Evaluating Federated Dino's performance on the segmentation task across diverse domains

Anonymous Author(s)

ABSTRACT

This study investigates the performance of the DINOv2 pretrained model within Federated Learning (FL) environments, focusing on its application to segmentation tasks across diverse domains. While DINOv2 has demonstrated high efficacy in centralized training scenarios, its capabilities under FL conditions—where data privacy and security are paramount—remain underexplored. Utilizing data sets spanning industrial, medical, and automotive sectors, we evaluated DINOv2's accuracy and generalization in decentralized settings. Our findings reveal that federated DINOv2 performs comparably to centralized models, effectively segmenting objects despite the decentralized and heterogeneous nature of the data. However, inherent biases in the pretrained model posed challenges, affecting performance across different domains. These results highlight the need for domain-specific fine-tuning and bias mitigation strategies to enhance the robustness of pretrained models in FL contexts. Future work should address these challenges to maximize the potential of FL in privacy-sensitive applications, ensuring high performance while maintaining data confidentiality.

CCS CONCEPTS

• **General and reference** → **Evaluation**; *Experimentation*; • **Computing methodologies** → **Distributed artificial intelligence**; *Image segmentation.*

KEYWORDS

Federated Learning, Data sets, Evaluation, Image segmentation

ACM Reference Format:

Anonymous Author(s). 2024. Evaluating Federated Dino's performance on the segmentation task across diverse domains. In *Proceedings of International Joint Workshop on Federated Learning for Data Mining and Graph Analytics (⟨KDD '24 FedKDD Workshop⟩).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The rapid advancements in machine learning and computer vision have led to sophisticated models capable of tasks like image classification, object detection, and semantic segmentation. One such model is DINOv2 [22], renowned for its effectiveness in identifying and segmenting objects within images. While extensively documented in centralized training environments, its performance under Federated Learning (FL) conditions remains underexplored.

FL offers a decentralized approach, where models are trained across multiple devices or servers holding local data samples without needing to exchange these samples [19]. FL is increasingly important in domains where data privacy and security are critical, such as industrial applications and the medical sector. The European Unions' European Commission's General Data Protection Regulation (GDPR) [25] enforces strict data handling requirements, highlighting the need for privacy-preserving techniques. FL aligns with these requirements by enabling robust model training without transferring sensitive data across entities or locations.

Pretraining significantly impacts the quality of computer vision models [7]. FL benefits from pre-trained models, leveraging global knowledge to improve accuracy even with distributed and heterogeneous data [6]. However, bias in existing data sets affects these models quality [14].

This study aims to evaluate DINOv2s' performance in FL settings and understand the implications of using pre-trained models in privacy-sensitive applications. Despite DINOv2s' claims to mitigate bias and enhance generalization, inherent biases in these models may lead to varied performance across domains. This analysis determines if pre-trained models can provide a competitive edge in decentralized training, leading to efficient and secure applications in critical domains. We used multiple data sets from different domains to measure DINOv2s' performance and generalization capabilities in FL, covering scenarios from industry, automotive and medical images.

2 RELATED WORK

Research in computer vision and FL has extensively investigated the impact of pretraining on various tasks. This section presents an overview of the scientific background of the experiments.

2.1 Supervised Segmentation with DINOv2

Semantic segmentation is a computer vision technique aimed at partitioning an image into multiple segments, each representing a distinct object class. Unlike object detection, semantic segmentation assigns a class label to every pixel in the image, enabling a precise understanding of its contents and spatial distribution [17]. DINOv2 [22], a self-supervised learning model in computer vision, uses Visual Transformers (ViTs) [11] as its core architecture. ViTs utilize self-attention mechanisms that allow DINOv2 to efficiently process a variety of visual information without the need for labeled data. This ability significantly enhances the model's adaptability and establishes DINOv2 as a versatile backbone for various vision tasks.

A key aspect of using DINOv2 in image segmentation is the use of "Ground DINO" and a header such as a linear classifier in this case. Ground DINO optimizes the integration process, while the

⟨KDD '24 FedKDD Workshop⟩, ⟨August 25th-29th, 2024⟩, ⟨Barcelona, Spain⟩

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

linear classifier interprets the CLS (Classification) tokens as input. These CLS tokens, crucial for capturing the contextual essence of the image, enable the linear classifier to effectively construct segmentation masks. This process is crucial for detailed feature analysis, including evaluating pixel intensity, color, texture similarities, and continuity. Such a strategic approach highlights the synergy between DINOv2's self-supervised learning capabilities and the supervised fine-tuning process, enabling the creation of accurate and precise segmentation tasks. In practical applications, the strategy involves adapting the pre-learned representations of DINOv2 to specific applications [5].

