CS 562 Presentation

Kung-Hsiang (Steeve) Huang

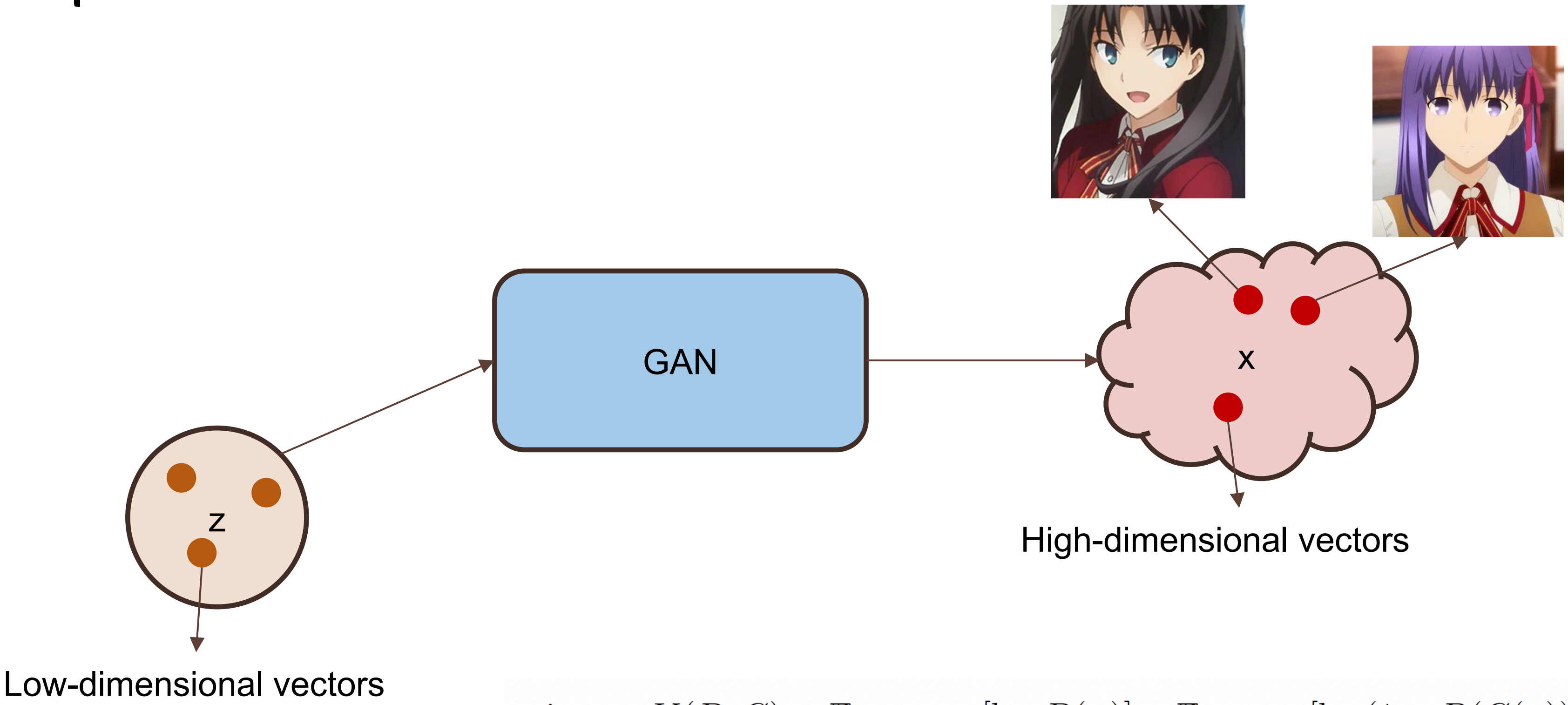
Papers to discuss

- Conditional Generative Adversarial Nets
- Video-to-Video Synthesis

