June-July Exam session Projects List

Objectives

The goal of the project is to tackle one of the proposed topics in the field of Computer Vision, developing non-trivial solutions. Students should explore the problem from an original perspective, applying methods that go beyond conventional solutions and demonstrating critical thinking and problem-solving skills.

Lines of conduct

- **Student groups:** The project can be carried out by a group, each consisting of a maximum of 3 people. Projects can also be completed by individual students, but we suggest to work in team.
- **Notebook format:** The project must be implemented using a notebook (e.g., Google Colab, Kaggle) or with an IDE (e.g., VSCode, PyCharm). The code should be optimized to support GPU usage and run without any error. The delivered code must follow the structure outlined below:
  - *Imports*: all the needed packages (for the notebook format)
  - *Globals*: useful variables on the whole code
  - *Utils*: code support functions
  - *Data*: everything related to data management
  - *Network*: code to structure the neural network
  - *Train*: part containing the training cycle elements
  - *Evaluation*: tests needed for the trained network

  You can find a sample template at this link. Try to maintain as much as possible this conceptual structure.
- **Deep learning framework:** All projects **MUST** be done in Python via the Pytorch framework.
- **Project assignment:** You are required to choose a project through this Google Form. In this form, you will provide information about your team and the project, and include the link to the project's GitHub repository. In this repository you have to upload:
  - Code (or notebook) implementing the project
  - Dataset (or a link to it)
  - Project presentation
  - Detailed README to provide a quick overview of the project and instructions on how to run it
- **Project submission:** The project must be presented on one of the exam dates. It can be presented at a different time than the written exam. Both the written exam and the project MUST be completed within the academic year (i.e., between the June 2025 session and the March 2026 session).

  October and March session are reserved to "categories of students referred to in Article 40, paragraph 6, of the General Study Manifesto, and out-of-school students enrolled for the A.Y. 2024-2025 in the third year of a Bachelor's degree and in the second year of a Master's degree".
- **Plagiarism:** Any attempt to plagiarize, whether by copying other students' work, directly replicating code from online resources, or submitting content highly retrived from generative AI models, will be strictly penalized. This course values originality and personal effort; therefore, students must submit independently developed solutions. On the other hand, it is acceptable to consult external resources for inspiration or guidance.

Project 1: Neuron Selectivity for Efficient Monocular Depth Estimation

Abstract: Monocular depth estimation (MDE) is a crucial computer vision task in many applications, such as augmented reality, robotics, and autonomous driving. However, MDE neural networks are often considered as black-boxes, related to the difficulty in explaining their decision-making process. Understanding how a depth map has been predicted is essential to increase reliability on MDE models and facilitate their adoption in real-world scenarios. Recently, neuron selectivity allows us to interpret the model behaviour by assigning to each neuron a specific depth range. This means there are neurons responsible for near depths and neurons responsible for far depths. While this strategy has been successfully applied to large-scale models, its impact on lightweight models optimized for mobile and low-resource environments remains unclear.

Dataset: NYU Depth V2

Task: Explainability is increasingly required in real-world applications to understand the behaviour of models and how their predictions are computed. This project proposes to explore the impact of neuron selectivity [1] on lightweight MDE models to verify if this explainability strategy can be also effective on efficient neural networks. Examples of lightweight MDE models are [3,4], but you can work with any efficient architecture for MDE.

Main objectives:

- *Baseline train and evaluation*: Train the lightweight architecture without any neuron selectivity constraints and evaluate its performance. This will be the baseline to compare against the proposed method.
- *Selectivity-based train*: Implement and apply the neuron selectivity training strategy to the lightweight model. This involves modifying the training process to encourage neurons to specialize in specific depth ranges, enhancing interpretability while preserving performance.
- *Performance-selectivity trade-off evaluation*: Measure performance metrics such as depth estimation error and neuron selectivity to assess the trade-off between interpretability and performance, determining the feasibility of applying this strategy to lightweight models.

References:

1. You, Z., Tsai, Y.-H., Chiu, W.-C., and Li, G. (2021). Towards Interpretable Deep Networks for Monocular Depth Estimation. arXiv.
2. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Yan, S. (2022). MetaFormer Is Actually What You Need for Vision. arXiv.
3. C. Schiavella, L. Cirillo, L. Papa, P. Russo, and I. Amerini, (2023). Optimize vision transformer architecture via efficient attention modules: a study on the monocular depth estimation task. In: International Conference on Image Analysis and Processing, Cham: Springer Nature Switzerland, pp. 383-394.
4. Papa, L., Russo, P., and Amerini, I. (2023). METER: A Mobile Vision Transformer Architecture for Monocular Depth Estimation. IEEE Transactions on Circuits and Systems for Video Technology, 33(10), 5882–5893.

