Trends and Applications of Computer Vision

Massimiliano Mancini, Giulia Boato,
Nicu Sebe

A.A. 2024/2025

Learning Goals

Hard skills

- Gain knowledge on advanced computer vision topics
- Deepen one topic through a project-based activity
- Learn to do a literature search on a given topic
- Fully understand the principles and limits of the techniques studied
- Get familiar with practicals tools and codes

Soft skills

- Learn to present formal arguments in talk and written form
- Learn to work in team and to communicate with supervisors

Main Areas

Multimedia Forensics
(Giulia Boato)

T1
Multimedia Forensics, deepfake detection, AI-generated media identification

Vision-Language Models
(Massimiliano Mancini)

T2
Vision-Language Models, test-time adaptation, vision-by-language

Team-based project activity

- Work on one subtopic among a list of proposed ones *(autonomously)*
  - work in small groups
  - get deeper knowledge and explore related work
  - set up experiments with pointed tools

- Present results in the form of presentations *(in class)* and written reports
  - the activity is supervised by one lecturer with fixed milestones
  - the activity is monitored together with Prof. **Nicu Sebe**

Timeline

Date	Plan
30/10	Presentation of projects list
06/11	T2 - Vision and Language
08/11	Lab on T1
13/11	Seminar on T1: **Journalism in the deepfake era**
15/11	Lab on T2
20/11	In progress meeting
22/11	Seminar on T2: **Davide Boscaini**
27/11	**Presentations "Related work and project status" I**
29/11	**Presentations "Related work and project status" II**
04/12	Final in progress meeting (Q&A on projects)
06/12	No lecture
11/12	**Final project discussion I**
13/12	**Final project discussion II**

Fix in your calendar the dates of the presentations!

Project milestones

Presentation on related work and project status *(in class)*:

- present the initial source(s) as well as the most important related work
- state how and why these sources are important for your topic
- expose the tools you have used and initial experiments you made
- max 15 minutes: 10 related work + 5 project status
- dates: November 27 and November 29

Intermediate meeting (only if needed - to help you, not an evaluation)

- write the lecturer and ask for an appointment in November

Related work write-up *(written report)*:

- summarize and categorize the relevant sources
- elaborate on why the sources are important
- be careful when formatting the bibliography
- 3 pages (excluding references)
- deadline December 3 (send the report via email to massimiliano.mancini@unitn.it and giulia.boato@unitn.it)

Project milestones

In-progress meetings (*to help you, not an evaluation*):

- present your progresses so far to the supervising lecturer
- explain the difficulties encountered so far, if any
- describe the open issues and discuss how you plan to address them
- date: November 20 and December 4

Final presentation and demonstration *(in class)*:

- introduce why your topic is relevant
- first, give the theoretical background
- then, give an overview and a small demonstration of your work
- ideally, conclude with an outlook
- max 20 minutes
- dates: December 11 and 13

Elements for grading

- Related work and project status *(presentation in class)*
- Related work write-up *(written report)*
- Final project discussion *(presentation in class)*

NOTES:

- When preparing the presentations, keep in mind that the audience is composed of lecturers and fellow students. The talks should be structured in such a way that everyone is able to follow.

- **All group members must contribute to the project work and to the presentations.**

Projects list and description:

10 projects are available, groups of 3 components are expected:

- P1 - Exploring the Adversarial Robustness of AI-generated Image Detectors
- P2 - Exploring the Adversarial Robustness of AI-generated Video Detectors
- P3 - Exploring Detector’s Robustness of AI-generated Multimedia shared on Social Networks
- P4 - Few-Shot and Continual Learning for Fake Image Detection
- P5 - AI generated Videos Detection
- P6 - Efficient test-time adaptation
- P7 - Vocabulary-free semantic segmentation
- P8 - Vision-by-language for bias identification
- P9 - Unlearning with SalUN
- P10 - CLIP on low-resource vision

Rules

Each group should send us via email:

- the components and name of the group
(**3 group members are expected**)

- the list of **ALL** projects ordered by preference (*from the most preferred to the least preferred*).