Conditional Generative Adversarial Nets

Mehdi Mirza, Simon Osindero

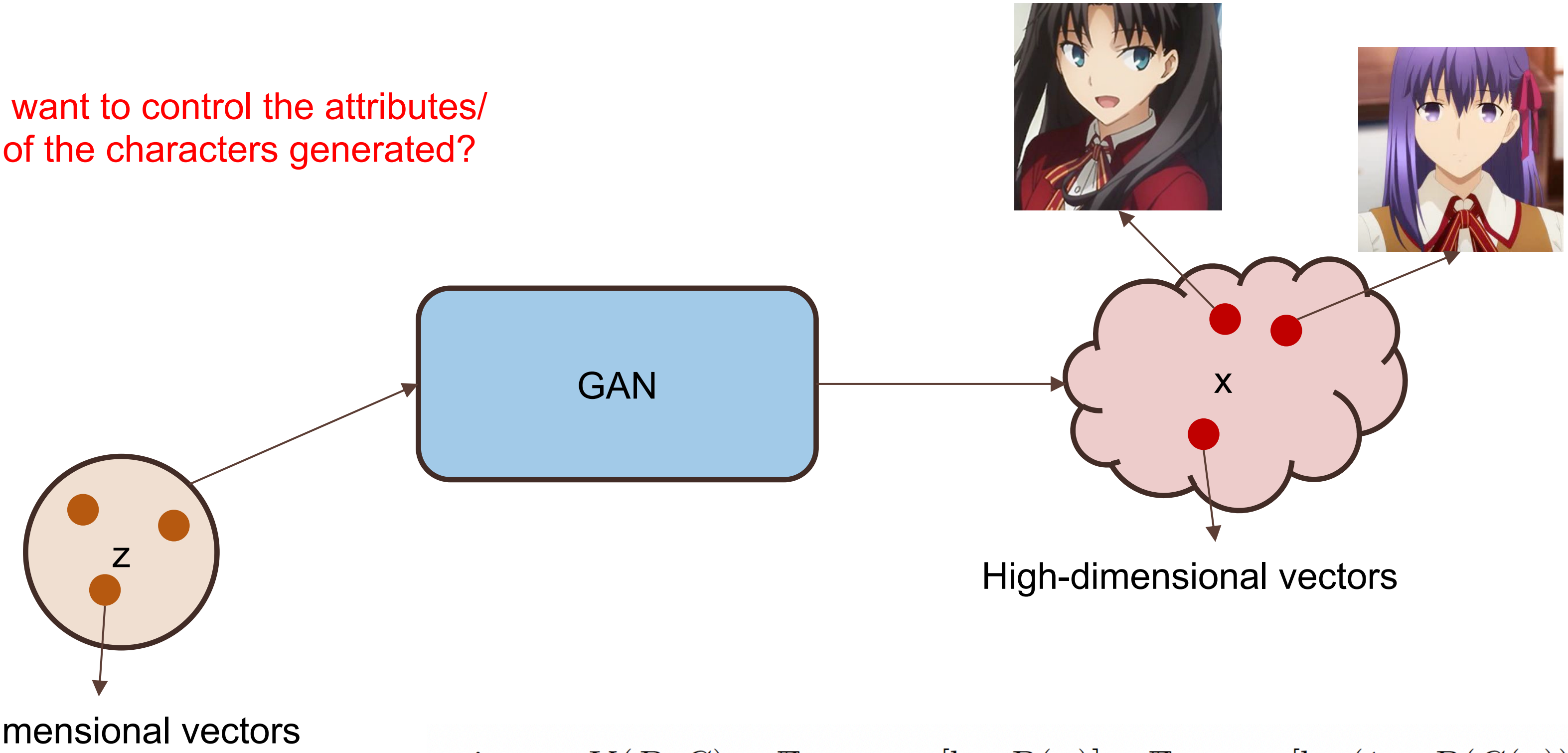
Recap: Unconditional GAN



$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

