



**KIDELTA**  
**LEARNING**

Scalable AI for Automated Driving

## Deliverable 03b

### Methodensammlung und Übersicht State of the Art

Version	1.0
Editor	Dr.-Ing. Amin Hosseini und Marius Bachhofer
Projektkoordination	Mercedes-Benz AG und ZF Friedrichshafen AG

Gefördert durch:



aufgrund eines Beschlusses  
des Deutschen Bundestages





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Project Description . . . . .	1
1.2	General Deliverables Overview . . . . .	2
1.3	Deliverable Context . . . . .	3
<b>2</b>	<b>Results</b>	<b>5</b>
2.1	E4.1.5.2a: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Automotive-Tauglichkeit Stand Ende PI4 . . . . .	5
2.1.1	Robustness of a pedestrian detection in corner cases (HSRT) . . . . .	5
2.2	E4.1.5.2.b: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Transfer Learning Stand Ende PI4 . . . . .	6
2.2.1	Federated Learning for Object Detection in Automated Driving (BMW) . . . . .	6
2.2.2	Methods to Prevent Catastrophic Forgetting after Class Redefinition with Minimal Relabeling (BMW) . . . . .	10
2.2.3	Auxiliary Task-Guided CycleGAN for Black-Box Model Domain Adaptation (HSRT) . . . . .	12
2.2.4	Incremental learning of an object detection model in the context of novel mobility concepts (HSRT) . . . . .	15
2.2.5	Environmental adaptations of an object detection model in the context of novel mobility concepts (HSRT) . . . . .	16
2.2.6	Survey on Unsupervised Domain Adaptation for Semantic Segmentation for Visual Perception in Automated Driving (TU Braunschweig) . . . . .	22
2.3	E4.1.5.2.c: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Didaktik Stand Ende PI4 . . . . .	26
2.3.1	M3: Monocular Self-Supervised Depth, Pose and Motion (BMW) . . . . .	26
2.3.2	Active Learning Methods Based on Consistency and Uncertainty (BMW) . . . . .	29
2.3.3	CARIAD . . . . .	33
2.3.4	Corner Case Detection (HSRT) . . . . .	37
2.3.5	Self attention techniques in the context of unsupervised domain adaptation for human pose estimation models (HSRT) . . . . .	37





## List of Figures

1.1	Timeline of the Deliverables and grouping to Milestones . . . . .	2
1.2	Follow up of the deliverables documenting the Methods activities. . . . .	4
2.1	HSRT-E4.1.5.2b-01: A sample frame of our paired dataset . . . . .	13
2.2	TUBS-E4.1.5.2b-01: Simplified diagram providing an overview of the adaptation paradigms. The red and green colors in the diagram represent the source and target domains, respectively. Dotted lines indicate datasets without available labels, while dashed lines indicate a subset of labels. For weakly-supervised DA, the available labels are noisy. . . . .	24
2.3	CARIAD-E4.1.5.2c-01: UDA taxonomy . . . . .	34
2.4	CARIAD-E4.1.5.2c-02: Performance (mIoU (%)) on the Cityscapes validation set after training on the source domains GTA5 (top row) or SYNTHIA (bottom row) with simultaneous adaptation to Cityscapes. The results are shown for models based on a ResNet-101 feature extractor (left column) or a VGG-16 feature extractor (right column). The reported values are taken from the respective papers. . . . .	35
2.5	CARIAD-E4.1.5.2c-03: Performance improvement (mIoU (% absolute)) on the Cityscapes validation set after training on the source domains GTA5 (top row) or SYNTHIA (bottom row) with simultaneous adaptation to Cityscapes. The results are shown for models based on a ResNet-101 feature extractor (left column) or a VGG-16 feature extractor (right column). The reported values are taken from the respective papers. Note that not all papers provide a baseline performance without adaptation. . . . .	36

## List of Tables

2.1	TUBS-E4.1.5.2b-01: ResNet- and transformer-based semantic segmentation results for baseline training and UDA methods. . . . .	23
-----	---	----





# 1

## Introduction

The document at hand provides information about the developments of the "Method"-Stream (see chapter 3) of KI Delta Learning, a research project funded by the German Federal Ministry for Economic Affairs and Climate Action. The chapters 1.1 to 1.4 of this document give a general introduction into the project and its deliverables. They show how the 18 deliverables contribute to the project aim and how they complement to deliver the project outcome. Chapter 2 gives the detailed content of:

- E4.1.5.2.a Methodensammlung, Katalog, Übersicht, SOTA zum Thema Automotive-Tauglichkeit Stand Ende PI4
- E4.1.5.2.b Methodensammlung, Katalog, Übersicht, SOTA zum Thema Transfer Learning Stand Ende PI4
- E4.1.5.2.c Methodensammlung, Katalog, Übersicht, SOTA zum Thema Didaktik Stand Ende PI4

### 1.1 Project Description

The goal of the KI Delta Learning project is the development of methods and tools for the the efficient extension and transformation of existing AI modules of autonomous vehicles to the challenges of new domains or more complex scenarios. AI modules are the core of the cognitive intelligence of automated vehicles and thus a key technology for ever higher levels of automation of assistance systems up fully to autonomous driving. Therefore, AI modules are of central importance to the future value creation of the German automotive industry. The market launch strategy of the German automotive industry for these assistance systems is proceeding step by step toward ever higher levels of automation and larger areas of application for automation. The focus of the project is the gradual expansion of the domains of application of assistance systems and the associated AI modules, which will be developed in parallel in six directions according to the most relevant use cases.

Thus, within the project, various deltas - the gaps between known domains of applications and new domain areas - are considered including deltas due to improved sensor technology, due to different traffic areas such as highways or construction sites, due to changes in country and corresponding traffic rules and signage, due to new forms of traffic and road users such as e-scooters, due to changing environmental conditions such as day, night, sun or rain, as well as deltas due to the advancement of neural network designs. A stepwise, structured extension of AI modules towards the six mentioned deltas is called "Delta Learning". This will not necessarily involve all six extensions simultaneously. Rather, the automation of assistance systems will gradually increase through efficient Delta Learning. KI Delta Learning aims to incrementally extend AI modules that have already been trained for limited areas and locations of use without completely re-executing the otherwise usual training and optimization process at a very high cost. So far, no sufficiently efficient and stable methods and tools exist for such Delta Learning.



## 1.2 General Deliverables Overview

Hence, the focus of the project is on method development. In two orthogonally operating but interlocked subprojects (TP2 and TP3), these methods are developed on the one hand starting from overcoming the deltas under the focus of transfer learning and on the other hand from the didactic approach. Furthermore, questions of the automotive suitability of the developed methods are investigated (TP4) as well as necessary data generated, recorded and processed (TP1).

The fundamental extension of current generations of AI algorithms expected to be achieved within the project enables a decisive leap towards the large-scale realization of autonomous vehicles. Thus, KI Delta Learning represents an important innovation building block for the competitiveness of the German automotive and supplier industry in an increasingly competitive economy.

## 1.2 General Deliverables Overview

The project worked on the goals over a period of three years and four months. It pursued the step-by-step improvement of the models and methods to be developed in four successive project steps, the Project Increments (PIs). At the end of each project increment, which goes hand in hand with the defined project milestones, the (interim) results achieved in the work packages were documented in the form of deliverables. A total of 18 deliverables were defined in the VHB, which served to communicate the results both internally and externally to the funding agency (BMW i, project sponsor). These 18 deliverables were distributed among the four project increments and can be grouped into following topics.

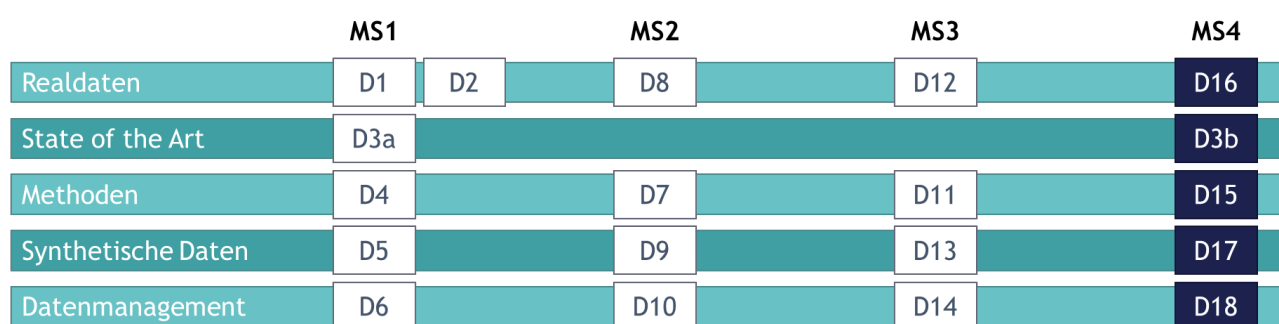


Figure 1.1: Timeline of the Deliverables and grouping to Milestones

The fourth project increment started with a delay of about three months in July 2023 and ended with the project runtime in April 2023. The results of fourth project increment are provided in the following deliverables, they represent Milestone 4:

1. D03b Methodensammlung und Übersicht SOTA
2. D15 Eine Validierung mit vollständigem Datensatz des Delta Learning-Projekts ist für die entwickelten Methoden des Delta Learning vorhanden
3. D16 Datensatz für die Realdaten aus PI4 mit definierten Annotierungen für unterschiedliche Domänen ist vorhanden & für Partner zugreifbar
4. D17 Bewertung der erzeugten synthetischen Daten und der verwendeten Verfahren in Bezug auf Qualität und Diversität ist erfolgt



### 1.3 Deliverable Context

5. D18 Aus- und Bewertung des integrierten Datenmanagements und der Datenverarbeitung für den Anwendungsfall Delta Learning ist in Form von Lessons Learned erfolgt

## 1.3 Deliverable Context

The Deliverable 03 give an overview of the state of the art and a collection of methods concerning delta learning approaches. Deliverable 03 is divided in part a and b, presenting the status at the beginning (a, 2021) and the end of the project lifetime (b, 2023).

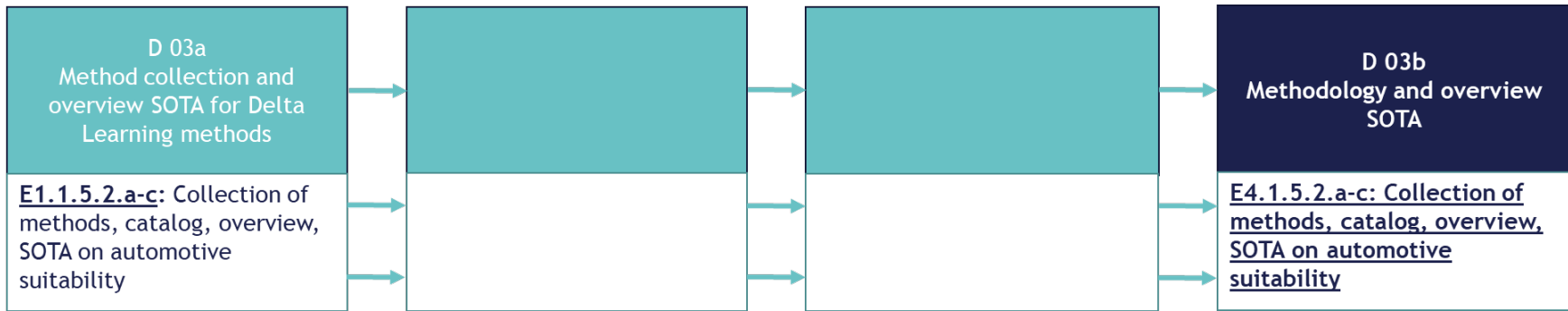


Figure 1.2: Follow up of the deliverables documenting the Methods activities.



# 2

## Results

### 2.1 E4.1.5.2a: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Automotive-Tauglichkeit Stand Ende PI4

Michael Brunner - HSRT

#### 2.1.1 Robustness of a pedestrian detection in corner cases (HSRT)

##### Introduction

Robust recognition and understanding of the environment are essential for the safety of autonomous vehicles.

In [HSRT-E4.1.5.2a-01] Ludl et al. investigate the behavior of a human pose estimation model in corner cases, e.g., a person performing a handstand, which is a situation usually not covered by standard datasets. They have shown that OpenPose trained on MS COCO dataset is not able to handle such corner cases very well. They use motion capture data and simulation to overcome this issue and improve performance in corner cases.

