Bridging the Syn-to-Real Gap in Microorganism Detection Using Blended Synthetic Data

Sebastian Jörz[1], Stefan Höreth[2], Antonio Jorba[3], Eva Bruckerseifer[1]

[1] Darmstadt University of Applied Sciences, Germany
[2] Entega Abwasserreinigung GmbH & Co. KG, Germany
[3] COUNT+CARE GmbH & Co. KG, Germany

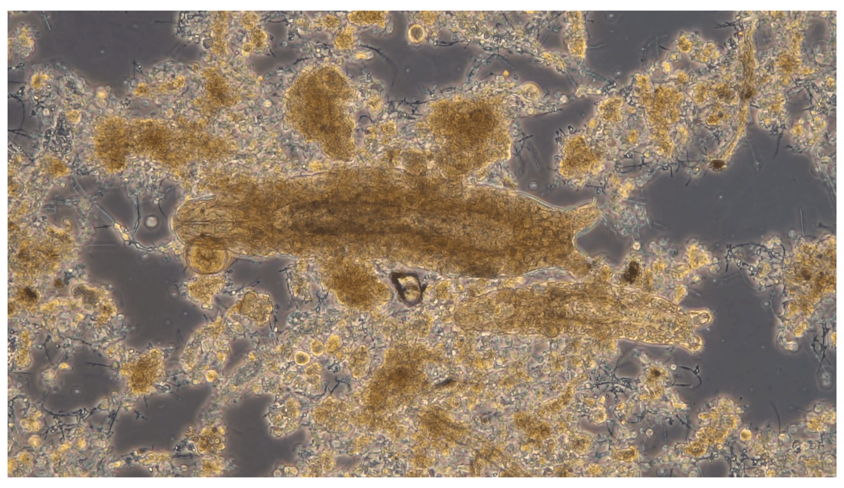
Monitoring Microorganisms

The number of different types of microorganisms is an indicator for the quality of the sewage plant process.

- **Overarching Goal:**
  - Activated Sludge Monitoring
  - Automated using YOLO object detection
- **Challenge:**
  - Limited Data
  - Homogeneous Data

The Domain Gap: Muddy vs. Clear

- **Idea:** Expand dataset with public data
- **Challenge:** Different domains
- **Generalisation:** Models struggle with real-world sewage data (our dataset) vs. clean public datasets [3]



Tardigrades in a muddy environment (e.g. in sewage plants). Source: own image

Tardigrade in a clear environment. Source: [2]

Our Solution: Synthetic Data Augmentation

- Generate synthetic data for our **'muddy' domain**
- Method: **Cut-Paste** (overlaying objects on varied backgrounds)
- Challenge: **Syn-to-Real Gap** (unrealistic composites, pixel artifacts)