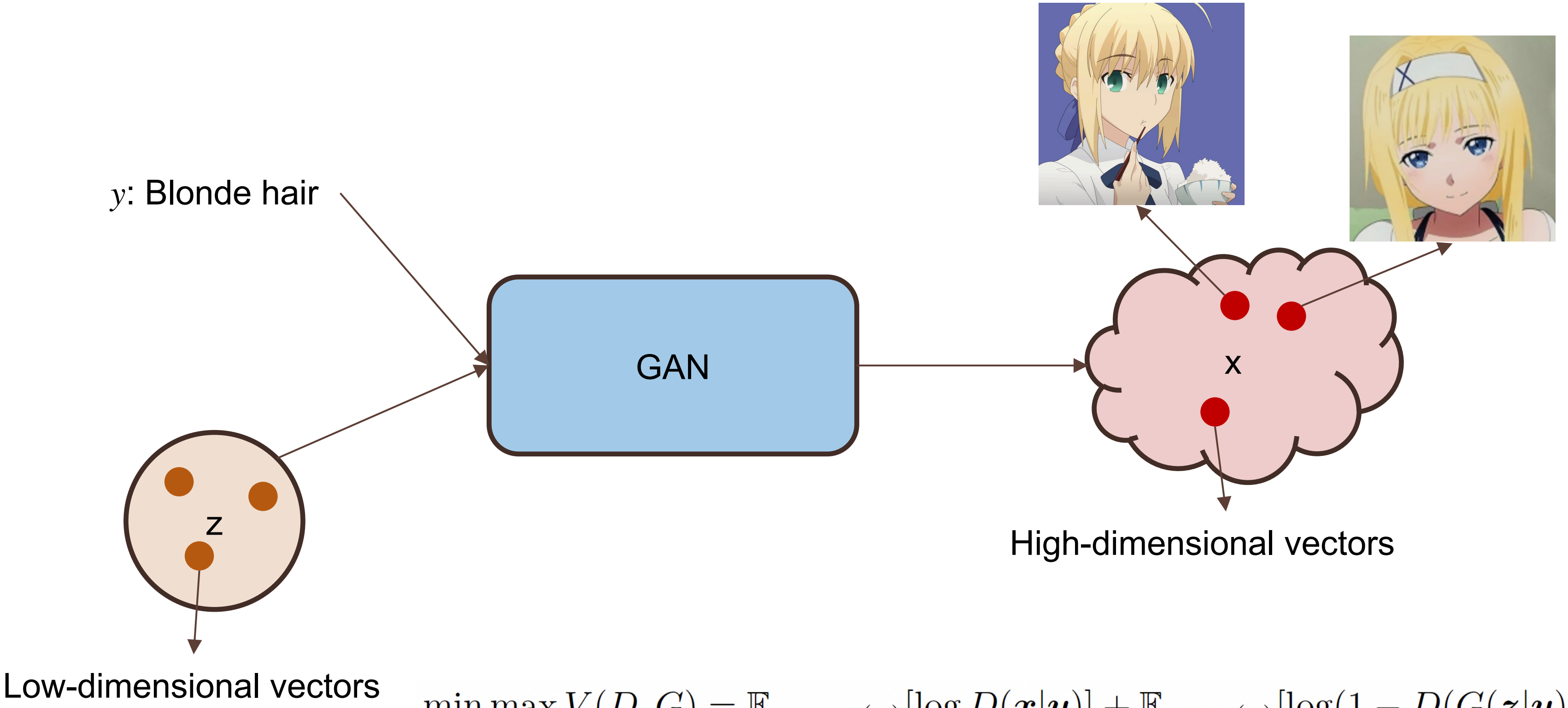
Recap: Unconditional GAN

What if we want to control the attributes/ properties of the characters generated?



