Video Super-Resolution using Diffusion Probabilistic Models

Dipesh Lamichhane (076MSIISE008)

Supervisor
Er. Dinesh Baniya Kshatri

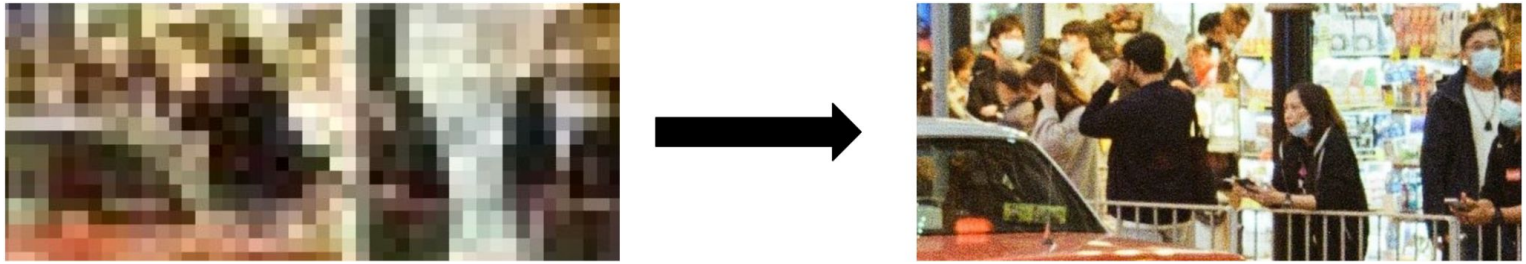
Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

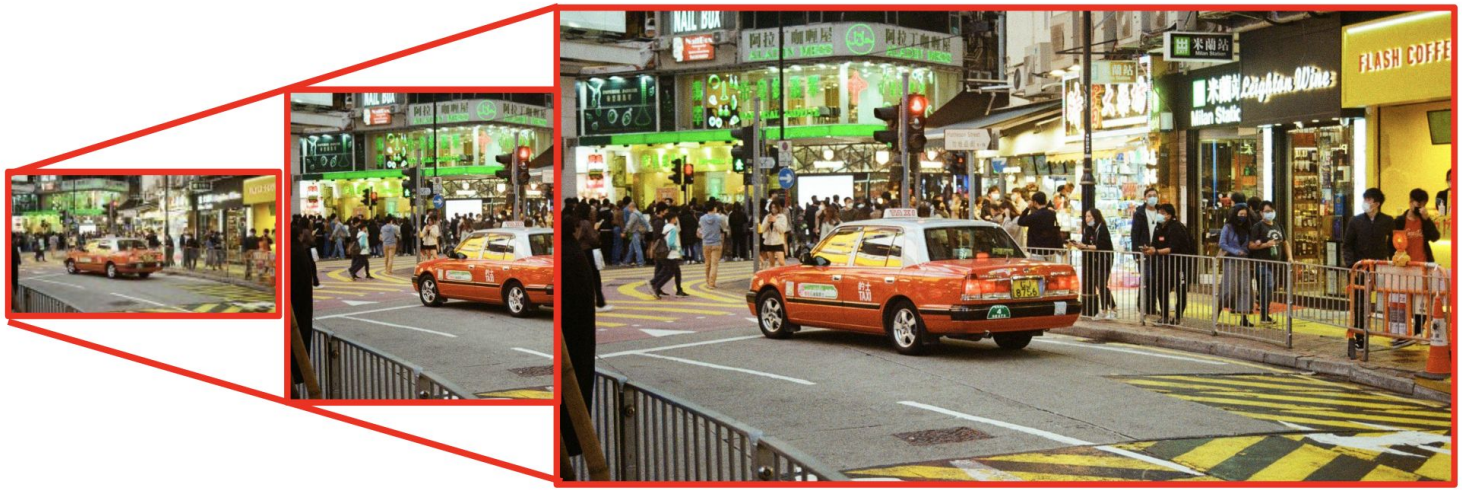
April 1, 2022

Presentation Outline

- Motivation
- Problem Definition
- Objectives
- Scope of Project
- Originality of Project
- Project Applications
- Literature Review
- Methodology
- Results
- Discussion and Analysis
- Future Enhancements
- Conclusion

Motivation





Problem Definition

- GANs are often difficult to train to convergence.
  - Undergo mode-collapse without very careful considerations.

- GANs capture less diversity than state-of-the-art likelihood-based models.

- Likelihood-based models are easier to train, but fall short in terms of visual quality.

- Flow-based methods require careful pre-processing of data before training and inference.

Objectives

- To enhance frame quality as well as remove motion-blur in videos using Diffusion Probabilistic Models.

- To compare the results with state-of-the-art Generative Adversarial Networks applied to Video Super-Resolution.

Scope of Project

- Project Capabilities:
  - Application of probabilistic models to video super-resolution.
  - Reduction of motion blur between output video frames.
  - Production of more detailed high-resolution videos with smooth flow.

- Project Limitations:
  - The model does not provide real-time capabilities.
  - Resulting model is not applicable to abrupt scene changes and cuts in videos.

Originality of Project

- Diffusion models are being actively developed for single image generation.
- Videos consist of separate image frames.
- The project does not apply to only individual frames.
- The model considers multiple neighboring temporal frames to reduce motion blur.

Project Applications

- Medical Imaging
  - Production of clearer results allowing improved and efficient diagnoses.
- Self-driving Vehicles
  - Processing videos with more details facilitate better control of autonomous vehicles.
- Drone Footage
  - Applying video super-resolution can produce high-quality footage from drones.
- Video Streaming
  - Streaming smaller LR videos at high bitrate and post-processing to high resolutions.
- Gaming Performance
  - Faster low-resolution renders can be super-resolved to high fidelity video frames, improving performance.

Literature Review - [1] (Based on Base Papers)

GANS (iSeeBetter) [1]	Diffusion Models
Use Generator that de-convolves from latent feature space to image space.	Use Generator that reverses a diffusion process that gradually adds noise to data.
Based on optimizing two competing neural networks, the discriminator, and generator.	Based on the idea of non-equilibrium thermodynamics.
Use four-fold loss function to improve high frequency results.	Loss functions incorporate minimizing likelihoods on conditional diffusion steps.
Can undergo mode-collapse when used without proper hyperparameters.	More stable than GANs in terms of training.

Literature Review - [2]
(Based on Base Papers)

- Super-Resolution using Repeated Refinement (SR3) [2]
  - Adapts denoising diffusion models to conditional single image generation.
  - Conditioning also includes variance, allowing fewer diffusion steps to achieve competent results.

- Cascaded Diffusion Models (CDMs) [3]
  - Conditional augmentation is essential for quality in diffusion model cascades.
  - Achieves better quality results with pure generative models that are not combined with any classifier, unlike GANs.

Methodology - [1]
(Theoretical Formulations - Diffusion Process)

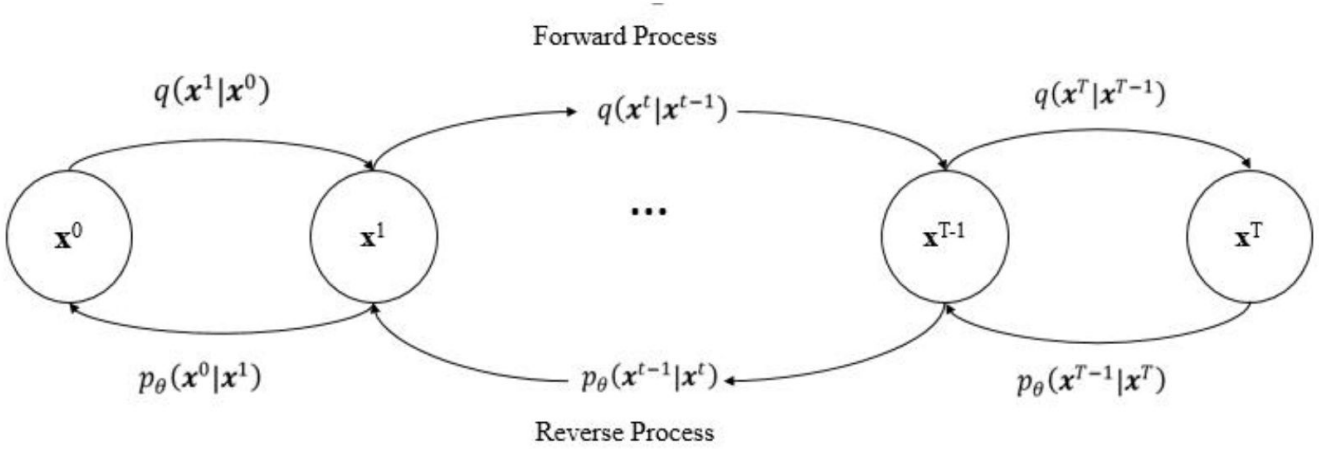
- Data distribution $\mathbf{x}^0$ undergoes small noise addition steps in forward process.
- These steps can be considered to form a Markov Chain.
  - Data at any time-step differs only slightly from the previous one.
  - It is only influenced by the previous time-step.
- Each forward time-step can be represented as,

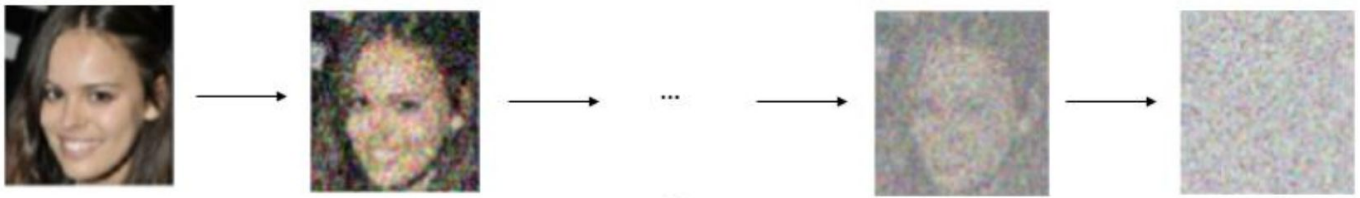
$$q(\boldsymbol{x}^t|\boldsymbol{x}^{t-1}) = T_\pi(\boldsymbol{x}^t|\boldsymbol{x}^{t-1};\beta)$$

  for a Markov diffusion kernel T_π, and diffusion rate β

- The model is trained to reverse the trajectory of this process.

