

**Providing Trust  
in  
Data-Driven Innovation  
with  
AI-Generated Synthetic Data**

[sanparith.marukatat@nectec.or.th](mailto:sanparith.marukatat@nectec.or.th)

- **NSTDA** (National Science and Technology Development Agency)
- **NECTEC** (National Electronics and Computer Technology Center)
- **AINRU** (Artificial INtelligence Research Group)
- **IPU** (Image Processing and Understanding)



## Defining data driven innovation

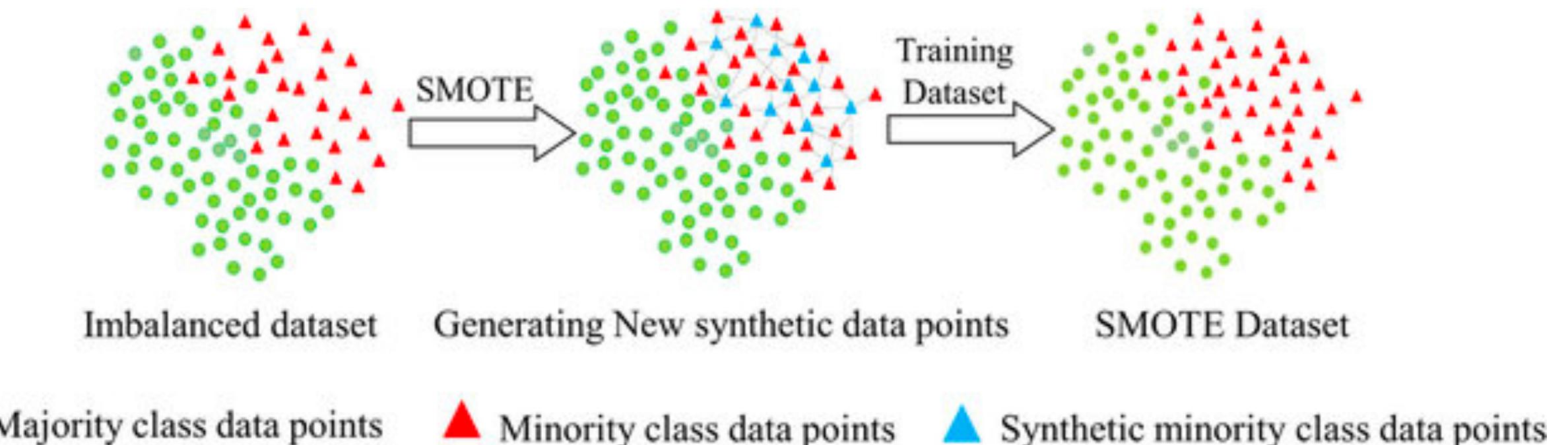
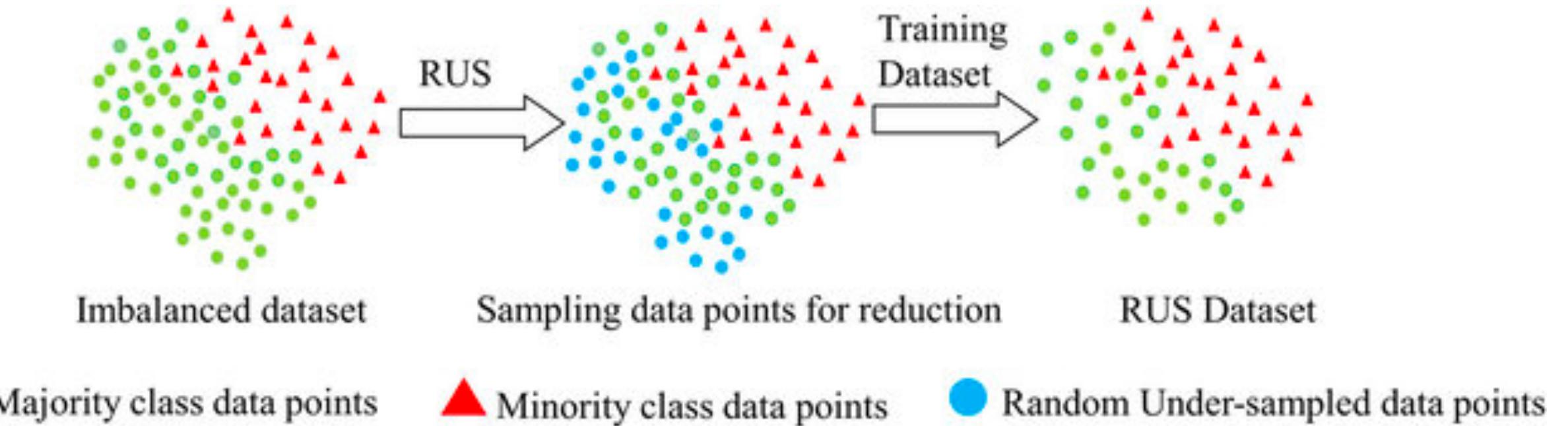
DDI is the innovation and consequent economic and social value that arises from the use of data analysis by private and public sector organisations to make better decisions and create new products and services. DDI can support a very wide range of innovations, including improving firm operational efficiency, developing new products and services, making better investment and strategic decisions, and more effective government interventions.