Deadline: November 4 (11.59 pm)

Special cases:

If you want to propose your own project, you are welcome! But:

1. Send us your plan by e-mail
   (i.e., title, repo you would start with, planned experiments/analyses)

2. Still share your preferences
   (we aim to keep a balanced group assignment across topics)

3. If the project is linked to another course (e.g., Advanced computer vision) share with us:
   a. The details of the project for the other course
   b. What will be the shared components and the different ones

Trends in CV

Projects presentation for T1

Giulia Boato

Project #1 • Exploring the Adversarial Robustness of AI-generated Image Detectors

The students working on this project will:

- Work on a large and diverse dataset of images generated by the MMLAB group with several generative techniques
- Evaluate the Adversarial Robustness of a Fake Image Detectors under different attacks
- Get familiar with deepfake detection State-of-the-Art (SoA)
  - CLIP-aided fake image detection
  - ResNet50 fake image detection
  - Robust Artifacts for Fake Image Detection
  - Exposing gan-generated profile photos from compact embeddings
  - Lighting (in) consistency of paint by text
  - Perspective (in) consistency of paint by text
- Get familiar with multimedia forensics adversarial attacks SoA
  - Mimicry attack (Boato2023)
  - Carlini Farid
  - Carrara
  - Verdoliva
  - Stable Diffusion Laundering
- Challenge modern fake image detectors by using the above mentioned adversarial techniques and develop solution robust to these latter.

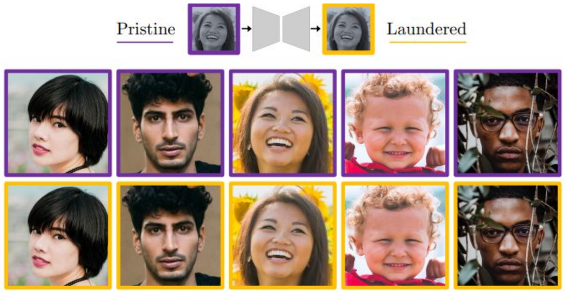


Fig. 1: Pristine images and their laundered copies obtained by passing pristine samples through SD autoencoder. Laundered samples look extremely realistic, with almost a total absence of notable generation artifacts, even in the case of uncommon patterns that could be harder to reproduce.

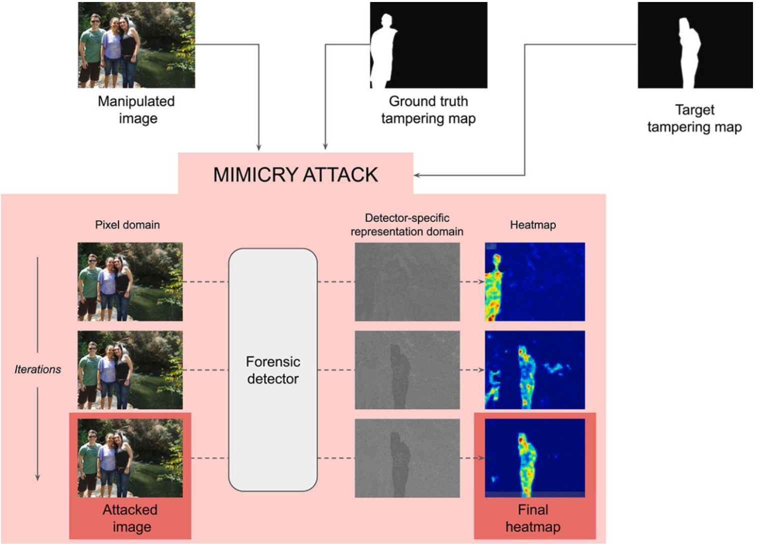
Project #2 • Exploring the Adversarial Robustness of AI-generated Video Detectors

The students working on this project will:

- Work on a large dataset of video generated by several generators
- Evaluate the Adversarial Robustness of a Fake Video Detectors under different attacks usually applied on images

- Get familiar with deepfake detection State-of-the-Art (SoA)
  - ResNet50
  - ResNet3D
  - MISLNet
- Invistigate novel solutions:
  - CLIP-aided fake media detection
  - LLMs
  - Method under submission of the MMLAB group

- Get familiar with multimedia forensics adversarial attacks SoA
  - Mimicry attack (Boato2023)
  - Carlini Farid
  - Carrara
  - Verdoliva
  - Stable Diffusion Laundering
- Challenge modern fake video detectors by using the above mentioned adversarial techniques and develop solution robust to these latter.



