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: Synthetic Dental X-Ray Generation and Segmentation Analysis

Abstract: Medical image analysis faces challenges like data scarcity, privacy restrictions, and ethical concerns. In dental radiology, obtaining diverse ortopanoramic X-ray (OPT) datasets is particularly difficult due to patient data sensitivity. This project trains a Generative Adversarial Network (GAN) to synthesize realistic OPT images and evaluates their effectiveness by training a CNN-based segmentation model on a public dental dataset. The final phase tests the segmentation model on both real and synthetic images to assess performance gaps, biases, and the usability of GAN-generated data in medical imaging.

Dataset: Public dental X-ray segmentation dataset (Teeth Segmentation)

Task: This project has two primary objectives: (1) developing a GAN model to generate synthetic ortopanoramic dental X-ray images, and (2) training and evaluating a YOLO-based segmentation model on real and synthetic datasets. The main focus is to assess how well a segmentation model trained on real data generalizes to synthetic images and to identify its potential limitations. The students will experiment with different GAN architectures and training strategies to improve the realism of the generated images. They will also analyze segmentation performance differences when applying CNN models to real vs. synthetic X-rays.

Main objectives:

- *Synthetic Data Generation:*
  - Train and fine-tune a GAN (e.g., StyleGAN, Pix2Pix, or CycleGAN) to generate realistic ortopanoramic X-ray images.
  - Apply image post-processing techniques to enhance quality and realism.
- *Teeth Segmentation:*
  - Train a CNN-based segmentation model (YOLO, U-Net, etc.) on a public dental X-ray dataset.
  - Fine-tune the model to correctly identify and segment teeth structures.
- *Evaluation and Generalization Study:*
  - Test the trained segmentation model on the synthetic dataset and compare its performance with real X-ray images.
  - Analyze whether segmentation accuracy is affected by synthetic data quality.
- *Ablation Study:*
  - Evaluate different GAN architectures to determine which produces the most realistic dental X-rays.
  - Explore domain adaptation techniques to bridge the gap between real and synthetic images.
  - Assess segmentation performance differences in various augmentation scenarios (e.g., training on mixed real + synthetic data).

References:

- S. Yang, K.-D. Kim, E. Ariji, N. Takata, and Y. Kise, "Evaluating the performance of generative adversarial network-synthesized periapical images in classifying C-shaped root canals," Scientific Reports, vol. 13, Oct. 2023, doi: 10.1038/s41598-023-45290-1.
- R. Varghese and S. M., "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024, pp. 1–6. doi: 10.1109/ADICS58448.2024.10533619.
- J. Chen et al., TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

Project 2: Out of Distribution Convolutional Response for Monocular Depth Estimation

Abstract: Recent advances in out-of-distribution (OOD) detection have introduced novel scoring methods based on the analysis of convolutional responses in deep neural networks. In particular, the CORES (Convolutional Response-based Score) method [1] exploits the magnitude and frequency of extreme activations in convolutional kernels to discriminate between in-distribution and OOD inputs without requiring access to the original training data. This project aims to explore the applicability of the CORES methodology within the context of Monocular Depth Estimation (MDE), a dense prediction task typically addressed through fully convolutional encoder–decoder architectures. This project aims to enable a clearer and more interpretable analysis of convolutional responses and their potential as OOD indicators.

Dataset: NYU Depth v2 (indoor scenes), KITTI (outdoor scenes)

Task: The aim of this project is to integrate the CORES scoring mechanism into a convolutional-base MDE network. The chosen network will be trained on an in-distribution dataset (e.g., NYU Depth v2) and subsequently evaluated on both in-distribution and out-of-distribution samples. The study will assess whether convolutional response patterns can serve as reliable OOD indicators in a dense prediction setting, using metrics such as AUROC and FPR95.

Main objectives:

- *Dataset preparation:* Preprocess and split the NYU Depth v2 dataset for training and validation. Collect OOD samples (e.g., KITTI images resized to match the input resolution). As an alternative, consider using a pretrained generative model to synthesize OOD samples.
- *Model implementation:* Implement and train a lightweight depth estimation model. Some recommended architectures are listed in references [2,3], but other models can also be considered. Training and comparing two or more models is encouraged.
- *CORES integration:* Integrate the convolutional response scoring method to extract layer-wise response statistics and compute the OOD score.
- *Evaluation:* Assess OOD detection performance (AUROC, FPR95) and depth estimation accuracy (RMSE, Abs_{Rel}, δ_1, δ_2, δ_3) on both ID and OOD datasets.
- *Ablation study:* Perform a comparative analysis across different layers and convolutional architectures, and evaluate the impact of model depth on the observed response patterns.

