at training and test time. For example, at test time we can generate guidance images using a graphics engine, which can provide ground truth 3D correspondences. This enables just-in-time colorization of a virtual 3D world with real-world colors and textures, as we move through the world.

Note that our guidance image also differs from the projected image used in prior works like Meshry *et al.* [55] in several aspects. First, in their case, the 3D point cloud is fixed once constructed, while in our case it is constantly being “colorized” as we synthesize more and more frames. As a result, our guidance image is blank at the beginning, and can become denser depending on the viewpoint. Second, the way we use these guidance images to generate outputs is also different. The guidance images can have misalignments and holes due to limitations of SfM, for example in the background and in the person’s head in Fig. 2. As a result, our method also differs from DeepFovea [37], which inpaints sparsely but accurately rendered video frames. In the following subsection, we describe a method that is robust to noises in guidance images, so it can produce outputs consistent over time and viewpoints.

**Framework for generating videos using guidance images.** Once the guidance images are generated, we are able to utilize them to synthesize the next frame. Our generator network is based on the SPADE architecture proposed by Park *et al.* [61], which accepts a random vector encoding the image style as input and uses a series of SPADE blocks and upsampling layers to generate an output image. Each SPADE block takes a semantic map as input and learns to modulate the incoming feature maps through an affine transform  $y = x \cdot \gamma_{\text{seg}} + \beta_{\text{seg}}$ , where  $x$  is the incoming feature map, and  $\gamma_{\text{seg}}$  and  $\beta_{\text{seg}}$  are predicted from the input segmentation map.

An overview of our method is shown in Fig. 3. At a high-level, our method consists of four sub-networks: 1) an input label embedding network (**orange**), 2) an image encoder (**red**), 3) a flow embedding network (**green**), and 4) an image generator (**gray**). In our method, we make two modifications to the original SPADE network. First, we feed in the concatenated labels (semantic segmen-

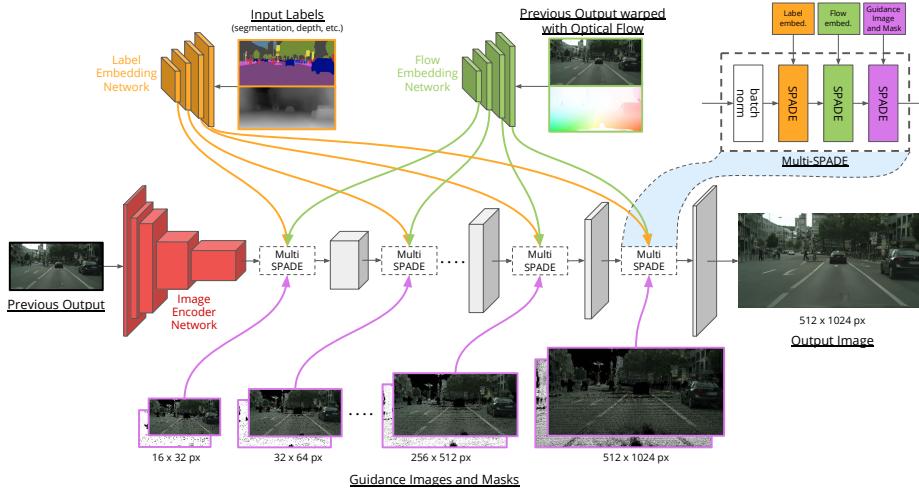


Fig. 3: Overview of our world consistent video-to-video synthesis architecture. Our Multi-SPADE module takes input labels, warped previous frames, and guidance images to modulate the features in each layer of our generator.

tation, edge maps, *etc.*) to a label embedding network (**orange**), and extract features in corresponding output layers as input to each SPADE block in the generator. Second, to keep the image style consistent over time, we encode the previously synthesized frame using the image encoder (**red**), and provide this embedding to our generator (**gray**) in place of the random vector<sup>2</sup>.

*Utilizing guidance images.* Although using this modified SPADE architecture produces output images with better visual quality than vid2vid [80], the outputs are not temporally stable, as shown in Sec. 4. To ensure world-consistency of the output, we would want to incorporate information from the introduced guidance images. Simply linearly combining it with the hallucinated frame from the SPADE generator is problematic, since the hallucinated frame may contain something very different from the guidance images. Another way is to directly concatenate it with the input labels. However, the semantic inputs and guidance images have different physical meanings. Besides, unlike semantic inputs, which are labeled densely (per pixel), the guidance images are labeled sparsely. Directly concatenating them would require the network to compensate for the difference. Hence, to avoid these potential issues, we choose to treat these two types of inputs differently.

To handle the sparsity of the guidance images, we first apply partial convolutions [48] on these images to extract features. Partial convolutions only convolve valid regions in the input with the convolution kernels, so the output features can be uncontaminated by the holes in the image. These features are then used

<sup>2</sup> When generating the first frame where no previous frame exists, we use an encoder which accepts the semantic map as input.