# Data related problem

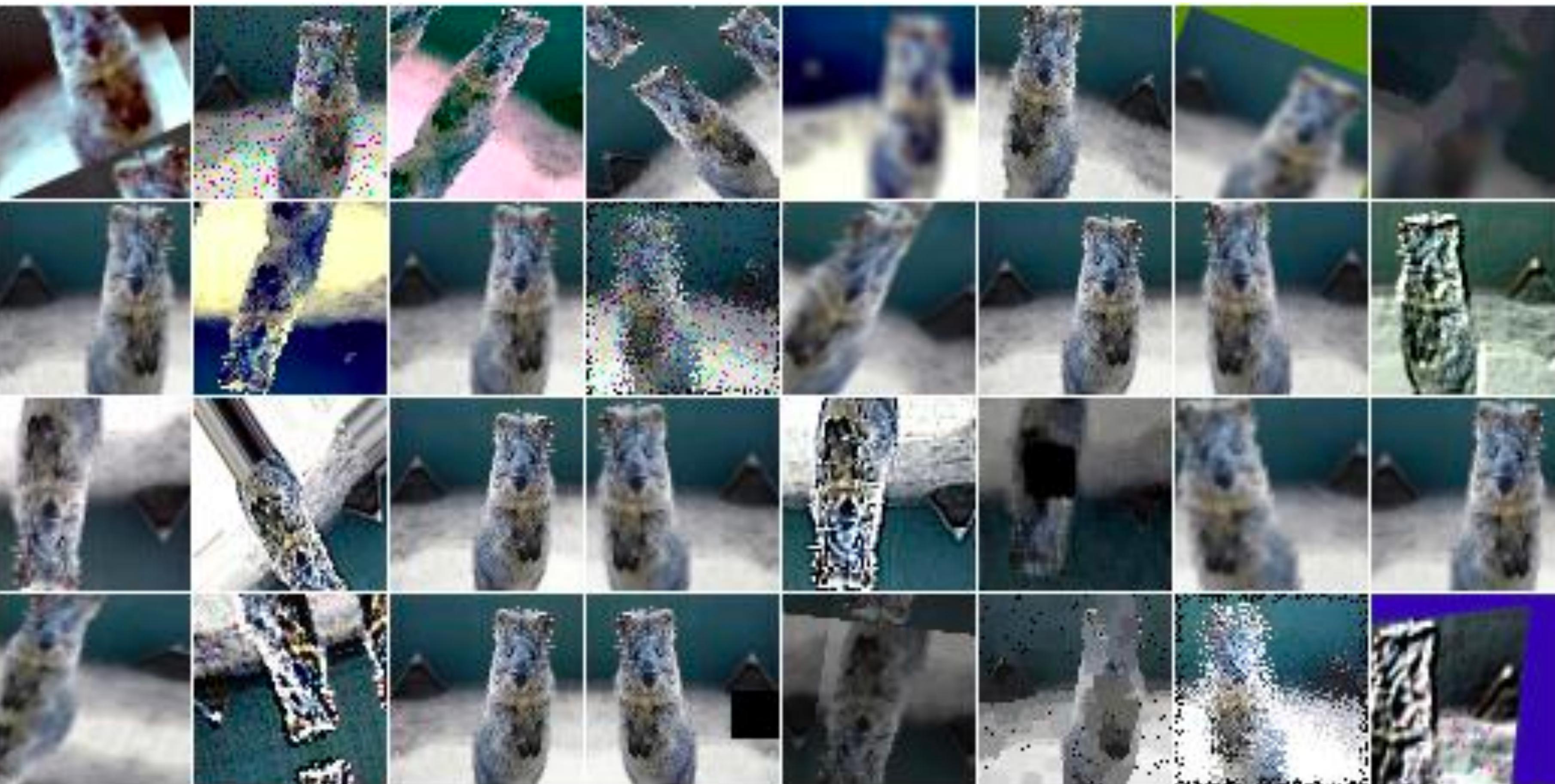
- Who own data?
- Privacy
- Biases
- Unbalanced data
- Data missing & imputation
- Data sovereignty
- ...

