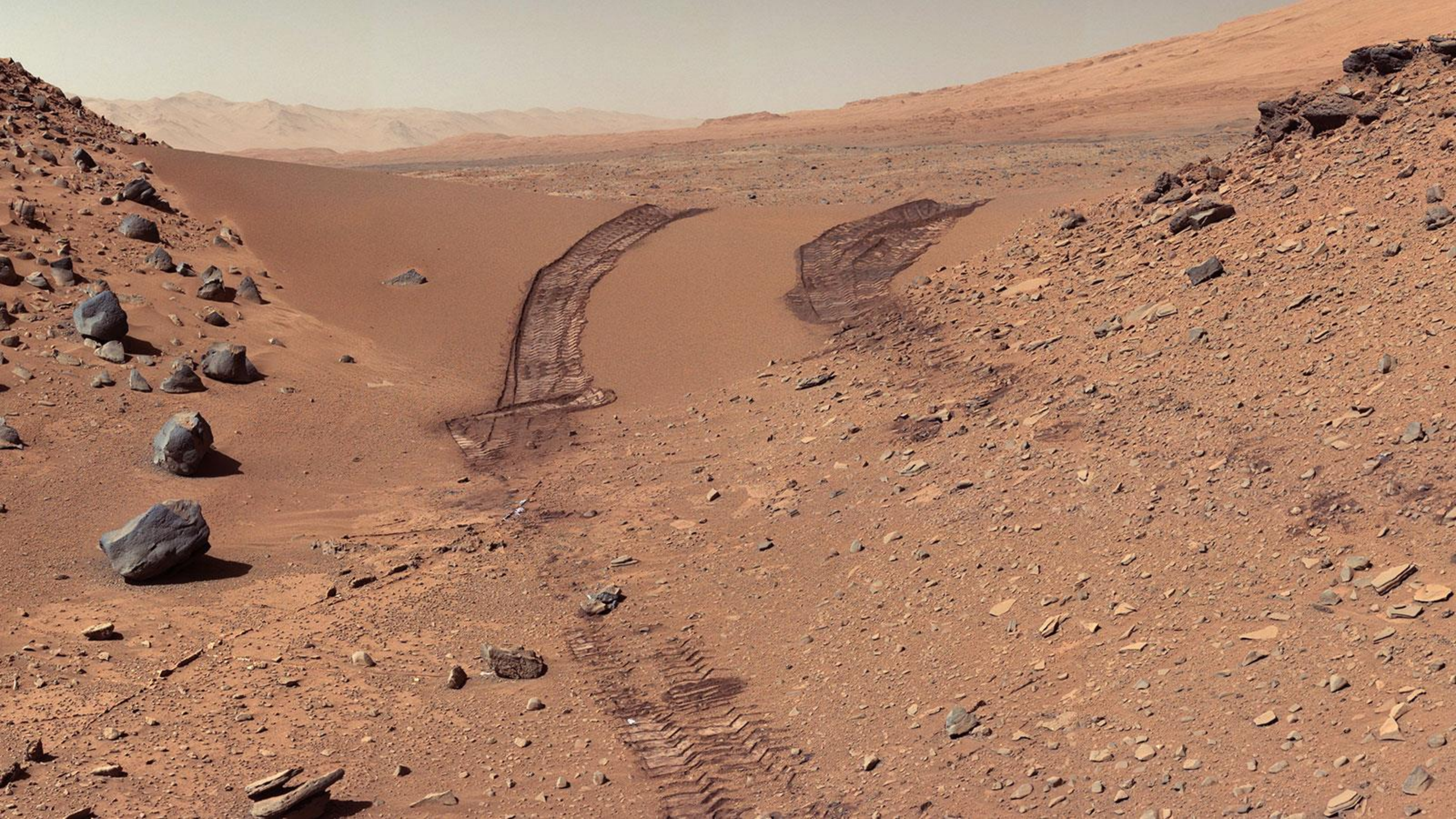
Assessing Synthetic Data Quality and Model Generalization for Planetary Imagery

Clara Salditt, Karan Molaverdikhani, Barbara Ercolano

Workshop on AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES'25), co-located with the 28th
European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy,
October 25, 2025



Outline

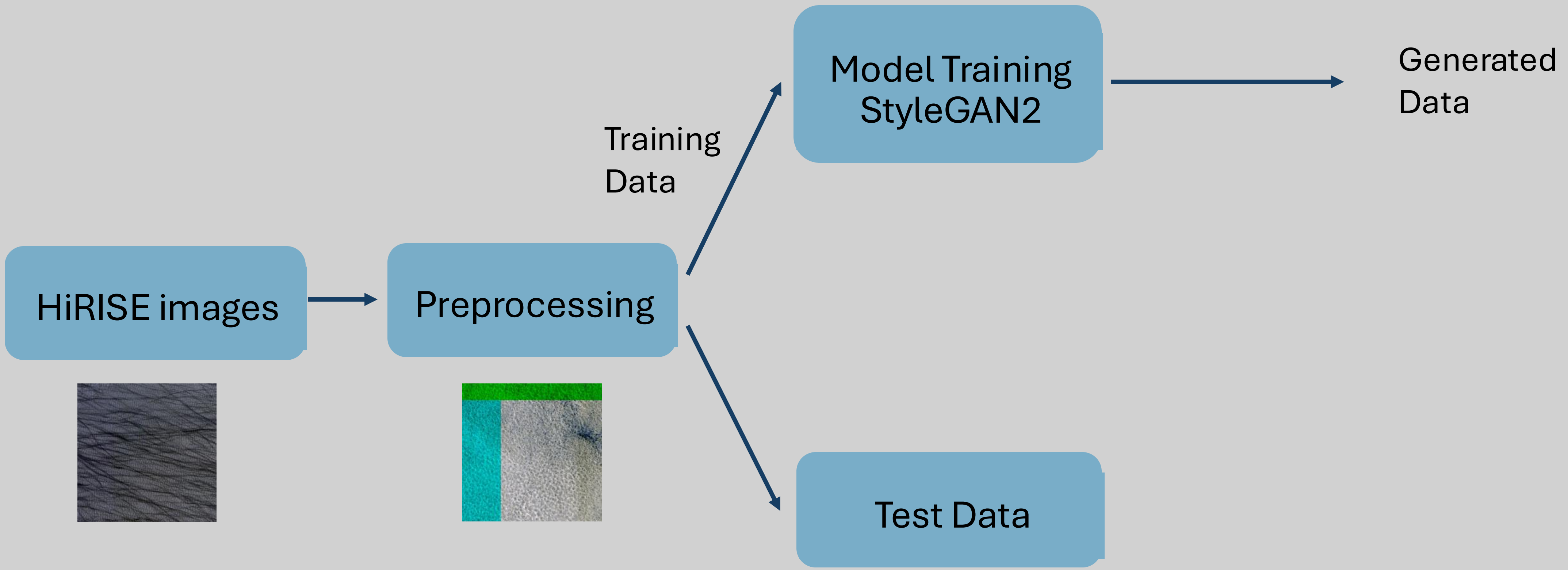
Motivation: Planets surfaces Rover landing and navigation