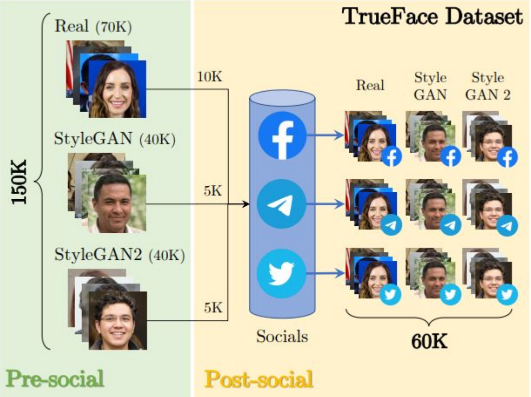
Project #3 • Exploring Detector's Robustness of AI-generated Multimedia shared on Social Networks

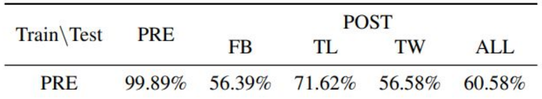
The students working on this project will:

- Work on a large and diverse dataset of multimedia generated by the MMLAB group using several diffusion models, proprietary and not.
- Evaluate the Implications of Social Network multimedia processing in multimedia forensics problems such as: real v. fake detection

- Get familiar with deepfake detection State-of-the-Art (SoA)
  - ResNet50
  - ResNet3D
  - Method under submission of the MMLAB group
  - CLIP-aided fake image detection
  - ResNet50 fake image detection
  - Robust Artifacts for Fake Image Detection
  - Exposing gan-generated profile photos from compact embeddings
  - Lighting (in) consistency of paint by text
  - Perspective (in) consistency of paint by text
  - TrueFace Dataset
  - Fake Image shared on Twitter

- Students can choose to work either on images or videos.
- If students decided to work on videos, the students will also develop tool for automatic upload/download of video contents.
- The goal of this project is to develop solutions that cope with social network multimedia rocessing, by implementing architectures robust to these social network media processing or that can easily learn how to adapt to these latter.





Train\Test	PRE	POST			
		FB	TL	TW	ALL
PRE	99.89%	56.39%	71.62%	56.58%	60.58%

Project #4 • Few-Shot and Continual Learning for Fake Image Detection

The students working on this project will:

- Work on a large and diverse dataset of images generated by the MMLAB group with several diffusion models
- Improve the generalization of real v. fake detectors, avoiding catastrophic forgetting phenomena.

- Get familiar with deepfake detection State-of-the-Art (SoA)
  - ResNet50 fake image detection
  - Few-Shot Learning
  - Continual Learning

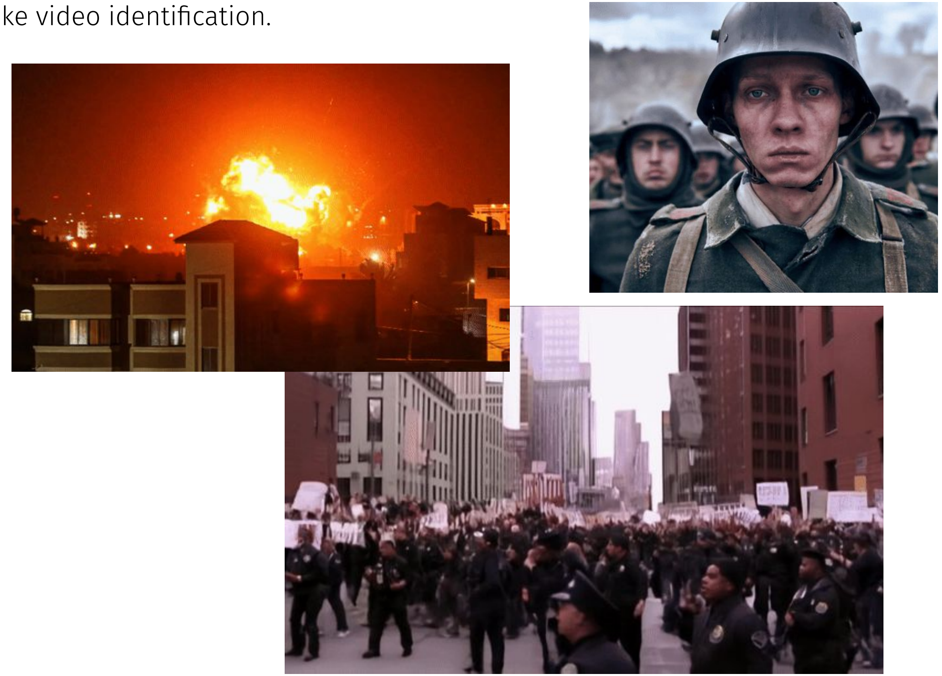
Project #5 · AI generated Videos Detection