References:

1. Tang, Keke, et al. "CORES: Convolutional Response-based Score for Out-of-distribution Detection." CVPR 2024.
2. Wofk, Diana, et al. "FastDepth: Fast Monocular Depth Estimation on Embedded Systems." ICRA 2019.
3. Papa, L., Russo, P., & Amerini, I. "METER: A Mobile Vision Transformer Architecture for Monocular Depth Estimation". IEEE Transactions on Circuits and Systems for Video Technology, 2023.

Project 3: 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 4: Efficent Training Process for GANs

Abstract: Generative Adversarial Networks (GANs) are one of the most popular topics in Deep Learning. They are types of Neural Networks used for Unsupervised learning. GANs consist of two distinct models: a Generator *G(x)* and a Discriminator *D(x)*, which are trained simultaneously in a competitive learning framework. The Generator aims to produce synthetic data that closely resembles the real data from the training set, to deceive the Discriminator. In contrast, the Discriminator is tasked with distinguishing between real and artificially generated data, attempting not to be misled. Through this adversarial process, both networks iteratively improve, enabling the system to learn and generate complex data structures such as audio, video, or image files. One of the main problems of this architecture is the unstable training due to the competition between *G(x)* and *D(x)*, and the large amount of data they need.

Dataset: ImageNet, CIFAR, and Stacked MNIST datasets

Task: The primary task of this project is to investigate techniques to make the training of GANs more efficient and avoid stability problems. This involves implementing a baseline on which then apply some techniques to increase its stability in the training phase. As a baseline, you can start from a simple GAN implementation or from a more recent version, R3GAN, that already has some efficient techniques applied. Another part that could be explored in the project is to make the model able to train with a limited amount of images in the dataset. The effectiveness of the generated images needs to be evaluated with some metrics such as Fréchet inception distance (FID).

Main objectives:

- *Baseline implementation*: Implement and train a baseline GAN, establishing performance benchmarks through metrics;
- *Efficient training techniques*: Apply some techniques to the baseline to make a more stable training;
- *Limited data*: Try to adapt the model to reach good performances with limited data in the training phase;
- *Evaluation*: Evaluate your changes with some metrics such as FID.

References:

- Tianlong Chen, Yu Cheng, Zhe Gan, Jingjing Liu, & Zhangyang Wang. (2021). Data-Efficient GAN Training Beyond (Just) Augmentations: A Lottery Ticket Perspective.
- Yifan Gong, Zheng Zhan, Qing Jin, Yanyu Li, Yerlan Idelbayev, Xian Liu, Andrey Zharkov, Kfir Aberman, Sergey Tulyakov, Yanzhi Wang, & Jian Ren. (2024). E^2GAN: Efficient Training of Efficient GANs for Image-to-Image Translation.
- Yiwen Huang, Aaron Gokaslan, Volodymyr Kuleshov, & James Tompkin. (2025). The GAN is dead; long live the GAN! A Modern GAN Baseline.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, & Yoshua Bengio. (2014). Generative Adversarial Networks.

Project 5: Efficient Attention Impact on Transformer Backbones for Perceptual Similarity

Abstract: Recent research on perceptual similarity metrics has introduced DreamSim [1], a method that leverages Vision Transformer (ViT) backbones to learn embeddings aligned with human judgments of image similarity. In parallel, efficient attention mechanisms have been proposed to reduce the quadratic complexity of standard self-attention, with promising results in different vision tasks, although showing some counter-effects on maintaining the informativeness of the embeddings when applied to an encoder module [2]. This project aims to investigate whether integrating these efficient attention modules into ViT backbones within the DreamSim pipeline leads to similar effects of information dispersion and performance trade-offs.

Dataset: NIGHTS Triplets Dataset

Task: The aim of this project is to integrate efficient attention modules into ViT backbones within a DreamSim-like pipeline. The encoder will be used to compute image embeddings, and cosine distance will be used to decide which image in each triplet is more similar to a reference. By replacing standard attention layers with Meta [3], Pyra [4], or MoH [5] variants, students will assess how these modifications affect both the perceptual similarity performance (agreement with human judgments) and computational efficiency.