##### Related Work

Micromobility vehicles like e-scooters and hoverboards are growing in popularity but are not part of standard datasets yet. This leads to the question how current models behave when faced with these unknown objects. Current research points out that neural networks can be overconfident in case of unknown data [HSRT-E4.1.5.2a-02, HSRT-E4.1.5.2a-03]. We therefore investigate the robustness of a reference model in terms of micromobility vehicles. We use the model proposed by Xiao et al. [HSRT-E4.1.5.2a-04] as our reference model for human pose estimation and Faster R-CNN [HSRT-E4.1.5.2a-05] implementation provided by torchvision for object detection.

##### Bibliography

[HSRT-E4.1.5.2a-02] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34<sup>th</sup> International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330, Sydney, NSW, Australia, 2017. JMLR.org.

[HSRT-E4.1.5.2a-03] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–50. IEEE, June 2019.



## 2.2 E4.1.5.2.b

[HSRT-E4.1.5.2a-01] D. Ludl, T. Gulde, S. Thalji, and C. Curio. Using simulation to improve human pose estimation for corner cases. In *Proc. 21<sup>st</sup> Int. Conf. Intelligent Transportation Systems (ITSC)*, pages 3575–3582, November 2018.

[HSRT-E4.1.5.2a-05] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017.

[HSRT-E4.1.5.2a-04] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 472–487, Cham, 2018. Springer International Publishing.

## 2.2 E4.1.5.2.b: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Transfer Learning Stand Ende PI4

### 2.2.1 Federated Learning for Object Detection in Automated Driving (BMW)

#### Introduction

The global superpowers USA, EU and BMW-E4.2.2.1-04 enact laws to protect their data assets and citizens' privacy. The trend towards stricter regulation complicates the training of machine learning based systems. In our contribution we study one possible solution: cross-silo federated learning. In federated learning models are trained locally, with reoccurring global weight update merging and redistribution to the local instances. We focus on the use case of automated driving and aim to study if federated learning can compete with joint training for object detection with the popular Faster R-CNN algorithm. Our main contribution is the in depth study of federated learning for challenging large scale object detection use cases. We show that federated learning can compete with conventional training at a merging frequency of 0.1 per epoch on the challenging BDD and Nulmages dataset. A noteworthy finding was that federated learning might be advantageous in the case of very long training times.

In September 2020 Facebook threatened to potentially discontinue it's services in the EU. On the 6th of April 2020 then US president BMW-E4.2.2.1-02 issued an executive order to ban the popular Chinese app Tiktok from US app stores [BMW-E4.2.2.1-02]. Meanwhile, automated driving companies operating in BMW-E4.2.2.1-04 cannot easily transfer vision data taken from public roads out of the country. All these events result from globally tightening privacy and data protection rules.

On the flip-side, modern computer vision algorithm perform better if trained on larger data sets, and with over 300 million images there is still no end in sight [BMW-E4.2.2.1-14]. For automated driving safety - and thus performance - is paramount and large, diverse datasets will be a must. The automated vehicle will not just have to work in the Bay Area, but also in the largest car market in the world: BMW-E4.2.2.1-04. Data acquisition and labeling is expensive, so automated driving companies will want to use the wealth of data they have acquired in their home location to provide the best system for foreign markets. This can be a problem as the transfer of data might be prohibited or require strong and thus expensive security measures.

One solution to this challenge is the conventional transfer learning: Pretraining with data from one country and subsequent fine-tuning with data from the target country. While simple, this approach can lead to catastrophic forgetting [BMW-E4.2.2.1-13]: During fine-tuning, connections made on the first



## 2.2 E4.1.5.2.b

dataset are lost. As certain scenarios might only be present in the the pretraining dataset, this can be a severe risk in safety critical applications such as automated driving.

This leads us to investigate another potential solution: federated learning [BMW-E4.2.2.1-17]. Federated learning is a decentralized approach to train a deep network model without the need to store all the training data in one location. Our study investigates federated learning for object detection in automated driving. We focus on the so called cross-silo use case, where only relatively few (in our case two) clients are put in federation. Focusing on the two clients case enables us to investigate the important use cases of extending operations to a foreign market or collaborating with another company.

Our **main contribution** is an in depth study of federated learning using two large scale real world object detection datasets. To our knowledge, this is the first time federated learning is being investigated for object detection on modern, large scale datasets. More specifically, we studied the effects of tuning different settings for cross-silo federated learning on model performance. Additionally, we explored the feasibility of applying multitask federated learning datasets. The results of our studies point to comparable results between federated learning and non federated learning approaches.

### Related Work

**Data Privacy** - Data transfer and ownership is getting more regulated throughout the globe. The EU for example implemented the General Data Protection Regulation (GDPR) [BMW-E4.2.2.1-10], significantly strengthening their citizens right to data protection and privacy. The GDPR also addresses data transfer to non EU countries. Until recently, transferring data to the USA was considered safe under the privacy shield agreement. However, on 16 of July 2020 the European Court of Justice declared this agreement invalid [BMW-E4.2.2.1-10]. This means companies which transfer data from the EU to the USA have to implement additional safeguards. As another example, in BMW-E4.2.2.1-04 guidelines in place for data privacy are likely soon translated into law. While BMW-E4.2.2.1-04 currently has no law on data privacy, it has an increasingly strict cyber security law in place [BMW-E4.2.2.1-04]. This law was enacted to protect important data whose misuse might be harmful to the country. Since 2019 it mandates strict rules and measures for cross-border data transfers. In this contribution we investigate one potential future measure for machine learning: federated learning.

**Federated Learning** - Introduced by [BMW-E4.2.2.1-17] in 2017 to address data privacy issue, federated learning is a technique where a server coordinates the participation of decentralized clients in the training of a machine learning model, without having access to the clients' data. In a single cycle, clients download a copy of the model from the server and train on their local data. Once trained, the updated models are uploaded back to the server to be merged before restarting the cycle. Unlike distributed training approaches by [BMW-E4.2.2.1-21], federated learning assumes that the client's data is non-IID.

According to [BMW-E4.2.2.1-20], federated learning can either have a cross-device setup with edge devices as clients or a cross-silo setup with data centers as clients. For example, cross-device federated learning was used by Google to train a keyboard text prediction model using smart phones [BMW-E4.2.2.1-22, BMW-E4.2.2.1-23], while cross-silo federated learning is applied to enable cross hospital collaboration in the medical domain [BMW-E4.2.2.1-09, BMW-E4.2.2.1-24]. Despite the different use cases, the data in both setups is always being kept privately on the clients and only the trained model is uploaded to the server. For our work, we will be focusing on cross-silo federated learning as it can potentially be used to train models using data from different companies or different countries without violating data privacy laws.

In the literature, there exist multiple works on different aspects of federated learning. [BMW-E4.2.2.1-17]



## Bibliography

introduced the federated averaging algorithm to reduce the communication rounds between client and server, while maintaining model accuracy. [BMW-E4.2.2.1-25] investigated the risk of adversarial attacks on federated learning. [BMW-E4.2.2.1-26] used secure aggregation to prevent the server from knowing the identity of the clients. [BMW-E4.2.2.1-06, BMW-E4.2.2.1-05] explored extensively the topic of multitask federated learning. However, to our knowledge no work exists where federated multitask learning has been evaluated for object detection. While these works and the works surveyed in [BMW-E4.2.2.1-20, BMW-E4.2.2.1-27, BMW-E4.2.2.1-28] give us a better insight into privacy, security and algorithmic design of federated learning, there is a lack of work focusing on model performance in federated learning for real world computer vision applications. This work attempts to address this by carrying out in depth study on how to improve the model performance in federated learning.

In the original work on federated learning [BMW-E4.2.2.1-17], the authors introduced the federated averaging algorithm, *FedAvg*. Instead of uploading the local model after each training step, the algorithm waits for the clients to train the local model over multiple training steps before uploading it. This reduces the number of communication rounds between the clients and the server, resulting in faster convergence of the master model. However, the authors only experimented using naive stochastic gradient descent (SGD) to train the local models. Recent works [BMW-E4.2.2.1-29, BMW-E4.2.2.1-07] improved upon *FedAvg* by incorporating server momentum to further speed up training and improve model performance. Here, the momentum is calculated on the server side using the current and previous merged model. Inspired by these works, we will look into the possibility of training clients' models with momentum to improve the model performance.

Work on federated learning on large scale vision problems is scarce. One of the few works known to the authors is from Hsu et al. [BMW-E4.2.2.1-07] who investigate federated learning for classification with datasets with more than half a million samples and 5000 classes. For object detection, the target use case of this study, federated learning has so far only been evaluated in rather small scale settings [BMW-E4.2.2.1-01, BMW-E4.2.2.1-11]. In [BMW-E4.2.2.1-01], federated learning for object detection was investigated on 26 street cameras. Their dataset contains about 1000 labeled images. [BMW-E4.2.2.1-11] evaluates federated learning by splitting across the task domain of the Pascal VOC dataset (less than 16K samples). In our work we investigate two current, practically relevant, large scale datasets with more than 75K image samples each.

## Bibliography

[BMW-E4.2.2.1-22] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

[BMW-E4.2.2.1-26] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[BMW-E4.2.2.1-04] Cybersecurity law of the people's republic of china (english translation). <https://iapp.org/resources/article/cybersecurity-law-of-the-peoples-republic-of-china-english-translation/>. Accessed: 2020-11-13.



## Bibliography

- [BMW-E4.2.2.1-02] Executive order to ban tiktok. <https://www.whitehouse.gov/presidential-actions/executive-order-addressing-threat-posed-tiktok/>. Accessed: 2020-11-13.
- [BMW-E4.2.2.1-23] Federated learning: Collaborative machine learning without centralized training data. <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>. Accessed: 2010-11-11.
- [BMW-E4.2.2.1-11] General data protection regulation (gdpr). <https://GDPR.eu/tag/GDPR/>. Accessed: 2020-11-13.
- [BMW-E4.2.2.1-10] Schrems2. <http://curia.europa.eu/juris/document/document.jsf?text=&docid=230683&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=13080182>. Accessed: 2020-11-13.
- [BMW-E4.2.2.1-24] What is federated learning? <https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/>. Accessed: 2010-11-11.
- [BMW-E4.2.2.1-25] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.
- [BMW-E4.2.2.1-21] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.
- [BMW-E4.2.2.1-07] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution, 2020.
- [BMW-E4.2.2.1-29] Zhouyuan Huo, Qian Yang, Bin Gu, Lawrence Carin Huang, et al. Faster on-device training using new federated momentum algorithm. *arXiv preprint arXiv:2002.02090*, 2020.
- [BMW-E4.2.2.1-20] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.
- [BMW-E4.2.2.1-27] Qinbin Li, Zeyi Wen, and Bingsheng He. Federated learning systems: Vision, hype and reality for data privacy and protection. *arXiv preprint arXiv:1907.09693*, 2019.
- [BMW-E4.2.2.1-01] Jiahuan Luo, Xueyang Wu, Yun Luo, Anbu Huang, Yunfeng Huang, Yang Liu, and Qiang Yang. Real-world image datasets for federated learning, 2019.
- [BMW-E4.2.2.1-28] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.
- [BMW-E4.2.2.1-17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.
- [BMW-E4.2.2.1-13] German Ignacio Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *CoRR*, abs/1802.07569, 2018.



## 2.2 E4.1.5.2.b

[BMW-E4.2.2.1-06] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4424–4434. Curran Associates, Inc., 2017.

[BMW-E4.2.2.1-09] Micah Sheller, Brandon Edwards, G. Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka Colen, and Spyridon Bakas. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12 2020.

[BMW-E4.2.2.1-05] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[BMW-E4.2.2.1-14] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era, 2017.

### 2.2.2 Methods to Prevent Catastrophic Forgetting after Class Redefinition with Minimal Relabeling (BMW)

#### Introduction

During the development of machine learning models, a common problem is that the requirements for the categories to be recognized may change. Either one wants to recognize additional classes to the already defined categories or their subclasses. In this work we analyze the latter case in the context of image segmentation. Since in the image segmentation task each pixel has to be assigned a class this is a likely scenario. For example, one would like to replace the class human with either person or rider depending on the context. This would incur additional labeling cost, since all occurrences of human need to be relabeled.

#### Related Work

In our literature search, we did not come across any work that specifically addresses the label refinement problem. There are, however, numerous papers on related areas. In the following, we summarize representative works in bullet points for the respective research fields.