Problems: Gaps and bias in real data, finished model training on real data

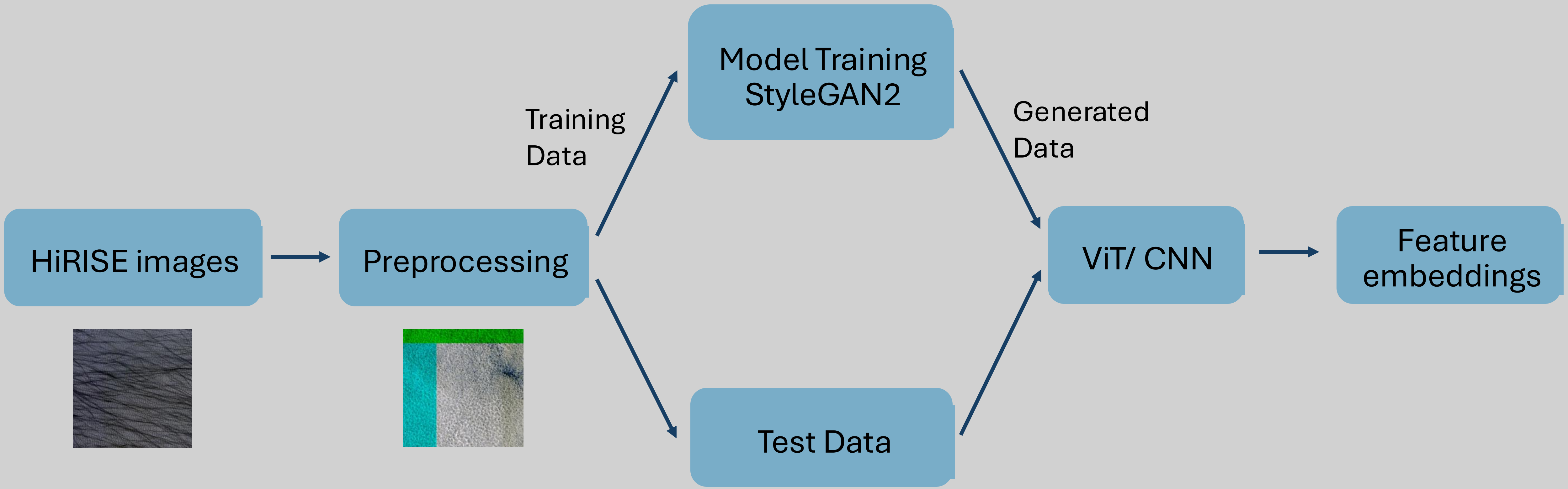
Solution: Synthetic data?

Conclusion: Data is currency in machine learning -> Data Quality assessment important

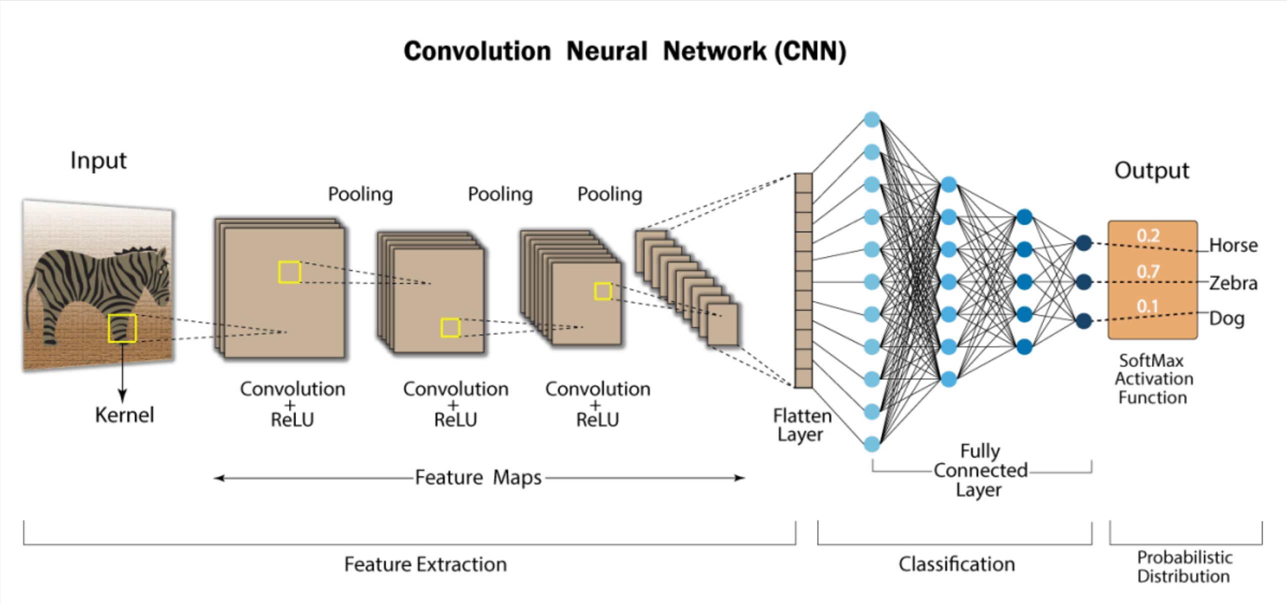
Summary of Methods

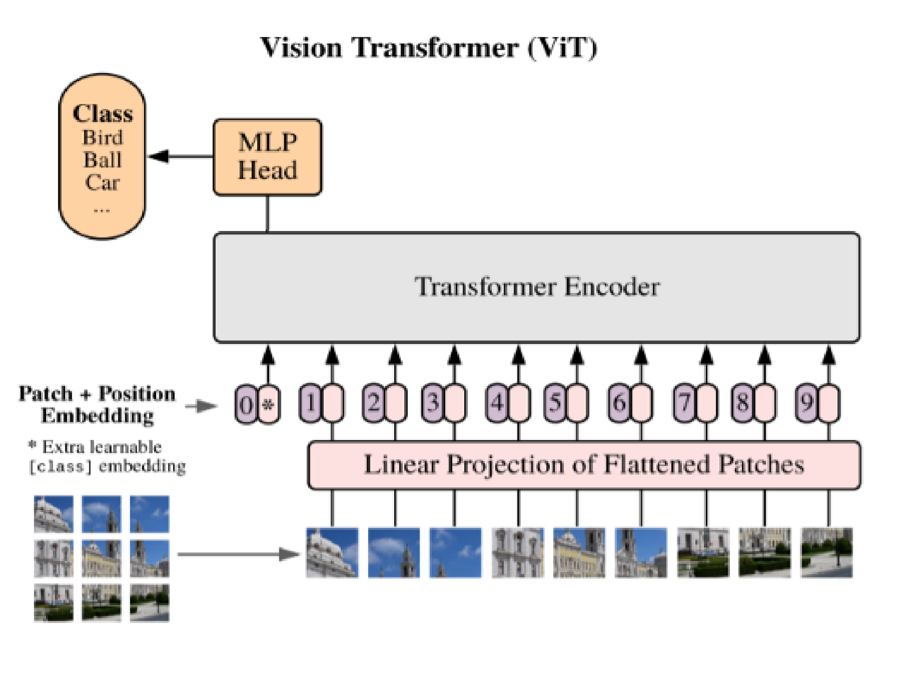


Summary of Methods



Feature extraction



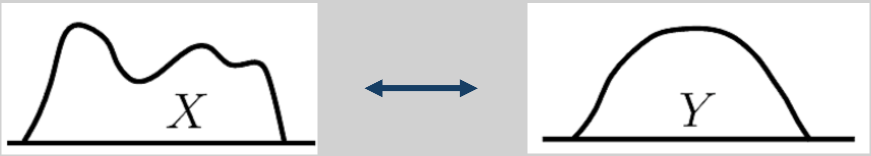


- Inceptionv3 (ImageNet: 1.2 M images 1000 classes)

