Video Reconstruction with Multimodal Information

Zhipeng Xie*, Yiping Duan*, Qiyuan Du*, Xiaoming Tao*, Jiazhong Yu†

* Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
† China Tower Corporation Limited, Beijing 100084, China

2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall) | 979-8-3503-2928-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/VTC2023-Fall60731.2023.10333705

***Abstract*—Video reconstruction refers to generate videos through the high-level representations (edge map, labels and so on), while the reconstruction quality is always unsatisfactory due to sparse high-level representations, especially on video data. In order to improve the video reconstruction quality, we proposed a novel approach that generates realistic video from its multimodal information including structure features and color features. To extract color features, we mainly apply the k-means algorithm to segment labels and the structure features are extracted by an edge detection network. Video generation is regarded as learning the mapping from multimodal representations to the original videos. So, a conditional GAN is applied with a learning objective that models the temporal video dynamics. We use a spatio-temporal generator with attention to model the inter-frame dynamics and video consistency is improved in this way. Moreover, we use a multiscale discriminator to improve the improve the intra-frame quality of the video. Experimental results on Cityscapes, Apolloscape datasets demonstrate that our proposed approach performs better in both traditional and generative evaluating indicators.**

***Index Terms*—Video reconstruction, multimodal representation, GAN, edge map, color features**

I. Introduction

Multimedia generation plays an important role in semantic communication, which mainly transmit semantic information to save bandwidth and improve service quality in communication process. As one of the most popular multimedia generation methods, video reconstruction is significant to enhance the quality of semantic communication. Video reconstruction aims to generate a realistic sequence of video frames given the semantic information of the original video. Generative models have achieved great advancement in generating realistic images. Unconditional image generation has been mainly realized by generative adversarial networks and can be applied in many fields. Conditional image generation using multimodal information includes generating images using text [1], image inpainting [2] and image-image translation [3], [4]. Ramesh *et al* [1] use a discrete variational autoencoder to get tokens of images and adopt autoregressive transformer to model text and image tokens as a single data stream. Nazeri *et al* [2] proposed an image inpainting model containing two generators.The edge generator completes edges in the missing part and image completion network creates the whole picture using edge as prior. The framework proposed by Hu *et al.* [5] encodes sparse edge maps and color representation to reconstruct realistic faces. All these methods build their own frameworks to reconstruct high visual quality images using multimodal representation. However, video generation has rarely been explored because spatio-temporal consistency is hard to maintain when generating frames, and inspire us to build a generative network that synthesises video through multimodal representation. To generate videos with high visual quality, video reconstruction must not only synthesise realistic frames but also model the temporal dynamics of objects. Thus the generator must be capable of maintaining the consistency of neighboring frames. We adopt the popular transformer structure [6] and flow model into our framework to learn the mapping from the multimodal representation to the original video. The popular transformer block is applied to model the feature correlation of multimodal features and the trained optical flow estimation model [7], [8] is used to model the temporal features and keep consistency maong frames. The color representation [9] corresponds to the style features of image in an explicit way and it can also improve the quality of generation. Thus the channel-wise k-means algorithm is used to segment an image and create its color representation in our model.

Xiaoming Tao is the corresponding author of this paper. Email: taoxm@mail.tsinghua.edu.cn. This work was supported by the National Key R&D Program of China with Grant number 2019YFB1803400, the National Natural Science Foundation of China (Nos. NSFC 61925105, and 62171257).

II. Related works

A. Image-to-image translation

Related works include image-to-image translation, conditional GANs and learning-based video synthesis. Image-to-image translation has made great progress with the development of deep generative models. A conditional GAN makes it possible to reconstruct images using multimodal information. Isola *et al.* [3] proposed the traditional pix2pix model, and it can generate high-visual-quality images with label maps as input, which makes it possible to complete different tasks, such as reconstructing objects with edge maps or colorizing sketches. Models similar to pix2pix need aligned image pairs for training. To adapt to the situation without aligned image pairs, CycleGAN [10] introduced cycle-consistency loss to learn cross-domain mapping through unsupervised training.