1. Fine-tuning strategies Fine-tuning is a well established technique in ML. The following two papers elaborate different variations and best practices.
  - a) Comparison of Fine-tuning and Extension Strategies for Deep Convolutional Neural Networks (MMM, 2016, [BMW-E4.1.5.2b-01])
    - Datasets: ImageNet, TRECVID SIN 2013 and PASCAL VOC
    - Comparison of three fine-tuning strategies: a) Replace last FC layer b) Replace last FC layer and reinitialize last N layers c) Last FC layer is replaced by MLP with N layers
    - In each case all weights are updated
    - Extensions strategy (c) works best (for specific configuration of the MLP)



- b) Best Practices for Fine-tuning Visual Classifiers to New Domains (2016[BMW-E4.1.5.2b-02])
- Datasets: Various datasets
  - As distance between source and target dataset increases, fine-tuning improves relative to freezing
  - Reinitializing more than last layer almost always leads to performance drop (notable exception: source and target datasets have high difference and there are a large number of labeled target examples available for fine-tuning)
2. Hierarchical classification Hierarchical classification approaches exploit similarity structures for the classes of interest and are most commonly employed when the number of classes is high.
- a) A survey of hierarchical classification across different application domains (KDD, 2011[BMW-E4.1.5.2b-03])
- Provides taxonomy and experimental comparisons
  - Any hierarchical classification approach (local or global) is overall better than the flat classification approach when solving a hierarchical classification problem
- b) A Survey of Hierarchical Classification Algorithms with Big-Bang Approach (ICST, 2019[BMW-E4.1.5.2b-04])
- Survey previous research on hierarchical classification algorithms using the big-bang approach, mostly in the domains of bioinformatics and text classification
3. Semi-supervised methods and self-supervised methods Semi-supervised and self-supervised learning are both large research areas. While the former uses unlabeled data explicitly in conjunction with labeled data, the latter can be used to learn semantic representations without any labeled examples. Thus, self-supervised losses can be readily used in conjunction with supervised losses to build semi-supervised approaches. The following paper gives an overview of semi-supervised approaches which minimize inconsistencies on the unlabeled data.
- a) Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision (CVPR, 2021[BMW-E4.1.5.2b-05])
- Datasets: PASCAL VOC 2012 and Cityscapes
  - They introduce an auxiliary loss which enforces consistency between the predictions of two (randomly initialized) networks. They feed an unlabeled image through both networks and use the prediction of one network as ground truth for the prediction of the other network and vice versa.
  - Could be applied to our problem by enforcing consistency on the fine-grained classes when only the coarse-grained label is known or in the total absence of labels.

## Bibliography

[BMW-E4.1.5.2b-02] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 435–442. Springer, 2016.



## 2.2 E4.1.5.2.b

[BMW-E4.1.5.2b-05] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[BMW-E4.1.5.2b-04] Sofi Defiyanti, Edi Winarko, and Sigit Priyanta. A survey of hierarchical classification algorithms with big-bang approach. In *2019 5th International Conference on Science and Technology (ICST)*, volume 1, pages 1–6. IEEE, 2019.

[BMW-E4.1.5.2b-01] Nikiforos Pittaras, Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In *Multi-Media Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*, pages 102–114. Springer, 2017.

[BMW-E4.1.5.2b-03] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72, 2011.

### 2.2.3 Auxiliary Task-Guided CycleGAN for Black-Box Model Domain Adaptation (HSRT)

Michael Brunner, Markus Rehmann and Cristóbal Curio - Hochschule Reutlingen

Copyright 2023 IEEE <https://ieeexplore.ieee.org/document/10030089>

Vinu Nair - Hochschule Reutlingen

#### Introduction

Deep learning has shown to be tremendously successful in complex tasks such as natural language processing [HSRT-E4.1.5.2b-01], computer vision [HSRT-E4.1.5.2b-02, HSRT-E4.1.5.2b-03], or content generation [HSRT-E4.1.5.2b-04] and is a key technology for autonomous driving [HSRT-E4.1.5.2b-05, HSRT-E4.1.5.2b-06]. However, if the training data on which such algorithms are trained diverges from the data they are supposed to operate on, which is commonly referred to as domain shift, performance degradation is to be expected. Thinking about the dynamic, diverse, and open world we are living in, it is clearly not possible to cover every possible scenario in a training dataset, showing the necessity to explicitly account for domain shifts. The active research area of domain adaptation (DA) works on methods to compensate for domain shifts and to improve model performance across domains.

DA can be attributed to transfer learning (TL), more specific to transductive TL [HSRT-E4.1.5.2b-07], where only labeled data from the source domain is available and existing knowledge is intended to be transferred between the source and target domain under the assumption that the tasks do not differ. We can further distinguish between semi-supervised DA, when some labeled data is available in the target domain, and unsupervised domain adaptation (UDA), when there is no labeled data available in the target domain [HSRT-E4.1.5.2b-08]. In this work, we follow a UDA approach, since UDA requires no time-consuming and expensive data labeling and therefore promises the greatest benefit. Moreover, while earlier shallow DA methods account for the domain shift by, e.g., instance re-weighting [HSRT-E4.1.5.2b-09] or simple feature augmentation [HSRT-E4.1.5.2b-10], deep DA methods are considered more promising due to better DA performance [HSRT-E4.1.5.2b-08, HSRT-E4.1.5.2b-11]. For this reason, we follow an



## 2.2 E4.1.5.2.b

unsupervised deep DA approach for cross-sensor adaptation based on CycleGAN [HSRT-E4.1.5.2b-12], a generative adversarial network (GAN) [HSRT-E4.1.5.2b-13] for unpaired image-to-image translation.

Existing DA methods are usually targeted and optimized for specific tasks or network architectures, e.g., image segmentation [HSRT-E4.1.5.2b-19, HSRT-E4.1.5.2b-20] or keypoint detection [HSRT-E4.1.5.2b-14, HSRT-E4.1.5.2b-15, HSRT-E4.1.5.2b-16], and as a major downside require modifications and thus access to the model and its parameters for which domain adaptation is to be performed. There is only little work on DA of black-box models, i.e., only having access to the non-differentiable predictions and having no access to the model parameters, and those methods are mainly targeted at image classification [HSRT-E4.1.5.2b-17, HSRT-E4.1.5.2b-18]. We also assume a black-box model and, in contrast to existing work, keep our UDA approach more general and do not rely on a specific task or architecture. Moreover, we focus on black-box UDA for regression instead of classification and thus train and evaluate our approach on the challenging task of 2D human pose estimation. For this purpose, we created a motion capture-based dataset consisting of paired real and synthetic images, cf. Fig. 2.1. While our approach does not rely on or make use of paired data, it allows us to evaluate the model's performance with and without DA on scenes with the same content. We have the same human pose configurations across domains, but the domain shift is either caused by different sensors, i.e., synthetic and real RGB images or synthetic RGB and synthetic depth images, or by variations in the person's appearance, i.e., clothing.

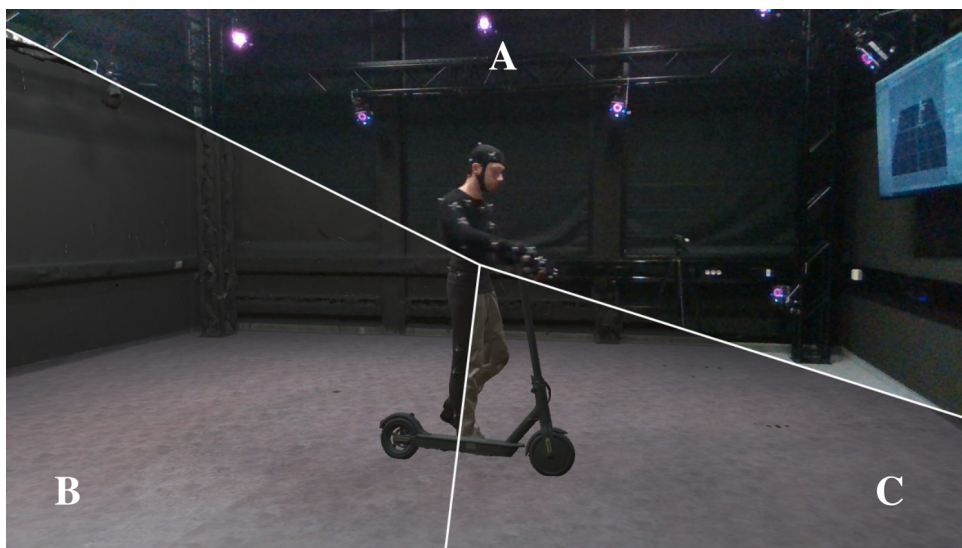


Figure 2.1: Figure HSRT-E4.1.5.2b-01: A sample frame of our paired dataset showing three different domains A) real RGB sensor data with motion capture suit, B) synthetic RGB sensor data with motion capture suit, C) synthetic RGB sensor data with casual clothing

We use the Transfer-Learning-Library developed by Jiang et al. [HSRT-E4.1.5.2b-21], an open-source library that includes various TL methods and reference models for various tasks, and extend it with our proposed method to conduct our experiments.

Our contributions can be summarized as follows:

- We analyze the performance of CycleGAN [HSRT-E4.1.5.2b-12] for unsupervised cross-sensor adaptation of a keypoint detection model, i.e., human pose estimation, across four different settings with varying domain shift.
- We show that unsupervised cross-sensor adaptation can be greatly improved by two simple modifica-



## 2.2 E4.1.5.2.b

tions to CycleGAN, namely switching to a cyclical learning rate [HSRT-E4.1.5.2b-22] and adding a task-related auxiliary loss inspired by multi-task learning [HSRT-E4.1.5.2b-23, HSRT-E4.1.5.2b-24, HSRT-E4.1.5.2b-25, HSRT-E4.1.5.2b-26] and self-supervision [HSRT-E4.1.5.2b-27], even under the assumption that we only have access to a black-box model but not to its parameters.

- We compare our method to a recent approach for unsupervised domain adaptation for keypoint detection (RegDA [HSRT-E4.1.5.2b-15]) and conclude by emphasizing the necessity for explicitly addressing sensor domain shift.

### Related Work

**Human pose estimation** is crucial for the safety of autonomous systems, e.g., in the area of autonomous driving or collaborative robots. While early deep learning-based methods such as DeepPose [HSRT-E4.1.5.2b-28] directly regressed the 2D coordinates of human body joints in an image, recent fully convolutional approaches usually generate heatmaps, where joint positions are retrieved with a non-differentiable argmax operation [HSRT-E4.1.5.2b-29, HSRT-E4.1.5.2b-30]. By replacing argmax with an integral operation, training can be performed in an end-to-end manner [HSRT-E4.1.5.2b-31]. Latest 3D pose estimation methods are able to jointly predict 2D and 3D poses as well as head and body orientation from images [HSRT-E4.1.5.2b-32]. Furthermore, we can distinguish between bottom-up and top-down human pose estimation. In the more commonly used top-down approaches, an additional object detection step is required to obtain bounding boxes of persons in an image. Pose estimation is performed afterward on every detected bounding box, as done in [HSRT-E4.1.5.2b-29, HSRT-E4.1.5.2b-30]. In contrast to top-down approaches, bottom-up approaches simultaneously predict the poses of multiple persons in a single step. In the case of OpenPose [HSRT-E4.1.5.2b-33], part affinity fields are predicted to associate joints with body parts and individuals.

There are already many different methods and architectures just for the task of human pose estimation that would need to be addressed in terms of domain adaptation. Therefore, the applicability of our UDA method is not targeted at a specific task or architecture, although in this work we focus on 2D human pose estimation and show the feasibility of our approach at the example of the pose estimation method proposed by Xiao et al. [HSRT-E4.1.5.2b-30].

**Unsupervised domain adaptation** is the setting where labeled source but only unlabeled target domain data is available and the goal is to transfer a model trained on the source domain to the target domain. We focus on deep domain adaptation (DDA) which is commonly categorized into discrepancy-, adversarial- and reconstruction-based methods [HSRT-E4.1.5.2b-08, HSRT-E4.1.5.2b-34].

*Discrepancy-based methods* aim to learn domain invariant features by reducing the discrepancy of intermediate network layers on the source and target domain, e.g., deep adaptation networks [HSRT-E4.1.5.2b-35] use a multi-kernel variant of maximum mean discrepancy.

