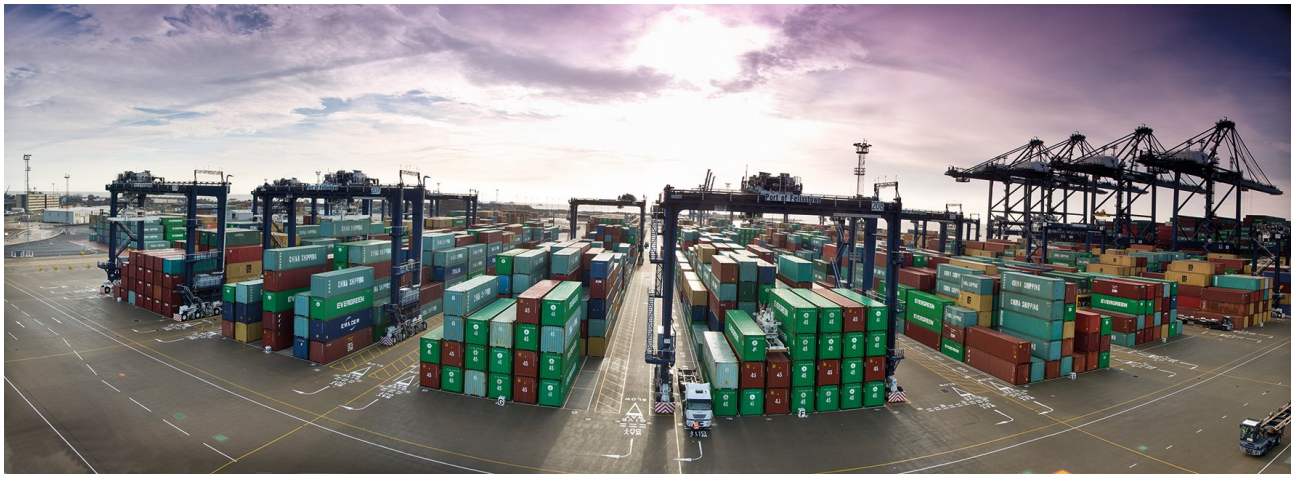
Master's Thesis at the Pattern Recognition Lab, FAU Erlangen-Nuremberg

Enhancing small-sized video object detection through temporal information and synthetic data



In recent years, there has been a steady increase in global container throughput [1], driving a high demand for automation within the harbor industry [2]. This rise in automation, however, introduces new safety risks, emphasizing the need for effective people detection systems to ensure safety in crane operations [3]. Due to the unique and elevated camera positions on cranes, individuals appear often small and in low resolution, which leads to significant challenges for accurate detection [4].

These camera mounting positions not only differ significantly from state-of-the-art setups, such as those used in datasets like COCO and ImageNet, but also vary widely across different terminals [4], making traditional supervised learning approaches costly and labor-intensive. While public datasets for people detection are widely available [5, 6], adapting pretrained models to these unique crane environments remains difficult due to domain differences. The collection and labeling of new data in these settings is resource-intensive, highlighting the need for alternative solutions.

Therefore, this study aims to investigate methods that can effectively overcome this domain gap, enabling accurate, real-time object detection without increasing the labeling workload.

To achieve this, two primary approaches will be explored. The first approach involves generating synthetic data within the target domain. Given the lack of annotated data for crane-specific settings, creating synthetic data offers a practical alternative to extensive manual labeling [7]. The second approach focuses on leveraging temporal information, which has shown promising results in other domains for enhancing video object detection accuracy. By aggregating features over time, object detectors can take advantage of information from previous frames to improve detections in the current frame, effectively addressing challenges posed by low resolution and small-scale targets. This can be achieved through techniques such as reusing features from earlier frames [8], incorporating optical flow [9], or applying simple frame subtraction [10].

1 Synthetic data generation

Techniques such as copy-pasting have demonstrated improvements in segmentation and detection tasks across various domains [11, 12]. Ghiasi et al.[7] simplified previous methods, such as that of Dwibedi et al.[11], by directly cutting out labeled objects from existing datasets and pasting these cutouts onto images of the target domain without further modeling the surrounding context. By using this simple approach the authors could achieve a significant improvement of the evaluated segmentation model and increase the data-efficiency on COCO. In the application field of small object detection, Hao et al. [13] further refined this approach by restricting the pasting locations, increasing its effectiveness in specific scenes.