Methodology - [2]
(Theoretical Formulations - Diffusion Processes)

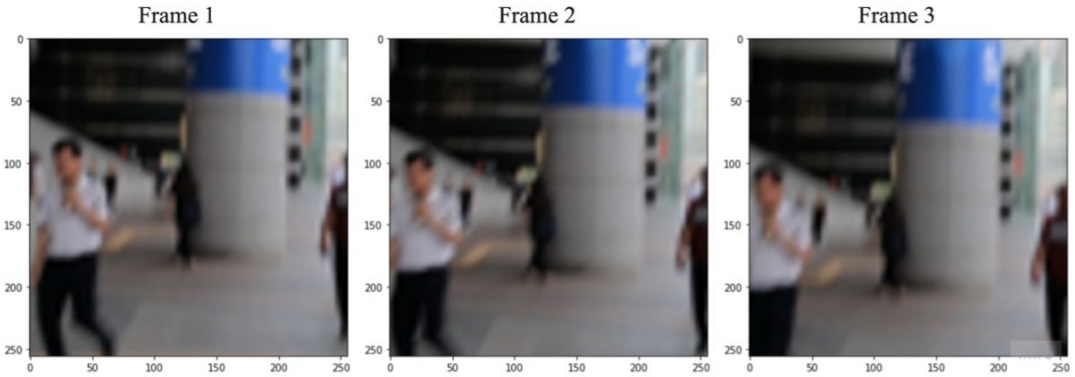




Methodology - [3]
(Theoretical Formulations - Quantifying Motion Blur)

- Motion blur causes smooth intensity along the direction of motion.
- Estimating areas with lowest directional high-frequency energy allows detection of motion blur.
- Can be quantified using reconstructed frame removing low frequency components in the frequency domain.

$$\text{FFT Blur} = \frac{1}{N \times N} \sum_{x,y} 20 \log \hat{f}(x,y)$$



Methodology - [4]
(Theoretical Formulations - Video Quality Assessment)

- Peak Signal-to-Noise Ratio (PSNR)
  - Measures maximum possible power of an image frame compared to the power of distortion that affects the quality of a restored image frame.

- Structural Similarity Index Measure (SSIM)
  - Provides a similarity measure between original and reconstructed image frames.
  - Incorporates Luminance, Contrast, and Structure parameters.

- Normalized Mean-Square Error (NMSE)
  - Measures the squared error of distortion between video frames.

Methodology - [5]
(Mathematical Modeling)

- A LR video frame can be considered a degradation of a HR frame,

$$I_i = \phi(\hat{I}_i, \{\hat{I}_j\}_{j=i-N_1}^{i+N_2}; \theta_\alpha)$$

$\{\hat{I}_j\}_{j=i-N_1}^{i+N_2}$ denotes the set of LR frames that influence the HR frame.

N_1 and N_2 denote the number of preceding and succeeding frames.

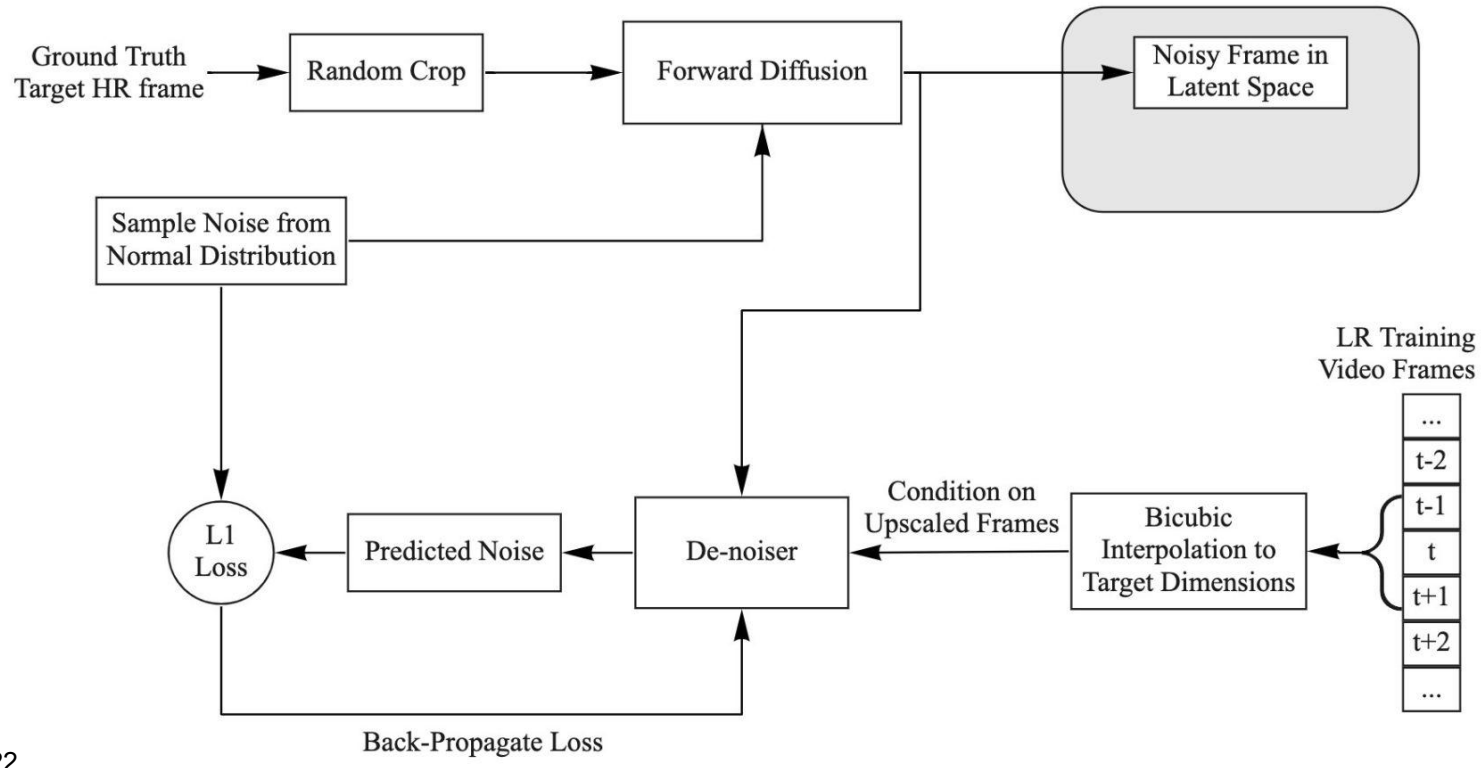
θ_α denotes the parameters of degradation.

- So, the video super-resolution process is the inverse of this degradation process.

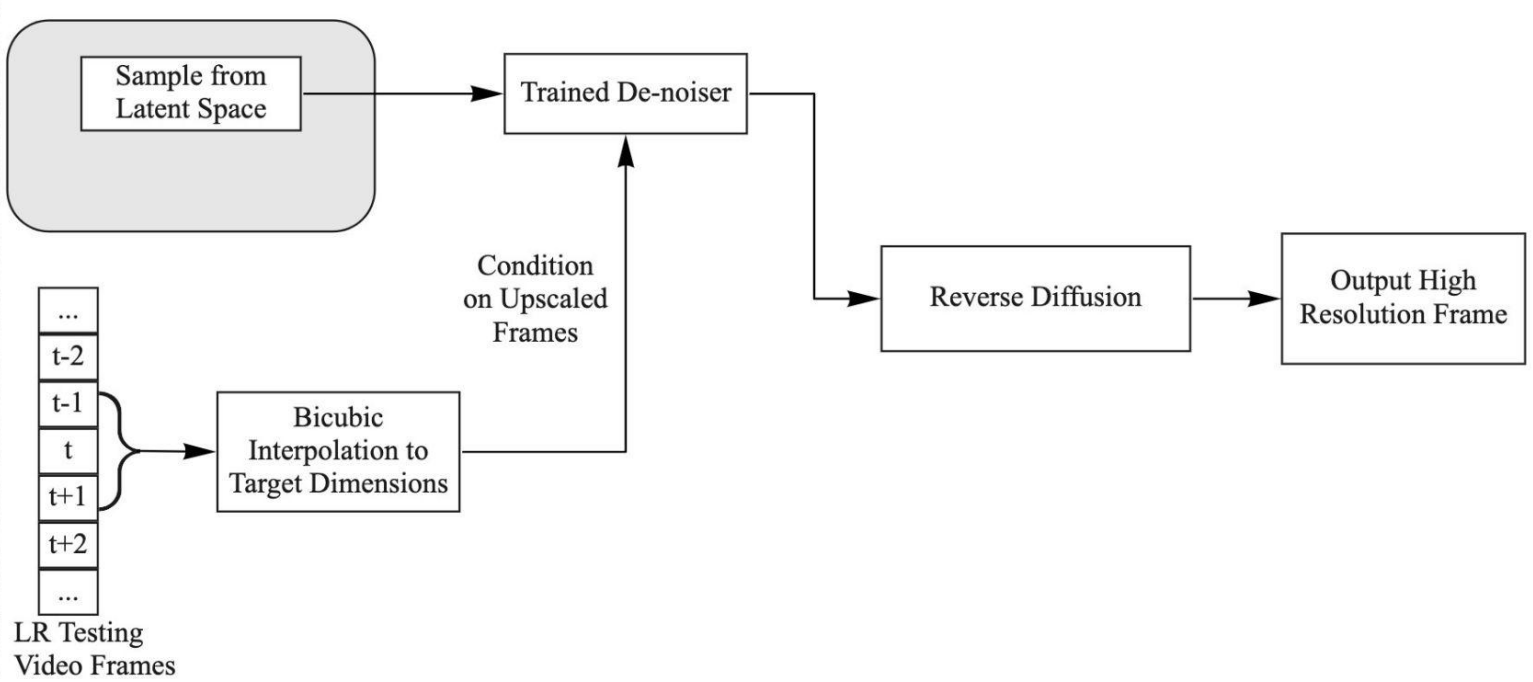
$$\hat{I}_i = \phi^{-1}(I_i, \{I_j\}_{j=i-N_1}^{i+N_2}; \theta_\beta)$$

θ_β contains the parameters for super-resolution.