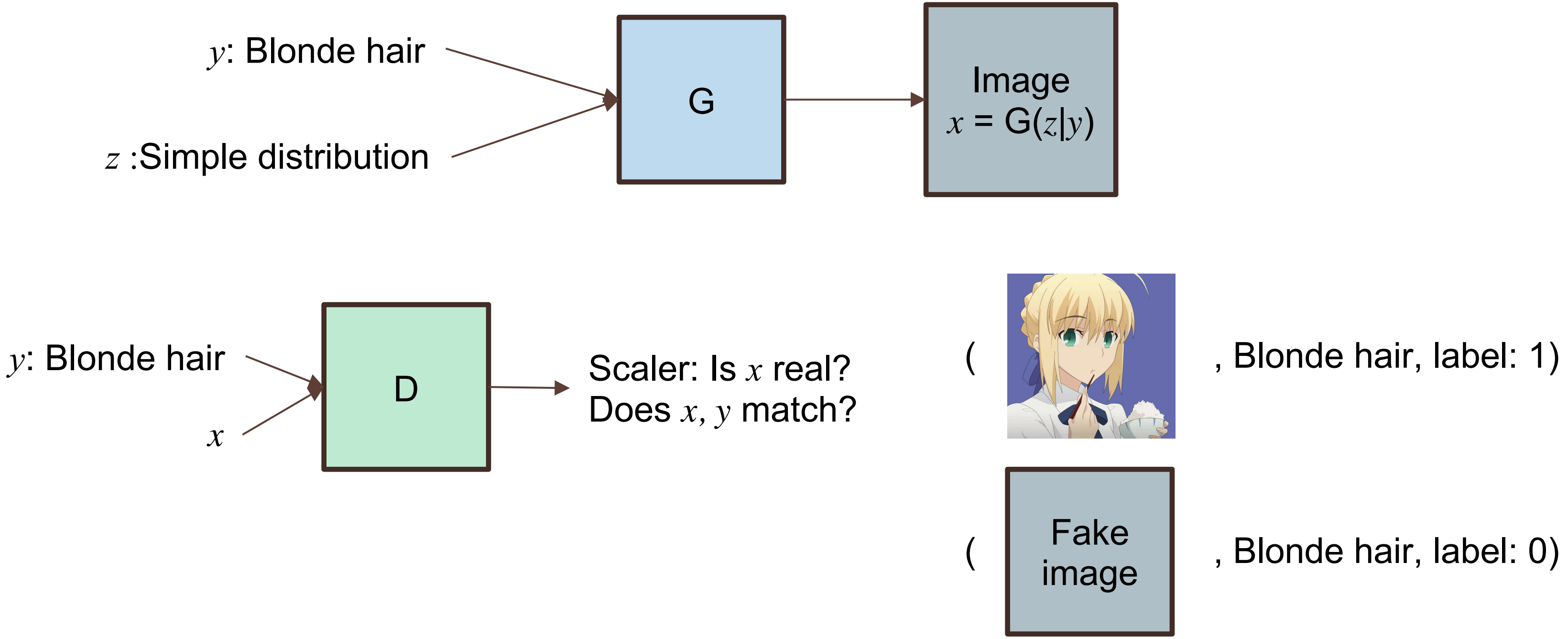
$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Conditional GAN



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))]$$

Training a conditional GAN

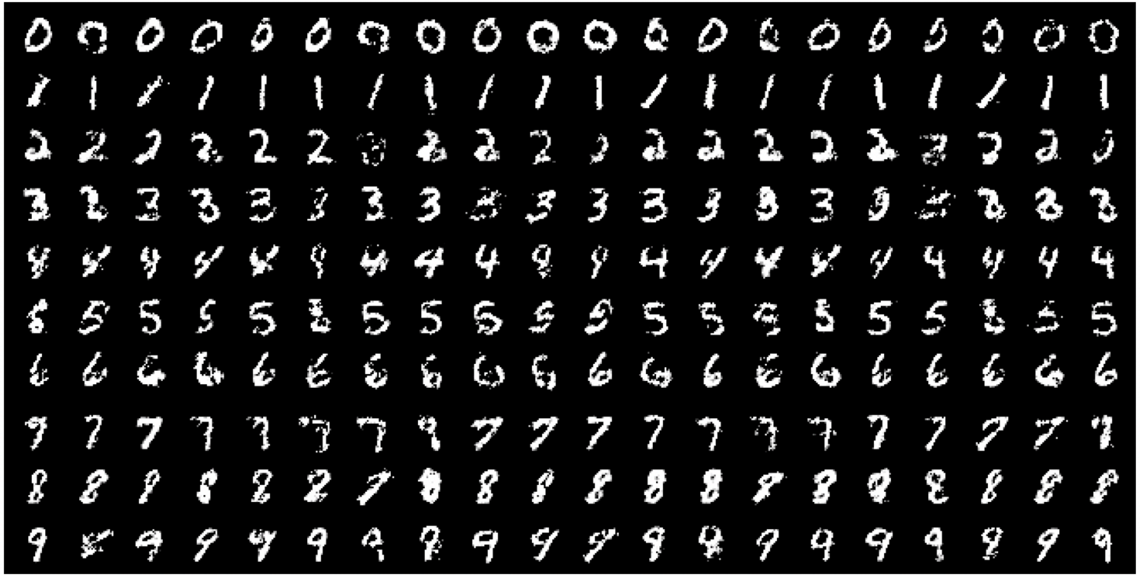


Quantitative results

Model	MNIST
DBN [1]	138 ± 2
Stacked CAE [1]	121 ± 1.6
Deep GSN [2]	214 ± 1.1
Adversarial nets	225 ± 2
Conditional adversarial nets	132 ± 1.8

