Towards Accurate Generative Models of Video: New Metric & Challenges

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, Sylvain Gelly

April 9, 2019

M. R. Karimi

Contents

Generative models did a great job on Images!

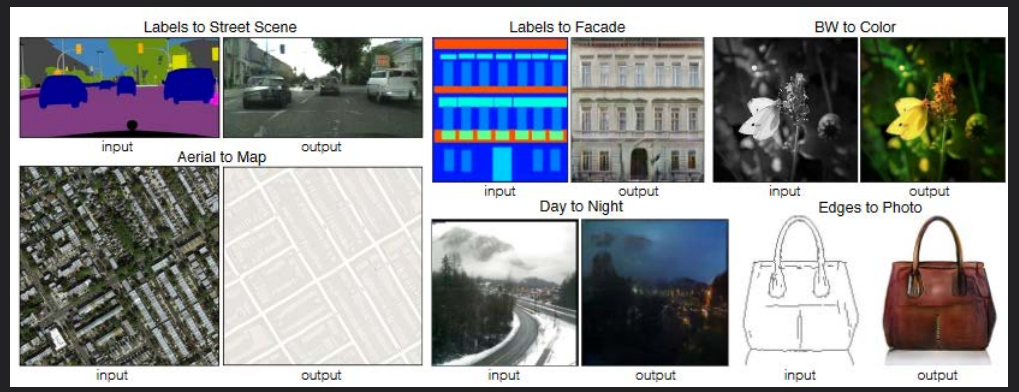
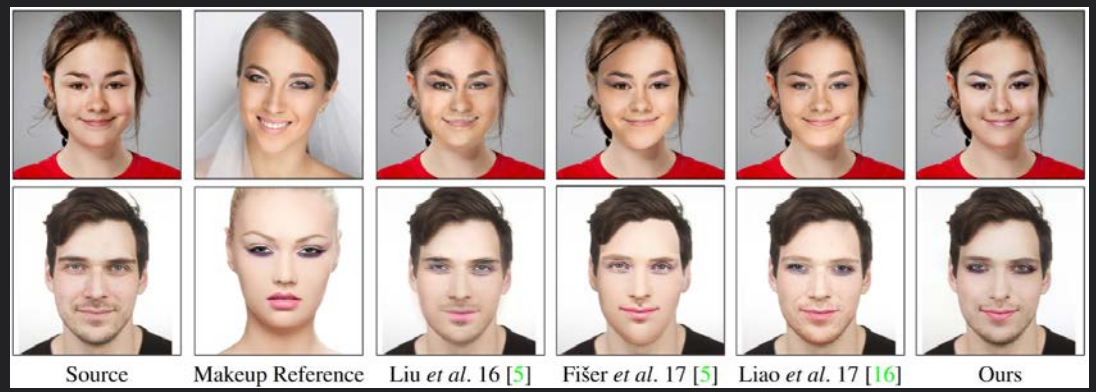


Image-to-image translation [Isola et al. 2017]



Style transfer [Chang et al. 2018]

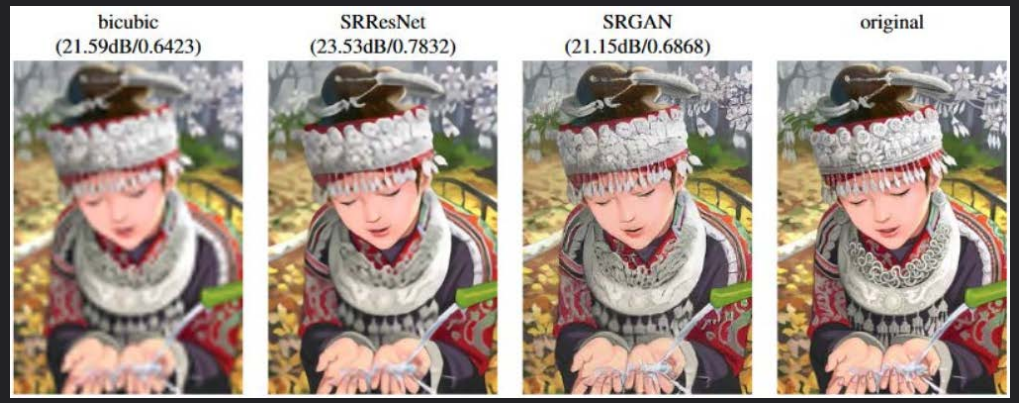


Image super-resolution [Lediget al 2016]

Next challenge is video generation!

