Responsible AI at the edge: towards privacy-preserving smart cities

Luca Zanella[1] , **Yiming Wang**[1] , **Nicola Dall'Asen**[2,3] , **Alberto Ancilotto**[1] , **Francesco Paissan**[1] , **Elisa Ricci**[1,2] , **Elisabetta Farella**[1] , **Alessio Brutti**[1] , **Marco Pistore**[1].
Digital Society Center, Fondazione Bruno Kessler[1] , University of Trento[2] , University of Pisa[3]
lzanella@fbk.eu

Abstract

With the massive amount of data produced by ambient environmental sensors, many AI-based solutions are emerging to support new smart cities' applications. However, these data may contain sensitive personal information, calling for responsible AI solutions. FBK proposes a privacy-preserving subsystem with a set of technological components that enable responsible AI and prevent unauthorised usage of personal data at the data storage and during data transmission under the context of Smart Cities. We demonstrate the proposed solution under an EU project MARVEL, where both video and audio anonymisation components are deployed at the edge level, enabled by a model compression component for complexity reduction. We discuss each component's technical challenges, current progress, and future directions.

1 Introduction

With the development of the Internet of Things (IoT) in modern society, the concept of *Smart Cities* is rapidly emerging. The large-scale city data captured from different sensors such as cameras and microphones, and smartphones provides inhabitants and policy-makers with the possibility of situational awareness, operational decision-making, and urban planning, empowered by Artificial Intelligence (AI) technologies [Zhang *et al.*, 2017; Khan *et al.*, 2017]. However, there are also growing concerns in security and privacy towards these AI technologies as the large-scale urban data contains sensitive personal information [Asghar *et al.*, 2019], calling for responsible *privacy-by-design* applications, with the safety and security of the final users at their core. Especially in the European territory, personal data must be processed in compliance with the General Data Protection Regulation [Commission, 2018] to ensure EU citizens' and visitors' rights and enforce a transparent way for handling data.

The ambition of the *Digital Society Center in Fondazione Bruno Kessler (FBK)* is to develop responsible, sustainable, inclusive, and secure digital technologies for smart cities via the excellence in interdisciplinary research and the strong partnerships with public administrations, primarily the Autonomous Province of Trento and the Municipality of Trento on strategic domains such as school, digital services, urban security, environmental sustainability. FBK has been involved in the EU project MARVEL[1] [Bajovic *et al.*, 2021], as well as a related project PROTECTOR[2], whose objective is to exploit ambient urban sensors, such as CCTV cameras and microphones, and employ multi-modal perception, AI analytics, software engineering, and the Edge-to-Fog-to-Cloud Computing Continuum paradigm to support data-driven real-time application workflows and decision making in modern cities.

As MARVEL uses both audio and visual data recorded in public spaces, it is essential to minimise as much as possible the risks that these persons are re-identified by analysing the physical or behavioural traits such as faces, gaits, and voices [Ross e Othman, 2010] and prevent their personal information being misused in the future. The role of FBK in MARVEL is to develop a set of technological components that enable responsible AI and prevent unauthorised re-identification during data storage and transmission. Specifically, as faces in the visual content and voices in the audio content are the most identity-informative content, we devise and develop face and voice anonymisation algorithms by redacting the raw content that minimises the likelihood of recognising an individual without hindering further AI analysis. Such anonymisation algorithms should by design run on the edge so that only anonymised data will proceed with the transmission. To this end, FBK also develops the supporting model compression techniques that enable deep learning at the edge with limited resources. We provide an overview of our proposed responsible AI solution with the components described in the following sections.

2 A snapshot: the FBK solution in MARVEL

Privacy preservation in audio and visual data is often achieved by means of redaction techniques, e.g. obfuscation, on the personally identifiable information (PII) of a data subject. Classical anonymisation techniques, such as removing any segment with speech content or speaker-specific features for audio, or blurring and pixelation for faces can successfully remove identifiable information. Nevertheless, this comes at a high cost of deteriorating further audio-visual analysis.

[1]https://www.marvel-project.eu/
[2]https://www.protector-project.eu/

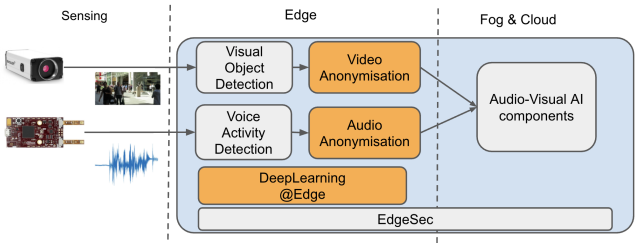

Figure 1: The overview of our responsible AI components (highlighted in **orange** blocks) in the MARVEL project following the Edge-to-Fog-to-Cloud paradigm.

