Vision-Assisted Medication Preparation: Improving Syringe Detection Models Using Synthetic Data

Bas Bruijnis
bruijnis.b@gmail.com
Supervisors: Nergis Tömen, Xucong Zhang

EEMCS, Delft University of Technology, The Netherlands

Can synthetic images improve syringe detection models when only limited real training data is available?

1. Introduction & Motivation

- **High-Risk Process:** Paediatric medication preparation relies on the manual handling of very small fluid volumes, making it highly vulnerable to **human error**.
- **The Solution: Vision-assisted verification systems** could support nurses by automatically monitoring these preparation steps.
- **The Bottleneck:** Developing reliable computer vision models is hindered by a **lack of training data** and severe **hand occlusions**.

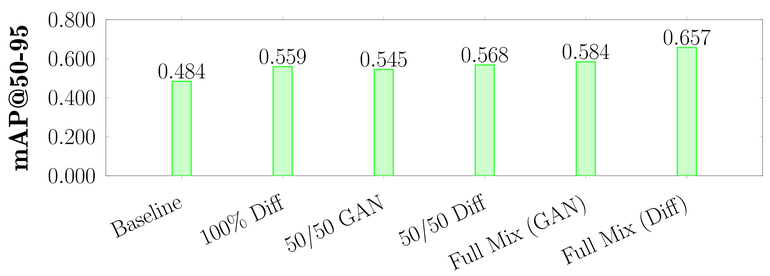
2. Methodology

Clinical Videos
(Erasmus MC)
↓
Real Dataset
(499 filtered frames)
↓
Synthetic Generation
(StyleGAN3 & SDXL + LoRA)
↓
Automatic Labelling
(SAM 3 → OBB)
↓
Detection Model
(YOLO11-OBB)
↓
Evaluation
(mAP)

3. Results

- Diffusion generated sharper, higher-resolution images (better FID scores) , but GANs maintained better structural consistency during hand interactions and provided vastly faster inference
- **Mixing = Highest Accuracy:** Combining real and synthetic datasets created the strongest detector, heavily boosting **strict spatial localization**.
- **Powerful Data Augmentation:** Even when dataset size remained constant (a 50/50 split of real/synthetic), mixing **improved model generalization** and **reduced overfitting** due to structural variation.

Main Contribution: Localization Accuracy



4. Conclusions

- Synthetic images **successfully improve** clinical syringe detection models.
- Merging **authentic textures** (real) with **structural variance** (synthetic) strengthens robustness.
- **Diffusion models** offer high fidelity and more control; **GANs** excel at rapid dataset expansion.

5. Limitations

- **Scope reductions:** High pipeline complexity and **infinite hyperparameter combinations** restricted the study to a subset of configurations.
- **Labelling Bias:** The automated SAM 3 pipeline occasionally drew **loose** bounding boxes.
- **Dataset Scale:** Sets were capped at **500 images/model**, leaving massive-scale generation unexplored.

6. Future Work

- Explore architectures like **ControlNet** for better structural control over hand occlusions.
- Generate **massive-scale synthetic datasets** and implement **explicit data splitting** for nuanced evaluation.
- Transition from static frames to **real-time action recognition** in live video streams.