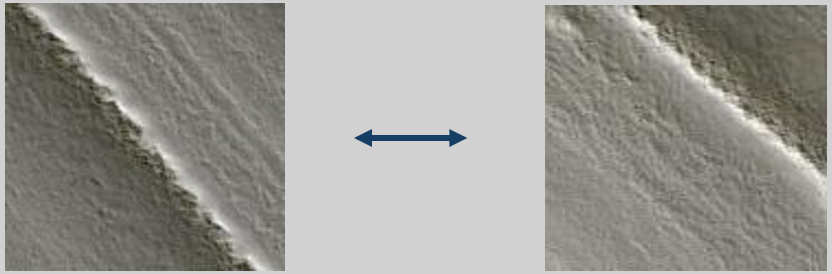
- CLIP (400 M images text pairs, supervised)
- DINOv2 (142 M images text pairs, self-supervised, Earth satellite Images)

Structure

1. Distribution based metrics



2. Pairwise image similarity metrics



3. Visualization techniques and qualitative feature space analysis

Evaluation metrics

Pairwise image similarity metrics

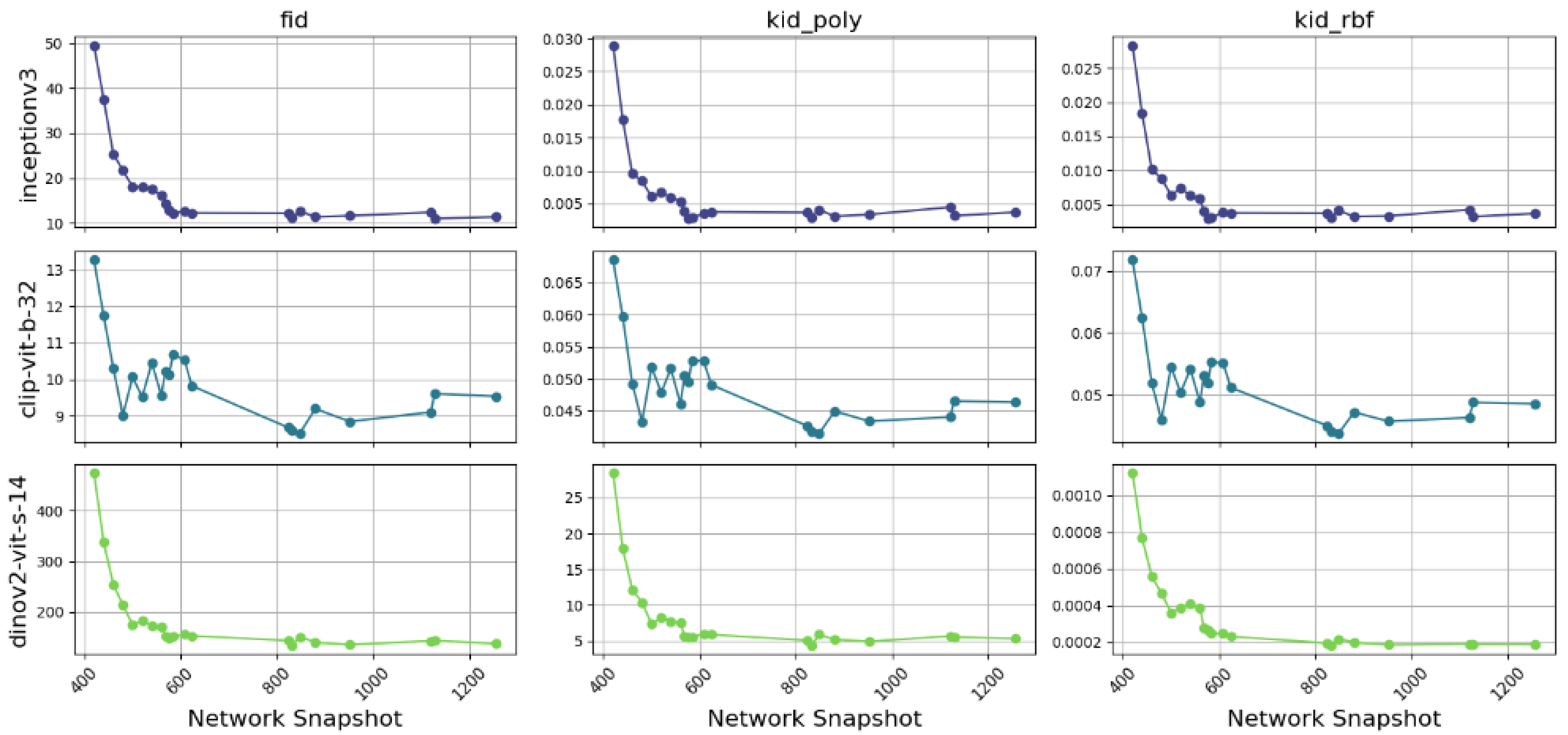
- PSNR (pixel-based)
- MS-SSIM (pixel-based)
- LPIPS
- DREAM-SIM

Distribution based metrics

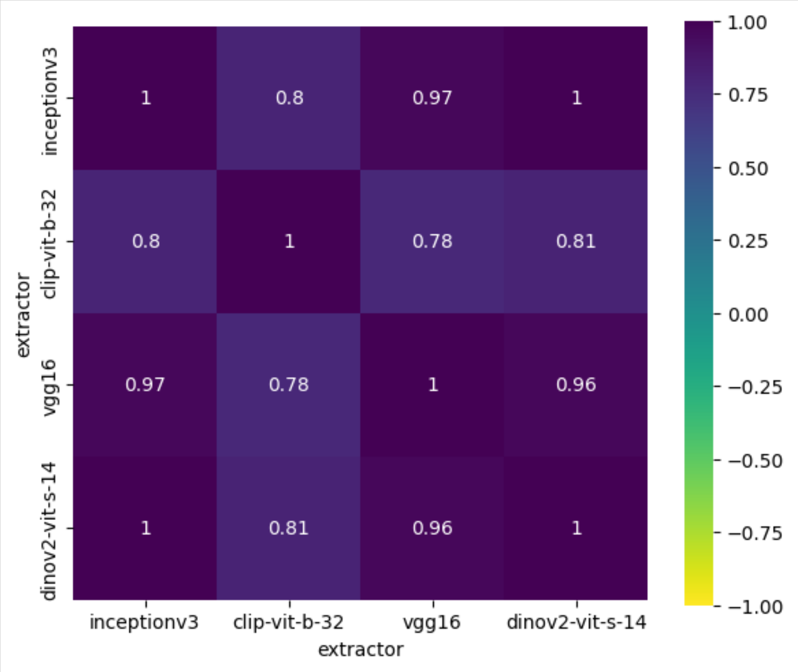
- ISC
- FID
- KID(poly), KID(rbf)
- CMMD
- Precision and Recall
- PPL