The students working on this project will:

- Generate a small dataset of videos fully generated by diffusion models.
- Evaluate the robustness and adaptability of method developed for fake video identification.

- Get familiar with deepfake detection State-of-the-Art (SoA)
  - ResNet50
  - ResNet3D
  - MISLNet
- Invistigate novel solutions:
  - CLIP-aided fake media detection
  - LLMs
  - Method under submission of the MMLAB group

- Get familiar with diffusion models for video generation
  - Stability-AI
  - LattE
  - Art-v
  - CogVideo



Trends in CV

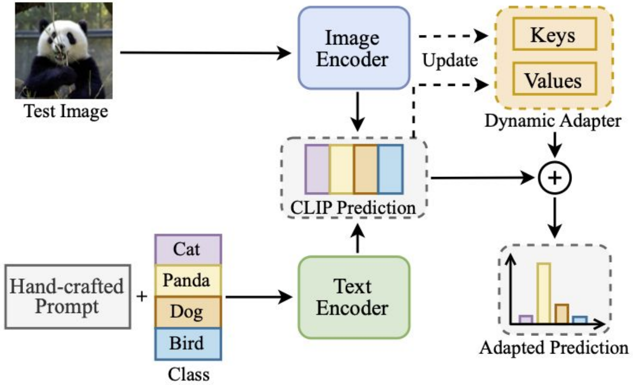
Projects presentation for T2

Massimiliano Mancini

Project #6 · Efficient test-time adaptation

Background: We have seen how caching can improve performance without training. What are its pros and cons for TTA?

1. Report on the main techniques proposed in the literature to perform continuous test-time adaptation.
   Focus on papers talking about challenges related to data streams (i.i.d. vs non i.i.d.). Potential starting points:
   a. https://proceedings.neurips.cc/paper_files/paper/2022/file/ae6c7dbd9429b3a75c41b5fb47e57c9e-Paper-Conference.pdf
   b. http://openaccess.thecvf.com/content/CVPR2022/html/Volpi_On_the_Road_to_Online_Adaptation_for_Semantic_Image_Segmentation_CVPR_2022_paper.html
2. Get familiar with TDA, a caching method for TTA and with its implementation available online
   a. Paper: https://openaccess.thecvf.com/content/CVPR2024/papers/Karmanov_Efficient_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2024_paper.pdf
   b. Code: https://kdiaaa.github.io/tda/
3. Benchmark them using standard datasets (e.g., natural distribution shift, Tab. 3)
   a. What are the failure cases?
4. How do its performance change w.r.t.:
   a. Hyperparameters
   b. Number of seen samples
   c. Data ordering/class distribution
5. Check how performance change w.r.t. budget-aware constraints
   a. What if predictions are made before vs after the cache update?
   b. What if cache-update time is considered during real-time inference?
6. (Bonus) mitigate the above issues
   a. Can you apply existing approaches to mitigate the above issues?



Karmanov, Adilbek, et al.
"Efficient test-time adaptation of vision-language models."
CVPR 2024

Project #7 · Vocabulary-free semantic segmentation

Background: We have seen how to remove label constraints for classification with CLIP. Can we do the same for segmentation?