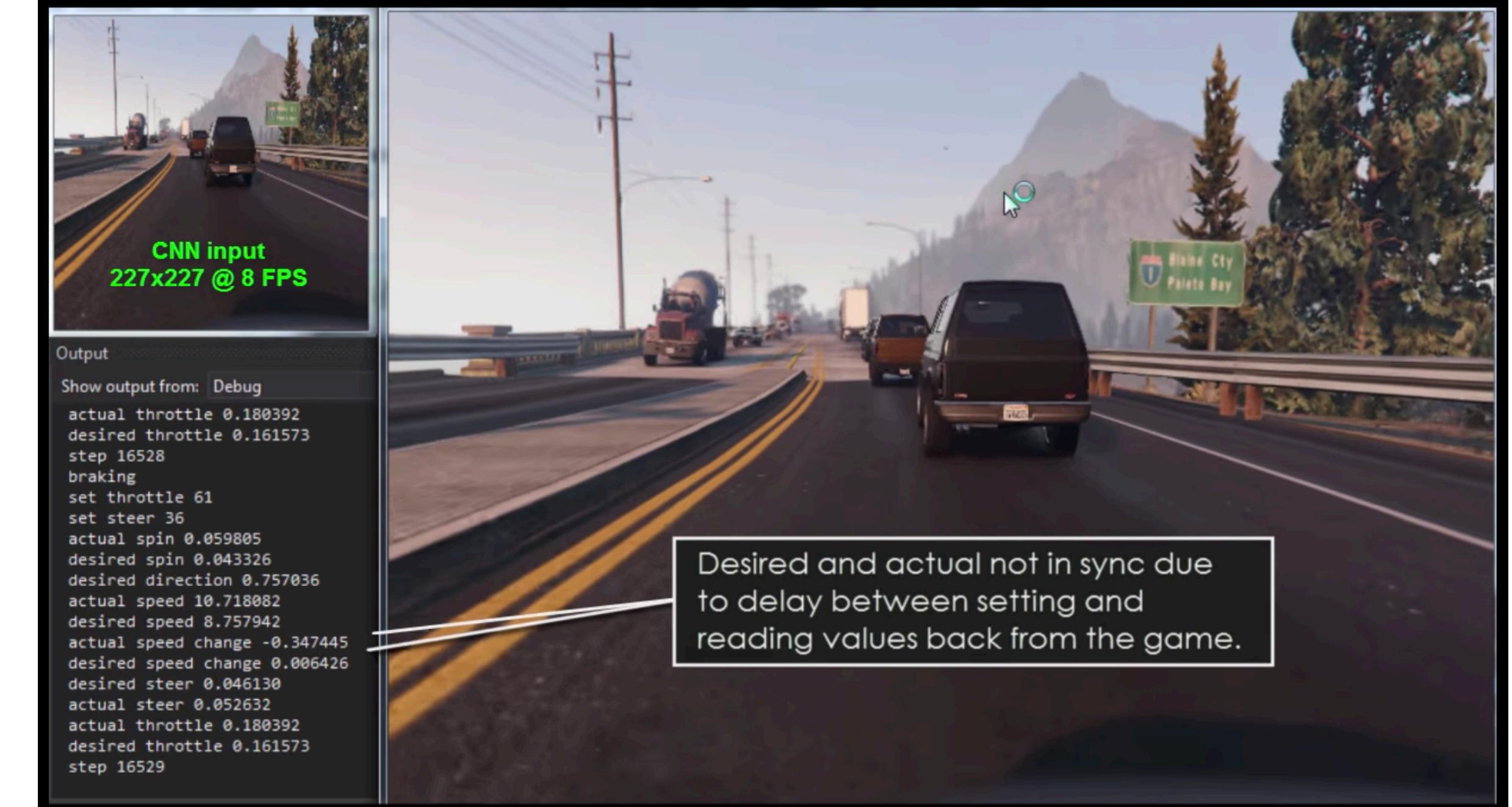
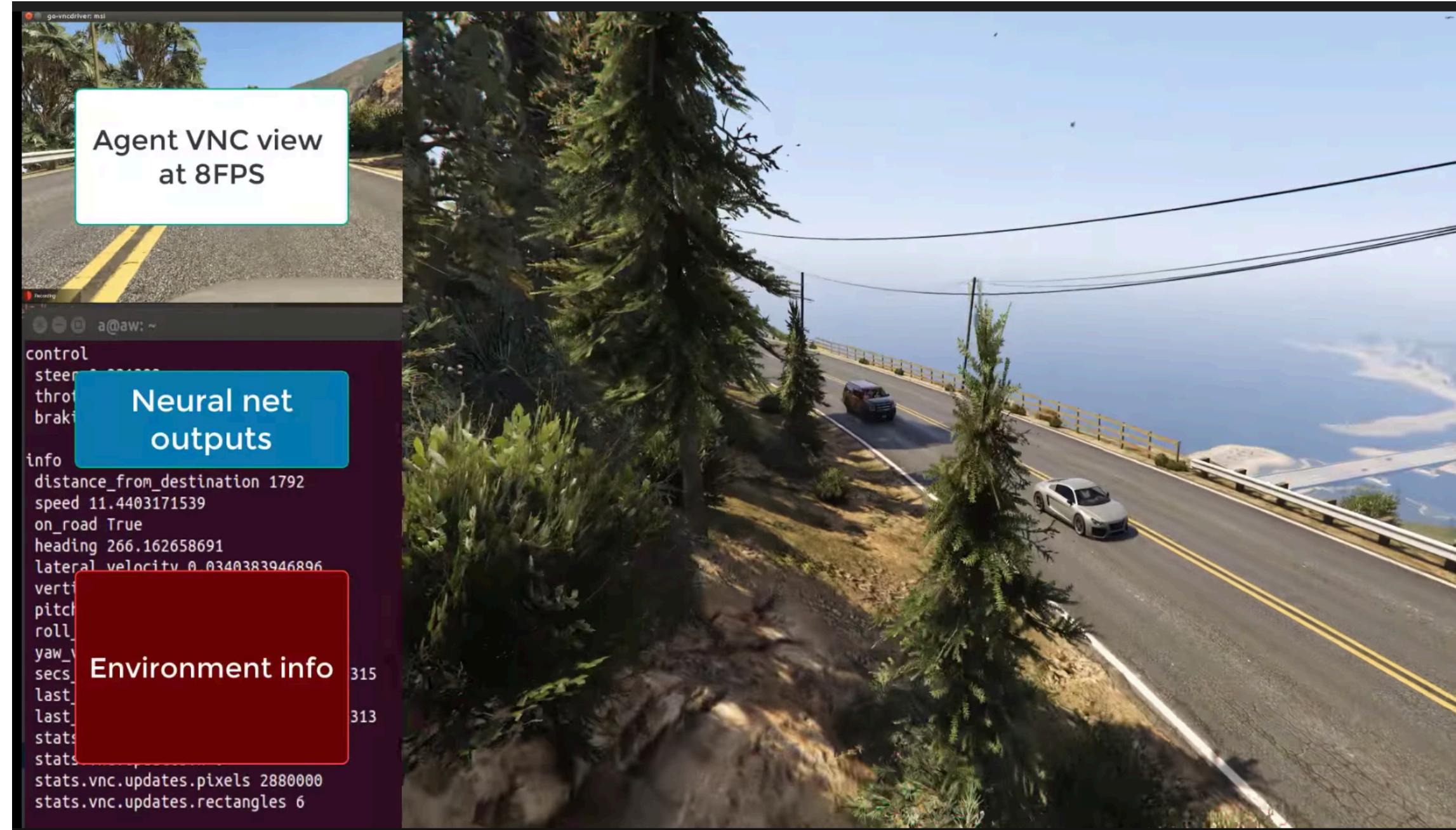
# "Classical" solutions