*Adversarial-based methods* use a discriminator, i.e., a domain classifier, that learns to classify data into source and target domain. The model is encouraged to learn domain invariant features through an adversarial goal, i.e., to fool the discriminator to misclassify the domains (domain confusion). It can further be distinguished between non-generative and generative methods [HSRT-E4.1.5.2b-08, HSRT-E4.1.5.2b-34]. Ganin and Lempitsky [HSRT-E4.1.5.2b-36] propose a non-generative adversarial-based method and introduce a gradient reversal layer combined with a domain classifier to learn domain-invariant features, which is also referred to as domain adversarial neural network (DANN) [HSRT-E4.1.5.2b-37]. The loss for the



## 2.2 E4.1.5.2.b

main task, i.e., classification, is minimized using labeled source data, while the features are forced to be discriminative for the source and unlabeled target domain by maximizing the domain classification loss through a gradient reversal layer prepended to the domain classifier. In the generative case, an additional generator is added leading to a GAN [HSRT-E4.1.5.2b-13] architecture. Such models can be used to, e.g., generate domain-specific images from ground-truth semantic segmentation masks [HSRT-E4.1.5.2b-38].

*Reconstruction-based methods* assume that reconstructing the source or target data is beneficial to encode domain-specific features. CycleGAN [HSRT-E4.1.5.2b-12] for unpaired image-to-image translation introduces a cycle consistency loss that forces the network to transfer the image style while preserving the image content. The cycle consistency loss is the L1 loss between the original image and the reconstructed image, i.e., after translation to the opposite domain and back again. CyCADA [HSRT-E4.1.5.2b-19] extends CycleGAN with additional semantic consistency, task, and feature losses for DA of a semantic segmentation model.

These DA methods usually require access to the model and its parameters because they rely on (intermediate) activations or model weights. Although generative adversarial- and reconstruction-based approaches, in particular, can be used without having access to the (differentiable) model and its parameters, additional task-specific losses are often based on the model's output [HSRT-E4.1.5.2b-20, HSRT-E4.1.5.2b-19]. In contrast, Zhang et al. [HSRT-E4.1.5.2b-17] focus their work on UDA of black-box source models for image classification. They use the source model to predict noisy labels for the target domain, estimate the noise rate, select good samples and train a new model for the target domain. This procedure is repeated with the updated model for the target domain. While [HSRT-E4.1.5.2b-17] requires predicted soft labels from the source model, DINE [HSRT-E4.1.5.2b-18] can perform UDA of black-box models with hard labels.

We follow the idea of DA of black-box models but focus on a regression task, more specifically the 2D human pose estimation model proposed by Xiao et al. [HSRT-E4.1.5.2b-30], in contrast to classification as done in [HSRT-E4.1.5.2b-17, HSRT-E4.1.5.2b-18]. Our approach is close to [HSRT-E4.1.5.2b-16], where DA for human pose estimation is performed by domain translation between synthetic and real depth data. However, we keep our method as general as possible and hence use a reconstruction-based method without making task-specific assumptions such that it can be easily adapted to other tasks. We compare our approach to RegDA [HSRT-E4.1.5.2b-15], a UDA method for 2D keypoint detection recently proposed by Jiang et al. RegDA is an adversarial-based method that trains an adversarial regressor to maximize and a feature regressor to minimize disparity on the target domain. Thus, the feature regressor learns domain invariant features. Ground false poses, which are required for the adversarial training procedure of RegDA, are generated under the assumption that wrong keypoint predictions are most likely located at the position of other keypoints. In contrast to RegDA, our method does not require access to the source model and its parameters but only to its predictions and is, therefore, suitable for black-box DA.

### 2.2.4 Incremental learning of an object detection model in the context of novel mobility concepts (HSRT)

#### Introduction

The emergence of micromobility vehicles like e-scooters and hoverboards and their growing popularity pose a challenge for computer vision. Current datasets do not cover these vehicle types and hence visual task models such as object detection or semantic segmentation are not able to detect them. Using simulated data, which is available relatively cheap, these models can be updated to detect new classes, but they will



## 2.2 E4.1.5.2.b

perform worse on real data due to domain shift. Furthermore, existing transfer learning approaches are not practical in an automotive environment where new unlabeled data is constantly available and joint training is expensive. To address these topics, we focus on the incremental learning(IL)(also known as continual learning) of object detection models for novel mobility classes and adapt the models to the target domain using unsupervised domain adaptation(UDA).

### Related work

**Object Detection** CNN-based object detectors fall into two categories. Single-stage detectors like SSD [HSRT-E4.1.5.2b-44] and YOLO [HSRT-E4.1.5.2b-45] and two-stage detectors like Faster-RCNN [HSRT-E4.1.5.2b-46]. We refer to [HSRT-E4.1.5.2b-47] for a comprehensive review of object detection models. We used Faster-RCNN based model for the object detection task due to its simplicity in implementation and the capability to distill knowledge from multiple stages. Faster-RCNN is a two-stage object detection network with a region proposal network and a Fast-RCNN detector. These networks use a common feature extraction backbone normally pre-trained on ImageNet.

**Incremental learning** Transfer learning has been extensively covered in the literature. We refer to [HSRT-E4.1.5.2b-48], [HSRT-E4.1.5.2b-49] for a comprehensive overview. Common approaches for class incremental learning are based on feature extraction, fine-tuning or joint training. Other specific methods are experimentally evaluated in [HSRT-E4.1.5.2b-50]. Li and Hoiem [HSRT-E4.1.5.2b-51] proposed learning without forgetting for incremental learning in image classification tasks. They use the original teacher model to make predictions for the existing tasks on the new data. Student model is an extended version of the teacher model for predicting the new task. Teacher model is then trained using knowledge distillation[HSRT-E4.1.5.2b-52]. A similar approach was introduced for object detection networks by Shmelkov et al., where they used a frozen copy of a detector network and distillation methods to perform incremental learning without using old labels [HSRT-E4.1.5.2b-53]. They use a distillation loss calculated from the response of the teacher network for images containing new classes. Inspired by this approach, Peng et al. proposed Faster Incremental Learning Object Detector (FILOD) which implements incremental learning for Faster-RCNN models[HSRT-E4.1.5.2b-54]. Faster-ILOD calculates adaptive distillation loss on the feature maps, Region Proposal Network(RPN) outputs together with the Shmelkov distillation loss to train the student model to alleviate the missing annotation problem in incremental learning. In this method, knowledge distillation is done at multiple stages of the faster-RCNN network.

## 2.2.5 Environmental adaptations of an object detection model in the context of novel mobility concepts (HSRT)

### Introduction

Simulated driving environments are ideal tools to produce large amounts of training data representing a wide range of conditions, including corner cases, without incurring any annotation cost. The effectiveness of using simulated data for environmental adaptation of object detection models with unsupervised domain adaptation needs further investigations especially in the context of novel mobility classes such as e-scooters. Since real world datasets does not contain many instances of novel mobility classes on different weather conditions, we generated the required training data using a driving simulation. The developed driving simulation included novel mobility classes such as E-Scooters and Hoverboards, as well as a variety of weather conditions. Training data was collected for each weather condition by driving through the simulated



## Bibliography

environment. Reference model for object detection were then trained using data from the sunny condition. Finally, unsupervised domain adaptation was performed from the sunny condition to the rainy, foggy, and snowy conditions.

### Related work

Weather adaptation is a challenging task in the field of computer vision. Various studies have been conducted to tackle this problem through domain adaptation and other techniques. In the paper Don't Worry About the Weather: Unsupervised Condition-Dependent Domain Adaptation, a domain adaptation system was proposed that uses light-weight input adapters to pre-processes input images, irrespective of their appearance [HSRT-E4.1.5.2b-55]. Dai et al. proposed a novel method to progressive adapt the semantic models trained on daytime scenes, along with large-scale annotations therein, to nighttime scenes via the bridge of twilight time [HSRT-E4.1.5.2b-56]. To adapt the task models for appearance changes based on lighting, seasonal, and weather conditions Wulfmeier et al. proposed an adversarial approach for lifelong, incremental domain adaptation which benefits from unsupervised alignment to a series of intermediate domains [HSRT-E4.1.5.2b-57]. Image transformation techniques which uses a denoising generator for adapting rainy images to clear images[HSRT-E4.1.5.2b-58].

For simulating different weather conditions we are using a driving simulator developed using the Unity® game engine. Game engine based solutions for environmental perception vehicle simulation are popular tools for generating synthetic data [HSRT-E4.1.5.2b-59], [HSRT-E4.1.5.2b-60]. The unsupervised domain adaptation model developed in WP2.2/2.3 was used to perform weather adaptation using the data generated from the unity simulation. This is a CycleGAN-based approach for cross-sensor and sim-to-real domain adaptation[HSRT-E4.1.5.2b-61]. For detailed review of related works of this method, see section 2.3.2. We are using this model for sim-to-sim domain adaptation with an object detection specific auxiliary task. First, we trained the object detection model on source domain. We generated network saliency maps[HSRT-E4.1.5.2b-62] for source domain images using the source trained object detection model. The auxiliary task for unsupervised domain adaptation of object detection model was to predict the saliency map of the target domain image. After the auxiliary ground truth generation, we trained the unsupervised domain adaptation model following the RegDa training settings using the transfer learning framework library with the source domain trained model [HSRT-E4.1.5.2b-63]. During training, the auxiliary task provides guidance for the domain adaptation network to focus on areas which are important for the object detection task. For object detection, we are using the Faster-RCNN[HSRT-E4.1.5.2b-46] model from the WP2.4.

### Bibliography

[HSRT-E4.1.5.2b-32] Dennis Burgermeister and Cristóbal Curio. PedRecNet: Multi-task deep neural network for full 3d human pose and orientation estimation. In *IEEE Intelligent Vehicles Symposium (IV)*, 2022.

[HSRT-E4.1.5.2b-01] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Lan-



## Bibliography

- guage Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [HSRT-E4.1.5.2b-45] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs, eess] type: article.
- [HSRT-E4.1.5.2b-23] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.
- [HSRT-E4.1.5.2b-33] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, January 2021.
- [HSRT-E4.1.5.2b-08] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, 2017.
- [HSRT-E4.1.5.2b-03] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, December 2021.
- [HSRT-E4.1.5.2b-05] You Dingyi, Wang Haiyan, and Yang Kaiming. State-of-the-art and trends of autonomous driving technology. In *2018 IEEE International Symposium on Innovation and Entrepreneurship (TEMIS-ISIE)*, pages 1–8, March 2018.
- [HSRT-E4.1.5.2b-10] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45<sup>th</sup> Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [HSRT-E4.1.5.2b-60] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st annual conference on robot learning*, pages 1–16.
- [HSRT-E4.1.5.2b-56] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 3819–3824. IEEE.
- [HSRT-E4.1.5.2b-16] Michael Essich, Dennis Ludl, Thomas Gulde, and Cristobal Curio. Learning to translate between real world and simulated 3d sensors while transferring task models. In *2019 International Conference on 3D Vision (3DV)*. IEEE, September 2019.
- [HSRT-E4.1.5.2b-61] Michael Essich, Markus Rehmann, and Cristóbal Curio. Auxiliary task-guided CycleGAN for black-box model domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pages 541–550.
- [HSRT-E4.1.5.2b-36] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 2015. PMLR.



## Bibliography

- [HSRT-E4.1.5.2b-13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [HSRT-E4.1.5.2b-37] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, January 2016.
- [HSRT-E4.1.5.2b-19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35<sup>th</sup> International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.
- [HSRT-E4.1.5.2b-52] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [cs, stat] type: article.
- [HSRT-E4.1.5.2b-38] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.
- [HSRT-E4.1.5.2b-21] Junguang Jiang, Baixu Chen, Bo Fu, and Mingsheng Long. Transfer-Learning-library, 2020.
- [HSRT-E4.1.5.2b-63] Junguang Jiang, Bo Fu, and Mingsheng Long. Transfer-learning-library.
- [HSRT-E4.1.5.2b-27] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. pages 8784–8794, Seattle, WA, USA, 2020. IEEE.
- [HSRT-E4.1.5.2b-15] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6776–6785, 2021.
- [HSRT-E4.1.5.2b-44] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. *SSD: Single Shot MultiBox Detector*, volume 9905, pages 21–37. Springer International Publishing. arXiv:1512.02325 [cs].
- [HSRT-E4.1.5.2b-35] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In Francis Bach and David Blei, editors, *Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, July 2015. PMLR.
- [HSRT-E4.1.5.2b-51] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947. arXiv: 1606.09282.
- [HSRT-E4.1.5.2b-18] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. DINE: Domain adaptation from single and multiple black-box predictors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.