Stylegan2-ada-pytorch

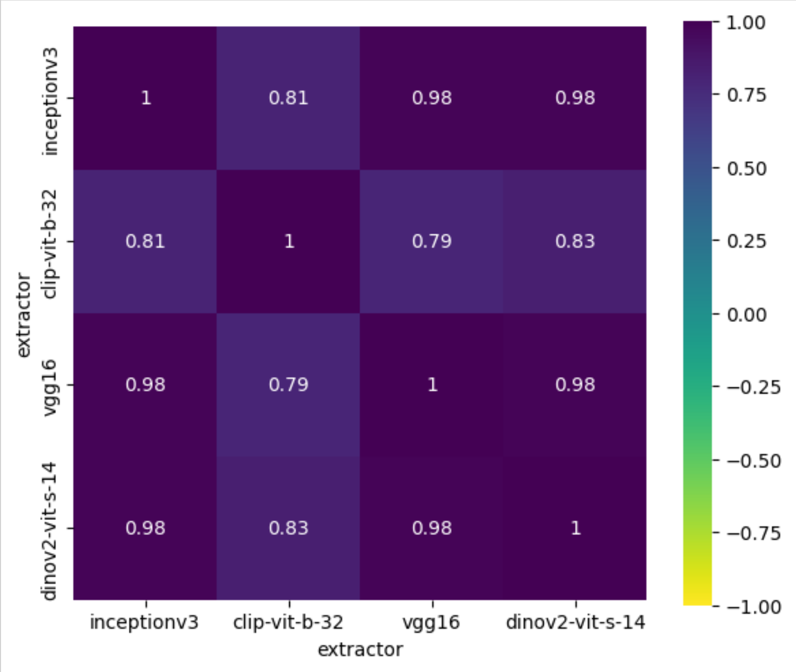
Metric	**StyleGAN2-ADA**
FID (fid50k)	9.94359
FID (fidelity)	11.669854
KID (fidelity poly)	0.003367
KID (fidelity rbf)	0.00329
ISC (fidelity isc)	5.3313
CMMD	2.172
PPL (pplzend)	42.2750
PPL (pplwend)	25.6877
PPL (pplzfull)	42.6170
PPL (pplwfull)	25.1294


fid
kid_poly
kid_rbf
inceptionv3
clip-vit-b-32
dinov2-vit-s-14
Network Snapshot
Network Snapshot
Network Snapshot

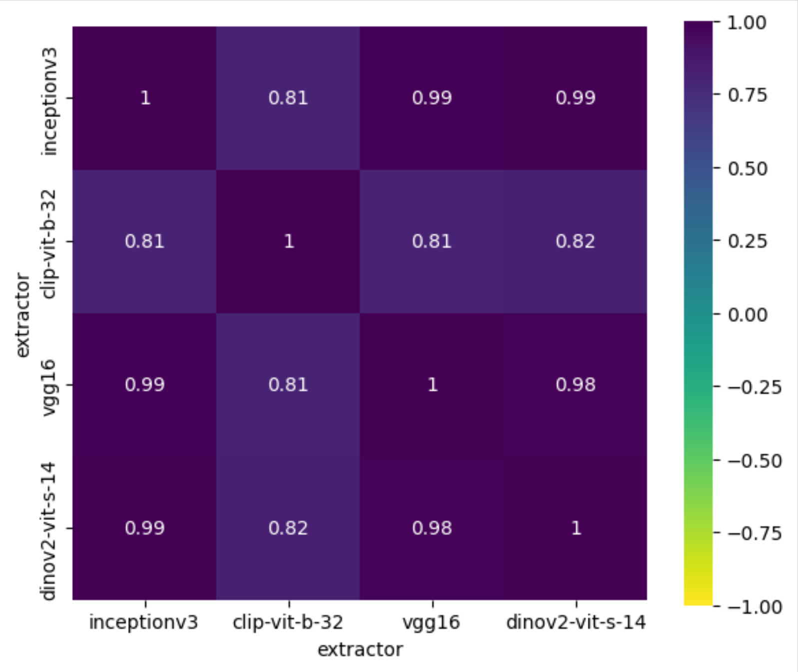
FID

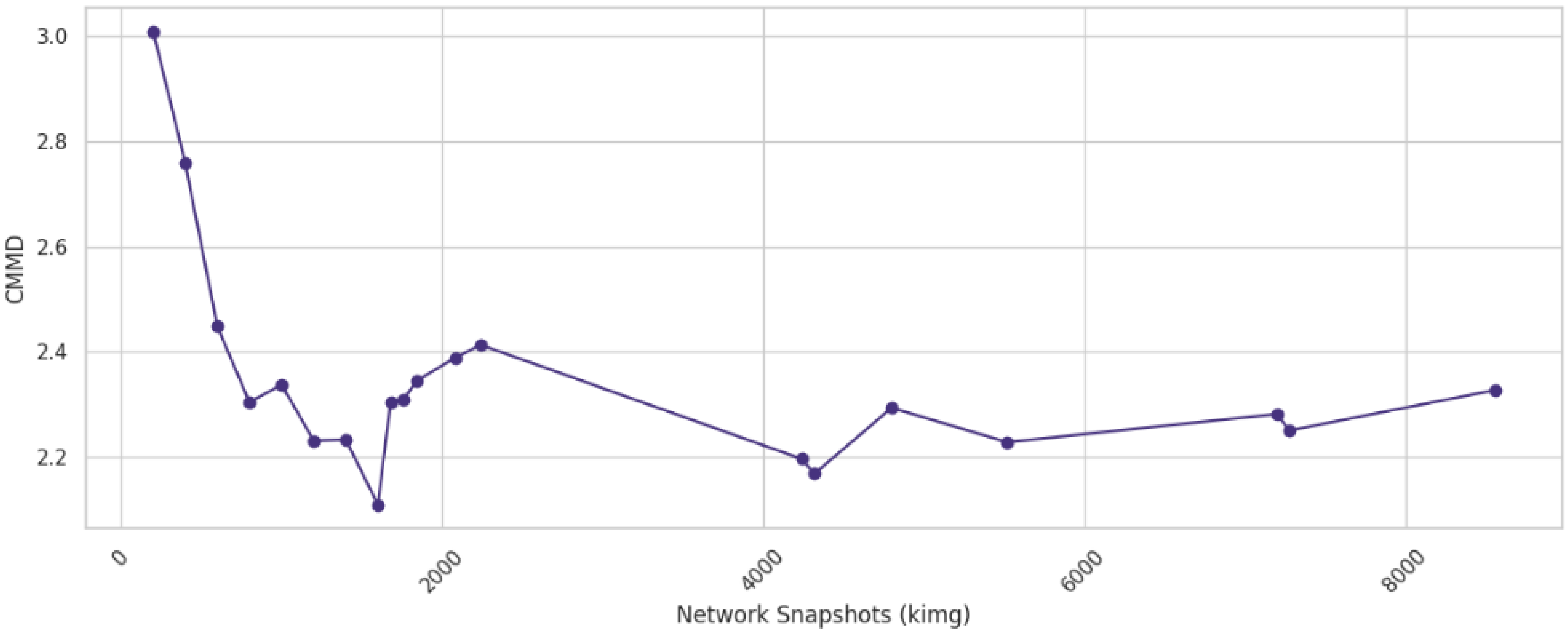


KID poly



KID rbf



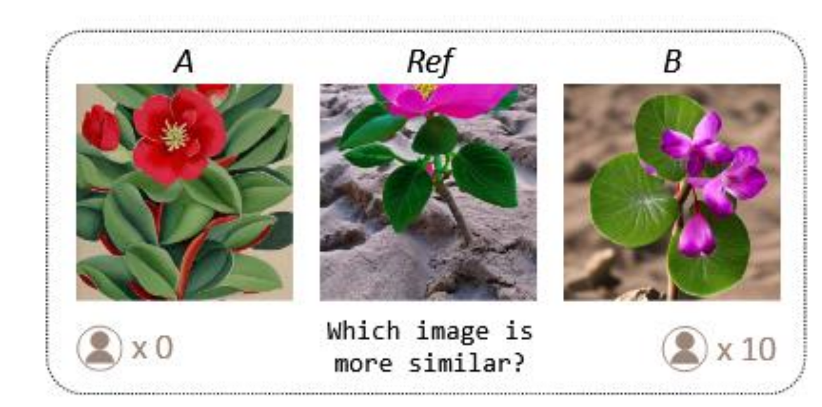


3.0
2.8
2.6
2.4
2.2
CMMD
0
2000
4000
6000
8000
Network Snapshots (kimg)

DreamSIM:

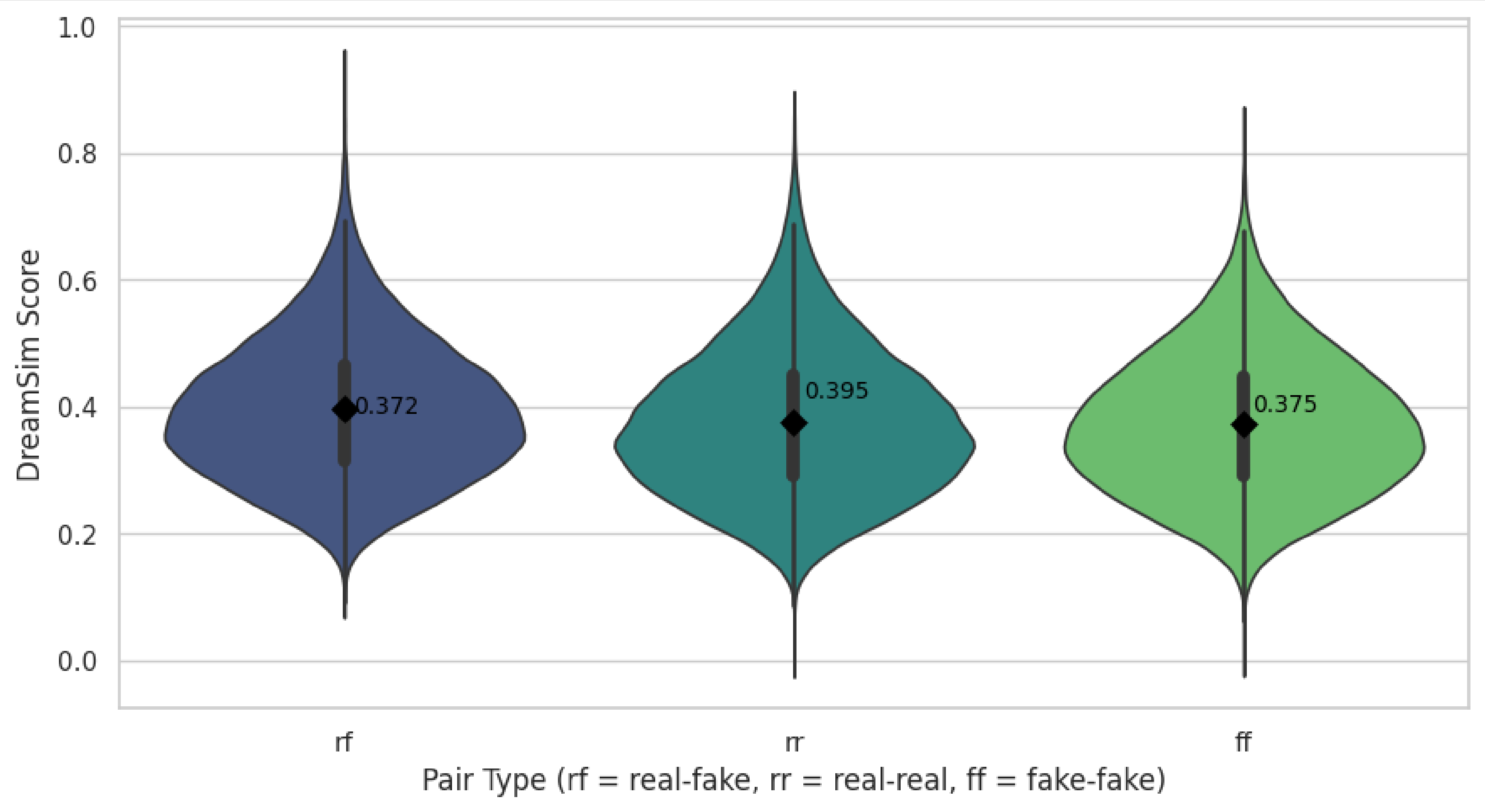
$$DreamSim(x, y) = \sum_{l \epsilon L} w_l \cdot sim(\theta_l(x), \theta_l(y))$$

- L is set of network layers used
- w_l are learned weights for each layer l
- $sim(\cdot,\cdot)$ is a similarity function (cosine-similarity, L2)
- $\theta_l(x)$ denotes feature representation of image x at layer l

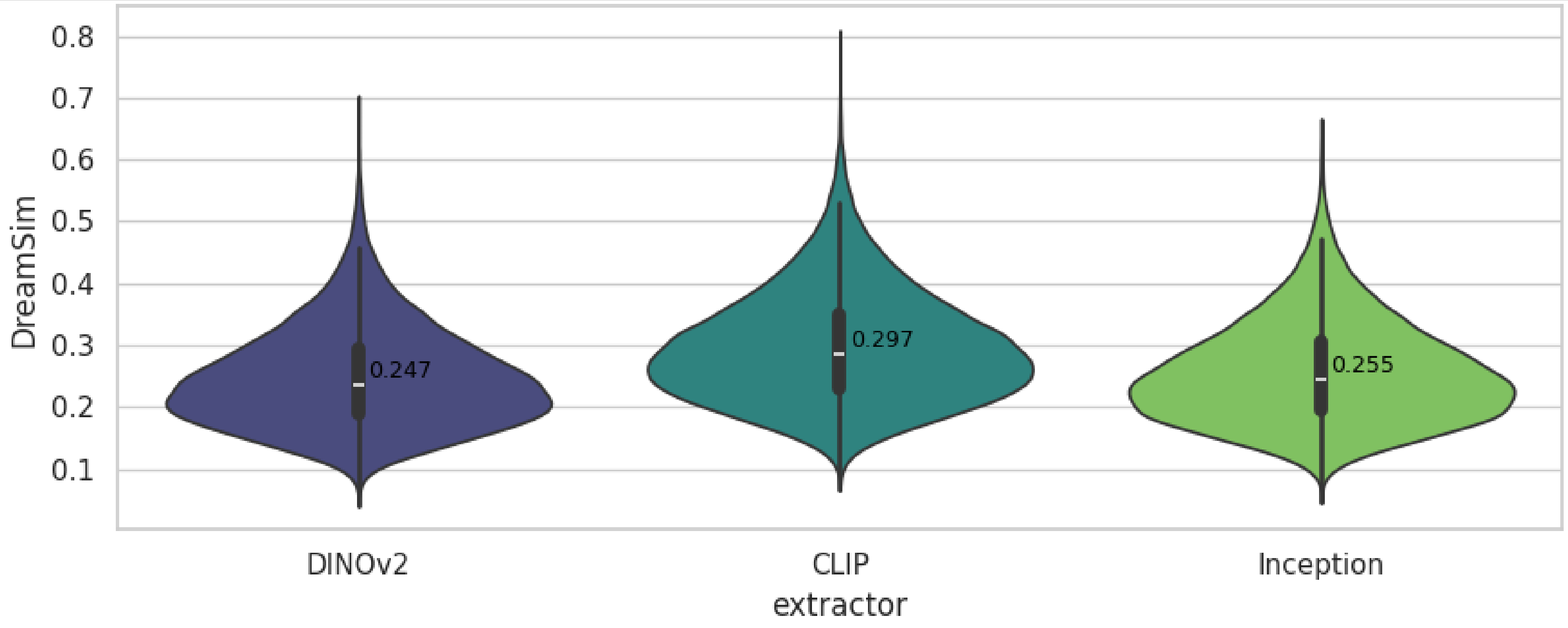


(Fu et al., 2023)