- Resampling
- SMOTE
- Image augmentation
- Use game and simulators



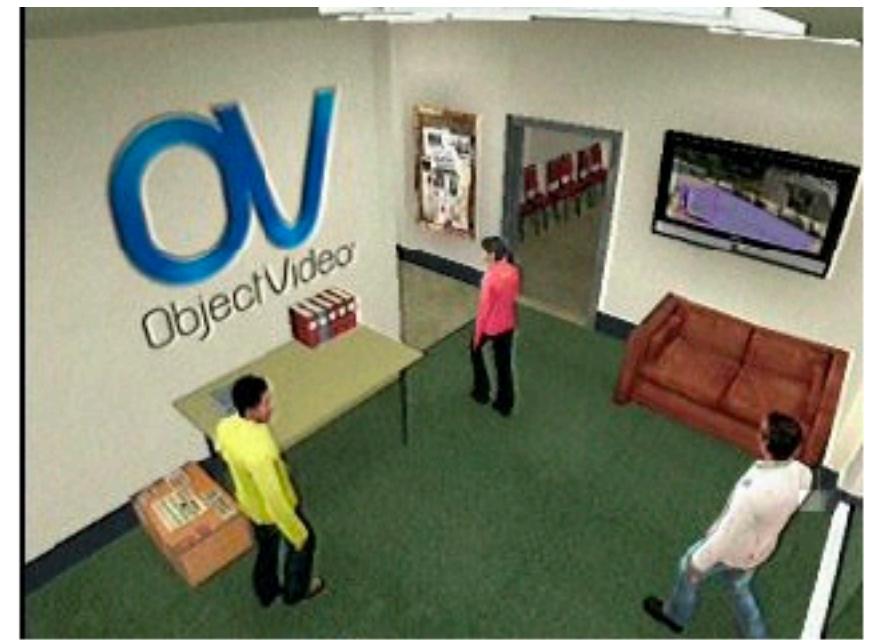
# Image augmentation





# GTA V + Open AI

# Half-life



(a) Office lobby