Methodology - [6]
(System Block Diagram - Training Phase)



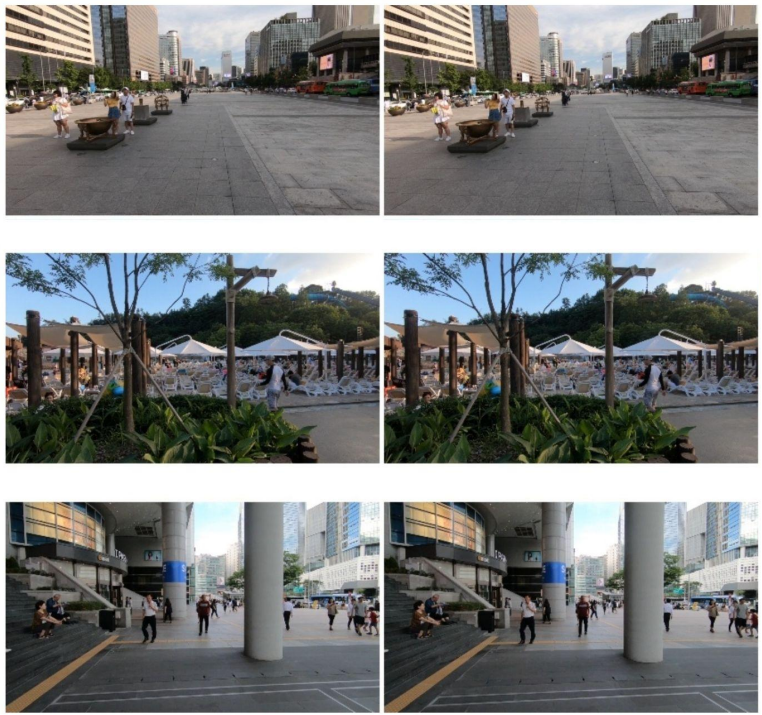
Methodology - [7]
(System Block Diagram - Inference Phase)



Methodology - [8]
(Dataset Description - REDS)

- The REalistic and Dynamic Scenes (REDS) dataset is designed for
  - Video De-blurring
  - Video Super-Resolution

Dimensions of each HR frame	1280 x 720
Dimensions of each LR frame	320 x 180
No. of frames per video	100
Total no. of videos	270
Total no. of frames	27000



First Frame Second Frame

Methodology - [9]
(Instrumentation Requirements)

- Python Programming Language.

- The PyTorch Library.
  - Uses Tensors for optimal processing.
  - Provides a myriad of functions, optimizers, and basic models.

- ffmpeg
  - Solution to record, convert, and stream audio and video.

- GPUs
  - Online cloud servers provide GPUs remotely that can be used for training.

Methodology - [10]
(Working Principle - Training Stage - Forward Diffusion)

- Training data is loaded and randomly cropped to 256 x 256 dimensions.
  - Large frames of dataset cannot be feasibly processed.
  - HR frames are randomly cropped once per iteration.

- Cosine variance schedule is used for adding noise to image.
  - Random noise is sampled from a normal distribution.
  - To move from intractable data distribution to a tractable normal distribution.

Methodology - [11]
(Working Principle - Training Stage - Denoising)

- LR frames in neighborhood are upscaled using bicubic interpolation.
  - Neighborhood includes the target frame, preceding frame, and succeeding frame.
  - Upscaled in order to match dimensions for concatenation.

- Upscaled, and noised frames stacked channel-wise, and input to the denoiser network.
  - Allows to condition the denoiser on upscaled, and noisy frames.

- L1 Loss is calculated between predicted noise $(\hat{\varepsilon})$ and actual noise (ε)

$$L = |\varepsilon - \hat{\varepsilon}|$$

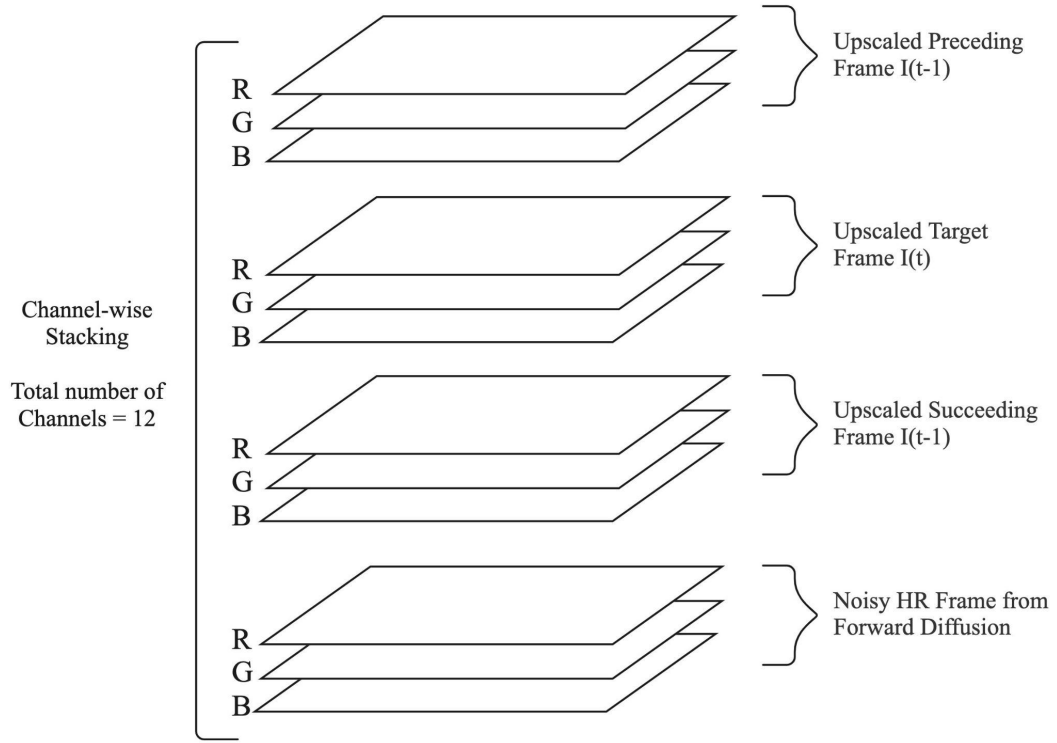
and normalized to per-pixel loss.

$$L_{pix} = \frac{L}{H \times W \times C}$$

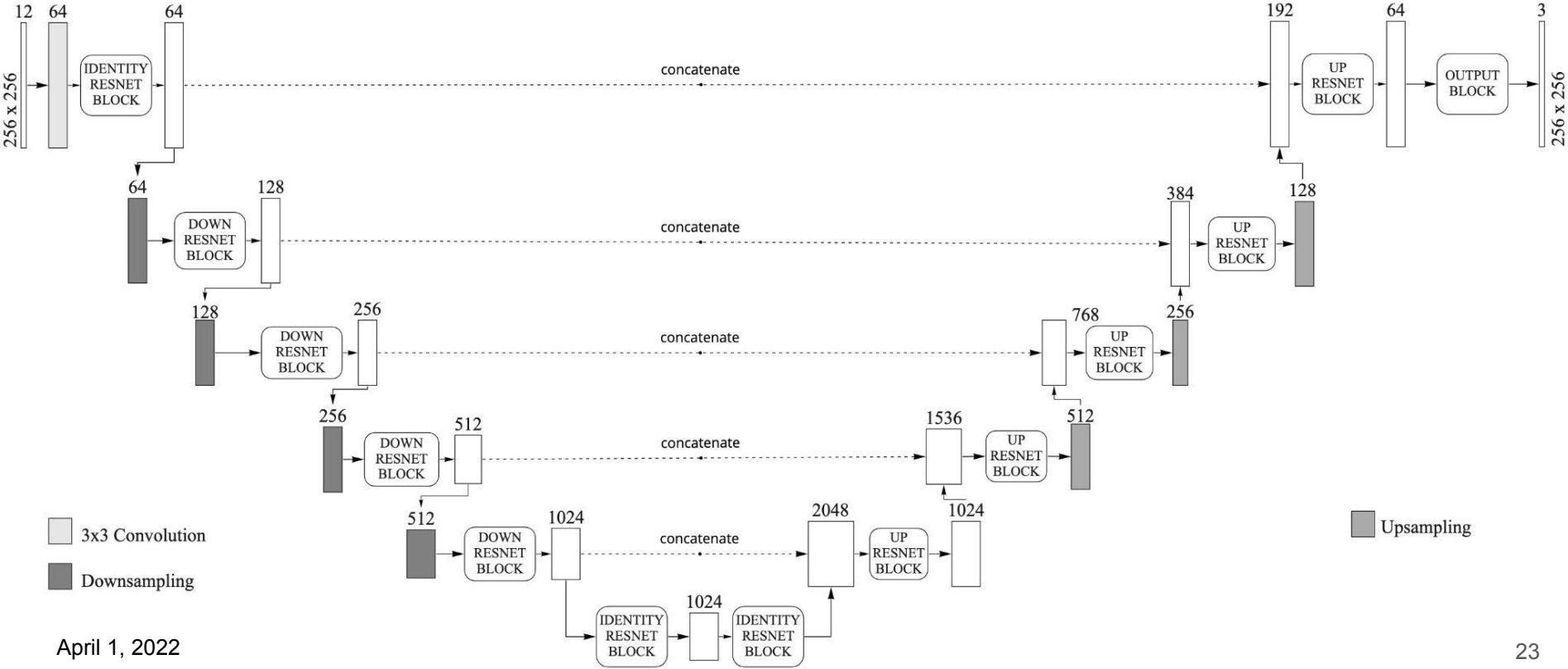
H, W, and C are height, width, and no. of channels respectively

Methodology - [12]
(Working Principle - Conditioning on Upscaled Frames)

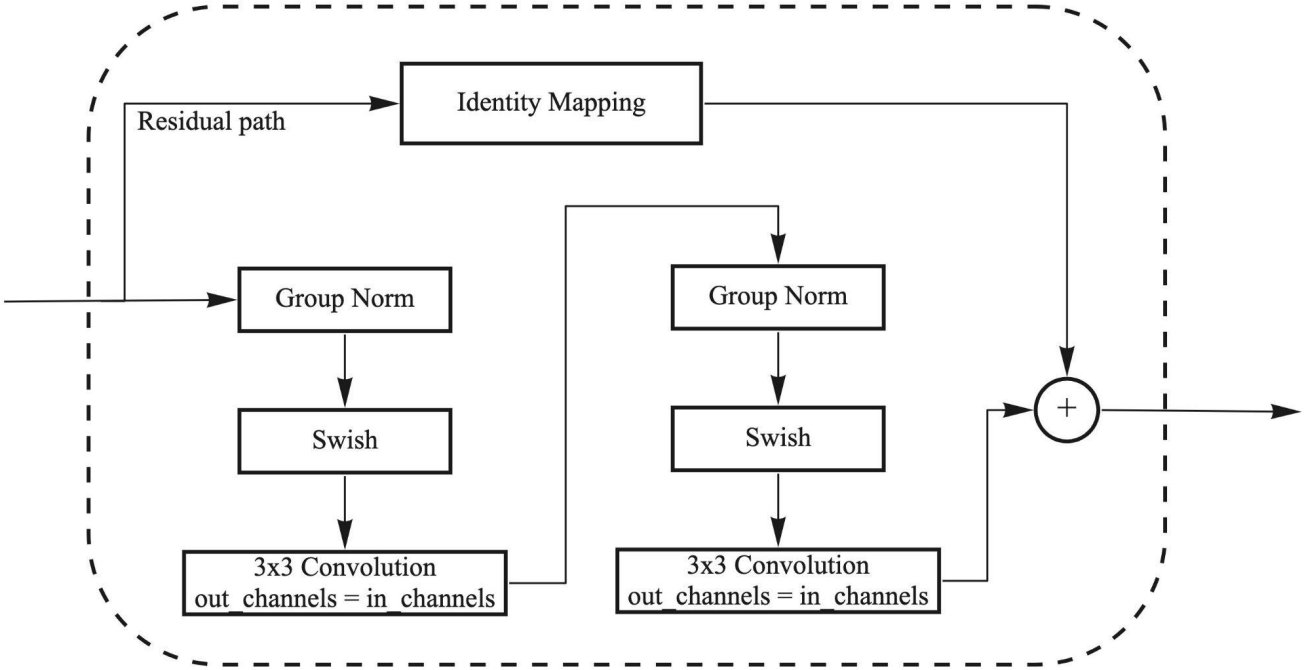
- R, G, and B denote the color channels of the video frames used.
- Stacks target frame, neighboring frames, and noisy frame channel-wise.
- Results in a 12-channel input for conditioning the denoiser network.



