Benchmarking the Alignment of Data-Quality Metrics, Human Judgment and Land-Cover Segmentation Performance for Earth Observation

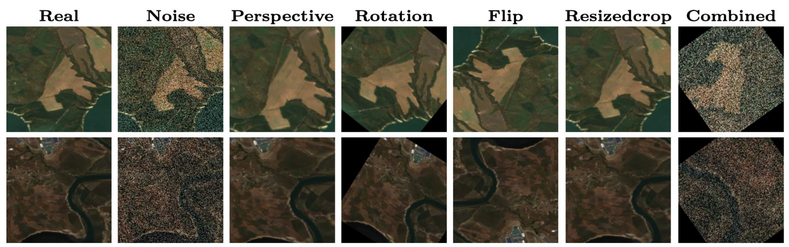
Ümit Mert Çağlar, Alptekin Temizel
Graduate School of Informatics, Middle East Technical University, Türkiye

ECCV 2026 | MALMÖ | SEPT 8–12

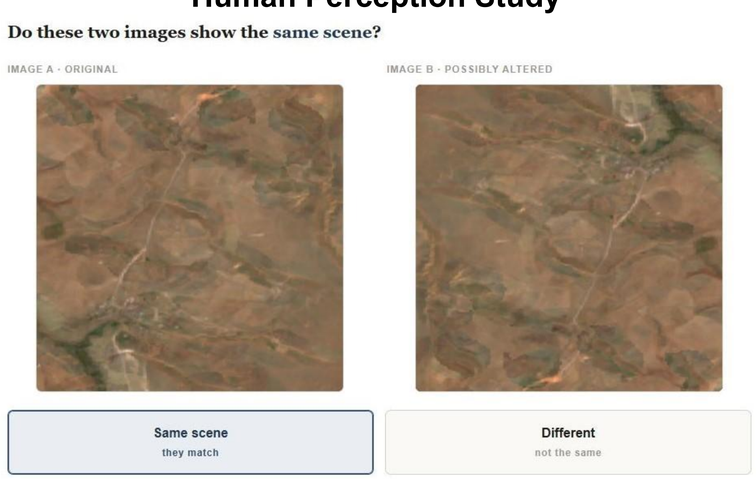
ABSTRACT

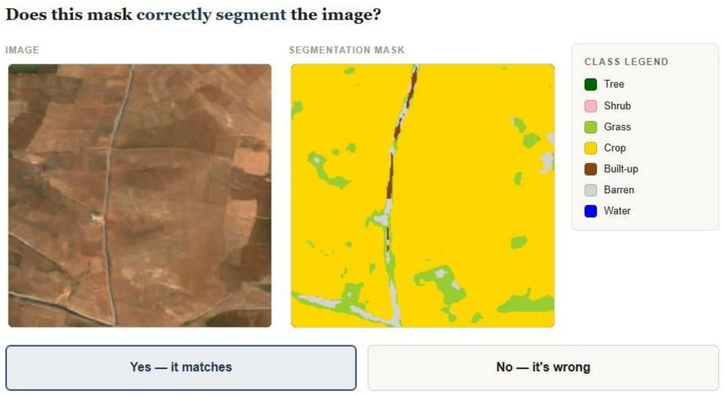
- Systematic comparison of automated metrics, human judgment and segmentation utility
- Perception study: 95 participants, 5,768 responses
- Some perturbations distort metric scores sharply
- Synthetics confused as real despite weaker scores
- Even weak scoring synthetic data can improve downstream performance when blended with real data
- Task performance unpredictable with quality-metrics
- Practical guidelines for synthetic data curation
- Used datasets across four geographic regions: Türkiye, Europe, Rep. of Korea, California-Nevada
- Synthetic data generated with conditional Stable Diffusion, UNet-GAN and unconditional StyleGAN3
- Perturbations (Noise, Perspective, Rotation, Flip, Resized Crop, and Combined) are applied to evaluate metric sensitivity against human perception limits.