Project 2: Improving Robustness of Deepfake Detectors through Gradient Regularization

Abstract: Recent generative methods have shown strong capabilities in producing high-quality deepfakes, posing a security threat in various domains, such as social networks and online platforms. Consequently, the demand for deepfake detectors has increased, with an emphasis on high accuracy in identifying deepfakes generated by different methods, as well as ensuring robustness against adversarial attacks to enhance their security. Building on that, this project proposes an approach to improve the robustness of deepfake detectors with a gradient regularization technique. This technique has shown promising results in improving model generalization, leveraging an approximation of the Hessian matrix in the gradient calculation to separate the feature space of the trained model by incorporating shallow feature statistics.

Dataset: DFFD: Diverse Fake Face Dataset [3] (you can also choose the FaceForensics++ dataset, but you will need to extract frames from videos)

Task: Increase the adversarial robustness of deepfake detectors exploiting the gradient regularization technique described in [1]. This method has demonstrated promising improvements in the generalization performance of deepfake detectors. As such, investigating its potential to enhance the security of these models against adversarial attacks [4] will be valuable for improving their robustness and reliability in real-world scenarios. An example of model that can be used for deepfake detection is the EfficientNetb0 [2], but you can choose any architecture you consider suitable for this task.

Main objectives:

- *Baseline train and evaluation*: Train or finetune a deepfake detector and evaluate its performance without gradient regularization.
- *Adversarially attack the baseline*: Use some adversarial attacks to evaluate the robustness of the baseline.
- *Baseline with gradient regularization train and evaluation*: Do the previous two steps on the model trained with the gradient regularization technique
- *Compare the performance of the two deepfake detectors*: Make an analysis and a comparison of the obtained results, trying to explain why and how they have been reached.

References:

1. W. Guan, W. Wang, J. Dong and B. Peng, (2024). Improving Generalization of Deepfake Detectors by Imposing Gradient Regularization, In IEEE Transactions on Information Forensics and Security, vol. 19, pp. 5345-5356.
2. M. Tan and Q. Le, (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proc. Int. Conf. Mach. Learn., pp. 6105–6114.
3. On the Detection of Digital Face Manipulation Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, Anil Jain, (2020), In: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, Jun. 2020
4. Abbasi, M., Váz, P., Silva, J. and Martins, P. (2025). Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. Applied Sciences, 15(3), 1225.

Project 3: Mechanistic Interpretability for Vision Models Optimization

Abstract: In recent years, vision transformers (ViTs) have shown very high performance on many vision tasks, such as image classification, object detection, and monocular depth estimation. However, their high computational cost makes them unsuitable for edge devices with limited hardware capabilities. Upon this, researchers are trying to find the optimal trade-off between the model performance and inference time. Therefore, this project proposes a way to reduce the inference time of ViTs with mechanistic interpretability. This very promising research field reverse-engineers a model to understand its decision-making process. In particular, the Automated Circuit Discovery (ACDC) technique tries to eliminate those edges from the model computational graph that are not important for the output computation. As a consequence, one can expect that a model that doesn't perform some computational operations is also more efficient in terms of inference time.

Dataset: Tiny-ImageNet

Task: Reduce the inference time of a ViT model by adopting the ACDC mechanistic interpretability technique [1] to remove those edges that are more irrelevant for the output computation. Original strategies to remove such edges from the model prediction computation should be addressed. A trade-off analysis between the accuracy and inference time performance must be carried out to verify how the ACDC method influences the accuracy metrics.

Main objectives:

- *Baseline train and evaluation*: Train the baseline ViT and evaluate its accuracy and inference time performance.
- *Model optimization through ACDC*: Identify irrelevant edges for the output prediction in the computational graph and exclude them to check if the inference time is reduced.
- *Optimized model training*: Once the irrelevant edges have been excluded by the model computational graph, train the new model without such edges.
- *Results comparative analysis*: Make a comparative analysis with the baseline on the obtained results. Focus on the trade-off between accuracy metrics and inference time, trying to explain in which scenario it is preferable to have a reduced inference time at the expense of an accuracy loss. You can take inspiration from [3] to make such an analysis.