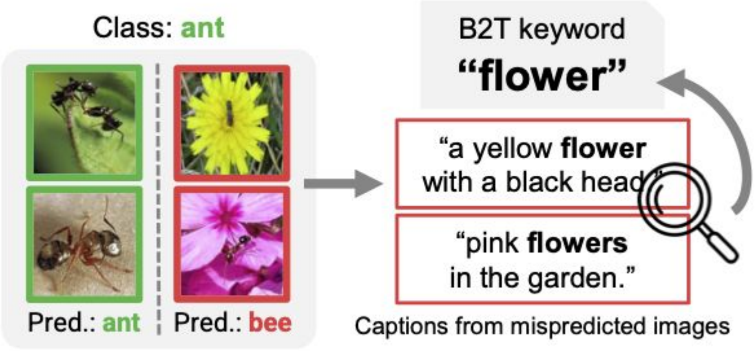
1. Report on the main techniques proposed in the literature for open vocabulary semantic segmentation. Potential starting points:
   a. https://github.com/Qinying-Liu/Awesome-Open-Vocabulary-Semantic-Segmentation
   b. https://arxiv.org/pdf/2112.01071
2. Get familiar with SAN, a state-of-the-art model for open-vocabulary semantic segmentation, and with its implementation
   a. Paper: https://openaccess.thecvf.com/content/CVPR2023/papers/Xu_Side_Adapter_Network_for_Open-Vocabulary_Semantic_Segmentation_CVPR_2023_paper.pdf
   b. Code: https://mendelxu.github.io/SAN
3. Benchmark it with different strategies to get the list of candidate labels:
   a. GT?
   b. Retrieval?
   c. Captioning?
4. Get familiar with SAM, a state-of-the-art model for image segmentation, and with its implementation available online
   a. Paper: https://openaccess.thecvf.com/content/ICCV2023/papers/Kirillov_Segment_Anything_ICCV_2023_paper.pdf
   b. Code: https://github.com/facebookresearch/segment-anything
5. Compare SAN and SAM
   a. What are their major performance differences?
   b. Do they show the same failure cases?
6. (Bonus) expand the experimental suite
   a. What happens if you test on rare domains and/or specific ones (e.g., sketches/cartoons, etc.)

Xu, Jiauri, et al.
"Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models." CVPR 2023.

Project #8 · Vision-by-language for bias identification

Motivation: The vision-by-language paradigm allows to tackle a variety of downstream tasks. Can it be applied also to analyze models?

1. Report on the main techniques proposed in the literature to identify biases. Potential starting points:
   - a. https://openaccess.thecvf.com/content/WACV2021/papers/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.pdf
   - b. https://dl.acm.org/doi/abs/10.1145/3637549 and/or (even better) related work in B2T
2. Get familiar with B2T, a simple model based on captioning for training-free bias identification
   - a. Paper: https://openaccess.thecvf.com/content/WACV2021/papers/Karkkainen_FairFace_Face_Attribute_Dataset_for_Balanced_Race_Gender_and_Age_WACV_2021_paper.pdf
   - b. Code: https://github.com/alinlab/b2t
3. Benchmark it using datasets and models of your choice
   - a. E.g., CelebA and Waterbirds as in the paper
   - b. Arbitrary classification datasets from torchvision
4. How do its performance change w.r.t.:
   - a. Captioner
   - b. Keywords extraction strategies
   - c. CLIP version
5. Check how things change w.r.t. different models
   - a. Do more powerful models lead to different biases?
   - b. Do models trained with different strategies (e.g., augmentation, dataset) lead to different biases?
6. (Bonus) bias mitigation
   - a. Can you apply apply a technique for bias mitigation?
   - b. What is the effect of bias mitigation on different models?

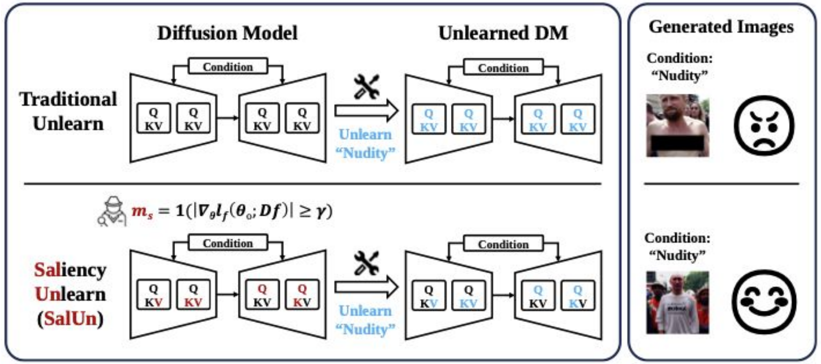

Class: ant
B2T keyword
"flower"
"a yellow flower with a black head"
"pink flowers in the garden."
Pred.: ant
Pred.: bee
Captions from mispredicted images

Kim, Younghyun, *et al.*
"Discovering and Mitigating Visual Biases through Keyword Explanation."
CVPR 2024.