Advances in Generative Adversarial Networks have allowed proposing different GAN-based audio [Kaneko *et al.*, 2019; Chen *et al.*, 2021] and video [Chen *et al.*, 2020; Li *et al.*, 2020] solutions. Voice anonymisation can be achieved by replacing the original voice signature with another identity while preserving the prosody and acoustic content, while video anonymisation can be achieved by swapping the original face with *natural-looking* faces of another identity with preserved facial pose and expression. However, as GAN-based solutions are notoriously computationally demanding, their deployment at the edge is a challenging problem. Recent advances in TinyML have allowed solving tasks that are normally performed on general-purpose GPUs, such as video object detection and tracking, on edge devices with limited resources using low-complexity networks. TinyML has been recently applied to several different classes of problems and devices [Venkatesh *et al.*, 2021]. Techniques such as model pruning [Yeom *et al.*, 2021] or knowledge distillation [Gou *et al.*, 2021] have been proposed to decrease the complexity of computationally expensive models that otherwise cannot fit on the edge.

Combining state-of-the-art solutions based on TinyML and GANs, we aim to develop lightweight GAN-based approaches for video and audio anonymisation on edge devices where data is collected. Figure 1 illustrates the proposed components (highlighted in orange) in the context of MARVEL following the Edge-to-Fog-to-Cloud paradigm. The GAN-based video anonymisation component redacts the identity-informative content by swapping the detected faces from the raw video data, while the GAN-based audio anonymisation component converts the identity of the detected speakers in the raw audio data. The Deep Learning@Edge component serves to reduce the computational load so that both video and audio anonymisation are performed at the edge.

In the following sections, we describe in detail each component together with its technical challenges and the proposed solutions.

3 Video Anonymisation

The goal of video anonymisation in MARVEL is to remove any personal information from a target identity, in particular the identity in the facial features, and substitute it with information of another, possibly non-existing source identity while preserving non-personal and geometric attributes, i.e. pose and expression for further analyses.

Technical Challenges. When anonymising CCTV videos captured by ambient cameras, there are some open challenges that existing works cannot yet address satisfactorily:

- *Pose Preservation*: preserving pose and expression in the generated faces is particularly important as we aim to feed the generated videos to downstream tasks such as emotion recognition.
- *Varying Resolution*: as people's faces can vary greatly in size, shape, and pose in surveillance videos, the anonymisation algorithm should work independently of these factors.
- *Temporal consistency*: the same face in neighbour frames should be anonymised coherently, without jittering in terms of position and lighting.

We aim to progressively address the above-mentioned challenges under the context of smart cities by proposing a GAN-based video anonymisation solution with better pose preservation.

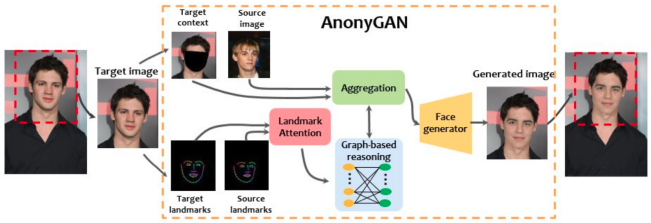

Figure 2: The architecture of **AnonyGAN**

Proposed Method. We developed **AnonyGAN** [Dall'Asen *et al.*, 2021] exploiting the landmarks of source and target images to perform face anonymisation with pose preservation. Facial pose preservation is achieved by first reasoning on the landmark to landmark relations, and then generating the image aggregating the appearance features, as depicted in Fig. 2. In this work, we proposed a landmark attention model to let the network learn on its own the importance among the landmarks to balance between visual naturalness and pose preservation.

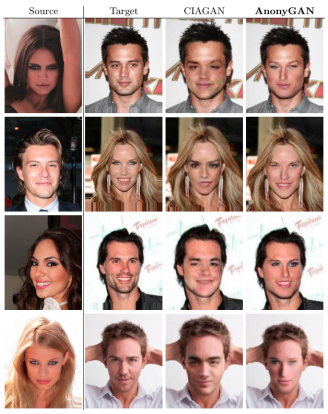

Figure 3: Qualitative results of our **AnonyGAN** in comparison with SOTA GAN-based method.