Main objectives:

- *Dataset preparation:* Prepare a manageable subset of the NIGHTS dataset (or similar triplets) for training and evaluation. Ensure images are preprocessed and resized to match model input requirements.
- *Model implementation:* Implement one or more ViT backbones. Integrate them into a DreamSim-like pipeline to extract embeddings and compute similarity decisions.
- *Efficient attention integration:* Replace standard self-attention layers in the backbone with efficient variants,
- *Evaluation:* Measure the agreement with human judgments (2AFC accuracy) and record inference speed on a GPU-limited environment (e.g., Colab). Compare results with the baseline backbone.
- *Ablation study:* Perform a comparative analysis across different efficient attention mechanisms and across encoder configurations (e.g., modifying all layers vs. only some layers). Discuss observed trade-offs between similarity accuracy and speed.

References:

1. Fu, S. et al. "DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data." NeurIPS 2023.
2. Schiavella, C. et al. "Efficient Attention Vision Transformers for Monocular Depth Estimation on Resource-Limited Hardware." Scientific Reports 2025.
3. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., ... Yan, S. (2022). MetaFormer Is Actually What You Need for Vision. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2111.11418
4. W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021.
5. Jin, P., Zhu, B., Yuan, L., & Yan, S. (2025). MoH: Multi-Head Attention as Mixture-of-Head Attention. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2410.11842

Project 6: Mitigating Bias in Vision-Language Models

Abstract: Vision-Language-Models (VLMs), such as CLIP, have demonstrated remarkable generalization capabilities, largely due to large-scale pre-training on uncurated web data. However, this reliance on this type of data introduces the risk of the models inheriting biases and stereotypes in the training corpus. These biases raise significant ethical concerns, especially in real-world applications where the model's outputs can influence decision-making and societal outcomes. Consequently, there is increasing interest in methods to mitigate biases in VLMs, enhancing their fairness and reliability.

Dataset: FairFace (for unbiasing evaluation), OxfordIIITPet (for zero-shot evaluation)

Task: This project aims to develop a technique to mitigate gender and ethnicity biases in VLMs. Using a provided evaluation dataset, the evaluation will focus on biases related to a predefined set of professions (e.g., doctor, nurse, etc. . .). Given that techniques can influence the model's generalization capabilities, students will also assess the zero-shot performance of the fine-tuned model on two domain-specific datasets. This dual evaluation will help identify trade-offs introduced by the debiasing process.

Main objectives:

- *Fine-tuning*: by leveraging a collected dataset (or any alternative dataset or strategy developed by the student), the fine-tuning process should reduce bias while preserving the model's overall capabilities.
- *Bias Evaluation*: Assess the representation of gender and ethnicity across professions using the provided evaluation dataset. The original (pre-fine-tuned) model will serve as the baseline for comparison. The expected result is a distribution of prediction flatter than the baseline.
- *Zero-Shot Generalization*: Compare the fine-tuned model with the original version on two domain-specific datasets to evaluate the zero-shot preservation capabilities of the proposed fine-tuning method.
- *Ablation study*: some key aspects to analyze
  - Do different versions of the same model exhibit the same biases?
  - What might these differences reveal about the data distributions used for pre-training?
  - How could variations in training data influence the direction and magnitude of biases in these models?

References:

1 Alec Radford, et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

2 Ibrahim Alabdulmohsin, et al. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? ICLR 2024.

3 M. D'Incà, et al. OpenBias: Open-set Bias Detection in Text-to-Image Generative Models. CVPR 2024.

4 Yi-Fan Zhang, et al. Debiasing Multimodal Large Language Models. ArXiv 2024.

5 Neale Ratzlaff, et al. Debiasing Large Vision-Language Models by Ablating Protected Attribute Representations. NeurIPS workshop on SafeGenAI 2024.

Project 7: Vehicle Re-Identification Using CNN-based Models

Abstract: Vehicle Re-Identification (Re-ID) is a critical component of Intelligent Transportation Systems and Urban Surveillance, focusing on matching vehicles across different non-overlapping camera views. In spite of progress in Deep Learning techniques, recognizing a vehicle with precision and accuracy remains a challenging task, aggravated by the presence of viewpoint and light variations, along with occlusion cases. This work involves investigating what are the best strategies to enhance the discrimination between vehicle models with very similar designs, aiming at building an efficient-based CNN model that is able to correctly separate multiple classes of vehicles. Various data augmentation techniques could be applied, and model efficiency should be taken into account.

Dataset: VRU & VeRi-776 (under request)

Task: The project's primary objective is to find an efficient and strong approach to extract meaningful spatial features from a Vehicle Re-ID Neural Network model *(often* a CNN), as well as a direct comparison of how different feature extraction methods and distance metrics could affect its performance. Students will showcase how several different feature augmentations, post-processing techniques, and network improvements could benefit the model in having a better distinction between different classes of vehicles. Finally, this work will create an effective Vehicle Re-ID system that is a good compromise between performance needs and practical deployability.