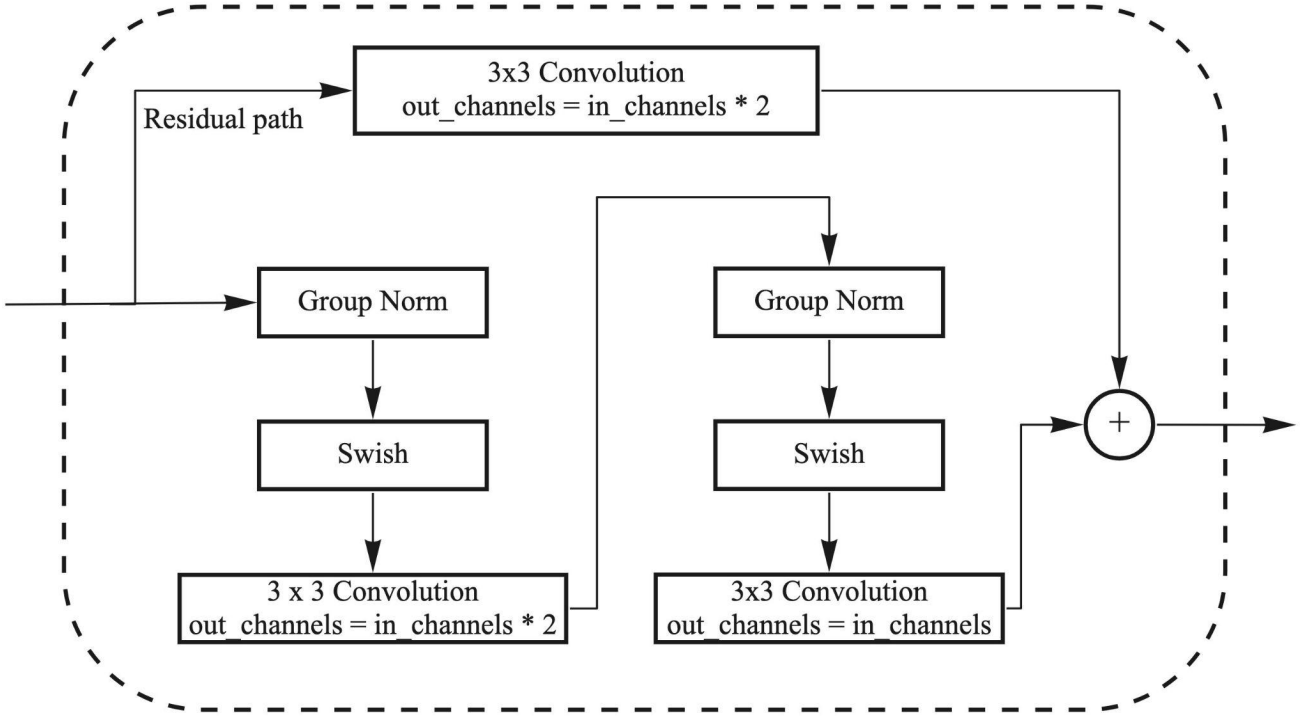
Methodology - [13]
(Working Principle - Denoiser Network)



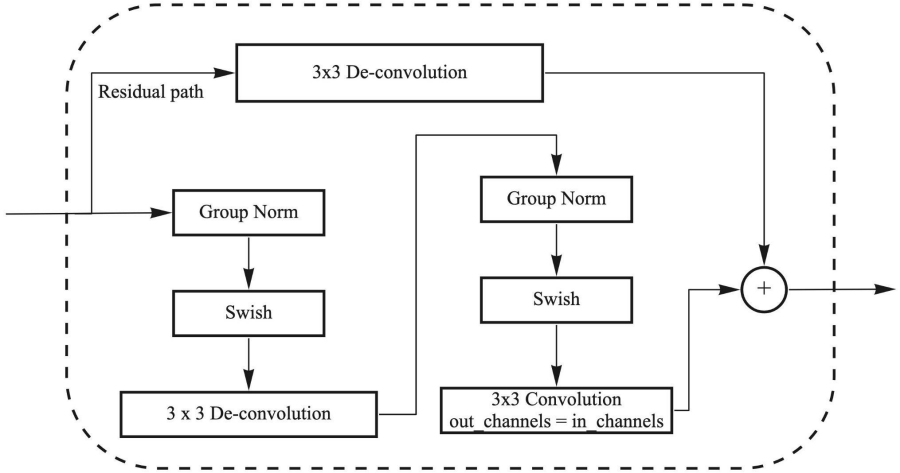
Methodology - [14]
(Denoiser Network - Identity ResNet Block)



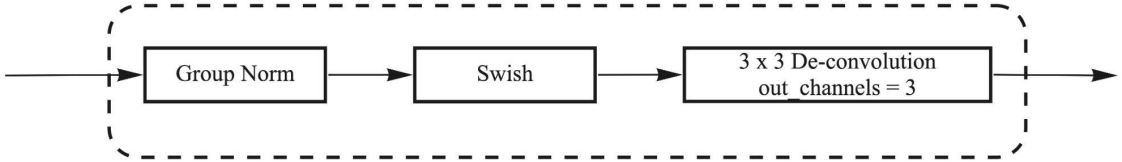
Methodology - [15]
(Denoiser Network - Down ResNet Block)



Methodology - [16]
(Denoiser Network - Up ResNet and Output Blocks)



Up ResNet Block



Output Block

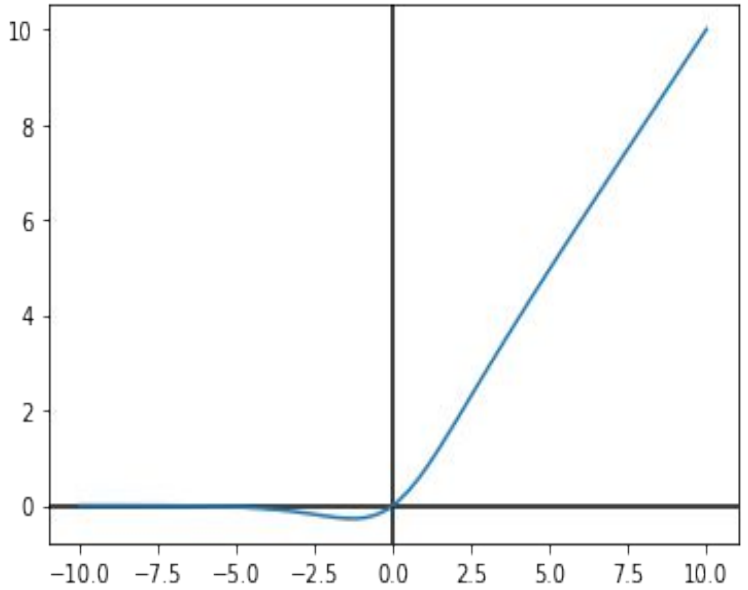
Methodology - [17]
(Denoiser Network - Swish Activation Function)

- The Swish activation function is given by,

$$f(x) = x \cdot \sigma(\beta x)$$

$\sigma(z) = (1 + \exp -z)^{(-1)}$ is the sigmoid function.

- It performs well on deeper networks.



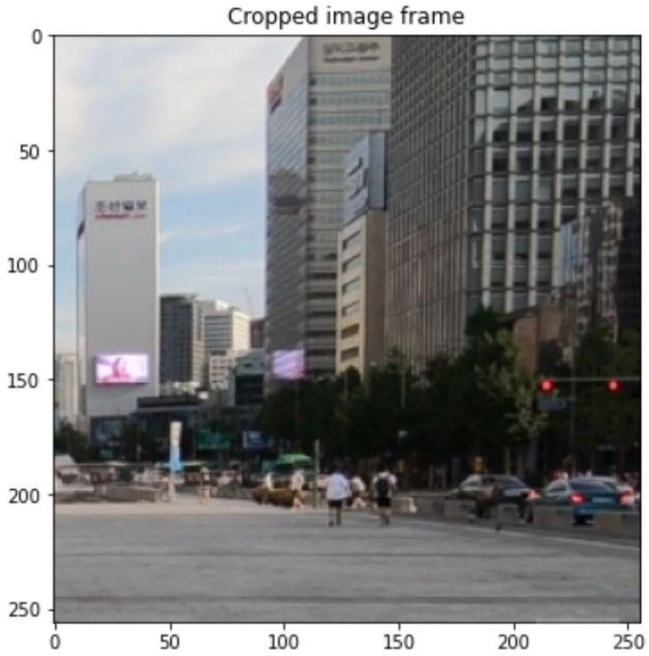
Activation Functions for Inception-ResNet-v2 on ImageNet

Activation Function	Top-1 Accuracy (%)		
Leaky ReLU	79.5	79.5	79.6
Parametric ReLU	79.7	79.8	80.1
Softplus	80.1	80.2	**80.4**
ReLU	79.5	79.6	79.8
Swish-1	**80.2**	**80.3**	**80.4**