References:

1. A. Conmy et al. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023)
2. A. Syed, C. Rager and A.Conmy, (2024). Attribution Patching Outperforms Automated Circuit Discovery, BlackboxNLP 2024.
3. C. Schiavella, L. Cirillo, L. Papa, P. Russo, and I. Amerini, (2023). Optimize vision transformer architecture via efficient attention modules: a study on the monocular depth estimation task. In: International Conference on Image Analysis and Processing, Cham: Springer Nature Switzerland, pp. 383-394.

Project 4: Car Plate Recognition and Reconstruction with Deep Learning

Abstract: Automatic car plate recognition is a crucial task in the field of computer vision with wide-ranging applications in intelligent transportation systems, traffic monitoring, law enforcement, and access control. The goal is to accurately recognize and reconstruct vehicle license plates from images or video streams, often captured under challenging real-world conditions such as varying lighting, occlusions, motion blur, and diverse plate formats. Deep learning models, particularly convolutional neural networks, have significantly advanced the performance and reliability of car plate recognition. This project aims to explore and implement deep learning-based approaches for license plate recognition, emphasizing practical challenges and the impact of robust solutions in modern urban infrastructure and mobility management.

Dataset: **Dataset:** CCPD [2]

Task: The objective of this project is to design and implement a deep learning-based system for license plate recognition, following the methodology outlined in [1]. The proposed solution is structured as a two-stage pipeline, leveraging the strengths of different neural network architectures to address the distinct subtasks involved in the recognition process. In the first stage, a YOLOv5 model is employed for license plate detection, allowing for fast and accurate localization of the plate region within vehicle images, even under challenging environmental conditions. In the second stage, the cropped plate region is passed to a specialized recognition model based on the PDLPR architecture. This model is responsible for decoding the sequence of alphanumeric characters on the plate, effectively treating the task as a sequence prediction problem. The integration of these two components aims to deliver a robust and efficient system for plates recognition and reconstruction suitable for deployment in real-world scenarios.

Main objectives:

- *Baseline implementation, training and evaluation:* Implement a simple baseline, train and evaluate it with the metrics used in [1].
- *YOLOv5 and PDLPR model implementation and evaluation:* Implement the proposed model in [1], composed by the YOLOv5 and PDLPR models, and evaluate it.
- *Comparison with the baseline:* Compare the performance of the proposed model with the baseline, underlining why the proposed model is working better or not on recognizing and recostructing the car plates.

References:

1. Tao, L., Hong, S., Lin, Y., Chen, Y., He, P. and Tie, Z. (2024). A Real-Time License Plate Detection and Recognition Model in Unconstrained Scenarios. Sensors, 24(9), 2791
2. Xu, Z.; Yang, W.; Meng, A.; Lu, N.; Huang, H.; Ying, C.; Huang, L. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
3. R. K. Prajapati, Y. Bhardwaj, R. K. Jain and D. Kamal Kant Hiran, "A Review Paper on Automatic Number Plate Recognition using Machine Learning : An In-Depth Analysis of Machine Learning Techniques in Automatic Number Plate Recognition: Opportunities and Limitations," 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, 2023, pp. 527-532

Project 5: Lightweight Convolutional Occupancy Networks for Efficient Virtual Scene Generation

Abstract: The generation of virtual scenes is a critical challenge in robotics, augmented reality, and simulation, where accurate and computationally efficient 3D representations are essential. However, existing solutions often involve a trade-off between reconstruction quality and execution speed, limiting their applicability in real-time scenarios. This project investigates using neural networks to generate 3D scenes, focusing on optimizing the model to reduce inference time while maintaining reasonable reconstruction fidelity. Various optimization strategies will be explored to achieve this, assessing their impact on efficiency and visual quality. The study will provide a comparative analysis against existing approaches, contributing to developing lightweight and effective models for virtual scene synthesis.

Dataset: Synthetic-Rooms (provided)

Task: This project explores using a Lightweight Convolutional Occupancy Network to generate virtual scenes composed of objects from the ShapeNet dataset. The primary objective is to optimize the network to achieve low inference time while preserving the quality of the reconstructed scene. Students will refine the network architecture, adjusting key components to improve computational efficiency without compromising accuracy.