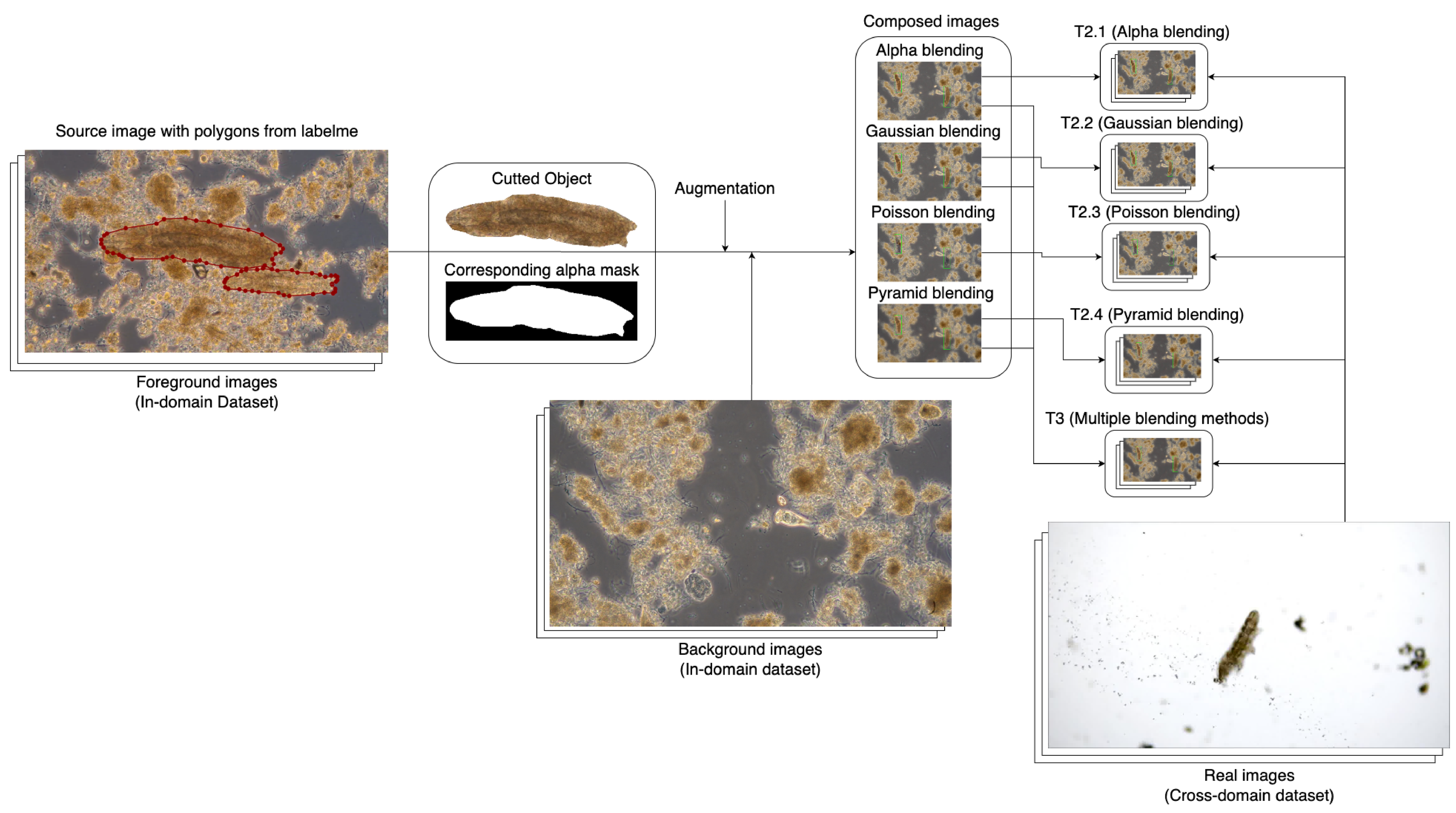
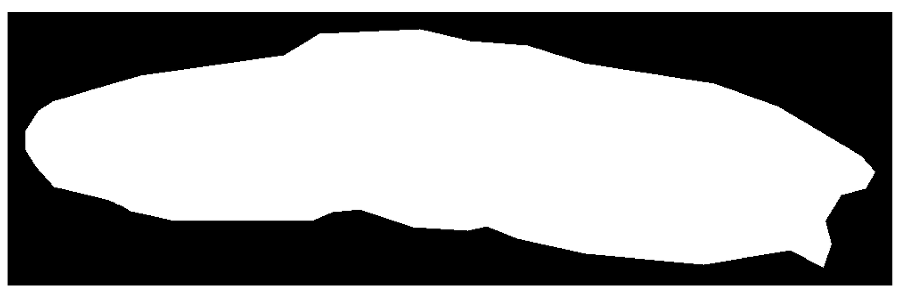


Abbildung 1: Procedure of the pipeline. Adapted from: [1]

Reduce the Syn-to-Real Gap

- **Blending:** Integrates objects seamlessly
- **Methods:**
  - **Alpha** blending
  - **Gaussian** blending
  - **Poisson** blending
  - **Pyramid** blending
  - **Multi-Blending** (Combination)
- Example of an Alpha Mask blended with Gaussian blending:



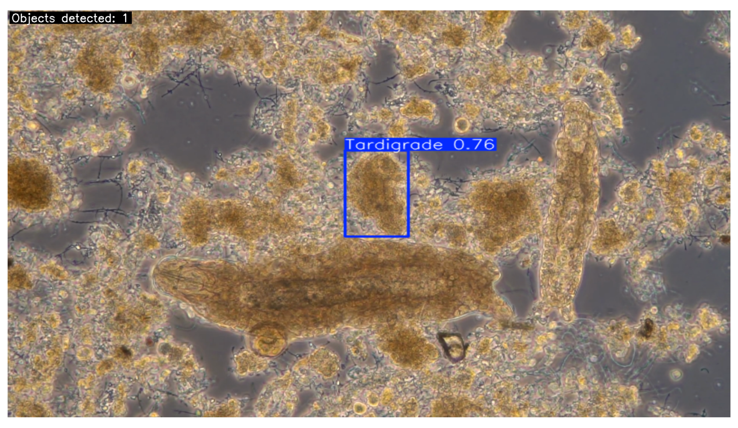
Alpha Mask

Blurred Alpha Mask with Gaussian blending

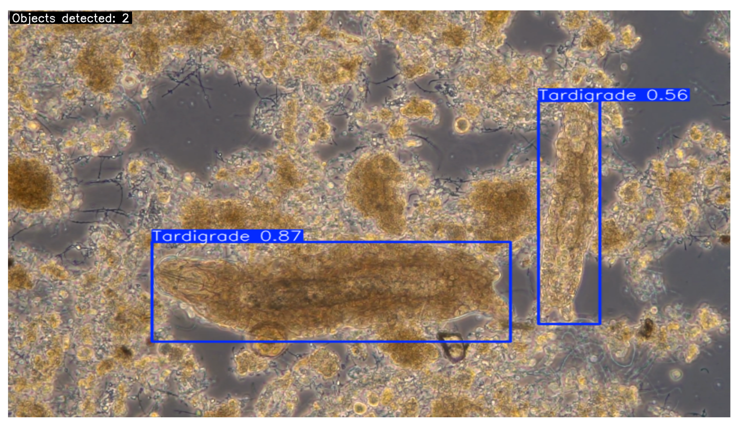
Evaluation & Datasets

- **Real Data:** Our In-Domain (**muddy**) & Public Cross-Domain (**clear**) datasets
- **Datasets for Evaluation:**
  - **T1: Baseline** (Real Data Only)
  - **T2: Single Blending** (Alpha, Gaussian, Poisson, Pyramid)
  - **T3: Multi-Blending** (Alpha + Gaussian + Pyramid)
- Varying **Synthetic Ratios**: 0%, 30%, 60%, 90%, 100%
- **Test Set:** 223 In-Domain images
- **Model:** YOLOv11m (25 Epochs)
- **Evaluation Metrics:**
  - mAP@[0.50:0.95] (Object detection performance)
  - FID & CMMD (Image Quality / Syn-Real Gap)

Improved Detection: Before vs. After



Prediction with a model trained on the T1 dataset (**Before**)



Prediction with a model trained on the T4 dataset (**After**)

Results

- **mAP Performance (YOLOv11m):**
  - Baseline (T1): **0.58 mAP**
  - Single Blending (T2): Up to **0.82 mAP** (Pyramid)
  - Poisson Blending: Underperformed significantly (0.27 mAP)
  - Multi-Blending (T3): **Best at 0.85 mAP**
  - **Optimal Ratio:** 60% Synthetic Data