## Bibliography

- [HSRT-E4.1.5.2b-26] Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. VALHALLA: Visual hallucination for machine translation. May 2022.
- [HSRT-E4.1.5.2b-50] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [HSRT-E4.1.5.2b-29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 483–499, Cham, 2016. Springer International Publishing.
- [HSRT-E4.1.5.2b-24] Bernardino Romera Paredes, Andreas Argyriou, Nadia Berthouze, and Massimiliano Pontil. Exploiting unrelated tasks in multi-task learning. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 951–959, La Palma, Canary Islands, 2012. PMLR.
- [HSRT-E4.1.5.2b-55] Horia Porav, Tom Bruls, and Paul Newman. Don't worry about the weather: Un-supervised condition-dependent domain adaptation. In *2019 IEEE intelligent transportation systems conference (ITSC)*, pages 33–40. IEEE.
- [HSRT-E4.1.5.2b-58] Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *2019 international conference on robotics and automation (ICRA)*, pages 7087–7093. IEEE.
- [HSRT-E4.1.5.2b-48] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.
- [HSRT-E4.1.5.2b-07] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.
- [HSRT-E4.1.5.2b-54] Can Peng, Kun Zhao, and Brian C. Lovell. Faster ILOD: Incremental learning for object detectors based on faster RCNN.
- [HSRT-E4.1.5.2b-46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv:1506.01497 [cs]*. arXiv: 1506.01497.
- [HSRT-E4.1.5.2b-11] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv*, April 2020.
- [HSRT-E4.1.5.2b-04] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38<sup>th</sup> International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.
- [HSRT-E4.1.5.2b-09] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000.



## Bibliography

- [HSRT-E4.1.5.2b-06] Weisong Shi and Liangkai Liu. *Computing Systems for Autonomous Driving*. Springer International Publishing, 2021.
- [HSRT-E4.1.5.2b-22] Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, March 2017.
- [HSRT-E4.1.5.2b-53] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting.
- [HSRT-E4.1.5.2b-62] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps.
- [HSRT-E4.1.5.2b-31] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Computer Vision – ECCV 2018*, pages 536–553. Springer International Publishing, 2018.
- [HSRT-E4.1.5.2b-28] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014. ISSN: 1063-6919.
- [HSRT-E4.1.5.2b-02] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018:1–13, 2018.
- [HSRT-E4.1.5.2b-25] Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv*, July 2020.
- [HSRT-E4.1.5.2b-57] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4489–4495. IEEE.
- [HSRT-E4.1.5.2b-34] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, October 2018.
- [HSRT-E4.1.5.2b-30] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 472–487, Cham, 2018. Springer International Publishing.
- [HSRT-E4.1.5.2b-59] Chi-Wen Yang, Tsung-Han Lee, Chien-Lung Huang, and Kuei-Shu Hsu. Unity 3D production and environmental perception vehicle simulation platform. In *2016 international conference on advanced materials for science and engineering (ICAMSE)*, pages 452–455.
- [HSRT-E4.1.5.2b-47] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*.
- [HSRT-E4.1.5.2b-14] Lei Zhang. Transfer adaptation learning: A decade survey. *arXiv*, March 2019.
- [HSRT-E4.1.5.2b-20] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.



## 2.2 E4.1.5.2.b

[HSRT-E4.1.5.2b-12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[HSRT-E4.1.5.2b-49] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

[HSRT-E4.1.5.2b-17] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. In *32<sup>nd</sup> British Machine Vision Conference (BMVC)*, 2021.

### **2.2.6 Survey on Unsupervised Domain Adaptation for Semantic Segmentation for Visual Perception in Automated Driving (TU Braunschweig)**

#### **Literature Research on Unsupervised Domain Adaptation**

In cooperation with DLR and CARIAD SE, TU Braunschweig has prepared a survey article with on topic of "Unsupervised Domain Adaptation (UDA) for Semantic Segmentation for Visual Perception in Automated Driving". In this article we proposed a straightforward classification system that can categorize various works for unsupervised domain adaptation. In addition, we conducted the first-ever survey of vision transformer networks for UDA. Our study is the most comprehensive to date, as we have examined three times as many papers as previous research. Furthermore, we provided a quantitative analysis of the different methods, which highlights methodical and performance trends over the past few years. We examined current approaches try to overcome domain shifts and identify common issues in the training process and evaluation of the adaptation approaches. Lastly, we identified promising research areas for future exploration in this field. For more information, we refer the reader to our article [TUBS-E4.1.5.2b-01]. Figure 2.2: TUBS-E4.1.5.2b-01 shows a brief overview of common adaptation paradigms.

In the following we will provide a brief explanation of various methods for domain adaptation. We will also explain the differences between domain adaptation, domain generalization, and closed- and open-set adaptation. Supervised training, which is shown in the first row of Figure 2.2: TUBS-E4.1.5.2b-01, requires full labels in the target domain, but it is not scalable due to the high cost of manual annotation.

Semi-supervised domain adaptation, as shown in the second row, has been studied in a limited number of works.

Weakly-supervised domain adaptation, shown in the third row, involves the use of samples with noisy labels or image-level labels [TUBS-E4.1.5.2b-02].

Source-free UDA, shown in the fifth row, refers to the task of UDA when no source domain data is available except for the pre-trained network weights [TUBS-E4.1.5.2b-03]. This can be further extended by continual/continuous UDA, shown in the sixth and seventh rows, which refers to a decoupling of training and adaptation, with the source domain being unavailable during adaptation and continuously changing target domains [TUBS-E4.1.5.2b-04, TUBS-E4.1.5.2b-05].

Domain generalization, shown in the last row, involves training networks to perform better in unseen domains without using any data from the target domain for adaptation [TUBS-E4.1.5.2b-06, TUBS-E4.1.5.2b-07].

Closed-set adaptation refers to the assumption that only the visual domain changes, while the number of pre-defined semantic classes remains unchanged, while open-set adaptation assumes that new classes may have to be learned.



Table 2.1: TUBS-E4.1.5.2b-01: ResNet- and transformer-based semantic segmentation results for baseline training and UDA methods.

Backbone	Method	mIoU (%) on				Mean
		$\mathcal{D}_{\text{test}^*}^{\text{CS}}$	$\mathcal{D}_{\text{test}^*}^{\text{MV}}$	$\mathcal{D}_{\text{test}^*}^{\text{BDD}}$	$\mathcal{D}_{\text{test}^*}^{\text{ACDC}}$	
ResNet	Baseline	41.0	46.0	39.2	32.1	39.6
ResNet	SAC [TUBS-E4.1.5.2b-15] ( $\rightarrow$ CS)	53.8	48.9	40.2	35.6	44.6
MiT-B5	SegFormer (Baseline)	44.5	49.8	42.6	36.8	43.4
MiT-B5	DAFormer [TUBS-E4.1.5.2b-13]( $\rightarrow$ CS)	<b>67.1</b>	<b>60.2</b>	<b>52.5</b>	<b>44.7</b>	<b>56.1</b>

We focused on the unsupervised domain adaptation paradigm in the survey. In this project we also investigated methods that perform source-free and continuous domain adaptation. Experimental results on different methods can be found in the result texts E4.2.2.1, E4.2.2.2, and E4.2.2.4. In the following we will give an additional brief overview of the latest methods for domain adaptation for semantic segmentation, which can also be found in the aforementioned survey paper [TUBS-E4.1.5.2b-01]. Particularly noteworthy are the methods based on vision transformer, which have recently emerged as a novel research direction in computer vision. Transformer networks with attention mechanisms were initially developed for language processing, and recently, they have gained much attention in computer vision as well. The basic vision transformer (ViT) [TUBS-E4.1.5.2b-08] is one of the foundational works, where the self-attention mechanism is the major change compared to standard convolutional architectures, such as the ResNet models. The self-attention mechanism, which learns the relations between the elements of a sequence of inputs and captures how the sequence elements influence each other, is replaced by image patches of  $16 \times 16$  pixels for images in ViT. Several new architectures such as Swin transformer [TUBS-E4.1.5.2b-09], pyramid transformer [TUBS-E4.1.5.2b-10], SegFormer [TUBS-E4.1.5.2b-11], and HRViT [TUBS-E4.1.5.2b-12] have been developed for semantic segmentation, using smaller image patches for this purpose.

These networks have been shown to have higher robustness against perturbations and better generalization capabilities than standard CNNs. DAFormer [TUBS-E4.1.5.2b-13] is the foundational employing vision transformers for unsupervised domain adaptation, proposing novel contributions to both method and architecture levels. It uses SegFormer as the encoder architecture and employs self-training with a teacher-student framework, strong augmentations, and confidence weighting, among other techniques. In addition, rare class sampling on the source domain and a feature distance loss to the pre-trained ImageNet features are part of the DAFormer approach. HRDA [TUBS-E4.1.5.2b-14] builds upon DAFormer and achieves state-of-the-art results in unsupervised domain adaptation.

In Table 2.1 we compare the ResNet-based SAC method [TUBS-E4.1.5.2b-15] with the transformer-based DAFormer method [TUBS-E4.1.5.2b-13]. It can be seen that the DAFormer methods achieves best performance in all tested target domains, even if it is only adapted to Cityscapes. In Chapter E4.2.2.4 we also show, that the DAFormer outperforms continuous and domain generalization methods on all domains.

### Training of Reference Models for Domain Adaptation

The results for the selected reference models and methods are presented in the results texts of the respective APs (E4.2.2.1, E4.2.2.2, and E4.2.2.4.). The models were trained on the GPU cluster of the Institute for Communications Technology at the TU Braunschweig. For some methods no implementations were

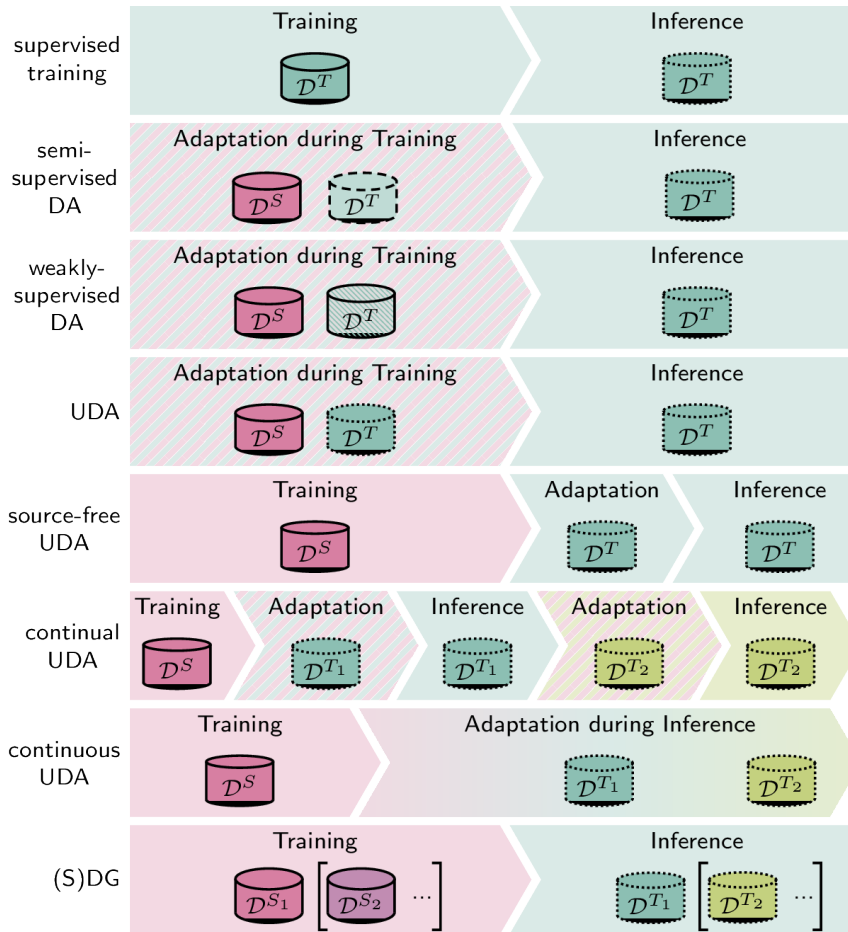


Figure 2.2: TUBS-E4.1.5.2b-01: Simplified diagram providing an overview of the adaptation paradigms. The red and green colors in the diagram represent the source and target domains, respectively. Dotted lines indicate datasets without available labels, while dashed lines indicate a subset of labels. For weakly-supervised DA, the available labels are noisy.

available, in this case the reported performance is cited from the respective publication, which is always indicated. For other methods model weights were provided that were evaluated on the GPU cluster, which is also always indicated.