- Capture the temporal dynamics of visual scene + their visual presentation
- Learning a good dynamics model remains a major challenge!

Having good metrics and dataset is also necessary for video generation!

- Metrics: Consider visual quality, temporal coherence, and diversity of generated frames!
- Datasets: Require corresponding data sets that test for specific capabilities.

Distance between data distributions

- Accurate generative model of videos
  - captures the data distribution from which the observed data was generated.
- Distance(P_R , P_G)
  - obvious evaluation metric
- No analytic expression of either distribution is available
  - which rules out straightforward application of many common distance functions.

Fréchet distance

$$d(P_r, P_G) = \min_{X,Y} E|X - Y|^2 \qquad (1)$$

- Difficult to solve for general case!
- It has a closed form solution!
  - When $P_r \; and \; P_G$ are multivariate Gaussians

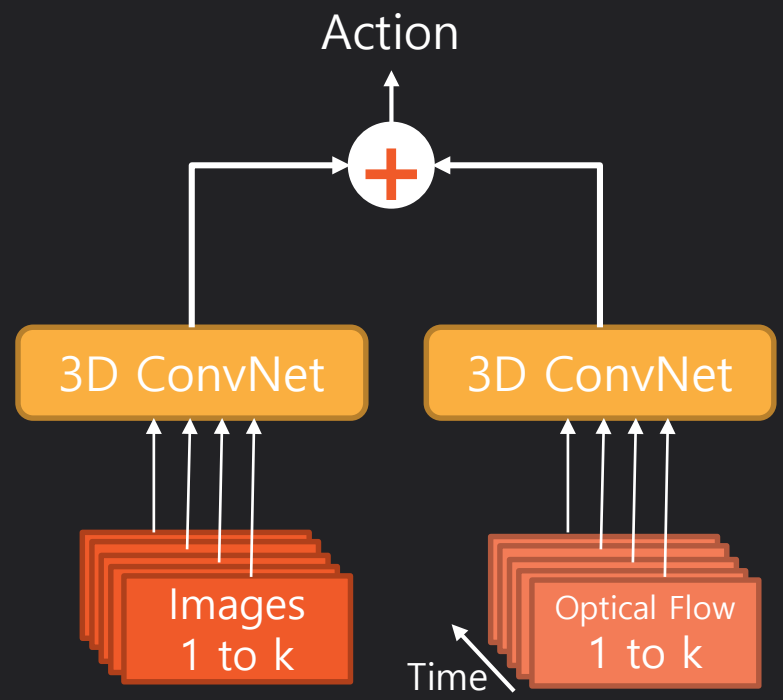
$$|\mu_R - \mu_G|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{\frac{1}{2}}) \qquad (2)$$

A multivariate Gaussian is NOT an accurate representation of the underlying data distribution!

- However, when using a suitable feature space, it is a reasonable approximation!
- For images: [Heusel et al. 2018]
  1. Train Inception network on ImageNet
  2. Record feature representation (activations) in one of the hidden layers from the network fed with P_R and P_G
  3. Computing Fréchet Inception Distance

Fréchet video distance

- Inflated 3D Convnet
- Action-recognition
  - Kinetics data set of human-centered YouTube videos
- Evaluation
  - Logits in the final layer
  - Output of the last pooling layer



Kernel Video Distance

- Downside of Eq. 2: Potentially large error in estimating Gaussian distributions over the learned feature space.
- Maximum mean discrepancy

$$\sum_{i \neq j}^{m} \frac{k(x_i, x_j)}{m(m-1)} - 2 \sum_{i}^{m} \sum_{j}^{n} \frac{k(x_i, x_j)}{mn} + \sum_{j \neq j}^{n} \frac{k(y_i, y_j)}{n(n-1)}$$

- Polynomial kernel

$$k(a, b) := (a^T b + 1)^3$$

Two main metrics

- Peak Signal to Noise Ratio (PSNR)
  - maximum attainable pixel value of the pixels in an image to its Mean Squared Error (MSE) with respect to a ground-truth image
- Structural Similarity Index Measurement (SSIM)
  - quality of an image as the perceived change in structural information.