Table 1: Parzen window-based log-likelihood estimates for MNIST. We followed the same procedure as [8] for computing these values.

Qualitative results



Generated MNIST digits

Limitation

y: Blonde hair

z :Simple distribution

G

Image
x = G(z|y)

y: Blonde hair

x

D

Scaler: Is x real?
Does x, y match?

( , Blonde hair, label: 1)

(Fake image , Blonde hair, label: 0)

Limitation

y: Blonde hair

z :Simple distribution

G

Image
x = G(z|y)

y: Blonde hair

x

D

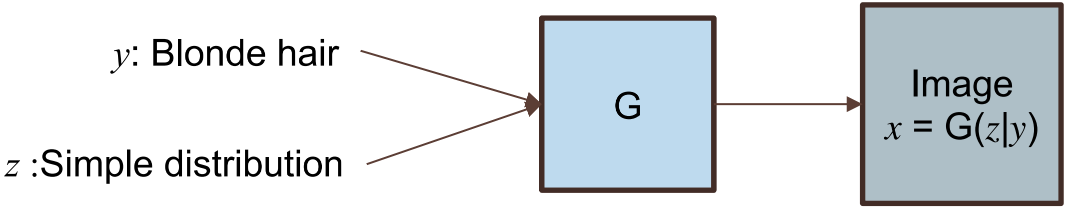
Scaler: Is x real?
Does x, y match?

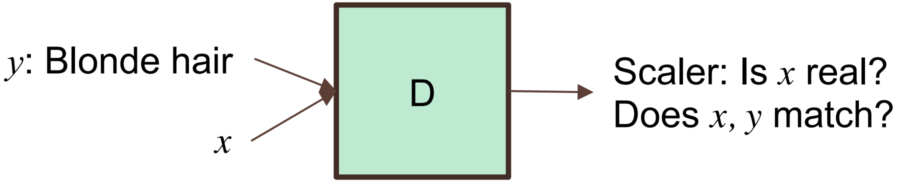
The generator and discriminator might not check if x, y match.

( , Blonde hair, label: 1)

(Fake image , Blonde hair, label: 0)

Limitation





( , Blonde hair, label: 1)

( , Blonde hair, label: 0)

(Fake image , Blonde hair, label: 0)

Applying (Conditional) GAN to NLP

- Issue: Textual outputs are discrete and non-differentiable.
- Solutions:
  - REINFORCE algorithm [1]
  - Gumbel-Softmax relaxation [2]

[1] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. EMNLP 2017
[2] Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. ICLR 2017.

Takeaway

- The paper presents a simple yet effective extension towards controllability of GAN by feeding additional information to it.
- There is still a great room for improvement for synthesizing images using conditional GAN.

Video-to-Video Synthesis

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu,
Andrew Tao, Jan Kautz, Bryan Catanzaro

Video-to-video synthesis

Main challenge: Temporal coherence

Problem formulation

The model is tasked to map from a given a sequence of video frames

$$\mathbf{s}_1^T \equiv \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T\}$$

to a target sequence of video frames

$$\tilde{\mathbf{x}}_1^T \equiv \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T\}$$

such that

$$p(\tilde{\mathbf{x}}_1^T | \mathbf{s}_1^T) = p(\mathbf{x}_1^T | \mathbf{s}_1^T)$$

Problem formulation

The model is tasked to map from a given a sequence of video frames

$$\mathbf{s}_1^T \equiv \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T\}$$