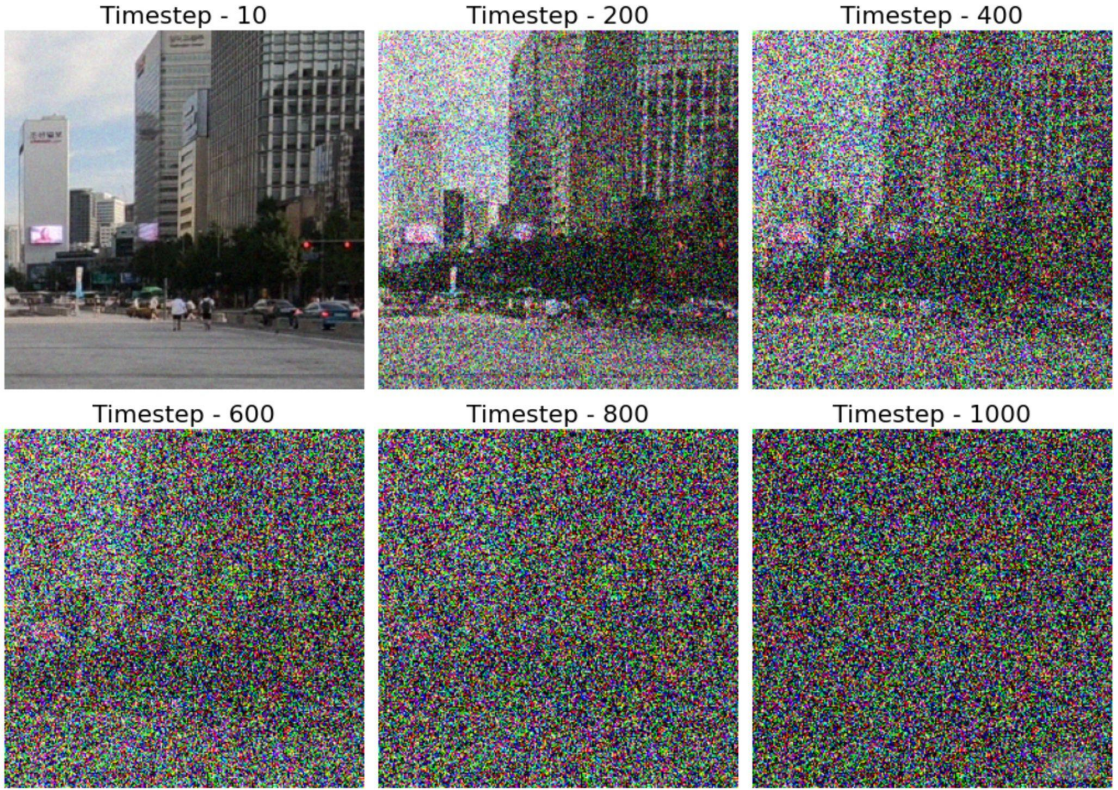
Methodology - [18]
(Working Principle - Reverse Diffusion Process)

- To start the reverse process, a noisy sample is taken from the Normal Distribution.
- Stack input upscaled frames channel-wise, with the noise.
- The frames are passed to the denoiser model to obtain frame for the next reverse time-step.
- Using the result from the previous time-step as the noisy input, all reverse diffusion steps are carried out.
- At the end of the reverse process, resulting SR frame is obtained.

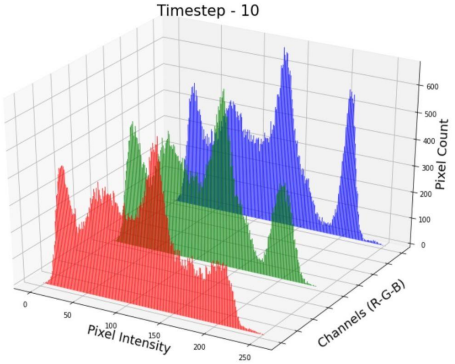
Results - [1]
(Random Cropping of Training Sample)

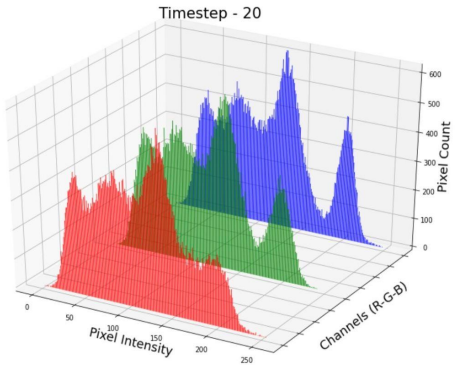


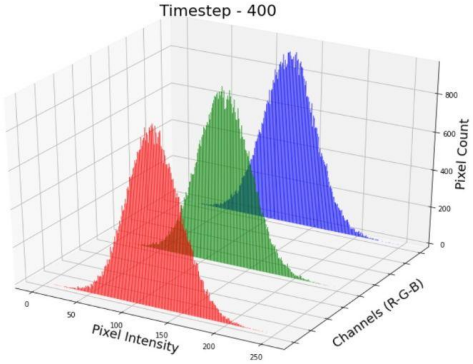
Results - [2]
(Noise Addition during the Forward Process)

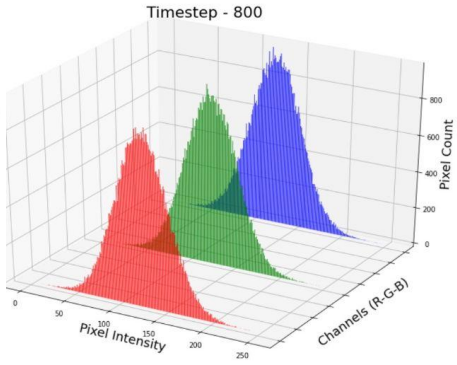


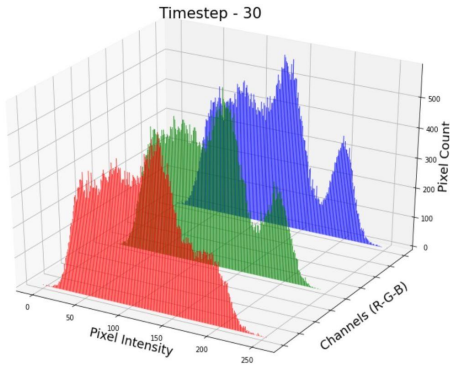
Results - [3]
(Transformation of Input Distribution from Noise Addition)

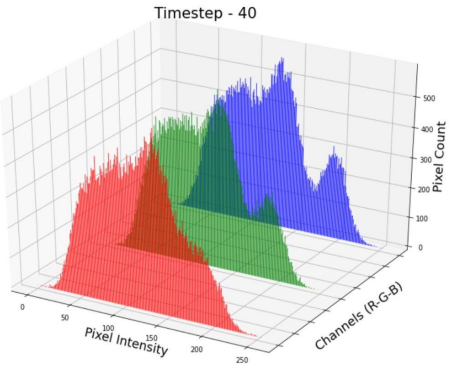


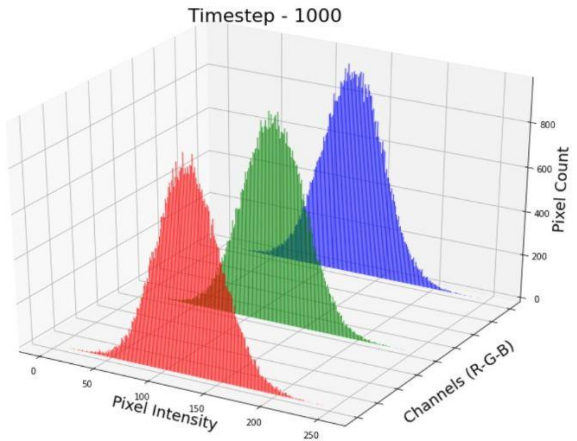




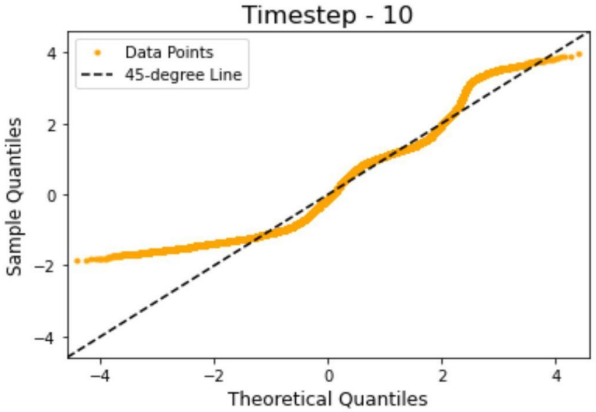


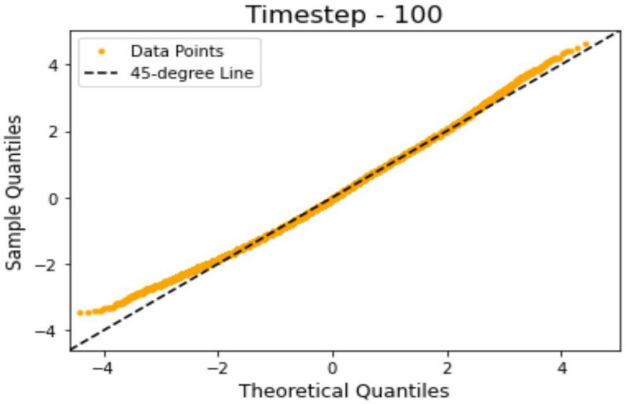


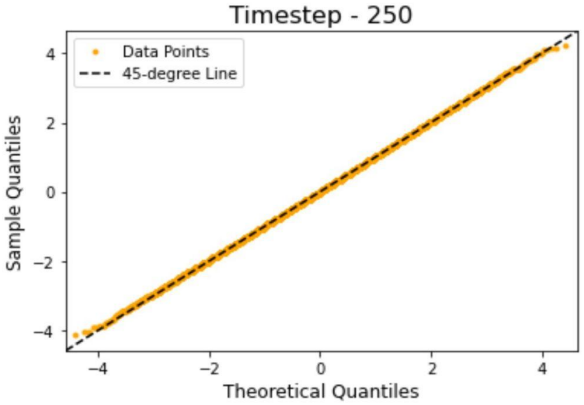


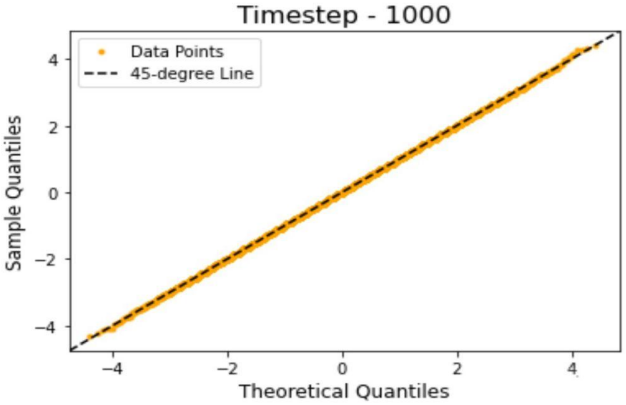


Results - [4]
(Q-Q Plots during Forward Process)