2 Temporal information

State-of-the-art object detection systems, such as YOLO, effectively identify and locate objects within images and can also be applied to video object detection. A common approach to enhance video object detection accuracy is the use of temporal information through feature aggregation [14, 15, 16]. Duan et al. [8] improve detection accuracy and speed of the YOLOx by fusing the features in the current frame with those from previous frames. A similar approach is suggested by Nans et al. [10], who propose a combination of frame difference and the RGB input image. This method raised the average precision compared to a baseline model. While it works particularly well when the camera position is static, it also shows potential in scenarios with moving cameras. Shi, Zhang, and Guo [16] take feature aggregation further by integrating it into a one-stage detector, introducing a Feature Selection Module to filter out low-quality predictions and a Feature Aggregation Module to combine relevant features from multiple frames. The Feature Selection Module minimizes computational load by selecting high-confidence, object-related features, while the Feature Aggregation Module aligns and aggregates features from reference frames based on similarity scores and confidence levels, thus boosting detector accuracy without sacrificing speed. Lastly, Yu Sun et al. [9] enhance the YOLOv5s mean Average Precision performance by adding an input layer that processes multiple frames along with inter-frame optical flow, significantly improving detection of small unmanned aerial vehicles in surveillance videos.

In this work, a one-stage still image object detection model is pretrained on public data and then adapted to the crane environment by fine-tuning with synthetic data and state-of-the-art augmentation methods. The resulting model architecture is then extended to use temporal information for video object detection, on basis of the pretrained backbone as shown in Figure 1. Finally, this adapted model is fine-tuned with crane-specific video data to maximize accuracy in real-time detection scenarios.

The thesis consists of the following milestones:

- Comparison of different methods for generating synthetic training data and evaluating augmentation strategies
- Extending the pretrained object detection model with temporal information.
- Performance evaluation in terms of Average Precision and inference time on a custom crane dataset sampled from multiple harbors.
- Comparison of the enhanced model's performance against a baseline model trained without temporal information.