Segmentation	Depth	Guidance Image	Generated Output

Fig. 4: Sample inputs and generated outputs on Cityscapes. Note how the guidance image is initially black, and becomes denser as more frames are synthesized. Click on any image to play video. *Please view with Acrobat Reader.*

to generate affine transformation parameters  $\gamma_{\text{guidance}}$  and  $\beta_{\text{guidance}}$ , which are *inserted* into existing SPADE blocks while keeping the rest of the blocks untouched. This results in a *Multi-SPADE* module, which allows us to use multiple conditioned inputs in sequence, so we can not only condition on the current input labels, but also on our guidance images,

$$y = (x \cdot \gamma_{\text{label}} + \beta_{\text{label}}) \cdot \gamma_{\text{guidance}} + \beta_{\text{guidance}}. \quad (1)$$

Using this module yields several benefits. First, conditioning on these maps generates more temporally smooth and higher quality frames than simple linear blending techniques. Separating the two types of input (semantic labels and guidance images) also allows us to adopt different types of convolutions (i.e. normal vs. partial). Second, since most of the network architecture remains unchanged, we can initialize the weights of the generator with one trained for single image generation. It is easy to collect large training datasets for single image generation by crawling the internet, while video datasets can be harder to collect and annotate. After the single image generator is trained, we can train a video generator by just training the newly added layers (*i.e.* layers generating  $\gamma_{\text{guidance}}$  and  $\beta_{\text{guidance}}$ ) and only finetune the other parts of the network.

*Handling dynamic objects.* The guidance image allows us to generate world-consistent outputs over time. However, since the guidance is generated based on SfM for real-world scenes, it has the inherent limitation that SfM cannot handle dynamic objects. To resolve this issue, we revert to using optical flow-warped frames to serve as additional maps in addition to the guidance images we have from SfM. The complete Multi-SPADE module then becomes

$$y = ((x \cdot \gamma_{\text{label}} + \beta_{\text{label}}) \cdot \gamma_{\text{flow}} + \beta_{\text{flow}}) \cdot \gamma_{\text{guidance}} + \beta_{\text{guidance}}, \quad (2)$$

where  $\gamma_{\text{flow}}$  and  $\beta_{\text{flow}}$  are generated using a flow-embedding network (**green**) applied on the optical flow-warped previous frame. This provides additional constraints that the generated frame should be consistent even in the dynamic regions. Note that this is needed only due to the limitation of SfM, and can potentially be removed when ground truth / high quality 3D registrations are available, for example in the case of game engines, or RGB-D data capture.

Figure 4 shows a sample set of inputs and outputs generated by our method on the Cityscapes dataset.

## 4 Experiments

**Implementation details.** We train our network in two stages. In the first stage, we only train our network to generate single images. This means that only the first SPADE layer of our Multi-SPADE block (visualized in Fig. 3) is trained. Following this, we have a network that can generate high-quality single frame outputs. In the second stage, we train on video clips, progressively doubling the generated video length every epoch, starting from 8 frames and stopping at 32 frames. In this stage, all 3 SPADE layers of each Multi-SPADE block are trained. We found that this two-stage pipeline makes the training faster and more stable. We observed that the ordering of the flow and guidance SPADEs did not make a significant difference in the output quality. We train the network for 20 epochs in each stage, and this takes about 10 days on an NVIDIA DGX-1 (8 V-100 GPUs) for an output resolution of  $1024 \times 512$ .

We train our generator with the multi-scale image discriminator using perceptual and GAN feature matching losses as in SPADE [61]. Following vid2vid [80], we add a temporal video discriminator at two temporal scales and a warping loss that encourages the output frame to be similar to the optical flow-warped previous frame. We also add a loss term to encourage the output frame to correspond to the guidance image, and this is necessary to ensure view consistency. Additional details about architecture and loss terms can be found in Appendix A and B. Code and trained models will be released upon publication.

**Datasets.** We train and evaluate our method on three datasets, Cityscapes [15], MannequinChallenge [45], and ScanNet [16], as they have mostly static scenes where existing SfM methods perform well.