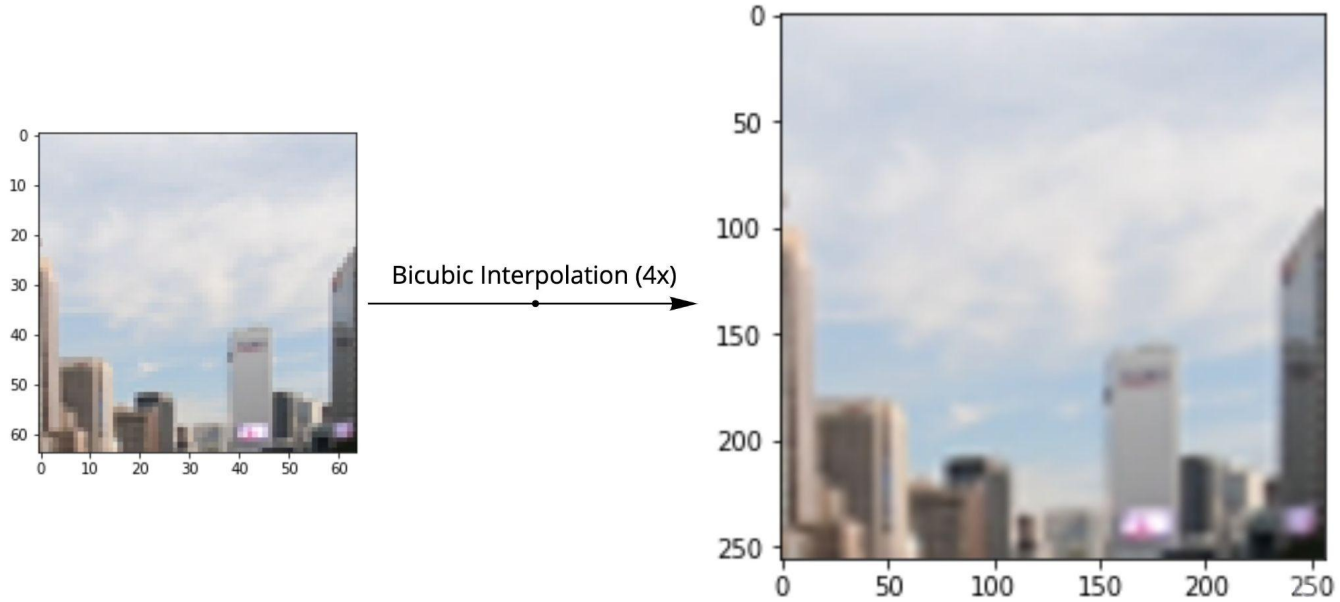




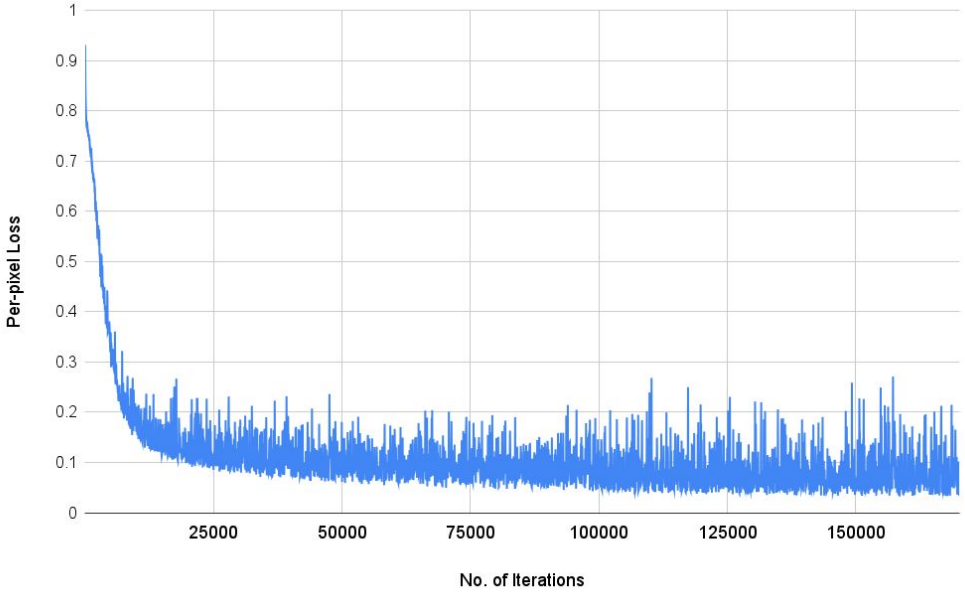


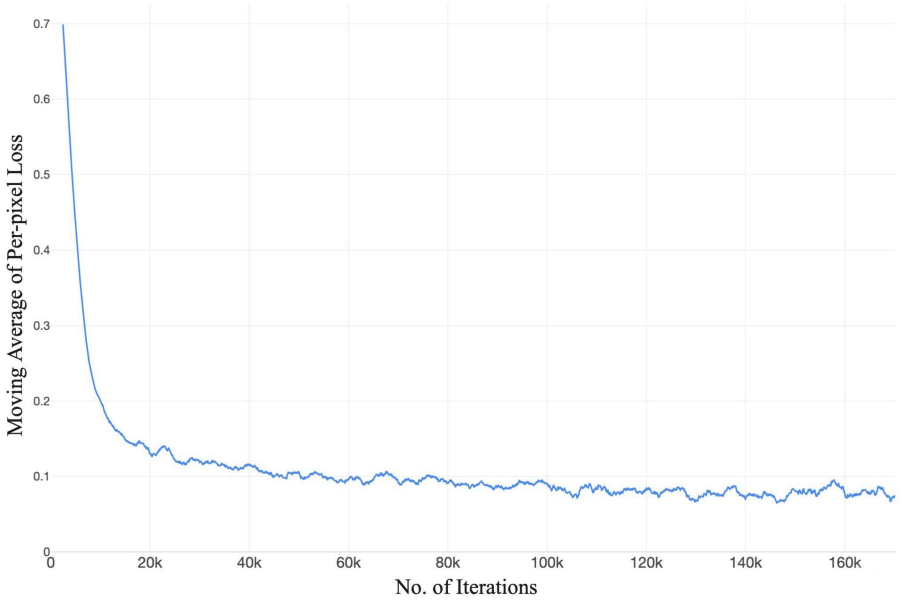


Results - [5]
(Bicubic Interpolation for 4x Upscaling)



Results - [6]
(Loss over Training Iterations)

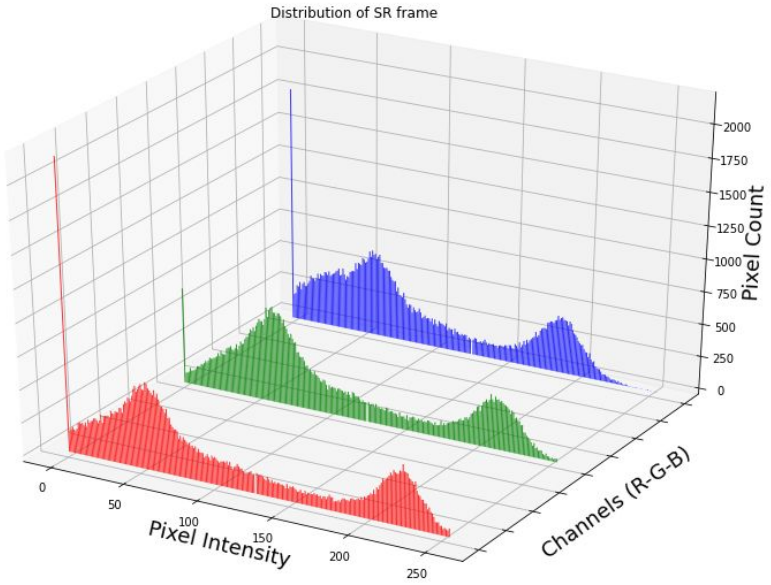


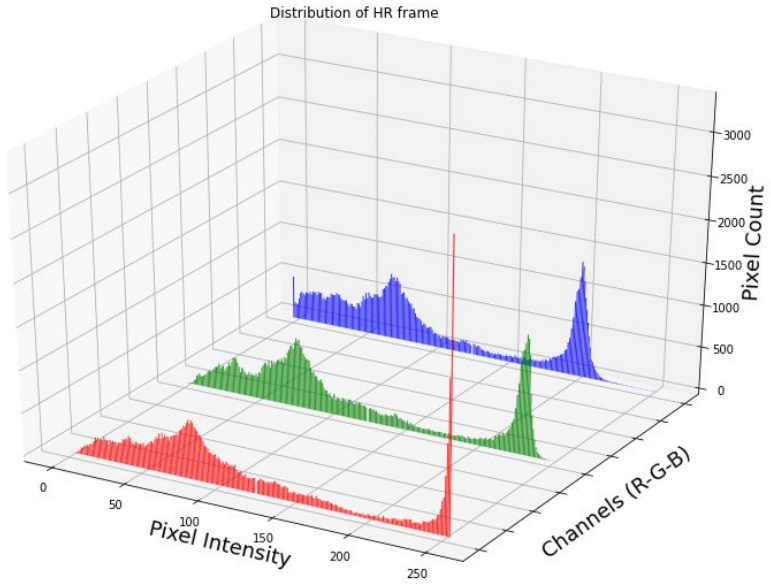


Results - [7]
(Reverse Diffusion Process on a Test Sample)