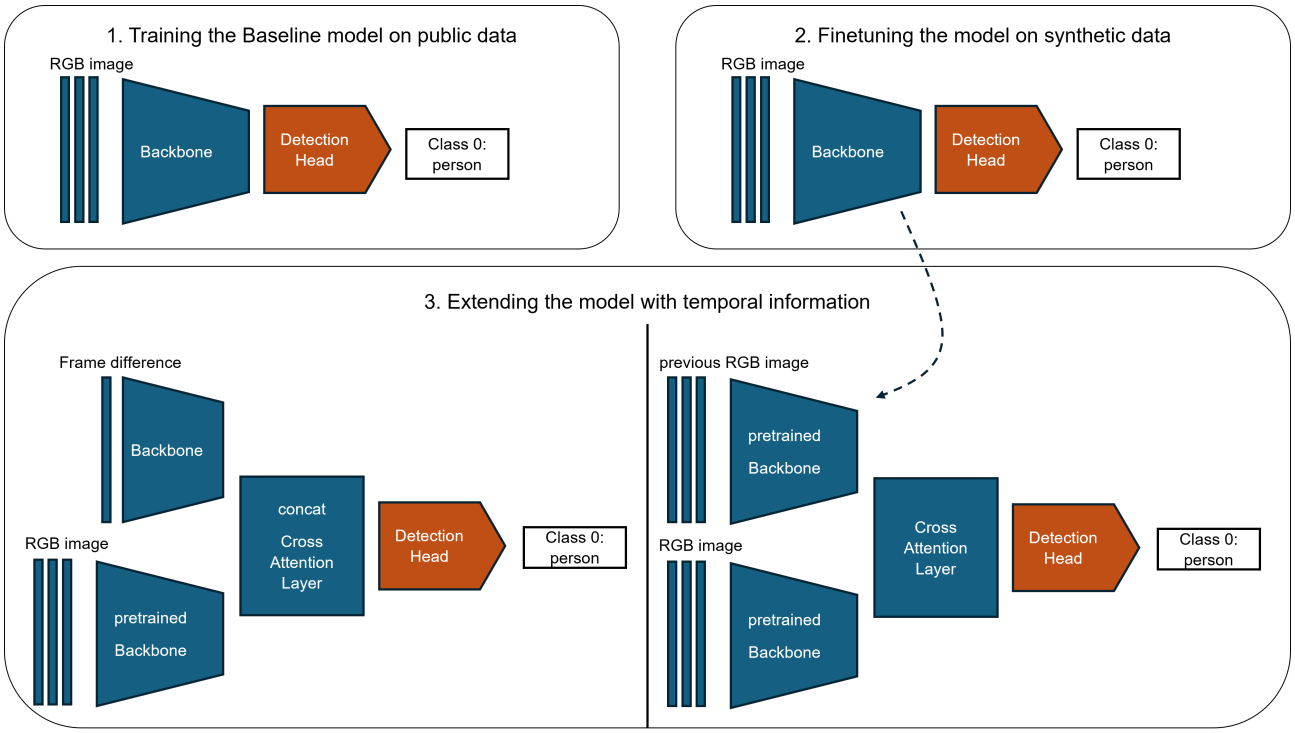


Figure 1: Overview of the three main steps in this study. First, a baseline model is trained on public data and then fine-tuned with synthetic data. Finally, the pretrained model is extended to incorporate temporal information. Two examples show how temporal information can be integrated in the lower part of the figure.

- Further experiments regarding augmentations and optimization of network architecure.

The implementation should be done in Python.

Supervisors:	M. Zinnen, Dr.-Ing. V. Christlein, J. Benkert, Prof. Dr.-Ing. habil. A. Maier
Student:	Philip Wagner
Start:	December 1st, 2024
End:	June 1st, 2025

References

[1] *Navigating Stormy Waters*. Number 2022 in Review of Maritime Transport / United Nations Conference on Trade and Development, Geneva. United Nations, Geneva, 2022.

[2] Ana María Martín-Soberón, Arturo Monfort, Rafael Sapiña, Noemí Monterde, and David Calduch. Automation in Port Container Terminals. *Procedia - Social and Behavioral Sciences*, 160:195–204, December 2014.

[3] Haibo Li. Research on safety monitoring system of workers in dangerous operation area of port. In *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pages 400–408, Banff, AB, Canada, August 2017. IEEE.

[4] Johannes Benkert, Robert Maack, and Tobias Meisen. Chances and Challenges: Transformation from a Laser-Based to a Camera-Based Container Crane Automation System. *Journal of Marine Science and Engineering*, 11(9):1718, August 2023.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham, 2014.

[6] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. CityPersons: A Diverse Dataset for Pedestrian Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4457–4465, Honolulu, HI, July 2017. IEEE.

[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2917–2927, Nashville, TN, USA, June 2021. IEEE.

[8] Liang Duan, Rongfei Yang, Kun Yue, Zhengbao Sun, and Guowu Yuan. Video object detection via space–time feature aggregation and result reuse. *IET Image Processing*, 18(12):3356–3367, October 2024.

[9] Yu Sun, Xiyang Zhi, Haowen Han, Shikai Jiang, Tianjun Shi, Jinnan Gong, and Wei Zhang. Enhancing UAV Detection in Surveillance Camera Videos through Spatiotemporal Information and Optical Flow. *Sensors*, 23(13):6037, June 2023.

[10] Lena Nans, Chelsea Mediavilla, Diego Marez, and Shibin Parameswaran. Leveraging motion saliency via frame differencing for enhanced object detection in videos. In Mohammad S. Alam and Vijayan K. Asari, editors, *Pattern Recognition and Tracking XXXIV*, page 35, Orlando, United States, June 2023. SPIE.

[11] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1310–1319, Venice, October 2017. IEEE.

[12] Jun Kimata, Tomoya Nitta, and Toru Tamaki. ObjectMix: Data Augmentation by Copy-Pasting Objects in Videos for Action Recognition. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, pages 1–7, Tokyo Japan, December 2022. ACM.

[13] Yunhe Hao, Wang Luo, Yingjie Li, Beibei Zhang, and Jia Bei. Copy and restricted paste: Data augmentation for small object detection in specific scenes. In Huimin Lu, Jintong Cai, and Yuchao Zheng, editors, *International Symposium on Artificial Intelligence and Robotics 2022*, page 16, Shanghai, China, December 2022. SPIE.

[14] Yuheng Shi, Naiyan Wang, and Xiaojie Guo. YOLOV: Making Still Image Object Detectors Great at Video Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2254–2262, June 2023.

[15] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-Guided Feature Aggregation for Video Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 408–417, Venice, October 2017. IEEE.

[16] Yuheng Shi, Tong Zhang, and Xiaojie Guo. Practical Video Object Detection via Feature Selection and Aggregation, July 2024.