Main objectives:

- *Training and Fine-Tuning*: Train the Lightweight Convolutional Occupancy Network on the Synthetic-Rooms dataset, ensuring accurate reconstruction of structured scenes. Fine-tune network parameters and apply geometrical modifications to optimize the balance between speed and quality.
- *Model Optimization*: Minimize inference time while preserving reconstruction quality by refining the architecture and applying optimization techniques such as model pruning, quantization, and geometrical simplifications.
- *Evaluation*: Compare the optimized model against state-of-the-art approaches using key metrics, including inference time, reconstruction accuracy, and visual fidelity, to assess trade-offs introduced by optimizations.
- *Trade-off analysis*: Analyze the impact of different optimization strategies on computational efficiency and reconstruction accuracy, evaluating whether the optimized model can outperform existing solutions in real-time scene generation.

References:

1 Lionar, S., Emtsev, D., Svilarkovic, D., and Peng, S. (2020). Dynamic Plane Convolutional Occupancy Networks. arXiv.

2 C. M. Tonti, L. Papa and I. Amerini, "Lightweight 3-D Convolutional Occupancy Networks for Virtual Object Reconstruction," in IEEE Computer Graphics and Applications, vol. 44, no. 2, pp. 23-36, March-April 2024, doi: 10.1109/MCG.2024.3359822.

3 Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., and Geiger, A. (2020). Convolutional Occupancy Networks. arXiv.

Project 6: Truth in Motion: Depth and Flow Enhanced DeepFake Detection

Abstract: Powerful tools and software for creating and processing multimedia content have been made available by recent advances in visual media technology. In particular, AI-driven methods made the production of Deepfakes videos simpler than ever. These fake and manioulated videos are dangerous because they can be used to alter public opinion or harm reputations. Considering these dangers, creating trustworthy techniques to identify Deepfakes is crucial to preserving data integrity. For this reason, accurately separating fake videos from authentic ones while maintaining a method that maximizes computational resources is the main goal of this work.

Dataset: FaceForensics++ (Real, Face2Face, Deepfakes, FaceSwap)

Task: This project's main goal is to create a reliable and effective Deepfake detection system that can differentiate real videos from those that have been altered. Starting with RGB frames, the project investigates a number of feature extraction strategies, such as depth-based and optical flow approaches. A Transformer-based model will be used in place of a traditional CNN to improve detection accuracy. This will allow for more efficient processing of optical flow frames and a more accurate evaluation of their effect on overall performance. Finally, in order to achieve a correct balance between detection efficacy and computational efficiency, the study will also look at post-processing methods and model compression strategies, including quantization, pruning, and distillation.

Main objectives:

- *Pre-processing*: Correctly download and split the FaceForensics++ dataset, developing a precise face detection mechanism that is able to extract face outlines.
- *Feature extraction generation*: Explore various feature extraction techniques, such as optical flow and depth-based methods, to enhance the detection process.
- *Training*: Develop and fine-tune a Transformer-based model to learn meaningful features from the extracted data, enabling reliable real/fake classification.
- *Model compression*: Investigate compression techniques such as quantization, pruning, and distillation to reduce computational costs while maintaining strong classification performance.

References:

1. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. CoRR, abs/1901.08971, 2019
2. I. Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1205-1207, doi: 10.1109/ICCVW.2019.00152.
3. Nassif AB, Nasir Q, Talib MA, Gouda OM. Improved Optical Flow Estimation Method for Deepfake Videos. 2022 Mar 24;22(7):2500. doi: 10.3390/s22072500. PMID: 35408114; PMCID: PMC9002804.

Project 7: Efficient Anomaly Detection in Industrial Images using Transformers with Dynamic Tanh

Abstract: In industrial environments, detecting anomalies in visual data is crucial for maintaining high standards of quality and operational safety. Traditional image analysis methods often face limitations when dealing with complex, high-dimensional data. Recent advancements in computer vision, particularly in transformer-based architectures such as Vision Transformers (ViTs), have shown great potential in capturing rich spatial features from images. At the same time, new techniques like Dynamic Tanh (DyT) offer promising solutions to improve the computational efficiency of these models. The combination of powerful feature extraction and efficient processing represents a significant opportunity for advancing anomaly detection systems in industrial applications.