Segmentation mask

to a target sequence of video frames

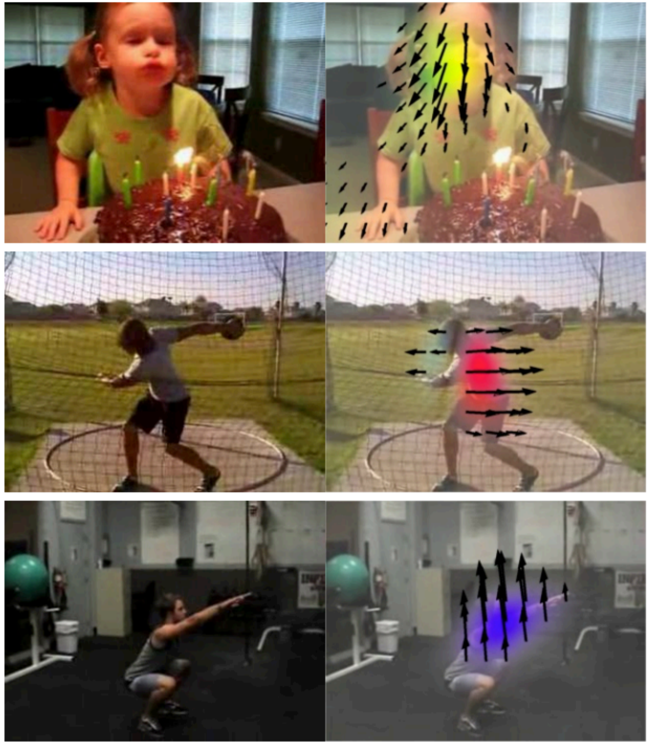
$$\tilde{\mathbf{x}}_1^T \equiv \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, ..., \tilde{\mathbf{x}}_T\}$$

such that

$$p(\tilde{\mathbf{x}}_1^T | \mathbf{s}_1^T) = p(\mathbf{x}_1^T | \mathbf{s}_1^T)$$

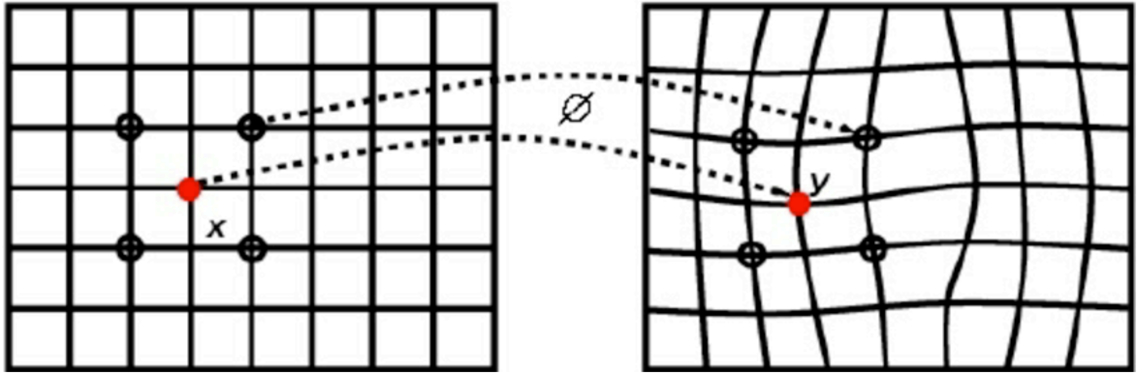
Real video clip

Optical flow



Walker, J., Gupta, A., & Hebert, M. (2015). Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision*

Image warping



Gilles, J., Dagobert, T., & De Franchis, C. (2008, October). Atmospheric turbulence restoration by diffeomorphic image registration and blind deconvolution. In International Conference on Advanced Concepts for Intelligent Vision Systems (pp. 400-409). Springer, Berlin, Heidelberg.