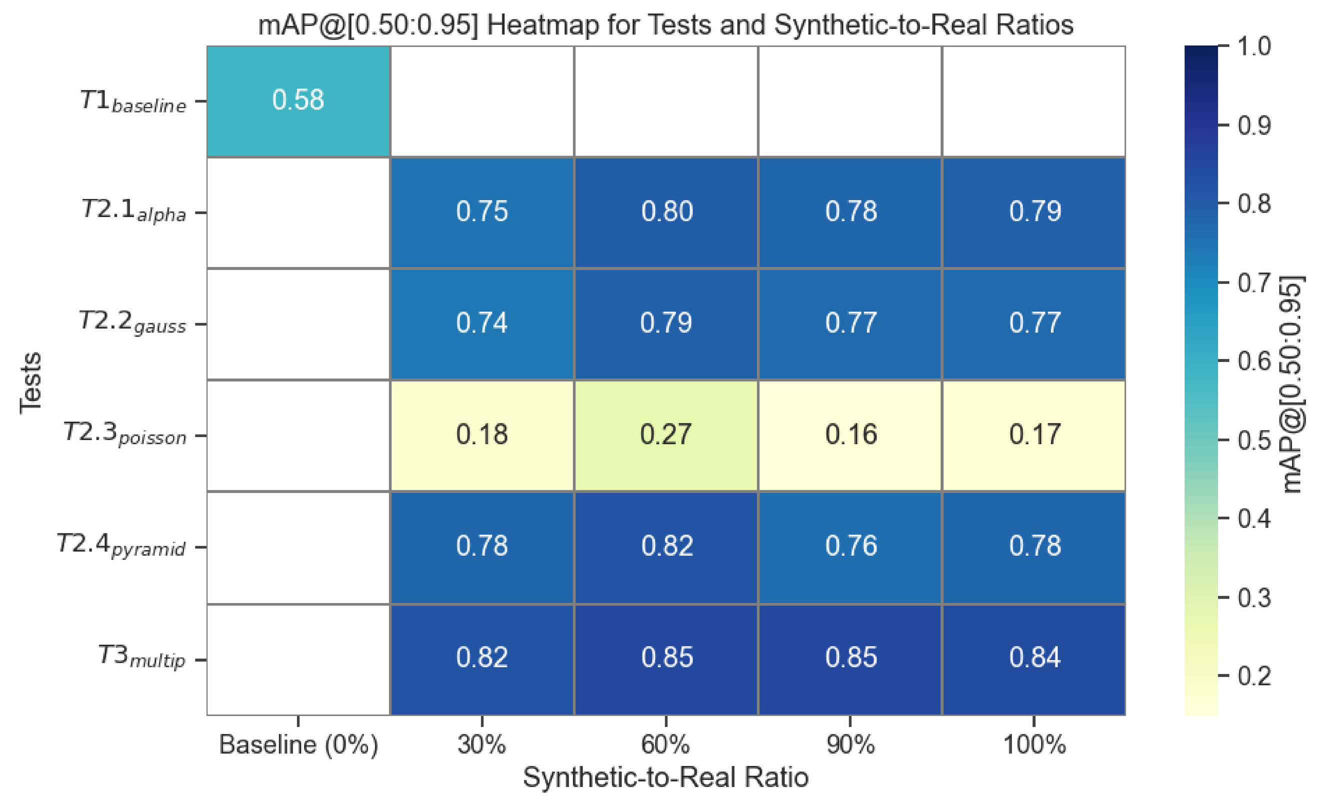


Abbildung 2: Heatmap of different models trained with different syn-to-real ratios and different blending methods

- **Image Quality (FID & CMMD):**
  - Lower scores = more realistic images
  - T1 (Baseline): Highest scores (least realistic).
  - **T3 (Multi-Blending): Lowest scores (most realistic)**
  - Correlates with mAP (smaller gap = better detection).