Dataset: BTAD, MVTec Anomaly Detection Dataset

Task: The aim of this project is to explore and implement an advanced approach to anomaly detection in industrial images by combining two cutting-edge techniques in computer vision: Vision Transformers (ViTs) for feature extraction and Dynamic Tanh (DyT) for improving transformer model efficiency. This proposal leverages the strengths of ViTs in capturing spatial dependencies alongside the efficiency enhancements offered by DyT as a replacement for traditional normalization layers. Students will focus on both anomaly detection and localization tasks, applying the DyT method to improve network speed and overall model performance on industrial datasets. The project will involve designing and testing an end-to-end system that integrates ViTs for image analysis with DyT-based efficient transformers. Additionally, students will evaluate the efficiency benefits of DyT by comparing runtime and performance metrics against models using traditional normalization layers.

Main objectives:

- *Implement DyT for Improved Efficiency*: Replace traditional normalization layers in transformer models with the Dynamic Tanh method to assess improvements in model efficiency
- *Evaluate Performance*: Evaluate the performance of the model in anomaly detection and localization tasks using appropriate metrics.
- *Baselines comparison*: Benchmark the proposed solution against existing and not efficient transformer models, to highlight the performance advantages of using DyT in Vision Transformers.

References:

1. Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., & Foresti, G. L. (2021, June). VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization. 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE), 01–06. doi:10.1109/isie45552.2021.9576231
2. Zhu, J., Chen, X., He, K., LeCun, Y., & Liu, Z. (2025). Transformers without Normalization. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2503.10622
3. Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., & Jin, Y. (2024). Deep Industrial Image Anomaly Detection: A Survey. Machine Intelligence Research, 21(1), 104–135. doi:10.1007/s11633-023-1459-z

Project 8: Efficient Computer Vision Models for Silkworm Feeding Prediction and Habitat Analysis

Abstract: Efficient silkworm rearing is essential for sustainable silk production, relying heavily on accurate monitoring of feeding conditions. Traditional methods of observation are labor-intensive and subject to human error, highlighting the need for automated solutions. Advances in computer vision, particularly lightweight convolutional neural networks, offer promising approaches for interpreting complex visual scenes such as silkworm rearing beds. Differentiating between silkworms, mulberry leaves, and background areas is a critical step in automating feeding management and optimizing farm operations. The increasing availability of visual data from rearing environments further supports the development of robust, efficient monitoring systems tailored to the needs of modern silk production.

Dataset: Silkwork rearing data (provided)

Task: The project focuses on binary classification, where the goal is to implement lightweight neural network architectures to determine whether the silkworms need feeding. Another objective involves unsupervised segmentation techniques designed to automatically distinguish and separate the three key elements present in the images: silkworms, mulberry leaves, and background. By applying non-supervised methods, the project seeks to extract more detailed insights from the rearing environment without requiring pixel-wise annotations for training. Both tasks will be complemented by data augmentation strategies, improving the models' ability to generalize across different conditions and making them more robust to variations in environmental factors.

Main objectives:

- *Rearing Classification*: Use lightweight architectures for binary classification (feeding vs. no feeding).
- *Unsupervised Segmentation of Rearing Beds*: Implement unsupervised methods to segment silkworms, mulberry leaves, and backgrounds without ground truth.
- *Performance Evaluation*: Assess classification models and segmentation outputs using quantitative metrics and qualitative analysis to determine their effectiveness in real-world conditions.

References:

1. Zhao, M., Luo, Y., and Ouyang, Y. (2024). RepNeXt: A Fast Multi-Scale CNN using Structural Reparameterization. arXiv.
2. Tan, M., and Le, Q. V. (2021). EfficientNetV2: Smaller Models and Faster Training. arXiv.
3. Mehta, S., and Rastegari, M. (2022). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. arXiv.
4. Rossetti, S., Samà, N., and Pirri, F. (2023). Removing supervision in semantic segmentation with local-global matching and area balancing. arXiv.
5. Niu, D., Wang, X., Han, X., Lian, L., Herzig, R., and Darrell, T. (2023). Unsupervised Universal Image Segmentation. arXiv.

Project 9: Advanced Out-of-Distribution Detection for Multi-Class Classification

Abstract: Detecting out-of-distribution (OOD) inputs is a fundamental challenge in building reliable deep learning systems, particularly for multi-class classification tasks. In real-world scenarios, classifiers often encounter data that differ significantly from their training distributions, leading to unpredictable and potentially unsafe outcomes. Recent advancements, including energy or gradient based approaches, offer new strategies for distinguishing between in-distribution and OOD samples. High-dimensional datasets with their complex visual features, provide a realistic testing ground for developing and evaluating improved OOD detection methods. Ensuring accurate separation between known and unknown inputs is increasingly critical for deploying models in practical applications.