2.2 Federated Learning

Training models for Artificial Intelligence (AI) require large volumes of data. However, companies and individuals have legitimate reasons to maintain the confidentiality of their data, making data collection a considerable challenge. FL addresses these aspects, as it enables private and secure training of modern AI models [19]. Proposed by Google as a decentralized machine learning paradigm, FL enables collaborative and distributed training of AI-models. The core components of a basic FL setup typically encompass three components:

- **AI model:** A predefined AI model which is adjusted iteratively and collaboratively during the training process.
- **Aggregator:** Serving as the coordination hub. It selects contributing nodes for each training round and combines their model adjustments into an update on the global model before distributing to the nodes.
- **FL-Nodes:** Nodes engage in a predetermined number of training steps with their local data. They forward then their updates to the aggregator. Each node can independently use its local model to perform inferences with local data.

In FL, the model's training is not performed centrally. Through collaboration, participants iteratively refine a global model. Figure 1 shows the training rounds of horizontal FL [26] consisting of three steps:

(1) **Model distribution** (Step 1): The central entity selects a set of nodes and transmits the current model to each node.
(2) **Model training** (Step 2): First, the model is trained with their local data (Step 2a), and then the model updates are transmitted back to the central entity (Step 2b).
(3) **Model aggregation** (Step 3): The central entity merges the received local updates into a global model update. In FedAVG for example, averaging the collected model parameters from each node weighted by its data amount results in the new model [19].

Furthermore, each participant possesses a locally accessible and thus globally refined model, which enables independent inferences without reliance on a central entity or network connectivity. As no training data – potentially millions of high-resolution images – exits the nodes, FL directly tackles two challenges inherent in distributed learning: privacy [15] and communication efficiency [24].

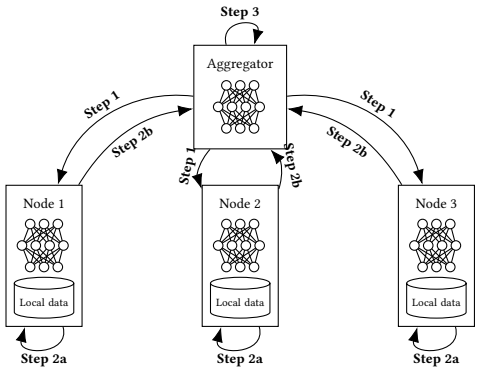

Figure 1: Typical training process of a horizontal Federated Learning system.

2.3 The benefits and challenges of pretraining

Pretraining has been recognized as highly effective in computer vision [23]. Models pre-trained on large-scale data sets serve as general-purpose feature extractors, significantly enhancing performance across various tasks without extensive fine-tuning [9]. This approach parallels the paradigm shift in Natural Language Processing (NLP) where task-agnostic pretrained representations have become the standard, achieving superior results compared to task-specific models [3]. Following this trend in NLP, the emergence of similar "foundation" models in computer vision is anticipated [2]. However, pretraining's and transfer learning's success in the image domain depends on the relevance of the source data set to the target data set, with optimal results when both domains closely align [20]. Despite these advantages, those models often perpetuate and amplify existing biases in training data, leading to unfair or wrong outcomes [4].

Recent studies indicate that pretraining can significantly alleviate accuracy drops caused by data heterogeneity in [21] environments. Pretraining stabilizes global aggregation in non-IID data scenarios, proving beneficial for FL model initialization [6]. However, while trained backbones offer marginal performance improvements in federated image segmentation, they are not indispensable, particularly for advanced tasks like medical imaging, where training from scratch might be more effective [12]. Conversely, for data sets with limited client images, pretraining is crucial for achieving state-of-the-art results, indicating significant benefits for small detection data sets [18].

3 EVALUATION ENVIRONMENT

With a the typically large number of nodes in a FL-network, evaluating such a system on independent devices is not feasible. A central system provided the environment with several simulated nodes. As the target of the experiments is to investigate the ability of DINOv2 to generalize over different domains in the FL-setting, simulating the FL-network is a feasible approach. Network factors – such as bandwidth limitations, latency, or other restrictions – have only an impact on the scalability of the network and not the accuracy.

The FL-environment was implemented in Python 3.9., and each node was deployed as a single docker container. The communication

between the aggregator and the nodes happened over the pub/sub mechanism provided by MQQT and Mosquitto in a docker container as a broker for the network. The central system contained an Intel Xeon E5-2695, 512 GByte RAM, and an NVIDIA A6000 with CUDA support.

