Machine Learning Video Upscaling

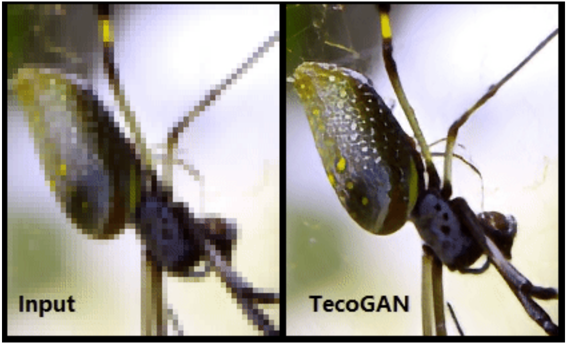
Taylor Carrington

Computer Science Senior Seminar
Division of Science and Mathematics
University of Minnesota, Morris

April 2021

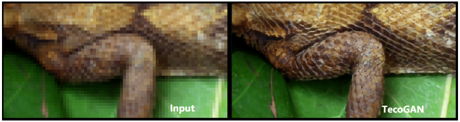
Introduction

- Video Upscaling
- TecoGAN



[1]

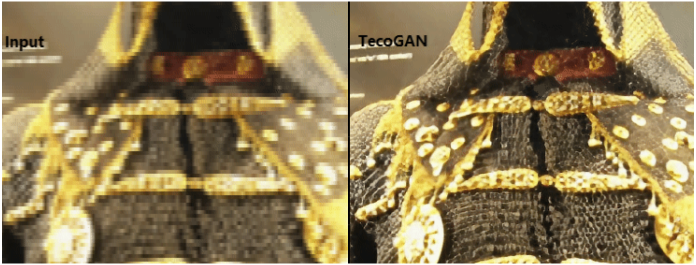
Table of Contents



[1]

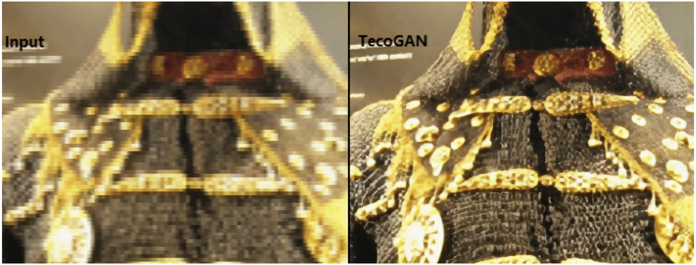
Background

- Videos
  - Video is split into frames, which are single images
  - Frames per Second (FPS): Video is made up of a variety of FPS but usually 24 for films.
    - 90 minutes film has 129,600 frames
  - Resolution: Low resolution (LR) is any video that is under 720p while high resolution (HR) is equal to or greater than that.



[1]

Background



[1]

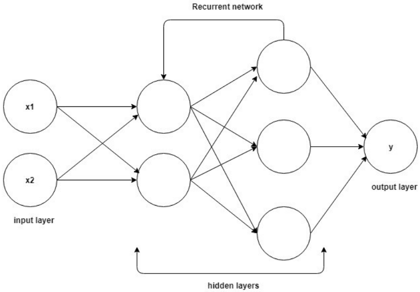
- Upscaling (VSR): Upscaling is taking a LR image or video and generating a HR image or video.
  - Spatial consistency: The objects in the frames stays the same
  - Temporal consistency: The motion between objects stay the same between frames.
  - Ground Truth (GT): Original HR image
  - Artifacts: Errors in the generated output that were not in the input

Background

- Machine Learning: Artificial intelligence which through the use of data and time improves the accuracy of a computer algorithm.
  - Supervised learning: The algorithm learns a way to match inputs with outputs by having training data and target data.

Background

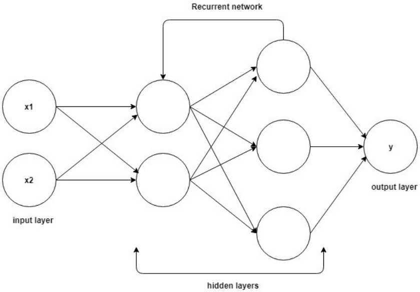
- Neural Networks (NN)
  - Training: Takes inputs and matches them to outputs through learning a data set.
  - Weights: Each edge has a weight which multiples the input data by a specific number.
  - Input data: LR frames
  - Target data: HR frames
  - Output data: Generated HR frames
  - Recurrent NN: Allows for nodes to send data back to a previous node.