Dataset: Food-101,SVHN

Task: The first step is to train a deep neural network for multi-class classification using the Food-101 dataset as the in-distribution data. Once the classifier is established, one or more OOD detection frameworks should be implemented in order to effectively distinguish between in-distribution and OOD inputs, with the Street View House Numbers (SVHN) dataset serving as the primary source of OOD examples. Additionally, it is possible to expand the experimental setup by including other OOD datasets to test the capabilities of their detection methods. The goal is to assess the model's ability to distinguish between in and out-of-distribution samples, and to evaluate its performance using appropriate metrics.

Main objectives:

- *Define a multi class classification model*: Establish a network architecture suitable for the problem.
- *Integrate OOD detection*: Implement an OOD detection method to discriminate between normal and anomalous data.
- *Evaluate the method's performance*: Analysing OOD detection capability metrics to determine the validity and effectiveness of the proposed solution.

References:

1. Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based Out-of-distribution Detection. NeurIPS 2020.
2. Sharifi, S. et al. (2024). Gradient-Regularized Out-of-Distribution Detection. arXiv [Cs.CV].
3. Tang, K., et al. (2024, June). CORES: Convolutional Response-based Score for Out-of-distribution Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10916–10925.

Project 10: Transformer-based Satellite Image and Segmentation Generation for Ground-to-Aerial Image Matching

Abstract: Ground-to-Aerial image matching is the problem of associating a query ground-view with the corresponding satellite image. The aim of this project is to create a reliable framework for this task by empowering state-of-the-art Transformers to generate synthetic images, in conjunction with older VGG-16 networks to efficiently extract relevant features. The ground-to-aerial image matching is performed in two steps: a Transformer takes in input the query ground-view, and generates the segmented and natural satellite images that better represent the original ones. Then, the ground-view along with the synthetic aerial image and its segmentation are passed to a neural network, which determines which satellite image is most likely associated with the query.

Dataset: CVUSA

Task: The project is split into two phases. The first phase involves fine-tuning a pre-trained Transformer model for image generation, aiming to synthesize aerial images and segmentation maps that resemble satellite images paired with a query ground view. The CVUSA dataset will be used, which provides ground-view images, satellite perspectives, and aerial images with segmentation labels. In the second phase, the focus is on building a neural network for feature extraction. The network includes two subnetworks: JointFeatureLearningNet and FeatureFusionNet. The former processes four input images (ground view, synthetic aerial, segmented synthetic aerial, and candidate satellite) to learn a joint representation that aligns the ground and candidate satellite views, with the segmented aerial image focusing the model on important image regions. The FeatureFusionNet combines representations from VGG16 networks for each input image. The VGGs for synthetic and candidate satellite images share weights, and their outputs are concatenated and passed through a Feed Forward Network (FFN). The FFNs are trained from scratch on the CVUSA dataset, while the VGGs are fine-tuned. The model uses triplet loss with a weighted soft margin to compare the generated representations and identify the correct candidate satellite image for the query ground view.

Main objectives:

- *Network development*: Fine-tune a pre-trained Transformer model for image generation to synthesize aerial images. Develop and implement two key subnetworks: JointFeatureLearningNet and FeatureFusionNet. This will involve extracting features from the ground view, synthetic aerial view, segmented aerial view, and candidate satellite images.
- *Image Matching*:Fine-tune pre-trained VGG16 networks, ensuring optimal performance for both image representation and feature extraction. Initialize and train FFN from scratch to learn appropriate feature combinations for accurate ground-to-aerial image matching.
- *Evaluate Performance on CVUSA Dataset*:Measure the performance of the model by evaluating the accuracy of the generated aerial images and segmentation maps, as well as the quality of the image matching for the ground-to-aerial image query task.

References:

1. Regmi, K., & Shah, M. (2019). Bridging the Domain Gap for Ground-to-Aerial Image Matching. arXiv.
2. F. Pro, N. Dionelis, L. Maiano, B. L. Saux and I. Amerini, "A Semantic Segmentation-Guided Approach for Ground-to-Aerial Image Matching," IGARSS 2024 - Athens, Greece, 2024, pp. 2630-2635
3. Mule, E., Pannacci, M., Goudarzi, A., Pro, F., Papa, L., Maiano, L., and Amerini, I. (2025). Enhancing Ground-to-Aerial Image Matching for Visual Misinformation Detection Using Semantic Segmentation. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops (pp. 795-803).