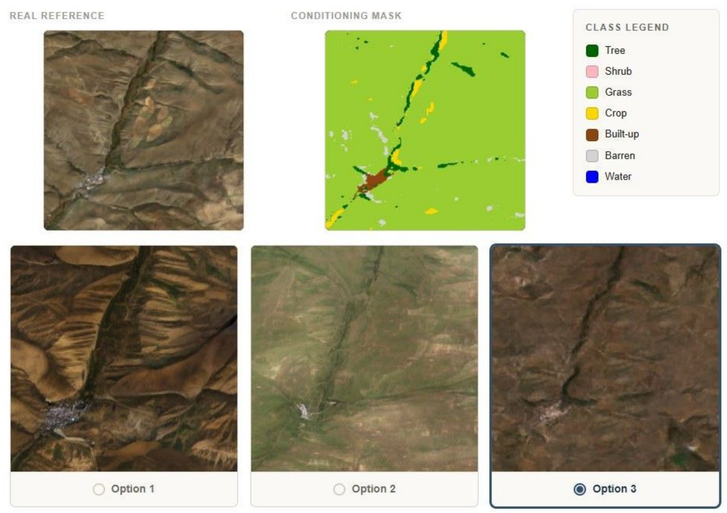
Perturbations

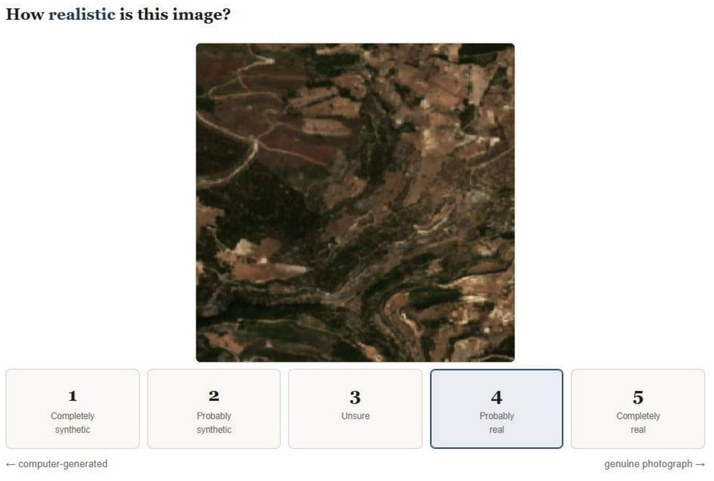


Human Perception Study





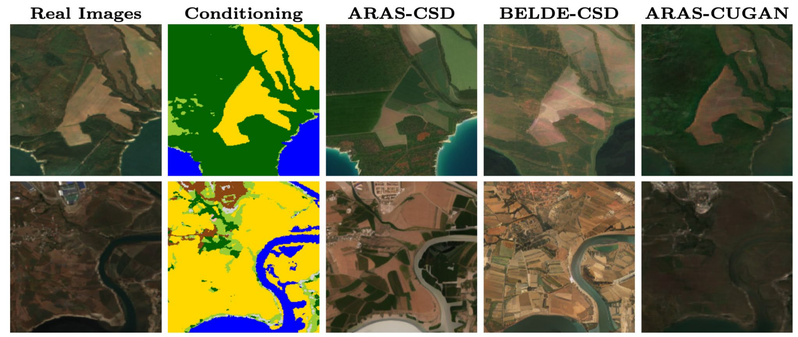




RESULTS

Train	Test	FID ↓	HPQS ↑	F1 (%) ↑
ARAS	ARAS	2.09	3.54	76.5 ± 0.4
SGAN3	ARAS	16.85	3.43	73.2 ± 0.4
SGAN3-D	ARAS	16.68	3.28	74.8 ± 0.2
CUGAN	ARAS	72.23	2.96	54.3 ± 0.7
ARAS, SGAN3, CUGAN	ARAS	13.19	3.31	77.2 ± 0.4
ARAS, CUGAN	ARAS	21.27	3.25	77.6 ± 0.7
ARAS, SGAN3	ARAS	12.12	3.49	77.7 ± 0.3
BELDE	BELDE	0.36	3.32	83.0 ± 0.5
BELDE	BELDE-CA-NV	20.4	3.09	66.2 ± 0.9
BELDE	BELDE-K	41.6	3.42	58.3 ± 0.2

Segmentation models trained and tested with different datasets to measure downstream performance and compared with the FID score (automated data quality) and Human Perception Quality Score (HPQS)



Synthetic data samples generated with DDPM and GAN models conditioned on the segmentation maps

CONCLUSION

- Automated metrics unreliable for synthetic data pruning
- Human perception also insufficient to predict utility
- Downstream task performance not correlated with neither automated metrics nor human perception
- Orientation-sensitive feature extractors linked to false pruning triggers, they can remove useful data!
- Task-aware evaluation suggested over metric curation
- Joint fidelity-and-utility evaluation framework proposed for synthetic Earth-observation data
- Better (low) FID does not guarantee better utility!