[2]

Background

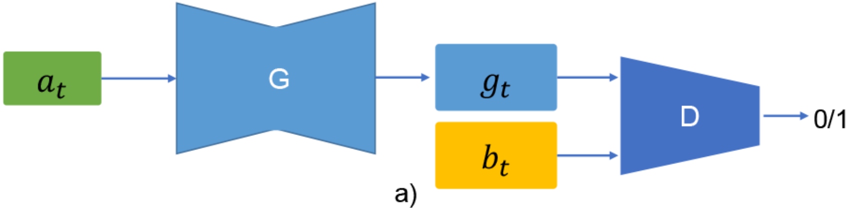
- Loss Terms
  - Used to evaluate the networks
  - Wants the lowest scores
  - Therefore the algorithm changes the weights number to improve the overall system



[2]

Background

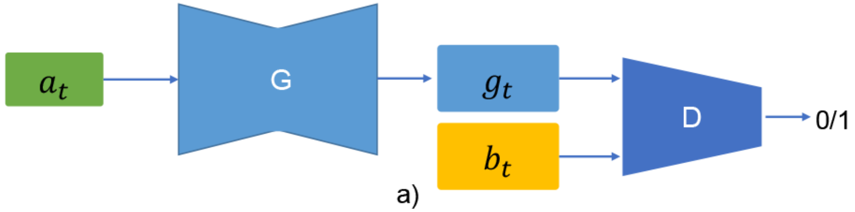
- Deep Learning: More advance than machine learning, uses multiple layers to mimic an human brain.
- Generative Adversarial Network (GANs)
  - Uses Two NN
  - Zero-Sum Game
  - Goal of tricking the Discriminator into thinking the generated data is the target data



[1]

Short Term Consistency

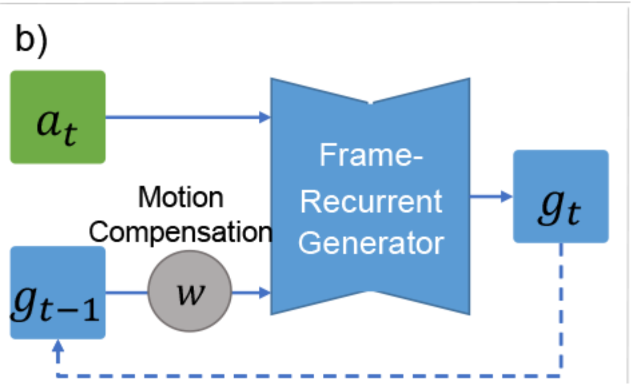
- a) Spatial GAN for image generation



[1]

Short Term Consistency

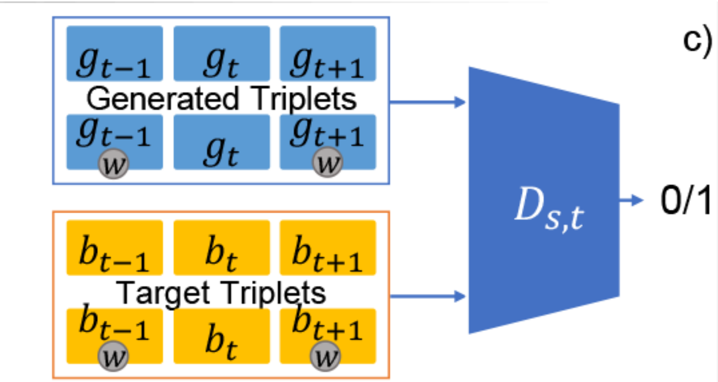
- b) Frame Recurrent Generator



[1]

Short Term Consistency

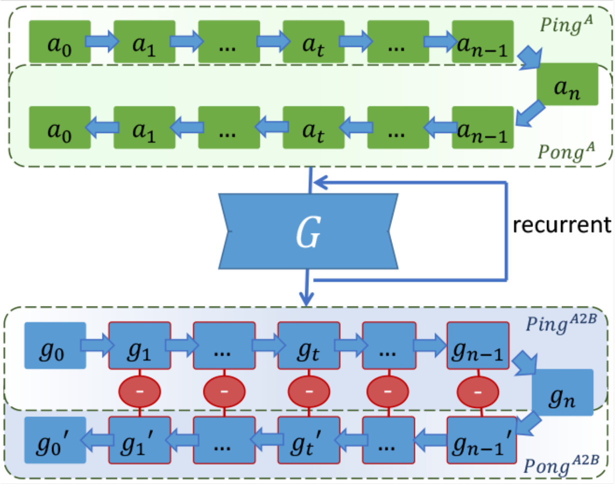
- c) Spatio-temporal Discriminator



[1]

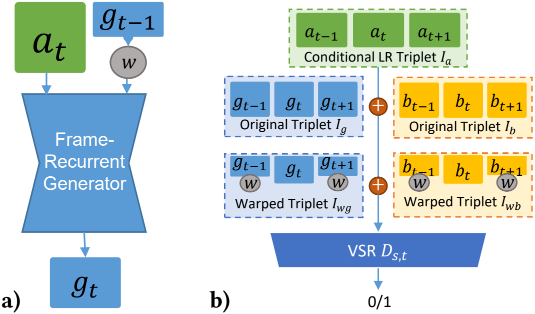
Long Term Consistency

- Ping-Pong Loss



[1]

Network Architecture for VSR



[1]

VSR Loss Term

- $\mathcal{L}_{G,F} = \lambda_w \mathcal{L}_{warp} + \lambda_a \mathcal{L}_{adv} + \lambda_\phi \mathcal{L}_\phi + \lambda_c \mathcal{L}_{content} + \lambda_p \mathcal{L}_{PP}$
    - $\mathcal{L}_{warp}$ is the warping loss which measures the difference between the input frame and the previous input frame.