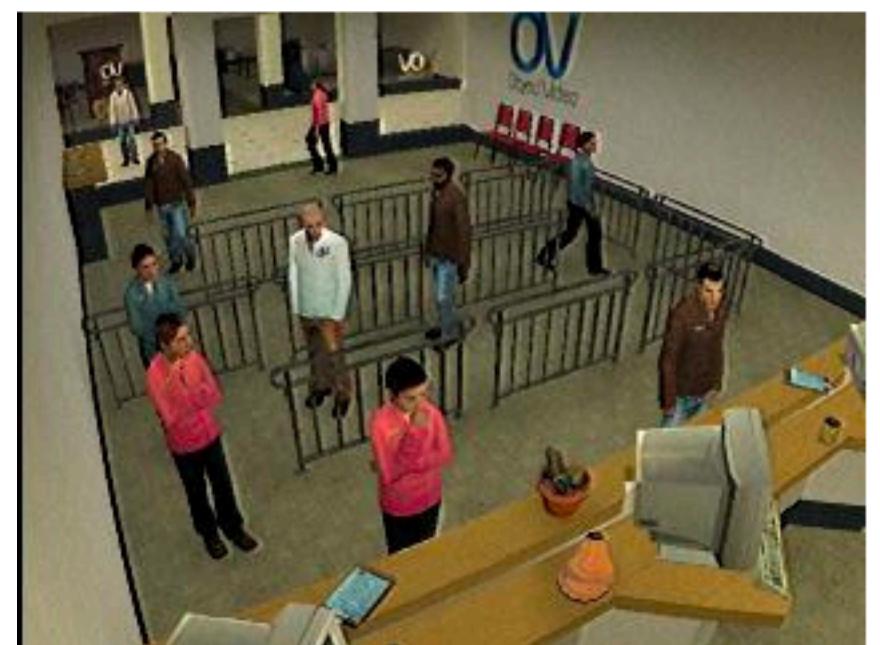
(b) Conference room



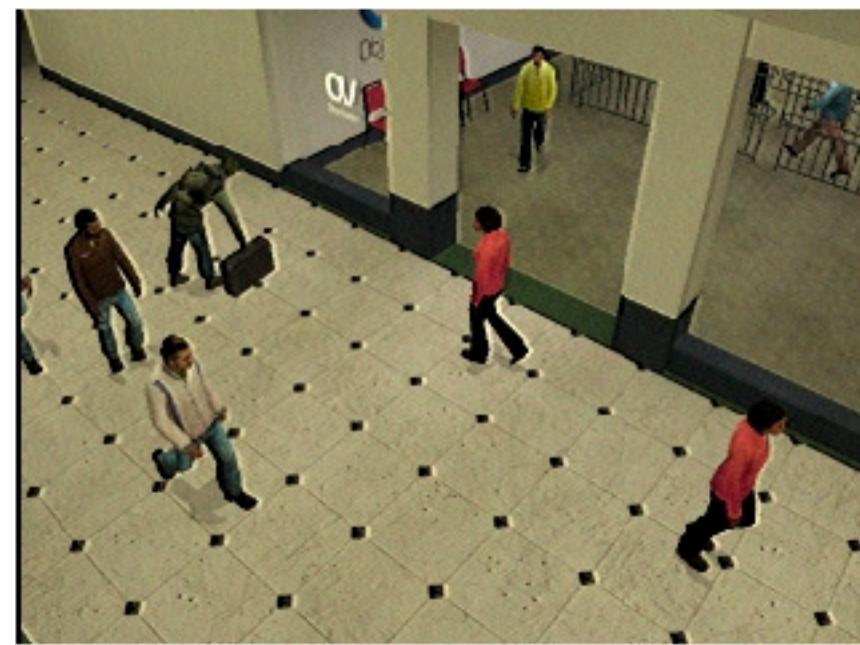
(a) Street view



(b) UAV view



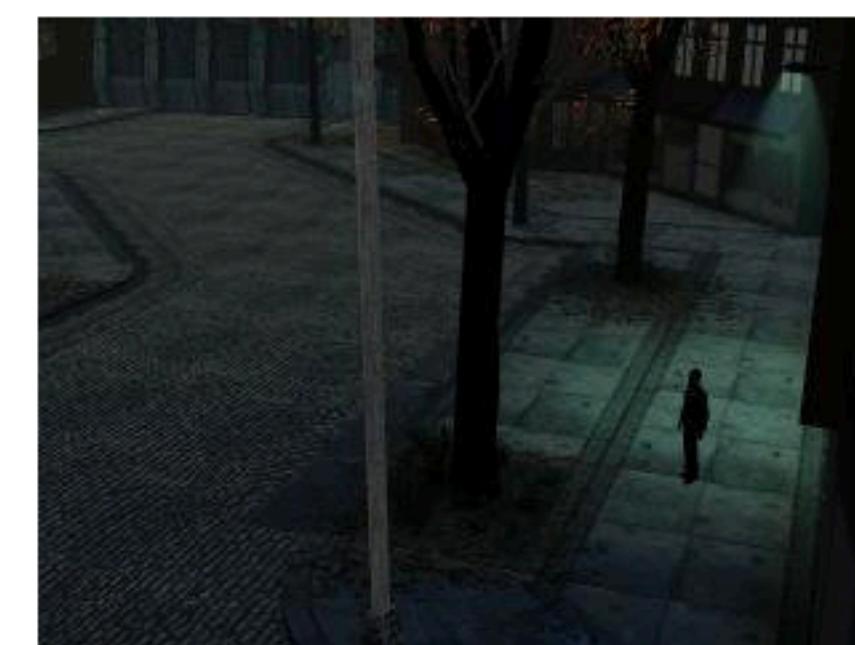
(c) Customer queue