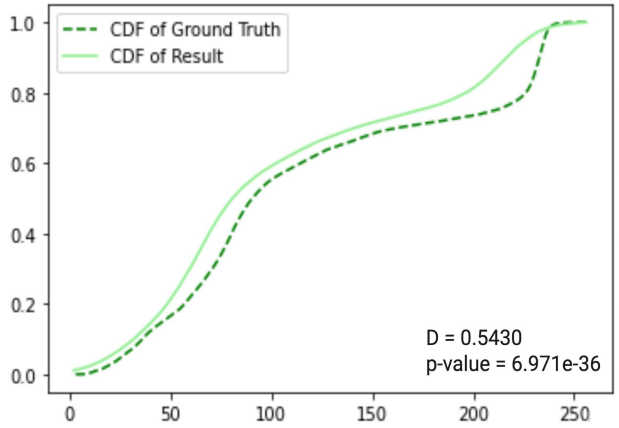
PSNR=20.12dB SSIM=0.47 NMSE=0.204

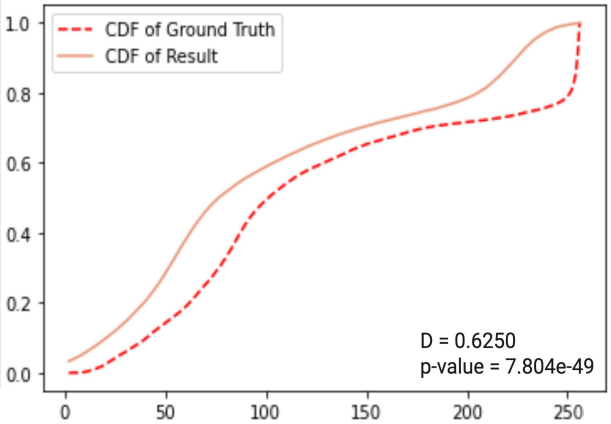
Results - [8]
(Distributions of SR Result and HR Ground Truth)

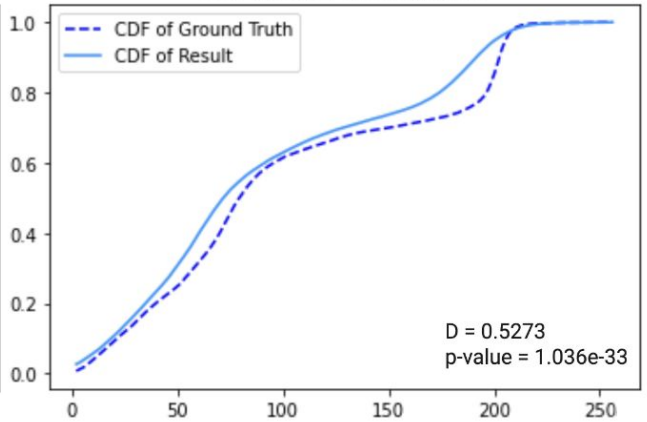




Results - [9]
(KS Test on Ground Truth and Result Distributions)



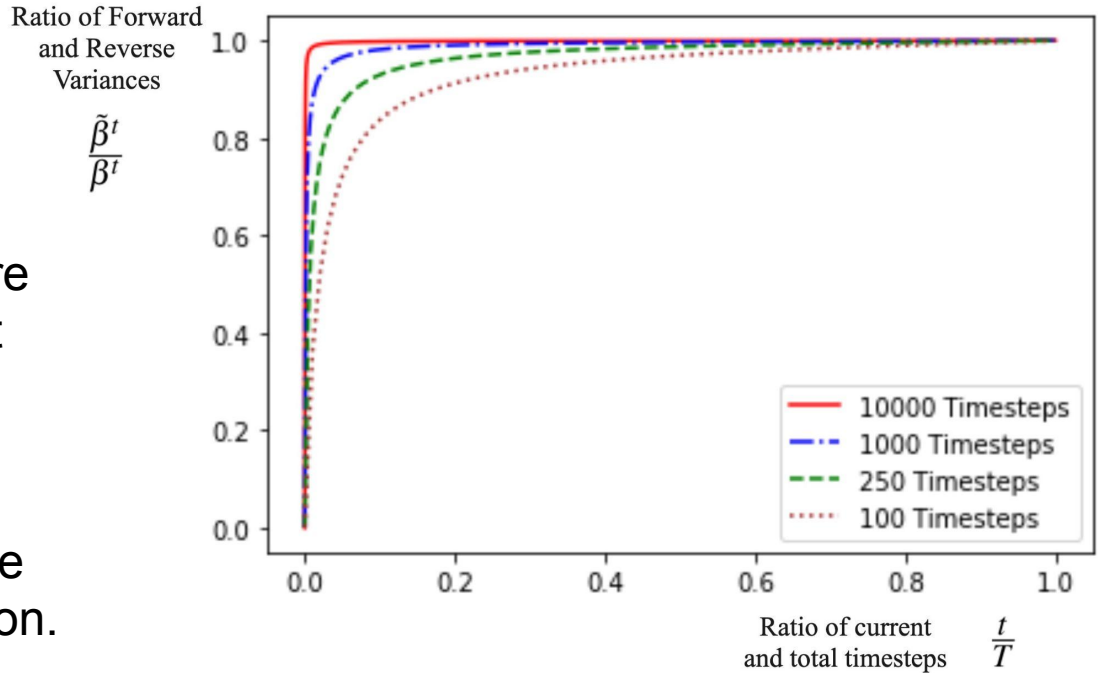




- Blue Channel shows the best results compared to the others.
- Reject Null Hypothesis that the distributions are identical.
- More training could provide closer distance measures.

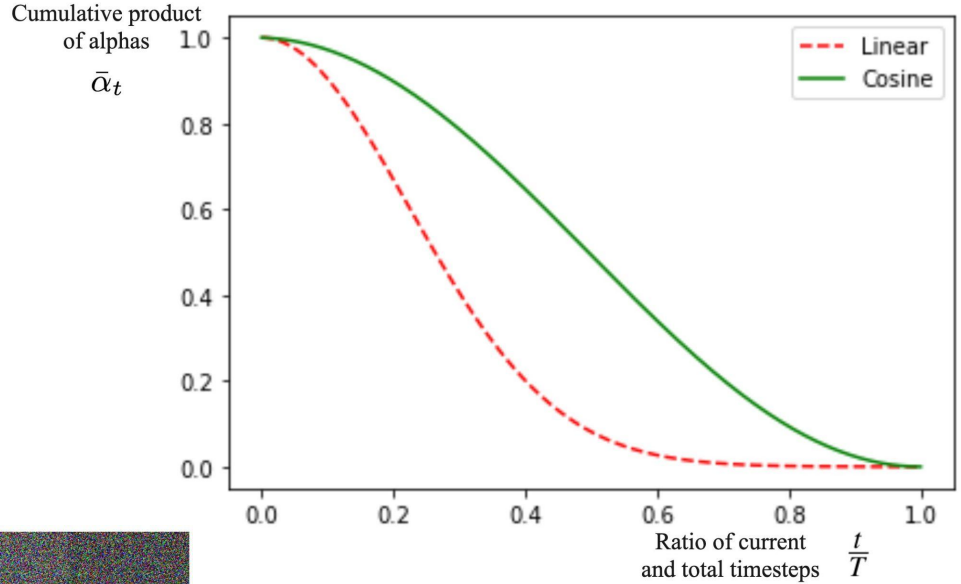
Discussion and Analysis - [1]
(Learning the Noise Distribution)

- Variance schedule (i.e. beta values) can be fixed.
- Different values of Beta for forward and reverse process are very close to each other except near t=0.
- Shows that model mean is more indicative of the noise distribution.



Discussion and Analysis - [2]
(Choice of Variance Schedule)

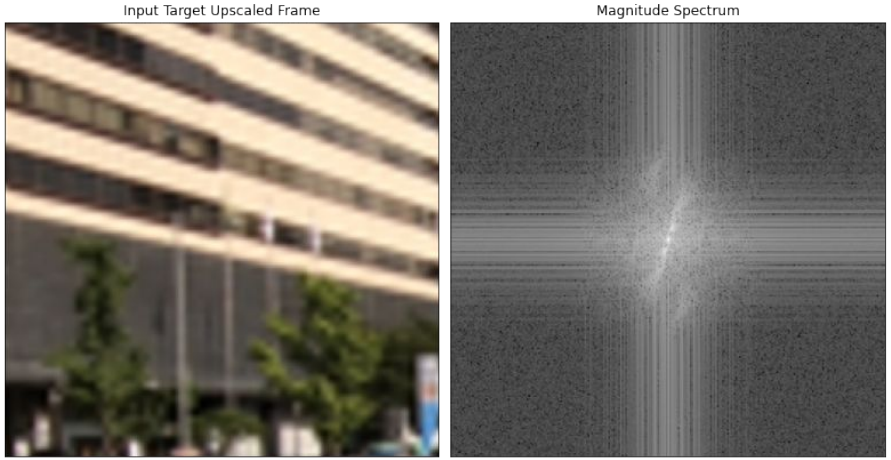
- In linear scheduling, noise gets added too quickly.
- Latter results contribute minimally to learning.
- Cosine schedulers provide a more smooth increase in variance.
- Allows smaller number of time-steps to be used.



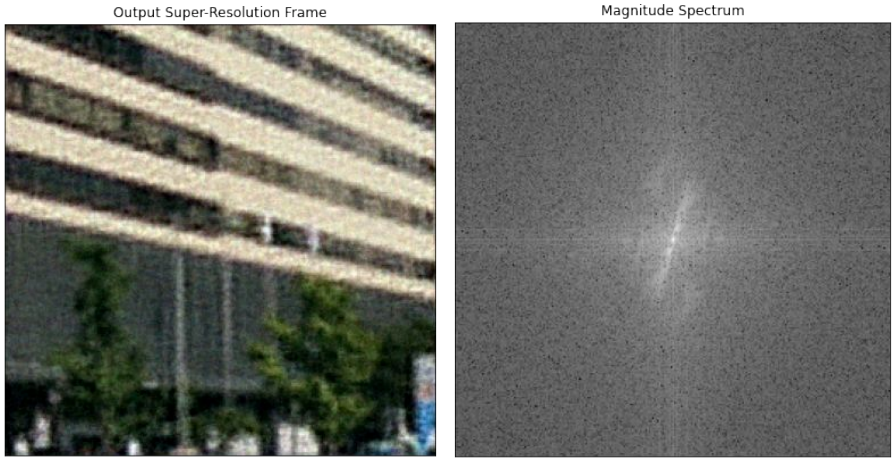
Cosine Schedule

Linear Schedule

Discussion and Analysis - [3]
(Evaluating Motion Blur)