TecoGAN - Method

VSR Loss Term

- $\mathcal{L}_{G,F} = \lambda_w \mathcal{L}_{warp} + \lambda_a \mathcal{L}_{adv} + \lambda_\phi \mathcal{L}_\phi + \lambda_c \mathcal{L}_{content} + \lambda_p \mathcal{L}_{PP}$
  - $\mathcal{L}_{adv}$ is the adversarial loss which measures how well the discriminator at judging the generated data.

TecoGAN - Method

VSR Loss Term

- $\mathcal{L}_{G,F} = \lambda_w \mathcal{L}_{warp} + \lambda_a \mathcal{L}_{adv} + \lambda_\phi \mathcal{L}_\phi + \lambda_c \mathcal{L}_{content} + \lambda_p \mathcal{L}_{PP}$
    - $\mathcal{L}_\phi$ is the perceptual loss which measures if the specific objects from the target triplets show up in the generated triplets.

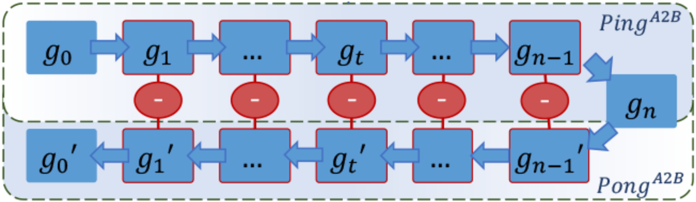
TecoGAN - Method

VSR Loss Term

- $\mathcal{L}_{G,F} = \lambda_w \mathcal{L}_{warp} + \lambda_a \mathcal{L}_{adv} + \lambda_\phi \mathcal{L}_\phi + \lambda_c \mathcal{L}_{content} + \lambda_p \mathcal{L}_{PP}$
  - $\mathcal{L}_{content}$ is the content loss which measures the difference between the generated frame and the target frame.

VSR Loss Term

- $\mathcal{L}_{G,F} = \lambda_w \mathcal{L}_{warp} + \lambda_a \mathcal{L}_{adv} + \lambda_\phi \mathcal{L}_\phi + \lambda_c \mathcal{L}_{content} + \lambda_p \mathcal{L}_{PP}$
  - $\mathcal{L}_{PP}$ is the "Ping Pong" loss
  - $\mathcal{L}_{PP} = \Sigma_{t=1}^{n-1} \|g_t - g_t'\|_2$
  - PP loss is the summation from frame (t) equals one to frame n-1 of the L2 loss of the forward generated frame (g_t) minus the reverse generated frame (g_t')



[1]

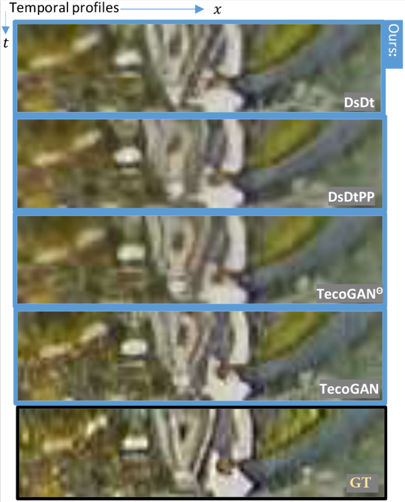
TecoGAN - Loss Ablation Study

Loss Ablation Study

- The study of an AI system that gets its components stripped down before each are adding back one by one
- With the goal of better understand how each component adds to the overall system's.