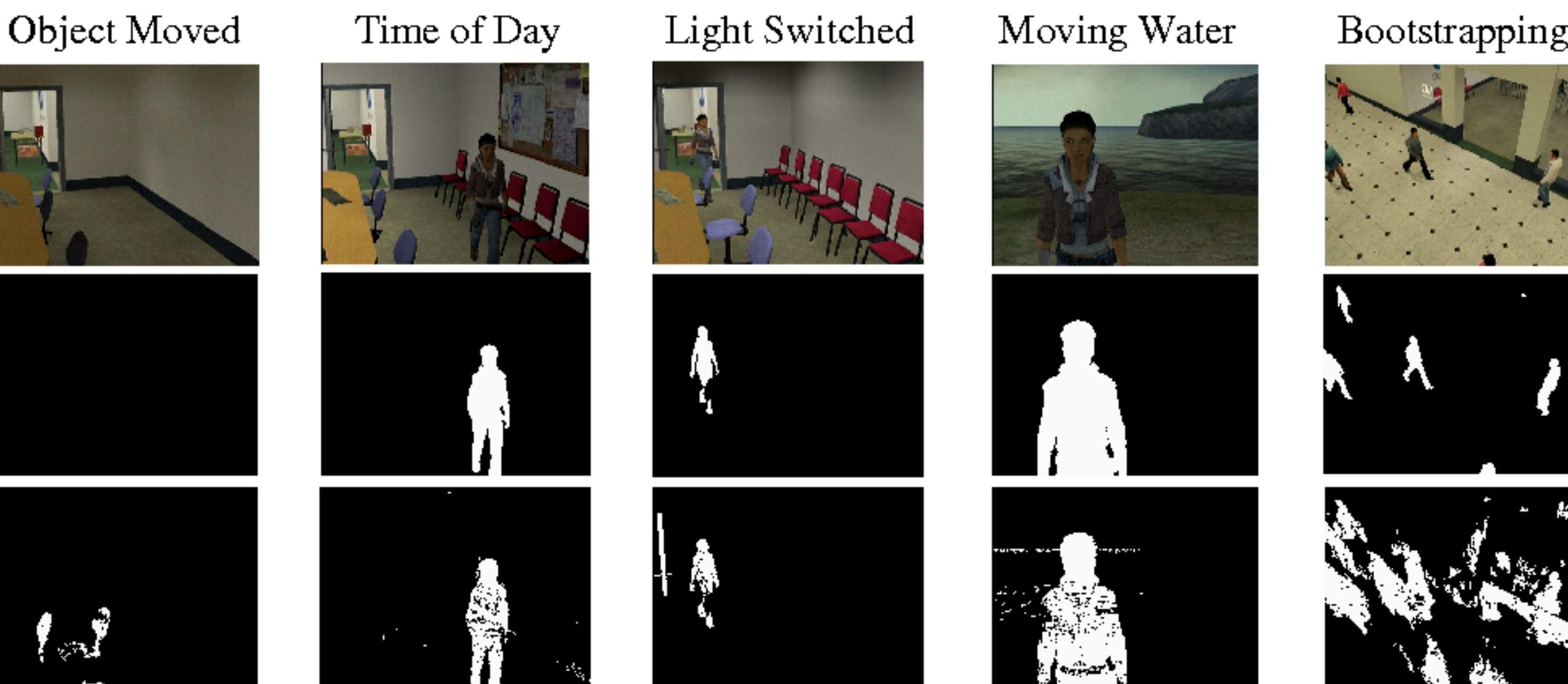
(d) Dropped bag



(c) Rain and fog

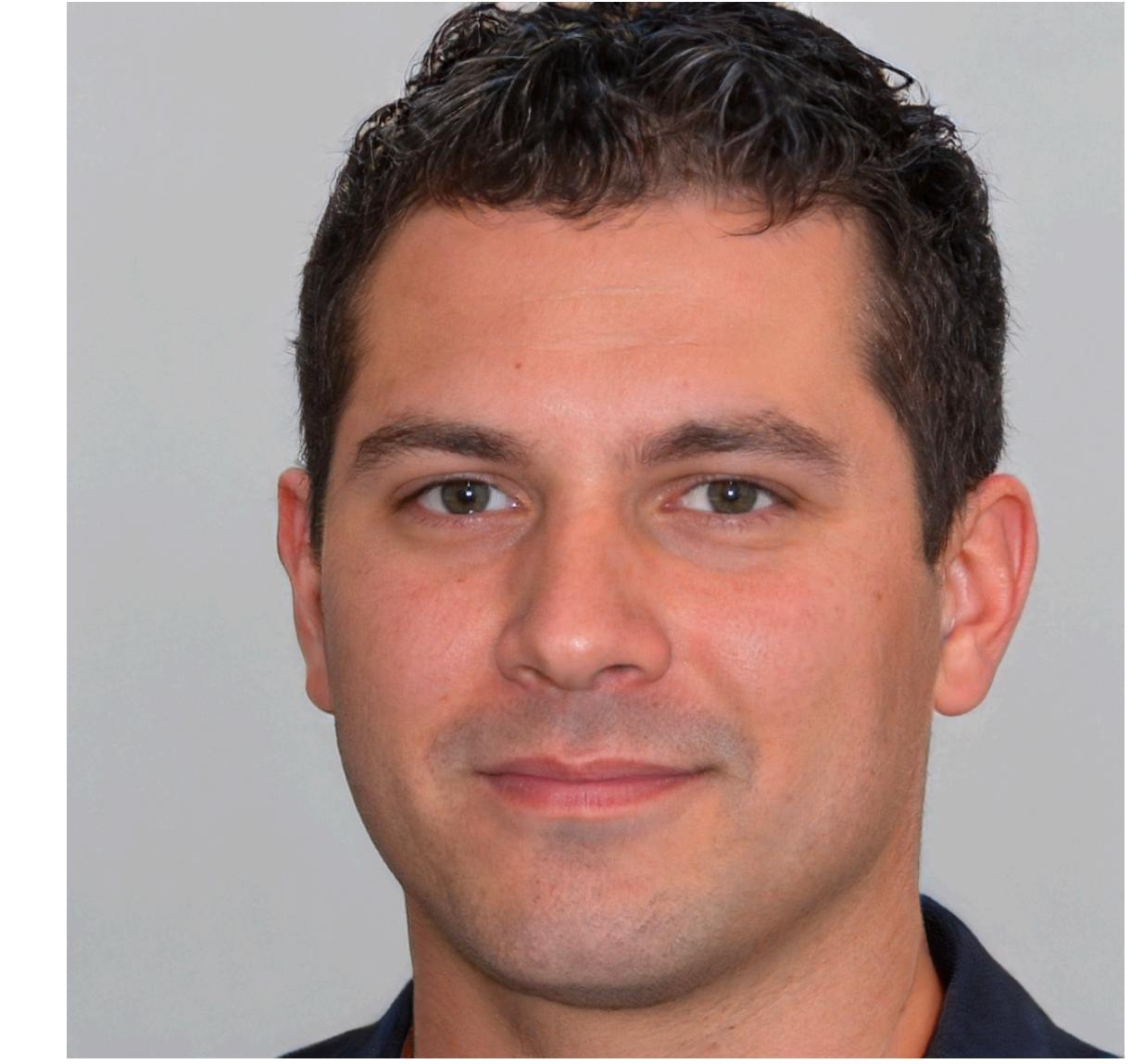


(d) Low light

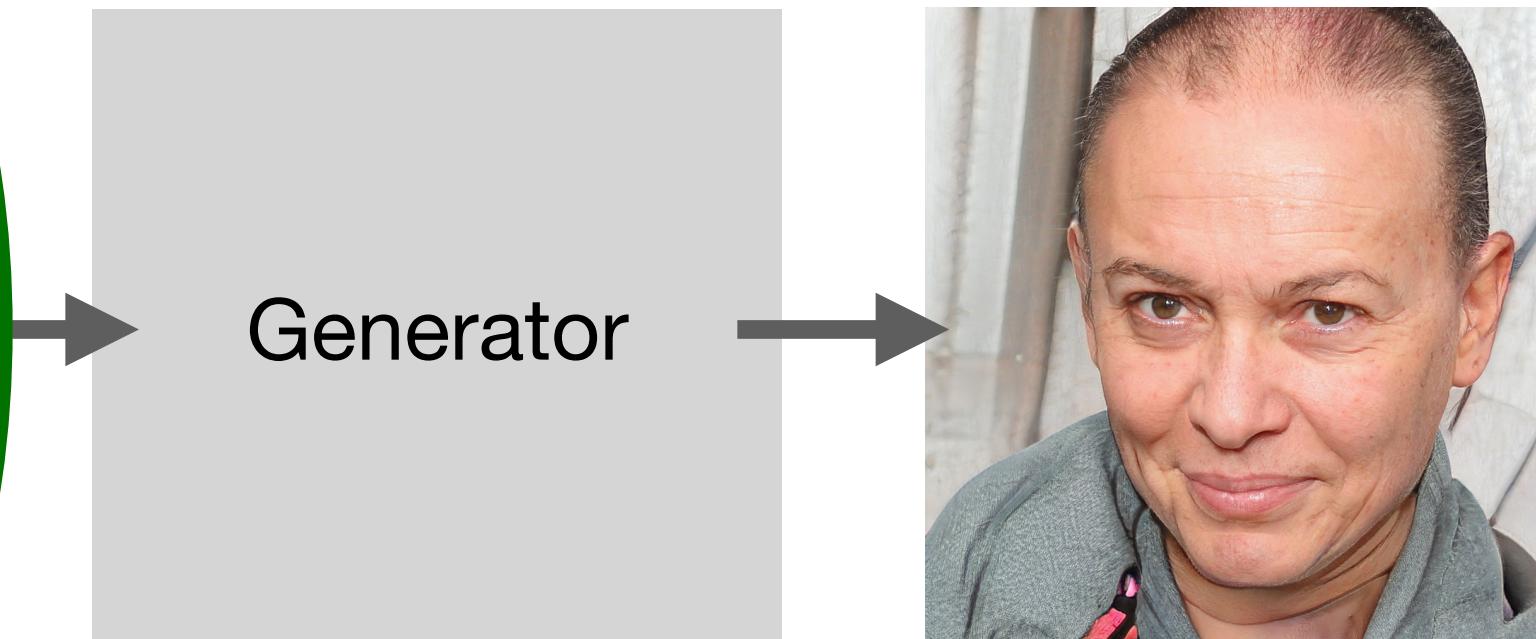
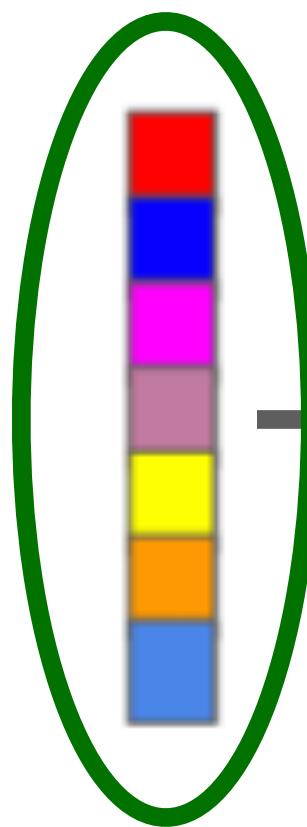