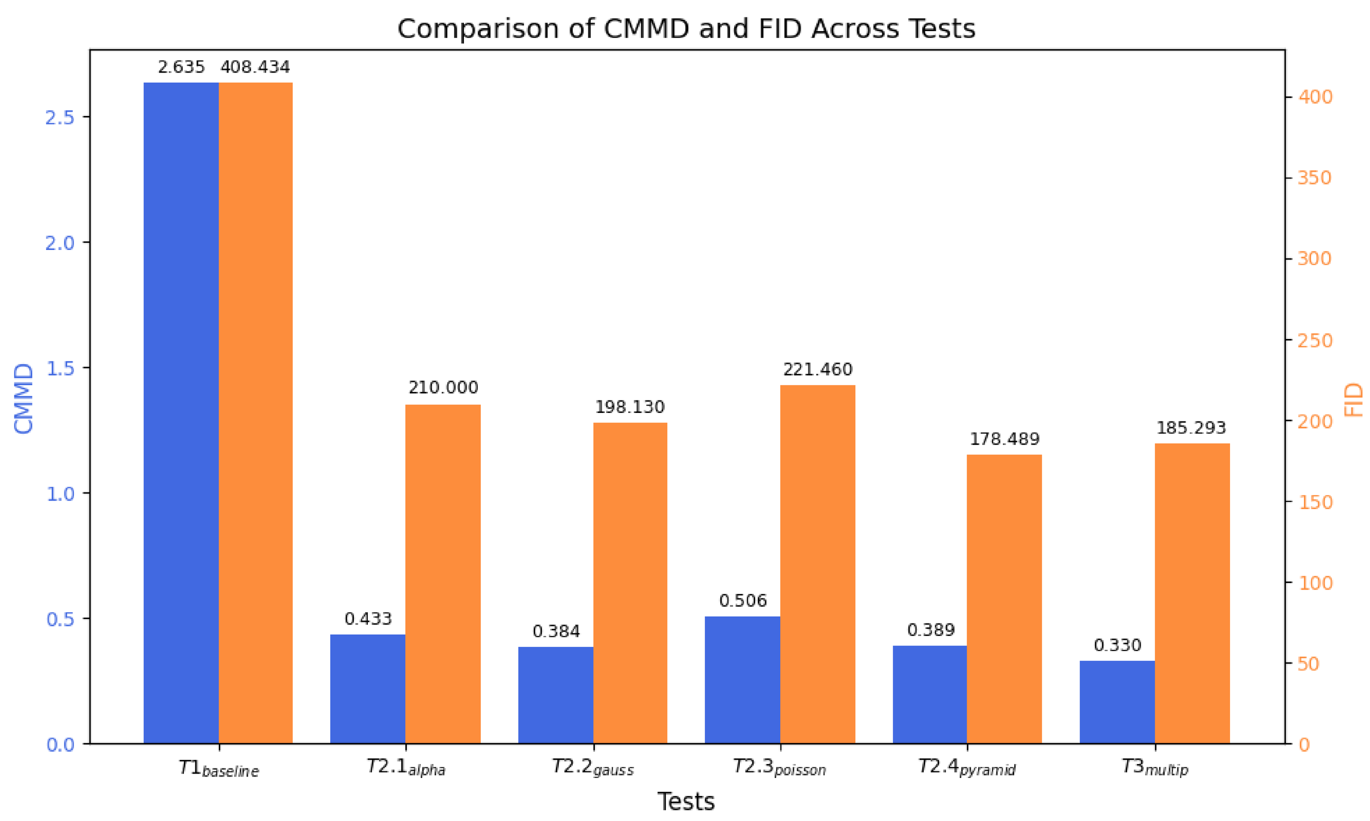


Abbildung 3: FID and CMMD, calculated on different test datasets in relation to the real dataset of the muddy environment

Conclusion & Future Directions

- **Synthetic data significantly boosts object detection in data-scarce domains**
- **Blending methods are crucial** for reducing the syn-real gap
- **Multi-blending excels**, yielding realistic images & highest mAP
- **Future Work:** Incorporate distractors (e.g., mud), noise, and explore more blending methods to enhance model resilience

References

[1] Jonas Dirr u. a. "Cut-paste image generation for instance segmentation for robotic picking of industrial parts". In: *The International Journal of Advanced Manufacturing Technology* 130.1 (1. Jan. 2024), S. 191–201. DOI: 10.1007/s00170-023-12622-4.

[2] Jasper John Jaso. *Tardigrade Dataset*. Roboflow. 2022. URL: https://universe.roboflow.com/jasper-john-jaso/tardigrade (besucht am 18.12.2024).

[3] Luis Rangel DaCosta u. a. "A robust synthetic data generation framework for machine learning in high-resolution transmission electron microscopy (HRTEM)". In: *npj Computational Materials* 10.1 (29. Juli 2024). Publisher: Nature Publishing Group, S. 1–11. DOI: 10.1038/s41524-024-01336-0.