3.1 DINO and FL Implementation

For pixel-wise segmentation, we employed a LinearClassifier as the segmentation head. This classifier includes a Conv2d layer to interpret the Dino's CLS token. The Conv2d layer conducts convolution operations to produce segmentation maps, assigning a class label to each pixel within the input images.

To evaluate the DINOv2 model's performance across various data sets, we adapted the segmentation head to the different used data sets by fine-tuning the model locally using the distributed FL approach. This adaptation process involved "freezing" the pre-trained DINOv2 model, enabling it to be fine-tuned under different training conditions. The model is pre-trained with the LVD-142M dataset. The experimental setup involved two primary targets:

- **Local:** In this scenario, the DINOv2 model, alongside the LinearClassifier, was fine-tuned on each data set individually. This approach aimed to assess the model's performance and robustness when trained in isolation on data from a single source.
- **Federated Learning:** Federated learning with configurations involving 2, 4, and 8 nodes to explore how distributed data affects the model's accuracy and robustness. Each node performed local updates on the model using its respective data subset, with these updates aggregated using the FedAVG algorithm to update the global model.

3.2 Data source

For the comprehensive evaluation of DINOv2s' performance in different domains and the FL environment, several data sets from diverse domains were utilized. Figure 2 shows example images taken from each data set.

(a) Diving (b) Medical (c) Industry

(d) Automotive (e) General

Figure 2: Example images from the used data sets

Prior to their use, each data set was translated from their original structure into the COCO image annotation format. This format consists of polygons surrounding particular areas in the image, and assigning a specific class to each pixel in the corresponding area. These annotations are used in the training process and enable a precise assessment of segmentation accuracy. Following domains were addressed during the experiments:

- **Diving:** The SUIM [13] data set is a comprehensive collection of underwater images comprising a wide variety of scenes such as coral reefs, wrecks, and diverse marine life.
- **Medical:** The ChestX-Det [16] data set comprises a large collection of chest X-ray images annotated with detailed information on various thoracic diseases.
- **Industry:** The VISION [1] data sets consist of images to replicate real-life industrial situations. The segmentations focus on outlining particular regions, like recognizing part boundaries or identifying surface defects needed for industrial quality control and automation.
- **Automotive:** The Cityscapes [8] data set was created for semantic and instance segmentation in urban environments. It consists of high-quality street images of humans, vehicles, and street furniture in city settings.
- **General:** The PASCAL Visual Object Classes (VOC) 2012 data set [10] is one of the most important resources used in evaluating algorithms for image segmentation in computer vision. It contains images of 20 object classes, like people, animals, vehicles, and everyday objects.

The characteristics of the five distinct data sets are detailed in table 1. It provides a comprehensive overview of each data set, including the number of images designated for training/testing and the number of distinct classes.

Data Set	#Classes	#Training	#Test
Diving(SUIM)	7	1525	110
Medical(ChestX-Det)	13	3578	553
Industry(VISION)	44	880	1014
Automotive(cityscapes)	7	2993	500
General(VOC2012)	20	17125	1500

Table 1: The size of the used data sets and the number of contained classes.

4 EVALUATION

Initially, DINOv2 was evaluated monolithically without FL and trained each on the entire data sets to establish a baseline. Subsequently, the data set underwent a split into several distinct segments (2, 4 and 8) in respect to the maximum number of FL-nodes, which could be simulated in the environment. The training duration varied: the monolithic systems underwent 20 epochs, while FL comprised 4 local epochs and 20 global rounds. For determining the accuracy, we applied the official test data sets from each data set.

We apply the in semantic segmentation commonly used metric Intersect over Union IoU to determine the quality of the predicted segmentation [17]. Let $n_{i,j}$ be the number of pixels of label i determined to belong to the label when there exist N_L different labels and let $t_i = \sum_j n_{i,j}$ be the total number of pixel belonging to label

i. Then IoU is defined as follow:

$$IoU = \frac{1}{N_L} \cdot \frac{\sum_i n_{i,i}}{t_i + \sum_j n_{j,i} - ni,i} \in [0, 1]$$

4.1 Impact of node count on DINOv2's performance

In Figure 3, the performance of a FL model across various application domains is illustrated using IoU metrics for different numbers of network nodes (2, 4, 8). The IoU values in the diving domain remain constant at values spanning from 7% to 18% across all node configurations. In the medical domain, IoU values exhibit slight fluctuations ranging from 0.99% to 1.99%, whereas in the automotive sector, a higher variability is observed with values spanning from 26.36% to 30.41%. The industry and general categories show minor to moderate variations in IoU, indicating a differential response of model performance to node count across these diverse application domains.