Project #9 · Unlearning with SalUN

Motivation: We have introduced unlearning as a growing field to remove data from a model. Here we delve into one of the methods.

1. Report on the main techniques proposed in the literature for unlearning. Potential starting points:
   a. https://github.com/tamlhp/awesome-machine-unlearning
   b. https://proceedings.neurips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html
2. Get familiar with the SalUN codebase, performing unlearning via pruning and random labeling
   a. Paper: https://arxiv.org/pdf/2310.12508
   b. Code: https://github.com/OPTML-Group/Unlearn-Saliency
3. Benchmark it on datasets of your choice (even the provided ones)
   a. Cifar-10
   b. Arbitrary datasets/models in torchvision
4. How do the performance of the model changes w.r.t.:
   a. Hyperparameters
   b. Pruning ratios
   c. Changing the objective
   (i.e., random labeling vs entropy maximization)
5. Study unlearning across different groups of data
   a. What are the easiest distributions of samples to unlearn?
   And what are the hardest ones?
   b. How does performance change w.r.t. number of unlearned samples?
6. (Bonus) explore one specific (other) application
   a. NSFW concept removal?
   b. Generative models?

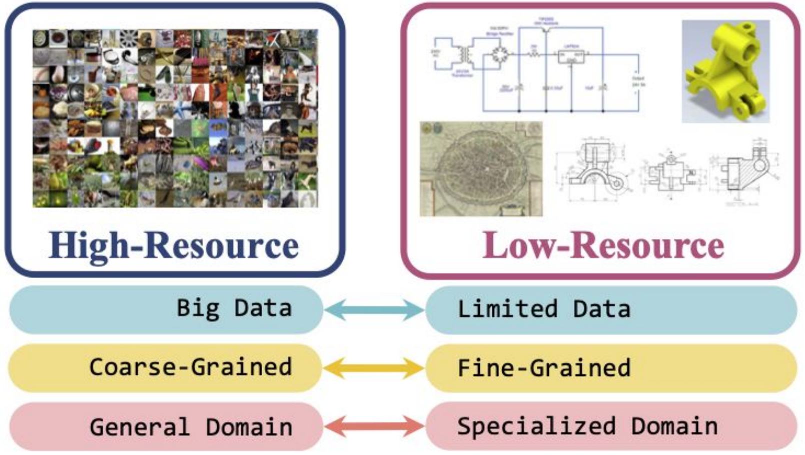


*Fan, Chongyu, et al.
"SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in both Image Classification and Generation."
ICLR 2024.*

Project #10 · CLIP on low-resource vision

Motivation: CLIP works quite well on a lot of settings. Does this hold also for low-resource scenarios?

1. Report on the main techniques proposed for low-shot learning and vision on rare domains. Potential starting points:
   a. https://github.com/Bryce1010/Awesome-Few-shot?tab=readme-ov-file
   b. https://link.springer.com/article/10.1007/s11263-022-01622-8
2. Get familiar with the baselines proposed in the reference above, used to tackle various low-resource challenges
   a. Paper: https://arxiv.org/abs/2401.04716
   b. Code: https://github.com/xiaobai1217/Low-Resource-Vision
3. Benchmark it using the provided datasets
   a. Circuit classification
   b. Drawing/maps retrieval
4. How do the performance of the model changes w.r.t.:
   a. Number of available samples
   b. Number of generated samples
   c. Other hyperparameters
5. Explore failure cases
   a. What are the easiest sample? And what are the hardest ones?
   b. Are there patterns on the embedding space that make an example easier/harder?
6. (Bonus) adding adapters
   a. Play with adapters (i.e., add LoRA/CoOp, whatever)
   b. Check their failure cases



Xhang, Yunhua, et al.
"Low-Resource Vision Challenges for Foundation Models."
CVPR 2024.