## Bibliography

- [TUBS-E4.1.5.2b-15] Nikita Araslanov and Stefan Roth. Self-Supervised Augmentation Consistency for Adapting Semantic Segmentation. In *Proc. of CVPR*, pages 15384–15394, virtual, June 2021.
- [TUBS-E4.1.5.2b-07] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening. In *Proc. of CVPR*, pages 11580–11590, virtual, June 2021.
- [TUBS-E4.1.5.2b-08] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*, 2020.
- [TUBS-E4.1.5.2b-12] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen



## Bibliography

- Lai, Vikas Chandra, and David Z Pan. Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation. In *Proc. of CVPR*, pages 12094–12103, 2022.
- [TUBS-E4.1.5.2b-13] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies For Domain-Adaptive Semantic Segmentation. In *Proc. of CVPR*, pages 9924–9935, New Orleans, LA, USA, June 2022.
- [TUBS-E4.1.5.2b-14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proc. of the ECCV*, 2022.
- [TUBS-E4.1.5.2b-02] Niklas Hanselmann, Nick Schneider, Benedikt Ortelt, and Andreas Geiger. Learning Cascaded Detection Tasks With Weakly-Supervised Domain Adaptation. In *Proc. of IV*, pages 532–539, virtual, July 2021.
- [TUBS-E4.1.5.2b-04] Marvin Klingner, Mouadh Ayache, and Tim Fingscheidt. Continual BatchNorm Adaptation (CBNA) for Semantic Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):20899–20911, 2022.
- [TUBS-E4.1.5.2b-03] Marvin Klingner, Jan-Aike Termöhlen, Jacob Ritterbach, and Tim Fingscheidt. Unsupervised BatchNorm Adaptation (UBNA): A Domain Adaptation Method for Semantic Segmentation Without Using Source Domain Representations. In *Proc. of WACV - Workshops*, pages 210–220, Waikoloa, HI, USA, January 2022.
- [TUBS-E4.1.5.2b-09] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proc. of ICCV*, pages 10012–10022, virtual, October 2021.
- [TUBS-E4.1.5.2b-06] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning Domain Generalized Semantic Segmentation From the Wild. In *Proc. of CVPR*, pages 9936–9946, New Orleans, LA, USA, June 2022.
- [TUBS-E4.1.5.2b-01] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. Schäfer, Nico M. Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *arXiv:2304.11928*, pages 1–40, April 2023.
- [TUBS-E4.1.5.2b-05] Jan-Aike Termöhlen, Marvin Klingner, Leon J. Brettin, Nico M. Schmidt, and Tim Fingscheidt. Continual Unsupervised Domain Adaptation for Semantic Segmentation by Online Frequency Domain Style Transfer. In *Proc. of ITSC*, pages 2881–2888, virtual, September 2021.
- [TUBS-E4.1.5.2b-10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *Proc. of ICCV*, pages 568–578, virtual, October 2021.
- [TUBS-E4.1.5.2b-11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation With Transformers. In *Proc. of NeurIPS*, pages 12077–12090, virtual, December 2021.



## 2.3 E4.1.5.2.c: Methodensammlung, Katalog, Übersicht, SOTA zum Thema Didaktik Stand Ende PI4

### 2.3.1 M3: Monocular Self-Supervised Depth, Pose and Motion (BMW)

#### Introduction

Accurately estimated depth and motion from images would be a low cost highly beneficial contribution to the automated driving (AD) sensing stack. Vision could provide redundancy or even replace typical range measuring sensors such as radar and lidar. The automated vehicle does not only need to know the current 3D geometry of the scene but also has to predict traffic participants' motion. Typically the only sensor which directly measures relative velocity is radar, providing only sparse measurements from a restricted field of view. An additional, dense 3D motion estimate from video data would thus be highly beneficial.

Supervised approaches require recording of ground truth with specialized settings such as lidar sensors or stereo cameras. Such settings are expensive and need to be well calibrated. This problem is especially prominent for motion estimation, where ground truth generation is hindered by the lack of reference sensors with direct, dense measurement capability. The only publicly available real-world scene flow dataset for AD is KITTI Scene Flow 2015 [BMW-E4.2.3.1-19]. Ground truth labeling a total of only 400 training and testing images required a multistage post-processing pipeline, including fitting of 3D CAD models to pointcloud measurements.

Recently, approaches to monocular self-supervised depth estimation drew interest in the research community (e.g. [BMW-E4.2.3.1-23, BMW-E4.2.3.1-26, BMW-E4.2.3.1-24]). These networks are able to estimate relative camera pose and pixel-wise depth from sequences of images with no further information besides camera intrinsics. Contrary to stereoscopic approaches, these methods have to resort to camera pose changes caused by vehicle motion and thus time. This violates the underlying assumption of a static scene and prevents the correct prediction and handling of motion in environments with dynamic objects. We argue that the assumption regarding static objects in a sequence of images is hurtful in this setting: mobility of objects is a key feature of driving imagery. Not accounting for it can result in vehicles being projected to infinity [BMW-E4.2.3.1-26, BMW-E4.2.3.1-14]. Such failure modes can have far reaching consequences in safety critical scenarios such as automated driving and might thus impede real life application altogether.

The presented work relaxes the static scene assumption and adds motion estimation to the self-supervised stack. This results in a dense, three dimensional motion field. Our design choices are guided by the assumptions that (1) most pixels in a scene are static and (2) velocity of moving objects is constant within the temporal context of the networks.

Our main contribution is a new method to jointly estimate depth, relative pose and object motion, called M3. The name derives from estimation of the mentioned three core properties of the 3D scene using only monocular image sequences. M3 is trained from sequences of images with known camera intrinsics. At inference time only one image for depth estimation and two images for pose and object motion are needed. Our method provides new state of the art results for this setting measured on the KITTI Scene Flow evaluation.



## Related Work

The breakthrough work of Eigen et al. [BMW-E4.2.3.1-07] estimated depth from single images by regressing directly to ground truth depth values using neural networks. This spurred a significant amount of follow up work (e.g. [BMW-E4.2.3.1-17, BMW-E4.2.3.1-15, BMW-E4.2.3.1-14]). Impressive results were achieved by making classical geometrical ideas differential and thus end-to-end trainable [BMW-E4.2.3.1-01]. This approach features iterative refinement of depth and motion to regress to supervised training samples. Current work by Ranftl et al. [BMW-E4.2.3.1-06] circumvents the hard task of collecting ground truth by developing mixing strategies for various data sources and further includes a dataset with depth information extracted from 3D movies.

The work of Zhou et al. [BMW-E4.2.3.1-27] was the first to alleviate the problem of scarce ground truth by learning from sequences of images alone. The method was quickly improved by follow up works (e.g. [BMW-E4.2.3.1-25, BMW-E4.2.3.1-13, BMW-E4.2.3.1-17]). A notable state of the art architecture is BMW-E4.2.3.1-26 [BMW-E4.2.3.1-26]: Two networks are used to estimate depth and relative pose respectively. As pose and depth can be used to synthesize one view from another, a pixel-wise photometric loss between reconstruction and original image can be calculated. To deal with occlusions, BMW-E4.2.3.1-26 pioneered taking the pixel-wise minimum of the photometric loss from adjacent frames with differing temporal order. BMW-E4.2.3.1-24-SfM [BMW-E4.2.3.1-24] further improves the accuracy of this method by replacing the ResNet [BMW-E4.2.3.1-12] architecture of the depth estimation network with 3D packing and unpacking operations which preserve fine grained spatial structures. Additionally this work addresses the arbitrary scale issue from which the self-supervised works typically suffer. They solve this issue by incorporating velocity into the loss, which constrains the estimated pose to lie on a circle. The arbitrary scale issue is also tackled in [BMW-E4.2.3.1-16] by introducing a loss forcing neighboring frames to be depth consistent.

Self-supervision in general was studied in a very recent survey paper, showing very promising results [BMW-E4.2.3.1-37]

Recently several methods explored self-supervised estimation of motion. Scene flow is the 3D motion of every point in an image. Optical flow is the projection of scene flow onto the camera plane. It is thus closely linked to scene flow, in fact scene flow can be recovered from optical flow and two depth estimates. This close link was e.g. exploited in [BMW-E4.2.3.1-17, BMW-E4.2.3.1-10] and [BMW-E4.2.3.1-15]. While [BMW-E4.2.3.1-17, BMW-E4.2.3.1-10] estimate optical flow to mask out moving regions and potentially estimate scene flow, [BMW-E4.2.3.1-15] adopts PWC-Net [BMW-E4.2.3.1-11] to estimate 3D scene flow directly, albeit with stereo supervision. Stereo images and 2D object detections are also used by Cao et al. [BMW-E4.2.3.1-13] to regress directly to 3D scene flow. Similarly to [BMW-E4.2.3.1-15] and [BMW-E4.2.3.1-13], Li et al. [BMW-E4.2.3.1-03] regress and regularize to a motion field directly, in their case without the need of stereo input. This is the only work known to the authors which estimates 3D object motion directly from monocular image sequences, but only evaluates depth estimation.

M3 builds upon and extends RAFT [BMW-E4.2.3.1-09], the current state of the art for optical flow estimation, to directly predict 3D object motion. We tightly integrate it in into a monodepth2-like depth and pose estimation pipeline.



## Bibliography

- [BMW-E4.2.3.1-37] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023.
- [BMW-E4.2.3.1-13] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik. Learning independent object motion from unlabelled stereoscopic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5594–5603, 2019.
- [BMW-E4.2.3.1-16] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [BMW-E4.2.3.1-07] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.
- [BMW-E4.2.3.1-24] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [BMW-E4.2.3.1-19] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020.
- [BMW-E4.2.3.1-14] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth, 2020.
- [BMW-E4.2.3.1-25] Clément Godard, Oisín Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.
- [BMW-E4.2.3.1-26] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.
- [BMW-E4.2.3.1-15] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7405, 2020.
- [BMW-E4.2.3.1-12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [BMW-E4.2.3.1-03] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, and Anelia Angelova. Unsupervised monocular depth learning in dynamic scenes, 2020.
- [BMW-E4.2.3.1-10] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019.



### 2.3 E4.1.5.2.c

- [BMW-E4.2.3.1-06] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [BMW-E4.2.3.1-11] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.
- [BMW-E4.2.3.1-01] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion, 2020.
- [BMW-E4.2.3.1-09] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.
- [BMW-E4.2.3.1-23] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.
- [BMW-E4.2.3.1-17] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.
- [BMW-E4.2.3.1-27] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

## 2.3.2 Active Learning Methods Based on Consistency and Uncertainty (BMW)

### Introduction

The goal of Active Learning (AL) is to improve labeling efficiency and thus reduce the annotation cost by designing a learning algorithm that queries the most valuable unlabeled data to be labeled. This work focuses on the semantic segmentation task which is especially annotation intensive. In particular, we investigate acquisition functions based on inconsistencies between different predictions and compare them to uncertainty based AL approaches.

### Related Work

In this section we summarize in bullet points representative works in consistency-based AL for semantic segmentation and object detection. We also highlight recent works in the emerging field of region-based AL for semantic segmentation. Consistency-based AL can be broadly divided into approaches which make use of multiple views, e.g., different augmented images, and works that rely on multiple models or model parts.

#### 1. Prediction Inconsistency between Different Views

##### a) Augmentations

Typical augmentations which can be used to compute inconsistencies are flipping, color transformations, blurring and noise injection. For other transformations such as translations and rotations, only the overlapping region can be scored.



## 2.3 E4.1.5.2.c

- i. Not All Labels Are Equal: Rationalizing The Labeling Costs for Training Object Detection (CVPR, 2022, [BMW-E4.1.5.2c-01])
  - Datasets: PASCAL VOC and MS-COCO
  - Unified strategy to choose which samples to manually label and which samples can be automatically labeled
  - Claim that robustness of the detector (consistency over augmentations) is more reliable than uncertainty of the detector as an acquisition function, especially for the low-performing classes
- ii. Localization-Aware Active Learning for Object Detection (2018, [BMW-E4.1.5.2c-02])
  - Datasets: PASCAL VOC and MS-COCO
  - Two metrics a) “localization tightness” of an object hypothesis, which is based on the overlapping ratio between the region proposal and the final prediction b) “localization stability” of an object hypothesis, which is based on the variation of predicted object locations when input images are corrupted by noise

### b) Pose Transformations

This approach requires camera pose and depth information which may be problematic on dynamic scenes.