III. Starcraft 2 Videos

Move Unit to Broder
(MUtB)

Collect Mineral Shards
(CMS)

Brawl

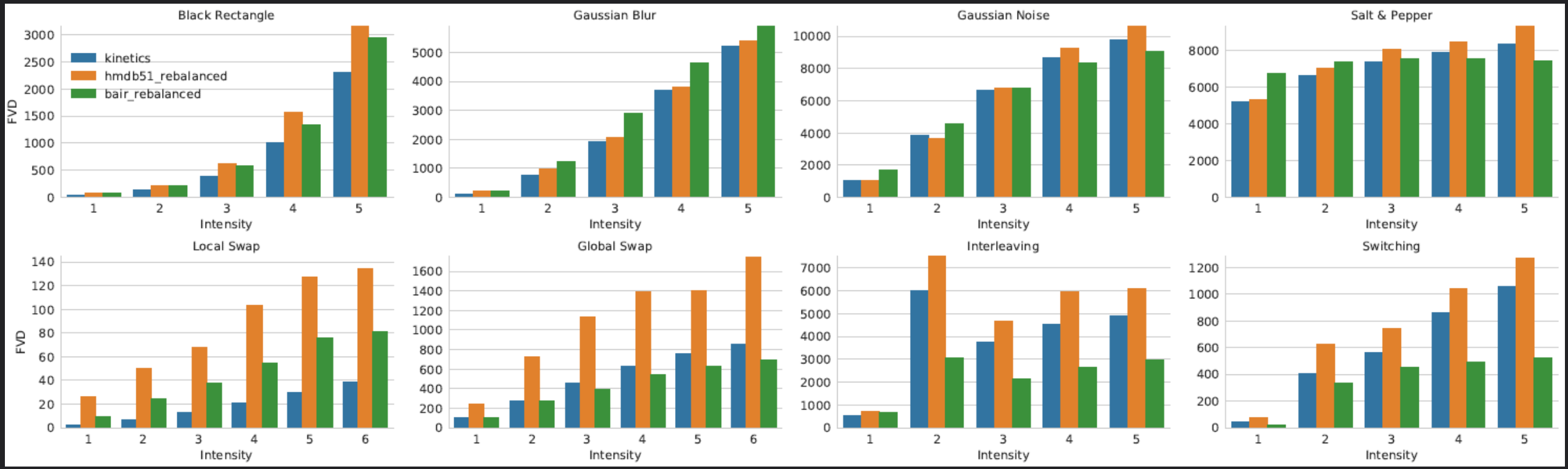
Road Trip with Medivac
(RTwM)