- **Cityscapes** [15]. This dataset consists of driving videos of  $2048 \times 1024$  resolution captured in several German cities, using a pair of stereo cameras. We split this dataset into a training set of 3500 videos with 30 frames each, and a test set of 3 long sequences with 600-1200 frames each, similar to vid2vid [80]. As not all the images are labeled with segmentation masks, we annotate the images using the network from Zhu *et al.* [97], which is based on a DeepLabv3-Plus [10]-like architecture with a WideResNet38 [84] backbone.
- **MannequinChallenge** [45]. This dataset contains video clips captured using hand-held cameras, of people pretending frozen in a large variety of poses, imitating mannequins. We resize all frames to  $1024 \times 512$  and randomly split this dataset into 3040 train sequences and 292 test sequences, with sequence lengths ranging from 5-140 frames. We generate human body segmentation and part-specific UV coordinate maps using DensePose [25,83] and body poses using OpenPose [7].
- **ScanNet** [16]. This dataset contains multiple video clips captured in a total of 706 indoor rooms. We set aside 50 rooms for testing, and the rest for training. From each video sequence, we extracted 3 sub-sequences of length at most 100, resulting in 4000 train sequences and 289 test sequences, with

images of size  $512 \times 512$ . We used the provided segmentation maps based on the NYUDv2 [58] 40 labels.

For all datasets, we also use MegaDepth [46] to generate depth maps and add the visualized inverted depth images as input. As the MannequinChallenge and ScanNet datasets contain a large variety of objects and classes which are not fully annotated, we use edge maps produced by HED [86] in order to better represent the input content. In order to generate guidance images, we performed SfM on all the video sequences using OpenSfM [1], which provided 3D point clouds and estimated cameras poses and parameters as output.

**Baselines.** We compare our method against the following strong baselines.

- vid2vid [80]. This is the prior state-of-the-art method for video-to-video synthesis. For comparison on Cityscapes, we use the publicly available pre-trained model. For the other two datasets, we train vid2vid from scratch using the public code, while providing the same input labels (semantic segmentation, depth, edge maps, etc.) as to our method.
- Inpainting [48]. We train a state-of-the-art partial convolution-based inpainting method to fill in the pixels missing from our guidance images. We train the models from scratch for each dataset, using masks obtained from the corresponding guidance images.
- Ours w/o W.C. (World Consistency). As an ablation, we also compare against our model that does not use guidance images. In this case, only the first two SPADE layers in each Multi-SPADE block are trained (label and flow-warped previous output SPADEs). Other details are the same as our full model.

**Evaluation metrics.** We use both objective and subjective metrics for evaluating our model against the baselines.

- *Segmentation accuracy and Fréchet Inception Distance (FID).* We adopt metrics widely used in prior work on image synthesis [11,61,81] to measure the quality of generated video frames. We evaluate the output frames based on how well they can be segmented by a trained segmentation network. We report both the mean Intersection-Over-Union (mIOU) and Pixel Accuracy (P.A.) using the PSPNet [93] (Cityscapes) and DeepLabv2 [9] (MannequinChallenge & ScanNet). We also use the Fréchet Inception Distance (FID) [29] to measure the distance between the distributions of the generated and real images, using the standard Inception-v3 network.
- *Human preference score.* Using Amazon Mechanical Turk (AMT), we perform a subjective visual test to gauge the relative quality of videos. We evaluate videos on two criteria: 1) *photorealism* and 2) *temporal stability*. The first aims to find which generated video looks more like a real video, while the second aims to find which one is more temporally smooth and has lesser flickering. For each question, an AMT participant is shown two videos synthesized by two different methods, and asked to choose the better one according to the current criterion. We generate several hundred questions for each dataset, each of them is answered by 3 different workers. We evaluate an algorithm by the ratio that its outputs are preferred.

Table 1: Comparison scores.  $\downarrow$  means lower is better, while  $\uparrow$  means the opposite.

Method	Cityscapes			MannequinChallenge			ScanNet		
	FID $\downarrow$	mIOU $\uparrow$	P.A. $\uparrow$	FID $\downarrow$	mIOU $\uparrow$	P.A. $\uparrow$	FID $\downarrow$	mIOU $\uparrow$	P.A. $\uparrow$
<b>Image synthesis models</b>									
SPADE [61]	48.25	0.63	0.95	29.99	0.13	0.63	31.46	0.08	0.54
<b>Video synthesis models</b>									
vid2vid [80]	69.07	0.55	0.94	72.25	0.05	0.45	60.03	0.04	0.35
Ours w/o W.C.	51.51	<b>0.62</b>	0.95	27.23	0.17	0.67	<b>20.93</b>	0.12	0.62
Ours	<b>49.89</b>	0.61	<b>0.95</b>	<b>22.69</b>	<b>0.19</b>	<b>0.69</b>	21.07	<b>0.13</b>	<b>0.63</b>

Table 2: Human preference scores. Higher is better.

Compared Methods	Cityscapes	MannequinChallenge	ScanNet
<b>Image Realism</b>			
Ours/vid2vid [80]	<b>0.73</b> /0.27	<b>0.83</b> /0.17	<b>0.77</b> /0.23
<b>Temporal Stability</b>			
Ours/vid2vid [80]	<b>0.75</b> /0.25	<b>0.63</b> /0.37	<b>0.82</b> /0.18

Table 3: Forward-backward consistency.  $\triangle$  means difference.

