MIDAS

Publik rapport

Författare: Martin Torstensson, Felix Rosberg, Yury Tarakanov, Boris Durán, Thanh Bui
Datum: 30-04-2023
Projekt inom Elektronik, mjukvara och kommunikation

TABLE OF CONTENTS

1 Sammanfattning

Det finns ett behov av att samla in stora mängder data för att stödja träning och validering av datahungriga modeller för maskininlärning. Alla de uppgifter som samlas in har gett upphov till oro när det gäller integritetsfrågor och lett till införandet av bestämmelser som GDPR. Förväntningarna på integritetsskyddande tekniska lösningar står i konflikt med behovet av högupplösta data om människors och fordons rörelser, vilket blir ett hinder för teknikutvecklingen i Sverige och Europa. Även om det är viktigt att möjliggöra datainsamling bör det ske på ett integritetsskyddande sätt.

Därför har detta projekt fokuserat på att utveckla anonymiseringsmetoder som gör det möjligt för industrin och den offentliga sektorn att samla in högupplösta data på ett socialt hållbart sätt genom att använda innovativa anonymiseringsalgoritmer som inte försämrar dataupplösningen. Syftet med dessa metoder är att ersätta den borttagna personliga informationen med datorgenererad information så att bilderna ser naturalistiska ut och att icke-personlig information, t.ex. blickriktning och ansiktsuttryck, bibehålls.

Arbetet har främst inriktats på anonymisering av ansikten, eftersom detta har ansetts vara den viktigaste delen, vilket ligger i linje med mer traditionella anonymiseringstekniker som innebär att ansikten pixlas eller suddas ut i bilder. Metoder för att anonymisera registreringsskyltar har också behandlats på grund av möjligheten att tydligt identifiera ägaren till ett fordon baserat på registreringsskyltarna. Dessutom har en serie workshops om GDPR och de krav som den ställer på algoritmerna hållits tillsammans med en juridisk expert. Fokus låg på att fastställa vad som måste anonymiseras för att algoritmen ska uppfylla GDPR, vilket ledde till en riktning mot fullständig kroppsanonymisering.

Som ett resultat av projektet har en ny toppmodern algoritm kallad FaceDancer skapats, som visar ett stort löfte när det gäller anonymisering. Båda exemplen som visar upp prestanda och utvärderingsresultat presenteras mer i detalj i rapporten och vidare i det officiella dokumentet.

Flera artiklar har publicerats om forskningen, bland annat två som publicerats i IEEE. Dessutom har många presentationer hållits för att sprida både på konferenser och i möten med industrin för att ytterligare sprida resultaten och påverka både forskarsamhället och industrin. Dessutom är deltagande företag små och medelstora företag och med datainsamling och dataanalys, vilket har ökat expertisen på området, vilket kan förbättra tekniken och framtida affärsmöjligheter.

2 Executive summary in English

There is a need for collection of large amounts of data to support the training and validation of data hungry machine learning models. All the data being collected has raised concerns regarding privacy and led to the introduction of regulations such as GDPR. The expectations on integrity protecting technical solutions conflict with the need for high resolution data of people and vehicles movements, which is becoming a hindrance for the technology development in Sweden and Europe. While it is important to allow data collection it should be done in an integrity protective way.

Therefore, this project has focused on developing anonymization methods to allow industry and public sector to collect high resolution data in a socially sustainable way by utilizing innovative anonymization algorithms that does not degrade data resolution. These methods aim to replace the removed personal information with computer generated information leaving images looking naturalistic and maintaining non-personal information, such as gaze direction and facial expressions.

The main focus of the work has been around anonymization of faces as this has been deemed as the most vital part, which is in line with more traditional anonymization techniques of pixelating or blurring faces in images. Methods to anonymize license plates have also been covered due to the ability to clearly identify the owner of a vehicle based on the plates. Furthermore, a series of workshops regarding GDPR and the requirements it poses on the algorithms have been held together with a legal expert. The focus was on determining what needs to be anonymized in order for the algorithm to comply with GDPR, which lead to a direction of full body anonymization.

As a result of the project a new state-of-art algorithm called FaceDancer has been created, which shows a great promise regarding anonymization. Both examples showcasing performance and evaluation results are presented in greater detail in the report and further in the official paper.

Several papers have been published regarding the research including two published in IEEE. Also, many presentations have been held to spread both in conferences and in meetings with the industry to further spread the results and impact the both the research community and industry. Additionally, participating companies are SME and with data collection and data analysis, which has increased the expertise level in the area, which can improve the technology and future business opportunities.

3 Background

3.1 State-of-the-Art

Artificial Intelligence (AI) and specifically machine learning (ML) are enablers for a number of traffic safety enhancing functions e.g. automated vehicles, active safety and connected infrastructure. AI and ML can also be used to understand how infrastructure is used by road users and what behavior patterns exist. The advantage of ML is that computers can be trained to recognize patterns, dangers and obstacles based on images of real events in real traffic. To train models, images of the environment are often required and there is a risk that the data collected will contain information covered by the GDPR[1]. ML networks, in particular Deep Neural Networks (DNN), are very effective in recognizing objects in images. They can be trained "end-to-end" i.e. without the help of an expert, the DNN can understand what is relevant in an image on its own. In traditional image analysis, filters of different kinds always had to be specially designed, e.g. edge filters, color and pattern filters etc. which required one as a programmer to always know what and how large the objects he/she were looking for. Therefore, in a traffic environment where new unknown objects are constantly appearing, a DNN-type ML network is better suited. For example: A vehicle with a forward-facing camera collects data on a street in a residential area to be able to practice recognizing objects in traffic. There, both adults and children may appear, and there are cars parked along the road. All this data is necessary for an ML network to learn about how people move near roads, how we interact with vehicles and how other vehicles look and move in relation to each other. But in the saved data there will be faces, clothes and number plates that are personal data. The immediate solution is to anonymize images by blurring faces and number plates, or placing a box over faces and number plates, thus securely storing data without being affected by GDPR. This solves the GDPR problem. Unfortunately, the data becomes useless because much of the information that ML algorithms can use is lost, e.g. facial expressions and attention indicators -eye contact is important for building trust and safe traffic environments[1]. In addition, the ML network risks using blurred fields or boxes in the image to recognize people and cars. Another more studied alternative is to replace the faces with a standard face and the number plates with a standard sign. Unfortunately, it doesn't work either: then the ML system will learn to recognize the exact standard face and standard sign. The purpose of this project is to investigate the possibility of creating anonymized but unique faces and number plates in video data to replace personal data in pictures. Thus, as much as possible of the real environment and interactions are retained in data collected in road safety-related research projects.

3.2 A short introduction to Generative Models

Generative models are one of the most promising approaches for endowing machines to analyze and make sense of the world that surrounds us. To train a generative model we first collect a large amount of data in some domain (e.g., think millions of images, sentences, or sounds, etc.) and then train a model to generate data like it. Generative models have the number of parameters significantly smaller than the amount of data we train them on, so the models are forced to discover and efficiently internalize the essence of the data in order to generate it.

[1] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*

A more formal definition

Given a set of $x_1, \ldots, x_n$ as samples from a true data distribution $p(x)$. In the example image below, the blue region shows the part of the image space that, with a high probability (over some threshold) contains real images, and black dots indicate our data points (each is one image in our dataset). Now, our model also describes a distribution $p_\theta(x)$ (green) that is defined implicitly by taking points from a unit Gaussian distribution (red) and mapping them through a (deterministic) neural network — our generative model (yellow). Our network is a function with parameters θ, and tweaking these parameters will tweak the generated distribution of images. Our goal then is to find parameters θ that produce a distribution that closely matches the true data distribution (for example, by having a small Kullback–Leibler divergence loss). Therefore, one can imagine the green distribution starting out random and then the training process iteratively changing the parameters θ to stretch and squeeze it to better match the blue distribution.