IV. Experiments

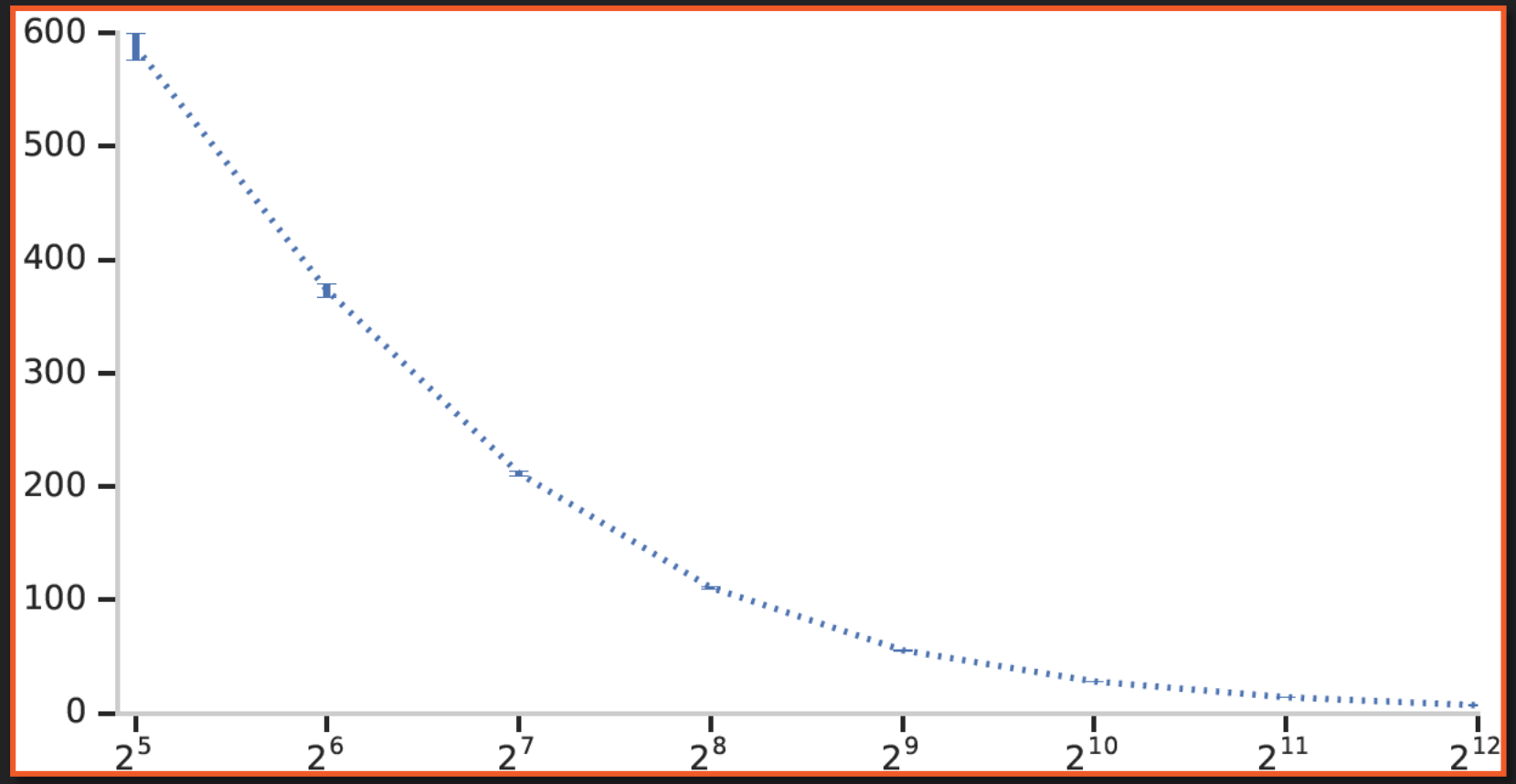
Noise Study

- Static Noise
  - Black rectangle drawn at random location
  - Gaussian blur
  - Gaussian noise
  - Salt & Pepper noise
- Temporal Noise
  - Locally swapping a number of randomly chosen frames with its neighbor
  - Globally swapping a number of randomly chosen pairs of frames selected
  - Interleaving the sequence of frames corresponding to multiple different videos
  - Switching from one video to another video after a number of frames