B. Conditional GANs

Various GANs make it possible to generate text, image and video with or without input. The input data of a GAN can take various forms, including images, semantic labels and text. Some prior models applied GANs unconditionally for image-to-image mappings, relying on latent space to learn

the mapping from the input domain to the output domain. To generate high-resolution image, Zhang *et al.* [11] proposed a self-attention GAN(SAGAN) using cues from global features to solve the problem that traditional generative models fail to reconstruct objects with more geometric or structural details. MoCoGAN [12] generates video by mapping a series of random vectors to a series of video frames. Each random vector is composed of a content part and a motion part.

C. Video generation

Video generation model not only needs to generate objects with good visual quality but also needs to learn the motion pattern of objects; otherwise the generated video is spatially temporally discontinuous. The temporal generative adversarial network(TGAN) [13] can learn the semantic representation of unlabeled video datasets. Its generation part consists of two generators, a temporal generator and an image generator. Vondrick *et al.* [14] proposed the video generator network(VGAN) that divides a video into foreground and background such that the generation of moving foreground and static background are decoupled. There are two independent data streams in VGAN framework: one is the moving foreground path of hierarchical space-time convolution, and the other is the static background path of hierarchical spatial convolution, both of which are upsampled.The two data streams and the mask of the motion path are combined to create the generated video.

III. VIDEO GENERATION BY MULTIMODAL FEATURES

A. Multimodal representation

1) Structure features extraction: The edge map mainly contains structure information and has been used in image synthesis. The pix2pix model can reconstruct shoes from their edge maps [3] and take only the edges as the input of its generative network, which provides the possibility of generating frames using edge maps. Sparse edges only contain the main contours of objects in images and lose texture information, while dense edge maps consist more detailed texture information of realistic scene. Traditional edge detectors mainly use pixel gradients to capture image brightness changes, while CNNs can extract features of different levels. In our proposed model, we adopt the PiDiNet to obtain dense edge maps because the PiDiNet architecture [15] captures more accurate and detailed edges by combining convolution networks and traditional edge detectors. A comparison between sparse edge and dense edge maps can be found in the experimental section.

2) Color representation: In adapt to different scenes, image segmentation has different scales, from coarse-grained street segmentation to fine-grained face segmentation, varying in semantic meaning. The most common segmentation approach labels each pixel according to its natural object class, and it has been applied in many image generating models [3]. However, color features cannot be maintained after object-based segmentation, and they correspond to the style of an image and often impact visual quality. The k-means algorithm can cluster pixels on a fine-grained level and extract pixel-pixel similarity of color domain. We apply the k-means algorithm as our segmentation method to obtain segmentation of video frame. To the final color representation. As a result, we combine the color features and edge map as the input of the generators.

Algorithm 1 Image segmentation

Initialize:

For a image consisting of N pixels, the values of each pixel is normalized to 0-1.
Let M be the collection of centers and randomly put a pixel into it.
Let the size of cluster centers be r.
while $r < k$ **do**
  define a candidate center set C
  Let sum_c be the sum of distances between the pixels in C and M.
  for each m_i in M **do**
    Let $dist_{m_i}$ be the minimum distance between m_i and other pixels.
    $dist_{m_i} \leftarrow inf$
    $c_i \leftarrow -1$
    for $j = 0$ to $N - 1$ **do**
      calculate the Euclidean distance $d_{m_i,j}$ of pixel i and j
      if $d_{m_i,j} < dist_{m_i}$ **then**
        $dist_{m_i} \leftarrow d_{m_i,j}$
        $c_i \leftarrow j$
      end if
    end for
    $sum_c \leftarrow sum_c + dist_{m_i}$
    if $i == r - 1$ **then**
      $sum_c \leftarrow sum_c * p$
    end if
    join c_i into C.
  end for
  for each c_i in C **do**
    $sum_c \leftarrow sum_c - dist_{m_i}$
    if $sum_c < 0$ **then**
      join c_i into M.
    end if
  end for
end while

B. Model architecture

Similar to image generation, video reconstruction aims to generate videos using different information provided by the original videos. However, due to the continuity among video frames, simply applying an image generative model in video generation may cause inconsistency and scene distortion. To solve the problem, we adopt temporal generator and multiscale discriminator to improve visual quality and generate temporally coherent video. The goal of our network is to find a mapping from multimodal representation of an original video to itself.