- i. ViewAL: Active Learning With Viewpoint Entropy for Semantic Segmentation (CVPR, 2020, [BMW-E4.1.5.2c-03])
  - Datasets: SceneNet-RGBD, ScanNet, and Matterport3D
  - Code: <https://github.com/nihalsid/ViewAL>
  - Approach requires 3D camera pose for each view and depth information for all pixels to cross-project predicted class distributions (for all pixels) to other views of the scene
  - Propose to label only superpixels (computed with the SEEDS algorithm)
  - Use the trained network to compute a view entropy score and a view divergence score for each unlabeled superpixel based on the projected class distributions.
  - The divergence score is computed as the mean KL divergence of the cross-projected class distributions
  - Use MC dropout (20 per view) to make probability estimates more robust to changes in the input

## 2. Prediction Inconsistency between Different Models or Model Parts

### a) Inconsistency within Model Ensemble

- i. Scalable Active Learning for Object Detection (IV, 2020, [BMW-E4.1.5.2c-04])
  - Datasets: PASCAL VOC and MS-COCO
  - Uses ensemble of models to compute 4 different informativeness scores (entropy, mutual information, gradient of the output layer, bounding boxes with confidence)



## 2.3 E4.1.5.2.c

- Scoring function outputs image heat map (informativeness score at each pixel and for each class), which are aggregated using max or average
  - ii. Advanced Active Learning Strategies for Object Detection (IV, 2020, [BMW-E4.1.5.2c-05])
    - Datasets: KITTI
    - Compute uncertainties with ensembles for 2D and 3D object detection (in contrast to MC dropout ensembles don't require changes in architecture)
    - Best performing score is calculated by finding the minimum IoU value for each RoI-match (max IoU) among all comparisons and taking the average (over RoIs)
    - Investigate different training strategies (continual learning) and imbalanced datasets
  - b) Disagreement between Predictions of Different Layers
    - i. Deep Active Learning for Object Detection (BMVC, 2018, [BMW-E4.1.5.2c-06])
      - Datasets: PASCAL VOC and KITTI
      - Query by committee with different convolution layers from SSD
  - c) Different Predictions of Multiple Heads
    - i. Multiple Instance Active Learning for Object Detection (CVPR, 2021, [BMW-E4.1.5.2c-07])
      - Datasets: PASCAL VOC and MS-COCO
      - Code: <https://github.com/yuantn/MI-AOD>
      - Use two adversarial instance classifiers and single bounding box regressor with shared backbone
      - The instance uncertainty is defined as the prediction discrepancy of the two classifiers
      - Re-weights the instance uncertainty scores by minimizing the image classification loss in an expectation-maximization fashion
  - d) Inconsistent Predictions when Using Different Input/Output Modalities
    - i. Deep active learning for efficient training of a lidar 3d object detector (IV, 2019, [BMW-E4.1.5.2c-08])
      - Datasets: KITTI
      - Leverages 2D region proposals generated from the RGB images to reduce the search space of objects
      - Use a “perfect” image detector to compare several ways of uncertainty estimation: MC-dropout and deep ensembles provide more reliable predictive uncertainties compared to the single softmax output
      - No location uncertainty
- ### 3. Uncertainty-based Active Learning
- i. Active Learning for Deep Object Detection via Probabilistic Modeling (ICCV, 2021, [BMW-E4.1.5.2c-09])
    - Datasets: PASCAL VOC and MS-COCO
    - Code: <https://github.com/NVlabs/AL-MDN>



- Use mixture density networks that estimate a probabilistic distribution for each localization and classification head's output
- Explicitly estimate the aleatoric and epistemic uncertainty in a single forward pass of a single model. They use a scoring function that aggregates these two types of uncertainties for both heads to obtain every image's informativeness score
- ii. Importance of Self-Consistency in Active Learning for Semantic Segmentation (BMVC, 2020, [BMW-E4.1.5.2c-10])
  - Datasets: Cityscapes and CamVid
  - Code: <https://github.com/isalirezag/EquAL>
  - EquAL: Use average entropy of predictions over original and flipped image.
  - EquAL+: During model training use inconsistency as additional self-supervised loss

#### 4. Region-based Active Learning

An emerging field for AL for semantic segmentation is region-based AL. In contrast to the conventional AL approach these methods query regions, i.e. rectangles, super pixels or individual pixels, of images for labeling.

- i. Revisiting Superpixels for Active Learning in Semantic Segmentation with Realistic Annotation Costs (CVPR, 2021, [BMW-E4.1.5.2c-11])
  - Datasets: Cityscapes and PASCAL VOC
  - Code: <https://github.com/cailile/Revisiting-Superpixels-for-Active-Learning>
  - Query superpixels for annotation
- ii. All you need are a few pixels: semantic segmentation with PIXELPICK (ICCV, 2021, [BMW-E4.1.5.2c-12])
  - Datasets: Cityscapes and PASCAL VOC and CamVid
  - Query single pixels for annotation (no localization required anymore)

## Bibliography

[BMW-E4.1.5.2c-09] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021.

[BMW-E4.1.5.2c-11] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10988–10997, 2021.

[BMW-E4.1.5.2c-01] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14492–14501, 2022.



### 2.3 E4.1.5.2.c

- [BMW-E4.1.5.2c-08] Di Feng, Xiao Wei, Lars Rosenbaum, Atsuto Maki, and Klaus Dietmayer. Deep active learning for efficient training of a lidar 3d object detector. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 667–674. IEEE, 2019.
- [BMW-E4.1.5.2c-10] S Alireza Golestaneh and Kris M Kitani. Importance of self-consistency in active learning for semantic segmentation. *arXiv preprint arXiv:2008.01860*, 2020.
- [BMW-E4.1.5.2c-04] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecy, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M Alvarez. Scalable active learning for object detection. In *2020 IEEE intelligent vehicles symposium (iv)*, pages 1430–1435. IEEE, 2020.
- [BMW-E4.1.5.2c-02] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-aware active learning for object detection. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 506–522. Springer, 2019.
- [BMW-E4.1.5.2c-06] Soumya Roy, Asim Unmesh, and Vinay P Nambodiri. Deep active learning for object detection. In *BMVC*, volume 362, page 91, 2018.
- [BMW-E4.1.5.2c-05] Sebastian Schmidt, Qing Rao, Julian Tatsch, and Alois Knoll. Advanced active learning strategies for object detection. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 871–876. IEEE, 2020.
- [BMW-E4.1.5.2c-03] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9433–9443, 2020.
- [BMW-E4.1.5.2c-12] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1687–1697, 2021.
- [BMW-E4.1.5.2c-07] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021.

**Author Deliverable:** Manuel Schwonberg

**Authors Survey:** Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. Schäfer, Nico M. Schmidt, Hanno Gottschalk, Tim Fingscheidt

Copyright IEEE: <https://arxiv.org/pdf/2304.11928.pdf> (not yet accepted but we expect this in the near future so copyright notice already included)

Parts of the following text are directly taken from the survey paper [CARIAD-E4.1.5.2.c-01].

### 2.3.3 CARIAD

#### Introduction

As the major part of our literature research and overview and in cooperation with TU Braunschweig and DLR, CARIAD contributed to a survey on unsupervised domain adaptation for semantic segmentation. Pre-



vious surveys on this topics were either relatively old [CARIAD-E4.1.5.2.c-04] in this dynamically emerging research field or had a smaller extend of covered research works[CARIAD-E4.1.5.2.c-03]. Despite the recent progress of Deep Neural Networks on a variety of different tasks several issues still need to be addressed that limit the applicability of DNNs in automated driving. The bad generalization of DNNs to new, unseen domains is a major problem on the way to a safe, large-scale application, because manual annotation of new domains is costly, particularly for semantic segmentation. For this reason, methods are required to adapt DNNs to new domains without labeling effort. We categorize and explain the different approaches for UDA. The number of considered publications is larger than any other survey on this topic. We conducted a critical analysis of the state-of-the-art and highlight promising future research directions. With this survey, we aim to facilitate UDA research further and encourage scientists to exploit novel research directions to generalize DNNs better.

We want to point out that we also created a website for our survey which contains a leaderboard and interactive plots and direct links to all papers. The website can be accessed here.

### Taxonomy

Based on our findings we proposed the taxonomy which can be seen in figure 2.3. Unsupervised domain adaptation approaches can be divided into largely four categories where the three major categories are build by the main machine learning spaces: input, feature and output space. When two or more of them are combined within one approach we refer to them as hybrid domain adaptation methods. For each of the spaces we also provide more fine-grained clustering but please refer to our paper for more details [CARIAD-E4.1.5.2.c-01].

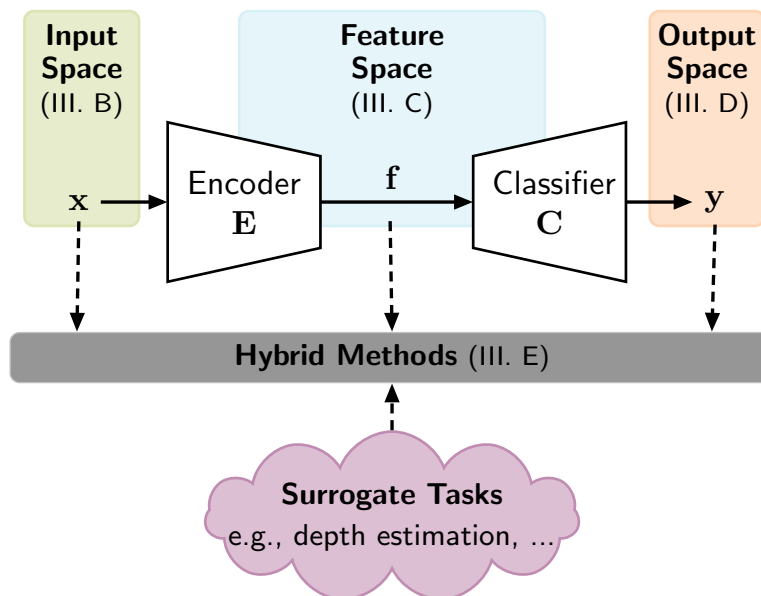


Figure 2.3: CARIAD-E4.1.5.2c-01: UDA taxonomy

### Quantitative Comparison

Based on the large database behind the survey with over 150 scientific works we were able to conduct a large-scale quantitative comparison leading to a meta analysis of the progress in the field of unsupervised



### 2.3 E4.1.5.2.c

domain adaptation for semantic segmentation in the past years.

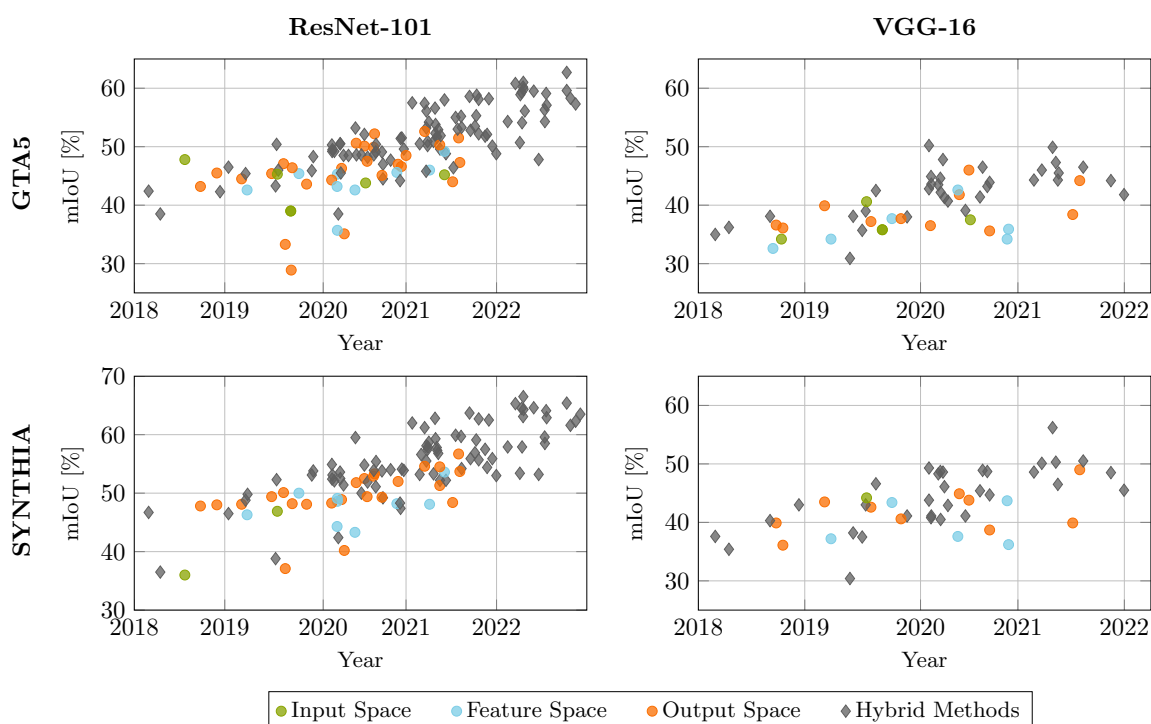


Figure 2.4: CARIAD-E4.1.5.2c-02: Performance (mIoU (%)) on the Cityscapes validation set after training on the source domains GTA5 (top row) or SYNTHIA (bottom row) with simultaneous adaptation to Cityscapes. The results are shown for models based on a ResNet-101 feature extractor (left column) or a VGG-16 feature extractor (right column). The reported values are taken from the respective papers.