Results of Noise Study



Effect of Sample Size on FVD



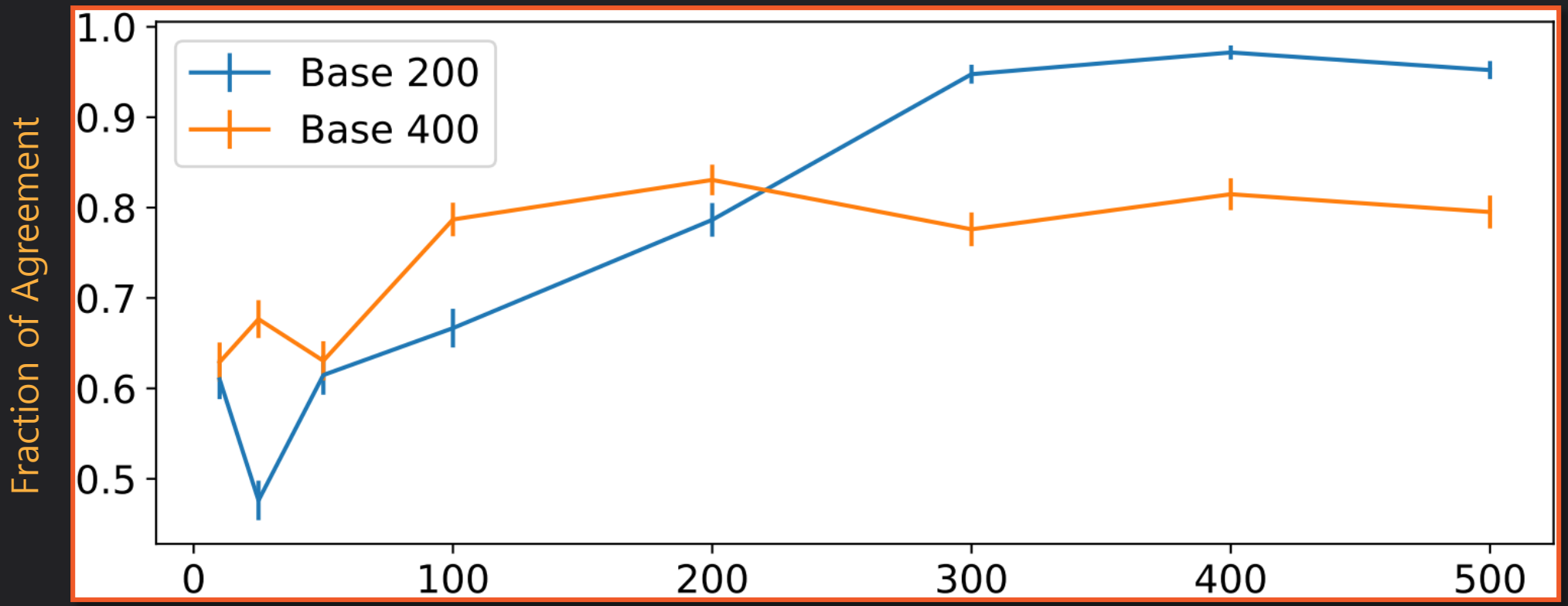
Human Evaluation

- If one metric is unable to distinguish between models, are those models truly equal in performance?
- If according to one metric there is a clear ranking among models, do humans/other metrics agree?

Results of Human Evaluation

Metric	eq. FVD	eq. SSIM	eq. PSNR	eq. KVD	spr. FVD	spr. SSIM	spr. PSNR	spr. KVD
FVD	N/A	**74.9** %	**81.0** %	**63.0** %	**71.9** %	58.4 %	63.5 %	**63.1** %
SSIM	51.5 %	N/A	44.6 %	43.6 %	61.8 %	51.2 %	45.9 %	50.2 %
PSNR	**56.3** %	21.4 %	N/A	48.8 %	54.1 %	37.0 %	44.8 %	54.1 %
KVD	40.6 %	70.4 %	73.8 %	N/A	69.4 %	56.8 %	**63.8** %	59.1%
Avg. FID	35.5 %	71.2 %	52.0 %	43.5 %	62.4 %	**62.7** %	57.6 %	51.2 %
Among raters	79.3 %	77.8 %	84.4 %	74.3 %	83.3 %	69.9 %	72.5 %	74.1 %

Resolution of FVD



Conclusion

Contributions

- Introducing Fréchet Video Distance (FVD)
- Investigation of FVD behaviour on temporal and frame-level perturbation
- Introducing StarCraft 2 Videos (SCV)
- Providing very comprehensive comparison of current state of the art models in terms of FVD.

However…

- Long-term memory learning and relational reasoning challenges are still open!

Q&A

Thank you! ☺

For more information
Please visit http://vml.kaist.ac.kr

Visual Media Lab
WHERE CREATIVITY MEETS TECHNOLOGY