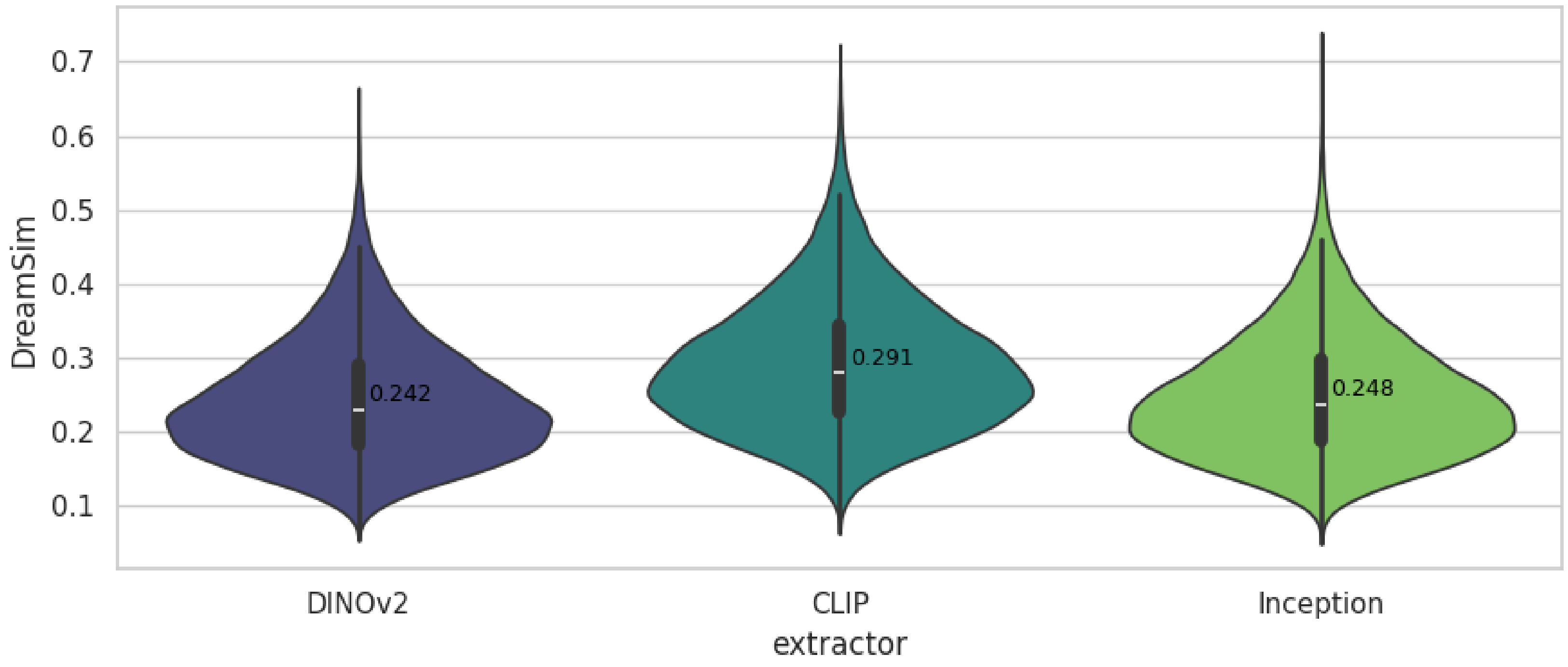
DreamSIM: Random matches



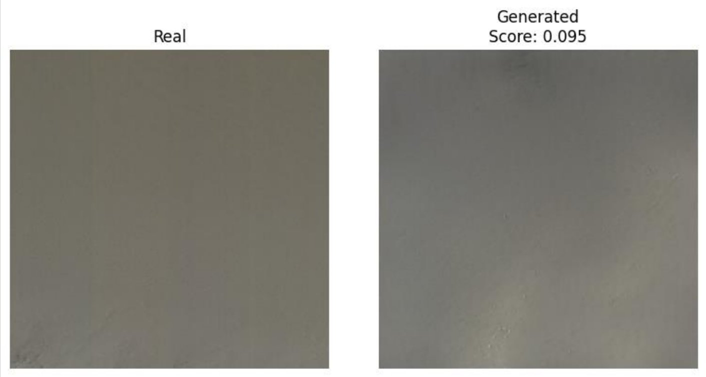
Nearest neighbor paring on **Test Data**



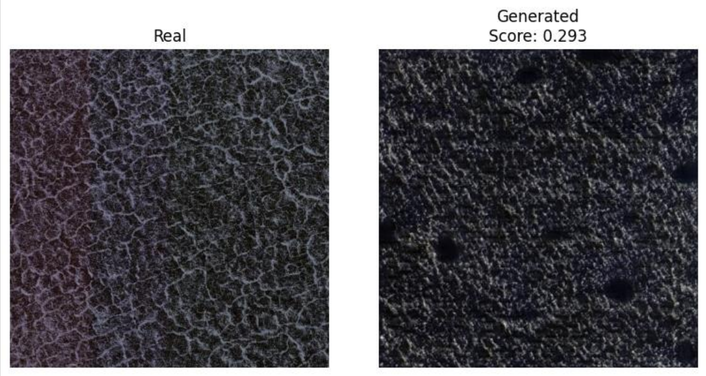
Nearest neighbor paring on **Training Data**