The two included plots in figure 2.4 and 2.5 reveal different novel findings. First we can clearly observe a strong domination of hybrid methods which represent the majority of the best performing unsupervised domain adaptation approaches. In 2022 only hybrid approaches were published. Second, we can observe that the performance boost over the 50% mIoU was mostly carried by the hybrid approaches and no stand-alone approach significantly crosses this performance line. For more details please refer to our survey paper.

### Future Research Directions

We identified interesting future research directions which may serve as a valuable input to other researchers. The standardization for benchmarking approaches is essential for a valid comparison since even slightly different setting may significantly impact the measured performance. Also the mIoU as a single scalar is used for evaluation of all the works in our database. While this performance criteria is essential it does not provide much information about what happens inside the network under domain shift so more meaningful metrics can be a useful research direction. Recently vision transformer architectures like DAFormer [CARIAD-E4.1.5.2.c-02] entered the area of unsupervised domain adaptation and researching their generalization capabilities will be an interesting direction. On the side of applicability domain adaptation still requires target data so domain generalization where no target data is required might be even

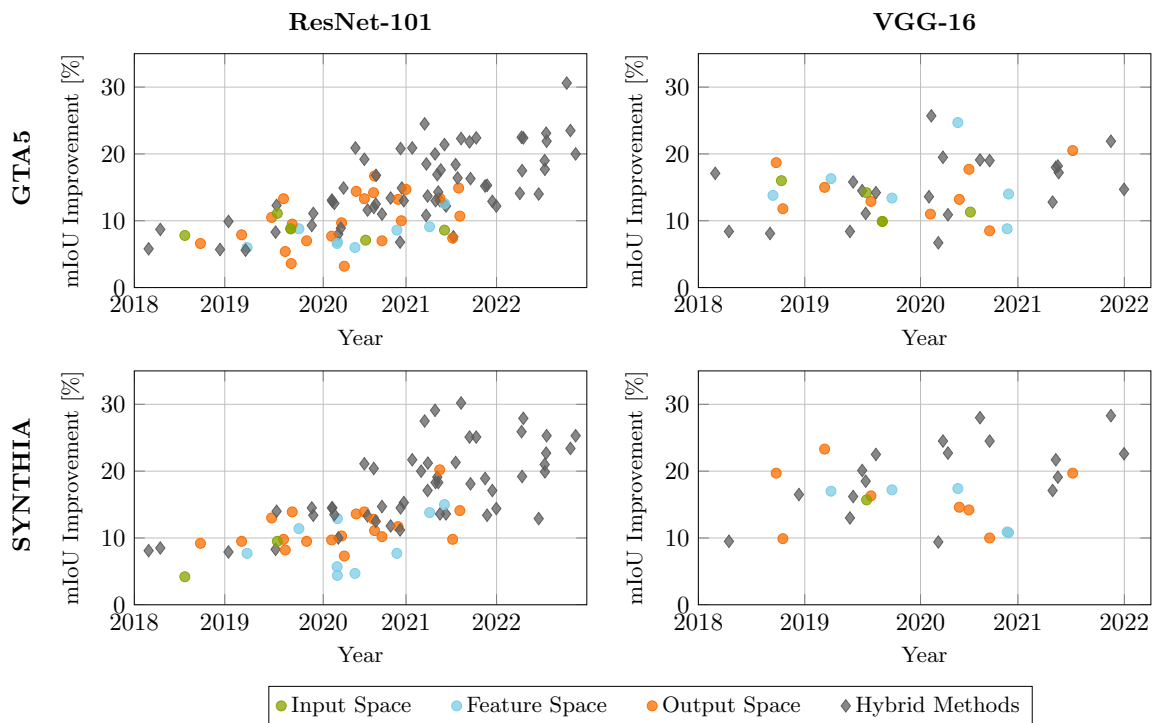


Figure 2.5: CARIAD-E4.1.5.2c-03: Performance improvement (mIoU (% absolute)) on the Cityscapes validation set after training on the source domains GTA5 (top row) or SYNTHIA (bottom row) with simultaneous adaptation to Cityscapes. The results are shown for models based on a ResNet-101 feature extractor (left column) or a VGG-16 feature extractor (right column). The reported values are taken from the respective papers. Note that not all papers provide a baseline performance without adaptation.

more interesting since the objective is to generalize to multiple unseen target domains. For more details also check out our survey paper.

## Bibliography

- [CARIAD-E4.1.5.2.c-03] G. Csurka. Domain Adaptation for Visual Applications: A Comprehensive Survey. *arXiv*, (1702.05374), March 2017.
- [CARIAD-E4.1.5.2.c-02] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies For Domain-Adaptive Semantic Segmentation. In *Proc. of CVPR*, pages 9924–9935, New Orleans, LA, USA, June 2022.
- [CARIAD-E4.1.5.2.c-01] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. Schäfer, Nico M. Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving, 2023.
- [CARIAD-E4.1.5.2.c-04] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised Domain Adaptation in Semantic Segmentation: A Review. *Technologies*, 8(2):1–35, 2020.



### 2.3.4 Corner Case Detection (HSRT)

Markus Rehmann - Hochschule Reutlingen

Vinu Nair - Hochschule Reutlingen

Hochschule Reutlingen refers to [HSRT-E4.1.5.2c-01] for an overview of deep learning methods applied to anomaly detection. Recently, generative adversarial networks have been used for anomaly detection in images [HSRT-E4.1.5.2c-02, HSRT-E4.1.5.2c-03, HSRT-E4.1.5.2c-04]. GANs learn to sample from the distribution of training data. This can be exploited for anomaly detection by making a GAN encode an image and reconstruct it again. Reconstruction should be successful as long as the image is from the known distribution but should otherwise fail. Anomalies can be identified by comparing the original and reconstructed image [HSRT-E4.1.5.2c-04], which allows the localization of anomalies in images.

GANomaly [HSRT-E4.1.5.2c-04] uses an Autoencoder for image reconstruction with the addition of using the encoder part of the model on the reconstructed image. Which allows the use of an encoder loss, calculated from the latent space of the original image and the reconstruction, in addition to the usual reconstruction loss. This encoder loss helps in the detection of anomalies. CycleGAN [HSRT-E4.1.5.2c-06] is an approach used for domain adaptation. And its generator is a useful baseline for an image reconstruction network, which already works well in other scenarios. The concept of Encoded Human Pose Images (EHPI) defined in [HSRT-E4.1.5.2c-07] can be used to transform sequential pose data into the image domain, intended for action recognition. Simple baselines for human pose estimation [HSRT-E4.1.5.2c-05] is a pose estimation network, which was utilized in the evaluation of the created methods.

The used methods are dependent on the input domain of the network. In the RGB image domain, an approach including an additional encoder loss, based on GANomaly [HSRT-E4.1.5.2c-04], has been combined with either the CycleGAN [HSRT-E4.1.5.2c-06] Generator or a simpler CNN Network for the reconstruction task. In the pose domain, linear and CNN Autoencoders were implemented for pose reconstruction. Additionally, the concept of EHPI [HSRT-E4.1.5.2c-07] was utilized to transform the pose data into an image, which allowed for the use of the models described for the RGB image domain. Furthermore, PCA [HSRT-E4.1.5.2c-08] was used as a traditional machine learning approach in the pose domain.

### 2.3.5 Self attention techniques in the context of unsupervised domain adaptation for human pose estimation models (HSRT)

#### Introduction

The training of machine learning algorithms for e.g., object detection, semantic segmentation, or human pose estimation usually requires huge labelled datasets. The process of creating these labelled datasets is very time-consuming and therefore expensive for data recorded on real sensors. Simulation, on the other hand, provides free ground truth annotations, but the models trained on synthetic data perform poorly on real data. Our goal is to improve the performance of a pose estimation model on the target (real) domain through the use of an Unsupervised Domain Adaptation(UDA) and self-attention networks. We did this by improving the model performance on the source domain using self-attention methods and using an UDA model to adapt the model to target domain.



### Related work

Attention mechanisms have shown to be effective for sequence processing, i.e., natural language processing and machine translation, because they allow modelling dependencies in sequences without regard to their distance in the input or output sequence [HSRT-E4.1.5.2c-9], [HSRT-E4.1.5.2c-10]. Vaswani et al. have shown that global dependencies can entirely be handled by attention mechanisms [HSRT-E4.1.5.2c-10]. Attention can also be applied to the image domain. The ability of CNNs to capture local dependencies in images relies on the size of the receptive field. To overcome this local limitation SAGAN [HSRT-E4.1.5.2c-11] uses a self-attention module in the discriminator as well as the generator of a GAN to capture long-range dependencies and generate more realistic images. Squeeze-and-Excitation blocks, adaptive recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels[HSRT-E4.1.5.2c-12]. Convolutional Block Attention Module (CBAM) infers attention maps sequentially by first applying feature-map attention and then spatial attention to find the refined feature-maps[HSRT-E4.1.5.2c-13]. CBAM-GAN applied CBAM after some convolution operators to adaptive rescale spatial and channel-wise features in GAN based models, which helped to enhance salient regions and extract more detailed features[HSRT-E4.1.5.2c-14]. Self-attention mechanisms including SAGAN suffer from quadratic increase in memory and compute requirements as the spatial and feature resolution increases. We used the EANET attention mechanism, capable of learning more representative features for the input while reducing computational cost [HSRT-E4.1.5.2c-15]. EANET computes the relation between self-queries and a much smaller learnable key memory, which captures the global context of the dataset.

### Bibliography

- [HSRT-E4.1.5.2c-04] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 622–637. Springer International Publishing.
- [HSRT-E4.1.5.2c-01] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. [abs/1901.03407](https://arxiv.org/abs/1901.03407).
- [HSRT-E4.1.5.2c-03] Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 3–17. Springer International Publishing.
- [HSRT-E4.1.5.2c-15] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks.
- [HSRT-E4.1.5.2c-12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. pages 7132–7141.
- [HSRT-E4.1.5.2c-07] Dennis Ludl, Thomas Gulde, and Cristobal Curio. Simple yet efficient real-time pose-based action recognition. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, oct 2019.
- [HSRT-E4.1.5.2c-9] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation.



## Bibliography

- [HSRT-E4.1.5.2c-02] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi. A survey on gans for anomaly detection. [abs/1906.11632](https://arxiv.org/abs/1906.11632).
- [HSRT-E4.1.5.2c-14] Bing Ma, Xiaoru Wang, Heng Zhang, Fu Li, and Jiawang Dan. Cbam-gan: Generative adversarial networks based on convolutional block attention module. In Xingming Sun, Zhaoqing Pan, and Elisa Bertino, editors, *Artificial intelligence and security*, pages 227–236, Cham. Springer International Publishing.
- [HSRT-E4.1.5.2c-08] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. 2(11):559–572.
- [HSRT-E4.1.5.2c-10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30.
- [HSRT-E4.1.5.2c-13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the european conference on computer vision (ECCV)*, pages 3–19.
- [HSRT-E4.1.5.2c-05] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- [HSRT-E4.1.5.2c-11] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. pages 7354–7363.
- [HSRT-E4.1.5.2c-06] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.