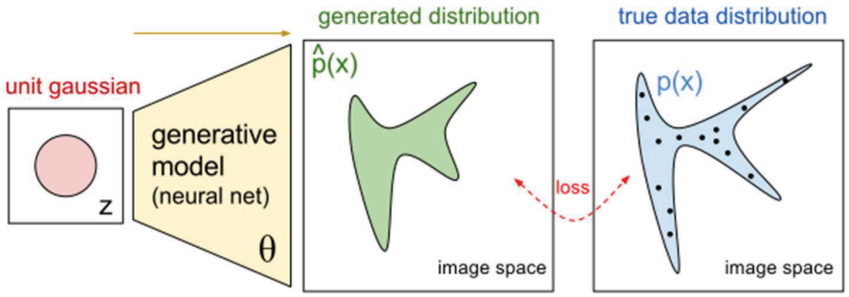

Figure 1: Diagram of how generative models learn and approximate a given data distribution

Generative algorithms learn a model of the joint probability, $p(x, y)$, of the inputs x and the label y, and make their predictions by using Bayes rules to calculate $p(y \parallel x)$, and then picking the most likely label y. Discriminative classifiers model the posterior $p(y \parallel x)$ directly or learn a direct map from inputs x to the class labels[2].
Discriminative algorithms map features to labels and they are concerned solely with that correlation. One way to think about generative algorithms is that they do the opposite. Instead of predicting a label given certain features, they attempt to predict features given a certain label.
The lists below present some of the most used generative and discriminative models in state-of-art algorithms:
Generative models

- Gaussian mixture model (and other types of mixture model)
- Hidden Markov model Probabilistic context-free grammar
- Bayesian network (e.g. Naive bayes, Autoregressive model)
- Averaged one-dependence estimators
- Latent Dirichlet allocation Boltzmann machine (e.g. Restricted Boltzmann machine, Deep belief network)
- Variational autoencoder
- Generative adversarial network
- Flow-based generative model
- Energy based model

Discriminative models

- k-nearest neighbors' algorithm
- Logistic regression
- Support Vector Machines
- Maximum-entropy Markov models
- Conditional random fields
- Neural networks

3.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were introduced first by Ian Goodfellow et al.[3]. In 2016, Yann LeCun described[2] adversarial training as "the most interesting idea in the last ten years". GANs are considered as the new players in the unsupervised learning arena and from the very beginning they have been able to achieve far better performance compared to traditional nets.

3.3.1 Architecture

At their core GANs are composed of two independent neural networks that act as adversaries competing continuously against each other. The first neural net is called the Discriminator (D) and is the one that has to undergo training. D is the classifier that will do the heavy lifting during the normal operation once the training is complete. The second network is called the Generator (G) and is tasked to generate random samples that resemble as close as possible to those inside the training space of D, Figure 2.

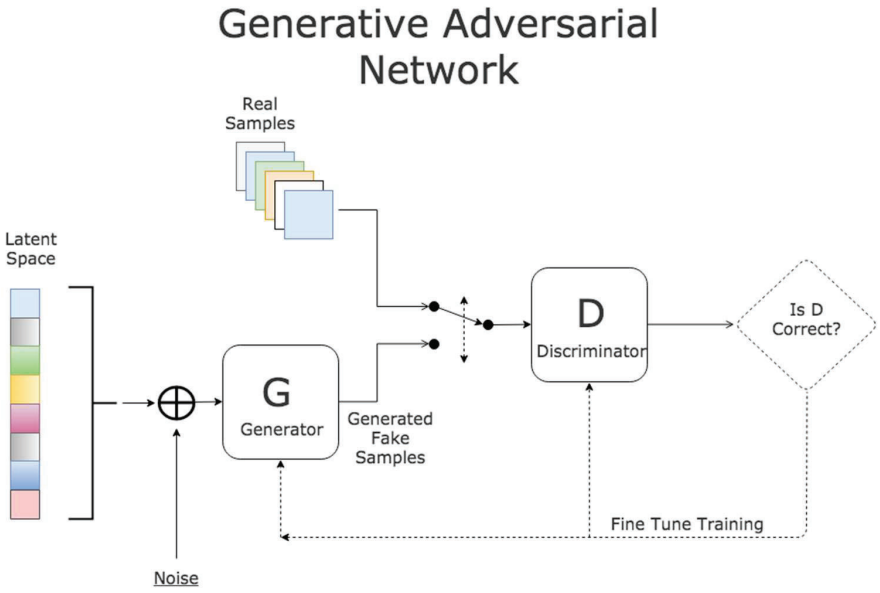

Figure 2: GAN architecture

One can think of a GAN as the opposition of a counterfeiter and a cop in a game of cat and mouse, where the counterfeiter is learning to pass false notes, and the cop is learning to detect them. Both are dynamic, i.e. the cop is in training, too (to extend the analogy, maybe the central bank is flagging bills that slipped through), and each side comes to learn the other's methods in a constant escalation.

During the training process, weights and biases are adjusted through backpropagation until the discriminator learns to distinguish real images of shoes from fake images. The generator gets feedback from the discriminator and uses it to produce images that are more 'real'. The discriminator network is a convolutional neural network that classifies the images as either fake or real. The generator produces new images through a de-convolutional neural network.

3.3.2 *Problems with GANs*

- Hard to achieve Nash Equilibrium: GANs represent a two-player non-cooperative game where each player updates its cost independently. This cannot guarantee a convergence.
- Low dimensional supports:
- Vanishing gradients: If the discriminator is bad then the generator does not have accurate feedback and it cannot represent reality. If the discriminator is too good, then the learning process becomes too slow or jammed.
- Mode collapse: The generator collapses and always produces the same outputs.

[2] *https://www.quora.com/What-are-some-recent-and-potentially-upcoming-breakthroughs-in-deep-learning*

- Lack of proper evaluation metrics: There is no good objective function that informs about the training progress.

3.3.3 GAN variants

Current GAN research focuses on how to improve their training process and on finding and improving the deployment of GANs in real-world applications. The main goals of improving the training process of this type of networks can be summarized in: (1) improving the diversity of generated images, (2) increasing the quality of the generated images, and (3) achieving a more stable training towards Nash equilibrium, [4].
Research on GANs has witnessed exponential growth during the last five years reporting dozens of new GAN-variants and GAN-related applications. The list below presents some of the most relevant examples of state-of-the-art GAN-variants attempting to improve the performance of this approach. These examples are grouped according to those working on different architectures and those working on improving the loss functions.

Architecture-variant GANs

- FCGAN[3] (Fully connected GAN)
- LAPGAN[5] (Laplacian Pyramid of Adversarial Networks)
- DCGAN[6] (Deep Convolutional GAN)
- BEGAN[7] (Boundary Equilibrium GAN)
- PROGAN[8] (Progressive GAN)
- SAGAN[9] (Self-Attention GAN)
- BigGAN [10]

Loss-function GANs

- WGAN[11] (Wasserstein GAN)
- WGAN-GP[12] (Wasserstein GAN with Gradient Penalty)
- LSGAN[13] (Least Square GAN) f-GAN (f-divergence GAN)
- f-GAN[14] (f-divergence GAN)
- UGAN[15] (Unrolled GAN)
- LS-GAN[16] (Loss Sensitive GAN)
- MR-GAN[17] (Mode Regularized GAN)
- Geometric GAN [18]
- RGAN[19] (Relativistic GAN)
- SN-GAN[20] (Spectral Normalization GAN)

A much larger list with tons of GAN flavors can be found at The GAN Zoo[21]

3.4 StyleGANs

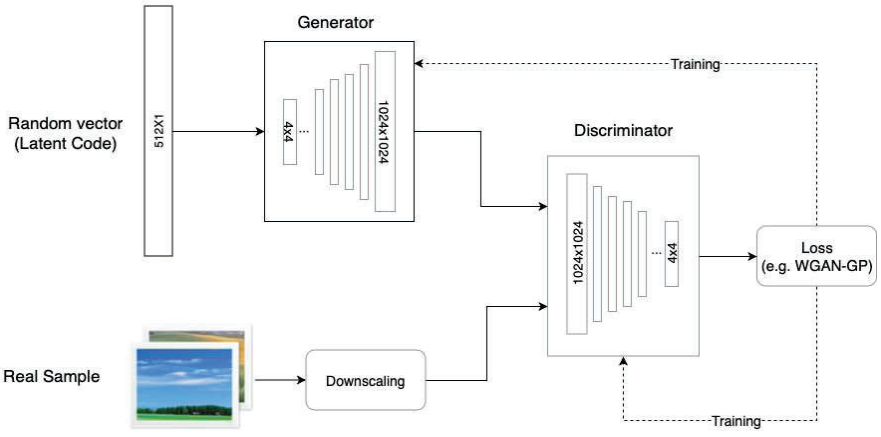

Figure 3: Overview of a PROGAN model.

The StyleGAN[22] model is arguably the state-of-the-art in the generation of high-quality images and is based on a previous model called PROGAN[8], Figure 3. The main features of this approach are the ability of generating high-resolution images using Progressive Growing and the ability of incorporating image styles into each layer using a normalization method for style transfer called AdaIN.

3.4.1 *StyleGAN v1*

The improvements made to the PROGAN model are constrained to the generator network. One of the main changes is the addition of a Mapping Network which task is to create an intermediate latent vector which in turn is fed to an also modified Generator network which is now labeled Synthesis Network, Figure 4.

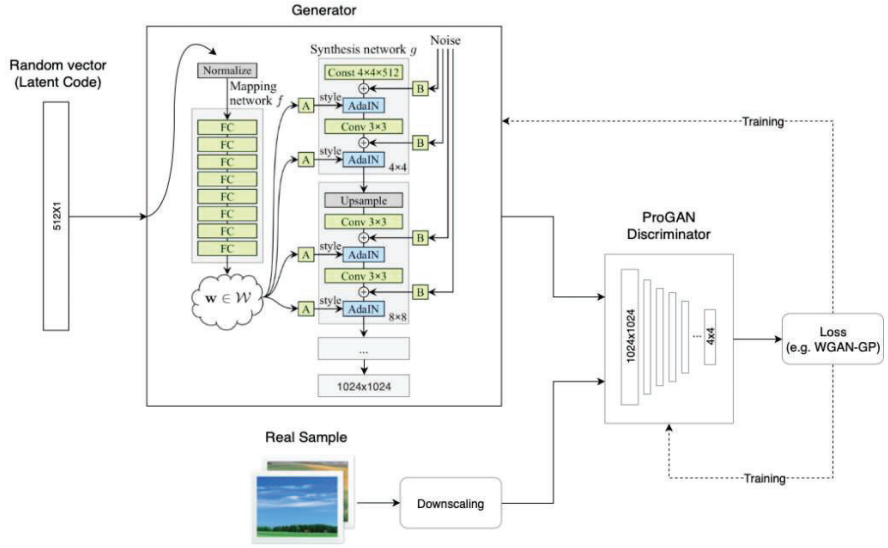

Figure 4: Overview of a StyleGANv1 model.

The Mapping Network consists of 8 fully connected layers and its output is of the same size as the input layer (512×1). The different elements in the intermediate vector between the Mapping and the Synthesis networks allow the user to control different visual features. The lower the layer (and the resolution), the coarser the features it affects: Coarse - resolution of up to 82 - affects pose, general hair style, face shape, etc. Middle - resolution of 162 to 322 - affects finer facial features, hair style, eyes open/closed, etc. Fine - resolution of 642 to 10242 - affects color scheme (eye, hair and skin) and micro features. The main problems observed in most of the images generated

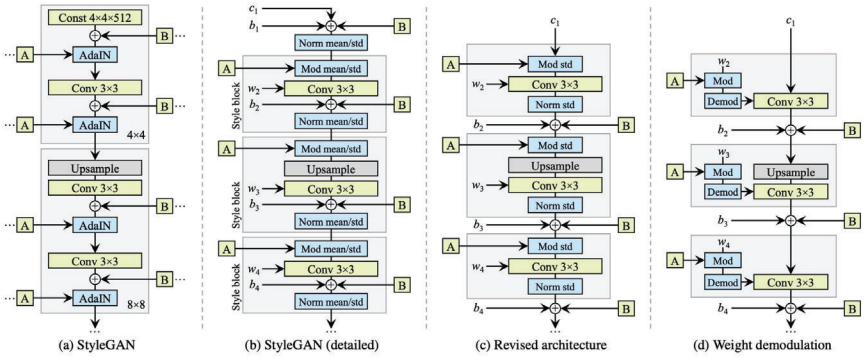

Figure 5: Style blocks for the StyleGAN models (original and improvement). (a) The simplified AdaIN block. (b) An expanded view of the AdaIN operations. (c) The operations proposed for improvement in StyleGANv2. (d) The revised architecture replacing instance normalization with a 'demodulation' operation. Source: https://arxiv.org/pdf/1912.04958.pdf

with StyleGANv1 are easily seen in the form of "water dropplets"-like noise. The authors believe these problems arise from the use of the normalization layer (AdaIN).

3.4.2 *StyleGAN v2*

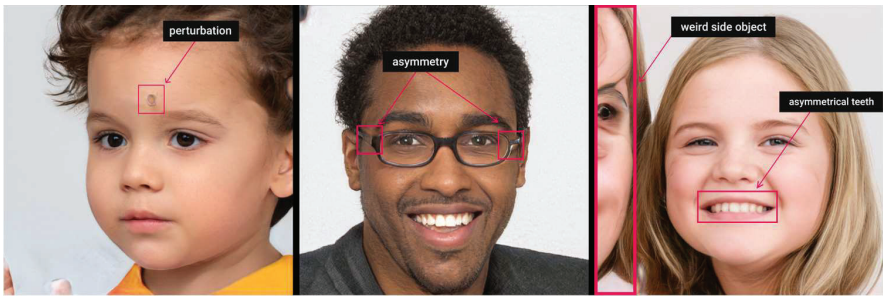

Figure 6: Examples of unwanted features in the images generated from StyleGANv1. Source: https://medium.com/towards-artificial-intelligence/stylegan-generated-face-classification-with-resnexts-19535ed1d91d

After careful observation the authors realized that the normalization layer (AdaIN) was creating several unwanted features in many of the generated images such as water-kind of droplets, asymmetry effects on the eyes and teeth, the inclusion of weird objects, non-natural hair styles and synthetic skin. Some of these examples can be seen in Figure 6.

In order to fix these problems, the authors decided to use a normalization of the convolution weights using estimated statistics rather than actual data statistics. Figure 5 shows a comparison of the style blocks used in StyleGANv1 and StyleGANv2.

3.5 FSGAN

The Face Swapping GAN[23] (FSGAN) is a deep learning-based approach for face swapping and reenactment in images and videos. The authors consider their model to be *subject agnostic* in the sense that the network does not need to be trained on images of a specific subject in order to accomplish the final results.
FSGAN consists of three main components. The first, (Figure 7 (a)), consists of a reenactment generator G_r and a segmentation CNN G_s. G_r is given a heatmaps encoding the facial landmarks of F_t, and generates the reenacted image I_r, such that F_r depicts F_s at the same pose and expression of F_t. It also computes S_r: the segmentation mask of F_r. Component G_s computes the face and hair segmentations of F_t.

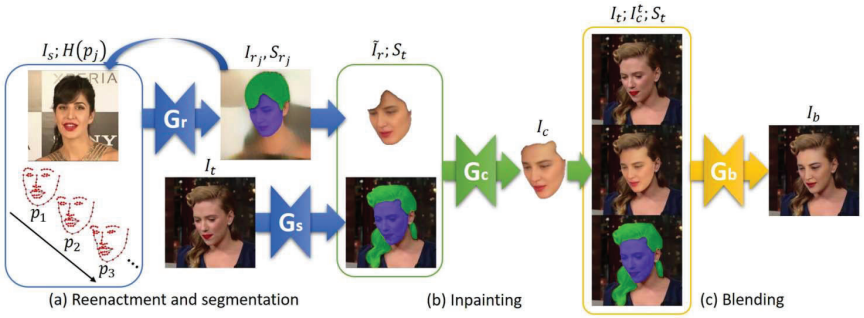

Figure 7: Overview of the proposed FSGAN approach. (a) The recurrent reenactment generator G_r and the segmentation generator G_s. G_r estimates the reenacted face F_r and its segmentation S_r, while G_s estimates the face and hair segmentation mask S_t of the target image I_t. (b) The inpainting generator G_c inpaints the missing parts of F_r based on S_t to estimate the complete reenacted face F_c. (c) The blending generator G_b blends F_c and F_t, using the segmentation mask S_t.

The reenacted image, I_r, may contain missing face parts, as illustrated in Figure 7(a) and (b). We therefore apply the face inpainting network, G_c using the segmentation S_t, to estimate the missing pixels. The final part of the FSGAN, shown in Figure 7 (c), is the blending of the completed face F_c into the target image I_t to derive the final face swapping result.

The face segmentation network, G_s, is based on U-Net[24], with bilinear interpolation for up sampling. All the other generators— G_r, G_c, and G_b — are based on those used by pix2pixHD[25], with coarse to-fine generators and multi-scale discriminators.

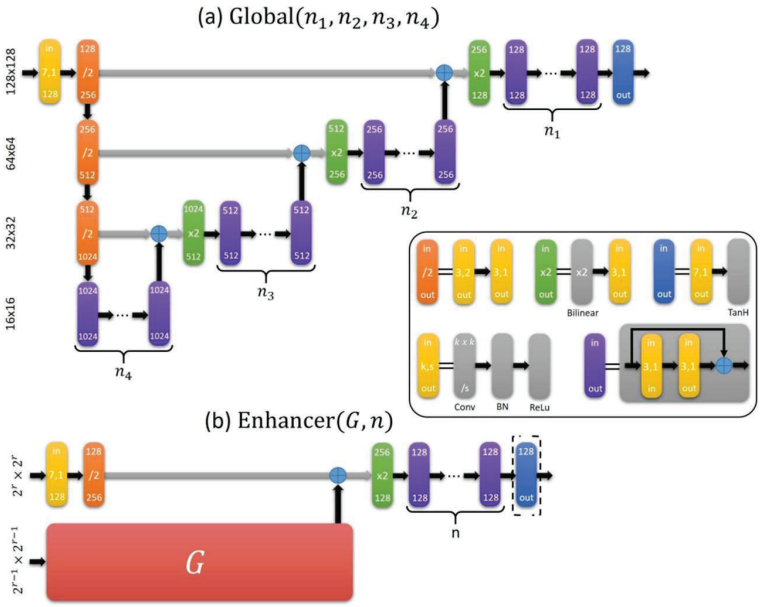

Figure 8: Generator architectures. (a) The global generator is based on a residual variant of the U-Net CNN, using a number of bottleneck layers per resolution. We replace the simple convolutions with bottleneck blocks (in purple), the concatenation with summation (plus sign), and the deconvolutions with bilinear up-sampling following by a convolution. (b) The enhancer utilizes a submodule and a number of bottleneck layers. The last output block (in blue) is only used in the enhancer of the finest resolution.

4 Purpose, research questions and methods

The MIDAS Project intends to bridge the research gap around methods that allow cameras, primarily needed for machine learning and active safety purposes, to operate in a non-intrusive and GDPR compliant way. To achieve this an anonymization method was developed that can automatically remove information from an image and replace it was a computer-generated alternative. Also, it is able to keep certain non-personal information like gaze direction and facial expressions. Three research questions were set up:

Research questions:

- RQ1: How can the generated face be maintained between frames in one video sequence?
- RQ2: How can we make sure that the generated face in one video sequence is different (unique)compared with the next time someone is captured in the video?
- RQ3: What measures can we use to guarantee that we have anonymized all personal details in a video frame?

5 Goal

This project enables the development of safe, efficient and environmentally friendly vehicles by facilitating safe and secure data collection from a GDPR perspective, which is key knowledge to be able to train machine learning algorithms that support the functionality in e.g. automated vehicles, active safety and connected infrastructure.

6 Results and Deliverables

6.1 Licentiate thesis from a PhD student

The PhD studies of Felix Rosberg have been supported by the project and his licentiate examination is preliminarily set to 31-10-2023.

6.2 Additional cases to anonymize

This chapter covers a combination of the deliverables:

- A report on possible additional cases to incorporate in the anonymizing process and possible tools/methods to use to handle them.
- Object detection
- Masking of objects with sensitive information

In order to evaluate what sources of information could result in a person being identified and thereby be considered personal information from the eyes of the GDPR regulation an initiative was started to arrange a series of workshops within the consortium. These workshops focused on discussions of what information could be used directly or indirectly to identify someone based on image data and how this relates to the algorithms we were developing. While the standard for most anonymization is mainly around faces and license plates, there are many other factors that could be used and in special cases identify someone. To facilitate the discussions and allow it to be concrete rather than just speculation a legal expert was participating, both to steer the discussion and to provide material.

Based on these discussions a list of potential identification sources was crafted:

- Eyes
- Tattoos
- Birth marks
- Hair style/color
- Clothing (Especially unique styles)
- Equipment
- Gait
- Scars
- Etc

From this list, which likely is only a portion of all sources, it is quickly made clear that while face and license plate anonymization is a good start it is not sufficient for a functional anonymization system. All the different sources are, however, not easy to handle individually. Firstly, his is due to the number of sources that need to be considered and also it is hard to motivate that all of them have individually been covered. Secondly, it is also a large challenge to successfully be able to cover some of these individually. Tattoo anonymization can be used as an example. There is a large variation of tattoos that are unique enough to identify a person and the datasets for this topic are rare and small. This makes it hard to create and train a model capable of detecting them let alone anonymizing them. Moreover, there is also a constant influx of new tattoos that could not possibly be covered by an already collected dataset. As a result, it would not be a feasible task.

As a conclusion from these discussions, there is an argument that individual anonymization is not possible to accomplish and would not provide full anonymization. Instead, the focus should be shifted towards full body

anonymization as it has an ability to cover both all detected issues as well as others that have not been considered during these discussions.

There is a recent work[26] looking into full-body anonymization using inpainting GANs. It is however limited in a temporal context, meaning each frame will produce a different body for the same person. A potential direction for dealing with this is using diffusion models[27] conditioned on previous frames. Diffusion models have shown to have exceptional expressive power, making them strong candidates for dealing with and controlling the complexity of realistic full-bodies. The drawback being inference time.

It would be possible to use a similar approach to the face anonymization method in full body anonymization on a high level. However, we expect that the underlying method must be noticeably tweaked. The expected main challenge with this task would be to create a proper encoding for the full body. Unlike faces, full body anonymization suffers from a greater variance due to e.g. clothing and body posture. These effects make it harder to achieve good training of the encoder network, which could result in a poor final result. This could potentially be alleviated with the help of a segmentation network capable of mapping the body up into separate parts, which could then be encoded individually.

Full body anonymization requires as a first step the detection and segmentation of human bodies in an image. A short study on the feasibility of this part was tested on surveillance video provided by Viscando. One of the main constraints for this study was the time required to compute a segmentation mask on each frame of the video stream. We tested a model called MoveNet[28] due to its high speed and accuracy on estimating body poses. The recommended implementation was slightly modified to create full-body masks on top of the people detected on each frame. The next step is to insert an artificially generated human body with the same pose into the provided mask.

Figure 9: Screenshot of a single frame from a surveillance video provided by Viscando. To the left, the original frame and to the right the automatic segmentation of human bodies in real-time.

6.3 Initial data collection

This iteration is mainly to allow development of object detection and recognition.

6.3.1 Viscando 3D&AI technology for traffic data collection and analysis

Viscando AB, a start-up from Gothenburg and a partner in MIDAS project, is specializing in collection of detailed traffic data and advanced traffic analysis – providing actionable insights on traffic flow and safety risks, real-time traffic control, as well as scenarios and reaction models for safer and more human-like automated vehicles.

This is enabled by Viscando in-house developed, patented infrastructure sensor OTUS3D (Figure 10). This sensor, based on 3D vision and artificial intelligence, detects, tracks and classifies all road users in its field of view, yielding accurate trajectories of different types of road users (pedestrians, 2-wheelers, light and heavy vehicles) sampled 10-20 times per second. Grids containing multiple sensors are used in measurements over larger areas.

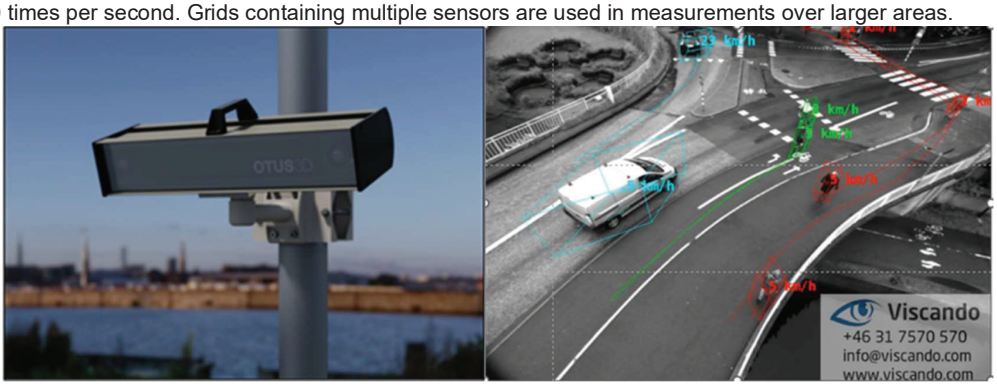

Figure 10: Left: a photo of Viscando OTUS3D sensor installed on a pole. Right: example of tracking of different road users in an urban environment.

The toolchain for the sensor data processing is depicted in Figure 11. In commercial operations, the video data is processed in the embedded computational unit and is removed directly after the classification step – that is, within 20 ms from being captured. Only the trajectories which are completely anonymous, are stored. Thus, personal data constituted by video, is neither stored nor transmitted.

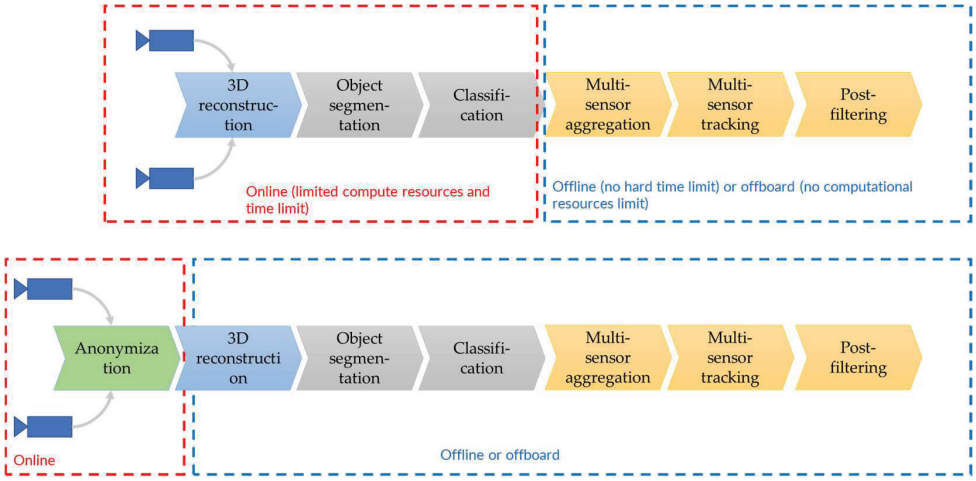

Figure 11: Current processing toolchain for Viscando (top) and envisioned toolchain enabled by online anonymization of video streams. The majority of data processing would be possible to run offline/offboard reducing the limitations of computation time and resources.

Possibility to store and share video, if anonymity can be insured, would be beneficial not only for Viscando but for all the customers and partners in traffic safety work, in at least foura ways:

1. Interesting or anomalous events identified in trajectory data could be possible to analyze using video. This would lead to more accurate metrics and insights extracted from the trajectories.
2. Certain metadata that is so far not possible to extract automatically with sufficient reliability, could be extracted and used for analytics and important research, such as: qualitative assessment of risk and perceived safety, attention, detailed interactions, eye contact etc.
3. New AI algorithms could be trained and tested on massive datasets - not only by Viscando but by expert research institutions like RISE.
4. Data processing could be done offline at Viscando servers of on cloud, reducing the need for embedded processing power. This would reduce the cost of sensing solutions and increase the quality of data (offboard processing allows for more complex, computationally heavy algorithms yielding better results)

Thus, the value of real-time anonymization without losing important features in behaviors and interactions between road users, which is the goal of MIDAS project, may lead to large benefits in traffic safety.
Viscando has both followed up in the development of anonymization algorithms by project partners, and provided datasets for training of these algorithms, both within and beyond the project scope. A certain amount of data is made public via a sharepoint (see chapter 6.9 for more information) for supporting the further work in broader scientific community. These datasets are described below.

6.3.2 3D-based anonymization technique

Viscando has developed its own simplified algorithm for data anonymization, utilizing stereovision algorithms. The first step of the stereovision algorithm searches the right and left images to find pairs of matching points, to calculate disparity (shift in the pixel index) and thus distance to the pixel. The points on the right video for which the algorithm has found a pair with a high confidence in the left one, are preserved in the anonymized video, and the rest are made black. Thus, a monochrome video with bright points corresponding to the most distinct 3D features (typically edges of objects) is generated. A snapshot from such video is shown alongside initial video in Figure 12.

Figure 12: A snapshot of video anonymized using 3D-based anonymization technique described in section 6.6.2. Left image shows original frame, right image the anonymized one.

Such anonymization has proven sufficient for visual confirmation of types of road users and to some extent their interactions (points 1 & 2 in the list above). However, desired features like gaze direction of people, are not possible to identify or use for training/testing of AI based detection tools.

6.3.3 *Dataset for initial face detection and recognition*

To collect personal information, in our case video of people, an explicit agreement from all persons participating is required. To ensure this, Viscando has performed a tailored collection of a small video dataset involving only Viscando personnel who gave their permission for this. The collection was done in a street setting that is similar to a typical location for a customer project, with a mixed traffic of pedestrians, vehicles and two-wheelers. One modification was to set the sensor at lower height (3 meters instead of typical height 5-8 meters) to allow for higher detail level of people on video, making it easier for initial algorithms to perform detection and anonymization. In MIDAS, face detection and swapping were in focus; the data allows also for full-body detection and swapping which will be in the scope of following projects.

Figure 13: Example from the initial video collected by Viscando, and the anonymization steps: segmentation of faces (bottom left) and face swap (bottom right).

6.4 Second data collection

In addition to the dataset described in the previous section, Viscando has extended certain customer projects to include collection of the data that can help in further development of anonymization algorithms.

Figure 14: A snapshot from the video sequence from a traffic measurement in Aarau which Viscando provides for open access in the scope of MIDAS project.

As several of the customer projects have been undertaken in Switzerland where the new data protection law, nFADP[29] (new federal act on data protection) has not yet come into force, it was possible to collect video data for research and development purposes in these projects. In total, over 50 hours of video were collected for the purpose of anonymization and AI algorithm development. A shorter sequence from a dataset collected as a part of a project in the city of Aarau is provided for open access in the scope of MIDAS project. Figure 14 provides a snapshot of the video.

It is worth mentioning that the resolution of the video and the sensor placement make the detail level of the faces of persons very low, so one may question whether these videos constitute personal data. However, many actors take a conservative standpoint in the question and consider the video where faces are visible (even with low detail level) personal data. Hence, there is still need for anonymization.
It was discovered during processing of the dataset using algorithms developed within MIDAS that the size and resolution of faces was too low for reliable detection and swapping of faces. Therefore, more development is needed for this type of data – which this open dataset contributes to.

In the projects within EU, GDPR ruled out the possibility of collecting video data. Viscando has collected 3D-based anonymized video data (described in section 6.6.2 above) in several projects to evaluate possibilities and gaps of this “secure but lossy” anonymization technique. It was confirmed that the presence and type of road users, as well as conflicts between them, are possible to confirm in this way; while more detailed information like sight direction, presence of helmet or pedaling for bicycle riders, gender and age of people are absolutely missing in the anonymized data. At the same time, these are often lifted as desirable or even necessary for better understanding of traffic and more directed actions towards increased traffic safety.

6.5 Face detection

The face detector of choice is an implementation of RetinaFace[30]. RetinaFace maintains state-of-the-art performance in accuracy, with a face inference time. The authors report 40 FPS at 4K resolution when deployed on a GPU (NVIDIA Tesla P40), 20 FPS at HD resolution on a multi-thread CPU (Intel i7-6700K) and 60 FPS at a 640x480 resolution on a single thread CPU. But what makes RetinaFace particularly useful for our face is the multi-task capabilities. The idea is to not only predict the location of the face, but also regress facial landmarks and a dense face mesh. Current implementation allows for a 5-point facial landmark prediction. These landmarks are then used to align the face so that the location of the landmarks is almost always in the same location before any manipulation or identity extraction. Both the facial recognition model and the face swapping model require aligned faces.

6.6 License plate detection

License plate detection (and recognition) is an integral part of current intelligent systems that are developed to locate and identify various objects. Machine learning algorithms have been used for license plate detection and recognition; however, it remains a challenging task due to various factors such as numerous shapes and designs of license plates, non-following of standard license plate templates, irregular outlines, angle variations, and occlusion, etc.
The history of license plate detection and recognition is long and has been evolving hand-in-hand with the development of artificial intelligence and even more rapidly with the advances in machine learning during the last decade. There are currently several private companies that offer among their products and services different solutions for license plate detection and recognition with high accuracy. It is important to remember that most of the proposed solutions are constrained to cameras fixed in the infrastructure which simplify greatly the challenges of angles, occlusions, light conditions, etc. The applications studied in MIDAS should include images taken from the point of view of moving vehicles and that alone represents a much more complex challenge assuming there is enough data for training the algorithms. Unfortunately, the datasets being used for training such systems are out of reach for public research like the one proposed in this project. Therefore, we decided to create a simple solution with a limited amount of data as a proof-of-concept for later improvement.

6.7 Masking of faces

This chapter covers the combination of the deliverables:

- Masking of faces
- GANs for anonymizing
- Demonstrator on a mobile device

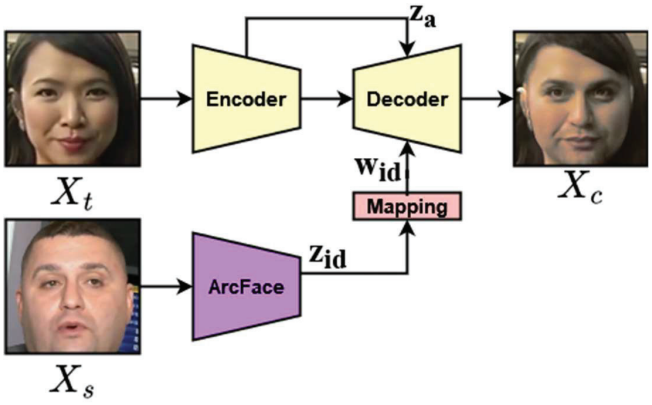

Figure 15: Simplified overview of the FaceDancer face swapping network. It consists of an encoder that encodes the target face, a facial recognition model that extracts the identity vector z_{id} of a source face, a mapping network that maps the identity vector z_{id} into a vector within a learned distribution space w_{id} and a decoder that reconstructs and manipulate the identity information by conditioning the information on the identity information w_{id}. The decoder also learns where to incorporate unconditioned information from the encoder (z_a).

To allow for anonymization of faces, we developed a face swapping method called FaceDancer[31]. FaceDancer is a target-oriented network architecture (See Figure 15), which means we manipulate the target face inside the network to output a face with same attributes and pose of the target face but with a new identity. The identity information is extracted from a source face, with a state-of-the-art facial recognition model called ArcFace[32]. FaceDancer achieved state-of-the-art performance for identity transfer compared to previous work, and because of how it was trained, it maintained strong pose- and occlusion-awareness without any costly post-processing. FaceDancer is also small, only consuming 1.18 to 1.27 GB of VRAM when face swapping a batch of 32 frames at the same time. The runtime for this batch lies between 64.6 to 78.3 *ms*. This translates to an FPS of 12.77 to 15.48. In the case of a batch size of 32, that means FaceDancer can face swap 409 to 495 images of faces at resolution 256x256 per second. This number is expected to go up as long as the batch size is able to increase, limited only by the VRAM of the GPU.

The run time profiling was done on a NVIDIA RTX 3090 and includes the computation of the facial recognition model. We expect, however, that potential bottlenecks may limit the throughput. Examples of potential bottlenecks are data loading between CPU and GPU, face detection, identity tracking and the process of inserting the manipulated face back into the correct frame which it was taken from.

FaceDancer is also consistent on video without any need for temporal refinement. Detailed results and methods can be found in the publication about FaceDancer (See Section 7.2).

So far, we have described a face swapping framework and will now detail how we can use FaceDancer for anonymizing faces. There are a few approaches one could consider depending on the needs. One is to keep a subset of fake faces generated by for example StyleGAN2[22] or StyleGAN3[33]. Then just sample from these faces to mask the identity of the target. One advantage of FaceDancer being target-oriented and using identity vectors

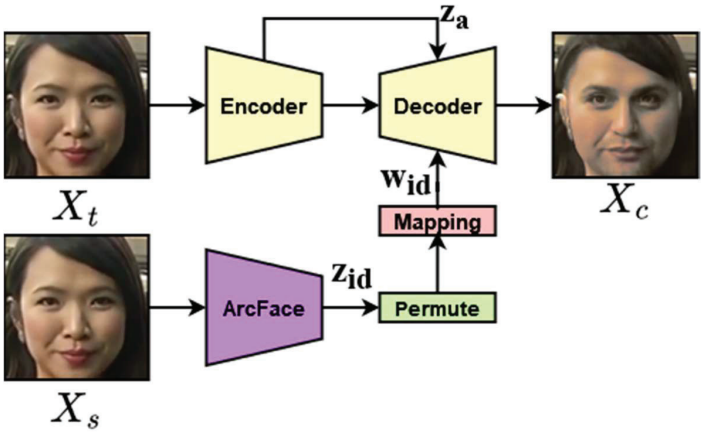

Figure 16: Simplified illustration of how the FaceDancer can be used as an anonymization tool when a identity permutation process is attached in the pipeline.

from a facial recognition model to manipulate the face, is that we do not have to save these faces as images, but instead as the extracted identity vectors. This is a rather naive and simple approach that limits variation.

A second approach is to use the same face as input into FaceDancer and as input into ArcFace. Then before manipulating the face based of the identity extracted from ArcFace (which is currently the same and would just yield a reconstruction), we apply some kind of identity permutation/manipulation on the identity vector (See Figure 16). One such approach is to attach a small Vector Quantized auto encoder[34] (VQAE), trained to reconstruct identity vectors. As seen in Figure 17, VQAEs work by having a learnable codebook where the closest codes if the encoded vector is decoded to a reconstruction. Given a vector z, we retrieve the quantized vector z_q by taking the taking the closest entry in the codebook z_c to a subpart of z, z_i:

$$z_q = \mathbf{q}(z) = \arg\min ||z_i - z_c||$$

After the VQAE is trained, we can invert the code retrieval operation from getting the closest codes to getting the codes that is the farthest away:

$$z_q = \mathbf{q}(z) = \arg\max ||z_i - z_c||$$

This guarantees anonymization that has a large identity distance from the unanonymized face and becomes consistent for if the same face would be anonymized at different locations. Due to the nature of quantization and the identity vector slightly changing from frame to frame, anonymized videos will exhibit jitter in the face. Possible future alleviations are having a temporal refinement network that fixes the inconsistencies frame to frame, tracking

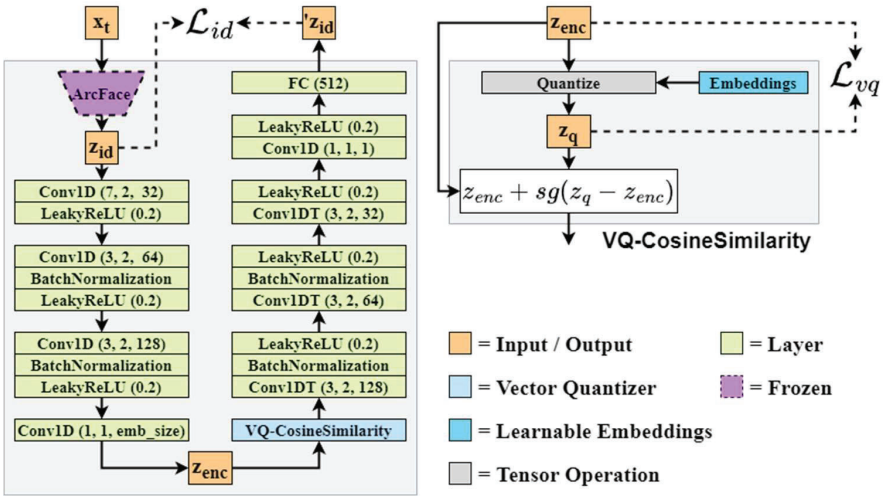

Figure 17: Detailed network overview of the Vector Quantized auto encoder that learns to reconstruct identity vectors z_{id} by retrieving the codes z_q in the codebook (Embeddings) closest to the encoded identity vector z_{enc}.

of which identities have been manipulated with what information recently or a combination.
The performance of this approach is determined in two steps. First the performance on the face swapping method used. This consists generally of identity retrieval, pose error, expression error and visual quality. These metrics are detailed in the FaceDancer paper (See Section 7.2) and the Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning paper (See Section 7.2). Secondly, the performance of the anonymization (Table 1). We did a preliminary study in evaluating anonymization and a paper was published called Towards Privacy Aware Data collection in Traffic: A Proposed Method for Measuring Facial Anonymity (See Section 7.2). This was later extended from naively just measuring the identity distance to a retrieval-based metric as described below. This study also investigated what parts of the human that can reveal the identity of celebrities. In short, features such as hair enabled the participants to identify the celebrity. However, it is important to note that using celebrities introduces unavoidable biases in the study that most likely exaggerate the capabilities of identifying people based of, for example, hair in a real case scenario.

We perform anonymization with a combination of FaceDancer[31] or SimSwap[35] together with the above-described identity permutation approach. We then try to retrieve the original identity in the dataset and count each failure of finding the correct identity as a success. We also evaluate the temporal consistency by computing the standard deviation of the distance between identity vectors, extracted from a facial recognition model CosFace[36], for 10 random frames in a video. Finally, we made preliminary research into the possibility of teaching a separate model to reconstruct the original identity from the anonymized face. We found out that this is possible for both FaceDancer and SimSwap and designed a third metric that quantifies how prone the anonymization is to this reconstruction attack. This is done by anonymizing faces, then pass it to the reconstruction attack model that tries to undo the anonymization and lastly try retrieve the original identity with the help of the facial recognition model CosFace in the data set. We compare this approach with real data and a previous facial anonymization method called DeepPrivacy[37] when possible. For example, identity retrieval and reconstruction attacks does not make sense for real data and due to DeepPrivacy masking out faces entirely it makes no sense to try performing a reconstruction attack. This is also an important aspect that makes this approach better because we do not mask anything and make changes to the face directly, we preserve important attribute information such as expression and eye gaze.

Table 1: Evaluation on real data, DeepPrivacy (CITE), and two face swapping methods (FaceDancer and SimSwap) in conjunction with above permutation method. ID is the identity retrieval metric, RA is identity retrieval after a reconstruction attack and Mtc is the temporal identity consistency.

Method	**ID ↓**	**RA ↓**	**M_{tc} ↓**
Real Data	-	-	0.074
DeepPrivacy	0.012	-	0.183
SimSwap + Permutations	0.002	0.958	0.199
FaceDancer + Permutations	0.020	0.999	0.087

In Figure 18 an example of the anonymization results from the project can be found. In the left side of the image is the original frame and to the right is the anonymized version.

Figure 18: An example of the anonymization results where the original image is to the left and the anonymized version is to the right.

6.8 Masking of license plates

6.8.1 License plate generator

GAN[3] based style transfer methods are able to generate new data that can closely resemble the domain of data used for training. These developments of GAN propose promising approaches for image-to-image translation challenges in computer vision.

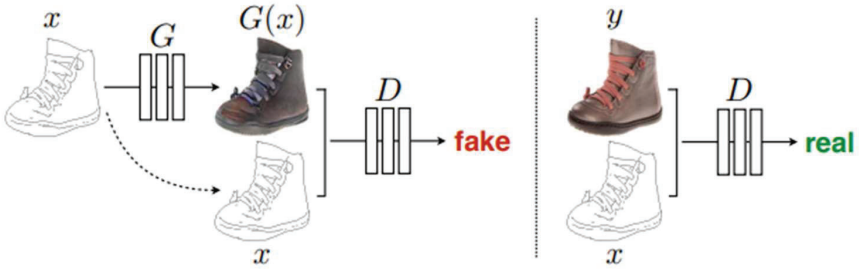

Figure 19: Conditional GAN to map edges - photo. Discriminator learns to classify fake and real {edges, photo} tuples while generator G learns to fool the discminator.(Image courtesy [2])

To generate “realistic” and anonymous license plates, image-to-image translation techniques have been applied as follows:

- Training a conditional GAN pix2pix[38] to map edges -> photo. The training dataset has been created as below:
    - The LP image dataset has first been collected from a public dataset openALPR-EU[39] consisting of 1347 EU license plate images.

- Edge images of the LP images had been generated to create a training dataset consisting of tuples {edges, photo} (Figure 20)
- Pix2pix model has been trained with the tuple dataset (Figure 23, Figure 21). The trained model can then be used to translate from any LP edge image to a GAN-based synthetic LP image.

- Generate Swedish license plate with valid number and template format[40] using a python script LP-Generator. Figure 25 illustrates an example of generated Swedish image by the LP-Generator script.
- Corresponding edges images were then extracted from the LP-Generator images, as the input for the trained pix2pix model (Figure 27).
- The trained pix2pix model will then translate the provided edge image into a realistic synthetic LP (Figure 28).

Figure 20: An example of generated training tuples {edges, photo} of an EU license plate in OpenAPLR dataset

Figure 21: ground truth LP

Figure 22: GAN-based synthetic LP at training epoch 1

Figure 23: ground truth LP

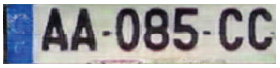

Figure 24: GAN-based synthetic LP at training epoch 25

Figure 25: Transportstyrelsen template for single-row license plates used for most types of vehicles

Figure 26: A Swedish license plate generated from the python script

Figure 27: Edge image of a generated LP image

Figure 28: Pix2pix synthetic LP image

6.8.2 *License plate replacement*

Once the anonymous GAN-based synthetic license plates are available, the next processing step to anonymize LP

Figure 29: Detected LP from the original image using WPOD algorithm.

in the original images (or video frames) is to detect the LPs and replace them with the anonymous synthetic ones. License plate detection algorithms used within MIDAS has been described in Section 6.6, Figure 29 illustrates a result of LP detection using WPOD algorithm[41].

Anonymous synthetic LP will then be affine transformed to match the shape of detected LP, and the transformed version will be blended into the original image to replace the LP using Poisson Image Blending[42], as in Figure 30.

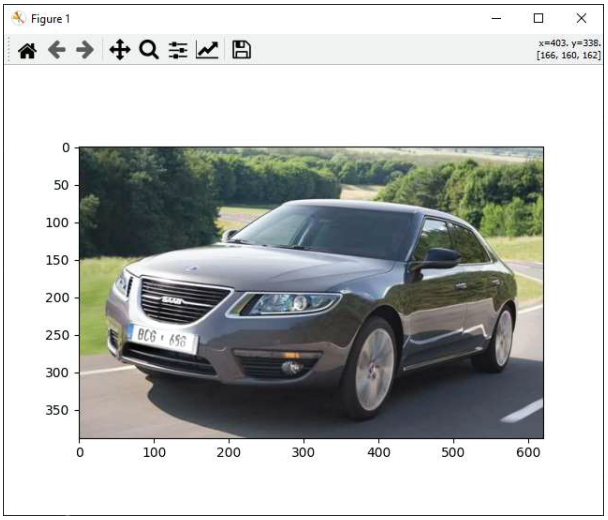

Figure 30: Replace the detected LP in original image with a pix2pix synthetic LP using Poisson image blending algorithm.

6.8.3 *Video data replacement*

The masking of license plates was also tested on video but with constraints since the database for training the detector was not created with Swedish license plates. We collected a small number of Swedish license plates to train a small detector based on YoloV8. This detector allowed us to work with sequences of images in real time and with high accuracy for the limited dataset. A larger dataset of images collected from Swedish vehicles and environments will improve the results observed in the output video (https://youtu.be/13w4k4c5RLk). The collection of training images will need to be compliant with the guidelines of GDPR. The implementation of this short test used 50 images publicly available of Swedish vehicles in traffic and demonstrated that it is possible to implement anonymization of license plates in real-time video sequences.

6.9 Public datasets

As a part of the project two datasets, which were presented earlier, have been collected. In order to make these results from the project available to the research community and industry data has been made available at (https://risecloud-my.sharepoint.com/:f:/g/personal/martin_torstensson_ri_se/EhkAP7u32ixDvuV9BVDAFEUB5Z9Wnjs8Zxa1bHRRVoxmNw?e=TYpLgV) or by contacting the contact persons within the project specified in chapter 9.

The initial dataset with personnel from Viscando was uploaded in two different versions. First the raw original dataset and also one where the images have been anonymized with the algorithms developed in this project. This provides an opportunity for others to try out their algorithms on the data and compare the results and to evaluate the performance of the current algorithm. Moreover, the two datasets can be used to evaluate how large the effects are of the anonymized images in other algorithms and whether it changes their performance or not.

From the larger databases collected in Schweiz up to a total of 8 hours of data was published. This data was only published in its raw format as the current anonymization algorithms do not work on data collected that far away. This is an interesting case where the data can still contain personal information, but the persons are small enough that the detection systems does not always apply. It is a good case for where full body anonymization would be a potential candidate in providing a solution.

Additionally, to provide an opportunity for both the research community and industry to test and evaluate the anonymization method it has been uploaded to Huggingface (https://huggingface.co/spaces/felixrosberg/face-swap). Where it is available for anyone to try out with their own images, whether they want to use the face swapping portion or the anonymization portion of the algorithm.

7 Dissemination and publications

7.1 Dissemination of knowledge and results

Hur har/planeras projektresultatet att användas och spridas?	Markera med X	Kommentar
Öka kunskapen inom området	X	Five papers have been published as a result of the project and several internal and external presentations have been held
Föras vidare till andra avancerade tekniska utvecklingsprojekt	X	A continuation project called MIDAS II has been applied for (Diarienr: 2023-00765)
Föras vidare till produktutvecklingsprojekt		
Introduceras på marknaden	X	Image-based demo is open available for anyone to try out both face swapping and anonymization at: https://huggingface.co/spaces/felixrosberg/face-swap
Användas i utredningar/regelverk/ tillståndsärenden/ politiska beslut		

7.2 Publications

FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping
F. Rosberg, E. Aksoy, C. Englund, F. Alonso-Fernandez
Published in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023
Arxiv: https://arxiv.org/abs/2210.10473

Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning

F. Rosberg, C. Englund
Published in 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG) 2021
IEEE Xplore: https://ieeexplore.ieee.org/document/9666958

Towards Privacy Aware Data collection in Traffic: A Proposed Method for Measuring Facial Anonymity
F. Rosberg, C. Englund, M. Torstensson, B. Durán
Published in FAST-zero 2021
Online access: https://tech.jsae.or.jp/paperinfo/en/content/conf2021-02.31/

Data Leakage in Anonymization Methods: Towards explainable machine learning
M. Torstensson, F. Rosberg, B. Durán, C. Englund
Published in FAST-zero 2021
Online access: https://tech.jsae.or.jp/paperinfo/en/content/conf2021-02.69/

Anonymising Data Collection for Traffic Safety: A first study for the MIDAS initiative
B. Durán, M. Torstensson, C. Englund
Published in FAST-zero 2021
Online access: https://tech.jsae.or.jp/paperinfo/en/content/conf2021-02.68/

7.3 Presentations

- 4 SAFER presentations
- 2 presentations at PhD conference ITEP
- Several presentations at PhD conferences within a smart industry research school
- 3 presentations in FastZero
- Poster in IEEE/CVF
- Poster in IEEE Xplore
- Presented at RISE housewarming with external attendants.
- Internal presentations

8 Conclusions and future research

Based on the results from the project it has been shown that it is possible to create new inventive ways of performing anonymization where important data does not have to be removed or the images look unrealistic. With the presented method it is possible to answer both RQ1 and RQ2 where the anonymization across frames in an image does still maintain the same identity and it is also possible to create new identities every time a person enters the camera range. While work has been done in this project towards answering RQ3, it is a hard question to definitively answer and the recommended approach to a next step is to create full body anonymization.

The results from MIDAS have been very encouraging and a continuation project (MIDAS II, Diarienr: 2023-00765) has already been applied for in the FFI program. Building upon the findings in this project some areas of future work has been identified in order to bring the research closer to a product that can be offered both to industry and the public sector:

- **Full body anonymization**: While this area was started within this project it is a challenging task with potential for improvement within a project with more dedication towards this specific area. More work is needed to create better methods to do full body anonymization in order to better remove all information that could potentially lead to identifying someone.
- **Legislative processes**: There is more necessary regarding legislative processes in order to make sure that a user of the system can be assured that they are fully complying with applicable regulations. This is both to make sure that the development process can further take this into consideration and change the algorithm design and evaluation to match the needs and to help users to be able to utilize the results.

- **Attack methods**: The evaluation of the methods should be further investigated. It would be beneficial to investigate different attack methods where someone has access to either a large amount of anonymized data or even a trained model. The question to be answered is if it is possible to reverse the anonymization with this additional information and in that case how it can be prevented.
- **More test scenarios**: In the continuation work tests on both large, annotated datasets and also the installation in test vehicles have been planned. This is to further increase the base of the evaluation and make sure the solution is functioning well in realistic test cases.

9 Participating parties and contact persons

RISE Research Institutes of Sweden AB (Coordinator)
Contact person: Martin Torstensson, martin.torstensson@ri.se

Berge Consulting AB
Contact person: Felix Rosberg, felix.rosberg@berge.io

Halmstad University
Contact person: Fernando Alonso-Fernandez, fernando.alonso-fernandez@hh.se

Viscando AB
Contact person: Yury Tarakanov, yury@viscando.com

References

[1] Z. Ren, X. Jiang, and W. Wang, “Analysis of the Influence of Pedestrians’ eye Contact on Drivers’ Comfort Boundary During the Crossing Conflict,” *Procedia Engineering*, vol. 137, pp. 399–406, Jan. 2016, doi: 10.1016/j.proeng.2016.01.274.

[2] A. Ng and M. Jordan, “On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes,” in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/7b7a53e239400a13bd6be6c91c4f6c4e-Paper.pdf

[3] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[4] Z. Wang, Q. She, and T. E. Ward, “Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy.” arXiv, Dec. 29, 2020. doi: 10.48550/arXiv.1906.01529.

[5] E. Denton, S. Chintala, A. Szlam, and R. Fergus, “Deep generative image models using a Laplacian pyramid of adversarial networks,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, in NIPS’15. Cambridge, MA, USA: MIT Press, Dec. 2015, pp. 1486–1494.

[6] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” arXiv, Jan. 07, 2016. doi: 10.48550/arXiv.1511.06434.

[7] D. Berthelot, T. Schumm, and L. Metz, “BEGAN: Boundary Equilibrium Generative Adversarial Networks.” arXiv, May 31, 2017. doi: 10.48550/arXiv.1703.10717.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” arXiv, Feb. 26, 2018. doi: 10.48550/arXiv.1710.10196.

[9] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-Attention Generative Adversarial Networks.” arXiv, Jun. 14, 2019. doi: 10.48550/arXiv.1805.08318.

[10] A. Brock, J. Donahue, and K. Simonyan, “Large Scale GAN Training for High Fidelity Natural Image Synthesis,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=B1xsqj09Fm

[11] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN.” arXiv, Dec. 06, 2017. doi: 10.48550/arXiv.1701.07875.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved Training of Wasserstein GANs.” arXiv, Dec. 25, 2017. doi: 10.48550/arXiv.1704.00028.

[13] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least Squares Generative Adversarial Networks.” arXiv, Apr. 05, 2017. doi: 10.48550/arXiv.1611.04076.

[14] S. Nowozin, B. Cseke, and R. Tomioka, “f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization.” arXiv, Jun. 02, 2016. doi: 10.48550/arXiv.1606.00709.

[15] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, “Unrolled Generative Adversarial Networks.” arXiv, May 12, 2017. doi: 10.48550/arXiv.1611.02163.

[16] G.-J. Qi, “Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities.” arXiv, Mar. 18, 2018. doi: 10.48550/arXiv.1701.06264.

[17] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, “Mode Regularized Generative Adversarial Networks.” arXiv, Mar. 02, 2017. doi: 10.48550/arXiv.1612.02136.

[18] J. H. Lim and J. C. Ye, “Geometric GAN.” arXiv, May 08, 2017. doi: 10.48550/arXiv.1705.02894.

[19] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN.” arXiv, Sep. 10, 2018. doi: 10.48550/arXiv.1807.00734.

[20] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral Normalization for Generative Adversarial Networks.” arXiv, Feb. 16, 2018. doi: 10.48550/arXiv.1802.05957.

[21] A. Hindupur, “The GAN Zoo,” *Medium*, Sep. 30, 2018. https://deephunt.in/the-gan-zoo-79597dc8c347 (accessed Jan. 14, 2021).

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 8107–8116. doi: 10.1109/CVPR42600.2020.00813.

[23] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject Agnostic Face Swapping and Reenactment,” *arXiv:1908.05932 [cs]*, Aug. 2019, Accessed: May 15, 2020. [Online]. Available: http://arxiv.org/abs/1908.05932

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation.” arXiv, May 18, 2015. doi: 10.48550/arXiv.1505.04597.

[25] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs.” arXiv, Aug. 20, 2018. doi: 10.48550/arXiv.1711.11585.

[26] H. Hukkelås and F. Lindseth, “DeepPrivacy2: Towards Realistic Full-Body Anonymization,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1329–1338.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[28] “MoveNet: Ultra fast and accurate pose detection model. | TensorFlow Hub,” *TensorFlow*. https://www.tensorflow.org/hub/tutorials/movenet (accessed Apr. 28, 2023).

[29] S. M. E. Portal, “New Federal Act on Data Protection (nFADP).” https://www.kmu.admin.ch/kmu/en/home/fakten-und-trends/digitalisierung/datenschutz/neues-datenschutzgesetz-revdsg.html (accessed Apr. 26, 2023).

[30] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[31] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, “FaceDancer: Pose-and Occlusion-Aware High Fidelity Face Swapping,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3454–3463.

[32] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694. doi: 10.1109/CVPR.2019.00482.

[33] T. Karras *et al.*, “Alias-free generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[34] A. Van Den Oord, O. Vinyals, and others, “Neural discrete representation learning,” *Advances in neural information processing systems*, vol. 30, 2017.

[35] R. Chen, X. Chen, B. Ni, and Y. Ge, “SimSwap: An Efficient Framework For High Fidelity Face Swapping,” in *Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020, pp. 2003–2011. doi: 10.1145/3394171.3413630.

[36] H. Wang *et al.*, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[37] H. Hukkelås, R. Mester, and F. Lindseth, “Deepprivacy: A generative adversarial network for face anonymization,” in Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14, Springer, 2019, pp. 565–578.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.

[39] openALPR, “OpenALPR-EU.” https://github.com/openalpr/train-detector/tree/master/eu (accessed Apr. 19, 2023).

[40] “License plates - Transportstyrelsen.” https://www.transportstyrelsen.se/en/road/Vehicles/license-plates/ (accessed Sep. 01, 2020).

[41] S. M. Silva and C. R. Jung, “License Plate Detection and Recognition in Unconstrained Scenarios,” in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., in Lecture Notes in Computer Science, vol. 11216. Cham: Springer International Publishing, 2018, pp. 593–609. doi: 10.1007/978-3-030-01258-8_36.

[42] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003, doi: 10.1145/882262.882269.