Combining DeepPrivacy with an Attribute-driven Generative Adversarial Network to Preserve Gender and Age in De-identified CCTV Footage

EMILIJA JASINSKAITE & ØYVOR YSTAD SKJEI

SUPERVISORS
Lei Jiao
Morten Goodwin

University of Agder, 2021
Faculty of Engineering and Science
Department of Engineering Sciences

Preface and Acknowledgements

This master thesis concludes the master's education in Information and Communication Technology at the University of Agder, Norway.

We want to thank our supervisors, Lei Jiao and Morten Goodwin, who have provided invaluable guidance throughout the project. They have always been available and provided us with insightful and valuable input on topics including ethical questions and research methods. A huge thank you for your time and support!

We also want to thank our better halves, Ketil Grandalen and Byamungu Kabiraba, and our families for being our rocks and providing understanding, love, and support during the project.

In addition we want to thank Bente Skattør, our contact person at the Oslo Police District, for providing us with this task of anonymization and preservation of certain attributes.

Abstract

A surveillance camera is an efficient solution to prohibit crimes for both small and big businesses, and is broadly utilized in big cities. Today, the police force can only access the camera footage for further investigation after an act of crime. In order to observe, find patterns, and react appropriately to an event, the Oslo Police wants to use its own CCTV cameras and analyze such footage in real-time. To investigate real-time CCTV footage and share such footage with a third-party for analyzing, the people in the footage need to be de-identified. In this thesis, we focus on de-identification of CCTV footage, preserving age and gender for more precise context information.

DeepPrivacy is a neural network model that creates new faces using image inpainting. It is found to be suitable for de-identification of CCTV footage but the creators did not intend to preserve age and gender. The thesis proposes combining DeepPrivacy and an attribute-driven network to enforce preservation of age and gender, and performs experiments on two state-of-the-art, attribute-driven Generative Adversarial Networks (GANs), AttGAN, and StarGAN v1. These networks are designed to keep the input image intact while changing specific attributes. The thesis also studies the option of changing the subjects' skin tone to a specific color to bypass potential ethnicity bias.

To preserve attributes in CCTV footage, AttGAN and StarGAN v1 should be trained on a dataset with diverse image quality and poses, with a good representation of different age groups and a balanced representation of gender. No such dataset exists, and thus, this thesis proposes a new dataset "Diverse Faces" of 223,548 images labeled with age and gender. To enable change of skin tone to bypass potential ethnicity biases, the thesis proposes an additional dataset named "Diverse Faces with Distinct Skin tones" containing 188,113 images, labeled with skin tone in addition to age group and gender.

The de-identification rate of DeepPrivacy is 97.40%, and the network originally preserves 77.50% gender and 42.25% age group. The proposed solution combines DeepPrivacy and AttGAN. For AttGAN, the thesis proposes instance normalization for better preservation of image background and combining Multi-scale Structural Similarity (MS-SSIM) and L1 norm for reconstruction loss for reducing image noise. "DeepPrivacy and AttGAN" (DP-ATT), trained on "Diverse Faces", preserves 89.00% gender and 79.78% age group. "DeepPrivacy and AttGAN with Skin tone" (DP-ATT-S), trained on "Diverse Faces with Distinct Skin tones", preserves 84.25% and 83.58% gender, and 62.42% and 63.50% age group for dark and light skin tone respectively.

Contents

Chapter 1

Introduction

Technology becomes more and more integrated in our everyday lives. In 1923, Samuel Shlafrock invented the first instant camera, which combined a camera and a dark room in a single compartment [9]. Now, nearly 100 years later, it is expected that we may reach one billion CCTV surveillance cameras globally by the end of 2021 [42].
In relation to the growth of use cases for technology, the term *smart city* has appeared. The European Commission defines a smart city as “a place where traditional networks and services are made more efficient with the use of digital and telecommunication technologies for the benefit of its inhabitants and business” [59]. The enormous amount of CCTV surveillance cameras produces huge volumes of data which, when gathered and analyzed, can be a great contributing factor to a smart city.

The Oslo Police district have decided to take part in the smart city movement and want to utilize CCTV footage to detect anomalies to react appropriately. Examples of anomalies in this context could be situations such as a fight, a person that have fallen and do not get up, theft, etc. Currently the Oslo Police are not allowed to real-time monitor the city of Oslo due to laws and regulations. The police force are only allowed to use data from legal and illegal CCTV cameras in relation to an investigation of a crime or event that has already taken place. Illegal CCTV cameras refer to cameras that are not approved for surveillance. For the Oslo Police to potentially gain access to deploy their own CCTV cameras and use them to capture real-time CCTV footage, the footage needs to be de-identified. Also, manually detection of anomalies using human operators is too resource intensive. Therefore, the Oslo Police want to use Artificial Intelligence (AI) to analyze and identify anomalies when they take place in real-time. They plan to use crowd sourcing and use external partners that can provide such anomaly detection network. In order to train the anomaly detection network, real CCTV footage needs to be used as training data. For the Oslo Police to be able to provide such training data, the footage, again, needs to be de-identified.

Traditional de-identification methods, such as blurring, are not sufficient for this scenario as they remove too much information, including all facial information, objects that may be in front of a face, and some of the surroundings. Face detection networks use facial landmarks like eyes, nose, ears, and other facial features to categorize an object as a human being. Such features may be lost when the images are blurred, and the imagery is altered to a point where it is hard to find useful information like age and gender. Age and gender can put context to a situation that may clarify the harmfulness in an act, like the difference between an adult and a child hitting another adult.

Although a person can be recognised by various factors, such as clothing, GPS data, the way they walk or the surroundings, the focus for de-identification of CCTV footage in this thesis is limited to de-identifying the face.

The *Black Lives Matter* movement originating in the United States 2020 and became a subject in Europe as well, and is proof of a biased world where different ethnicity may lead to different treatment by the federal state. CCTV footage does not differentiate on ethnicity, however an anomaly detection system based on a potentially biased training set might do. In the thesis we suggest an option of choosing a collective color unrelated to ethnicity that is persistent throughout the set of de-identified images. Because ethnicity holds more than just a color, it is also necessary to change the appearance of the original face, which is a process needed for de-identification either way.

1.1 Motivation

In this thesis we seek to find a neural network design meant to compete with state-of-the-art solutions for de-identification with preservation of age and gender. We believe it will make a valuable contribution to the ongoing de-identification research and add another tool to the collection of de-identification methods for images and videos. This field of research helps the realization of Smart Cities, a concept that most likely is going to improve the use of surveillance cameras and make them a standby alert system rather than a box of information that is rarely used.

The value of de-identified images and videos goes beyond the Smart City and government usage. In a more general perspective, such a solution enables any person to upload images and videos of themselves while de-identified. Their audience would still be able to see the gender and age group of the creator but would not be able to identify the original face, nor the ethnicity. This makes it possible to create and share content without being afraid that it will be related to yourself personally.

The method of de-identification can also be used in research when observing human behavior to create possibly more realistic statistics. If one were to film in a public space, consent from the people that may appear in the video would be needed. Asking for such consent may also include revealing the behavioral study, which may result in less natural behavior. Using the method of de-identification while preserving age and gender makes it possible to use said footage without the consent as the people in it would not be identifiable. As the people that are being studied are unaware of the study itself, they will presumably behave more naturally, while age and gender are still attributes that the study may include. An example of such behavior study where this would be relevant, can be a study performed in grocery stores to find statistics on which time of the day different genders and age groups go shopping. Envisioning a store chain that uses TV screens for marketing in their stores, such study would enable more targeted marketing, targeting different age groups and genders on different times of the day. A different example may be a study for a pub that wants to know which ages and genders their client base consists of, or a nutrition study, where footage from several restaurants is used to estimate the nutritional values that the different ages and genders get when eating out.

1.2 Problem Statement

The problem this thesis focuses on is de-identifying a person by only changing the face, while still preserving gender and age group, and also changing skin color to avoid racial bias. It is also important to maintain trackability as it is assumed that the anomaly detection network will need to track people in order to find an anomaly event. Although time is an important aspect of real-time de-identification, it will not be in focus for this thesis.

1.3 Research Question and Hypotheses

The research question of this thesis is:
Can we remove all recognizable features from a face and still generate a new face with same gender and approximately same age?

It is accompanied by the following hypotheses:

1. *H1: We can remove all recognizable features from a face and still generate a new face with same gender.*
   This hypothesis focuses on preserving gender. A confirmation of this hypothesis would be if, after using the proposed solution, a gender estimation network is able to estimate the original gender in most of the de-identified images. The hypothesis will be confirmed if we are able to preserve the original gender in 10% more images than using DeepPrivacy alone.

2. *H2: We can remove all recognizable features from a face and still generate a new face with approximately same age.* The second hypothesis focuses on preserving age. This hypothesis is confirmed if, after using the proposed solution, an age estimation network is able to estimate the original age group in most of the de-identified images. Again, the hypothesis will be confirmed if we are able to preserve the original age group in 10% more images than using DeepPrivacy alone.

3. *H3: We can change all skin colors to one color in order to avoid bias towards certain skin tones.* The third hypothesis focuses on changing skin tone. This hypothesis is confirmed if the skin tone of most of the images is closer to the average color than it is using DeepPrivacy alone. The deviation of the average skin color compared to the target skin color should not be more than 20% per Red (R), Green (G) and Blue (B) value.

1.4 Contributions

The main contributions of this thesis are:

- We propose a novel method cascading DeepPrivacy and an attribute-driven GAN in order to preserve gender and age group in de-identified CCTV footage[1]

- Using the proposed scheme, we present “DeepPrivacy and AttGAN” (DP-ATT) trained on “Diverse Faces” with the attributes of gender and multiple age groups. It has the de-identification rate of 98.47%, and is able to preserve gender with an accuracy of 89.00% and age group with an accuracy of 79.78%[2].

- Again using the same proposed scheme, we present “DeepPrivacy and AttGAN with Skin tone” (DP-ATT-S) trained on “Diverse Faces with Distinct Skin tones” with the attributes of gender, skin tone and multiple age groups. It also has the de-identification rate of 98.47%. Gender is preserved with the accuracy of 84.25% and 83.58%, and age group is preserved with the accuracy of 62.42% and 63.50% for dark and light skin tone respectively[3].

- We prepare the dataset “Diverse Faces”, consisting of 180,492 images for training and 43,119 images for validation. All images have label files with age group and gender[4].

[1] The code can be found here.
[2] The pretrained AttGAN model for DP-ATT can be found here.
[3] The pretrained AttGAN model for DP-ATT-S can be found here.
[4] Diverse Faces can be found here.

- We prepare the dataset “Diverse Faces with Distinct Skin tones”, consisting 174,334 images for training and 13,779 images for validation. All images have label files with skin tone, age group and gender [5].

1.5 Outline of the Thesis

This thesis is organized as stated below:
Chapter 2 contains the technical background and state-of-the-art solutions.
Chapter 3 includes information about the network architecture and training details of the proposed method, as well as the proposed datasets.
Chapter 4 starts with a de-identification experiment for DeepPrivacy to find the de-identification rate. Then, a trackability experiment is conducted on de-identified videos. The next part focuses on preserving the age group and gender when de-identifying faces, with experiments testing different networks, and how changes to the training dataset and parameter changes affects the results. Then, we look into some options for changing skin tone in the de-identified images, and in addition conduct some experiments that have the goal of improving the results. The chapter ends with an evaluation of our proposed networks and comparison to state-of-the-art.
Chapter 5 contains the conclusions of this thesis and answers the research question.

[5] Diverse Faces with Distinct Skin tones can be found here.

Chapter 2

Background

In this chapter the reader will find state-of-the-art solutions to solve similar task to the stated research question and hypothesis. The mentioned work provides methods to answer questions like how to de-identify faces, how to preserve information like age and gender, and how one can track people in CCTV footage. Today, these questions are best solved using neural networks. For image generation in particular, a type of network has become increasingly popular the last few years, namely a network construction known as *Generative Adversarial Network* (GAN). Being a generative network, GAN is able to output multiple, never-seen-before samples from a training set. It can be used for generating larger datasets, however, has also been able to fulfill other image-applications like style transfer, attribute transfer, and realistic image inpainting of missing pixels. The concept of GAN and variations are described further in section 2.1.

Section 2.2 looks into laws of surveillance, ethics of AI, and the meaning behind *de-identification*. This is followed by research on the predecessor of GAN, the *k-Same algorithm*. The next subsection focuses on state-of-the-art de-identification methods using GAN and the subsequent subsection on age and gender preservation using GAN. The last part looks into tracking de-identified people in CCTV footage.

Section 2.3 discusses four different approaches based on state-of-the-art. Each approach is evaluated before one is chosen for further exploration. Additional information concerning the chosen state-of-the-art is given to help the reader better understand how they work.

2.1 Generative Adversarial Networks (GANs)

Goodfellow et al. [22] described a network called Generative Adversarial Network in June 2014. Generative models enable multi-model output, where a single output may produce multiple acceptable results [21]. It is based on a game theoretic scenario where a Generator (G) competes against an adversary known as a Discriminator (D). The goal of G is to convince D that the generated samples, in our case images, are real. D receives samples from both G and a real dataset, and emits a probability indicating whether the sample is real or fake. If the game goes as planned, D will learn the difference of real and fake images, however, as G improves, the generated samples will become indistinguishable from the real data and D emits a probability of 0.5 everywhere. At this stage, the game comes to a halt. G can now be used to generate data similar to the real input, or for other application purposes described later in this section.

2.1.1 Vanilla GAN

In original GAN design, also known as Vanilla GAN [65], the generator and discriminator take part in a zero-sum game where a function $v(G, D)$ determines the payoff of D, and

$-v(G, D)$ the payoff of G. D will try to maximize v and G minimize as shown in Equation 2.1,

$$\theta^* = \arg\min_{g}\max_{d} v(G, D). \tag{2.1}$$

To make the generated samples indistinguishable from the real data, the generated sample z should be drawn from a data distribution $\mathbb{P}_\theta$ similar to the distribution of real data $\mathbb{P}_r$. The proposed loss function is

$$v(D, G) := \mathbb{E}_{x\sim\mathbb{P}_r}\left[\log D(x)\right] + \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[\log(1 - D(G(z)))\right], \tag{2.2}$$

where x is a sample from the real images, z a generated sample, $\mathbb{E}_{x\sim\mathbb{P}_r}$ and $\mathbb{E}_{z\sim\mathbb{P}_\theta}$ is the expected value of x and z given distribution $\mathbb{P}_r$ and $\mathbb{P}_\theta$ respectively. D will reject generated samples by assigning high values to samples from $\mathbb{P}_r$ and low values for samples from $\mathbb{P}_\theta$. Since this is a $minmax$ problem, it is not necessary to define the base of the logarithm, any base will be sufficient. To measure the discrepancy between two probability distributions, it is possible to use f-divergence.

Kullback Leibler (KL) divergence [31] can be viewed as the relative entropy between two probability density functions $f(x)$ and $g(x)$,

$$KL(F||G) \overset{def}{=} \int f(x)\log\frac{f(x)}{g(x)}dx. \tag{2.3}$$

Jensen-Shannon (JS) divergence [65] is defined for any probability density functions $P_r(x)$ and $P_m(x)$

$$\begin{aligned} JS(\mathbb{P}_r, \mathbb{P}_\theta) = \frac{1}{2}KL(\mathbb{P}_r, \mathbb{P}_m) + \frac{1}{2}KL(\mathbb{P}_\theta, \mathbb{P}_m),\\ where\ \mathbb{P}_m = \frac{\mathbb{P}_r + \mathbb{P}_\theta}{2}. \end{aligned} \tag{2.4}$$

2.1.2 Wasserstein GAN (WGAN)

Density replication can fail to recreate the dimensional manifolds of the true data. Adding Gaussian noise with high bandwidth can help to overcome this issue, however, such method adds noise to the outputted data. Arjovsky et al. [6] introduced the use of Wasserstein-1, or Earth Mover distance, which enabled more learning stability and is more resistant to mode collapse, a problem where all input images map to the same output image, disabling progress of the optimization. Instead of estimating the density of $\mathbb{P}_r$ which may not exist, Wasserstein GAN defines a random Gaussian variable Z with a fixed distribution $p(z)$ and passes it through a parametric function $g_\theta : Z- > X$ (typically a neural network) that directly generates samples following a certain distribution $\mathbb{P}_\theta$. By adjusting θ one can derive a distribution close to $\mathbb{P}_r$. Equation 2.5 describes Wasserstein 1, where $\prod(\mathbb{P}_r, \mathbb{P}_\theta)$ is all joint distributions $\gamma(x, y)$ whose marginals are $\mathbb{P}_r$ and $\mathbb{P}_\theta$ respectively.

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \inf_{\gamma\epsilon\prod(\mathbb{P}_r,\mathbb{P}_\theta)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||]. \tag{2.5}$$

The method enforces weight clipping for the discriminator, cutting the weights that does not comply within the space of 1-Lipschitz functions. Note that the objective of WGAN is to minimize the adversarial loss as opposed to Vanilla GAN, and is stated as

$$\min_{G}\max_{||D||L\leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}\left[D(x)\right] - \mathbb{E}_{z\sim\mathbb{P}_\theta}\left[D(G(z))\right] \tag{2.6}$$

2.1.3 Improved Wasserstein GAN: WGAN with gradient penalty (WGAN-GP)

Gulrajani et al. [26] proposed an alternative way of clipping weights of WGAN by penalizing the norm of gradient of the discriminator with respect to its input. This method is more stable for a broader variety of GAN architectures, requiring less hyper-parameter tuning.

2.1.4 Deep Convolutional GAN (DCGAN)

Radford et al. [53] proposed DCGAN to bridge the gap between Convolutional Neural Networks (CNNs) and GANs in unsupervised learning. The architecture replaces pool layers with strided convolutions in the discriminator, and fractional-strided convolutions in the generator. Stride is the distance between two consecutive positions of the pooling window, where fractional striding add zero-padding between input vaues [17]. The design uses batch normalization in both the generator and the discriminator and removes fully connected hidden layers for deeper architectures.

2.1.5 Progressive GAN

Karras et al [41] proposed a GAN model that grows both the generator and the discriminator progressively from a small resolution, increasing for each layer. The generator and discriminator grow synchronously, thus mirroring each others structure. The trainable layers are smoothly faded in to avoid sudden changes to the already trained resolution layers. Such techniques stabilize as well as speed up the training phase. The incremental nature of the architecture allows the network to detect large scale structures of the image distribution early in the process, and then increasingly fine-tune the image as it scales, instead of learning all scales simultaneously. The progressive architecture is more likely to converge at the smaller resolutions, so the last rounds of scaling are only tasked with refining the image representation. By enabling the possibility of scaling the weights during training, Karras et al. make sure the learning speed is the same for all weights. The authors also present some implementation details that discourage competition between the generator and the discriminator and a higher quality dataset based on CelebA, which they have named CelebA-HQ.

2.1.6 MSG-GAN

GANs can be difficult to adapt to different datasets and their training phase suffers from mode collapse and training instability, the latter being a problem that Animesh Karnewar and Oliver Wang proposes a solution for with a new training technique; Multi-Scale Gradient Generative Adversarial Network (MSG-GAN) [38]. MSG-GAN provides a stable approach for high resolution image synthesis by allowing the flow of gradients from the discriminator to the generator at multiple resolutions simultaneously. Meaning that instead of the discriminator only looking at the final output with the highest resolution from the generator, it can also look at the outputs of the intermediate layers.

It is found that this method is robust to different loss functions, datasets of different sizes, resolutions and domains, and architectures. It is also shown that MSG-GAN is able to converge stably using the same set of fixed hyper parameters for all the variations. This training technique can be used in place of the progressive growing technique, offering similar training time, fewer hyper parameters, and easier generalizing to different datasets. Also, the generated images from the experiments conducted by Karnewar et al., do not show any traces of the phase artifacts that are visible in progressively grown GANs.

Pidhorskyi and Gianfranco [52] propose a Auto Encoder (AE) network that aims to combine the generative and representational properties training a encoder-generator map. Opposed to general GAN designs, they believe the AE architecture is able to learn a less entangled representation of the latent space. They design two Adversarial Latent Autoencoders, one derived from StyleGAN and the other with progressive growing. The design learns the latent space distribution while the data distribution is learned in adversarial settings, thus the learning becomes less entangled. The network generates images comparable to the quality of StyleGAN generated images, however, it can in addition produce face reconstructions and manipulations based on real images.

2.1.7 CycleGAN

Image-to-image translation is a concept based upon the assumption that it is possible to translate an image into another, in the same manner one can translate words from one language to another. Jun-Yan Zhu et al. proposed an approach for image-to-image translation without the need of paired examples, named CycleGAN [72]. The approach uses two separate generators, namely G and F. Given an input domain X and an output domain Y, $G(x)$ =y, and likewise $F(y)$ = x. They define **cycle consistency loss** which is meant to encourage $F(G(x))$ = x and $G(F(y))$ = y. In addition of introducing a translation factor, F provides an inverse mapping that avoids mode collapse.

2.1.8 Data Normalization

Normalization is a technique used to ensure that data has certain statistical properties, removing magnitudes between different features [33]. In neural networks, normalization is used to shift and scale activations by using the mean μ and standard deviation σ (see Equation 2.7), making features more equally represented. The activation x at any layer exists within four dimensions, the batch size N, the number of channels (filters) C, height H, and weight W, $x \epsilon \mathbb{R}^{N\times C\times H\times W}$ [64]. In this thesis, we describe two techniques using normalization, namely **instance normalization (IN)** and **batch normalization (BN)**.

$$\hat{x} = \frac{x - \mu}{\sigma}. \tag{2.7}$$

Instance Normalization

In instance normalization, mean and variance are calculated for each individual channel for each individual sample across both H and W (see Equation 2.8). This indicates that each training sample is reflected in the normalization process.

$$\hat{x} = \frac{x - \mu_{nc}}{\sqrt{\sigma_{nc} + \epsilon}},\ \mu_{nc} = \frac{1}{HW}\sum_{j=1}^{H}\sum_{k=1}^{W} x_{ncjk},\ \sigma_{nc}^2 = \frac{1}{HW}\sum_{j=1}^{H}\sum_{k=1}^{W}(x_{ncjk} - \mu_{nc})^2. \tag{2.8}$$

Batch Normalization

In batch normalization, mean and variance are calculated for each individual channel for all samples, H, and W (see Equation 2.9). Thus, contrarily to instance normalization, the mean and variance are a representation of all samples combined, awarding more similar features throughout the dataset.

$$\hat{x} = \frac{x - \mu_{c}}{\sqrt{\sigma_{c}^2 + \epsilon}},\ \mu_{c} = \frac{1}{NHW}\sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W} x_{icjk},\ \sigma_{nc}^2 = \frac{1}{NHW}\sum_{i=1}^{N}\sum_{j=1}^{H}\sum_{k=1}^{W}(x_{icjk} - \mu_{c})^2. \tag{2.9}$$

2.2 Related Work and the State-of-the-art for De-identification and Tracking

This section contains information about relevant regulations and ethics, the K-Same algorithm, other approaches of using GAN for de-identification, and attribute-driven GANs. It also focuses on the state-of-the-art of tracking, presenting Simple Online and Realtime Tracking (SORT) and You Only Look Once (YOLO).

2.2.1 Regulations of Camera Surveillance

In Norway, multiple laws are regulating the use of surveillance technology.
"Menneskerettsloven" [47], focusing on preserving human rights, entered into force 21th of May, 1999 and was last changed on the 9th of May, 2014. Protocol 8 protects privacy and family life, saying that everyone has the right to respect for their privacy and family life, and that there shall be no interference by public authority in the exercise of this right, except where this is in accordance with the law and is necessary in a democratic society for reasons of national security, public security or the economic welfare of the country, in order to prevent disorder or crime, to protect health or morals, or to protect the rights and freedoms of others.

"Politiloven" [46] entered into force on the 4th of August 1995, and was last changed recently, on the 16th of April, 2021. Paragraph 6a focuses specifically on camera surveillance, stating that the police force can use camera surveillance if it is necessary in order to carry out certain tasks, including protecting person, property and public goods and legal activities, prevent crime and other violations of public order and security, detect and stop criminal activity and prosecute criminal offenses, and provide citizens with assistance when the circumstances indicate that it is required.

"Personopplysningsloven" [45] entered into force on 15th of June, 2018, and was last changed on the 14th of April, 2000. This law focuses on the processing and storage of personal data. Personal data is defined as any information about an identified or identifiable person. An image of a person can be considered as personal data if it is possible to recognise people in the image [15].
Protocol 5a states that personal data shall be processed in a lawful, fair and transparent manner with respect to the data subject, while Protocol 6 elaborates legality of such data processing. Protocol 6.1 announces different terms, of which at least one should be fulfilled to accomplish legal processing of personal data. The listed terms include consent from the subject, and the processing being necessary to, amongst others, fulfill legal obligations, protect the vital interests of the subject or another person, and/or perform a task in the public interest.
Protocol 25 is about built-in privacy and privacy by default. Protocol 25.1 states that the responsible for the processing should carry out appropriate technical and organizational measures, e.g. pseudonymisation, designed accomplish an effective implementation of the principles of protection of personal data, and to integrate the necessary guarantees into the processing to meet the requirements of this Regulation and to protect the rights of subjects. Pseudonymization being the processing of personal data in such a way that the personal data can no longer be linked to a specific data subject without the use of additional data, provided that the said additional data is stored separately and covered by technical and organizational measures that ensure the personal data cannot be linked to the subject.

2.2.2 Ethics and Regulations of AI Systems

“Towards Responsible AI Innovation” [48] is a report on AI for law enforcement written as a collaboration between the international police organization (INTERPOL) and United Nations Interregional Justice Research Institute (UNICRI), together with partners from business and academia, to discuss advancements in AI and how such tool can help in the mission of fighting crime, claiming that it can be a powerful tool with a game-changing potential. On the other hand, AI is a double-edge-sword that need careful wielding to avoid infringing of fundamental human rights, such as the presumption of innocence and protection against self-incrimination. As for most technologies, AI can also be used for malicious activities. In 2019, the voice of a CEO of an energy company was successfully imitated by AI, resulting in a transfer of a substantial sum of money to a private account [48] .

To keep up with the growing technology of AI, law enforcement needs to cooperate with other stakeholders like the public sector, industry, academia, related security entities, intelligence agencies, counter-terrorism bodies and so on. AI systems created for use of law enforcement should comply with the principles of human rights, democracy, justice and rule of law.

In April 2018, 24 members of EU signed “Declaration on Cooperation on Artificial Intelligence” followed by the establishment of High-Level Expert Group on AI (AI-HLEG). This group has representatives from academia, industry and civil society, and was tasked with elaborating recommendations on future-related policy development and on ethical, legal and societal issues related to AI. They worked out seven Ethical Guidelines [5] for trustworthy AI, which are loosely translated to the following bullet points:

- **Human agency and oversight**: AI systems should empower people, allowing them to make informed decisions and foster fundamental rights,
- **Technical Robustness and safety**: AI systems must be resilient and secure, and include a fall back plan,
- **Privacy and data governance**: AI systems must ensure full respect for privacy protection and adequate data governance,
- **Transparency**: AI systems and AI businesses models should inform people about the system’s capabilities and limitations,
- **Diversity, non-discriminant and fairness**: AI systems should avoid unfair bias, foster diversity and be accessible to everybody,
- **Societal and environmental well-being**: AI systems should benefit all human beings, including future generations,
- **Accountability**: AI systems should have mechanisms that ensure responsibility and accountability.

Many of these requirements involve non-technical methods like regulations, codes of conduct, standardization, certification, and participating in terms of accountability. In February 2020, the European Commission released a White Paper “On Artificial Intelligence - A European Approach to Excellence and Trust” [13] that builds upon the Ethic Guidelines and also suggests specific legal requirements such as AI systems being trained on representative data, keeping detailed information on AI development and informing citizen when they are interacting with an AI system.

There are also other organizations taking effort to assert AI ethics. In May 2018, Organisation for Economic Co-operation and Development established an expert group that created “Principles on Artificial Intelligence” [18] which is adapted in 42 states [48]. Some countries have also established national committees on the ethical dimensions of AI, such as the Committee on Artificial Intelligence on the House of Lords of the United Kingdom and the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore. Large industries like Google, IBM and Microsoft have established ethical principles for exploration of AI. According to Google, AI should be socially beneficial, avoid creation or re-enforcement of bias, be built and tested for safety, and be accountable to people [18].

2.2.3 De-identification

“Towards Responsible AI Innovation” defines *visual processing* in technological context to be the mimicry of the human visual system by a computer system. It involves the extraction, analysis and understanding of information from images. Law enforcement has been supported by visual information like pictures, videos, vehicles and locations for a long time. Surveillance technology, CCTV in particular, allows for quick identification of victims, perpetrators, or people of interest. Today, surveillance systems are often combined with machine learning algorithms, which in the majority of cases are more efficient and effective than human labour. The integration of machine learning has revolutionized the areas of image processing and object recognition, making it possible to detect and track human faces and bodies. It is possible to identify abnormal behavior and black- or white-listing people into buildings or events such as concerts and festivals. The Oslo Police District collaborates with partners within the police force and externally with industry and academia to create a *non-intrusive* surveillance system. The non-intrusive part is to be achieved by *de-identifying* faces of people in videos. It is then, hopefully, possible to both view real-time CCTV footage to assess situations, and share the data with police partners that can use pattern recognition to identify acts such as vandalism, street fighting and other abnormalities without braking any law or regulation.

National Institute of Standards and Technology (NIST) [20] defines **de-identification** as a tool that organizations can use to remove personal information from data they collect, use, archive, and share with other organizations. The term is not restricted to a single technique but should rather be understood as a collection of approaches, algorithms, and tools used to protect privacy. Often there is a high correlation between utility and protection; the more privacy protection the less information can be gathered from the data. NIST states that the use of de-identification is of special importance for government agencies, businesses, and other organizations that share data to outsiders. Re-identification of de-identified data removes the protection and makes the de-identification method useless. It is difficult to foresee the re-identification risk of a de-identification method, thus an adaption to new re-identification methods should be a part of the de-identifying process. Even if a person is obfuscated to the degree to no recognition, it is possible to link the person using other information available to find the true identity of the person. Oh et al. [51] show that it is possible to train a person recognition system that only need a handful of images in order to threaten a person’s privacy.

For still photographs, consumer videos and surveillance videos, the de-identification process should remove identifiable information. ICT Cost Action IC1206 [60] defined three terms of identifiers of people in multimedia content:

- **Biometric identifiers** are distinctive, measurable, unique and permanent personal characteristics like face, iris, ear, fingerprint, voice, gait, gesture, lip motion, and typing style;
- **Soft biometrics** are vague physical and behavioral and not necessarily permanent and distinctive like height, weight, eye color, silhouette, age, gender, race, moles,

tattoos, birthmarks, and scars;

- **Non-biometric identifiers** are hairstyle, dressing style, and the context of text, speech, and social-political activity.

If all biometric identifiers, soft biometric, and non-biometric identifiers are de-identified one has achieved multimodal de-identification, which should be the goal for multimedia de-identification.

2.2.4 K-Same algorithm

The earliest use of de-identification in images mostly revolved around ad-hoc techniques such as blurring, pixelating, or totally removing facial information from images. Newton et al. [49] showed that pixelation and blurring failed against parrot recognition attacks where the attacker invokes the same de-identification technique on its own dataset and match the identified image with the original image. The remaining method, removing all face information, does provide privacy but little information can be used for information processing. Thus, the existing techniques were not sufficient for de-identification purposes.

At Carnegie University in Pittsburgh 2005, Newton et al. [49] introduced the *k-Same* algorithm. This algorithm finds similarity in faces from a collection of faces by measuring distance and then makes a new face by averaging the image components. Through testing, they found that their method, "total image blackout", and the process of changing a huge volume of pixels is equally effective as randomly guessing the identity. Of these methods, k-Same was the only one preserving information.

The same year, two advancement of this algorithm were proposed, *k-Same-Select* [24] in 2005 and *k-Same-M* [25] in 2006. K-Same-Select divides images based on their attributes onto mutual exclusive subsets before applying the same-K algorithm. k-Same-M use an Active Appearance Model to better align the faces before averaging the image components in order to overcome ghosting effects.

In 2009, Gross [23] pointed out that the k-Same algorithm becomes weak in presence of multiple images from the same person, as they are likely to have less distance which degrades the level of privacy in the generated image. Since a surveillance video likely has multiple frames of the same subject, they concluded that the algorithm does not provide sufficient protection. They instead propose a multifactor model that accounts for both identity and non-identity factors when constructing new faces. They showed that this method could preserve expression from the original image, and therefore in practice, other types of attributes.

Slobodan and Nikola Pavesic [58] are also sceptic to the performance of k-Same based algorithms, since the approach focus on still frontal images. This has a degrading visual quality and does not preserve naturalness.

In 2015, Amin Jourabloo et al. [36] looked into de-identification of face images while preserving a large set of attributes. Like the k-Same algorithm, they select k images that share similar attributes of the input image but use a gradient decent instead of averaging. Their method performs better having lower recognition rate while preserving more attributes.

Using k-Same for image de-identification significantly improved the prior de-identification methods, and have throughout the years been improved to overcome initial problems. We believe it is capable of de-identifying people in certain settings, however, when presented with real CCTV footage, the people to be de-ientified are not likely to have a forward pose or to be closely aligned. Finding the average of the presented faces only gives one new identity, whereas the video will contain multiple people that should be de-identified separately from each other. We therefore need another technology or mechanism to de-identify CCTV footage.

2.2.5 De-Identification using GAN

Compared to the k-Same algorithm, GAN is a more prominent solution, generating new, unique samples every time and can be less tangled to certain poses or image quality. There exist multiple types of GANs solving a broad variety of simple and more complicated tasks. In this section we look deeper into GANs that solve de-identification.

In 2018, Wu et al. [68] presented a GAN for face de-identification called Privacy-Protective-GAN (PPGAN) which focused on keeping the generated face natural and preservation of the attributes skin color and age group. The person is de-identified but share similar luminance, contrast and structure with the original image. The reported de-identification rate jumps between 84.7% for white, senior male to 100% for black youth. The generated sample images in the report look similar to the original images. For these reasons we do not find this solution suitable to this thesis.

Ren et al [57] use adversarial training with videos as input. The task of the network is to de-identify faces while maximizing action detection performance. The network does not focus on age and gender preservation, requiring further research for such purposes. Their results on binary face verification show that the verification is successful two of three times, meaning only 1/3 of the images are de-identified enough to not be recognized. This de-identification rate is too low for the problem statement of this thesis.

Hukkelaas et al. [35] designed an architecture called DeepPrivacy that de-identifies faces by exclusively generating new faces on annotations that keeps the original pose and background using image inpainting, which is the task of filling in missing areas of an image. This architecture was further improved to overcome known inpainting issues using imputed convolutions and MSG-GAN [34]. This network does not intend to preserve age or gender. To de-identify a variety of images, they generated a dataset crawling images from Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [63], a dataset with approximatly 99 million photos which is collected from the online tool Flickr. These images are of diverse quality and poses, and is therefore more suitable for de-identifying CCTV footage than traditional image datasets like CelebA.
Li et al [43] designed a network called AnonymousNet which comprises facial attribute estimation, privacy-metric-oriented face obfuscation, directed natural image synthesis, and adversarial perturbation. Their method makes it possible to manipulate the facial privacy in a photo-realistic fashion that can be used in different application scenarios. The reported results show that their method has less structural similarity than blurring, pixelation, masking and inpainting, thus higher diversity from the original image. They focus on 40 different attributes, using random forest fed by deep image features to learn each attribute and use this attributes to provide a new identity. AnonymousNet is found to be less suitable to solve the research question because attribute alteration to create a new identity does not ensure de-identification.

Gafni et al. [19] propose a method for face de-identification that can modify a video at high frame rates, minimizing the correlation of identity while preserving the pose, illumination, and expression. They do not provide a singular de-identification rate but from the reported results there seem to be high variety of de-identification in the videos, and the identities are mostly recognizable after de-identification. Observations of reported images and videos [1] confirms this statement.

In 2020, Chen et al. [10] proposed a novel image privacy preservation method to balance privacy protection and utility by generating realistic faces that matches the key face attributes of the original image. They suggest a model which start by a DCGAN generating a new image. Then both the original and generated image are labeled using a detection algorithm. To transfer the original labels to the generated image, they use an attribute-driven GAN named StarGAN. More about StarGAN is found in the next Section (2.2.6). The reported results suggest 89.1% accuracy for correct labeling of 13 attributes and that the de-identified images have a larger distance between original and created images than blurring and pixelation. Unfortunately, images generated by DCGAN contain a new background in addition to a new face. For the problem statement of this thesis, it is simpler if the background stay the same during the de-identification process because it requires less post-processing.

Croft et al. [14] looked into design and applications meant to preserve privacy, both in a database-setting and the obfuscation of facial images while preserving some utility like age and gender. The research suggests that GAN exceeds k-Same algorithm for de-identification, and implements a model using an attribute-driven GAN named AttGAN, which they call Differential Privacy. AttGAN is altered to gain more control over each attribute change. The authors question shortcut connections in encoder-decoder networks, stating that shortcut connections can leak sensitive information about images. They therefore make the encoding stochastic. This compromises the task of generating realistic images thus noise is injected into the encoding before it is passed to the decoder. To keep pose information, pose metrics are added to the classification network as a part of the loss-function. For training they extract the head from the background using Mask R-CNN [29] so the network only learns images of a head with a white background. To handle white spots due to image overlaying they implement a GAN based pluralistic image completion using facial inpainting. From the presented results, this method is better at de-identifying people than DeepPrivacy because it changes the whole head area, not only the masked part of the head.

All mentioned solutions were evaluated for this thesis, some of them found to be more relevant than others. Differential Privacy is the closest to what this thesis want to achieve. The alterations to AttGAN make it safer to use for de-identification as the neural network does not contain information needed to restore an image. Since the generated images are based on an attribute-driven GAN, one can build upon the design to include age groups as well as gender. Due to unavailable source code, this direction was not further investigated [2]. DeepPrivacy does provide similar privacy security in the network since the generator never sees the sensitive face data. It does preserve more information than Differential Privacy, like hair and ears, and because the inpainting use the remaining pixels within the face to generate the new image, it most likely will use the same skin tone. The biggest problem using this network is preserving age and gender, to accomplish this we therefore suggest using an attribute-driven GAN in addition.

2.2.6 Attribute-Driven GANs

Attribute-driven GAN is a terminology we use to describe GAN architectures which focus on changing specific attributes within an image. This subsection will contain information about and rationale of different attribute-driven GANs.

StyleGAN [39] introduced by Karras et al. in December 2018, is created to enable control

[1] Live Face De-Identification Video Samples by Gafni et al. [19] can be found here.

[2] We contacted Croft, which declined to share the code due to commercial use

of the image synthesis process in image generators and can be used to create realistic-looking faces with the ability to fine-tune facial attributes. Creating such images with selected facial attributes is accomplished by re-designing the generator architecture in a way where the generator starts from a learned constant input and adjusts the "style" or attributes of the image at each convolution layer based on two latent codes entering before and after a randomly selected cross point. This technique regularizes the network from assuming certain attributes are correlated. Noise is injected into the network in order to automatically and unsupervised separate attributes from stochastic variation while simultaneously enable scale-specific mixing of interpolation operations. Such mixing decorrelates neighboring styles and allows for better control of the generated images. The input is mapped to an intermediate latent space W, which controls the generator through Adaptive Instance Normalization (AdaIN). Before each AdaIN, Gaussian noise is added to improve the quality of the generated image.

In 2020, Kerras et al. published a new version of StyleGAN, StyleGAN v2 [40], which redesigned and improved StyleGAN. The new version improve image quality issues by redesigning the generator normalization, the progressive growing has been changed, and the generator has been regularized to encourage good conditioning in the mapping from latent codes to images. The normalization is redesigned to remove these characteristic artifacts observed in StyleGAN. The new, alternative normalization method bases the normalization on the expected statistics of the incoming feature maps. By using a skip generator and a residual discriminator, they achieve the same goals as progressive growing without changing the topology during training. Also, the synthesis network has been regularized to favor smooth mappings, improving the image quality.

In 2019, Abdal et al. proposed a technique, Image2StyleGAN [1], for inputting custom images into the latent space of StyleGAN. They use an embedding algorithm to map a given image into the extended latent space of StyleGAN, which in this case has been pre-trained on the FFHQ dataset. Three operations are used on vectors in the latent space: linear interpolation, crossover, and the addition of a vector and a scaled difference vector. These operations correspond to the image processing applications of morphing, style transfer and expression transfer.

In 2020, Abdal et al. proposed Image2StyleGAN++ [2], extending Image2StyleGAN. In this version, noise space optimization is used to restore high frequency features in images to increase the quality of reconstructed images. It is found that stable noise optimization can only be conducted if the optimization is done sequentially with the local latent space, not jointly. The global latent space embedding algorithm has also been extended in order to enable local modifications, and embedding has been combined with activation tensor manipulation in order to perform high quality local and global semantic edits in images. Image2StyleGAN++ does not appear to be focusing on solving or improving the limitations stated in Image2StyleGAN, such as time use and inheriting artifacts from StyleGAN.

Image2StyleGAN does have some limitations, including inheriting image artifacts that are present in the pre-trained StyleGAN and latency which may limit interactive editing. Although the experiments conducted in Image2StyleGAN do not contain the specific functionality that is needed for our use case (change of age and gender), it is noted that it enables control in latent space.

Abdal et al. [3] designed a model, named StyleFlow, that allows attribute-conditioned semantic edits on real images and StyleGAN generated images. They explore the latent spaces of a pretrained GAN and utilizes normalization flow maps to sample an n-dimensional distribution conditioned on the target attributes. In addition to the paper, StyleFlow presents a user interface that allows the user to perform editing on both real and generated images.

In 2020, Härkönen et al. described a technique, named GANSpace [28], to analyze GANs and, in an unsupervised manner, create controls for image synthesis. GANSpace can be applied to existing, trained GANs and consists of applying Principal Component Analysis in either latent or feature space to identify important latent directions. By layer-wise perturbation along the principal directions, a large number of controls can be defined. GANSpace is applied to both BigGAN [8] and StyleGAN [39] to demonstrate control over elements such as viewpoint, lighting, time of day and aging. GANSpace contains a user interface where it is possible to select layers and change different components to create changes in the image. These can be labeled and saved. This makes it possible to control existing general-purpose image representations rather than train a new model for each different task. For this thesis, adding GANSpace to the existing image generation method in DeepPrivacy may allow change of age and gender inside the process, instead of using a separate tool after the de-identification process.

Existing image-to-image translation has limited scalability in handling more than two image domains, making it necessary to build different models independently for every pair of domains. Choi et al. propose StarGAN [12], which is a new approach enabling image-to-image translations for multiple domains using a single model. The generator takes both image and domain information as inputs and generates a fake image as output according to the chosen domain. The fake image is then inputted into the discriminator, which, in addition to deciding whether the image is real or fake, also classifies the targeted domain. In addition, domain classification loss is used to make sure that the output image is classified to its target domain. Reconstruction loss is also used, which focuses on preserving the content of the input image while only the domain-related part is changed. StarGAN learns the mappings between multiple domains using a single generator, but it does not capture the multi-modal nature of the data distribution because each domain is indicated by a label. In 2020, Choi et al. released version 2 of StarGAN [11] which focuses on generating diverse images across multiple domains. In the new version, the domain label is replaced with a domain specific style code representing diverse styles of a specific domain. Their results show that StarGAN v2 give move diversity and better visual quality.

He et.al. designed AttGAN [30] in order to manipulate single or multiple attributes of faces while preserving the original face. They share some similar objective functions of StarGAN [12], which was developed independently in parallel, releasing their report five days ahead. Their method is based on unsupervised learning, strengthening the weakness of finding a training set containing faces correctly labeled for the combination of attributes of interest. Their model consists of a generator with an encoder-decoder structure, an attribute classifier, and a discriminator. The model utilizes two decoders, one decodes an image with a classification constraint using a desired label while the other does reconstruction learning using the original label. The first process ensures attribute change while the other ensures identity preservation. The classifier estimates the attributes, and the decoder classifies real and fake images. The network has three trainable components, namely attribute classification, reconstruction learning, and adversarial learning.

For our use, adding GANSpace to the existing image generation method in DeepPrivacy may provide the possibility of choosing a specific age and gender when generating the image. Although the experiments conducted in Image2StyleGAN do not contain the specific functionality that is needed for our use case, it seems like it has the potential. Both GANSpace and StyleFlow are methods that can provide control of attribute changes in StyleGAN. Using either GANSpace or StyleFLow together with Image2StyleGAN may result in the ability of using a custom input image as well as controlling the changes that are applied in latent space. The simplicity of using only one single model, puts Star-

GAN and AttGAN at advantage compared to be an easier alternative for changing age and gender than combining Image2StyleGAN with StyleFlow/GANSpace. As StarGAN v1 and AttGAN preserve background, these seem to be the two solutions that deliver better at what we are requesting in our application.

2.2.7 Tracking People in CCTV Footage

Some of the utility that should be preserved, is the ability to track people and follow their actions. In order to preserve information about the people's actions and whereabouts it is necessary to preserve or generate some information that makes re-identification possible. This contradicts the goal of de-identification but can be hold to the multimodel de-identification standard as long as biometric, soft biometrics, and non-biometric identifiers are removed.

There exist different methods for tracking people, one example is face tracking, while another is whole-body-tracking. For whole-body-tracking one can use a neural network that is capable of re-identifying a person based on it its whole body from bottom to head, both front and back. Such an algorithm could be more suitable on CCTV footage because the faces can be unclear, have too low resolution, and also work on people that is not facing the camera. Ahmed et. al [4] studied person re-identification and found it to be a difficult task because the changeability of pose, lighting, and camera view cause a broad set of images. A re-identification network is often presented two images, and should, based on the content, be able to guess if it is the same person (true) or two different people (false).

Simple Online and Realtime Tracking (SORT)

Alex Bewley et al. proposed a new, minimalistic approach to multiple object tracking (MOT) in 2017 called SORT [7]. This approach focuses on providing both accuracy and speed for online and real-time applications where the tracker is only presented with detections from the current and previous frame, ignoring re-identification and issues regarding occlusion as this introduces undesired complexity. SORT consists of four parts: a detection part, an estimation part, a data association part, and a creation/destruction of track identity part. The detection part uses the Faster Region CNN detection framework with two stages; the first stage extracts features and proposes regions, whereas the second stage classifies the object in the proposed regions. Only person detection results with output probabilities greater than 50% are passed to the tracking framework. The estimation part consists of propagating the detections from the current frame to the next using a linear constant velocity model. A Kalman filter framework is used to solve the velocity components to update target state when a detection is associated to a target. In cases where no detections are associated to the target, the state is predicted without correction using the linear velocity model.

The data association part of SORT uses the intersection-over-union (IOU) distance between each detection and all predicted bounding boxes from the existing targets in order to assign detections to existing targets. The assignment is solved using the Hungarian algorithm. In situations where the overlap is less than IOUmin, the assignment is rejected. The last part of SORT is track identity creation and deletion for when objects enter and leave the image. In situations where a detection has an overlap less than IOUmin, the existence of an untracked object is signified, and a new track identity is created. If a track identity is not detected for TLost frames, it gets terminated in order to prevent an unbounded growth of track identities. The performance of SORT has been evaluated using a set of testing sequences set by the MOT benchmark database, containing both moving and static camera sequences. FrRCNN(VGG16) is used for detection, and the experiments are run on a single core of an Intel i7 2.5GHz computer with 16 GB RAM. Results show that the tracker runs at 260Hz and achieves best in class performance, balancing speed and accuracy.

In 2017, Alex Bewley et al. published a second tracking report, introducing SORT with Deep association metric, namely DeepSORT [67]. DeepSORT replaces the association metric from SORT with a new metric that combines motion and appearance information. This is done by implementing a CNN that is trained on a large person re-identification dataset. Their results show that the number of identity switches is reduced by 45% compared to the original SORT, showing that it is more robust against occlusion. As occlusion can be expected in CCTV footage, it is probably more suitable for video surveillance footage.

You Only Look Once (YOLO)

In 2016, Joseph Redmon et al. presented a real-time object detector named YOLO [56]. YOLO works by dividing an input image into an SxS grid. For each grid cell, bounding boxes, confidence of the bounding boxes, and class probabilities are predicted. The confidence score reflects how confident YOLO is that the bounding box contains an object and how accurate the box is. The grid cell that contains the center of an object is responsible for detecting that object. Regardless of how many bounding boxes there are in a grid cell, one grid cell only predicts one set of class probabilities. For evaluation, YOLO is pretrained on the ImageNet 1000-class competition dataset and evaluated on the PASCAL VOC detection dataset. It is shown that YOLOs processing time is 45 frames per second using a Titan X GPU, enabling real-time detection with low latency. It is also able to achieve more than twice the mean average precision compared to other (at that time) state-of-the-art real-time systems.

YOLOs limitations include spatial constraints on bounding box predictions because each grid cell only predicts two boxes and is limited to one class, limiting the detection of multiple objects in a small space. It also struggles to generalize to objects in new aspect ratios or configurations, and the loss function creates limitations by treating errors in small bounding boxes and large bounding boxes in the same way, despite a small error in a small box having a larger impact than a small error in a large box. Also, compared to the 2016 state-of-the-art detectors, YOLO struggles with small objects.

In the end of 2016, Joseph Redmon et al. introduced **YOLOv2** and **YOLO9000** [54]. The focus of the goal was to improve the recall and localization for YOLO, while maintaining the classification accuracy. YOLOv2 improved the architecture, introducing batch normalization in order to improve convergence and eliminate the need of other regularization methods, increasing mAP more than 2%, and increasing the classifier resolution. The changes resulted in almost 4% increase in mAP. Instead of predicting the coordinates of bounding boxes directly using fully connected layers as done in YOLOv1, the fully connected layers are removed and replaced with anchor boxes to predict bounding boxes. Following this, instead of predicting the class from the spatial location, class is predicted for every anchor box. This change lowers the mAP by 0.3, but increases the recall from 81% to 88%. YOLO is also tested using dimension clusters with directly predicting the bounding box center location, which is shown to improve YOLO by almost 5% more than the version using the anchor boxes. The detector has access to finer-grained features to improve the localization of smaller objects, increasing the performance by 1%.

The original YOLO used a custom network based on the Googlenet architecture as a base feature extractor, which in YOLOv2 has been replaced by Darknet-19, improving the accuracy. The authors also propose a method to jointly train on object detection and classification, and use this to train YOLO9000, which consists of the YOLOv2 architecture but limits the output size by using 3 priors instead of 5, on the COCO detection dataset and the top 9000 classes from the ImageNet classification dataset. The new training method allows for prediction of detections for object classes that are lacking labeled data. YOLO9000 is evaluated on the ImageNet detection task and is able to achieve 19.7 mAP overall and 16.0 mAP on the unlabeled data.

In 2018, Joseph Redmon et al. released a paper for **YOLOv3** [55], containing several improvements to the previous YOLOv2/YOLO9000 architecture. Changes in the YOLOv3 include that the objectness score is now predicted for each bounding box using logistic regression, prediction across three different scales has been implemented, and softmax has been replaced by independent logistic classifiers in order to support multiple classes in a bounding box. Also, the Darknet-19 network for feature extraction has been replaced by Darknet-53, consisting of 53 convolutional layers. YOLOv3 struggles with medium to large objects.

2.3 Potential Approaches for this Thesis

Here we evaluate the possibilities of answering the research question. There exist different approaches that will solve the problem, some already mentioned in the previous section. For solutions similar to our own application, like PP-GAN, AnonymousNet, and “Design and Applications of Differentially Private mechanisms”, we were not able to find or receive code.

We discussed multiple approaches for preserving age and gender group when de-identifying footage, including the following four approaches. For all approaches, age and gender must be detected before the de-identification, and when the new face is ready it needs to replace the original face in the frame. Note that the following approaches are theoretical approaches that we evaluated after researching the field. Only one of the approaches will be used in this thesis.

2.3.1 Only StarGAN/AttGAN

The first approach would be using only StarGAN or AttGAN to both de-identify and preserve age and gender, where one would choose the domain/attribute corresponding to the original age and gender.

As this is the simplest approach, it is also the easiest to implement and test. However, the face may not get de-identified enough as it seems like StarGAN/AttGAN does not always apply enough changes. Also, the changes that the GAN networks apply are reversible, which is not at all optimal for de-identification.

2.3.2 Tracker, DeepPrivacy and StarGAN v2

The second approach consists of using a person tracking network, DeepPrivacy, and StarGAN v2. The method would first be using a tracker to link an ID to a person throughout the video clip and assigning a reference face to the ID, which contains the original age and gender. Then, a new face would be generated using DeepPrivacy, which afterwards would be merged with the assigned reference photo using StarGAN v2. The reason why the reference face is merged with the face generated from DeepPrivacy, is to keep the original pose of the face (which DeepPrivacy preserves when de-identifying).

This approach would minimize potential flickering when de-identifying a video, as well as ensure trackability, as one person would always keep the same face. It also ensures de-identification. A limitation of this approach is de-identifying crowds, as this would require a big database of reference faces with different combinations of ages and genders. Also, a crowd may complicate the tracking process.

2.3.3 DeepPrivacy and StarGAN/AttGAN

The third approach consists of generating a new face using DeepPrivacy and use StarGAN/AttGAN on the generated face with a domain/attribute that corresponds to the original age and gender.

This approach seems fairly simple and ensures de-identification. However, it does not include tracking and will therefore not ensure that a person keeps the same face throughout the video clip.

2.3.4 DeepPrivacy - Changes in latent space

The fourth approach consists of applying the method used in StyleFlow/GANSpace to DeepPrivacy v1, in order to enable control of age and gender during the generation of an image (in latent space).

The first version of DeepPrivacy and first version of StyleGAN is built upon the same generator structure (progressive GAN), so it is possible that some of the changes may be convertible. This would be a very interesting approach and should be investigated further. However, as we have little to none experience in latent space changes, and the time period for this thesis is limited, this approach may not be optimal now.

2.3.5 Architectural description of DeepPrivacy, StarGAN and AttGAN

In this subsection we will present the architectures for the chosen approach; DeepPrivacy, StarGAN, and AttGAN.

DeepPrivacy

We incorporate the newest version of DeepPrivacy [34] illustrated in Figure 2.1. The architecture builds upon a MSG-GAN design but differentiates from the traditional MSG-GAN by summarizing RGB outputs from each resolution instead of matching each resolution to the discriminator. z is a latent variable. Empty circles indicate a encoder-decoder connection (U-net connection), while circles with a plus sign indicate a residual connection. The pose information is pre-processed into a feature bank using two Fully-Connected Neural Network (FCNN)layers and concatenated to the features from the encoder.

Instead of using traditional convolution layers, they use *Imputed Convolution* (IConv) which replace uncertain values with an estimate from spatially close features. This ensures proper handling of the masked area and generates images that are visually pleasing. They combine multiple feature maps from different layers in order to only propagate features with a high certainty from shallow layers[3].

The authors have also incorporated a way of de-identificating videos which follows the pipeline process illustrated in Figure 2.2. Each frame is individually processed, starting with a face detection network that detects all faces that are present in each particular frame. The faces are then annotated with face location and pose information. Before further processing, each face is rotated and scaled to fit the generator size, which in this

[3] If the reader is interested in more details on this architecture, please take a look at “Image Inpainting with Learnable Feature Imputation” [34].

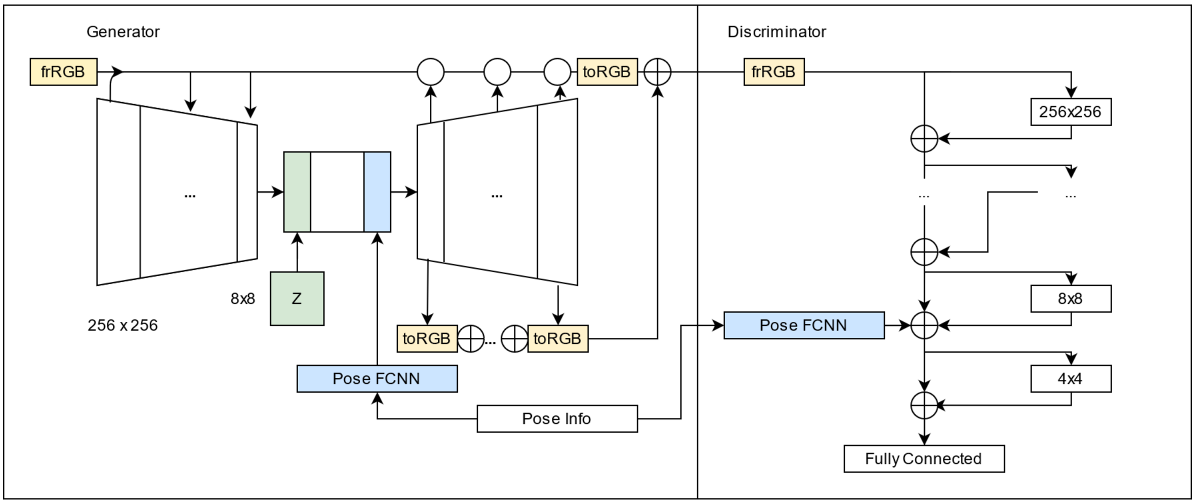

Figure 2.1: DeepPrivacy architecture inspired by figure found in [34, p.7].

case is 128×128. The faces are masked before inpainting. It is possible to do different types of masking, the default being a black rectangle over the whole face. After inpainting, the generated face is pasted upon the previous face, and then rotated and scaled back to its original parameters. When each face is given a new identity, the next frame is processed. If no faces are detected in a frame, the process goes directly to the next frame.

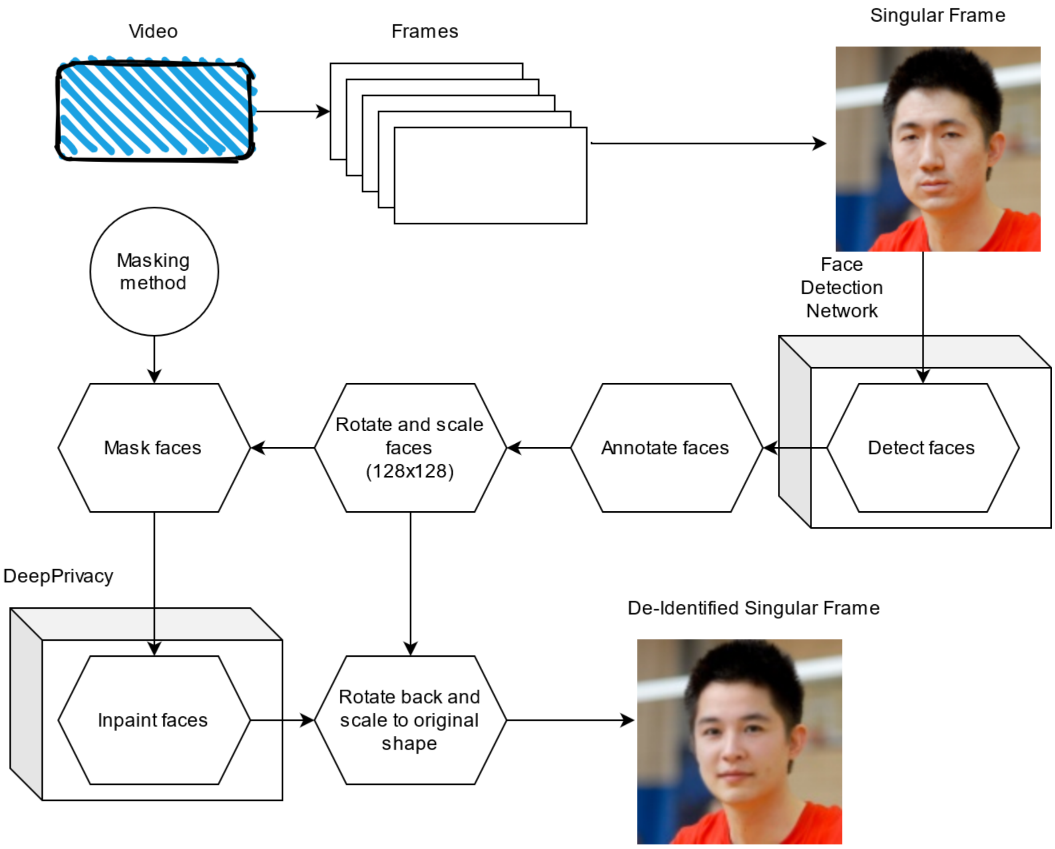

Figure 2.2: Inpainting process of DeepPrivacy retrieved from reading DeepPrivacy's source code.

StarGAN v1

Figure 2.3 shows the architecture of the StarGAN V1 [12]. The architecture consists of an discriminator and a generator, where the generator takes image and target domain as input and generates a fake image, and then tries to reconstruct the image based on the fake image and its original domain. This process ensures that the new target domain is learned while the background information is kept. The generator network of StarGAN uses instance normalization for data normalization.

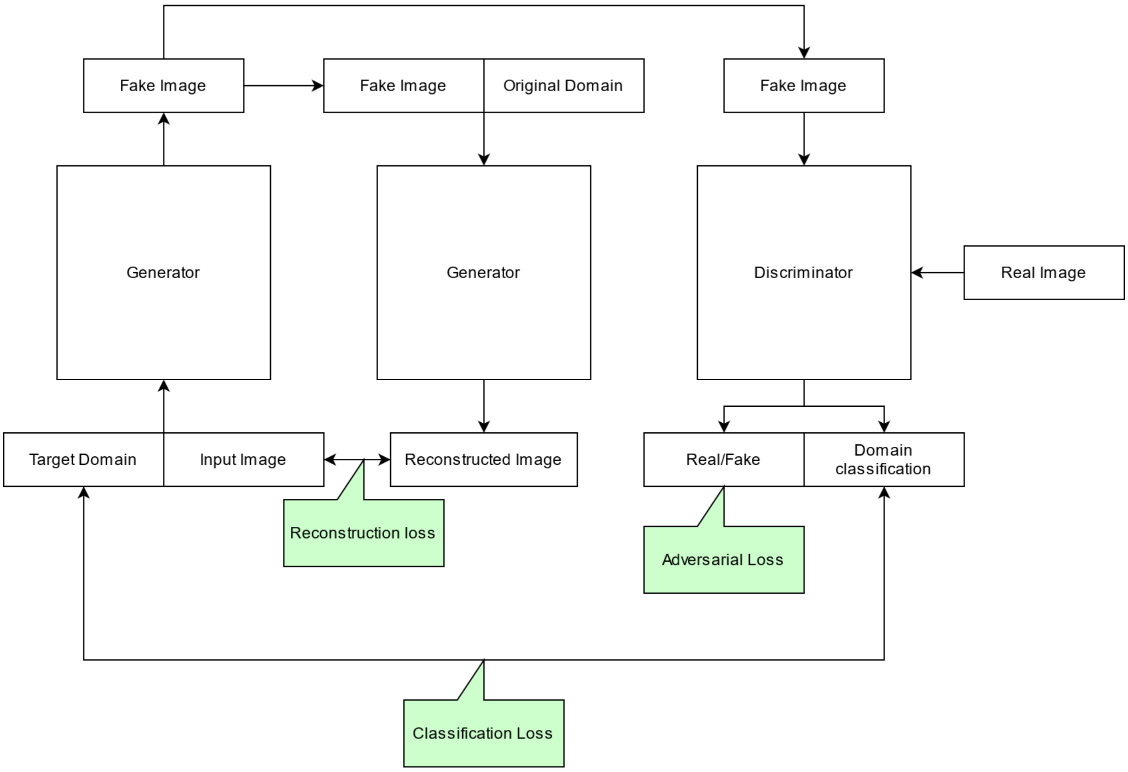

Figure 2.3: The architecture of StarGAN v1 inspired by the figure found in [12, p.8791].

The figure illustrates two additional losses to the adversarial loss, namely classification loss to ensure that the face has the right attributes and a reconstruction loss to keep as much as possible true to the original image. Each loss is defined below.

The adversarial loss uses the maximum-likelihood method from adversarial training,

$$L_{adv} = \mathbb{E}_x[\log D_r(x)] + \mathbb{E}_x[\log(1 - D_r(G(x,b))], \tag{2.10}$$

where G generates an image G(x,b) conditioned on both input image x and target domain label b.
An auxiliary classifier is used on top of the discriminator to ensure the generated image have the correct target domain. The classification process should optimize both the discriminator and generator. The discriminator receives classification penalty based upon how well real images are classified L^r_{cls}. The generator is penalized based on the classification loss of fake images L^f_{cls}.

$$L^r_{cls} = \mathbb{E}_{x,a}[-\log D_{cls}(a|x)], \tag{2.11}$$

where $D_{cls}(a|x)$ denotes a probability distribution over the original domain label a.

$$L^f_{cls} = \mathbb{E}_{x,\hat{b}}[-\log D_{cls}(b|G(x,b))], \tag{2.12}$$

where $D_{cls}(b|G(x,b))$ denotes a probability distribution over the target domain label a.

The last loss, is the reconstruction loss L_{rec}, which is computed by the following formula

$$L_{rec} = \mathbb{E}_{x,b,a}[||x - G(G(x,b),a)||_1], \tag{2.13}$$

where $G(G(x,b),a)$ should translate the image back with the original target domain. The reconstruction loss is found using the L1 norm.
The training objective for StareGAN is

$$\begin{aligned} D_{loss} &= -L_{adv} + \lambda_{cls} L^r_{cls}, \\ G_{loss} &= L_{adv} + \lambda_{cls} L^f_{cls} + \lambda_{rec} L_{rec}, \end{aligned} \tag{2.14}$$

where λ_{cls} and $\lambda_{rec} L_{rec}$ are hyper-parameters that control the importance of the corresponding losses. The default value of λ_{cls} is 1 and λ_{rec} is 10.

AttGAN

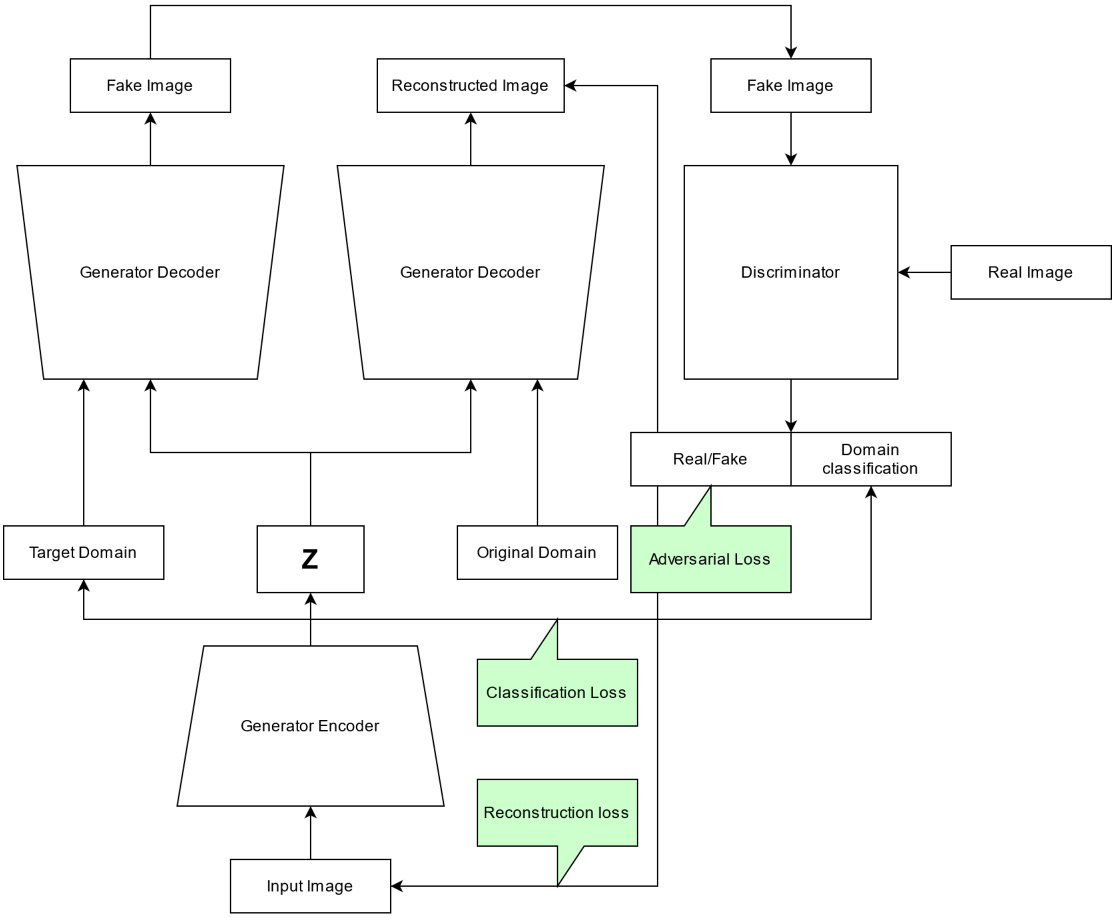

Figure 2.4: AttGAN Architecture inspired by the figure found in [30, p.5467].

The architecture of AttGAN [30] is quite similar to StarGAN, however, uses a encoder-decoder with latent space z instead of a cycle-dependency (see Figure 2.4). Instead of two generators it has two decoders, one for generating a new image with target label and one for reconstructing the image using the original label. The network uses WGAN-GP as adversarial loss, where the discriminator has the following loss

$$\min_{||D||_{L\leq 1}} L_{adv_d} = -\mathbb{E}_{x^a} D_r(x^a) + \mathbb{E}_{x^a} D_{r,b}(x^{\hat{b}}), \tag{2.15}$$

where x^a denotes original image x with original label a, $D_{r,b}$ is the distribution over the real images and label b, and $x^{\hat{b}}$ is $G_{dec}(G_{enc}(x^a), b)$. The generator network of StarGAN uses batch normalization for data normalization.
The generator should minimize L_{adv_g} for the encoder and both decoders:

$$\min_{G_{enc}, G_{dec}} L_{adv_g} = -\mathbb{E}_{x^a} D_{r,b}(x^{\hat{b}}). \tag{2.16}$$

Like StarGAN v1, the architecture has an auxiliary classifier, however, in this architecture both discriminator and generator are penalized with L^r_{cls} and L^f_{cls} respectively.

$$\begin{aligned} L^r_{cls} &= \mathbb{E}_{x,a}\left[\sum_{i=1}^{n} -a_i \log C_i(x^a) - (1-a_i)(1 - C_i(x^a))\right], \\ L^f_{cls} &= \mathbb{E}_{x,b}\left[\sum_{i=1}^{n} -b_i \log C_i(x^{\hat{b}}) - (1-b_i)(1 - C_i(x^b))\right], \end{aligned} \tag{2.17}$$

where $C_i(x^a)$ indicates the prediction of the i^{th} attribute.
L1 norm is used for reconstruction loss L_{rec},

$$\mathbb{E}_{x,b,a}[||x^a - x^{\hat{a}}||_1]. \tag{2.18}$$

The full training objective is then

$$\begin{aligned} D_{loss} &= L_{adv} + \lambda_{cls_d} L^r_{cls}, \\ G_{loss} &= L_{adv} + \lambda_{cls_g} L^f_{cls} + \lambda_{rec} L_{rec}, \end{aligned} \tag{2.19}$$

where λ_{cls_d}, λ_{cls_g}, and λ_{rec} are hyper-parameters for balancing losses. Their default values are 1, 10, and 100 respectively.

Chapter 3

Proposed Method and Datasets

This chapter contains information about the proposed method, including its network architecture and training details, as well as the datasets proposed in this thesis together with labeling details.

3.1 Proposed Method

The proposed method uses the data sets, attributes, network, and other settings that are found to be best for the use case of this thesis. Details and results of testing different approaches can be found in Chapter 4.

The first step of the proposed method uses a separate age and gender estimation network to find the original age and gender, which is used to create target labels. These labels will later be used to inform the attribute-driven GAN of the original attributes, so that it makes sure that the de-identified face contains the same attributes. In this thesis, the existing Age-gender-estimation network has been used[1], which is a CNN trained on IMDB-WIKI[2]. Another step in the pre-prosessing is to erase the face in the original image. This is done so that the neural networks that are to de-identify the face and preserve attributes are unaware of the original face, creating a new face independently of the original face.
DeepPrivacy is used for de-identification, as it is robust against occlusions, different poses, lighting and image quality, and is found to have a de-identification rate of 97.40% (found in 4.1). AttGAN is used for preservation of age and gender, and change of skin color. After researching different attribute-driven GANs, StarGAN and AttGAN was seen as the most prominent alternatives for the purpose of changing age and gender. Results in 4.4.1 and 4.5.2 show that AttGAN is most suitable.
The attribute preservation is done by inputting the de-identified image into AttGAN together with the target labels reflecting the original age and gender, as well as the desired skin tone. AttGAN then changes the attributes in the image to the target attributes, meaning if the target gender is male, AttGAN will try to change the gender of the person in the image to male. The resulting image is a de-identified image with original gender and age group, and desired skin color.

We propose two versions of the network, one without skin tone, "DeepPrivacy and AttGAN" (DP-ATT), and one with skin tone, "DeepPrivacy and AttGAN with Skin tone" (DP-ATT-S). Figure 3.1 shows a simplified illustration of the versions. The black arrows shows steps that are used in both versions, while the blue arrow shows steps specifically for DP-ATT and the red arrow shows the steps specifically for DP-ATT-S. x is the original image, x^m is the original image with a masked face, x^d is the the identified image, x^{da} is the original

[1] The code for Age-gender-estimation can be found here.

[2] The IMDB-WIKI dataset can be found here.

image with preserved age group and gender, and x^{db} is the original image with preserved age group and gender, and skin tone changed to light.

The same mechanisms as DeepPrivacy are utilized for image pre-processing, but here age and gender information are also detected and sent into the proposed network together with the erased face. In the version with skin tone, information about the desired skin tone is also included. After being de-identified by DeepPrivacy, AttGAN uses the target labels to reconstruct the de-identified image with original gender, age group and, if using the version with skin tone, a chosen target skin tone. The de-identified image on top of the figure is the result after DeepPrivacy alone, and the right-most image is the overall result. The right-most top image being the output using the version without skin tone and the right-most bottom image being the output using the version with skin tone.

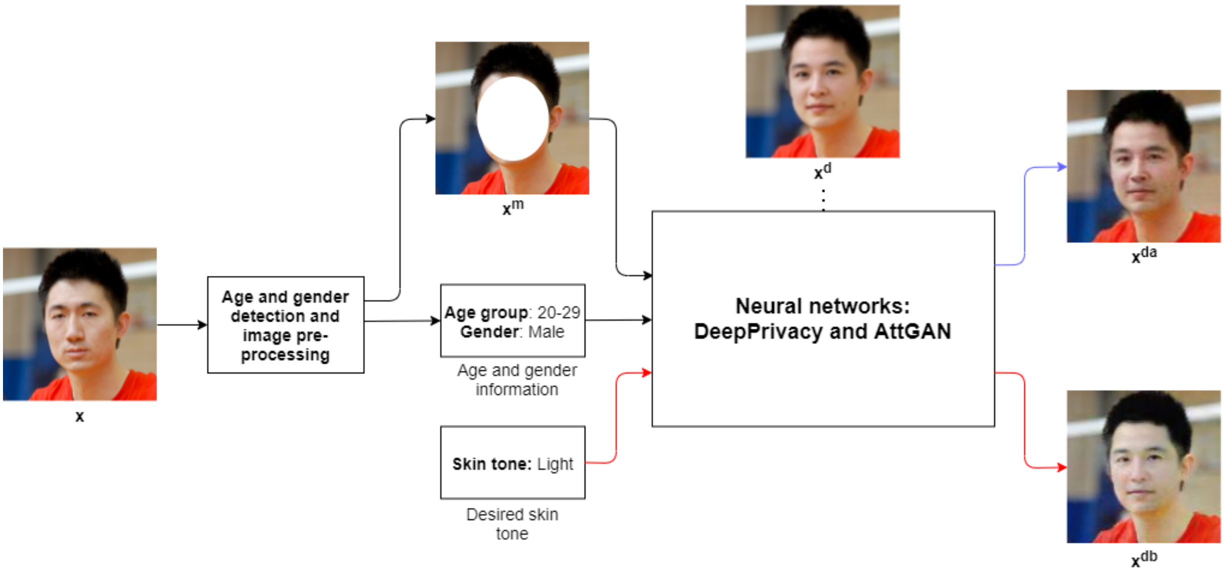
x
Age and gender detection and image pre-processing
x^m
Age group: 20-29
Gender: Male
Age and gender information
Skin tone: Light
Desired skin tone
x^d
Neural networks: DeepPrivacy and AttGAN
x^da
x^db

Figure 3.1: Illustration of the proposed method.

3.1.1 Network Architecture

Table 3.1 shows the network architecture of the AttGAN encoder and decoder for DP-ATT-S, while table 3.2 shows the network architecture of the AttGAN encoder and decoder for DP-ATT. Table 3.3 shows the architecture of the AttGAN discriminator and classifier, and is identical for both the implementation with skin tone and the implementation without skin tone.

The network architectures are very similar to the architecture of the original AttGAN [30]. The only difference is that Instance Normalization (IN) is used instead of Batch Normalization (BN) for the version that includes skin tone (see 4.5.5 for the experiment where the use of IN instead of BN is tested).

Table 3.1: Network architecture of the encoder and decoder for the DP-ATT-S.

Encoder	Decoder
Conv(64,4,2), IN, Leaky ReLU	DeConv(1024,4,2), IN, ReLU
Conv(128,4,2), IN, Leaky ReLU	DeConv(512,4,2), IN, ReLU
Conv(256,4,2), IN, Leaky ReLU	DeConv(256,4,2), IN, ReLU
Conv(512,4,2), IN, Leaky ReLU	DeConv(128,4,2), IN, ReLU
Conv(1024,4,2), IN, Leaky ReLU	DeConv(3,4,2), Tanh

Table 3.2: Network architecture of the encoder and decoder for DP-ATT.

Encoder	Decoder
Conv(64,4,2), BN, Leaky ReLU	DeConv(1024,4,2), BN, ReLU
Conv(128,4,2), BN, Leaky ReLU	DeConv(512,4,2), BN, ReLU
Conv(256,4,2), BN, Leaky ReLU	DeConv(256,4,2), BN, ReLU
Conv(512,4,2), BN, Leaky ReLU	DeConv(128,4,2), BN, ReLU
Conv(1024,4,2), BN, Leaky ReLU	DeConv(3,4,2), Tanh

Table 3.3: Network architecture of the discriminator and classifier for the proposed network.

<table>
<tr><th>Discriminator</th><th>Classifier</th></tr>
<tr><td colspan="2">Conv(64,4,2), LN/IN, Leaky ReLU</td></tr>
<tr><td colspan="2">Conv(128,4,2), LN/IN, Leaky ReLU</td></tr>
<tr><td colspan="2">Conv(256,4,2), LN/IN, Leaky ReLU</td></tr>
<tr><td colspan="2">Conv(512,4,2), LN/IN, Leaky ReLU</td></tr>
<tr><td colspan="2">Conv(1024,4,2), LN/IN, Leaky ReLU</td></tr>
<tr><td>FC(1024), LN/IN, Leaky ReLU</td><td>FC(1024), LN/IN, Leaky ReLU</td></tr>
<tr><td>FC(1)</td><td>FC(13), Sigmoid</td></tr>
</table>

3.1.2 Training Details

In this subsection, the training details of the two versions of the proposed network are presented. For simplicity, the details that are the same as for the original AttGAN are not mentioned.

DP-ATT-S

The version with skin tone is trained using “Diverse Faces with Distinct Skin tones” with the gray light skin tone and the attributes female, dark_skin, 0-19, 20-29, 30-39 and 40+. Related to the change from BN to IN, the number of discriminator updates per generator update has been changed to 1 (instead of 5 which the original AttGAN use).
A combination of MS-SSIM and L1 is used for reconstruction loss instead of L1 alone (see 4.5.6 for experiment). The alpha value of this loss is set to 0.84 and compensation is set to 0.2 (see 4.6 for more details).

DP-ATT

The version without skin tone is trained using the extended version of “Diverse Faces” with the attributes female, 0-19, 20-29, 30-39 and 40+.
Similar to the version with skin tone, the combination of MS-SSIM and L1 is adopted instead of L1. The same compensation and alpha values are employed here as in the version with skin tone.

3.2 Datasets

To our knowledge, there are no datasets of same size as CelebA containing single-face images with a variety of image quality and pose together with labels on age and gender, and a good representation of all age groups. Therefore, to train the proposed schemes, two different datasets are created. One contains only original skin tones, namely “Diverse Faces”, and the other one contains two distinct skin tones, i.e., “Diverse Faces with Distinct Skin tones”. In what follows, we will present in detail the two created data sets.

3.2.1 Diverse Faces

Both AttGAN and StarGAN v1 use CelebA or versions of CelebA for training - training sets that do not represent the diversity of CCTV-footage as these images are normally frontal images of good quality. There were several reasons to believe that AttGAN and StarGAN v1 would perform quite well on images of more diversity, because both focus on preserving as much of the original image as possible. We had a theory that a more diverse dataset would complement the diverse images generated by DeepPrivacy, being more robust against small abnormalities in these images (e.g., pose, how close the face is, and lighting).

CelebA has a training set of size 202,599 [44], thus we decided to acquire the same amount of data. We looked for datasets with the same diversity as DeepPrivacy's **Flickr Diverse Faces (FDF)** dataset. FDF, as is, does not contain gender and age labeling, so we looked for other datasets that might have such information and found **FairFace**. Due to its diversity and size, we labeled some of the images from FDF so they could be used for training as well.

The FairFace dataset [37] focuses on balancing the distribution of different races (specifically White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino). It consists of 108,501 labelled images that are collected from the YFCC-100M Flickr dataset, and is divided into validation and training sets. The labels for each image include gender, age-group, and race. Each image contains only one face.

The FDF dataset [35] focuses on variety in poses, occlusions, backgrounds, and people. It consists of 1.47M images with bounding box and key point annotation for each face in the image.

Original Version

The original version of our proposed dataset, Diverse Faces, consists of 158,635 images for training and 10,954 images for validation. The FairFace validation set is used for validation, while the FairFace training set is combined with images from FDF to form a larger training set.

Expanded Version

The expanded version is Diverse Faces expanded with additional FDF images. Previously, the validation set consisted of only Fairface images, while in this version the previous Fairface images have been added to the training set and the validation set has been replaced by new FDF images. The validation set consists of 43,119 images, and the training set consists of 180,492 images.

Test Dataset

The test dataset consists of FDF images that have been de-identified using DeepPrivacy. Age-gender-estimation, a network that estimates the age and gender of faces in images, is used to find the age and gender on the images both before and after de-identification. There are several attribute target files created for the test dataset, reflecting the attribute label files created for the different datasets. The age and gender that were detected before de-identification are configured as the target age and gender.

The target files are made with the restriction of maximum 150 images per the age groups of 0-19, 20-29, 30-39 and 40+ in order to balance the distribution, which consists of 1200 images in total.

3.2.2 Diverse Faces with Distinct Skin tones

This subsection will describe how the Diverse Faces dataset was prepared in order to include skin tone.

Beige Light Skin Tone

It was created two copies of the Diverse Faces dataset, one for dark skin tone and one for light skin tone. "Skin detection" was used to change the skin color, with RGB 227,208,202 for the light skin color and 111,55,55 for the dark skin color.
As the skin color changing network was unable to process some of the images, such images were removed from the dataset. In total, the validation set ended up with 12,363 images while the training set ended up with 163,252 images.

Grey Light Skin Tone

Random sampling during training showed that the beige light skin tone removed facial details, making the face look less natural. In some samples the skin color was seen on parts of the background as well. With some testing and random sampling, it seems like 150, 150, 150 is better for preserving the facial details and changing only the skin.
It was therefore created a different version of Diverse Faces with Distinct Skin tones with RGB 150, 150, 150 for the light skin tone. The validation set ended up with 13,779 images, and the training set with 174,334 images. It can be observed that a higher number of images survived the data preparation process for the new skin tone compared to the beige one, which presumably is due to better preservation of facial details, making more faces detectable.

Test Dataset

The only difference for this test dataset compared to the test set of "Diverse faces", is that there are two copies of the target files with one where the target for all the images is light skin, and one where the target for all the images is dark skin.

3.2.3 Labeling the Datasets

The Diverse Faces and Diverse Faces with Distinct Skin Tones datasets are labelled using the same network that is used to detect the age and gender of the original image in our proposed scheme, namely Age-gender-estimation. For images where multiple ages/genders are detected, it is assumed that the image contains multiple people. Such images are removed from the datasets. Also, the Age-gender-estimation network did have some trouble detecting faces in some of the images. To make sure we only kept the images that have detectable facial landmarks, we removed the images where the landmarks were not detected. For Diverse Faces with Distinct Skin tones, binary labels for dark/light skin tone are also added.

Chapter 4

Evaluation and Discussion

This chapter starts by looking at DeepPrivacy and how well it performs de-identification (Section 4.1).

Section 4.2 contains a traceability experiment that is conducted in order to determine how much the de-identification affects traceability and whether the observed flickering affect after de-identification using DeepPrivacy has any effect on it.

Section 4.3 introduces the methods used for evaluation of age, gender, skin color and image quality in the following experiments.

Section 4.4 focuses on the first two hypotheses, H1 and H2, and contains experiments with the main goal of preserving gender and age in de-identified CCTV footage. First, two different attribute-driven GAN networks are compared, then some changes is applied to the training dataset in order to improve the training, and at the end some parameter changes are made, also with the goal of improving training.

Section 4.5 contains experiments conducted with the additional focus of changing skin tone (H3). First a sequential run of GAN to preserve age and gender and a network to change skin tone, is tested. Then skin tone is added as an attribute to GAN. Again, experiments where changes are applied to the training dataset in order to improve training, is conducted. Towards the end of this section, the focus shifts to minimizing background noise and improving image quality. In relation, experiments are conducted where the reconstruction loss and normalization function are changed.

The last section (Section 4.6) evaluates the final, proposed solution. There are two versions of the solution, one that focuses on preserving gender and age as well as changing skin color (H1, H2 and H3), while the other version does not change the skin color (only H1 and H2).

4.1 DeepPrivacy on the matter of De-Identification

If we are to use DeepPrivacy in combination with AttGAN or StarGAN, it is important that DeepPrivacy has a good de-identification rate, because both AttGAN and StarGAN are potentially reversible given the nature of their architecture. To find the identification rate, we use a combination of a face detector, MTCNN [69], and a face verification model, LResNet100E-IR [27], trained on the MS1M-Arcface dataset [16]. We evaluate a threshold based on the samples with low distances and conclude that 0.3 should be sufficient for de-identification. Note that images with distances right below 0.3 seem to be de-identified as well, but the in-painted faces are too similar to the person that is to be de-identified. There are also some images where DeepPrivacy is unable to detect the face and the face is therefore never de-identified.

We use in total 41,094 de-identified images generated by DeepPrivacy and verify them against the corresponding original images and find the de-identification rate to be 97,40%.

4.2 Traceability in De-Identified CCTV Footage

A tracking experiment was conducted in order to determine whether and to which degree the de-identification affects traceability. Recall that in order to detect anomaly situations and maximize the utility of the de-identified CCTV video, person traceability is important. By taking a closer look at the de-identified videos, it seems like the flickering phenomenon in DeepPrivacy does not change the whole face; a person seems to keep the same facial characteristics consistent in multiple frames. It looks like the flickering is a result of rapid change of lighting. It has to be noted that under inferior conditions (such as a lot of occlusion), the facial characteristics can be inconsistent. This makes sense due to the inpainting technique of DeepPrivacy, which uses nearby pixels in order to reconstruct/in paint a face. If such neighbor pixels are changed, the in painted face may change as well. The flickering effect may be visually unpleasing, but as the de-identified videos are to be further processed by computers for the incident detection, not humans, we conclude that it is unnecessary to work on methods to prevent the effect in this thesis.

An existing, pre-trained object tracking network was used, with YOLOv3 for object detection and classification, and DeepSORT for tracking over multiple frames[1].
The dataset used for this experiment is Anomaly-Detection-Dataset, released by Waquas Sultani et al. in conjunction with their paper "Real-world Anomaly Detection in Surveillance Videos" [61]. The dataset consists of 1900 real-world surveillance videos, and contains both videos of normal events and unwanted incidents, including fighting, road accidents and burglary [2].

We used 50 videos of normal events and 50 videos of anomaly events, were the anomaly events where an even mix of abuse, arrest, arson, and assault. Note that not all videos contain visible faces. Each video is processed by the object tracking network with the class to be tracked being "person", and the number of different people found in the video is noted. The videos are then de-identified by using DeepPrivacy, and again processed by the object tracker network in order to note the number of different people in the de-identified video. The number of people in an original video is compared to the number of people in the same video after de-identification.

Before conducting the experiment, we had the following theories:

- If more people are detected in the original video than the de-identified video, it could mean that the de-identifier creates distorted faces that are untraceable.

- If the same number of people are detected in the original and de-identified video, it can mean that the de-identification did not affect the traceability at all.

- If a higher number of people are detected in the de-identified video than the original video, it could be an indication that the flickering affects the tracking process and multiple people are detected instead of one.

Table 4.1 displays the number of people found in original and de-identified videos containing normal behavior, while table 4.2 shows the same comparison for videos containing

[1] The object tracking network together with a link to the used YoloV3 weights can be found here.
[2] The Anomaly-Detection-Dataset can be found here.

abnormal behavior. The sum shows that, in both types of videos, a slightly higher number of people are found in the original videos. Looking at individual videos alone, we do not observe a prominent pattern indicating any direct conclusion about our theories. Also, after manually checking the videos that have the largest deviation, it can be seen that the network has trouble re-identifying people when filmed straight from above. For instance, the video “365.mp4” contains a total of 3 people that almost constantly are in frame, filmed straight from above. The tracking network detected as many as 49 people in the original video, and 28 in the de-identified video.

After some suspicion, we found that the de-identification done by DeepPrivacy drastically decreases the bitrate of the video. With the video “Abuse001.mp4” as an example, the original total bitrate is 1882kbps, while the total bitrate of the de-identified version is 56 kbps. This change in video quality is likely the reason for the difference in number of people found in the original and de-identified videos. From observations, it seems like some people at further distance become more unclear. These are people that originally appear unclear or so far away from the camera that they do not have to be de-identified by DeepPrivacy. The lowered bitrate is not optimal for the further processing of the de-identified videos and it should therefore be increased after de-identification in order to restore the original quality of the video.

Table 4.1: Comparing the traceability of people in original and de-identified videos containing normal behaviour.

Filename	Original	De-identified	Original minus De-Identified
003.mp4	1	1	0
006.mp4	1	1	0
010.mp4	8	7	1
014.mp4	5	5	0
015.mp4	5	5	0
018.mp4	30	28	2
019.mp4	13	12	1
024.mp4	41	42	-1
025.mp4	11	9	2
027.mp4	33	33	0
...	...	...	...
312.mp4	7	5	2
317.mp4	6	6	0
345.mp4	4	3	1
352.mp4	14	17	-3
360.mp4	17	16	1
365.mp4	49	28	21
401.mp4	24	23	1
417.mp4	3	3	0
439.mp4	72	74	-2
452.mp4	40	35	5
SUM	1562	1406	156
Standard deviation			8.83
Average			3.12

4.3 Evaluation Methods

The generated images are evaluated on their capability of preserving gender and age group, changing skin color and retaining image quality. This section will present the

Table 4.2: Comparing the traceability of people in original and de-identified videos containing abnormal behaviour.

Filename	Original	De-identified	Original minus de-identified
Abuse001.mp4	17	16	1
Abuse002.mp4	25	22	3
Abuse003.mp4	18	14	4
Abuse004.mp4	72	74	-2
Abuse005.mp4	14	10	4
Abuse006.mp4	39	33	6
Abuse007.mp4	7	7	0
Abuse008.mp4	37	45	-8
Abuse009.mp4	11	11	0
Abuse010.mp4	11	9	2
...	...	...	...
Assault004.mp4	32	30	2
Assault005.mp4	22	23	-1
Assault006.mp4	153	164	-11
Assault007.mp4	12	12	0
Assault008.mp4	175	175	0
Assault009.mp4	60	61	-1
Assault010.mp4	149	124	25
Assault011.mp4	90	84	6
Assault012.mp4	31	32	-1
Assault013.mp4	45	37	8
SUM	2439	2300	139
Standard deviation			10.75
Average			2.78

evaluation methods used for the following experiments in this thesis.

Evaluating Gender

To evaluate the preservation of gender, the following steps were used:

1. The trained GAN model was used to change the gender of the test dataset, with the target gender being the gender detected by age-gender-estimation before DeepPrivacy de-identification of the test dataset.

2. Age-gender-estimation is then used to find the new gender of the images in the test dataset.

3. The total estimated is found by summarizing the number of images where a new gender is found. The percentage of total estimated is found by dividing total estimated by total test images and multiplying by 100. A face may not be estimated if it is missing facial landmarks, making the face undetectable for the age-gender-estimation network.

4. The new gender is checked against the target gender. For the percentage of correct gender, we divide the number of correct genders by the number of estimated images and multiply the result with 100.

Evaluating Age

To evaluate the preservation of age group, the following steps were used:

1. The trained GAN model was used to change the age of the test dataset, with the target age group being the group that the age detected by age-gender-estimation before DeepPrivacy de-identification of the test dataset belongs to.

2. Age-gender-estimation is then used to find the new age of the images in the test dataset.

3. The total estimated is found by summarizing the number of images where a new age is found. The percentage of total estimated is found by dividing total estimated by total test images and multiplying by 100.

4. The new age is checked against the target age group. Percentage of correct age group is found by dividing the number of images where the target group contains the new age by the number of estimated images, then multiplying by 100.
   For the Mean Square Error (MSE), if the new age is within the original age group, the result is 0. Else, the result is calculated by squaring the distance from the new age to the nearest age in the target age group. The penalty is exponential resulting in higher penalty for larger deviations. As an example, a 10-year deviation equals a penalty of 100, while a 20-year deviation are equal to 400. Figure 4.1 illustrates this method, with the red dashed lines being the distance from the new age to the nearest age in the targeted age group, age groups being the colored blocks.

Evaluating Skin Color

To evaluate the change of skin color, the following steps were used:

1. The trained GAN model was used to change the skin color of the test dataset. This is done in two batches, one where the skin color is changed to light and the other to dark.

2. Skin-detection is then used to find the RGB values of the new skin colors of the images in the test dataset [3]. The skin-detection code is run three times for each test

[3] The code for Skin-detection can be found here.

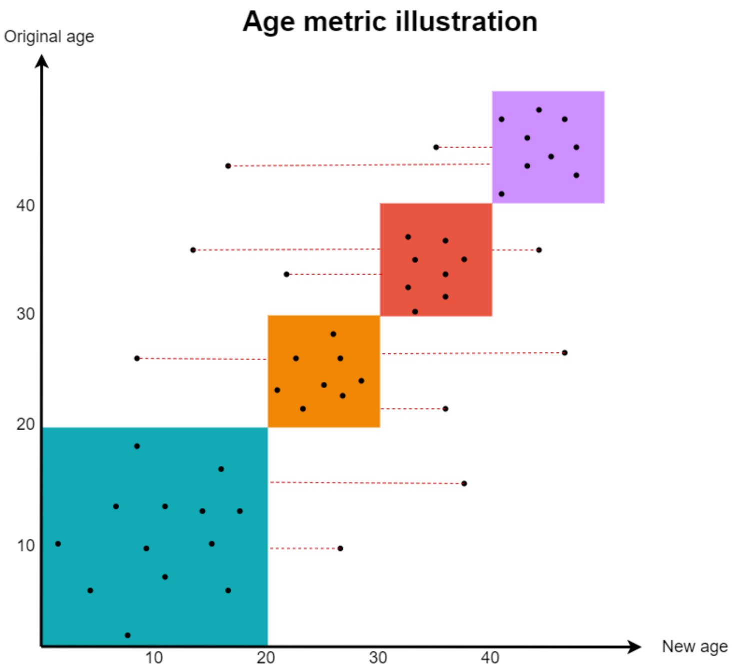

Figure 4.1: Illustration of the age metric.

dataset in order to accommodate irregularities.

3. The total estimated is found by summarizing the number of images where the skin-detection network is able to find a new RGB value. The percentage of total estimated is found by dividing total estimated by total test images and multiplying by 100.

4. The average skin color is found from the new skin colors. The new skin colors are compared to the average color in order to evaluate how close the colors are. This is done by finding the MSE between average and the estimated colors. Figure 4.2 illustrates this method for the R value, with the target value being the average value, the gray line being the correct skin tone and the dashed lines being the distance from the new skin tone to the target skin tone. The same is done to the G and B values.

Evaluating Image Quality

There are several methods of both objectively and subjectively determining image quality. A common objective method is measuring the difference, or more precise the distance between the original and the generated image. To capture a broad variety of measurements, we have used six different methods of measuring distance: L1, L2, SSIM, PSNR, FID, and LPIPS. For all mentioned distance methods, the original images and generated images are denoted A and B respectively.

L1, L2, and Smooth L1

In **L1**, also called **Absolute-value norm**, the distance is found by calculating the distance between each point from one vector to another, and then summarized. **L2**, also called **Euclidean norm**, differentiate from L1 in that distance is powered by two and then the root of the summation of the distances is found. Both distances, and other norms can be found by Equation 4.1, where p denotes the norm, a_i is a singular vector point in A, and b_i singular vector point in B. Both summarize the pixel-wise distance per image, however

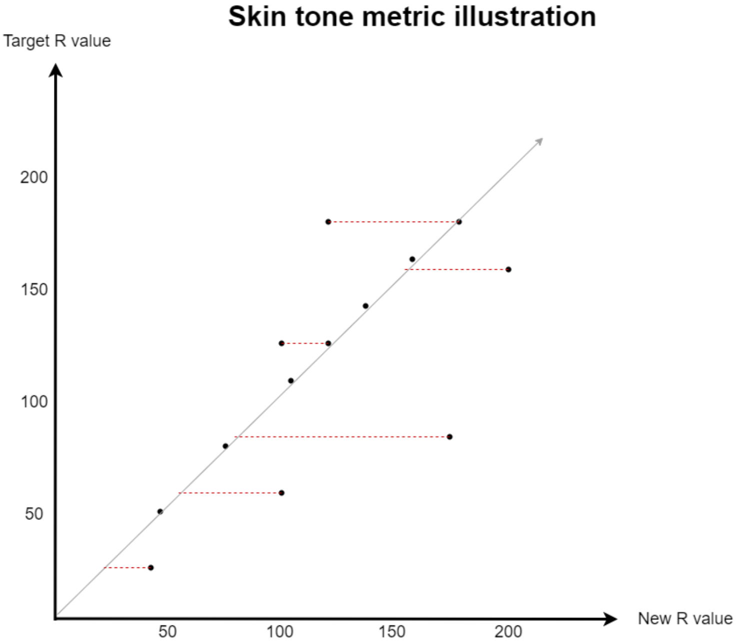

Figure 4.2: Illustration of the skin tone metric for the R value.

the L2 norm penalizes bigger distances harder.

$$\|x\|_p = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{\frac{1}{p}} \tag{4.1}$$

In **Smooth L1** also called **Huber Loss** [62], it is possible to set a threshold β, where pixels below β are more penalized than those above (see Equation 4.2).

$$loss(A, B) = \frac{1}{n} \sum z_i, \;\; z_i = \begin{cases} \frac{(a_i - b_i)^2}{2\beta}, \; if \, |a_i - b_i| < \beta, \\ |a_i - b_i| - \frac{\beta}{2}, \; otherwise. \end{cases} \tag{4.2}$$

Peak Signal-to-Noise Ratio (PSNR)

PSNR measures in general the ratio between the maximum power of signal and the power of corrupting noise. A low ratio is an indication of high noise, thus the higher ratio the better image quality. The ratio can be found following Equation 4.3, where a_i is a singular vector point in A, b_i singular vector point in B, and MAX_i is the maximum pixel value of the image which in our case is 255.

$$PSNR = 10 \times \log_{10} \left(\frac{MAX_I^2}{MSE} \right), \;\; MSE = \frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2 . \tag{4.3}$$

Structural Similarity Index Measure (SSIM)

SSIM measures the similarity between to images, using one of the images as the reference [50]. The image degradation is understood as the perceived change in structural information, and is found using three components, luminance l masking for image distortions in bright regions (Equation 4.4), contrast c masking for areas of texture (Equation 4.5), and structure s (Equation 4.6). μ_A and σ_A is the mean and variance of image A respectively, C_1, C_2, and C_3 are constants that can be altered based on image range and individual component evaluation. SSIM is the multiplication of each component as shown in Equation 4.7, where α, β, and γ make it possible to adjust how much of each component should be present.

$$l = \frac{2\mu_A \mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1}. \tag{4.4}$$

$$c = \frac{2\sigma_A\sigma_B + C_2}{\sigma_A^2 + \sigma_B^2 + C_2}. \tag{4.5}$$

$$s = \frac{2\sigma_A B + C_3}{\sigma_A\sigma_B + C_3}. \tag{4.6}$$

$$SSIM(A,B) = [l(A,B)]^{\alpha}\,[c(A,B)]^{\beta}\,[s(A,B)]^{\gamma}\,. \tag{4.7}$$

Multiscale SSIM (MS-SSIM)
MS-SSIM [66] adds more flexibility to the SSIM, making it possible to scale the contrast and structure (see Equation 4.8).

$$MS\text{-}SSIM(A,B) = [l_M(A,B)]^{\alpha M} \prod_{j=1}^{M} ([c_j(A,B)]^{\beta j}\,[s_j(A,B)]^{\gamma j}). \tag{4.8}$$

MS-SSIM + L1
MS-SSIM + L1 [71] is a combination of MS-SSIM and L1, evaluating both measures in order to find a distance as shown in Equation 4.9. α decides which one of MS-SSIM and Gaussian L1, $\acute{L}1$, should have the most impact, if $alpha$ is larger than 0.5, MS-SSIM has the greatest impact on the measured distance.

$$MS\text{-}SSIM + L1 = \alpha MS\text{-}SSIM + (1-\alpha)\acute{L}1. \tag{4.9}$$

Learned Perceptual Image Patch Similarity (LPIPS)
Zhang et al [70] found that the deep features of a trained network can be used to find a distance metric that correlates to what the human eye find to be realistic. The distance is found using a network that first computes deep embeddings, then normalizes the activations in the Channel dimension, scale each channel with vector w and finds the L2 distance which is averaged across all the layers (see Figure 4.3).

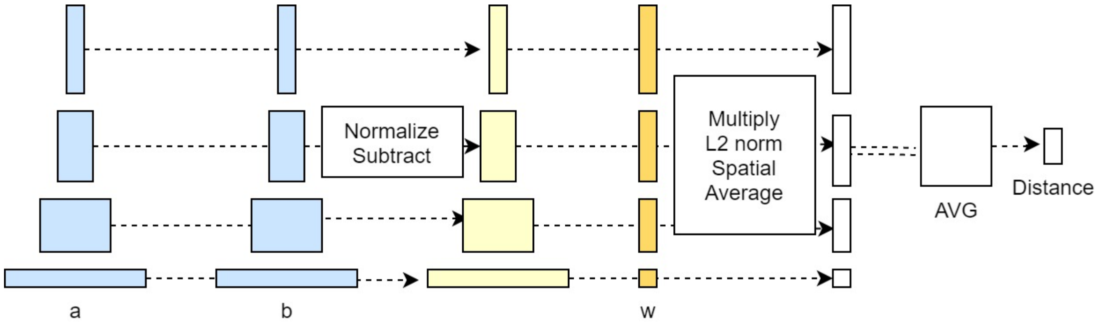

Figure 4.3: Computing distance from a network.

Fréchet Inception Distance (FID)
Heusel et al. [32] introduce a distance measure that could portray how well generative images fit to the observed data, involving the distribution of real data $\mathbb{P}_r$ and the distribution of the generated data $\mathbb{P}_\theta$. FID can be found using Equation 4.10, μ_r, μ_θ and C_r, C_θ is the mean and covariance-matrices of $\mathbb{P}_r$ and $\mathbb{P}_\theta$ respectively. The lower the score, the more similar the generated distribution is to the distribution of real images.

$$FID = ||\mu_\theta - \mu_r||_2^2 + tr\left(C_\theta + C_r - 2(C_\theta C_r)^{\frac{1}{2}}\right), \tag{4.10}$$

where tr is the trace of the squared matrix summing the diagonal from upper left do lower right.

Visual comparison of image quality metrics
Figure 4.4 visually compares the different metrics for evaluation of image quality. It can

be seen that low L1 and L2 values reflect subtle changes in the image, while high L1 and L2 values reflects more aggressive changes. There can also be seen a correlation between L1 & L2 and PSNR, as images with low L1 & L2 seem to have a high PSNR value and vice versa. A low PSNR value shows that the new image has more noise compared to the original image, while a high PSNR value shows that the new image has less noise. For SSIM, it can be seen that a low value reflects low similarity for luminance, contrast and structure, and vice versa for a high value. Recall that LPIPS and FID values are the lower the better. This is also reflected in the figure, as the images with low values have considerably more natural changes than the images with high values.

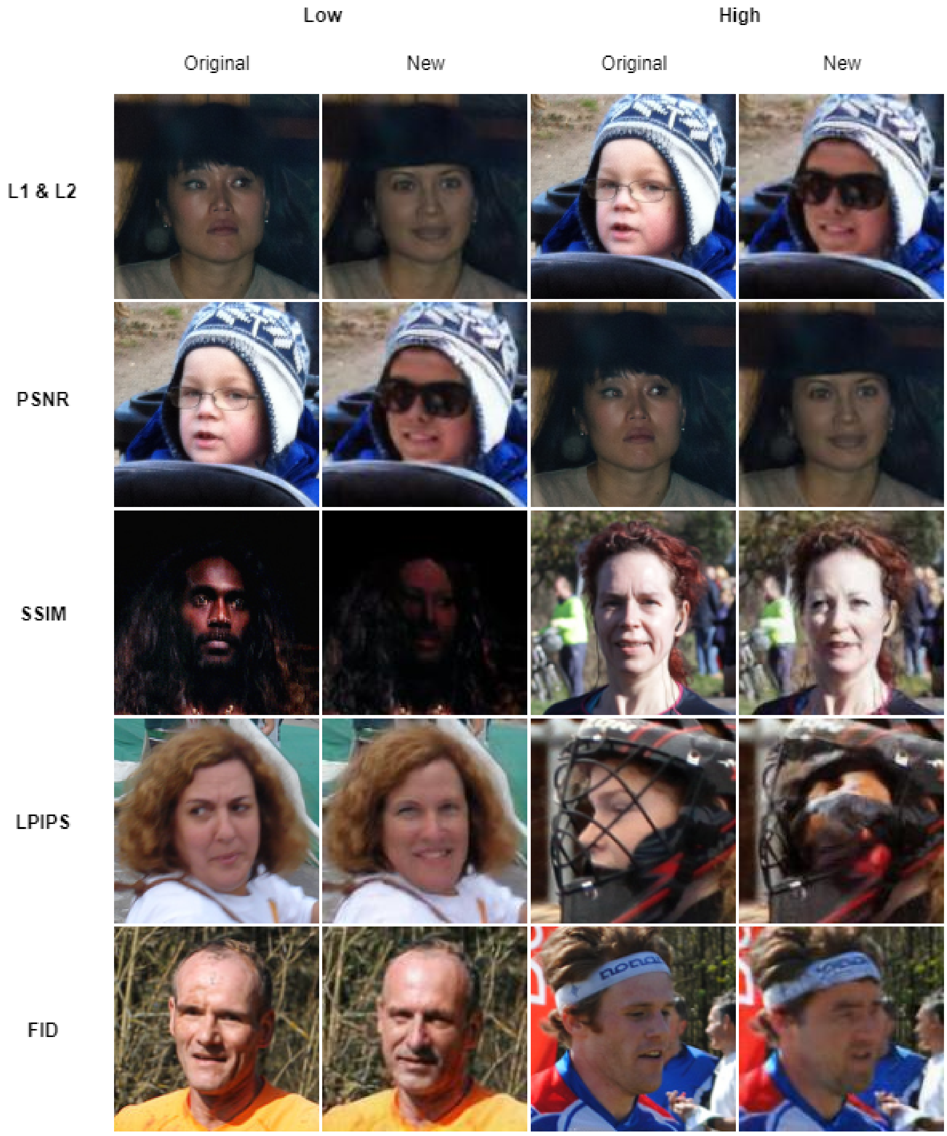

Figure 4.4: Visually comparing the different evaluation metrics.

4.4 Preserving Age and Gender in De-Identified CCTV Footage

Recall the different approaches in rationale. Ensuring that one person has the same face throughout the video is not in the scope anymore, the third approach using only DeepPrivacy and StarGAN/AttGAN will be used in this thesis.

This section focuses on preserving age group and gender in footage that have been de-identified by DeepPrivacy. It should be noted that some of the experiments in the next section, “Removing Skin tone Information in De-Identified CCTV Footage”, have been conducted simultaneously as some experiments in this section. The experiments have been divided into two separate direction because of their different goals. Because of this, some of the lessons learned in the two sections were learned at the same time.

StarGAN and AttGAN do not focus on de-identification, in fact their processes are designed to be reversible. Recall from Subsection 2.3.5, StarGAN has two generators, one for generating new images based on target attributes, and one for reconstructing original images using generated images and original attributes. AttGAN has a similar property for one of the decoders. For these reasons the de-identification rate of DeepPrivacy is considered to be the maximum rate when combined with an attribute-driven GAN, even if the combination results in a higher de-identification rate.

4.4.1 AttGAN vs StarGAN

For the first round, the goal was to find whether StarGAN v1 or AttGAN is the best alternative for our purpose. The results of StarGAN v1 are from iteration 200000, which is the default stopping point, while the results of AttGAN are from epoch 60. The default stopping point for the used AttGAN implementation is 200 epochs. However, it was observed that the model did not get better after 60 epochs. Omitting 140 epochs was also significantly timesaving.

Distinct age groups with balanced dataset

In this experiment, the original version of Diverse Faces is used, with the FDF images labeled by age-gender-estimation. The images where the facial landmarks were undetectable or that contained multiple people are not used.

Label files with the attributes female, 0-9, 20-29 and 50+ were used in order to have visually distinct age groups. To balance the distribution of the different age groups in the training part of the dataset, each combination of age and gender is limited to have no more than 5000 images. Table 4.3 and Table 4.4 show the distribution of the training and validation datasets. It can be seen that females that are 50 years and older are slightly underrepresented.

Table 4.3: Distribution of the training images for “Distinct age groups with balanced dataset”.

Age group	Female	Male	Summary
0-9	5000	5000	10000
20-29	5000	5000	10000
50+	4635	5000	9635

Table 4.4: Distribution of the validation images for “Distinct age groups with balanced dataset”.

Age group	Female	Male	Summary
0-9	1323	1413	2736
20-29	1832	1468	3300
50+	472	763	1235

Table 4.5 shows the evaluation of “Distinct age groups with balanced dataset”. It can be seen that StarGAN has trouble preserving facial landmarks when reconstructing the image, resulting in a lower number of estimated images. Overall gender results show that

AttGAN performs better than using DeepPrivacy alone. The percentage of estimated and correct age group is nearly the same whether DeepPrivacy is used alone or together with AttGAN. However, MSE is considerably lower using AttGAN, meaning that AttGAN less frequently has large deviations from the correct age groups. The image quality evaluation shows that StarGAN combined with DeepPrivacy alters multiple pixels compared to DeepPrivacy alone, while DeepPrivacy + AttGAN only seem to change the required pixels. One can observe that much of the image quality is lost when using AttGAN together with DeepPrivacy. Note that image quality is evaluated for all images, not only the estimated images.

Table 4.5: Evaluation results for “Distinct age groups with balanced dataset”.

Evaluation criteria	DeepPrivacy	DeepPrivacy +StarGAN	DeepPrivacy +AttGAN
Total estimated (%)	92.92	50.75	**88.17**
Gender			
Correct gender (%)	83.41	54.19	**90.26**
Estimated & Correct (%)	77.50	27.50	**79.58**
Age group			
Correct age group (%)	45.47	19.38	**48.30**
Estimated & Correct (%)	42.25	9.83	**42.58**
MSE	75.49	290.83	**55.65**
Image quality			
L1	0.0294	0.3562	0.04865
L2	0.0601	0.4256	0.0753
PSNR	25.0227	7.6371	**22.8402**
SSIM	0.7411	0.0043	**0.5327**
LPIPS	0.0677	0.5267	**0.1466**
FID	29.5835	267.4002	**89.4002**

Figure 4.5 reflects the results in terms of noise, similarity, and visual perception. It is impressive that some evaluation can be done on images by StarGAN at all. We do not have a theory to why the images have become so dark, other than that StarGAN, as it is learning the attributes through cycle-loss, believes that the dark color is a part of the style. Other observations are that the boy at the bottom row is in painted as an older person by DeepPrivacy. Nor AttGAN or StarGAN is really capable of doing anything about this.

Distinct age and gender groups with balanced dataset

This experiment used Diverse Faces with the FDF images labeled by age-gender-estimation. The images where the facial landmarks were undetectable or that contained multiple people, were removed.

Label files with the attributes female_0-9, female_20-29, female_50+, male_0-9, male_20-29, and male_50+ were used, with the same distribution and limitation as for the previous experiment.

Table 4.6 shows the evaluation of “Distinct age and gender groups with balanced dataset”. It can be observed that the new attributes combining age and gender do not have a positive effect on gender, as the overall gender results show that DeepPrivacy gets better results used alone than with any of the GANs. Similarly to the age evaluation of the

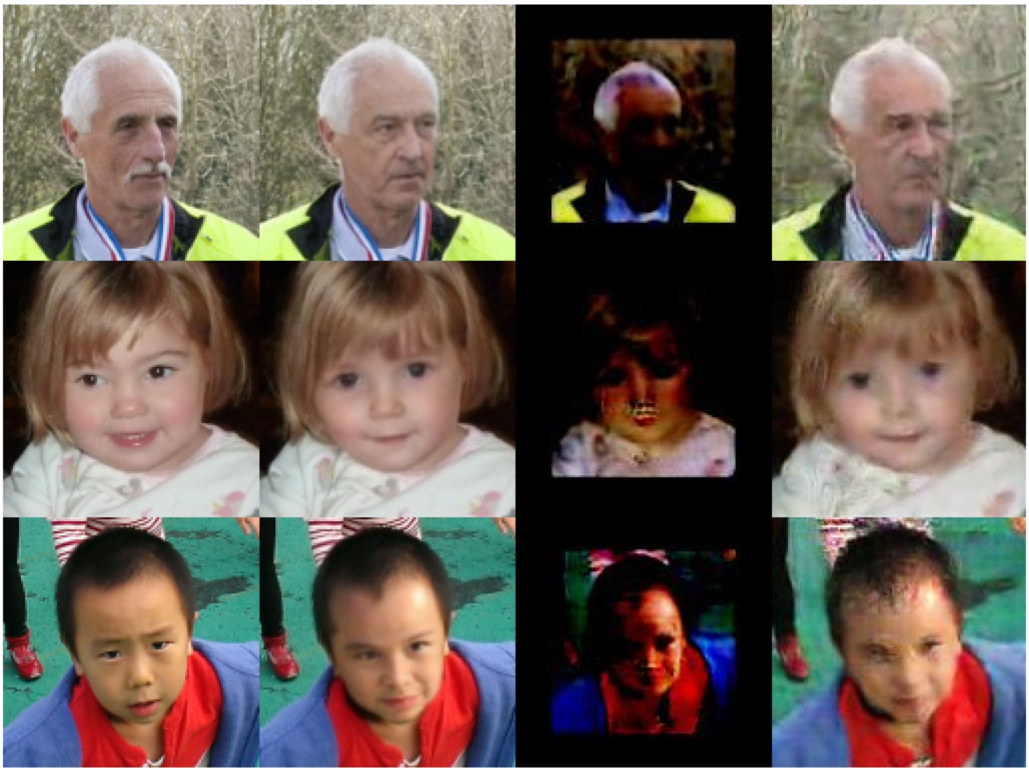

Figure 4.5: Comparing (from left) original, DeepPrivacy, StarGAN and AttGAN.

previous experiment, the percent of total estimated and correct age group shows that DeepPrivacy and AttGAN are very similar, while MSE indicates that the ages that AttGAN got wrong are closer to their target age group than the ages that DeepPrivacy got wrong. The image quality evaluation values are similar to those of Table 4.5. However, the FID score is marginally better for “Distinct age groups with balanced dataset” implying less natural changes in the images for this experiment.

Discussion: Does AttGAN or StarGAN preserve gender and age best?

Generally, it can be seen that the image quality of the reconstructed images from both of the GANs are significantly worse than image quality of the input images. It is also found that separate attributes for gender age groups provide better results than the combination of gender and age groups.
From both experiments we can see that DeepPrivacy in general produce samples with around 83% correct gender and 40-45% correct age group. Overall, as AttGAN provides better than or equal results to DeepPrivacy, AttGAN will be used for future experiments in this section.

Additional findings

After the previous experiments, we gained a suspicion about whether combining labels from two different sources was optimal. Recall that the FairFace images already had labels while the FDF images were labelled using age-gender-estimation. Therefore, in the rest of the experiments in this section, the FairFace images were relabeled by age-gender-estimation.

We also observed that as many as 60 epochs were unnecessary for AttGAN in most of the previous experiments and will therefore in upcoming experiments compare the epoch that has the lowest generator loss. Although GANs normally do not have a clear convergence and one have to observe the sample images in order to find the best epoch, the

Table 4.6: Evaluation results for “Distinct age and gender groups with balanced dataset”.

Evaluation criteria	DeepPrivacy	DeepPrivacy +StarGAN	DeepPrivacy +AttGAN
Total estimated (%)	80.33	49.25	**77.00**
Gender			
Correct gender (%)	83.20	55.50	**85.28**
Estimated & Correct (%)	66.83	27.33	**65.67**
Age group			
Correct age group(%)	43.57	14.38	**45.35**
Estimated & Correct (%)	35.00	7.08	**34.92**
MSE	156.23	157.53	**77.43**
Image quality			
L1	0.02897	0.3570	0.0479
L2	0.0589	0.4263	0.0739
PSNR	25.1385	7.6252	**22.9742**
SSIM	0.7452	0.0037	**0.5354**
LPIPS	0.0660	0.5274	**0.1396**
FID	31.2932	269.0175	**91.6393**

generator loss of Wasserstein GAN, which is used in this thesis, does show properties of convergence [6].
Also, in order to save time, early stopping was implemented for all upcoming experiments. The implementation makes sure that if the generator loss does not improve during the next 20 epochs after last improvement, the training stops.

4.4.2 Improving Training by Changes to the Dataset

This subsection focuses on improving the results of AttGAN by testing different changes to the dataset and labelling used for training.

Using the existing dataset without balance restrictions

This experiment uses Diverse Faces with the attributes female, 0-19, 20-29, 30-39 and 40+. The new age groups are selected based on observations in the previous experiments in this section and the first experiment in the next section. Recall that some of the experiments in both sections were conducted simultaneously.
Both the images from the FDF dataset and the Fairface dataset have been labeled using age-gender-estimation. As some images were lost due to undetectable facial landmarks and/or multiple people in an image, the validation set consists of 7,969 images and the training set consists of 135,283 images.
Table 4.7 and Table 4.8 show the distribution of the training and validation datasets. It can be seen that without the restriction of maximum 5000 images per combination of age and gender, the datasets are imbalanced. However, removing the restriction enables the use of a significantly larger dataset.

Table 4.9 shows the evaluation of “Using the existing dataset without balance restrictions”. As AttGAN gained better results in “Distinct age groups with balanced dataset” than in “Distinct age and gender groups with balanced dataset”, this experiment is compared to the formal results. It can be seen that removing the restriction of the dataset results in a higher number of images being estimated, and also a higher number of images with the correct gender. It can also be seen that the results for age group have been

Table 4.7: Distribution of the training images for "Using the existing dataset without balance restrictions".

Age group	Female	Male	Summary
0-19	7443	7861	15304
20-29	20031	16313	36344
30-39	20695	35777	56472
40+	5431	21732	27163

Table 4.8: Distribution of the validation images for "Using the existing dataset without balance restrictions".

Age group	Female	Male	Summary
0-19	280	270	550
20-29	1120	658	1778
30-39	1608	2416	4024
40+	318	1299	1617

improved significantly, as the network that have been trained without the data restriction is more than 30% better at preserving the correct age group, as well as having a tremendously lower MSE. The PSNR, SSIM, LPIPS and FID results have increased, indicating that the model without the balance restrictions produces images that are both in higher quality and look more natural for the human eye.

Table 4.9: Evaluation results for "Using the existing dataset without balance restrictions".

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 60): Distinct age groups with balanced datset	DeepPrivacy +AttGAN (Epoch 25): No balance restrictions
Total estimated (%)	88.17	**90.58**
Gender		
Correct gender (%)	90.27	**93.56**
Estimated & Correct (%)	79.58	**84.75**
Age group		
Correct age group (%)	48.30	**80.50**
Estimated & Correct (%)	42.58	**72.92**
MSE	55.65	**6.81**
Image quality		
L1	0.0487	0.0427
L2	0.0753	0.0693
PSNR	22.8402	**23.5994**
SSIM	0.5327	**0.5917**
LPIPS	0.1466	**0.1214**
FID	89.4002	**68.5411**

Extended dataset

This experiment uses the expanded version of Diverse Faces with all images labeled by age-gender-estimation. The images where the facial landmarks were undetectable or that contained multiple people, were removed. The attributes female, 0-19, 20-29, 30-39 and

40+ were used.
Table 4.10 and Table 4.11 show the distribution of the training and validation datasets. It is observed that the distribution is largely imbalanced. However, as seen in the previous experiment, it seems like the size of the dataset is more important than the distribution being balanced.

Table 4.10: Distribution of the training images for "Extended dataset".

Age group	Female	Male	Summary
0-19	10585	11308	21893
20-29	27063	22876	49939
30-39	25970	46703	72673
40+	7248	28739	35987

Table 4.11: Distribution of the validation images for "Extended dataset".

Age group	Female	Male	Summary
0-19	2973	3343	6316
20-29	6692	6616	13308
30-39	4843	10291	15134
40+	1816	6545	8361

Table 4.12 shows the evaluation of this experiment compared to the previous one. It can be seen that even though the "best epoch" in this experiment only is 1/5 of the epochs for the previous experiment, both the gender and age results are slightly better. It can also be seen that "Extended dataset" does not have as good image quality results as the previous experiment, which may be related to the low number of epochs. Improving the training in order to increase the number of epochs may be the way to go in order to produce better quality images. Overall, for the use case of this thesis, the focus on preservation of age group and gender is higher than the quality.

Introducing more age groups

This experiment uses the expanded version of Diverse Faces with all images labeled by age-gender-estimation. The images where the facial landmarks were undetectable or that contained multiple people, were removed. The attributes female, 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70+ were used.
Table 4.13 and Table 4.14 show the distribution of the training and validation datasets. As previously, the distribution is very uneven. However, using the entire dataset is prioritized rather than restricting age groups to a maximum number of images in order to create a balanced distribution.

Table 4.15 shows the evaluation for "Introducing more age groups". It can be seen that the percentage of correct gender and age group has decreased, while the number of estimated images have increased slightly. The latter, however, may be related to the number of epochs. It can also be seen that the image quality has increased slightly, which again may be more related to the number of epochs rather than the age groups.

Discussion: Did the changes done to the training set improve the results?

The experiments in this subsection shows that the size of the dataset used to train AttGAN is of higher importance than the balance of the distribution for the different

Table 4.12: Evaluation results for “Extended dataset”.

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 25): No balance restrictions	DeepPrivacy +AttGAN (Epoch 5): Extended Dataset
Total estimated (%)	**90.58**	89.08
Gender		
Correct gender (%)	93.56	**95.60**
Estimated & Correct (%)	84.75	**85.17**
Age group		
Correct age group (%)	80.50	**81.95**
Estimated & Correct (%)	72.92	**73.00**
MSE	6.81	**5.75**
Image quality		
L1	0.0427	0.0486
L2	0.0693	0.0750
PSNR	**23.5994**	22.8722
SSIM	**0.5917**	0.5427
LPIPS	**0.1214**	0.1363
FID	**68.5411**	88.2059

Table 4.13: Distribution of the training images for “Introducing more age groups”.

Age group	Female	Male	Summary
0-9	934	365	1299
10-19	9651	10943	20594
20-29	27063	22876	49939
30-39	25970	46703	72673
40-49	5228	20595	25823
50-59	1467	5886	7353
60-69	399	1910	2309
70+	154	348	502

Table 4.14: Distribution of the validation images for “Introducing more age groups”.

Age group	Female	Male	Summary
0-9	310	125	435
10-19	2663	3218	5881
20-29	6692	6616	13308
30-39	4843	10291	15134
40-49	1265	4497	5762
50-59	407	1452	1859
60-69	111	496	607
70+	33	100	133

Table 4.15: Evaluation results for “Introducing more age groups”.

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 5): Extended Dataset, 4 age groups	DeepPrivacy +AttGAN (Epoch 8): Extended Dataset, 7 age groups
Total estimated (%)	89.08	**90.83**
Gender		
Correct gender (%)	**95.60**	94.95
Estimated & Correct (%)	85.17	**86.25**
Age group		
Correct age group (%)	**81.95**	79.63
Estimated & Correct (%)	**73.00**	72.33
MSE	**5.75**	7.17
Image quality		
L1	0.0486	0.0424
L2	0.0750	0.0701
PSNR	22.8722	**23.5104**
SSIM	0.5427	**0.5887**
LPIPS	0.1363	**0.1212**
FID	88.2059	**72.2156**

attributes. Also, it is observed that introducing more age groups did not increase the results in regard to preserving age groups and gender.
Another observation is that it seems like the quality of the generated images as well as the percentage of estimated images are correlated to the number of epochs.

In conclusion, the increase of the dataset did improve the results. However, as the AttGAN seems to converge very early when using the expanded dataset, it may be worth looking into improvement of training.

4.4.3 Improving Training by Parameter Changes

To stabilize the training, we tried to change the number of discriminator updates per generator update (n_d). The default value in the AttGAN implementation used for this thesis is 5. This experiment uses the same dataset and attributes as the experiment “Extended dataset”.

Table 4.16 shows the evaluation of “Improving training by parameter changes”. It can be seen that the overall results for both gender and age groups are slightly improved, as well as MSE for age groups. It can also be seen that the image quality has increased, thus fine tuning the number of discriminator updates per generator update seem to have the potential to increase the stability of training.

Although, the number of epochs did not increase too much in this scenario, one can view a better trend from the training summary (see Figure 4.6). This figure compares the total generator loss and the total discriminator loss for 4 and 5 discriminator updates per generator update. Thus, for future training it is probably wise to increase the number of generator updates compared to discriminator updates.

Table 4.16: Evaluation results for “Improving training by parameter changes”.

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 5): Extended Dataset n_d 5	DeepPrivacy +AttGAN (Epoch 7): Extended Dataset, n_d 4
Total estimated (%)	89.08	**90.01**
Gender		
Correct gender (%)	**95.60**	95.10
Estimated & Correct (%)	85.17	**85.67**
Age group		
Correct age group (%)	**81.95**	81.12
Estimated & Correct (%)	73.00	**73.08**
MSE	5.75	**5.02**
Image quality		
L1	0.0486	0.0452
L2	0.0750	0.0718
PSNR	22.8722	**23.2639**
SSIM	0.5427	**0.5836**
LPIPS	0.1363	**0.1280**
FID	88.2059	**75.9000**

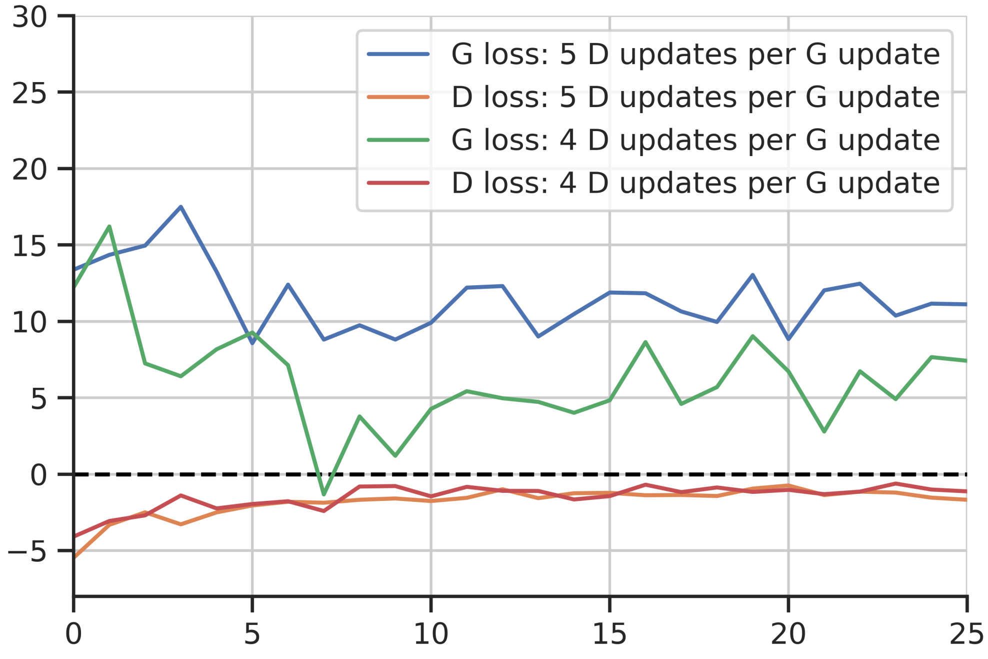

Figure 4.6: Training details from the different generator updates, where n_d is the number of discriminator updates compared to generator updates.

4.5 Changing Skin Color to Avoid Bias in De-Identified CCTV Footage

Recall that we wish to provide the option to change the skin colors of all people into one skin tone in order for the de-identified CCTV footage to bypass any potential ethnicity biases that an anomaly detection network may have. As DeepPrivacy seems to remove other ethnicity factors than the skin tone, we will only focus on changing the skin tone (not other factors such as eye color, eye shape, nose shape etc).
In this section, the focus is on not only preserving the age group and gender, but also changing the skin tone to either a light tone or a dark tone.

4.5.1 Using a Separate Network to Change Skin Tone after Image Generation by AttGAN

An experiment was conducted in order to check what the results look like when using AttGAN for preserving age group and gender, and skin-detection for changing the skin tone.
As the model for “Extended dataset” with 4 discriminator updates per generator update provided the best results in the previous section, that model was chosen for this experiment. The test dataset was first changed to the target gender and age groups by AttGAN, then the images went through skin-detection in order to change the skin tone. RGB values 111,55,55 was chosen for this experiment.

Table 4.17 shows the evaluation for “Using a separate network to change skin tone after image generation by AttGAN”. The table compares the results before and after using the skin-detection network for changing the skin tone. It can be seen that the gender and age group results decrease tremendously. As a much lower percentage of images are estimated, it seems like facial landmarks have become less detectable. In the skin color results it can be seen that of all the 1200 test images, only 54.41 % could be detected by the same skin-detection code after the color was changed. The code is supposed to give all images the same skin color which is reflected in the low MSE-value. DeepPrivacy has a high MSE value, which shows that DeepPrivacy does not try to change all skin colors to one color. The average RGB values after the use of AttGAN and skin-detection are 157,86 and 79, all within 20% of the original target value. It can also be seen that the quality has decreased considerably, with the resulting images being less natural to human eyes and containing more noise.

These observations are confirmed by the last column of Figure 4.7. AttGAN seems to have added some noise to the image, and then the skin-detection code adds even more blurriness and the face almost disappears in the background as the background becomes more disrupted.
The boxes in the top left corner of the images show the age and gender found by Age-gender-estimation. It can be seen that DeepPrivacy changes the age significantly, and that AttGAN is able to restore the age close to the original. Also note that DeepPrivacy seems to have given the man in the last row some feminine characteristics and that AttGAN has made the person more masculine, proving that it to some degree is able to change gender.

Table 4.17: Evaluation results for “Using a separate network to change skin tone after image generation by AttGAN”.

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 7): Extended Dataset, n_d 4	DeepPrivacy +AttGAN (Epoch 7): Extended Dataset, n_d 4, RGB [111,55,55]
Total estimated (%)	**90.01**	67.33
Gender		
Correct gender (%)	**95.10**	91.71
Estimated & Correct (%)	**85.67**	61.75
Age group		
Correct age group (%)	**81.12**	78.59
Estimated & Correct (%)	**73.08**	52.92
MSE	**5.75**	7.68
Skin color		
Total estimated[4] (%)	76.17	54.41
MSE	19768	5728
Image quality		
L1	0.0452	0.0947
L2	0.0718	0.1307
PSNR	**23.2639**	18.5502
SSIM	**0.5836**	0.4749
LPIPS	**0.1280**	0.2034
FID	**75.9000**	105.6964

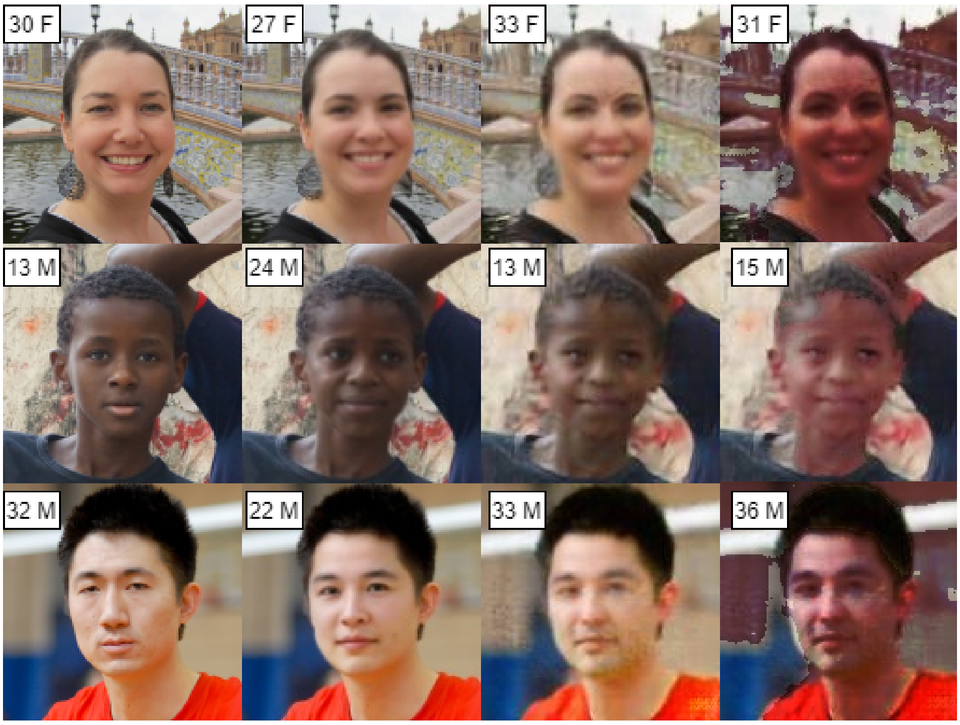

Figure 4.7: Comparing (from left) original, DeepPrivacy, AttGAN and AttGAN+Skin-detection.

Discussion: How does the skin tone affect the results?

It can be seen that using AttGAN and skin-detection sequentially in order to preserve age group and gender, as well as change skin tone, decreases image quality considerably. This may also be the reason why the percentage of estimated images decreased, as lower image quality likely results in less detectable facial landmarks.

Because of the decrease in quality and the assumption that using only one network rather than two increases usability, we decided to rather look at the skin tone as an attribute, thus the attribute-driven GAN will change gender, age, and the skin tone all in one process.

4.5.2 Changing Skin Tone as an Attribute: AttGAN vs StarGAN

This experiment focuses on whether AttGAN or StarGAN is better at changing the skin tone. The results of StarGAN v1 are from iteration 200000, which is the default stopping point, while the results of AttGAN are from epoch 60.

This experiment used Diverse Faces with Distinct Skin tones with the beige light skin tone. The attributes female, dark_skin, 0-19, 20-29, 30-39 and 40+ were used. The FDF images were labelled by age-gender-estimation, and the images where the facial landmarks were undetectable or that contained multiple people, were removed.
Table 4.18 and Table 4.19 show the distribution of the training and validation images.
From Table 4.18 one can observe that the elderly females have the lowest representation, almost half of elderly male which is the category having the most data representation. This may confuse the GANs to believe that elderly people tend to be male. In general, one can see that females are less representative in all combinations except in the age 20-29. This results in an unequal distribution of gender, resulting in a 22,572 image difference of genders. This may affect the gender attribute, making male distribution more dominant that female.

Table 4.18: Distribution of the training images.

Group	FDF dark/light	Fairface dark/light	Total dark/light	Summary
Male 0-19	3012/3000	8131/8141	11143/11141	22284
Male 20-29	4951/4924	8457/8457	13408/13381	26789
Male 30-39	6648/6611	8131/8131	14779/14742	29521
Male 40+	5370/5351	10062/10059	15432/15410	30842
Female 0-19	2596/2579	7799/7818	10395/10397	20792
Female 20-29	5040/4999	10692/10681	15732/15680	31412
Female 30-39	3312/3294	6401/6398	9713/9692	19405
Female 40+	1602/1584	6043/6041	7645/7625	15270
Total male	19981/19886	34781/34803	54762/54674	109451
Total female	12550/12456	30935/30938	43485/43394	86879

In Fairface, the distribution between light and dark is minimal, having a few representatives more for the light attribute, whereas in FDF it is the opposite and about 100 images in difference. This yields that the darker color is slightly more represented, however, not in a marginal scale. We observed that the data is not equally represented but believed that each domain is represented sufficient.

Table 4.19: Distribution of the validation images.

Group	Fairface dark	Fairface light	Summary
Male 0-19	1061	1061	2122
Male 20-29	1138	1132	2270
Male 30-39	997	993	1990
Male 40+	1253	1249	2502
Female 0-19	1006	999	2005
Female 20-29	1352	1342	2694
Female 30-39	764	761	1525
Female 40+	746	740	1486
Total male	4449	4434	8883
Total female	3868	3842	7710

The validation images, all gathered from Fairface, represent each attribute in a variation from 740 images for light elderly females to 1352 images of young, dark females. These data imply somewhat the same distribution as the training set, the exception being adult male which has a higher, relative representation in the training set.

Table 4.20 shows the results of the evaluation for “Changing Skin tone as an Attribute: AttGAN vs StarGAN”. It is observed that age-gender-estimation is only able to detect very few faces in the images in the dataset that have been changed to dark skin tone by StarGAN. This may be because StarGAN is more aggressive when making the attribute changes, possibly meaning that it adds too much dark color to the face and therefore makes the facial landmarks less visible. Using the light skin tone, a higher number of images retains a detectable face for both GAN types. Overall, StarGAN does not provide good gender results, and it is better to use DeepPrivacy alone than together with StarGAN. However, AttGAN with the light skin color provides significantly better gender results than DeepPrivacy alone, despite having a lower percentage of total estimated images. It can be observed that using AttGAN there is a relatively small difference between the age group results for light skin tone and dark skin tone, but using the light skin tone does seem to preserve the facial landmarks better, resulting in fewer images with undetectable faces. It can also be seen that AttGAN, in both skin tone situations, provides significantly better age group results than DeepPrivacy alone. Using MSE, the results of

DeepPrivacy together with AttGAN seem to be three times better than with DeepPrivacy. Note that the average RGB values that are in bold are the ones within 20% deviation of the target RGB values. It can be observed that the light skin tone results in better preserved facial landmarks due to the number of detected faces. However, the dark skin tone seems to come closer to the average skin tone. Looking at the average RGB values for each experiment, the darker color is closer to the target color. The dark average color generated by AttGAN almost hits target where the R value is 21% off target. It can also be seen that using DeepPrivacy alone achieves a lower MSE than using StarGAN in order to make the skin color light. If the goal is a dark skin color, adding StarGAN to DeepPrivacy does provide a lower MSE but makes the facial landmarks in most images undetectable. From the image quality results we observe that L1 and L2 are higher in both DeepPrivacy combined StarGAN and AttGAN, thus changing more of the images compared to DeepPrivacy alone. Because DeepPrivacy only has altered the face-part of the image, a very high score is not necessarily better in this case. StarGAN with dark skin color has the most altered pixels, thus probably provide more de-identification but at the same time has changed the most from the original image. DeepPrivacy has the images with less noise, followed by DeepPrivacy combined with AttGAN. StarGAN combined with DeepPrivacy is inferior. The results follow the same order as before, where StarGAN is at the bottom, with a terrible FID score. Images with dark skin do in general perform worse than those of light skin color.

Looking at Figure 4.8, it is not hard to understand why a face detection algorithm would fail to find face contours in the StarGAN dataset with dark skin tone. Although some of the errors probably are related to the black frames around the images, they do not look promising for further exploration. AttGAN does perform better, which is confirmed by the images. However, it has not restricted the color change to the face alone, and it seems like much of the background has changed color as well. Since the skin-detection code also changes parts of the background sometimes, it is not that strange that AttGAN performs color change in most of the image.

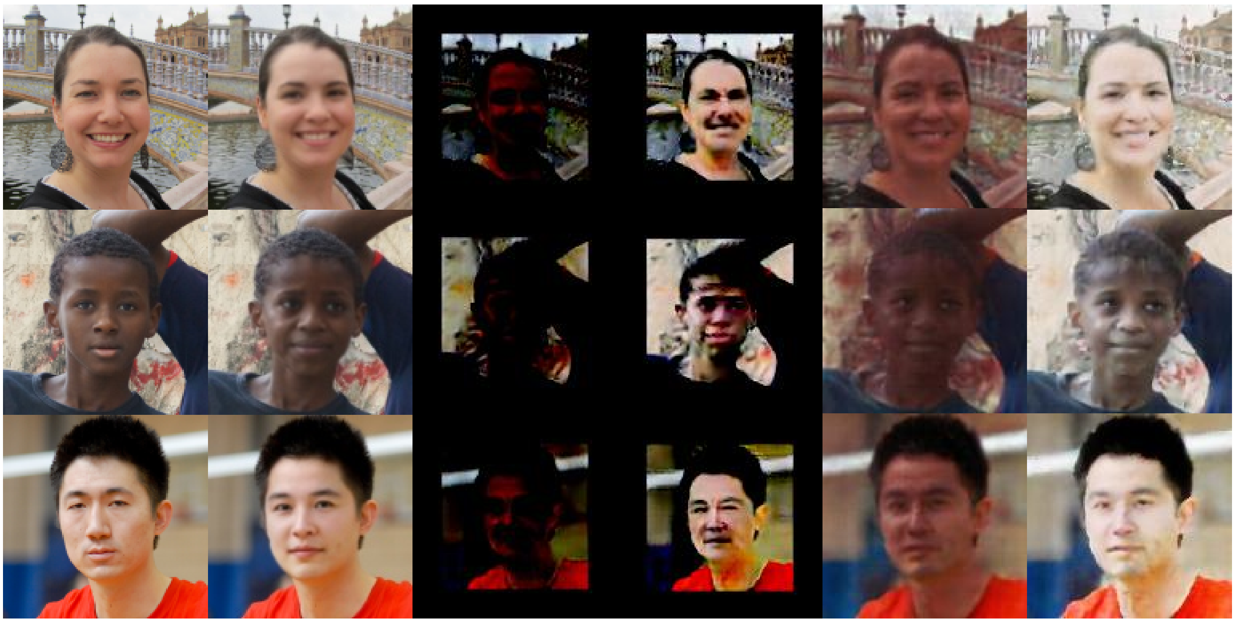

Figure 4.8: Comparing (from left) original, DeepPrivacy, StarGAN Dark, StarGAN Light, AttGAN Dark and AttGAN light.

Table 4.20: Evaluation results for “Changing Skin tone as an Attribute: AttGAN vs StarGAN”.

Evaluation criteria	DeepPrivacy	DeepPrivacy +StarGAN Dark	DeepPrivacy +StarGAN Light	DeepPrivacy +AttGAN Dark	DeepPrivacy +AttGAN Light
Total estimated (%)	93.08	11.67	84.17	88.42	**91.67**
Gender					
Correct gender (%)	85.32	55.00	54.46	87.28	**91.73**
Estimated & Correct (%)	79.42	06.42	45.83	77.17	**84.08**
Age group					
Correct age group (%)	54.16	20.00	27.43	71.54	**71.64**
Estimated & Correct (%)	50.42	2.33	23.08	63.25	**65.67**
MSE	36.99	81.04	123.62	12.80	**10.97**
Skin color					
Total estimated (%)	76.17	3.25	64.42	66.42	**76.58**
MSE	19768	9038	49001	**7622**	16763
Target RGB	-	[111,55,55]	[227,208,202]	[111,55,55]	[227,208,202]
Average RGB	[222,166,145]	[20,**11**,**13**]	[**175**,141,124]	[165,**94**,**84**]	[302,275,259]
Image quality					
L1	0.02933	0.3731	0.3226	0.1004	0.1208
L2	0.0606	0.4390	0.3980	0.1342	0.1610
PSNR	24.9555	7.3851	8.2324	**17.9679**	16.2332
SSIM	0.7423	0.0009	0.0066	0.5072	**0.5176**
LPIPS	0.0690	0.5865	0.4740	0.1910	**0.1737**
FID	30.7108	270.1649	251.0998	90.5100	**80.6915**

Discussion: Which is better of AttGAN and StarGAN?

When we observed sample images during training, it was noticed that when changing skin tone, StarGAN seemed to be preserving the background of the original image. This sparked an idea that perhaps some combination of the two networks could be able to make the necessary changes without changing the background.
As AttGAN comes closer to the target skin tone as well as having best results preserving the gender and age group, AttGAN will be used in the upcoming experiments in this section. Also, the following experiments will be compared by “best epoch”, being the epoch with lowest generator loss, rather than epoch 60.

4.5.3 Relabeling Data for Training Consistency

This experiment used Diverse Faces with Distinct Skin tones with the beige light skin tone. The attributes female, dark_skin, 0-19, 20-29, 30-39 and 40+ were used. The FairFace and the FDF images were labelled by age-gender-estimation, and the images where the facial landmarks were undetectable or that contained multiple people, were removed. Since the only difference from the data used in the previous experiment is that the FairFace images are labeled by age-gender-estimation, the distributions are very similar to the previous experiment.

Table 4.21 shows the evaluation for “Relabeling Data for Training Consistency”. It can be seen that the results of the gender preservation have increased considerably. For the age group preservation, the results for the dark skin tone have increased, while the results for the light skin tone have decreased slightly. However, overall age group results have gotten better, as the gap between the results for the light skin tone and the dark skin tone has decreased significantly. It can be seen that even though the average RGB value is further from the target RGB value than before the relabelling of data, the MSE is considerably lower. As the goal is to have all skin colors as close to one color as possible, the version using the relabelled data has the leading skin color results. Also, if the goal is changing to a dark skin tone, using DeepPrivacy alone gives a slightly lower MSE than together with the previous AttGAN model. However, the new AttGAN with relabelled data provides lower MSE than DeepPrivacy alone. The image quality results suggests a give and take scenario between the two colors, as the PSNR has become better, SSIM worse, LPIPS better, and FID worse.

4.5.4 Improving Data Quality: Changing Light Color

This experiment used Diverse Faces with Distinct Skin tones with the gray light skin tone. The attributes female, dark_skin, 0-19, 20-29, 30-39 and 40+ were used. The FairFace and the FDF images were labelled by age-gender-estimation, and the images where the facial landmarks were undetectable or that contained multiple people, were removed. Since the only difference from the data used in “Changing Skin tone as an Attribute: AttGAN vs StarGAN” is that the FairFace images are labeled by age-gender-estimation and the light skin tone has been changed to a different color, the distributions are very similar.

Table 4.22 shows the evaluation for “Improving Data Quality: Changing Light Color”. It can be seen that the new color has increased the correctness of gender preservation. Since we only have changed the light color, this suggests that the overall training has become better with the new training images. Interestingly, the age-gender-estimation network was not able to estimate a higher number of images. The new average RGB values are further away from the target values. It can also be seen that MSE for the dark and new light color deviates considerably less, as previously the MSE for the dark color was less than half of the MSE for the light color. It seems like the new light color makes

Table 4.21: Evaluation results for “Relabeling Data for Training Consistency”.

Evaluation criteria	DeepPrivacy +AttGAN (Epoch 21) D	DeepPrivacy +AttGAN (Epoch 21) L	DeepPrivacy +AttGAN (Epoch 36) D Relabelled	DeepPrivacy +AttGAN (Epoch 36) L Relabelled
Total estimated (%)	86.25	89.67	88.75	**91.83**
Gender				
Correct gender (%)	84.64	90.99	88.54	**91.29**
Estimated & Correct (%)	73.00	81.58	78.58	**83.83**
Age group				
Correct age group (%)	70.43	**75.65**	71.08	73.32
Estimated & Correct (%)	60.75	**67.83**	63.08	67.33
MSE	16.37	**8.50**	12.90	10.60
Skin color				
Total estimated (%)	76.33	**78.00**	69.17	76.41
MSE	7540	19903	6521	14989
Target RGB	[111,55,55]	[227,208,202]	[111,55,55]	[227,208,202]
Average RGB	[**161**,**91**,**80**]	[294,208,**245**]	[170,**93**,**83**]	[305,271,**251**]
Image quality				
L1	0.1073	0.1249	0.1012	0.1288
L2	0.1421	0.1676	0.1357	0.1676
PSNR	17.3745	15.7996	**17.7669**	15.9127
SSIM	0.4756	**0.4948**	0.4928	0.4855
LPIPS	0.2136	0.1924	0.1945	**0.1886**
FID	97.9805	**85.3047**	92.7679	91.2781

the skin color change more stable. The overall larger MSE for this experiment may be due to the lower number of epochs. The image quality results show that the new color have incorporated less noise into the test samples, and overall improved the visual data to human eyes.

Table 4.22: Evaluation results for “Improving Data Quality: Changing Light Color”.

Evaluation criteria	DeepPrivacy +AttGAN Epoch(36) Dark	DeepPrivacy +AttGAN Epoch(36) OLD Light	DeepPrivacy +AttGAN Epoch(18) Dark	DeepPrivacy +AttGAN Epoch(18) NEW Light
Total estimated (%)	88.75	**91.83**	87.83	91.08
Gender				
Correct gender (%)	88.54	91.29	90.89	**91.87**
Estimated & Correct (%)	78.58	79.83	81.33	**83.83**
Age group				
Correct age group (%)	71.08	73.32	**79.70**	75.16
Estimated & Correct (%)	63.08	67.33	**70.00**	68.58
MSE	12.90	10.60	10.32	**7.61**
Skin color				
Total estimated (%)	69.17	76.42	78.42	79.67
MSE	6521	14989	13096	16107
Target RGB	[111,55,55]	[227,208,202]	[111,55,55]	[150,150,150]
Average RGB	[170,**93**,**83**]	[305,271,**251**]	[182,**103**, **87**]	[261,235,220]
Image quality				
L1	0.1012	0.1288	0.0821	0.0877
L2	0.1357	0.1676	0.1123	0.1214
PSNR	17.7669	15.9127	**19.2900**	18.5915
SSIM	0.4928	0.4855	0.5169	**0.5376**
LPIPS	0.1945	0.1886	**0.1719**	0.1766
FID	92.7679	91.2781	84.4071	**83.2994**

Discussion: Did the new color improve the results?

Overall, the new color seems to have increased light results and decreased the dark results. For the gender classification, the new color gives better results for both colors. For age, the new results are more even between the colors, which is why we choose to go further with the new light color. It should be noted that a more detailed study of different colors may have improved the overall results further, however, due to time limitation, it is not further explored in this thesis.

4.5.5 Removing Background Noise from Skin tone Change

For this experiment, the same setup is used as for “Improving Data Quality: Changing Light Color”. The only difference is that batch normalization for the generator encoder and decoder has been replaced with instance normalization. This was done because of the observation that StarGAN was better at preserving the background than AttGAN when changing skin tone, and when comparing the two networks, the type of normalization was one of few differences.

The training was quite unstable, and we ended up updating the generator as often as the discriminator, although in separate iterations.

For the following experiments, we focus only on the image metrics as the task at first hand is to see if we are able to only change the skin tone. Table 4.23 compares the distance metrics of the model design with Batch Normalization (BN) to the model design with Instance Normalization (IN), and Table 4.24 compares the same metrics, however, only for everything that is not skin. This is accomplished by using the skin-detection code to extract everything from an image that is not skin, resulting in the image being black where the skin was, and one should therefore be careful to put too much attention of the PSNR, SSIM, LPIPS, and FID score, as the images certainly have some noise and are not meant to be pleasant to the human eye.

What can be observed from Table 4.23, is that instance normalization has decreased noise levels, is more structurally similar to the original image and more pleasant to the human eye. Table 4.24 shows that instance normalization alters less background than the batch normalization, but the difference is minor. Both L1 and L2 are larger for the lighter color independently of BN and IN.

Table 4.23: Image quality evaluation results for “Removing Background Noise from Skin tone Change”.

Evaluation criteria	DeepPrivacy +AttGAN Epoch(18) Dark BN	DeepPrivacy +AttGAN Epoch(18) NEW Light BN	DeepPrivacy +AttGAN Epoch(18) Dark IN	DeepPrivacy +AttGAN Epoch(18) NEW Light IN
L1	0.0821	0.0877	0.0570	0.0820
L2	0.1123	0.1214	0.0826	0.1129
PSNR	19.2900	18.5915	**22.0893**	19.5913
SSIM	0.5169	0.5376	**0.5892**	0.5604
LPIPS	0.1719	0.1766	**0.1208**	0.1507
FID	84.4071	83.2994	**71.5074**	75.6025

Figure 4.9 confirms the objective measurements, perhaps to a more satisfying degree than the numbers indicate as one can easily observe that the color change is more restricted to the skin area.

Table 4.24: Image quality evaluation results for "Removing Background Noise from Skin tone Change, comparing background only".

Evaluation criteria	DeepPrivacy +AttGAN Epoch(18) Dark BN	DeepPrivacy +AttGAN Epoch(18) NEW Light BN	DeepPrivacy +AttGAN Epoch(18) Dark IN	DeepPrivacy +AttGAN Epoch(18) NEW Light IN
L1	0.0714	0.1129	0.0636	0.0989
L2	0.1533	0.2146	0.1441	0.1957
PSNR	17.0695	14.1961	**17.8068**	15.1294
SSIM	0.3380	0.3048	**0.3739**	0.3140
LPIPS	0.2954	0.3376	**0.2636**	0.3245
FID	87.3035	139.2901	**77.6610**	116.1191

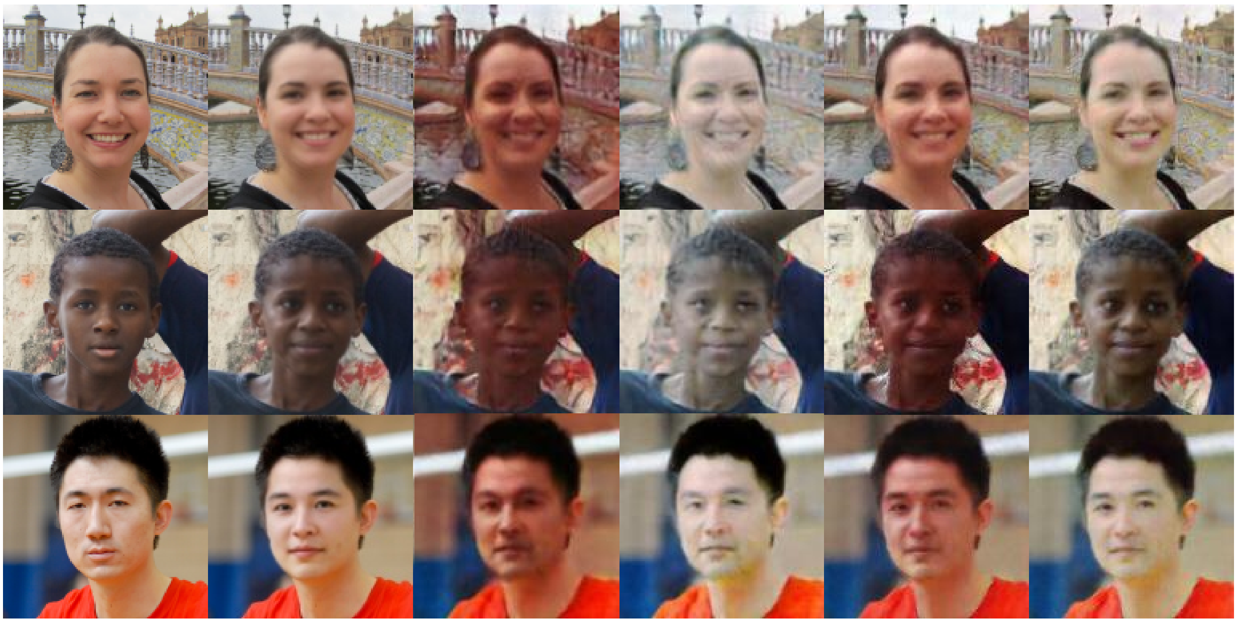

Figure 4.9: Comparing (from left) original, DeepPrivacy, AttGAN BN Dark, AttGAN BN Light, AttGAN IN Dark and AttGAN IN Light.

Discussion: How did instance normalization affect the images?

From the quantitative and visual clues, instance normalization seems to have kept more of the background from the original image. Since IN focuses more on each individual image, it is more likely to learn the skin tone change in each image, while in BN it is more likely to learn a summation of several images put together. As the presence of skin appears at different x and y positions in the images the summary is likely to have a representation of the skin tone in more places than where skin is present. The quality of images is better; thus, it is perhaps better at image reconstruction, but could also simply be related to how often the generator is updated, which is way more often that the one using BN.

4.5.6 Improving Image Quality: Reconstruction Loss

The following experiments use the same setup as for "Improving Data Quality: Changing Light Color", only changing one loss related detail at a time with the goal of improving the image quality. As this subsection focuses solely on the image quality, only the image quality results will be evaluated.

Increasing and decreasing λ_{rec}

The “Increasing and decreasing λ_{rec}” experiment tests different values for λ_{rec}. In the used AttGAN implementation, the λ_{rec} value is multiplied by the reconstruction loss value when forming the total generator loss. It is therefore assumed that a larger value equals an increased focus on the reconstruction. The default value for λ_{rec} is 100.

Table 4.25 compares the quality of the result images produced using different λ_{rec} values. In order to be able to compare all the results side by side, DeepPrivacy has been shortened to “DP”, and AttGAN has been shortened to “A”. It can be observed that the image quality remains best using the default λ_{rec}, as that is when PSNR and SSIM are best.

Table 4.25: This table shows the image quality of “Increasing and decreasing λ_{rec}”. Each measurement is found by measuring the distance between the original test images, D stands for dark and L for light.

Evaluation criteria	DP	DP +A L50 D	DP +A L50 L	DP +A L100 D	DP +A L100 L	DP +A L150 D	DP +A L150 L
L1	0.0293	0.0736	0.0833	0.0711	0.0727	0.0809	0.0797
L2	0.0606	0.1010	0.1139	0.0928	0.0956	0.1101	0.1125
PSNR	24.9555	20.1996	19.1312	**20.9463**	20.7326	19.4587	19.2823
SSIM	0.7423	0.5433	0.5576	0.6693	**0.6892**	0.5302	0.5504
LPIPS	0.0690	0.1630	0.1761	**0.1226**	0.1250	0.1609	0.1784
FID	30.7108	79.2361	78.7052	**44.7957**	45.1745	79.7804	78.1779

L1 Smooth

“L1 Smooth” changes the reconstruction loss from L1 to L1 Smooth.

Table 4.26 shows the image quality results of using L1 compared to using L1 Smooth. It can be seen that more pixels are changed using L1 Smooth. Changing from using L1 to L1 Smooth did not seem to improve the quality, as the PSNR and SSIM results have worsened.

Table 4.26: This table shows the image quality of “L1 Smooth”. Each measurement is found by measuring the distance between the original test images, D stands for dark and L for light.

Evaluation criteria	DeepPrivacy	DeepPrivacy +AttGAN L1 D	DeepPrivacy +AttGAN L1 L	DeepPrivacy +AttGAN L1 Smooth D	DeepPrivacy +AttGAN L1 Smooth L
L1	0.0293	0.0711	0.0727	0.0936	0.0913
L2	0.0606	0.0928	0.0956	0.1244	0.1243
PSNR	24.9555	**20.9463**	20.7326	18.3730	18.3124
SSIM	0.7423	0.6693	**0.6892**	0.4653	0.4997
LPIPS	0.0690	**0.1226**	0.1250	0.1987	0.2064
FID	30.7108	**44.7957**	45.1745	87.4730	90.3135

MS-SSIM + L1

"MS-SSIM + L1" changes the reconstruction loss from L1 to a combination of MS-SSIM and L1[5]. Recall from Chapter 2 that it is found that combining MS-SSIM and L1 increases the reconstruction quality.

Table 4.27 shows the image quality evaluation of using L1 compared to the combination of MS-SSIM and L1. One can observe that the pixel distance is lower using the new reconstruction loss. Also, PSNR has increased, although SSIM has decreased.

Table 4.27: This table shows the de-identificaton and image quality of "MS-SSIM + L1". Each measurement is found by measuring the distance between the original test images, D stands for dark and L for light.

Evaluation criteria	DeepPrivacy	DeepPrivacy +AttGAN L1 D	DeepPrivacy +AttGAN L1 L	DeepPrivacy +AttGAN MS-SSIM+L1 D	DeepPrivacy +AttGAN MS-SSIM+L1 L
L1	0.0293	0.0711	0.0727	0.0646	0.0622
L2	0.0606	0.0928	0.0956	0.0929	0.0949
PSNR	24.9555	20.9463	20.7326	**20.9501**	20.8414
SSIM	0.7423	0.6693	**0.6892**	0.6113	0.6359
LPIPS	0.0690	**0.1226**	0.1250	0.1288	0.1405
FID	30.7108	**44.7957**	45.1745	64.3933	63.3115

Discussion: How did the changes in reconstruction loss affect the quality of the images?

It was found that testing the values 50 and 150 for λ_{rec} did not increase the quality of the image. Although it may be worth fine-tuning the value using minimal changes to see whether a different value than the default value of 100 may provide better results, as it is visible that changing the λ_{rec} value does affect the image quality.
It was also found that L1 Smooth did not increase the image quality for the use case of this thesis.
Changing from L1 to a combination of MS-SSIM and L1 did, however, result in changes of interest. Recall that the PSNR increased, while SSIM decreased. As PSNR focuses on comparing the noise, while SSIM compares luminance, contrast and structure, we chose to focus on as high PSNR as possible to increase the image quality for the use case of this thesis. Therefore, MS-SSIM+L1 will be used as the reconstruction loss function for the final solution.

4.6 Final Solution

We have chosen to create two versions of the final solution, one with skin tone, "DeepPrivacy and AttGAN with Skin tone" (DP-ATT-S), and one without, "DeepPrivacy and AttGAN" (DP-ATT). The reason is that we observed that the age group and gender preservation results were better in the experiments without skin tone. Making both alternatives of the solution available, the Oslo Police and others can make a choice on whether they prefer as good age group and gender preservation results as possible, or to have skin tone included.

DP-ATT-S is trained using "Diverse Faces with Distinct Skin tones" with the gray light

[5] The implementation of MS-SSIM+L1 used in this thesis can be found here.

skin tone and the attributes female, dark_skin, 0-19, 20-29, 30-39 and 40+, and the λ_{rec} value set to 100. Instance normalization is used instead of batch normalization for the generator encoder and decoder in order to reduce noise in the background and preserve the colors of the background when changing skin tone. The number of discriminator updates per generator update is set to 1, as we found that this seemed to improve learning when using instance normalization. The combination of MS-SSIM and L1 is used for reconstruction loss instead of L1 alone, in order to reduce noise in the generated image, increasing the image quality. The alpha value 0.84 is used, as suggested by the paper proposing this loss function. Compensation is set to 0.2, as the implementation default of 200 made the reconstruction loss out of proportion compared to the other losses in AttGAN. The compensation value is only used for multiplying with the reconstruction loss in order to adjust the loss value in proportion to the other losses used.

DP-ATT is trained using the extended version of "Diverse Faces" with the attributes female, 0-19, 20-29, 30-39 and 40+, and the λ_{rec} value set to 100. The number of discriminator updates per generator update is set to 5, which is the default for the AttGAN implementation used in this thesis. Similarly as for DP-ATT-S, the combination of MS-SSIM and L1 is used instead of L1 in order to increase image quality. The same compensation and alpha values are used here as in DP-ATT-S.

For the final round we also changed the base for finding the best epoch, testing some other variants based on different loss parameters in the generator. Since the adversarial loss can be negative and the reconstruction loss and classification loss always are positive, it can be difficult to say if the total loss is high because the adversarial loss is very low or because the other loss functions are large. As the adversarial loss eventually should converge it is not bad if it falls below zero, however, large deviations of either positive or negative values do not imply good adversarial training. It is therefore beneficial if all of the loss distances are as close to zero as possible, so instead of using the total loss of the generator, we decided to use the epoch that had the lowest sum for $abs(g_{adv}loss)$ + $\lambda_{rec}\ g_r loss$ + $\lambda_{cls_g}\ g_c loss$.

NVIDIA Tesla V100 SXM3 32 GB was used for the training. The version with skin tone used approximately 21 minutes per epoch, while the version without skin tone used approximately 22 minutes.

Table 4.28 shows the evaluation of gender, age group and image quality for the final solutions. As expected, DP-ATT is able to achieve considerably higher accuracy in both gender and age group. It can also be seen that DP-ATT is both better at preserving luminance, contrast and structure, and minimizing noise in the generated image.

Table 4.29 shows the results from the skin detection for DP-ATT-S. It can be seen that the number of total images estimated after the use of AttGAN are similar to the number after only using DeepPrivacy. The MSE results are not as low as achieved in some previous experiments. However, one of the focuses for DP-ATT-S has been preserving the background, which seems to have resulted in less change in skin color. It can also be seen that regardless of the target color, using AttGAN to change the skin color results in lower MSE than DeepPrivacy alone.

Figure 4.10 shows a comparison of original images, de-identified by DeepPrivacy, changed to dark skin tone, and target age group and gender by DP-ATT-S, changed to light skin tone, and target age group and gender by DP-ATT-S, and changed to target age group and gender by DP-ATT.

Table 4.28: Evaluation results for the final solutions.

Evaluation criteria	DP-ATT-S Epoch(147) Dark IN MS-SSIM+L1	DP-ATT-S Epoch(147) Light IN MS-SSIM+L1	DP-ATT Epoch(40) MS-SSIM+L1
Total estimated (%)	91.33	91.67	92.33
Gender			
Correct gender (%)	92.24	91.18	96.39
Estimated & Correct	84.25	83.58	89.00
Age group			
Correct age group (%)	68.34	69.27	79.78
Estimated & Correct	62.42	63.50	73.67
MSE	13.82	14.65	6.70
Image quality			
L1	0.0614	0.0805	0.0413
L2	0.0853	0.1098	0.0683
PSNR	21.7285	19.5064	23.7440
SSIM	0.6038	0.5866	0.6353
LPIPS	0.1213	0.1433	0.0960
FID	64.3426	65.6125	60.4838

Table 4.29: Skin evaluation for DP-ATT-S.

Evaluation criteria	DeepPrivacy	DP-ATT-S +AttGAN Epoch(147) Dark IN MS-SSIM+L1	DP-ATT-S Epoch(147) Light IN MS-SSIM+L1
Total estimated (%)	76.17	70.33	77.33
MSE	19768	14765	15559
Target RGB	-	[111,55,55]	[150,150,150]
Average RGB	[222,166,145]	[210,131,119]	[253,229,208]

It can be seen that the final solutions are not perfect, as DP-ATT-S still does affect the background, and the image quality does still get reduced when the image is reconstructed by AttGAN. It is also seen that achieving a specific skin color is difficult as the original skin tone highly affects the outcome of the new skin color.

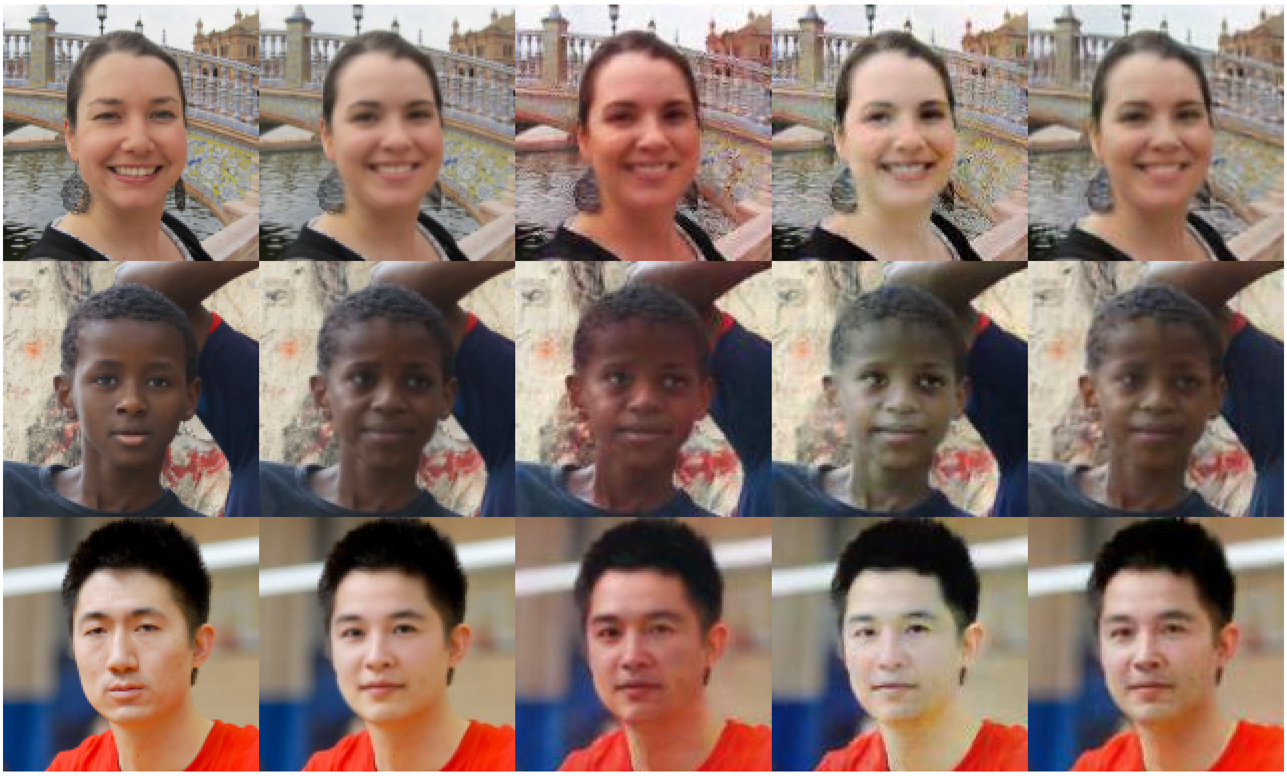

Figure 4.10: Comparing (from left) original, DeepPrivacy, DP-ATT-S dark, DP-ATT-S light and DP-ATT.

DP-ATT and DP-ATT-S were also tested on an image containing a large group of people, as CCTV footage can often contain multiple people in one frame. Figure 4.11 shows the original image, found using Google Images. Figure 4.12 shows the same image using DP-ATT, Figure 4.13 shows the same image using DP-ATT-S with dark skin tone and Figure 4.14 shows the same image using DP-ATT-S with light skin tone. It can be seen that using DP-ATT-S dark results in more smoother, natural appearing faces than using DP-ATT. DP-ATT-S with light skin is visually better at providing a similar skin color for all the faces in the image. However, as the previous results indicated, it does produce less natural images. Noise can be seen in the faces, as well as a dark color in the mouth region. This can possibly make the faces less detectable.

4.6.1 Discussion: Comparing the final solutions to state-of-the-art

Privacy-Protective-GAN for Face De-identification [68] achieves the age group classification accuracy of 86.9% for black skin tone and 87.3% for white skin tone. Note that this solution does not change the skin tone, they preserve it. Compared to our best age group classification result of 73.67%, their solution may be better at preserving age. They also convey some gender results, but it is difficult to extract the classification information from the de-identification information, especially as it is combined with ethnicity as well. It is debatable whether the results are comparable, as “Privacy-Protective-GAN for Face De-identification” has trained separate classifiers for each attribute that they use to find the classification accuracy for the generated images. Also, different datasets are used as

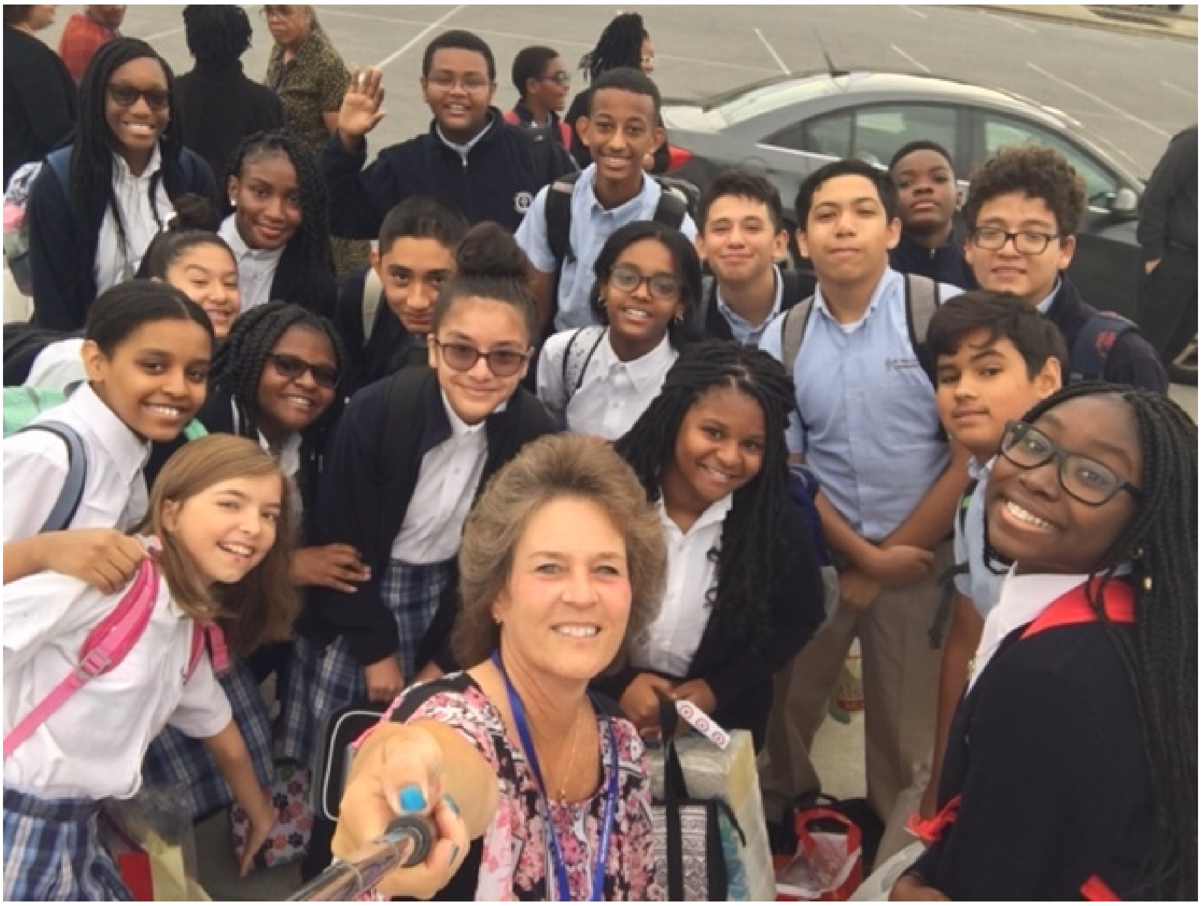

Figure 4.11: Original image of a large group of people.

they use the MORPH dataset. Since the solution that is proposed in “Privacy-Protective-GAN for Face De-identification” is not publicly available, and they do not provide more details on the network for finding classification accuracy, we are not able to either test their solution with our test set and age-gender-estimation for finding classification accuracy, nor are we able to use the same network for finding classification accuracy as they use.

’AnonymousNet: Natural Face De-Identification with Measurable Privacy’ [43] achieves PSNR value of 20.079 and SSIM value of 0.7894. Compared to our best values of 23.7440 for PSNR and 0.6353 for SSIM, it can be assumed that their solution is better at preserving luminance, contrast and structure, while our solution is better at minimizing noise in the generated image.

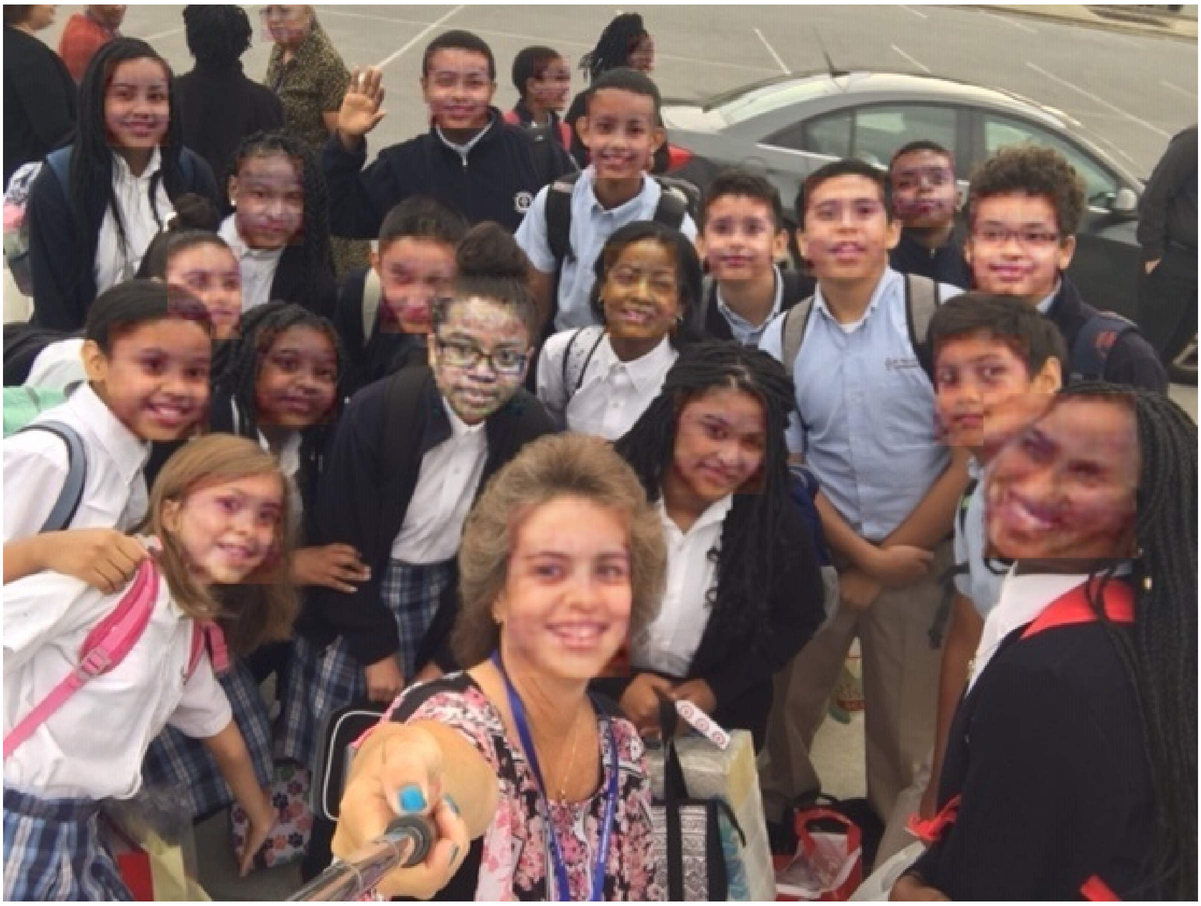

Figure 4.12: Image of a large group of people de-identified using DP-ATT.

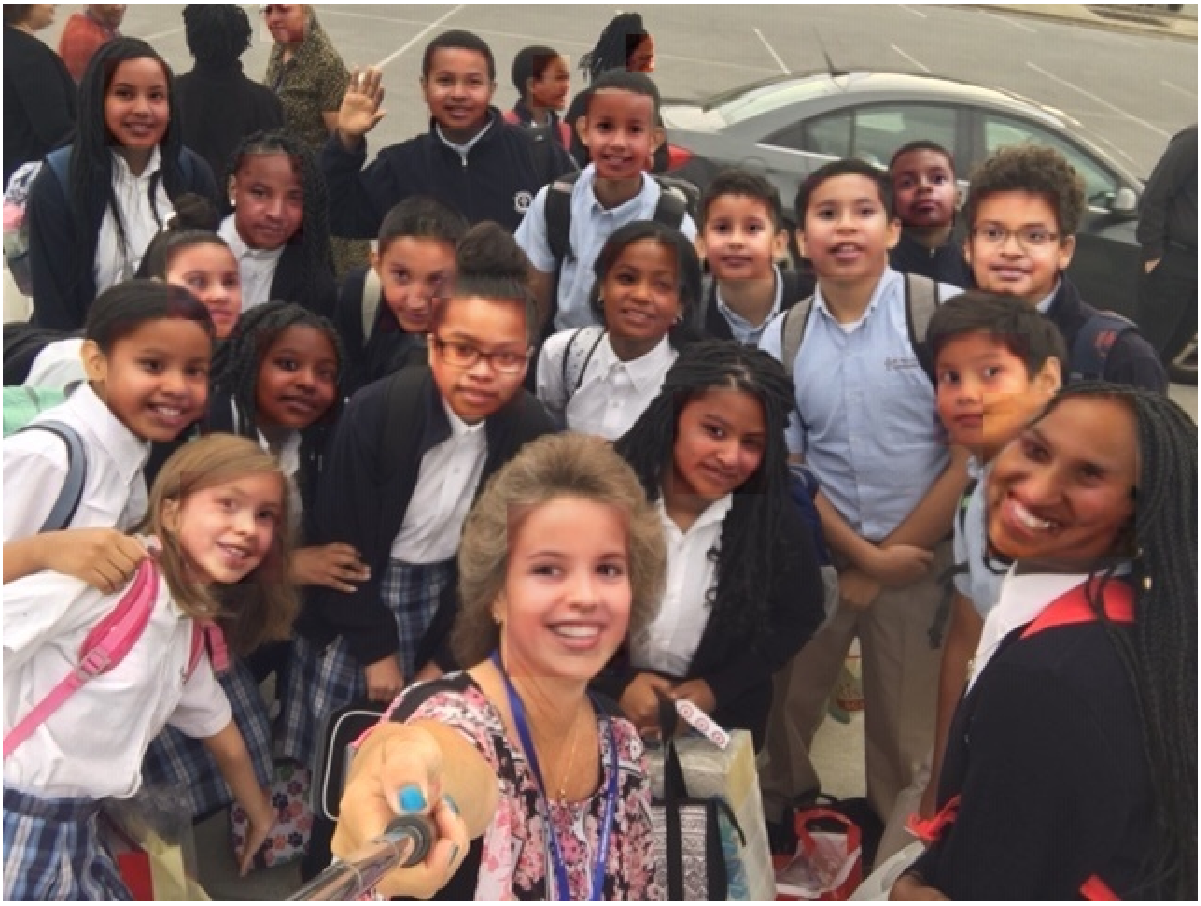

Figure 4.13: Image of a large group of people de-identified using DP-ATT-S dark.

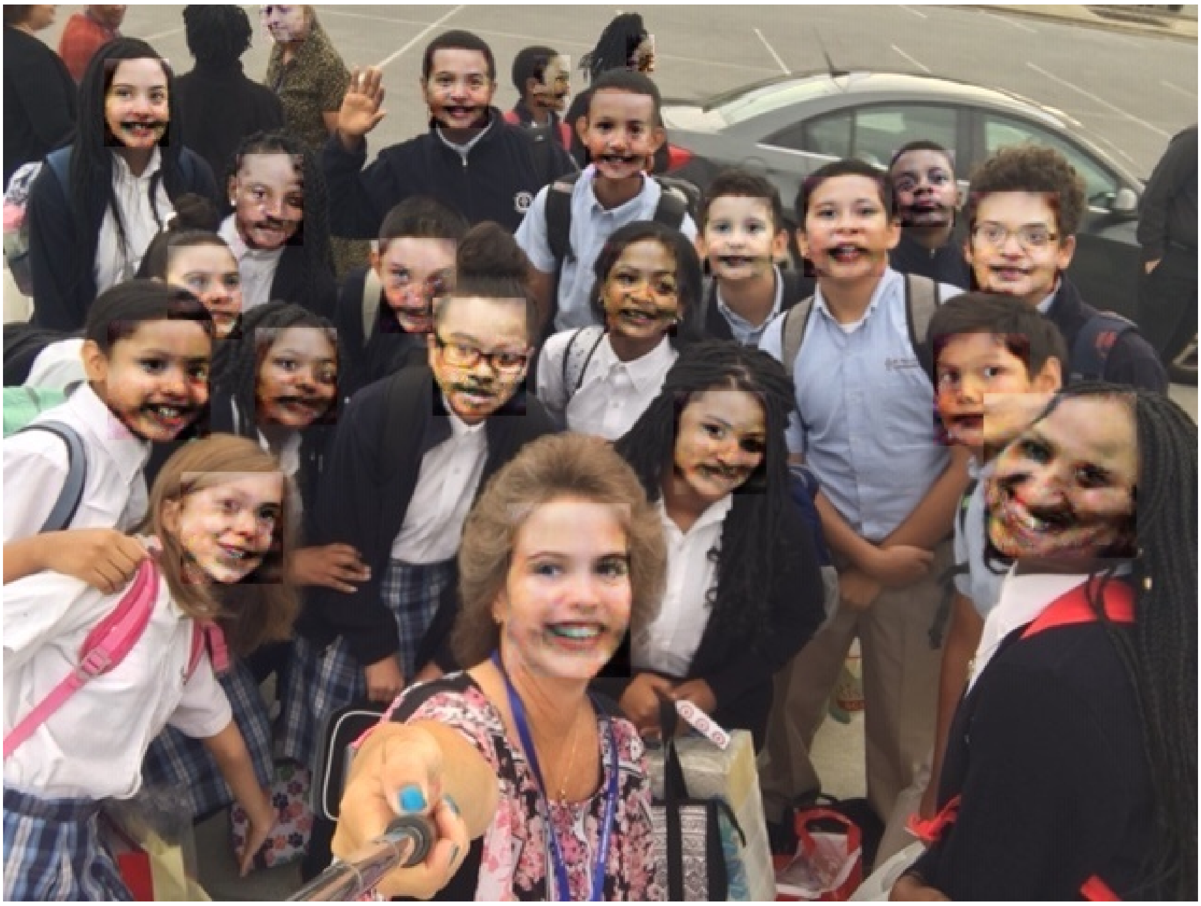

Figure 4.14: Image of a large group of people de-identified using DP-ATT-S light.

Chapter 5

Conclusions and Future Work

Section 5.1 contains the findings and conclusions of this thesis, while Section 5.2 contains some suggestions on how to take this research further.

5.1 Conclusions

In this thesis, DeepPrivacy is combined with AttGAN in order to preserve gender and age group in de-identified CCTV footage. Another GAN is also tested, namely StarGAN, which falls short to AttGAN given our chosen evaluation methods. It is found that AttGAN is able to achieve considerably better results in preservation of both attributes, as well as producing visually better images. In our experimental testing, we found that MS-SSIM and L1 offer less noise in GAN generated images than using L1 or smooth L1. It is also found that replacing batch normalization with instance normalization for the generator encoder and decoder resulted in less changes in the image background.

The first hypothesis of this thesis, “We can remove all recognizable features from a face and still generate a new face with same gender”, is shown to be correct. We are able to de-identify a face with the accuracy of 97.40%, and preserve the gender with our best accuracy of 89.00%. Compared to the gender preservation results of DeepPrivacy alone, being 77.50%, we manage to increase gender preservation with 11.50%.
The second hypothesis, “We can remove all recognizable features from a face and still generate a new face with approximately same age”, is also shown to be correct, as we are able to preserve the age group with our best accuracy of 73.67%. Comparing to DeepPrivacy’s age preservation accuracy of 42.25%, we manage to increase age preservation with 31.42%.
The correctness of the third hypothesis, “We can change all skin colors to one color in order to avoid bias towards certain skin tones”, is however debatable for the proposed algorithm. DP-ATT-S does provide images where the skin color is closer to the average color than it is using DeepPrivacy alone. However, the deviation of the average skin color compared to the target skin color is more than 20% for each of R, G and B. We are able to change the skin tone, but the underlying original skin tone does have a major effect on the final, achieved skin tone. The proposed scheme is therefore not optimal for the purpose of changing skin tone in order to avoid bias in the anomaly detector towards certain skin tones.

Overall, we show through comprehensive testing that our proposed method is able to remove all recognizable features from a face and still generate a new face with same gender and approximately same age.

5.2 Future Work

Mentioned below are some tasks that we would have looked further into if we had more time on this thesis.

- Implement a new generator loss where the face is masked and the combination of MS-SSIM and L1 is used on the input image versus the reconstructed image where skin tone is changed, in order to penalize change of the background. A different solution to the issue may be to, after de-identifying the face and using AttGAN to change to original age group and gender, crop out only the face and place it back in the original frame instead of the rectangular face image which contains parts of the background.
- Train DP-ATT with IN for the AttGAN generator encoder and decoder instead of BN, as the group images indicate that this could result in visually smoother faces.
- Fine-tune alpha for the MS-SSIM and L1 combination, as well as fine-tuning the number of discriminator updates per generator update.
- Work on increasing image quality further, with the goal for the quality being the same as after only using DeepPrivacy.
- It has been noticed that the dataset got a different age distribution after relabeling the FairFace images using Age-gender-estimation. The new age distribution is closer to the distribution of the dataset that Age-gender-estimation is trained on (IMDB-WIKI). Thus, it can be beneficial to use a different network to detect age and gender if one wish a broader age distribution. This was not an issue for the final proposed networks in this thesis due to an even distribution in the chosen age groups.
- As the Oslo Police want to de-identify in real-time, the proposed algorithm should be tested in regard to time and optimized to lower delay.
- Restoring bitrate of videoes after using DeepPrivacy for de-identification.

Bibliography

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2stylegan: How to embed images into the stylegan latent space?” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4432–4441.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2stylegan++: How to edit the embedded images?” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8296–8305.

[3] Rameen Abdal et al. “Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows.” In: *ACM Transactions on Graphics (TOG)* 40.3 (2021), pp. 1–21.

[4] Ejaz Ahmed, Michael Jones, and Tim K Marks. “An improved deep learning architecture for person re-identification.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3908–3916.

[5] High-Level Expert Group on AI. *Ethics guidelines for trustworthy AI*. 2018. URL: `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai` (visited on 06/01/2021).

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein generative adversarial networks.” In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[7] Alex Bewley et al. “Simple online and realtime tracking.” In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468.

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large scale GAN training for high fidelity natural image synthesis.” In: *arXiv preprint arXiv:1809.11096* (2018).

[9] *Camera patent*. URL: `https://patents.google.com/patent/US1559795A/en`.

[10] Zhenfei Chen et al. “GAN-Based Image Privacy Preservation: Balancing Privacy and Utility.” In: *International Conference on Machine Learning for Cyber Security*. Springer. 2020, pp. 287–296.

[11] Yunjey Choi et al. “Stargan v2: Diverse image synthesis for multiple domains.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8188–8197.

[12] Yunjey Choi et al. “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8789–8797.

[13] European Commission. *White paper: On Artificial Intelligence - A European approach to excellence and trust*. Tech. rep. 2020.

[14] William Lee Croft. “Design and Applications of Differentially Private Mechanisms: Adherence to Query Range Constraints and Obfuscation of Facial Images.” PhD thesis. Carleton University, 2020.

[15] *Deling av bilder*. URL: `https://www.datatilsynet.no/personvern-pa-ulike-omrader/internett-og-apper/bilder-pa-nett/`.

[16] Jiankang Deng et al. “Arcface: Additive angular margin loss for deep face recognition.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.

[17] Vincent Dumoulin and Francesco Visin. *A guide to convolution arithmetic for deep learning*. 2018. arXiv: 1603.07285 [stat.ML].

[18] Organisation for Economic Co-operation and Development. *OECD Principles on AI*. 2019. URL: https://www.oecd.org/going-digital/ai/principles/ (visited on 06/01/2021).

[19] Oran Gafni, Lior Wolf, and Yaniv Taigman. “Live face de-identification in video.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 9378–9387.

[20] Simson L Garfinkel et al. “De-identification of personal information.” In: *National institute of standards and technology* (2015).

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[22] Ian Goodfellow et al. “Generative adversarial networks.” In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[23] Ralph Gross et al. “Face de-identification.” In: *Protecting privacy in video surveillance*. Springer, 2009, pp. 129–146.

[24] Ralph Gross et al. “Integrating utility into face de-identification.” In: *International Workshop on Privacy Enhancing Technologies*. Springer. 2005, pp. 227–242.

[25] Ralph Gross et al. “Model-based face de-identification.” In: *2006 Conference on computer vision and pattern recognition workshop (CVPRW’06)*. IEEE. 2006, pp. 161–161.

[26] Ishaan Gulrajani et al. “Improved training of wasserstein gans.” In: *arXiv preprint arXiv:1704.00028* (2017).

[27] Jia Guo and Jiankang Deng. *deepinsight/insightface*. URL: https://github.com/deepinsight/insightface.

[28] Erik Härkönen et al. “Ganspace: Discovering interpretable gan controls.” In: *arXiv preprint arXiv:2004.02546* (2020).

[29] Kaiming He et al. “Mask r-cnn.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[30] Z. He et al. “AttGAN: Facial Attribute Editing by Only Changing What You Want.” In: *IEEE Transactions on Image Processing* 28.11 (2019), pp. 5464–5478. DOI: 10.1109/TIP.2019.2916751.

[31] John R Hershey and Peder A Olsen. “Approximating the Kullback Leibler divergence between Gaussian mixture models.” In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*. Vol. 4. IEEE. 2007, pp. IV–317.

[32] Martin Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium.” In: (2017).

[33] Lei Huang et al. “Normalization Techniques in Training DNNs: Methodology, Analysis and Application.” In: *arXiv preprint arXiv:2009.12836* (2020).

[34] HÃ¥kon HukkelÃ¥s, Frank Lindseth, and Rudolf Mester. *Image Inpainting with Learnable Feature Imputation*. 2020. arXiv: 2011.01077 [cs.CV].

[35] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. “Deepprivacy: A generative adversarial network for face anonymization.” In: *International Symposium on Visual Computing*. Springer. 2019, pp. 565–578.

[36] Amin Jourabloo, Xi Yin, and Xiaoming Liu. “Attribute preserved face de-identification.” In: *2015 International conference on biometrics (ICB)*. IEEE. 2015, pp. 278–285.

[37] Kimmo Karkkainen and Jungseock Joo. “FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 1548–1558.

[38] Animesh Karnewar and Oliver Wang. “Msg-gan: Multi-scale gradients for generative adversarial networks.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7799–7808.

[39] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[40] Tero Karras et al. “Analyzing and improving the image quality of stylegan.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8110–8119.

[41] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. 2018. arXiv: 1710.10196 [cs.NE].

[42] Avery Koop. *Mapped: The Top Surveillance Cities Worldwide*. 2021. URL: https://www.visualcapitalist.com/mapped-the-top-surveillance-cities-worldwide/.

[43] Tao Li and Lei Lin. “Anonymousnet: Natural face de-identification with measurable privacy.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[44] Ziwei Liu et al. “Deep Learning Face Attributes in the Wild.” In: *Proceedings of International Conference on Computer Vision (ICCV)*. 2015.

[45] *Lov om behandling av personopplysninger (personopplysningsloven)*. URL: https://lovdata.no/dokument/NL/lov/2018-06-15-38.

[46] *Lov om politiet (politiloven)*. URL: https://lovdata.no/dokument/NL/lov/1995-08-04-53.

[47] *Lov om styrking av menneskerettighetenes stilling i norsk rett (menneskerettsloven)*. URL: https://lovdata.no/dokument/NL/lov/1999-05-21-30/emkn/ARTIKKEL_8#emkn/ARTIKKEL_8.

[48] Odhran McCarthy et al. *Towards Responsible AI Innovation Second INTERPOL-UNICRI Report in Artificial Intelligence for Law Enforcement*. Tech. rep. INTERPOL and UNICRI, 2020.

[49] Elaine M Newton, Latanya Sweeney, and Bradley Malin. “Preserving privacy by de-identifying face images.” In: *IEEE transactions on Knowledge and Data Engineering* 17.2 (2005), pp. 232–243.

[50] Jim Nilsson and Tomas Akenine-Möller. “Understanding ssim.” In: *arXiv preprint arXiv:2006.13846* (2020).

[51] Seong Joon Oh et al. *Faceless Person Recognition; Privacy Implications in Social Media*. 2016. arXiv: 1607.08438 [cs.CV].

[52] Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. “Adversarial Latent Autoencoders.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[53] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” In: *arXiv preprint arXiv:1511.06434* (2015).

[54] Joseph Redmon and Ali Farhadi. “YOLO9000: better, faster, stronger.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271.

[55] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement.” In: *arXiv preprint arXiv:1804.02767* (2018).

[56] Joseph Redmon et al. “You only look once: Unified, real-time object detection.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[57] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. “Learning to anonymize faces for privacy preserving action detection.” In: *Proceedings of the european conference on computer vision (ECCV)*. 2018, pp. 620–636.

[58] Slobodan Ribaric and Nikola Pavesic. “An overview of face de-identification in still images and videos.” In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 4. IEEE. 2015, pp. 1–6.

[59] *Smart cities*. 2020. URL: https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en.

[60] Igor Stojanovic. “De-identification for privacy protection in multimedia content.” In: (2013).

[61] Waqas Sultani, Chen Chen, and Mubarak Shah. “Real-world anomaly detection in surveillance videos.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6479–6488.

[62] Arief Rachman Sutanto and Dae-Ki Kang. “A Novel Diminish Smooth L1 Loss Model with Generative Adversarial Network.” In: *Intelligent Human Computer Interaction*. Ed. by Madhusudan Singh et al. Cham: Springer International Publishing, 2021, pp. 361–368.

[63] Bart Thomee et al. “YFCC100M: The new data in multimedia research.” In: *Communications of the ACM* 59.2 (2016), pp. 64–73.

[64] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. *Instance Normalization: The Missing Ingredient for Fast Stylization*. 2017. arXiv: 1607.08022 [cs.CV].

[65] Yang Wang. “A Mathematical Introduction to Generative Adversarial Nets (GAN).” In: *arXiv preprint arXiv:2009.00169* (2020).

[66] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. “Multiscale structural similarity for image quality assessment.” In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[67] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. “Simple online and realtime tracking with a deep association metric.” In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

[68] Yifan Wu, Fan Yang, and Haibin Ling. *Privacy-Protective-GAN for Face De-identification*. 2018. arXiv: 1806.08906 [cs.CV].

[69] K. Zhang et al. “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks.” In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. ISSN: 1070-9908. DOI: 10.1109/LSP.2016.2603342.

[70] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[71] Hang Zhao et al. “Loss functions for image restoration with neural networks.” In: *IEEE Transactions on computational imaging* 3.1 (2016), pp. 47–57.

[72] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks.” In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.