Loss Ablation Study

- DsOnly
- DsDt
- DsDtPP
- *TecoGAN*$^{\ominus}$
- TecoGAN



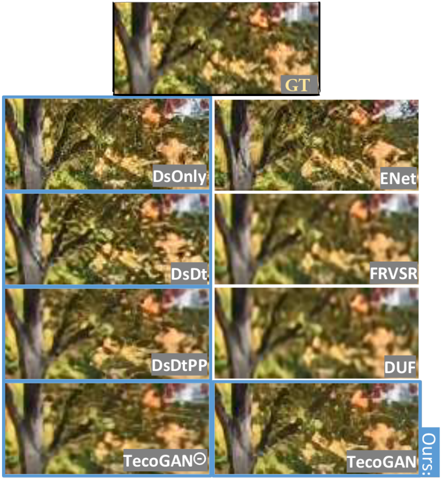
[1]

TecoGAN - Loss Ablation Study

Other Methods that TecoGAN is tested against are:

- ENet: Upscales images only, does not pay attention to temporal changes
- FRVSR: Upscales videos, does not have adversarial loss
- DUF: Also upscales videos, does not have adversarial loss

All are compared to the GT

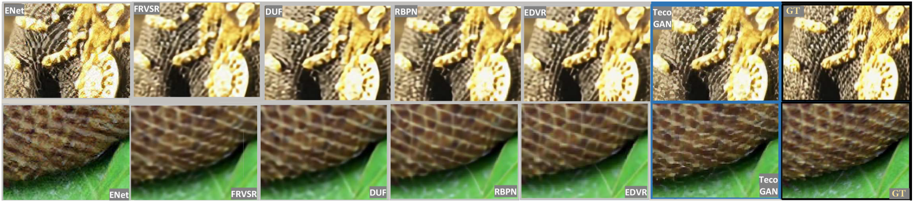


[1]

TecoGAN - Metrics Evaluation

More Methods that TecoGAN is tested against

- TecoGAN: 3 million weights
- RBPN: 20 million weights
- EDVR: 12 million weights



[1]

TecoGAN - Results

- LPIPS: Perceptional Distance to the GT
- tOF: Pixel-wise distances of estimated motion
- tLP: Perceptional Distance of consecutive frames

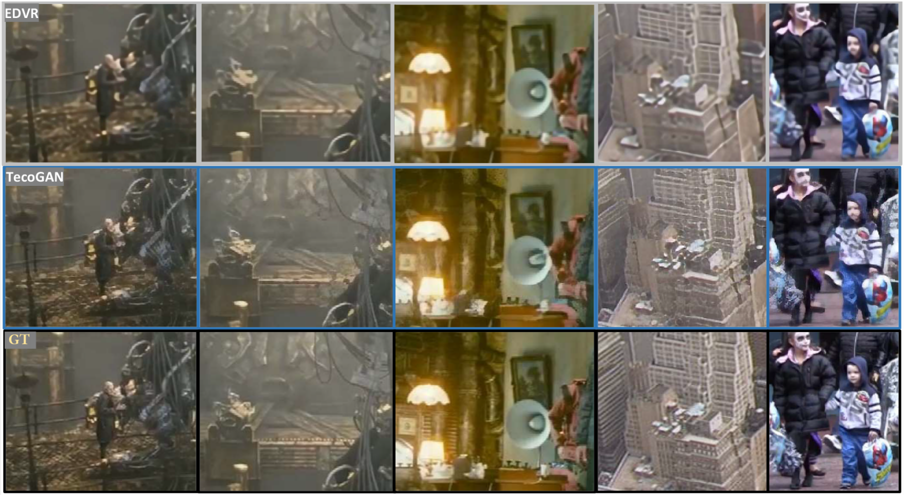
Methods	**LPIPS↓ ×10**	**tOF↓ ×10**	**tLP↓ ×100**
TecoGAN	**1.623**	1.897	**0.668**
ENet	2.458	4.009	4.848
FRVSR	2.506	2.090	0.957
DUF	2.607	1.588	1.329
RBPN	2.511	1.473	0.911
EDVR	2.356	**1.367**	0.982

TecoGAN - Results

- PSNR: Pixel-Wise Accuracy
- User Study: 50 participants who made 1000 votes
- Processing Time: How long each low resolution frame took to be upscaled

Methods	PSNR↑	User Study↑	Processing Time↓ (ms/frame)	PT for 90 minutes film (HR)↓
TecoGAN	25.57	**3.258**	41.92	1.5
ENet	22.31	1.616	-	-
FRVSR	26.91	2.600	36.95	1.33
DUF	**27.38**	2.933	942.21	33.92
RBPN	27.15	-	510.90	18.39
EDVR	27.34	-	299.71	10.79

Conclusion



[1]

Acknowledgements

- Senior Seminar Advisor: Nic McPhee
- Senior Seminar Professor: Elena Machkasova
- External Reviewer: Paul Friederichsen

Questions?

Questions?

Bibliography

Mengyu Chu et al. "Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation". In: *ACM Trans. Graph.* 39.4 (July 2020). ISSN: 0730-0301. DOI: 10.1145/3386569.3392457. URL: https://doi.org/10.1145/3386569.3392457.

Debarko. *RNN or Recurrent Neural Network for Noobs*. [Online; accessed 9-April-2021]. 2018. URL: https://hackernoon.com/rnn-or-recurrent-neural-network-for-noobs-a9afbb00e860.