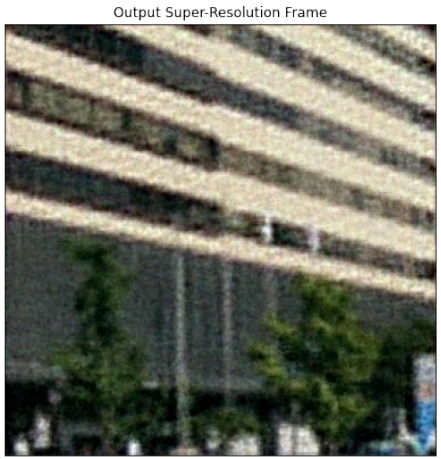
FFT Blur: -11.89

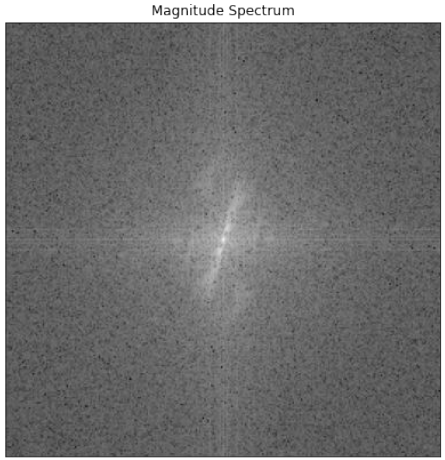


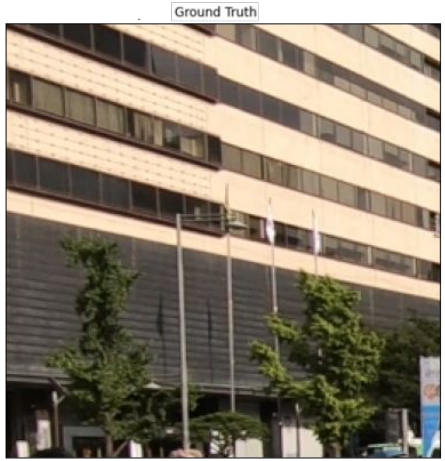
FFT Blur: 28.21

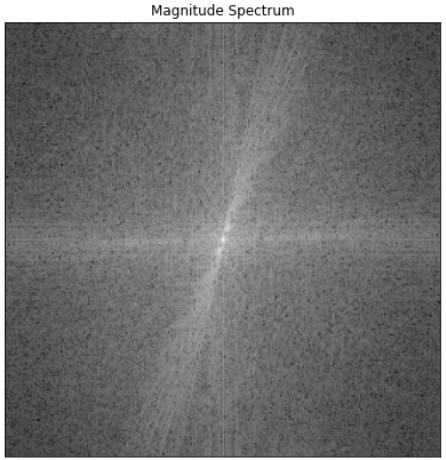
- Magnitude Spectrum of result show more spread out of high frequency components.
- Signifies the reduction of Motion Blur (Higher FFT Blur Metric is better)

Discussion and Analysis - [4]
(Evaluating Motion Blur)







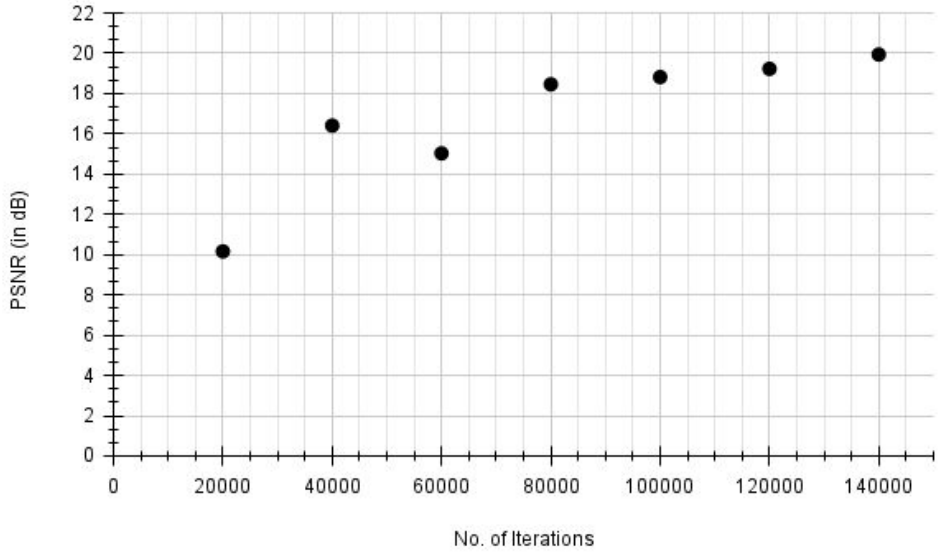


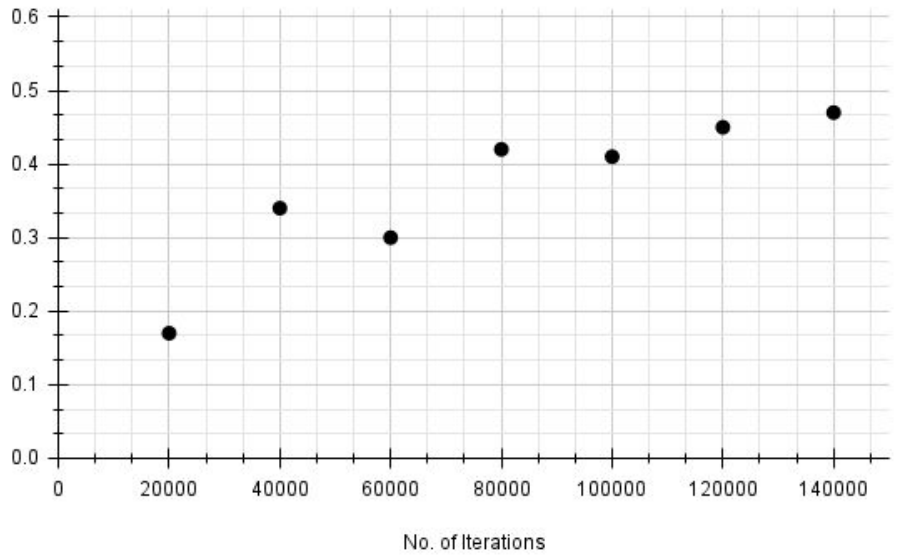
FFT Blur: 28.21

FFT Blur: 32.80

- FFT Blur Value is closer to the Ground Truth HR frame.
- Values show that minute details have not been fully recovered

Discussion and Analysis - [5]
(Image/Video Quality Assessment during Training)





Discussion and Analysis - [6]
(Comparison with State-of-the-art Methods)

- Applied the model on a test sample provided from MSU Video Super-Resolution Benchmark.
  - Averaged the PSNR, and SSIM values obtained.
- Results don’t come close to state-of-the-art methods.
- Possible Reasons:
  - Model hasn’t been fully trained.
  - Random cropping of video frames.
  - Use of a different dataset.

Model	PSNR (in dB)	SSIM	Dataset Used for Training
RBPN [8]	31.407	0.899	Vimeo-90k
iSeeBetter [14]	31.104	0.896	Vimeo-90k, SPMCS, Vid4
ESRGAN [12]	27.33	0.808	Set5, Set14, BSDS100
This Project	19.94	0.427	REDS

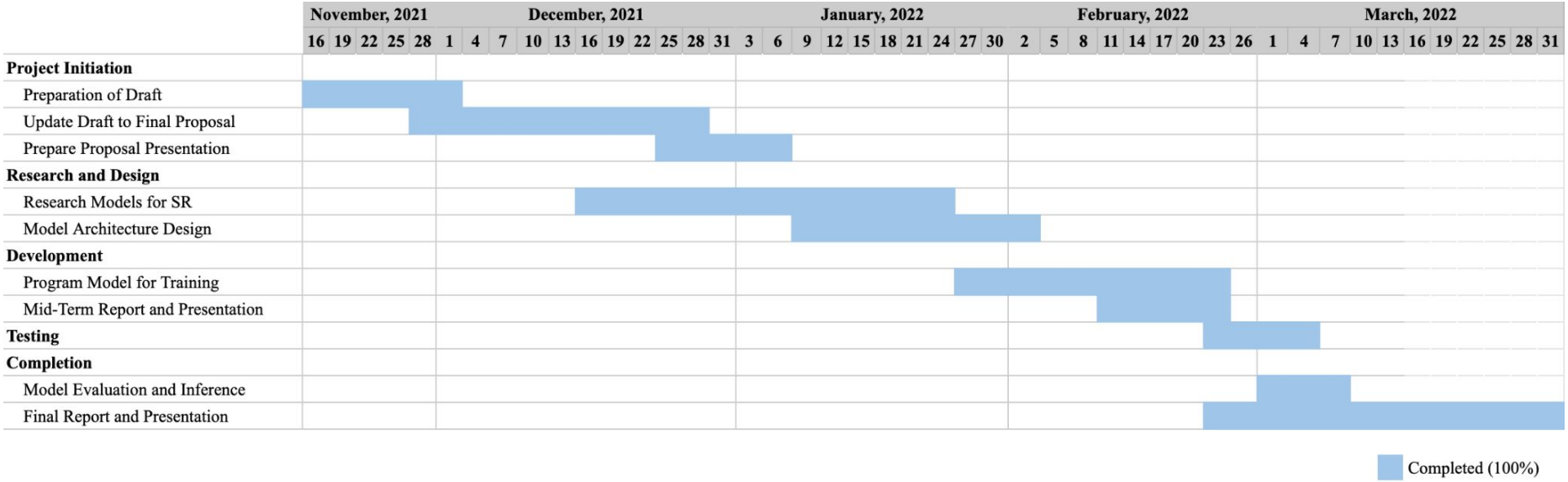
Future Enhancements

- Conditional augmentation on LR inputs could improve results.
- Could use entire video frames instead of random cropping.
- Usage of noisy LR target frames at the start of the reverse process.
- Optimizing, or revamping the Denoiser Network Architecture.
- Enhancing efficiency for real-time processing capabilities.

Conclusion

- Diffusion models were successfully adapted for video super resolution.
  - Results showed visual improvement to the LR input.
  - Showed reduction in motion blur as well.
- Comparison of results to state-of-the-art generative models were carried out.
  - Results are poor compared to the standards of the state-of-the-art.
- The objectives set during the project's conception were met.

Project Schedule



References - [1]

[1] A. Chadha, J. Britto, and M. M. Roja, “iSeeBetter: Spatio-temporal video super-resolution using recurrent generative back-projection networks,” Comput. Vis. Media, vol. 6, no. 3, pp. 307–317, 2020, doi: 10.1007/s41095-020-0175-7.

[2] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image Super-Resolution via Iterative Refinement,” 2021, [Online]. Available: http://arxiv.org/abs/2104.07636.

[3] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded Diffusion Models for High Fidelity Image Generation,” 2021, [Online]. Available: http://arxiv.org/abs/2106.15282.

References - [2]

[4] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” 2021, [Online]. Available: http://arxiv.org/abs/2105.05233

[5] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Adv. Neural Inf. Process. Syst., vol. 2020-December, Jun. 2020, Accessed: Jan. 01, 2022. [Online]. Available: https://arxiv.org/abs/2006.11239v2.

[6] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, 2017. arXiv: 1710.05941 [cs.NE].

THANK YOU