# New solution

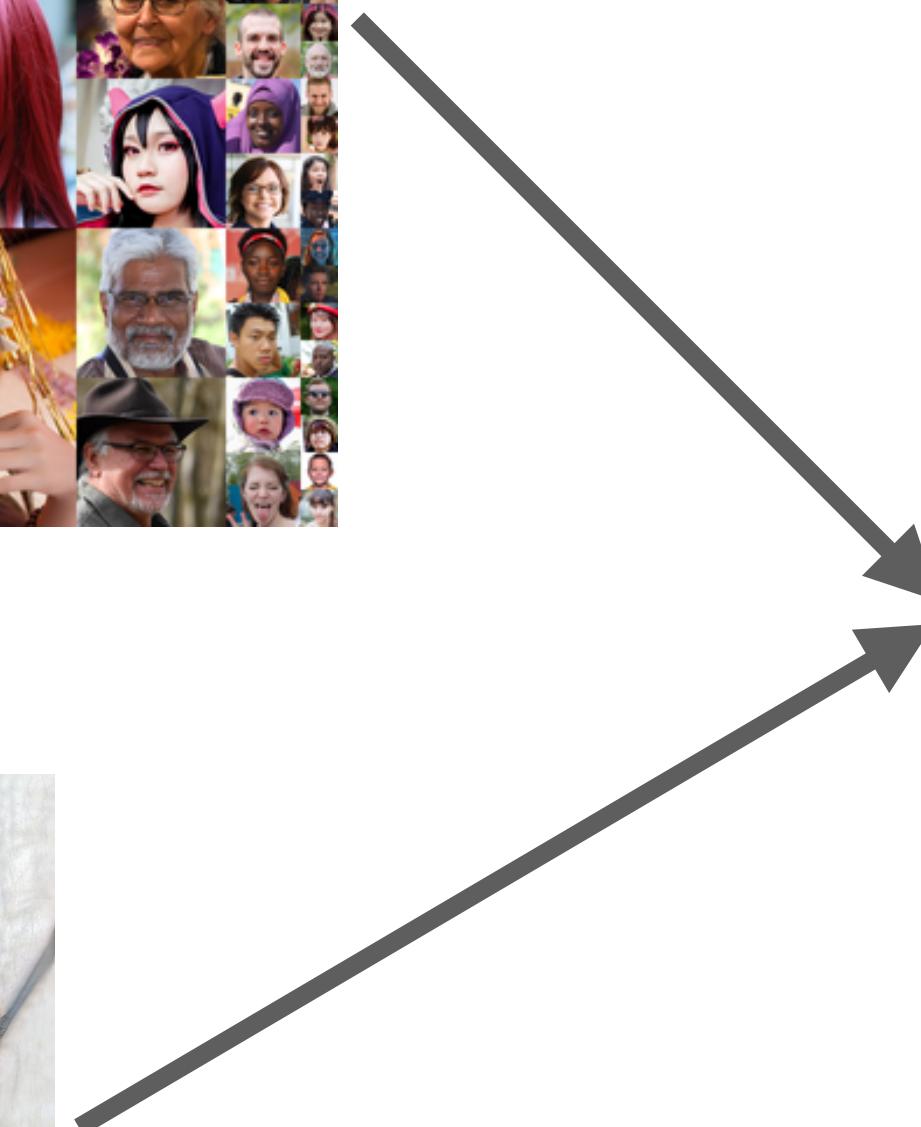
## "GAN"