Approach

- Sequential generator: generate future frames
- Conditional image discriminator: ensure each frame is photorealistic
- Conditional video discriminator: ensure temporal consistency

Sequential generator

- Markov assumption

$$p(\tilde{\mathbf{x}}_1^T|\mathbf{s}_1^T) = \prod_{t=1}^{T} p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t})$$

- Use a feed-forward network *F* to approximate $p(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t})$

$$F(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}) = (\mathbf{1} - \tilde{\mathbf{m}}_t) \odot \tilde{\mathbf{w}}_{t-1}(\tilde{\mathbf{x}}_{t-1}) + \tilde{\mathbf{m}}_t \odot \tilde{\mathbf{h}}_t$$

  - Estimated optical flow $\tilde{\mathbf{w}}_{t-1} = W(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t})$
  - Hallucinated image $\tilde{\mathbf{h}}_t = H(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t})$
  - Occlusion mask $\tilde{\mathbf{m}}_t = M(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t})$

Conditional image discriminator

- Distinguish true pairs $(\mathbf{x}_t, \mathbf{s}_t)$ from fake $(\tilde{\mathbf{x}}_t, \mathbf{s}_t)$.
- Objective function:

$$\mathcal{L}_I = E_{\phi_I(\mathbf{x}_1^T, \mathbf{s}_1^T)}[\log D_I(\mathbf{x}_i, \mathbf{s}_i)] + E_{\phi_I(\tilde{\mathbf{x}}_1^T, \mathbf{s}_1^T)}[\log(1 - D_I(\tilde{\mathbf{x}}_i, \mathbf{s}_i))].$$

Conditional video discriminator

- Enhance temporal consistency by ensuring consecutive output frames resemble that of real frames **given the gold optical flow**.
- Distinguish true pairs $(\mathbf{x}_{t-K}^{t-1}, \mathbf{w}_{t-K}^{t-2})$ from fake $(\tilde{\mathbf{x}}_{t-K}^{t-1}, \mathbf{w}_{t-K}^{t-2})$.
- Objective function:

$$\mathcal{L}_V = E_{\phi_V(\mathbf{w}_1^{T-1}, \mathbf{x}_1^T, \mathbf{s}_1^T)}[\log D_V(\mathbf{x}_{i-K}^{i-1}, \mathbf{w}_{i-K}^{i-2})] + E_{\phi_V(\mathbf{w}_1^{T-1}, \tilde{\mathbf{x}}_1^T, \mathbf{s}_1^T)}[\log(1 - D_V(\tilde{\mathbf{x}}_{i-K}^{i-1}, \mathbf{w}_{i-K}^{i-2}))]$$

Final objective function

$$\min_{F}\left(\max_{D_I}\mathcal{L}_I(F, D_I) + \max_{D_V}\mathcal{L}_V(F, D_V)\right) + \lambda_W\mathcal{L}_W(F)$$

$$\mathcal{L}_W = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_1 + \|\tilde{\mathbf{w}}_t(\mathbf{x}_t) - \mathbf{x}_{t+1}\|_1\right)$$

Quantitative Results

Table 1: Comparison between competing video-to-video synthesis approaches on Cityscapes.

Fréchet Inception Dist.	I3D	ResNeXt
pix2pixHD	5.57	0.18
COVST	5.55	0.18
vid2vid (ours)	**4.66**	**0.15**

Human Preference Score	short seq.	long seq.
vid2vid (ours) / pix2pixHD	**0.87** / 0.13	**0.83** / 0.17
vid2vid (ours) / COVST	**0.84** / 0.16	**0.80** / 0.20

Quantitative Results

Table 2: Ablation study. We compare the proposed approach to its three variants.

	Human Preference Score
vid2vid (ours) / no background-foreground prior	**0.80** / 0.20
vid2vid (ours) / no conditional video discriminator	**0.84** / 0.16
vid2vid (ours) / no flow warping	**0.67** / 0.33

Quantitative Results

Table 3: Comparison between future video prediction methods on Cityscapes.

Fréchet Inception Dist.	I3D	ResNeXt
PredNet	11.18	0.59
MCNet	10.00	0.43
`vid2vid` (ours)	**3.44**	**0.18**

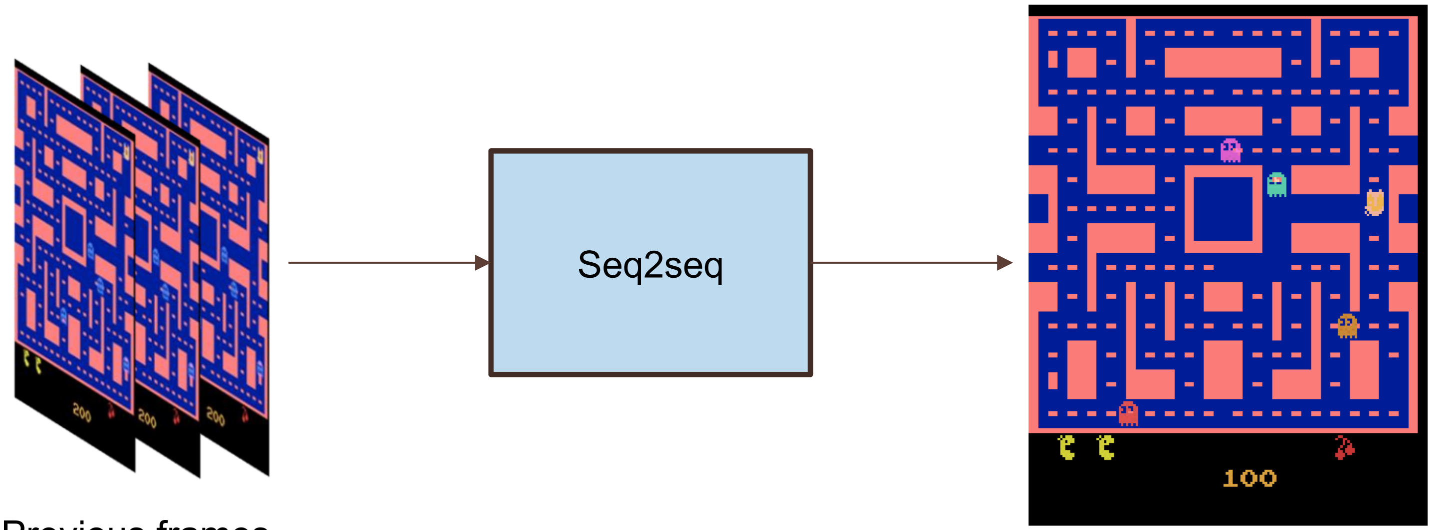
Human Preference Score			
`vid2vid` (ours)	/	PredNet	**0.92** / 0.08
`vid2vid` (ours)	/	MCNet	**0.98** / 0.02

Qualitative Results

Figure 2: Apolloscape results. Left: `pix2pixHD`. Center: `COVST`. Right: proposed. The input semantic segmentation mask video is shown in the left video. *Click the image to play the video clip in a browser.*

Why not regular seq2seq models?

A toy example

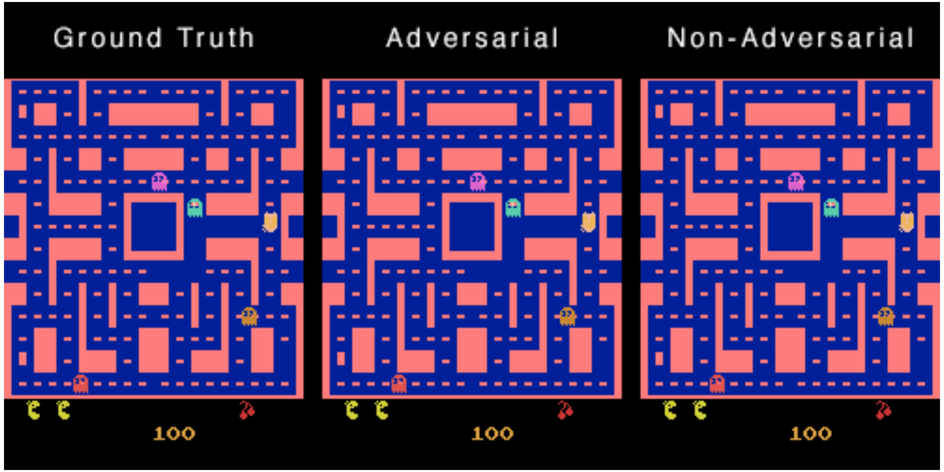


Previous frames

Next frame

Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *ICLR 2016*

A toy example



Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *ICLR 2016*

Takeaway

- Enforcing temporal consistency is key to high quality video synthesis.
- By adding a video conditional discriminator, the work successfully shows improvement in temporal consistency of the videos generated.
- The proposed approach can be applied to various tasks, such as future video prediction.

References

- https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/gan_v10.pdf