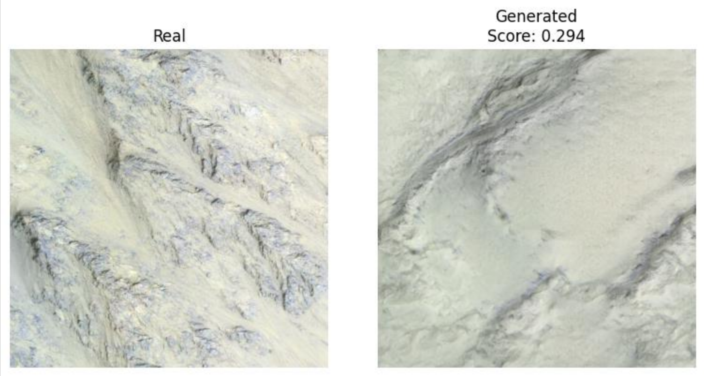


Best scores

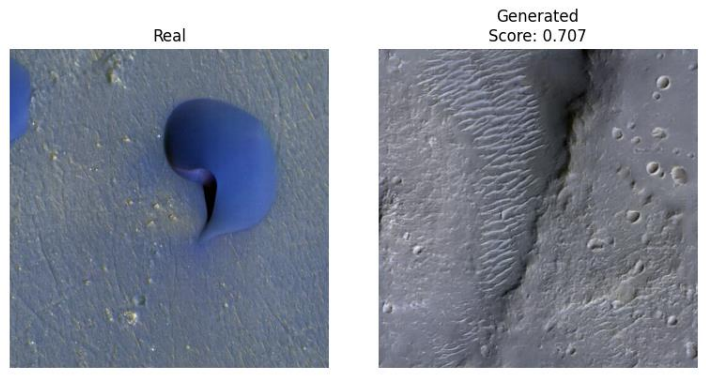


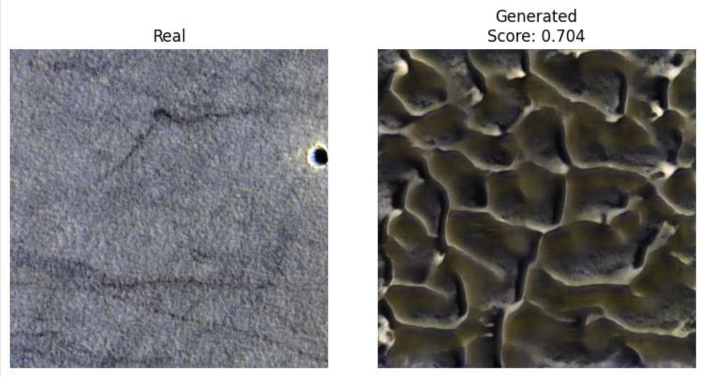
Average scores





Worst scores





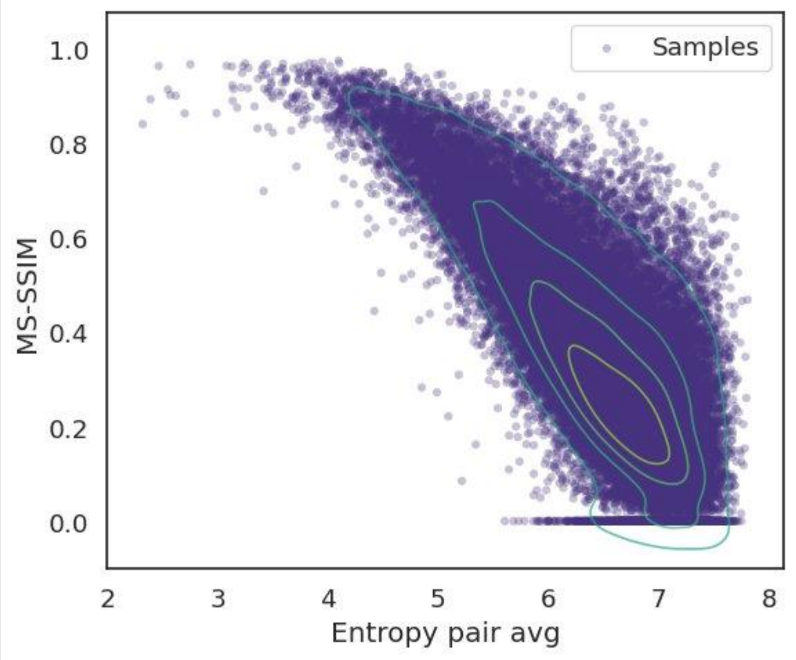
Entropy Bias:

$$H(p) = -\sum_{i=1}^{N} p_i \times \log p_i$$

Average over RBG and image pair

Entropy Bias:

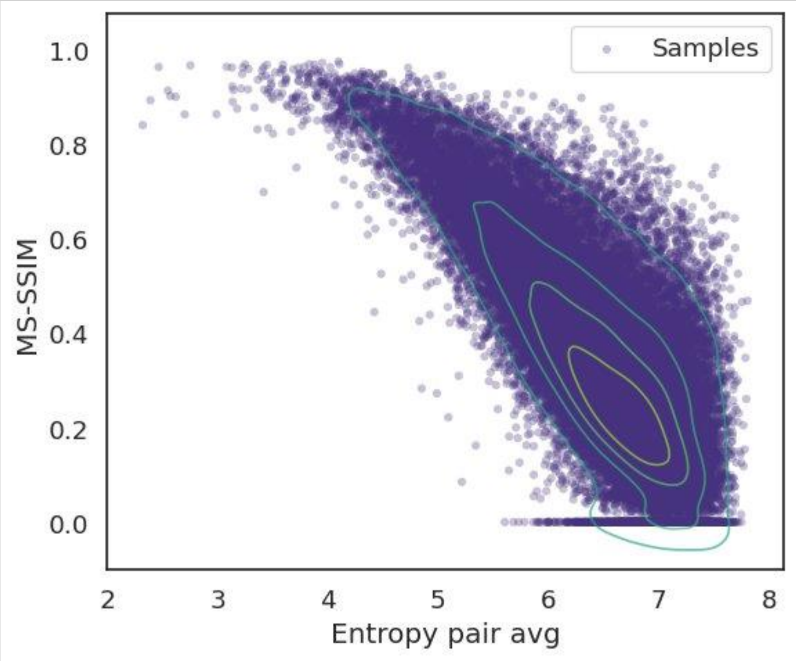
$$H(p) = -\sum_{i=1}^{N} p_i \times \log p_i$$ Average over RBG and image pair

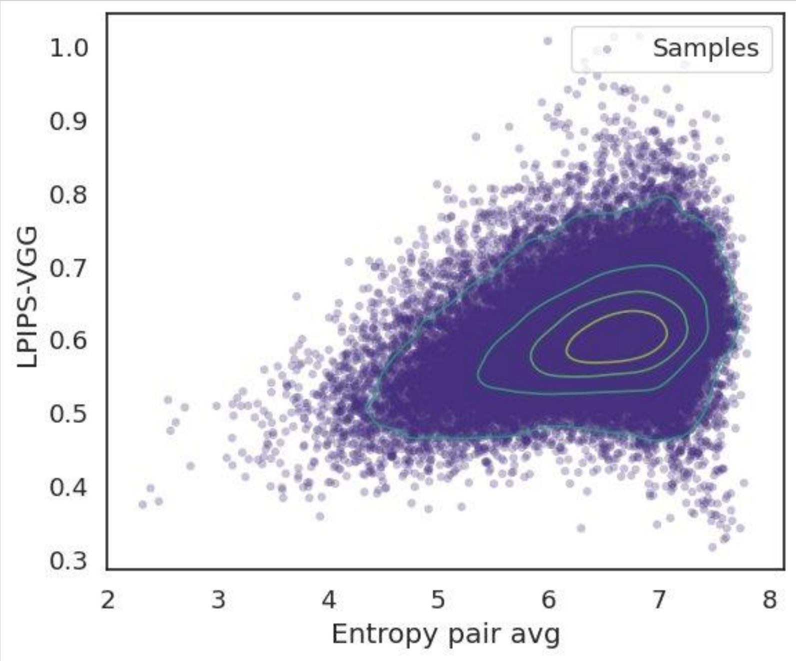


Entropy Bias:

$$H(p) = -\sum_{i=1}^{N} p_i \times \log p_i$$

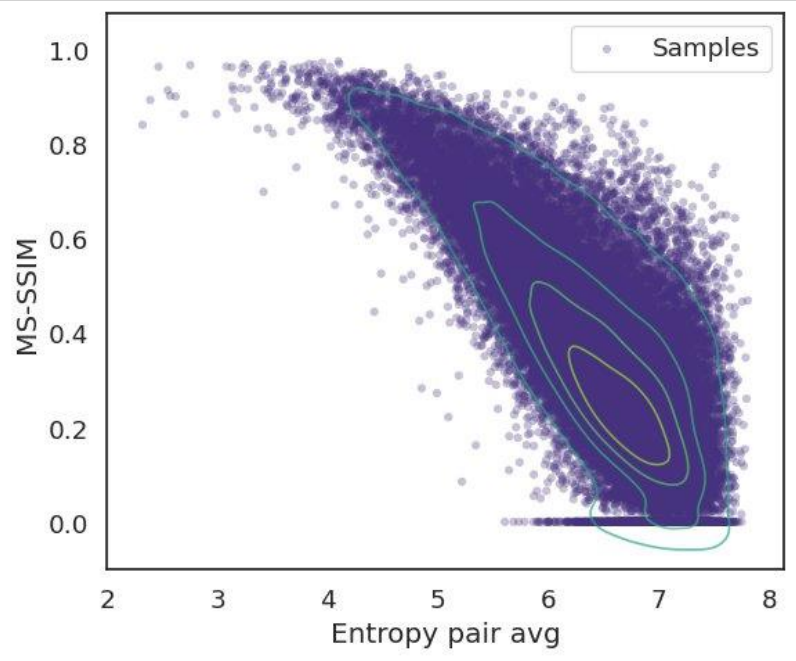
Average over RBG and image pair

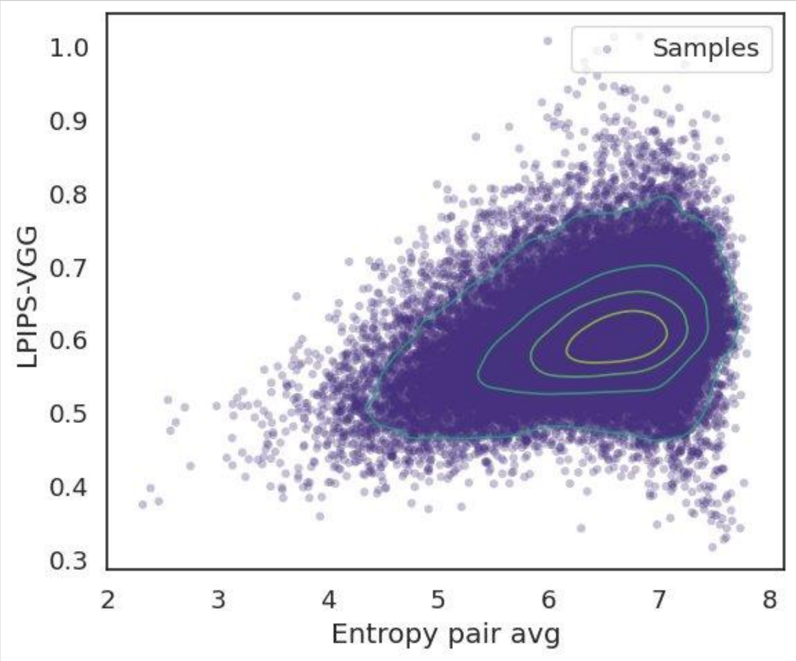


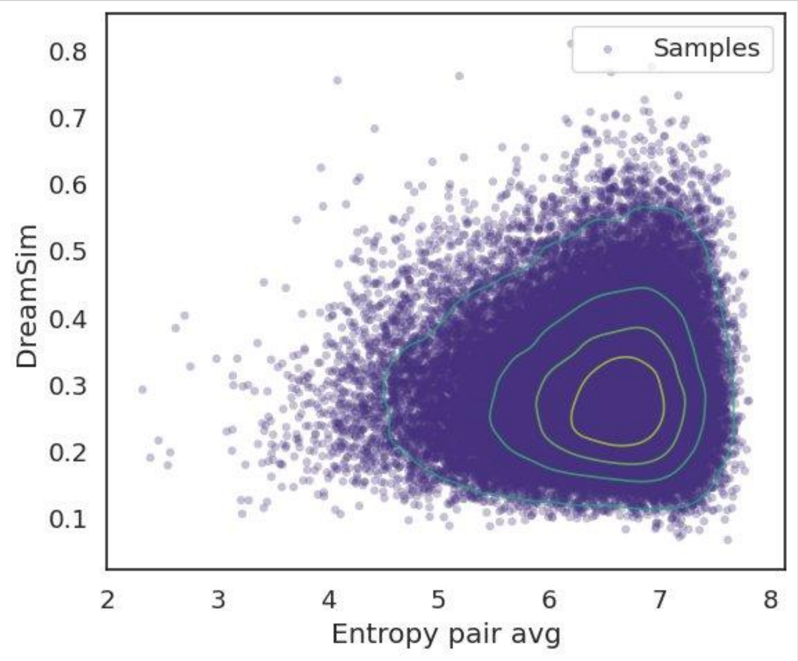


Entropy Bias:

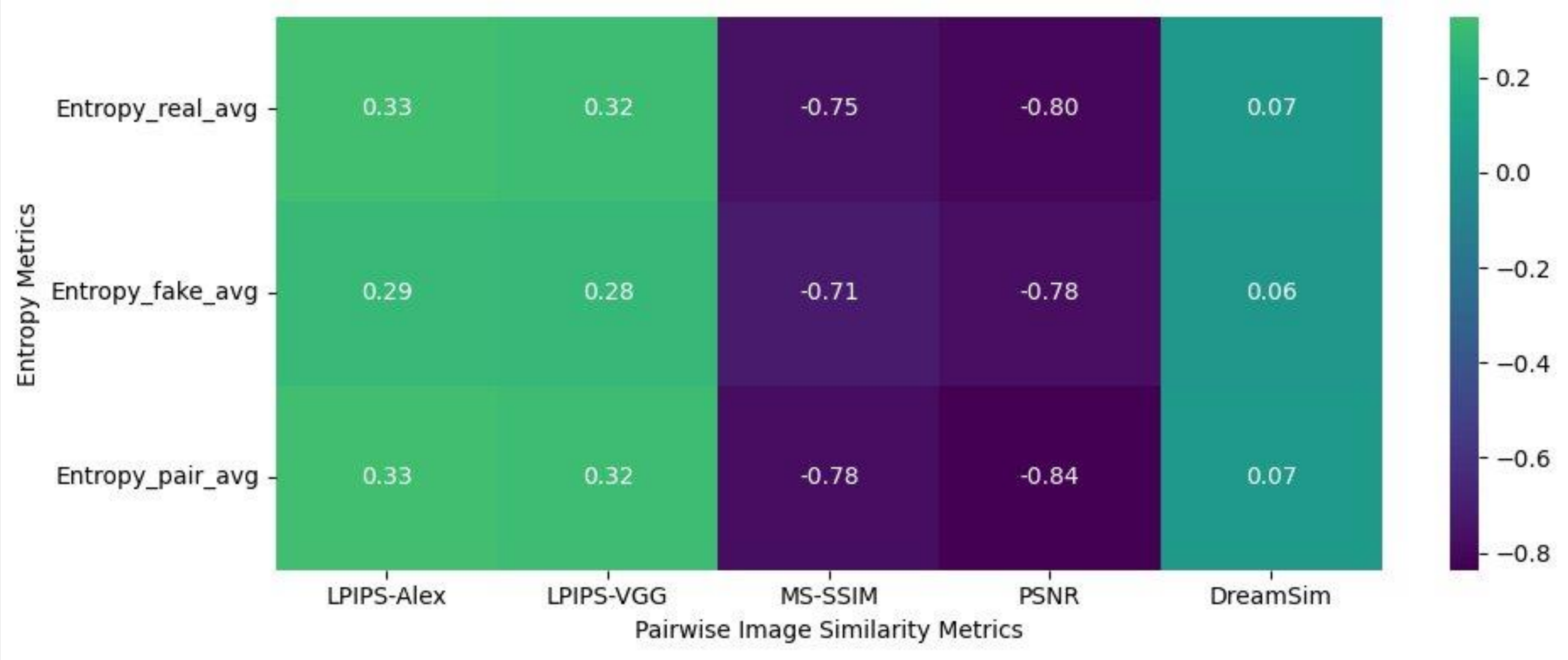
$$H(p) = -\sum_{i=1}^{N} p_i \times \log p_i$$ Average over RBG and image pair







Pearson Correlation between Metrics & Entropy



Conclusion

- Fidelity and diversity images compared to benchmarks on popular datasets
- Choice of feature extractor: CLIP << Inception < DINO
- Pixel-based metrics have entropy biases, reduces for features-based metrics, vanishes when human aligned.
- No single reliable metric -> evaluation domain and application specific.

Outlook and open questions

Clip’s “Bump” -> sensitive to color?

Image similarity metrics bias toward simple structure?

Better nearest neighbor matches?

Are models aligned with human perception?


Thank you for listening!