# GAN



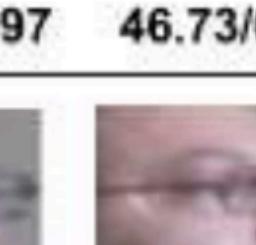
**latent variable**



Check on this website

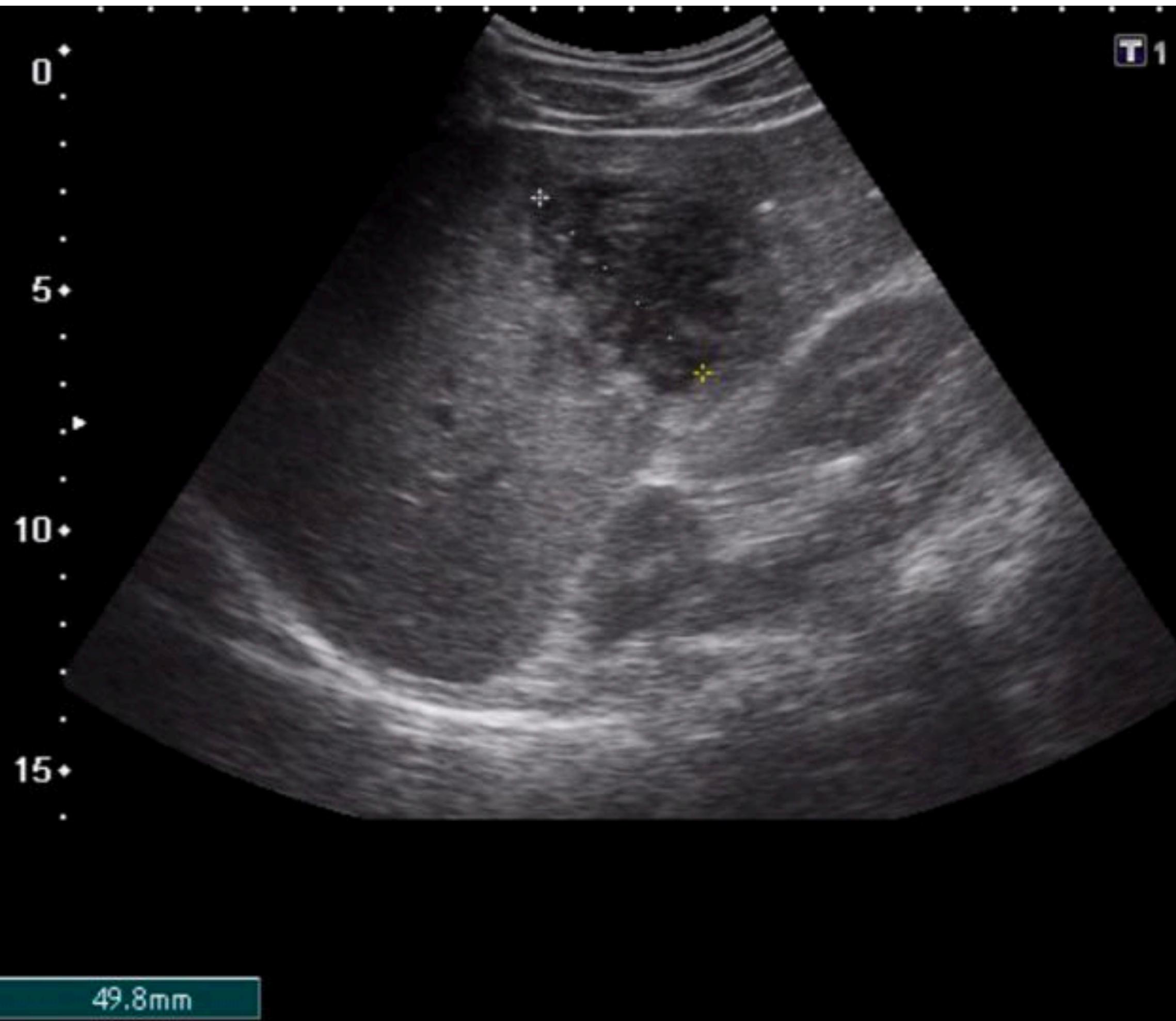
<https://thispersondoesnotexist.com/>

# GAN for head pose synthesis

QUERY	Samples generated by PosIX-GAN									39.92/0.927
	∞/1	26.70/0.817	48.99/0.997	47.88/0.996	47.71/0.997	24.29/0.782	46.13/0.997	48.88/0.996	49.23/0.997	
										HPIID
	∞/1	47.97/0.998	47.02/0.997	47.21/0.997	47.92/0.996	46.52/0.998	46.89/0.997	47.59/0.997	47.62/0.997	46.73/0.996
										47.27/0.997
	∞/1	47.97/0.998	47.02/0.997	47.21/0.997	47.92/0.996	46.52/0.998	46.89/0.997	47.59/0.997	47.62/0.997	46.73/0.996
										37.62/0.987
	∞/1	36.72/0.989	33.31/0.987	32.04/0.983	42.56/0.991	34.09/0.985	41.95/0.990	34.04/0.982	41.59/0.989	42.31/0.989
										Multi-PIE
	∞/1	42.81/0.991	41.55/0.989	41.61/0.991	42.91/0.991	41.62/0.987	42.36/0.990	41.37/0.988	41.69/0.989	43.09/0.989

Samples generated by PosIX-GAN

# GAN for removing markers



# GAN for tabular data

The screenshot shows the GitHub organization page for 'The Synthetic Data Vault Project'. The page has a dark theme. At the top, there's a search bar with placeholder text 'Search or jump to...', a '/’ button, and navigation links for 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the header, there's a logo for 'SDV' (The Synthetic Data Vault) and contact information: a link to https://sdv.dev, an email address sdv@sdv.dev, and a 'Verified' badge. The main navigation bar includes 'Overview' (which is selected), 'Repositories' (11), 'Packages', 'People' (3), and 'Projects'. The 'Pinned' section contains six cards, each representing a project:

- SDV**: Synthetic Data Generation for tabular, relational and time series data. Includes a Jupyter Notebook, 482 stars, and 88 forks.
- CTGAN**: Conditional GAN for generating synthetic tabular data. Written in Python, with 429 stars and 126 forks.
- Copulas**: A library to model multivariate data using copulas. Includes a Jupyter Notebook, 213 stars, and 62 forks.
- RDT**: A library of Reversible Data Transforms. Written in Python, with 32 stars and 14 forks.
- SDGym**: Benchmarking synthetic data generation methods. Written in Python, with 114 stars and 40 forks.
- SDMetrics**: Metrics to evaluate quality and efficacy of synthetic datasets. Written in Python, with 39 stars and 11 forks.

**Thank you  
Q & A**