Method	Cityscapes		MannequinChallenge		ScanNet	
	$\triangle$ RGB $\downarrow$	$\triangle$ LAB $\downarrow$	$\triangle$ RGB $\downarrow$	$\triangle$ LAB $\downarrow$	$\triangle$ RGB $\downarrow$	$\triangle$ LAB $\downarrow$
vid2vid [80]	14.90	3.46	37.56	9.42	46.30	12.16
Ours	<b>8.73</b>	<b>2.04</b>	<b>12.61</b>	<b>3.61</b>	<b>11.85</b>	<b>3.41</b>

- *Forward-Backward consistency.* A major contribution of our work is generating outputs that are consistent over a longer duration of time with the world that was previously generated. All our datasets have videos that explore new parts of the world over time, rarely revisiting previously explored parts. However, a simple way to revisit a location is to play the video in forward and then in reverse, *i.e.* arrange frames from time  $t = 0, 1, \dots, N-1, N, N-1, \dots, 1, 0$ . We can then compare the first produced and last produced frames and measure their difference. We measure the difference per-pixel in both RGB and LAB space, and a lower value would indicate better long-term consistency.

**Main results.** In Table 1, we compare our proposed approach against vid2vid [80], as well as SPADE [61], which is the single image generator that our method builds upon. We also compare against a version of our method that does not use guidance images and is thus not world-consistent (Ours w/o W.C.). Inpainting [48] could not provide meaningful output images without large artifacts, as shown in Fig 5. We can observe that our method consistently beats vid2vid on all three metrics on all three datasets, indicating superior image quality. Interestingly, our method also improves upon SPADE in FID, probably as a result of reducing temporal variance across an output video sequence. We also see improvements over Ours w/o W.C. on almost all metrics.

In Table 2, we show human evaluation results on metrics of image realism and temporal stability. We observe that the majority of workers rank our method better on both metrics.

vid2vid [80]

SPADE [61]

Inpainting [48]

Ours w/o W.C.

Ours

Fig. 5: Comparison of different video generation methods on the Cityscapes dataset. Note that for our results, the textures of the cars, roads, and signboards are stable over time, while they change gradually in vid2vid and other methods. Click on an image to play the video. *Please view with Acrobat Reader.*

In Fig. 5, we visualize some sequences generated by the various methods (please zoom in and play the videos in Adobe Acrobat). We can observe that in the first row, vid2vid [80] produces temporal artifacts in the cars parked to the side and patterns on the road. SPADE [61], which produces one frame at a time, produces very unstable videos, as shown in the second row. The third row shows outputs from the partial convolution-based inpainting [48] method. It clearly has a hard time producing visually and semantically meaningful outputs. The fourth row shows Ours w/o W.C., an intermediate version of our method that uses labels and optical flow-warped previous output as input. While this clearly improves upon vid2vid in image quality and SPADE in temporal stability, it causes flickering in trees, cars, and signboards. The last row shows our method. Note how the textures of the cars, roads, and signboards, which are areas we

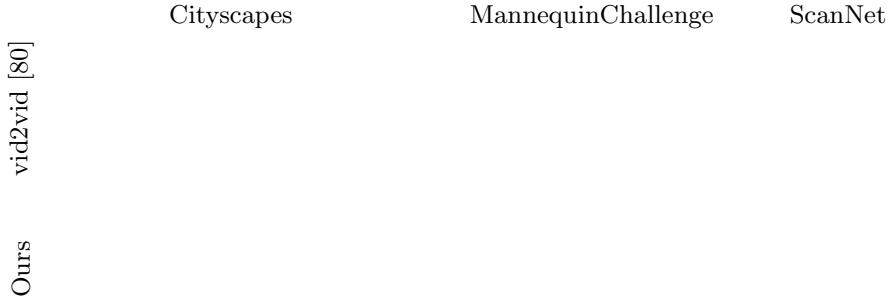


Fig. 6: Forward-backward consistency. Click on each image to see the change in output when the viewpoint is revisited. Note how drastically vid2vid results change, while ours remain almost the same. *Please view with Acrobat Reader.*

Fig. 7: Qualitative results on the MannequinChallenge and ScanNet datasets. Click on an image to play video. Note the results are consistent over time and viewpoints. *Please view with Acrobat Reader.*

have guidance images, are stable over time. We also provide high resolution, uncompressed videos for all three datasets on our website.

In Table 3, we compare the forward-backward consistency of different methods, and it shows that our method beats vid2vid [80] by a large margin, especially on the MannequinChallenge and ScanNet datasets (by more than a factor of 3). Figure 6 visualizes some frames at the start and end of generation. As can be seen, the outputs of vid2vid change dramatically, while ours are consistent. We show additional qualitative examples in Fig. 7. We also provide additional quantitative results on short-term consistency in Appendix C.