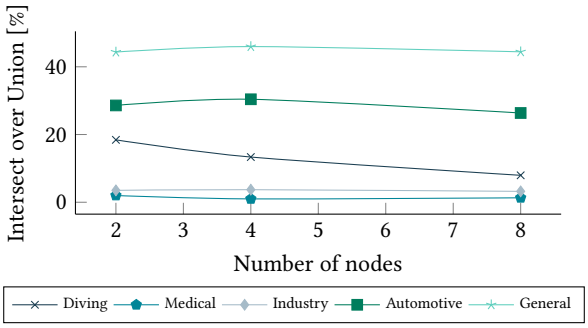

Figure 3: Intersect over Union for each data set in regard of a growing number of nodes

4.2 Impact of federalization of DINOv2

The comparative analysis of semantic segmentation performance using FL and monolithic (M) training approaches has yielded insightful results. The bar chart for IoU across all labels 4 indicates that the FL approach closely approximates the performance of the monolithic setup. In 4, the performance of monolithic models versus feature learning models across various application domains is presented. The monolithic model demonstrates higher performance in the categories of diving, automotive, and general, whereas the feature learning model excels in the medical sector. Notably, there are significant disparities in performance, with the monolithic model achieving substantially higher values in diving, automotive, and general domains, while showing relatively lower scores in medical and industrial categories.

In the diving category, the monolithic model achieves an IoU of 58.99 compared to 18.43 for the FL model. In the medical category, the FL model outperforms the monolithic model with an IoU of 1.99 versus 0.37. In the automotive category, the monolithic model shows significantly better performance with an IoU of 57.43 compared to 30.42 for the FL model. The results in the industrial category are closer, with the monolithic model achieving an IoU of 4.08 and the FL model an IoU of 3.51. In the general category, the monolithic model exhibits an IoU of 64.99, while the FL model achieves 45.98.

These results highlight the varying strengths and weaknesses of the two training approaches across different application domains. The monolithic model shows strong performance in most general categories, while the FL model performs better in the medical sector. This could be attributed to the different nature of data sources and the training methodologies employed by each approach.

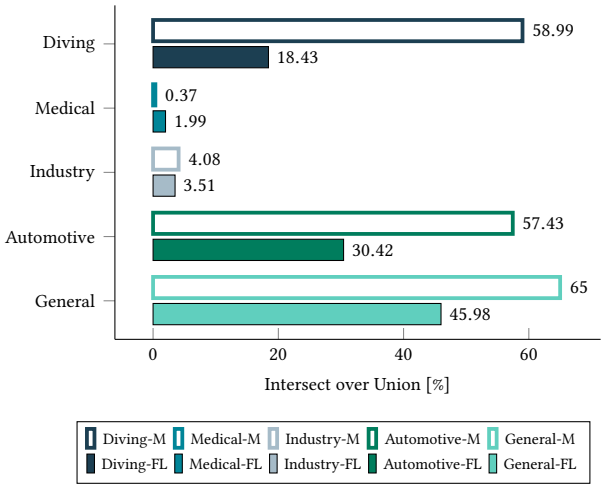

Figure 4: Monolithic Intersect over Union vs FL Intersect over Union

5 CONCLUSION

This study explored the performance and implications of using the DINOv2 pre-trained model within FL environments. Through a comprehensive evaluation across diverse data sets, we demonstrated that DINOv2 can function effectively under FL conditions and achieves comparable accuracy and performance to the monolithic setup, where all data is centralized and available. Our findings indicate that the model retains its ability to segment objects accurately, even when trained on decentralized and heterogeneous data, underscoring the potential of FL to maintain high performance without compromising data privacy.

The results show that DINOv2 was able to extract the classes well from the General(VOC2012) and Automotive(Cityscapes) data sets, which are part of its original training data (LVD-142M). It has also generalized to the extent that it can perform well on related data sets such as Diving (SUMI). However, DINOv2 cannot perform on data (in our evaluation with medical and industry data), which has no similarity to the data with which it was originally trained. While DINOv2 exhibits strong generalization capabilities within related domains, its performance significantly drops when applied to domains entirely dissimilar and containing unknown features.