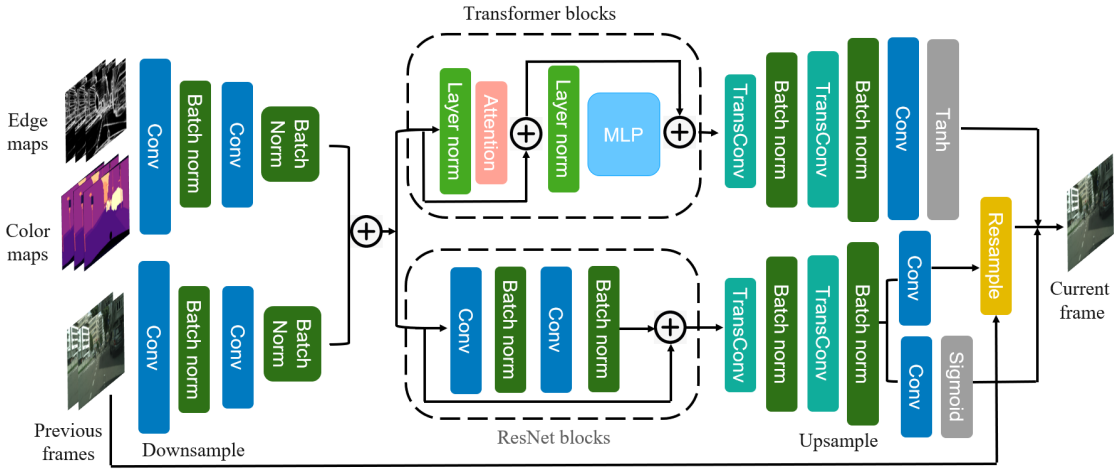

Fig. 1. The framework of video frame generator. The transformer blocks are adopted to model the correlation of structure features and color features, represented by edge and color maps respectively. When generating the current frames, both the previously generated frames and the input maps are combined to model the temporal features to generate consecutive video.

In our model architecture, we use conditional GAN to learn the mapping. Let G denote the generator function of our conditional GAN and D denote the discriminator function. Let $\mathbf{S} = \{s_1, s_2, ...s_n\}$ be a sequence of real continuous video frames, and $\mathbf{X} = \{x_1, x_2, x_3, ..., x_n\}$ be the sequence of the multimodal maps corresponding to the original frames. The output of our model is an generated sequence $\mathbf{Y} = \{y_1, y_2, y_3, ..., y_n\}$ given $\mathbf{X}$. To make the output $\mathbf{Y}$ as similar to $\mathbf{S}$ as possible, the objective of our model is to train a GAN that can learn the mapping from $\mathbf{X}$ to $\mathbf{S}$, which means that the conditional distribution $p(\mathbf{Y}|\mathbf{X})$ matches the conditional distribution of $p(\mathbf{S}|\mathbf{X})$. To find the objective mapping, we convert it to optimizing the min-max problem as follows:

$$\max_D \min_G E_{(\mathbf{S},\mathbf{X})}[\log D(\mathbf{S},\mathbf{X})] + E_{(\mathbf{Y},\mathbf{X})}[\log(1 - D(\mathbf{Y},\mathbf{X}))] \tag{1}$$

In comparison with image generation, the output of the video generative model must be temporally coherent, so the generator cannot take only the multimodal representation of the current frame into account. To simplify the problem, a Markov assumption is built to model the temporal relation of coherent frames in the same video, which means that when generating the current i-th frame, the network also takes the $L-1$ previously generated frames and L input frames as input:

$$y_i = G(\{x_{i-L}, x_{i-L+1}, ..., x_i\}, \{y_{i-L}, y_{i-L+1}, ..., y_{i-1}\}) \tag{2}$$