**Generating consistent stereo outputs.** Here, we show a novel application enabled by our method through the use of guidance images. We show videos rendered simultaneously for multiple viewpoints, specifically for a pair of stereo viewpoints on the Cityscapes dataset in Fig. 8. For the strongest baseline, Ours w/o W.C., the left-right videos can only be generated independently, and they clearly are not consistent across multiple viewpoints, as highlighted by the boxes. On the other hand, our method can generate left-right videos in sync by sharing the underlying 3D point cloud and guidance maps. Note how the textures on

Ours w/o W.C.  
Ours

Fig. 8: Stereo results on Cityscapes. Click on an image to see the the outputs produced by a pair of stereo cameras. Note how our method produces images consistent across the two views, while they differ in the highlighted regions without using the world consistency. *Please view with Acrobat Reader.*

roads, including shadows, move in sync and remain consistent over time and camera locations.

## 5 Conclusions and discussion

We presented a video-to-video synthesis framework that can achieve world consistency. By using a novel guidance image extracted from the generated 3D world, we are able to synthesize the current frame conditioned on all the past frames. The conditioning was implemented using a novel Multi-SPADE module, which not only led to better visual quality, but also made transplanting a single image generator to a video generator possible. Comparisons on several challenging datasets showed that our method improves upon prior state-of-the-art methods.

While advancing the state-of-the-art, our framework still has several limitations. For example, the guidance image generation is based on SfM. When SfM fails to register the 3D content, our method will also fail to ensure consistency. Also, we do not consider a possible change in time of the day or lighting in the current framework. In the future, our framework can benefit from improved guidance images enabled by better 3D registration algorithms. Furthermore, the albedo and shading of the 3D world may be disentangled to better model the time effects. We leave these to future work.

**Acknowledgements.** We would like to thank Jan Kautz, Guilin Liu, Andrew Tao, and Bryan Catanzaro for their feedback, and Sabu Nadarajan, Nithya Natesan, and Sivakumar Arayandi Thottakara for helping us with the compute, without which this work would not have been possible.

## References

1. OpenSfM. <https://github.com/mapillary/OpenSfM>
2. Aliev, K.A., Ulyanov, D., Lempitsky, V.: Neural point-based graphics. arXiv preprint arXiv:1906.08240 (2019)
3. Benaim, S., Wolf, L.: One-shot unsupervised cross domain translation. In: Conference on Neural Information Processing Systems (NeurIPS) (2018)
4. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. ACM Transactions on Graphics (TOG) (2015)
5. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR) (2019)
7. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008 (2018)
8. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: IEEE International Conference on Computer Vision (ICCV) (2019)
9. 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017)
10. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (ECCV) (2018)
11. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
12. Chen, Y., Pan, Y., Yao, T., Tian, X., Mei, T.: Mocycle-gan: Unpaired video-to-video translation. In: ACM International Conference on Multimedia (MM) (2019)
13. Choi, I., Gallo, O., Troccoli, A., Kim, M.H., Kautz, J.: Extreme view synthesis. In: IEEE International Conference on Computer Vision (ICCV) (2019)
14. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Denton, E.L., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
18. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Conference on Robot Learning (CoRL) (2017)
19. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Conference on Neural Information Processing Systems (NeurIPS) (2016)

20. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
21. Flynn, J., Neulander, I., Philbin, J., Snavely, N.: Deepstereo: Learning to predict new views from the world's imagery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
22. Gafni, O., Wolf, L., Taigman, Y.: Vid2game: Controllable characters extracted from real-world videos. In: International Conference on Learning Representations (ICLR) (2020)
23. Golyanik, V., Kim, K., Maier, R., Nießner, M., Stricker, D., Kautz, J.: Multiframe scene flow with piecewise rigid motion. In: International Conference on 3D Vision (3DV) (2017)
24. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2014)
25. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense human pose estimation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
26. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
27. Hao, Z., Huang, X., Belongie, S.: Controllable video generation with sparse trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
28. Hedman, P., Philip, J., Price, T., Frahm, J.M., Drettakis, G., Brostow, G.: Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (TOG) (2018)
29. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
30. Hu, Q., Waelchli, A., Portenier, T., Zwicker, M., Favaro, P.: Video synthesis from a single image and motion stroke. arXiv preprint arXiv:1812.01874 (2018)
31. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. European Conference on Computer Vision (ECCV) (2018)
32. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
33. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
34. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
35. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG) (2016)
36. Kalchbrenner, N., Oord, A.v.d., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K.: Video pixel networks. International Conference on Machine Learning (ICML) (2017)
37. Kaplanyan, A.S., Sochenov, A., Leimkühler, T., Okunev, M., Goodall, T., Rufo, G.: DeepFovea: neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. ACM Transactions on Graphics (TOG) (2019)

38. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR) (2018)
39. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
40. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
41. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: European Conference on Computer Vision (ECCV) (2018)
42. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523 (2018)
43. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision (ECCV) (2018)
44. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Flow-grounded spatial-temporal video prediction from still images. In: European Conference on Computer Vision (ECCV) (2018)
45. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
46. Li, Z., Snavely, N.: MegaDepth: Learning single-view depth prediction from internet photos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
47. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion GAN for future-flow embedded video prediction. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
48. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: European Conference on Computer Vision (ECCV) (2018)
49. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2017)
50. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: IEEE International Conference on Computer Vision (ICCV) (2019)
51. Liu, X., Yin, G., Shao, J., Wang, X., et al.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2019)
52. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. Nature (1981)
53. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. In: International Conference on Learning Representations (ICLR) (2017)
54. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (ICLR) (2016)
55. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: Neural rerendering in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

56. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* (2019)
57. Miyato, T., Koyama, M.: cGANs with projection discriminator. In: International Conference on Learning Representations (ICLR) (2018)
58. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: European Conference on Computer Vision (ECCV) (2012)
59. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: International Conference on Machine Learning (ICML) (2017)
60. Pan, J., Wang, C., Jia, X., Shao, J., Sheng, L., Yan, J., Wang, X.: Video generation from single semantic label map. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
61. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
62. Qi, X., Chen, Q., Jia, J., Koltun, V.: Semi-parametric image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
63. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning (ICML) (2016)
64. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: IEEE International Conference on Computer Vision (ICCV) (2017)
65. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
66. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
67. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
68. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4D RGBD light field from a single image. In: IEEE International Conference on Computer Vision (ICCV) (2017)
69. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: International Conference on Machine Learning (ICML) (2015)
70. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: International Conference on Learning Representations (ICLR) (2017)
71. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* (2019)
72. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)* (1992)
73. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: MoCoGAN: Decomposing motion and content for video generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
74. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (1999)

75. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: International Conference on Learning Representations (ICLR) (2017)
76. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Conference on Neural Information Processing Systems (NeurIPS) (2016)
77. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: European Conference on Computer Vision (ECCV) (2016)
78. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: IEEE International Conference on Computer Vision (ICCV) (2017)
79. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2019)
80. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2018)
81. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
82. Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J.: Synsin: End-to-end view synthesis from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
83. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
84. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition (2019)
85. Xie, J., Girshick, R., Farhadi, A.: Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In: European Conference on Computer Vision (ECCV) (2016)
86. Xie, S., Tu, Z.: Holistically-nested edge detection. In: IEEE International Conference on Computer Vision (ICCV) (2015)
87. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
88. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: Conference on Neural Information Processing Systems (NeurIPS) (2016)
89. Yao, C.H., Chang, C.Y., Chien, S.Y.: Occlusion-aware video temporal consistency. In: ACM International Conference on Multimedia (MM) (2017)
90. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning (ICML) (2019)
91. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
92. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
93. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

94. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. In: ACM SIGGRAPH (2018)
95. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Dance dance generation: Motion transfer for internet videos. arXiv preprint arXiv:1904.00129 (2019)
96. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
97. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

## A Objective functions

Our objective functions contain five losses: an image GAN loss, a video GAN loss, a perceptual loss, a flow-warping loss, and a world-consistency loss. Except for the world-consistency loss, the others are inherited from the vid2vid [80]. Note that we replace the least square losses used in the vid2vid for GAN losses with the hinge losses as used in SPADE [61]. We describe these terms in details in the following.

**GAN losses.** Let  $\mathbf{s}_1^T \equiv \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_T\}$  be a sequence of input semantic frames. Let  $\mathbf{x}_1^T \equiv \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$  be the sequence of corresponding real video frames, and  $\tilde{\mathbf{x}}_1^T \equiv \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_T\}$  be the synthesized frames by our generator. Define  $(\mathbf{x}_t, \mathbf{s}_t)$  as one pair of frames at a particular time instance where  $\mathbf{x}_t \in \mathbf{x}_1^T$  and  $\mathbf{s}_t \in \mathbf{s}_1^T$ . The image GAN loss ( $\mathcal{L}_I^t$ ) and the video GAN loss ( $\mathcal{L}_V^t$ ) for time  $t$  are then defined as

$$\mathcal{L}_I^t = E_{(\mathbf{x}_t, \mathbf{s}_t)} [\min(0, -1 + D_I(\mathbf{x}_t, \mathbf{s}_t))] + \quad (3)$$

$$E_{(\tilde{\mathbf{x}}_t, \mathbf{s}_t)} [\min(0, -1 - D_I(\tilde{\mathbf{x}}_t, \mathbf{s}_t))] \quad (4)$$

$$\mathcal{L}_V^t = E_{\mathbf{x}_{t-K+1}^t} [\min(0, -1 + D_V(\mathbf{x}_{t-K+1}^t))] + \quad (5)$$

$$E_{\tilde{\mathbf{x}}_{t-K}^{t-1}} [\min(0, -1 - D_V(\tilde{\mathbf{x}}_{t-K}^{t-1}))] \quad (6)$$

where  $D_I$  and  $D_V$  are the image and video discriminators, respectively. The video discriminator takes  $K$  consecutive frames and concatenates them together for discrimination. For both GAN losses, we also accompany them by the feature matching loss ( $\mathcal{L}_{FM}^t$ ) as in pix2pixHD [81],

$$\mathcal{L}_{FM, I/V}^t = \sum_i \frac{1}{P_i} \left[ \|D_{\{I/V\}}^{(i)}(\mathbf{x}_t) - D_{\{I/V\}}^{(i)}(\tilde{\mathbf{x}}_t)\|_1 \right], \quad (7)$$

where  $D_{\{I/V\}}^{(i)}$  denotes the  $i$ -th layer with  $P_i$  elements of the discriminator network  $D_I$  or  $D_V$ .

**Perceptual loss.** We use the VGG-16 network [66] as a feature extractor and minimize L1 losses between the extracted features from the real and the generated images. In particular,

$$\mathcal{L}_P^t = \sum_i \frac{1}{P_i} \left[ \|\psi^{(i)}(\mathbf{x}_t) - \psi^{(i)}(\tilde{\mathbf{x}}_t)\|_1 \right], \quad (8)$$

where  $\psi^{(i)}$  denotes the  $i$ -th layer of the VGG network.

**Flow-warping loss.** We first warp the previous frame to the current frame using optical flow. We then encourage the warped frame to be similar to the current frame by using an L1 loss,

$$\mathcal{L}_F^t = \|\tilde{\mathbf{x}}_t - \mathbf{w}_t(\tilde{\mathbf{x}}_{t-1})\|_1 \quad (9)$$

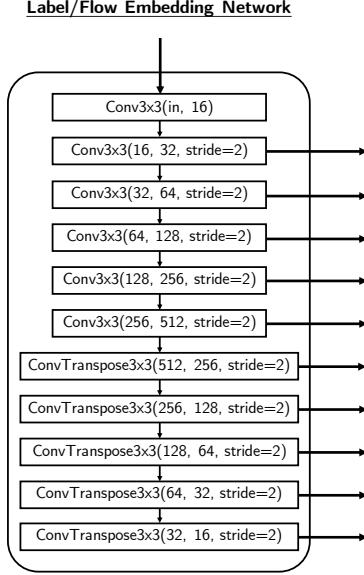


Fig. 9: Label / flow-warped image embedding network.

where  $\mathbf{w}_t$  is the warping function derived from optical flow.

**World-consistency loss.** Finally, we add the world consistency by enforcing the generated image to be similar to our guidance image. It is achieved by

$$\mathcal{L}_{WC}^t = \|\tilde{\mathbf{x}}_t - \tilde{\mathbf{g}}_t\|_1 \quad (10)$$

where  $\tilde{\mathbf{g}}_t$  is our estimated guidance image.

The overall objective function is then

$$\mathcal{L} = \sum_t \min_G \left( \max_{D_I, D_V} (\lambda_I \mathcal{L}_I^t + \lambda_V \mathcal{L}_V^t) \right) + \quad (11)$$

$$\min_G (\lambda_{FM} \mathcal{L}_{FM}^t + \lambda_P \mathcal{L}_P^t + \lambda_F \mathcal{L}_F^t + \lambda_W \mathcal{L}_{WC}^t) \quad (12)$$

where  $\lambda$  are the weights for each individual terms, which are set to 1, 1, 10, 10, 10 in all of our experiments.

**Optimization details.** We use the ADAM optimizer [40] with  $(\beta_1, \beta_2) = 0, 0.999$  for all experiments and network components. We use a learning rate of 1e-4 for the encoder and generator networks (which are described below) and 4e-4 for the discriminators.

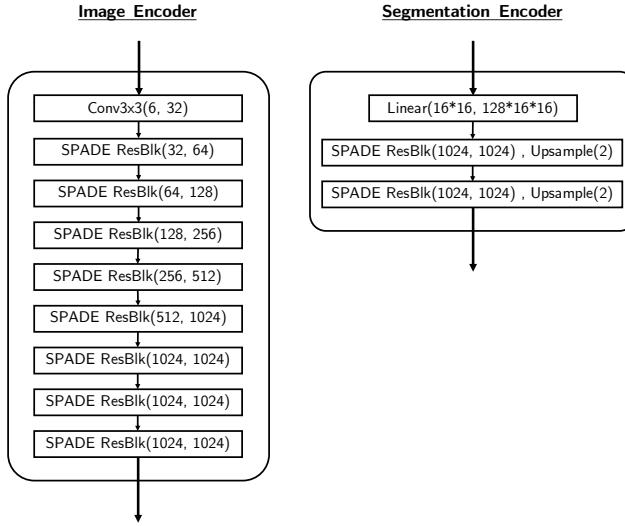


Fig. 10: Previous image / segmentation encoder.

## B Network architecture

As described in the main paper, our framework contains four components: a label embedding network (Fig. 9), an image encoder (Fig. 10), a flow embedding network (Fig. 9), and an image generator (Fig. 11).

**Label embedding network (Fig. 9).** We adopt an encoder-decoder style network to embed the input labels into different feature representations, which are then fed to the Multi-SPADE modules in the image generator.

**Image / segmentation encoder (Fig. 10).** These networks generate the input to the main image generator. The segmentation encoder is used when generating the first frame in the sequence, while the image encoder is used when generating the subsequent frames. The segmentation encoder encodes the input semantics of the first frame, while the image encoder encodes the previously generated frame.

**Flow embedding network (Fig. 9).** It is used to embed the optical flow-warped previous frame, which adopts the same architecture as the label embedding network except for the number of input channels. The embedded features are again fed to the Multi-SPADE layers in the main image generator.

**Image generator (Fig. 11).** The generator consists of a series of Multi-SPADE residual blocks (M-SPADE ResBlks) and upsampling layers. The structure of each M-SPADE Resblk is shown in Fig. 12, which replaces the SPADE layers in the original SPADE Resblks with Multi-SPADE layers.

**Discriminators.** We use the same image and video discriminators as vid2vid [80].

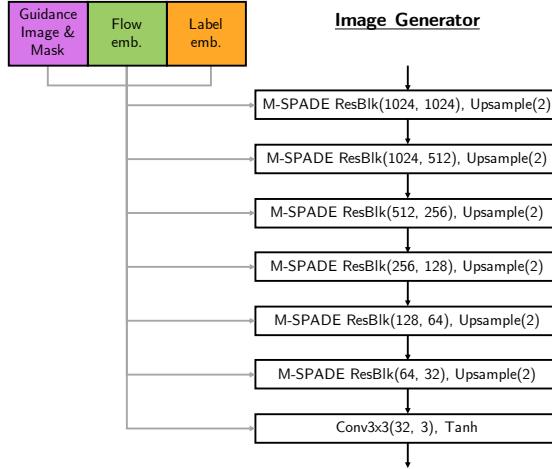


Fig. 11: Image generator.

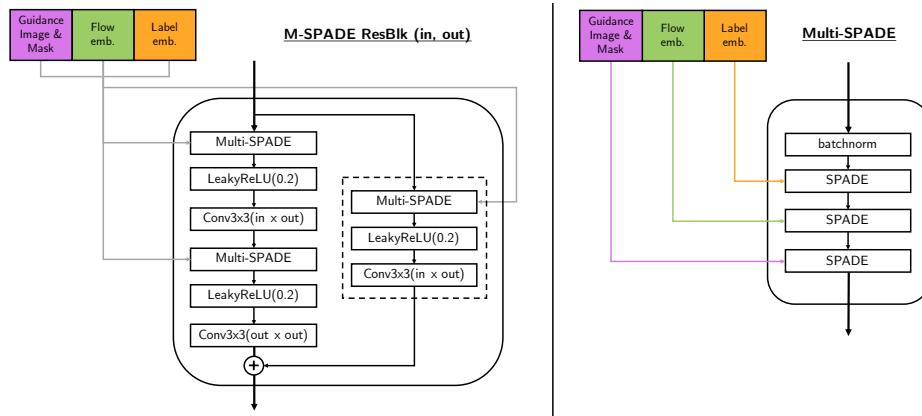


Fig. 12: Multi-SPADE Residual Block and Multi-SPADE module.

## C Additional Results

**Short-term temporal video consistency.** For each sequence, we first take two neighboring frames from the ground truth images to compute the optical flow between them using FlowNet2 [32]. We then use the optical flow to warp the corresponding synthesized images and compute the L1 distance between the warped image and the target image, in RGB space, normalized by the number of pixels and channels. This process is repeated for all pairs of neighboring frames in all sequences and averaged. The result is shown in below in Table 4. As can be seen, Ours w/o World Consistency (W.C.) consistently performs better than

vid2vid [80], and Ours (with world consistency) again consistently outperforms Ours w/o W.C.

Table 4: Short-term temporal consistency scores. Lower is better.

Dataset	vid2vid [80]	Ours w/o W.C.	Ours
Cityscapes	0.0036	0.0032	<b>0.0029</b>
MannequinChallenge	0.0397	0.0319	<b>0.0312</b>
ScanNet	0.0351	0.0278	<b>0.0192</b>