Project 11: Uncertainty-Aware Road Obstacle Identification

Abstract: Reliable road obstacle identification is a critical requirement for the safe operation of autonomous driving systems. Traditional object detection methods often struggle to recognize unexpected or unknown obstacles, as they are typically limited to predefined categories. The ability to detect obstacles beyond known classes, particularly in dynamic and complex environments, is essential for the safety of autonomous vehicles. Recent advancements in semantic segmentation, anomaly detection, and uncertainty quantification offer new avenues to improve detection accuracy and reliability, enabling systems to recognize both known and unknown road obstacles. Such uncertainty-aware methods provide formal statistical guarantees on the reliability of predictions, a crucial aspect for ensuring safe and robust decision-making in real-world driving conditions.

Dataset: Cityscapes, LostAndFound, Fishyscapes

Task: The aim of this project is to develop a general, model-agnostic framework for road obstacle identification, starting from the outputs of any semantic segmentation network. The system will focus on anomaly-aware semantic segmentation to detect obstacles outside the predefined classes. This will allow for the identification of unknown obstacles as part of the segmentation output. To ensure that each identification is accompanied by a reliable measure of confidence, the framework will integrate uncertainty quantification through Conformal Prediction methods. By combining these components, the system will not only recognize potential obstacles but also provide formal statistical guarantees regarding the reliability of its predictions.

Main objectives:

- *Anomaly-Aware Obstacle Segmentation:* Integrate into a semantic segmentation model techniques to detect obstacles that fall outside known classes.
- *Statistical Uncertainty Quantification:* obtain semantic segmentation outputs and obstacle proposals guarantees on detection reliability.
- *Comprehensive Evaluation:* Benchmark the system using both detection performance metrics and uncertainty metrics.

References:

1. Noguchi, C., Ohgushi, T., & Yamanaka, M. (2024). Road Obstacle Detection based on Unknown Objectness Scores. arXiv [Cs.CV].
2. Mossina, L., Dalmau, J., & Andéol, L. (2024). Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty. arXiv [Cs.CV].
3. Angelopoulos, A. N., & Bates, S. (2022). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv [Cs.LG].

Project 12: Grounded Multimodal Named Entity Recognition

Abstract: In recent years, Multimodal Named Entity Recognition (MNER) has emerged as an important task in multimodal information extraction and multimodal deepfake detection, especially when applied to social media content. MNER aims to extract named entities and corresponding categories from image-text pairs sourced from social media. However, first approaches to MNER only aim to extract the entity-type pairs in text. Grounded Multimodal Named Entity Recognition (GMNER) addresses this issue: given a text-image social post, GMNER aims to identify the named entities in text, their entity types, and their bounding box groundings in the image. The aim of the project is to experiment with different ways to fuse textual information with visual information.

Dataset: GMNER Dataset (Twitter10000 v2.0)

Task: The task of this project is to implement a Grounded Multimodal Named Entity Recognition model that given the text of a tweet and the image of the tweet is able to identify the Named Entity in the text and identify the entity in the image with a bounding box. Each token in the sentence should be tagged following the IOB2 tagging scheme. The project should correctly implement the architecture proposed by Yu et al. and evaluate its performance by means of Recall, Precision and F1-score. Later, the aim is to reduce the model size using one or more model compression techniques, to optimize the training/inference time while keeping the same performance of the original model.

Main objectives:

- *Dataset download:* Download the images here and the preprocessed texts here;
- *Model implementation:* Train the GMNER model introduced by Yu et al. and evaluate its performance and training/inference time.
- Ablation study: Evaluate the performance of the model when considering only text vs text and image.
- *Model compression:* Use any kind of model compression technique to reduce the model size, while keeping the performance of the compressed model as good as the original model.

References:

- Yu, Jianfei, et al. "Grounded multimodal named entity recognition on social media." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.
- Moon, Seungwhan, Leonardo Neves, and Vitor Carvalho. "Multimodal named entity recognition for short social media posts." arXiv preprint arXiv:1802.07862 (2018).;
- Wang, Dongsheng, et al. "2M-NER: contrastive learning for multilingual and multimodal NER with language and modal fusion." Applied Intelligence 54.8 (2024): 6252-6268.
- Yu, Jianfei, et al. "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer." Association for Computational Linguistics, 2020.