Preliminary Result. In Figure 3, we can observe that **AnonyGAN** is able to generate more natural-looking faces with a correct attribute transfer from source to target. We can see that the pose of the target is better preserved, confirming the effectiveness of the proposed Landmark Attention module.

Accurate representation of landmarks is not always possible, for example in CCTV contexts with small faces, extreme poses, and blurred images, making the proposed solution ineffective under these circumstances. Our current research focuses on a landmark-free solution that can implicitly capture the target pose and transfer it to the resulting image regardless of the resolution of the input image, to work effectively on video without preprocessing steps.

4 Audio Anonymisation

The goal of audio anonymisation is to remove information about the speaker identity from the audio streams captured by environmental microphones, while maintaining the speech content untouched for further analysis.

Technical Challenges In the literature, voice conversion is typically applied to clean speech signals without accounting for noises and reverberation effects that are typical for real-world scenarios, which are the conditions of the MARVEL where audio data are captured by distant outdoor microphones. There are a set of open technical challenges yet to be addressed by the community:

- *Acoustic background preservation*: the preservation of the acoustic scene background both in speech segments as well as when speech is not present is not a requirement for state-of-the-art voice conversion techniques, and it calls the development of specific solutions.
- *Generalisation capability*: as the environmental conditions and the recording setups considerably vary over time and across recording sites, how to achieve a generalisation solution is a key challenge.

Proposed Method Following the same strategy adopted for the video data anonymisation, the information about the speaker identity is removed using GAN-based voice conversion techniques [Kaneko *et al.*, 2019; Chen *et al.*, 2021]. Preliminary results have confirmed that many-to-many approaches generate realistic speech signals but cannot generalise to unseen speakers. Our current research focuses on novel implementations of the voice conversion paradigms, particularly suitable for privacy preservation, as non-parallel any-to-many voice conversion [Liu *et al.*, 2021]. For what concerns the preservation of the acoustic background, we will further investigate the use of signal separation methods [Luo e Mesgarani, 2019; Macartney e Weyde, 2018].

5 Deep Learning@Edge

Deep Learning@Edge is a supporting component that aims to reduce the computational load required by GAN-based video and audio anonymisation algorithms, so that they can be deployed on edge devices with limited processing resources.

Technical Challenges Bringing AI algorithms to peripheral devices is a non-trivial task and it requires the design of innovative solutions, as current neural network-based approaches are highly demanding in terms of memory footprint and computational complexity. Some technical challenges to be solved to bring the anonymisation pipeline to edge platforms are the following:

- *Limited Resources*: typical GAN-based approaches [Park *et al.*, 2019] require orders of magnitude more resources than what is available on edge devices.
- *Limited resolution*: the low RAM available on edge devices limits the size of the original image that can be processed, calling for the approach to offer good performance even with a lower input resolution;
- *Resource variability*: different platforms offer different sets of resources. The approaches should scale optimally to different hardware configurations, achieving high efficiencies regardless of computational constraints.

Proposed Method. We develop the model compression techniques with a classic generative neural network, CycleGAN, as a proof-of-concept. CycleGAN targets the problem of unpaired image to image translation, a broader formulation of the anonymisation task. As the convolutional blocks used in the original network present way higher computational complexity than what can be achieved on the edge devices, we replaced them with the blocks shown in Figure 4 [Paissan *et al.*, 2021]. The original decoder was replaced

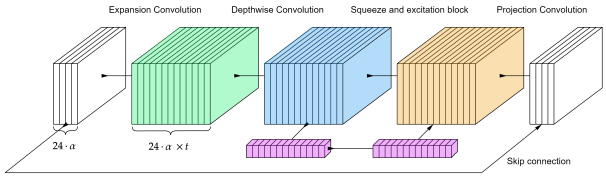

Figure 4: The architecture of a *Phinet* convolutional block.

by a sequence of upsampling blocks based on the same convolutional blocks coupled with a depth-to-space transformation to increase the resolution of the feature map with each block, resulting in a high-performance decoder that can be implemented efficiently on small edge devices, while requiring only a fraction of the original parameters and operations. Moreover, the hardware-aware scaling approach presented in [Paissan *et al.*, 2021] can be used to generate networks with good performance for certain hardware resources, avoiding the need to perform an extensive search for the network hyper-parameters, saving the time that would otherwise be required for the network architecture search step.

Preliminary Results The modified CycleGAN network reduces the number of architecture parameters to 1% of the original network that requires $11.3M$, without a noticeable drop in the quality of the generated image. It could run at 21 frames per second working on a QVGA input on a `Kendrite K210` MCU while requiring about $300mW$ of power. This demonstrated the possibility of bringing GAN-based approaches to edge devices. Future works will target video and audio anonymisation problems, by working

on compression of recent approaches for removing identity-informative content.

6 Conclusion

In this paper, we introduced the privacy-preserving subsystem for responsible AI in smart cities adopted by FBK in the EU MARVEL project, which safeguards the data captured by surveillance cameras and microphones. Regarding privacy in visual data, our GAN-based component can swap the identity of detected faces while preserving facial pose and expression. Regarding privacy in audio data, our GAN-based component can replace the identity of detected speakers while preserving prosody and acoustic content. Both anonymisation components can be deployed at the edge after the reduction of computational load, ensuring that identity-informative content never leaves the edge devices. As a result, subsequent AI analysis can be performed on anonymised data with minimal personally identifiable information, protecting the privacy of the individuals involved in the recordings.

Acknowledgement

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 957337.

References

[Asghar *et al.*, 2019] M. N. Asghar, N. Kanwal, B. Lee, M. Fleury, M. Herbst, e Y. Qiao. Visual surveillance within the eu general data protection regulation: A technology perspective. *IEEE Access*, 7:111709–111726, 2019.

[Bajovic *et al.*, 2021] Dragana Bajovic, Arian Bakhtiarnia, George Bravos, Alessio Brutti, Felix Burkhardt, Daniel Cauchi, Antony Chazapis, Claire Cianco, Nicola Dall'Asen, Vlado Delic, et al. Marvel: Multimodal extreme scale data analytics for smart cities environments. In *2021 International Balkan Conference on Communications and Networking (BalkanCom)*, pages 143–147. IEEE, 2021.

[Chen *et al.*, 2020] Renwang Chen, Xuanhong Chen, Bingbing Ni, e Yanhao Ge. Simswap. *Proceedings of the 28th ACM International Conference on Multimedia*, Oct 2020.

[Chen *et al.*, 2021] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, e Hung-yi Lee. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5954–5958, 2021.

[Commission, 2018] European Commission. Reform of eu data protection rules, 2018.

[Dall'Asen *et al.*, 2021] Nicola Dall'Asen, Yiming Wang, Hao Tang, Luca Zanella, e Elisa Ricci. Graph-based generative face anonymisation with pose preservation, 2021.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, e Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[Kaneko *et al.*, 2019] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, e Nobukatsu Hojo. StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion. In *Proc. Interspeech 2019*, pages 679–683, 2019.

[Khan *et al.*, 2017] Zaheer Khan, Zeeshan Pervez, e Abdul Ghafoor Abbasi. Towards a secure service provisioning framework in a smart city environment. *Future Generation Computer Systems*, 77:112–135, 2017.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, e Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020.

[Liu *et al.*, 2021] Yufei Liu, Chengzhu Yu, Wang Shuai, Zhenchuan Yang, Yang Chao, e Weibin Zhang. Non-Parallel Any-to-Many Voice Conversion by Replacing Speaker Statistics. In *Proc. Interspeech 2021*, pages 1369–1373, 2021.

[Luo e Mesgarani, 2019] Yi Luo e Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.

[Macartney e Weyde, 2018] Craig Macartney e Tillman Weyde. Improved speech enhancement with the wave-u-net. *CoRR*, abs/1811.11307, 2018.

[Paissan *et al.*, 2021] Francesco Paissan, Alberto Ancilotto, e Elisabetta Farella. Phinets: a scalable backbone for low-power AI at the edge. *CoRR*, abs/2110.00337, 2021.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, e Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. pages 1–1, 07 2019.

[Ross e Othman, 2010] Arun Ross e Asem Othman. Visual cryptography for biometric privacy. *IEEE transactions on information forensics and security*, 6(1):70–81, 2010.

[Venkatesh *et al.*, 2021] Ganesh Venkatesh, Alagappan Valliappan, Jay Mahadeokar, Yuan Shangguan, Christian Fuegen, Michael L. Seltzer, e Vikas Chandra. Memory-efficient speech recognition on smart devices. *CoRR*, abs/2102.11531, 2021.

[Yeom *et al.*, 2021] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Alexander Binder, Simon Wiedemann, Klaus-Robert Müller, e Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, 115:107899, 2021.

[Zhang *et al.*, 2017] Kuan Zhang, Jianbing Ni, Kan Yang, Xiaohui Liang, Ju Ren, e Xuemin Sherman Shen. Security and privacy in smart city applications: Challenges and solutions. *IEEE Communications Magazine*, 55(1):122–129, 2017.