Main objectives:

- *Training*: Implement and fine-tune various CNN architectures (*e.g. ResNet-18, ResNet-50, etc.*) to extract discriminative features from vehicle images and develop comparison methods using different distance metrics to match vehicle identities.
- *Evaluation and Benchmarking*: Assess the performance of those models using standard metrics (*mAP, Rank@K accuracy, CMC curves*) and compare against baseline methods to quantify improvements.
- *Feature Enhancement*: Explore techniques to enhance feature discrimination, such as data augmentation, network optimizations, robust layer mechanisms, analyzing their impact on Re-ID accuracy.
- *Ablation Study*:
  - Correctly analyze how and why different post-processing and comparison techniques affect matching precision
  - Study the effect of using a deeper CNN architecture, showcasing what's the impact on balancing feature quality versus computational efficiency
  - Carry tests with the additional VeRi-776 Dataset for a more comprehensive and reward-shaping grade

References:

- Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle re-identification: An efficient baseline using triplet embedding. CoRR, abs/1901.01015, 2019.
- Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and A strong baseline for deep person re-identification. CoRR, abs/1903.07071, 2019.
- Su V. Huynh, Nam H. Nguyen, Ngoc T. Nguyen, Vinh TQ. Nguyen, Chau Huynh, and Chuong Nguyen. A strong baseline for vehicle re-identification. CoRR, abs/2104.10850, 2021.
- Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li and Wei Jiang. TransReID: Transformer-based Object Re-Identification. CoRR, abs/2102.04378, 2021.

Project 8: Invisibility Cloak for Depth Deception

Abstract: Monocular depth estimation is a fundamental computer vision task with critical applications in autonomous driving, augmented reality, and robotics. While substantial progress has been made in improving the accuracy of depth estimation models, their vulnerability to adversarial attacks remains relatively unexplored. Adversarial examples have been extensively studied in image classification and object detection domains, demonstrating the susceptibility of deep learning models to imperceptible perturbations. This project investigates the transferability of physical adversarial attack techniques, inspired by the "Invisibility Cloak" approach for object detectors, to the domain of monocular depth estimation. By analyzing how carefully crafted physical objects can manipulate depth predictions in real-world scenes, we aim to assess the robustness of current depth estimation architectures and highlight potential security concerns in depth-dependent applications.

Dataset: NYU Depth V2

Task: The primary task of this project is to investigate how adversarial patterns, designed to be physically realizable, can manipulate the output of state-of-the-art monocular depth estimation models. This involves implementing a baseline depth estimation model trained on the NYU Depth v2 dataset, then developing an adversarial attack framework that can generate patterns to cause targeted depth prediction errors. The attack methodology will be adapted from the "Invisibility Cloak" approach, focusing on creating printable patterns that can be placed on objects to alter their perceived depth. The project will explore different adversarial objectives, including making objects appear closer or farther than their actual position, completely "erasing" objects from the depth map, or creating false depth artifacts. The effectiveness of these attacks will be evaluated across different lighting conditions and viewing angles to assess their robustness in real-world scenarios. Additionally, the project will investigate potential defense mechanisms, such as adversarial training or input preprocessing, to mitigate the impact of such attacks on depth estimation systems.

Main objectives:

- *Baseline implementation*: Implement and train a state-of-the-art monocular depth estimation model on the NYU Depth v2 dataset, establishing performance benchmarks for normal operating conditions;
- *Attack adaptation*: Modify the adversarial attack methodology from the "Invisibility Cloak" paper to target depth estimation networks, developing techniques for generating physically realizable patterns that manipulate depth predictions;
- *Effect characterization*: Analyze and quantify the effectiveness of different adversarial patterns across various scenes, object types, and viewing conditions, identifying the factors that influence attack success rates;
- *Physical validation*: Create a small set of physical adversarial patterns and test their effectiveness in real-world settings, comparing the results with those predicted by the digital simulation;
- *Defense exploration*: Investigate potential countermeasures against the developed attacks, including adversarial training, input preprocessing techniques, and model architectural modifications to improve robustness.

References:

- Thys, S., Van Ranst, W., & Goedemé, T. (2019). Fooling automated surveillance cameras: adversarial patches to attack person detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops;
- Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In International conference on machine learning.
- Wu, Z., Lim, S.-N., Davis, L., & Goldstein, T. (2020). Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1910.14667

Project 9: Blind Blur Estimation for Image Restoration

Abstract: Image restoration is a fundamental task in computer vision, with image blur being one of the most common types of degradation. Blind blur estimation, which aims to identify the type and parameters of a blur without access to the original image or the blur kernel, is particularly challenging due to the diversity and ambiguity of real-world degradations. This project focuses on the problem of blind image blur estimation by designing a two-stage learning-based framework. In the first stage, a neural model classifies the type of blur affecting an image region. In the second stage, a regression module estimates the corresponding blur parameters to support further restoration. The study also explores the effectiveness of different architectures for feature extraction, comparing traditional deep neural networks with more recent lightweight transformer-based encoders.

