PROJECT TITLE

Design and Development of algorithms for Face Privacy Protection and Deepfakes Prevention

PROJECT DESCRIPTION

This proposal pertains to a research grant funded by the MUR for the PRIN project “AdVVent: Adversarial Venture, the Mixed Blessing of Adversarial Attacks”.
The research project involves the design, development, and experimental validation of algorithms for face privacy protection and deepfakes prevention. More in detail, the project will focus on the development of novel deep learning-based computer vision algorithms for (i) securing and protecting face imagery against other AI methods that automatically scrape faces images using a face recognition engine, i.e., web scraper threat; (ii) preventing the creation of Deepfakes, or else, in case the perturbation fails in disrupting the Deepfake generation model, to inject non-visible noise in the Deepfaked output so that it is possible to test if the image was generated through AI.
The project will involve the use of state-of-the-art algorithms for image generation, based on Generative Adversarial Networks, Diffusion Models, or other deep learning models considered interesting to explore. After a first phase of analysis of the state-of-the-art, the primary objectives guiding the algorithms development and performance evaluation activities will be delineated.

ACTIVITY PLAN

The project described above will unfold through the following main phases:

State of the art analysis:
In the initial phase, the Researcher will conduct a detailed analysis of the state of the art in computer vision algorithms for the specific problem at hand. A review of the scientific literature will be carried out regarding the topics related to adversarial attacks for privacy preservation and pro-active scheme of defense for detecting and localizing the manipulation of face images.

Definition of the research and development goals:
Following the analysis of the state of the art, the activity will focus on the definition of the research and development goals. One potential research focus could involve the exploration of generative adversarial networks or denoising diffusion probabilistic models. These models could serve a dual purpose: firstly, to introduce noise into the image that can interfere with the performance of recognition algorithms, and secondly, to embed a template within the image, enabling the detection and localization of potential image manipulations.

Development and test of the algorithms
Once the objectives are defined, the algorithm development activity will start. During this phase, the performance evaluation will be conducted using publicly available dataset, such as standard datasets commonly used as benchmarks by the international scientific community. The tests will allow for assessing whether the developed algorithms provide adequate results in comparison to the state of the art.