For the objective loss function, we use GAN loss and feature matching loss to improve the quality of generation. When designing the generator, we input the sum of the features of the L input maps and $L-1$ previously generated frames into 4 transformer blocks instead of the original ResNet blocks in the vid2vid model. Let D_I denote the conditional image discriminator and D_V be the video discriminator. The objective function to be optimized can be written as follows:

$$\min_G(\max_{D_I} \mathcal{L}_\mathcal{I}(G, D_I) + \max_{D_V} \mathcal{L}_\mathcal{V}(G, D_V)) + \lambda \mathcal{L}_\mathcal{W}(G) \tag{3}$$

where $\mathcal{L}_\mathcal{I}$ is the conditional GAN loss with respect to image generation. The patch resampling ϕ is applied during calculation, and $\mathcal{L}_\mathcal{I}$ can be defined as:

$$\mathcal{L}_\mathcal{I} = E_{\phi(\mathbf{S},\mathbf{X})}[\log D_I(\mathbf{S},\mathbf{X})] + E_{\phi(\mathbf{Y},\mathbf{X})}[\log(1 - D_I(\mathbf{Y},\mathbf{X}))] \tag{4}$$

$\mathcal{L}_\mathcal{V}$ represents the conditional GAN loss with respect to video synthesis, and its definition is similar to the image conditional GAN loss with estimated optical flow. The optical flow of consecutive frames is important for generating temporally coherent frames in video synthesis, so we use the traditional optical flow estimation model [7] and the flow estimation function is denoted as F. The previous frame is warped by the flow to generate the current frame using resampling operation denoted as R. $\mathcal{L}_\mathcal{F}$ is the flow estimation loss and defined as:

$$\mathcal{L}_\mathcal{F} = \frac{1}{n-1} \sum_{i=1}^{n-1} (\|F(y_i, y_{i+1}) - F(s_i, s_{i+1})\| + \|R[F(y_i, y_{i+1}), s_i] - s_{i+1}\|) \tag{5}$$

To stabilize adversarial training, the VGG loss denoted as $\mathcal{L}_{\text{VGG}}$ is added to the learning objective and it is defined as:

$$\mathcal{L}_{\text{VGG}} = \sum_i \frac{1}{P_i} (\|\psi^i(s) - \psi^i(y)\|) \tag{6}$$

where ψ^i denotes the i-th layer with P_i elements of the VGG network.

Fig. 2. The visualized comparison of our method and other video synthesis models. It shows that the results produced by vid2vid model have color distortion. Fast-vid2vid performs better than vid2vid model. However, it produces texture distortion when generating objects like cars. The results generated by our method have better visual quality compared with the original video sequence.

TABLE I
IMAGE RECONSTRUCTION RESULTS WITH DATA FROM DIFFERENT SINGLE MODALITY.

dataset	input	PSNR	SSIM	MSE
Cityscapes	edge map	117.7	0.68	537
	semantic label	112.6	0.5	1869
	sketch	114.3	0.566	1242
	color domain	123.3	0.73	232
Apolloscape	edge map	93.2	0.49	1786
	semantic label	86.9	0.31	2343
	sketch	89.4	0.45	1956
	color domain	101.9	0.53	697

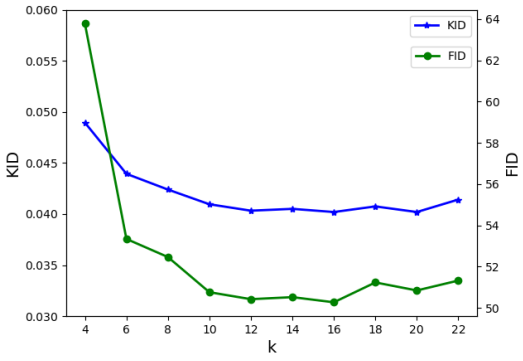

Fig. 3. The k has effects on the granularity of color feature map that is significant for generation. Excessive k can create more detailed color feature map but it may cause over-fitting.

IV. EXPERIMENTS

In our experiments, multiple quantitative evaluations are used to verify the effectiveness of the proposed model. To better evaluate our method, we also applied Learned Perceptual Image Patch Similarity(LPIPS) [16], Fréchet Inception Distance(FID) [17] and Kernel Inception Distance(KID) [18] that are widely used in generative models to evaluate perceptual quality . We used 1886 videos (89250 images) from the Cityscapes [19] dataset for training and evaluation on 4 NVIDIA RTX 3090 GPUs. To further evaluate our method, the Apolloscape dataset [20] consisting of 915 videos is also used with data augmentation including spatial reversal. The objective function is optimized using Adam with exponentially weighted iterate averaging.

TABLE II
VIDEO RECONSTRUCTION RESULTS ON CITYSCAPES DATASET.

method	PSNR	SSIM	FID	KID
vid2vid	16.997	0.494	77.892	0.0593 ± 0.0032
fast-vid2vid	14.356	0.450	69.352	0.0623 ± 0.0029
pix2pixHD	19.375	0.494	205.168	0.2411± 0.0082
ours without color	21.916	0.630	66.574	0.0343 ± 0.0039
ours	23.182	0.700	53.341	0.0439 ± 0.0026

To compare the effects of different modals when generating objects, we conducted experiments using only the image generation part on Apolloscape and Cityscapes datasets: we take labels, sketches, edge maps and color domains as the input of the image generator and the results indicated that the color domain is the most useful one and it results from that the visual feature is significant when generating complex objects. The reason why the result of edge map has higher SSIM than that of sketch is that multilevel structure information is more effective when producing changeable scenes with many details. The label modal has the worst result is because it only contains the pixel meaning on objects level. The label maps are created by the popular semantic segmentation method Deeplabv3 [21]. In terms of quantitative evaluation, we also

compared with other models including vid2vid [22], fast-vid2vid [23] and other video generation models and the results can be found in the table 2 and it showed that our model performs better.

To evaluate out framework, we evaluate the quality based on a perceptual approach for the generated videos, and reflect the overall quality of the videos with the average LPIPS/FID/KID metrics of the synthesized results. We also evaluate the performances on traditional pixel-level metrics like PSNR and SSIM for more analysis of the strengths and weaknesses of our method. It can be seen from Fig.2 that the output of our model has a better visual quality. The granularity of color representation depends on the k of segmentation algorithm. Its impact on generation results is shown in Fig.3.

The results in Table II demonstrate that our model can reconstruct more realistic video, and the ablation study about color representation shows that color features can improve it. The reason that the standard deviation of KID of pix2pixHD model is higher is because it is a image generation model, and could not keep the consistency among the generated frames. However, one of the disadvantage of our framework is that the output cannot be manipulated by human. In the generation model [4], objects in semantic label map can be replaced by others to create more diverse results.

V. CONCLUSIONS

The final goal of our model is to train a generative network that can generate a temporally coherent video using multimodal features. The structure information represented by the edge map is extracted by deep network, and the improved image segmentation method with k-means algorithm is applied to segment images to obtain the color representation of image. To choose the best representation of video frame, we conducted experiments on image generation and found the color feature and edge map have the best potential when synthesizing changeable scenes with many details and objects with complicated texture. To better model the correlation of different modals, the transformer structure is adopted in our framework. The existing image generating networks often output incoherent frames from a given sequence of images because the temporal dynamics of the source video are not considered in the learning objective. To generate video from multimodal representations, both the spatio-temporal generator and multiscale discriminator are applied in the generation part to learn the mapping from the multimodal domain to the original domain. Results of quantitative evaluation and visualized comparison show that our model generated higher visual-quality videos compared to other video generating models.

REFERENCES

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[2] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[3] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[4] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Y. Hu, S. Yang, W. Yang, L. Duan, and J. Liu, "Towards coding for human and machine vision: A scalable image coding approach," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[7] F. Reda, R. Pottorff, J. Barker, and B. Catanzaro, "flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks," https://github.com/NVIDIA/flownet2-pytorch, 2017.

[8] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1647–1655.

[9] S. You, N. You, and M. Pan, "Pi-rec: Progressive image reconstruction network with edge and color domain," *arXiv preprint arXiv:1903.10146*, 2019.

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, 2017.

[11] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[12] S. Tulyakov, M. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1526–1535.

[13] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2830–2839.

[14] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *Advances in neural information processing systems*, vol. 29, 2016.

[15] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu, "Pixel difference networks for efficient edge detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5097–5107.

[16] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018.

[17] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[18] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[19] 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.

[20] Miao Liao, Feixiang Lu, Dingfu Zhou, Sibo Zhang, Wei Li, and Ruigang Yang, "Dvi: Depth guided video inpainting for autonomous driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.

[21] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[22] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[23] Long Zhuo, Guangcong Wang, Shikai Li, Wanye Wu, and Ziwei Liu, "Fast-vid2vid: Spatial-temporal compression for video-to-video synthesis," in *European Conference on Computer Vision (ECCV)*, 2022.