This underscores the necessity of training with domain-specific data sets to cover that domain's characteristics. The other approach is to include a more diverse range of domains and their particular images in the data sets used for pre-training foundation models like DINOv2 to enhance the robustness and generalization capabilities of such pre-trained models.

6 OUTLOOK

Future work will focus on several key areas to enhance our understanding of DINOv2's performance in the FL setting. We will conduct a comprehensive hyperparameter search to optimize model configurations. Additionally, training DINOv2 from scratch in FL settings without a pre-trained model with its original LVD-142M data set will establish a performance baseline on the suitability of DINOv2 to be deployed in a federated setting. Extending the LVD-142M data set to include more diverse domains will enable a better generalization capability. Finally, training from scratch will create further insights into DINOv2's overall performance in diverse domains. These steps aim to improve model accuracy, generalization, and effectiveness in privacy-sensitive applications.

REFERENCES

[1] Haoping Bai, Shancong Mou, Tatiana Likhomanenko, Ramazan Gokberk Cinbis, Oncel Tuzel, Ping Huang, Jiulong Shan, Jianjun Shi, and Meng Cao. 2023. VISION Datasets: A Benchmark for Vision-based InduStrial InspectiON.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. https://doi.org/10.48550/arXiv.2108.07258

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

[4] Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models. https://doi.org/10.48550/arXiv.2310.17530

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[6] Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han-Wei Shen, and Wei-Lun Chao. 2023. On the Importance and Applicability of Pre-Training for Federated Learning. https://doi.org/10.48550/arXiv.2206.11488

[7] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Michael Carbin, and Zhangyang Wang. 2021. The Lottery Tickets Hypothesis for Supervised and Self-supervised Pre-training in Computer Vision Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16301–16311. https://doi.org/10.1109/CVPR46437.2021.01604

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[9] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: a deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, Beijing, China, I–647–I–655.

[10] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. 2021. Location-aware Single Image Reflection Removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4997–5006. https://doi.org/10.1109/ICCV48922.2021.00497

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929

[12] Chaoyang He, Alay Dilipbhai Shah, Zhenheng Tang, Di Fan1Adarshan Naiynar Sivashunmugam, Keerti Bhogaraju, Mita Shimpi, Li Shen, Xiaowen Chu, Mahdi Soltanolkotabi, and Salman Avestimehr. 2021. FedCV: A Federated Learning Framework for Diverse Computer Vision Tasks. https://doi.org/10.48550/arXiv.2111.11066

[13] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. 2020. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1769–1776. https://doi.org/10.1109/IROS45743.2020.9340821

[14] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. 2012. Undoing the Damage of Dataset Bias. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer, Berlin, Heidelberg, 158–171. https://doi.org/10.1007/978-3-642-33718-5_12

[15] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2023. A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection. *IEEE Transactions on Knowledge and Data Engineering* 35, 4 (April 2023), 3347–3366. https://doi.org/10.1109/TKDE.2021.3124599

[16] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. 2021. A Structure-Aware Relation Network for Thoracic Diseases Detection and Segmentation. *IEEE Transactions on Medical Imaging* 40, 8 (Aug. 2021), 2042–2052. https://doi.org/10.1109/TMI.2021.3070847

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

[18] Jiahuan Luo, Xueyang Wu, Yun Luo, Anbu Huang, Yunfeng Huang, Yang Liu, and Qiang Yang. 2021. Real-World Image Datasets for Federated Learning. https://doi.org/10.48550/arXiv.1910.11089

[19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 1273–1282.

[20] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. 2022. Factors of Influence for Transfer Learning Across Diverse Appearance Domains and Task Types. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (Dec. 2022), 9298–9314. https://doi.org/10.1109/TPAMI.2021.3129870

[21] John Nguyen, Jianyu Wang, Kshitiz Malik, Maziar Sanjabi, and Michael Rabbat. 2023. Where to Begin? On the Impact of Pre-Training and Initialization in Federated Learning. https://doi.org/10.48550/arXiv.2206.15387

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. https://doi.org/10.48550/arXiv.2304.07193

[23] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

[24] Abhishek Singh, Praneeth Vepakomma, Otkrist Gupta, and Ramesh Raskar. 2019. Detailed comparison of communication efficiency of split learning and federated learning. https://doi.org/10.48550/arXiv.1909.09145

[25] Paul Voigt and Axel von dem Bussche. 2017. The EU General Data Protection Regulation (GDPR): A Practical Guide. *Springer Nature eBook* (2017). https://doi.org/10.1007/978-3-319-57959-7

[26] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (March 2019), 1–19. https://doi.org/10.1145/3298981