Dataset: BSDS500, DIV2K

Task: The goal of this project is to develop and compare two models for blind image blur estimation. The first replicates the original approach based on a deep neural network (DNN) followed by a general regression neural network (GRNN) for parameter prediction. The second replaces the DNN with a lightweight transformer encoder (e.g., MobileViT, TinyViT) trained on Fourier-domain representations of degraded image patches. The task includes training, evaluation, and analysis of classification accuracy, regression precision, and computational performance.

Main objectives:

- *Reproduce a Deep Learning Baseline*: Implement the original two-stage approach for blur type classification (DNN) and blur parameter estimation (GRNN) based on the 2016 paper by Yan and Shao.
- *Introduce a Transformer-based Encoder*: Replace the DNN with a lightweight transformer model for feature extraction and blur type classification.
- *Perform Quantitative Evaluation*: Compare the two systems based on classification accuracy, parameter estimation error, and computational metrics such as training time and memory usage.

References:

- R. Yan and L. Shao, "Blind Image Blur Estimation via Deep Learning," in IEEE Transactions on Image Processing, vol. 25, no. 4, pp. 1910-1921, April 2016, doi: 10.1109/TIP.2016.2535273.
- Howard, A., et al. (2019). Searching for MobileNetV3. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/1905.02244
- Zheng W, Lu S, Yang Y, Yin Z, Yin L. 2024. Lightweight transformer image feature extraction network. PeerJ Computer Science 10:e1755 https://doi.org/10.7717/peerj-cs.1755

Project 10: Lightweight Occlusion Removal

Abstract: Image inpainting is a fundamental computer vision task that aims to restore missing or damaged regions in images with visually plausible content. Effective occlusion removal requires understanding complex scene structures, textures, and contextual relationships to generate coherent visual results. While recent deep learning approaches have shown significant progress, balancing reconstruction quality with computational efficiency remains challenging. Lightweight architectures like PEPSI++ offer promising solutions by achieving competitive results with fewer parameters, yet there remains room for improvement in handling diverse occlusion patterns and preserving structural consistency. This project explores the effectiveness of the PEPSI++ architecture for occlusion removal across various scenarios and investigates potential enhancements to improve its performance without significantly increasing computational complexity.

Dataset: Places2, COCO

Task: The primary task of this project involves implementing and evaluating the PEPSI++ network for occlusion removal in natural images. This entails implementing the baseline PEPSI++ model as described in the original paper, then exploring modifications to its loss function by incorporating perceptual and adversarial components to enhance visual quality. The training process will involve experimenting with different mask generation strategies to improve the model's ability to handle various occlusion patterns. Evaluation will be conducted using both quantitative metrics such as PSNR, SSIM, and FID, and qualitative assessments focusing on perceptual quality and structural coherence. Particular attention will be given to challenging occlusion scenarios where the background contains complex textures or structural patterns that are typically difficult to reconstruct using existing lightweight approaches. The project will also include a comparative analysis against other inpainting methods to contextualize the performance of the enhanced PEPSI++ model.

Main objectives:

- *Baseline implementation*: Implement the original PEPSI++ architecture as described in the paper, establishing performance benchmarks on standard datasets with various occlusion configurations;
- *Loss function exploration*: Experiment with different combinations of reconstruction, perceptual, and adversarial losses to identify the optimal balance for visually coherent occlusion removal;
- *Mask strategy optimization*: Investigate the impact of different mask shapes, sizes, and distributions during training on the network's ability to handle real-world occlusion patterns;
- *Cross-dataset evaluation*: Assess the model's generalization capabilities by training on one dataset (Places2) and evaluating on another (COCO), identifying potential domain-specific limitations;
- *Comparative analysis*: Perform detailed quantitative and qualitative comparisons against other relevant inpainting approaches across different occlusion scenarios, highlighting the strengths and limitations of the PEPSI++ architecture.

References:

- Li, W., Lin, W., Chen, L., Li, J., & Tang, C. (2022). PEPSI++: Fast and Lightweight Network for Image Inpainting. arXiv preprint arXiv:2202.08949;
- Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV);
- Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition;