Balancing Privacy and Utility: An Evaluation of Generative Models for Car Anonymization in Street Scene Images

Miriam Louise Carnot[1], Robert Hassler[2], Eric Peukert[1], André Ludwig[2,3], Bogdan Franczyk[2,4]

[1] ScaDS.AI (University of Leipzig)
[2] University of Leipzig
[3] Kühne Logistics University
[4] Wrocław University of Economics

KEY WORDS: urban imagery, privacy, generative models

ABSTRACT:

When recording street scenes, the privacy of the people concerned must always be taken into account. Previous work has focused heavily on directly identifiable features such as faces or license plates. However, the visibility of indirectly identifiable objects can also limit the protection of citizens. This is especially true for cars. Regular recordings could be used to infer when a person is at home or where else they have been. The obvious solution would be to simply make cars unrecognizable. However, previous work has shown that this has a negative impact on downstream tasks such as traffic sign recognition. Therefore, in this work, we use generative models to synthetically modify cars. We compare models from the generative adversarial network (DP-GAN, OASIS) and diffusion model communities (Kolors, SDXL) to see which ones are best suited in terms of anonymity, image integrity, and performance. For the GAN-based models, we use an image-to-image translation approach to modify only image sections with cars, while for the diffusion models, we develop two methods that use text-guided image inpainting. We compare the developed methods using established metrics and perform a survey with test subjects. In terms of anonymity, all models achieve convincing results, while diffusion models that generate each car mask individually produce particularly realistic images. In return, GAN-based methods process images more than twice as fast creating a trade-off between image integrity and performance.

1. INTRODUCTION

In every city today, imagery is being collected by surveillance cameras, map providers, or intelligent vehicles. Whenever such data is recorded, the privacy rights of individuals in public spaces must be protected. Many countries have enacted regulations to protect privacy, such as the General Data Protection Regulation in Germany (Datenschutz-Grundverordnung (DSGVO), 2016). Faces and license plates must be blurred so that data users do not know in which location people have been and at what time. However, extracting personal information can also be indirectly possible, even with blurred faces and license plates (Adeboye et al., 2022). Vehicles may have, for example, distinctive stickers or interior features. Information about a car can also be obtained from its position. If a certain car is always parked in front of the same house, it can be concluded that it belongs to the resident. In particular, for data sets with regular recordings, there is a risk of inferring when a person is at home, when visitors are there, or if this specific car can be seen in other places. This type of information is very sensitive.
Traditionally, image regions are anonymized using methods such as pixelation, blurring, or masking (Liu et al., 2024). However, such deterministic methods can produce artifacts that may have a negative impact on downstream models, such as the localization of objects (Uittenbogaard et al., 2019). Also, blurring an entire image still gives information whether a person is at home or not, in the case of single houses or cars that can be identified even when blurred. We therefore believe that increasing privacy by replacing vehicles with different ones or letting them disappear by generating background should be investigated. In this work, we compare different generative models for this task and aim to maintain image integrity while meeting data protection requirements (Lee and You, 2024)(Liu et al., 2024).

Today, 79 % of synthetic images were created using Generative Adversarial Networks (GAN) (Goodfellow et al., 2020), as found in the review study by Zulfiqar et al. (Zulfiqar et al., 2024). Since 2021, diffusion models have become increasingly important (Ho et al., 2020), as they outperform GANs in many scenarios in terms of image quality and diversity, but require more computational resources (Croitoru et al., 2023)(Dhariwal and Nichol, 2021). We therefore think it is worthwhile to compare the two architecture types for the case of modifying cars. The objective of this work is not only to anonymize the cars in the images as far as possible (color, model, etc.) but also to preserve the image integrity so that it is not immediately obvious that the image has been modified. In the best case, the process should involve minimal computational effort. For this purpose, we first focus on identifying which generative models are suitable for the task and how they can be applied. We then want to take a closer look at the relationship between anonymization effectiveness, visual integrity, and performance, and evaluate these for the selected models.
With this work, we contribute the following:

- a pipeline for the synthetic modification of cars in 2D images using two GAN and two stable diffusion models,
- two proposed integration methods for diffusion models,
- a systematic comparison of the models regarding anonymity, image integrity, and performance based on a quantitative and a qualitative evaluation.

The study is structured as follows: In Section 2, we analyze the basics of generative image synthesis and the current state of research in anonymization with generative models. Based on that

foundation, Section 3 describes the design of our anonymization methodology, and in Section 4, we evaluate the methods developed. We extend the trade-off between anonymization effectiveness and image integrity described by Lee and You (Lee and You, 2024) by the dimension of performance. Finally, we discuss the results and give an outlook for future work in Section 5 and summarize our findings in Section 6.
The code is available under: `github.com/miriamcarnot/generative-car-anonymization`

2. RELATED WORK

In this Section, we first discuss two central methods of image generation: image inpainting and image-to-image translation. We then look into existing approaches to image anonymization and explain the functionalities of GANs and diffusion models.

Image Inpainting. Through image inpainting, missing or damaged image areas are filled in. A segmentation mask defines the area that marks the image section to be replaced (Szeliski, 2022). Then a neural network reconstructs the masked area. GANs, diffusion models, or transformers are frequently used for this task (Malm et al., 2023). While simple approaches rely on context-based "hole filling" (Zhao et al., 2021), advanced methods enable targeted control of image synthesis through conditioning information such as prompts (Reed et al., 2016), also called "Text-Guided Image Inpainting" (Wang et al., 2023). In this work, we use image painting to insert vehicles into previously defined segmentation masks with diffusion models.

Image-to-Image-Translation (I2IT). I2IT addresses the transfer of images from one domain to another without losing the underlying content of the image. Typical applications include the transformation of label maps into photorealistic images or style transfer between different works of art (Szeliski, 2022). In label maps, each pixel of an image is assigned a specific semantic label. They are used to segment an image and thus allow for the identification of specific areas such as vehicles, roads, or buildings for further processing. GAN-based models are frequently used in I2IT (e.g., Pix2Pix (Isola et al., 2017)). I2IT forms the basis for our GAN-based methods.

Image anonymization methods. Anonymization solutions for personal image content differ in terms of the generative models used, the image areas to be anonymized and the anonymization methods applied. Sensitive image areas can be fundamentally divided into two categories: directly identifiable objects (e.g., people, vehicles), and indirectly identifiable objects (e.g., prominent building facades or street names) (Liu et al., 2024). Different areas of the image can be exploited in different ways. Georeferencing attacks aim to retrieve the exact location of the person being photographed using indirectly identifiable objects (Adeboye et al., 2022). Surveillance systems can be maliciously used to track people (Malm et al., 2023), which can be prevented by replacing vulnerable image areas.
For vehicle anonymization, different methods have been proposed. Uittenbogaard et al. use a Wasserstein GAN to anonymize vehicles (Uittenbogaard et al., 2019), while DeepClean uses a private Gaussian Mixture Model to pixelate image areas (Adeboye et al., 2022). Other approaches for anonymizing directly identifiable objects include the processing of people (Hukkelås and Lindseth, 2023)(Brkic et al., 2017)(Hukkelås et al., 2023)(Wu et al., 2018)(Klemp et al., 2023)(Malm et al., 2023)(Liu et al., 2024). While DeepPrivacy synthetically replaces faces using a GAN and already achieves high image quality (Hukkelås and Lindseth, 2023), LDFA shows that even higher visual integrity can be achieved by using a Latent Diffusion Model (LDM) (Klemp et al., 2023). Others anonymize indirectly identifiable objects such as vehicles or buildings by pixelation, removal, or replacement (Adeboye et al., 2022). While the SVIA pipeline is based on a diffusion model (Liu et al., 2024), ADGAN uses a GAN-based approach (Xiong et al., 2020). As Lee and You have shown, anonymization methods can negatively influence the performance of downstream models (Lee and You, 2024). In their study, blurring certain image areas led to a strong performance loss in downstream instance segmentation and object detection tasks. The preservation of image integrity is thus a central criterion when selecting methods (Malm et al., 2023)(Klemp et al., 2023). Also, deterministic anonymization methods such as blurring do not offer absolute certainty. For example, a blurred area can in some cases be partially reconstructed using digital deblurring techniques, which constitutes a privacy leak (Piano et al., 2024). Thus, we decided to compare generative models for this anonymization task.

Diffusion Models. Diffusion models are based on a two-stage process consisting of a forward and a reverse diffusion process (Ho et al., 2020). In forward diffusion, noise is gradually added to an input image until it ultimately corresponds to a pure noise distribution, whereby the model learns relevant image features such as edges, textures, and structural details. In reverse diffusion, the noise is gradually removed to reconstruct a synthetic image from the noise distribution. Models like stable diffusion (Rombach et al., 2022) make it possible to specifically condition the reverse diffusion process using prompts. Stable diffusion is a latent diffusion model (LDM) in which the image generation does not take place in the high-dimensional pixel space, but in a compressed, latent representation space which significantly reduces the computational complexity. Diffusion models can generate previously unseen content by using prompts that describe the scenario (Croitoru et al., 2023). Thus, they do not have to be trained specifically for generating cars to do so. Also, a large number of inpainting solutions are freely available. Though diffusion models may take long processing times, their ability to flexibly generate high-quality and detailed images is promising for this work.

Generative Adversarial Networks (GANs). GANs are based on contrastive training, in which a generator tries to produce realistic images, while a discriminator learns to distinguish between real and generated images (Goodfellow et al., 2020). Both models improve each other: the generator learns to generate more realistic data, while the discriminator learns to recognize forgeries better (Szeliski, 2022). Upon training, their inference time is very short (Wang, 2024). However, GANs may suffer from "mode collapse" where the generator no longer covers the entire diversity of the training data, but is restricted to a limited number of patterns that reliably deceive the discriminator. This leads to reduced image diversity (Wang, 2024). Also, generated images are often of lower quality and detail compared to those created with diffusion models (Dhariwal and Nichol, 2021). There are GAN-based approaches for image inpainting (Zhao et al., 2021)(Zheng et al., 2022), as well as conditional variants in which prompts are used to control image generation (Reed et al., 2016). The application of GAN-based inpainting is mostly limited to so-called "hole filling", in which missing image areas are primarily reconstructed using the surrounding context (Szeliski, 2022). To the best of our knowledge, there are currently no text-driven inpainting GANs suitable for replacing vehicles in segmentation masks.

3. METHODOLOGY

We designed different implementation methods for GANs and diffusion models. For GAN-based approaches, we use image-to-image translation (I2IT), since we have not found a GAN model that can specifically draw vehicles into predefined masks. Diffusion models offer more flexibility and allow us to generate vehicles directly in the corresponding image regions with a text-driven inpainting approach. In this Section, we give details on the chosen data set and the different implementations.

The Data Set. For our comparison, we chose the Cityscapes dataset (Cordts et al., 2016), which contains high-resolution images (2048 × 1024 pixels) of urban street scenes taken in several German cities. They are annotated with pixel accuracy and provide segmentation masks of typical objects in the urban environment, such as cars, pedestrians, or buildings. It is widely used and suitable for semantic segmentation and instance segmentation tasks. The study can be extended to any other street view imagery dataset.

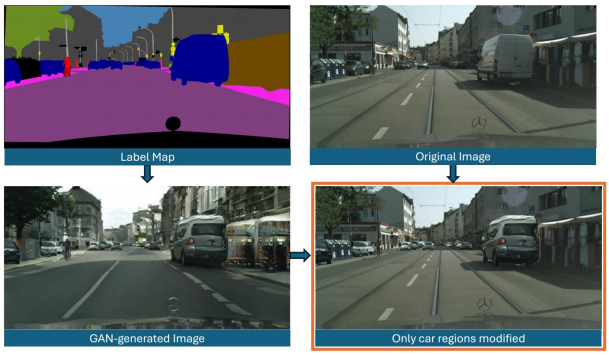
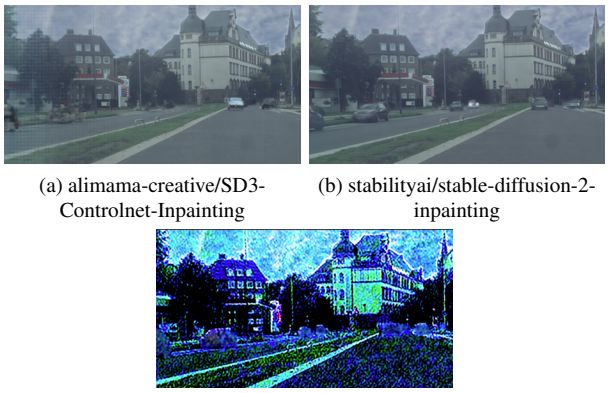


Figure 1. GAN-based Image-to-Image-Translation (I2IT).

GAN-based I2IT. Figure 1 shows how we can anonymize cars in images using GANs. First, we convert the label maps from the dataset (a) into photorealistic images (b). The vehicle regions are then extracted from the generated images and merged with the remaining image regions of the original images (c) to create an image in which only cars are modified (d).
We identified two high-performance commonly used I2IT GANs with publicly available source code that are capable of translating label maps from the Cityscapes dataset into photorealistic images: OASIS (Sushko et al., 2020) and DP-GAN (Li et al., 2022). OASIS generates images from label maps without having to rely on complex loss functions as in previous models. Instead, it uses a discriminator that assigns each pixel to a class, ensuring a precise match between the label map and the generated image. By inserting random noise for each layer, OASIS can also quickly generate different variants of the same label map. However, OASIS and previous models struggle with creating both small and large objects realistically. The authors of DP-GAN addressed this issue by designing an architecture in which image information is processed simultaneously on several size scales. Thus, large objects do not break up into fragmented areas and details are displayed sharply. Compared to previous models, DP-GAN generates more natural and detailed images from the same label maps.
Both models have pre-trained instances for the I2IT of the Cityscapes dataset. If another urban image dataset is to be anonymized with a GAN, the label maps resulting from a segmentation model need to be adapted to the Cityscapes format.



(a) alimama-creative/SD3-Controlnet-Inpainting (b) stabilityai/stable-diffusion-2-inpainting

(c) playgroundai/playground-v2.5-1024px-aesthetic

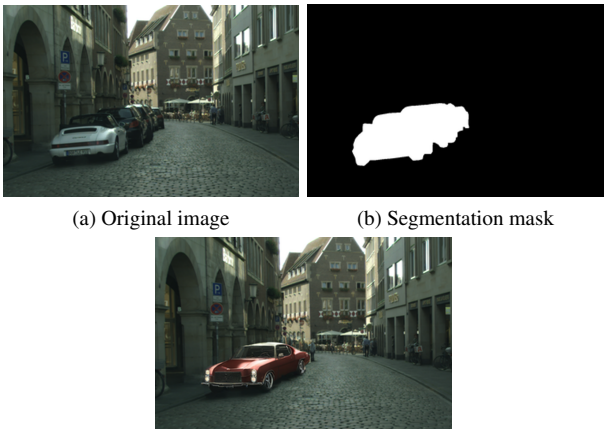
Figure 2. Examples of unsuccessful vehicle generation with diffusion models.

Diffusion-based Text-Guided-Image-Inpainting. Not all diffusion models are equally suitable for modifying cars. Some models delivered inadequate results in previous test runs, others are not accessible free of charge. Some diffusion models offer inpainting instances (e.g., SDXL (diffusers, 2023), Kandinsky (Shakhmatov, 2023), Kolors (Kwai-Kolors, 2024)), while others do not (e.g., Playground v2.5 (Li et al., 2024), Stable Diffusion 3 (AI, 2024)). The models with inpainting instances can be guided through prompts and are thus especially promising. Figure 2 shows images generated by models that are not considered further due to their evident deficits in our initial test runs. The model named "SD3-Controlnet-Inpainting" from alimama-creative was excluded due to clear artifacts that are visible at the edges of the images. These probably arise because the model was originally trained on square images (1024 × 1024 pixels) and the resolution is forced to increase (alimama-creative, 2024). The "stable-diffusion-2-inpainting" model from StabilityAI generated not convincing image regions with mostly blurred cars (Ho et al., 2020). The "playground-v2.5-1024px-aesthetic" model proposed by Li et al. (Li et al., 2024) generated images with strong blue noise. The vehicles are not visible or remain unchanged, indicating that the model did not perform the inpainting process as expected. The inpainting instances of the Kolors and SDXL models showed realistic and high-quality generated vehicle areas in the test runs (see Figure 3). SDXL is a latent diffusion model that was specially developed for the generation of high-resolution photorealistic images. The underlying U-Net architecture comprises 2.6 billion parameters which is a significant increase over the previous versions of Stable Diffusion 1.5 (860 million parameters) and Stable Diffusion 2.5 (865 million parameters) and contributes to greater detail and precision (Podell et al., 2023). Kolors is a latent diffusion model based on the architecture of SDXL. A key innovation is the extended training plan for high-resolution im-

(a) Kolors (b) SDXL

Figure 3. Examples of successful vehicle generation with diffusion models.

ages, which increases the number of diffusion steps from 1000 to 1100. This allows Kolors to reproduce finer detail and reduce artifacts (Kolors Team, 2024).
During the generation of cars in the segmentation masks, directly adjacent vehicle areas can be merged (e.g., in a traffic jam or a row of parking spaces) resulting in a coherent large-scale mask. The diffusion models then often generate a vehicle in a single inference step that extends over several originally separate mask areas. This can lead to unrealistic representations, for example oversized vehicles as in Figure 4.



(a) Original image (b) Segmentation mask

(c) Result

Figure 4. Example for fusion effects during vehicle generation.

To prevent this merging, we have implemented a second method, which generates each car separately in addition to the initial method that considers all regions as one mask. We call the initial method "whole" while we refer to the new approach as "steps". Figure 6 shows the schematics of both approaches. With the "whole" method, all vehicle areas are first merged into a one mask and then generated in a single inference step. This keeps processing times low but increases the risk of fusion effects. With the "steps" method, vehicle areas are generated one by one. For this purpose, the original image and a mask that only covers the area of a single car are passed as input in the respective inference step. To avoid unwanted effects between the generation steps, the unmodified image is used for each inference step. In step *n*, the original image together with the corresponding mask is used as input instead of the output image from step *n-1*. Finally, the generated vehicles are combined to form an overall image. This method can generate cars more precisely and avoid fusion artifacts. However, this iterative process increases the processing time of an image.

(a) Original image (b) Modified image

Figure 5. Unwanted changes in the background.

Regardless of the selected implementation method ("steps" or "whole"), the generated cars are then extracted from the manipulated image and combined with the original image. This is necessary as diffusion models may also change unmasked image areas which impairs the image's value for downstream

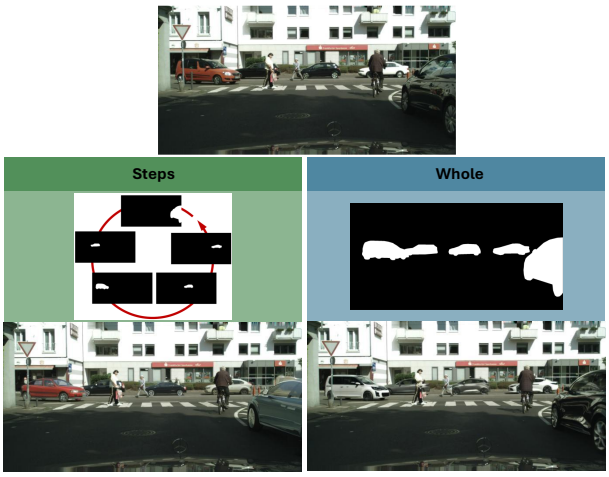


Figure 6. Outline of the implementation methods ("whole" and "steps") for the diffusion models.

tasks, for example, traffic signs (see Figure 5). In both implementation methods, the images are generated with the diffusion models SDXL and Kolors with identical hyperparameters (see Listing 1). The *prompt* and *negative_prompt* guide the diffusion model by specifying desired features, such as a photorealistic car in an urban scene, and undesired artifacts like low resolution and blurry details. The *num_inference_steps* determines the number of denoising steps, balancing computational load and image quality. The *guidance_scale* controls the influence of the prompt, ensuring the output closely matches the desired specifications while allowing for creative freedom. The *strength* parameter dictates the extent of transformation, enabling significant alterations to the input image.

Listing 1. Hyperparameters for the diffusion models

```
prompt: "A photorealistic car seamlessly
    integrated into an urban street scene,
     with consistent texture and lighting.
     Replace the original car with a
    modern, neutral design, ensuring no
    logos, license plates, or text. The
    new car should have a generic color,
    blending naturally with realistic
    reflections and shadows.",
negative_prompt: "worst quality, low
    resolution, overexposed, blurry,
    distorted shapes, text artifacts,
    unrealistic shadows.",
num_inference_steps: 25,
guidance_scale: 6.0,
strength: 0.999,
```

4. RESULTS

Following our methodology, we obtain six different pipelines (two with GAN and four with diffusion architectures): I2IT with DP-GAN, I2IT with OASIS, Kolors with the *steps* method, SDXL with the *steps* method, Kolors with the *whole* method, SDXL with the *whole* method. All diffusion models use Text-Guided-Image-Inpainting with the Hyperparameters listed in Listing 1. To evaluate the anonymization methods, each method replaces the cars of the 500 images of the Cityscapes validation dataset. We first compare the proposed pipelines in a quantitat-

Figure 7. Qualitative evaluation, part 1: method selection (1 out of 6).

Figure 8. Qualitative evaluation, part 2: individual assessment using the Likert scale (1:very good, 5:very bad).

ive evaluation using calculated metrics. In the following qualitative evaluation, we give the generated images to test subjects and let them evaluate their anonymity and integrity. Both analysis approaches quantify the degree of anonymization and the image quality. In addition, the image processing times and the utilization of the GPU memory (VRAM) are recorded.

Quantitative Evaluation. To evaluate the anonymization performance of an anonymization method, two dimensions must be taken into account: the quality of the resulting images and the effectiveness of the anonymization (Liu et al., 2024). We decided on two established metrics: the Ferchet Inception Distance (FID) (Heusel et al., 2017) and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). The FID quantifies the difference between two image data sets. Features are extracted from the original image and the generated image. Then, the deviations in the mean values and covariances of these feature distributions are calculated. A low FID value means that the distribution of the generated images is more similar to that of the real images indicating a higher image quality and better suitability for downstream tasks. FID has been shown to correlate with human preference, making it a suitable metric for evaluating image quality in this study. LPIPS compares images by measuring the differences in the feature maps of selected layers of a pre-trained neural network. A higher LPIPS value indicates that the features of the two images are less similar. As a result, it signals changes in the affected image regions and thus enables a quantitative evaluation of the anonymization effectiveness. Since vehicle areas in the images often only make up a small part of the total area and the remaining image content remains unchanged by the anonymization, a comparison of entire images would lead to distorted metric results. Therefore, the vehicle areas are extracted based on their bounding boxes and only these patches are compared with each other. If vehicles are close to each other, their bounding boxes are merged into a combined patch.

Modell	**FID↓**	**LPIPS-Score↑**
DP_GAN	100,40	0,4571
OASIS	106,49	**0,4644**
Kolors_steps	**37,39**	0,3804
Kolors_whole	42,42	0,3936
SDXL_steps	46,20	0,3971
SDXL_whole	48,37	0,4098

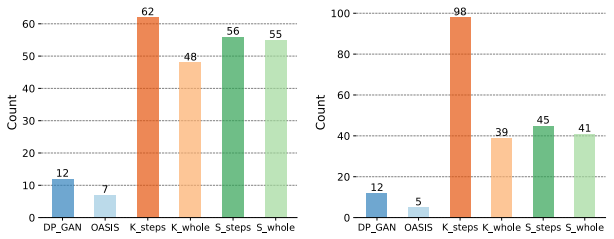
Table 1. Results for FID (measures image quality) and LPIPS (anonymization effectiveness). Best results are bold, poorest results are underlined.

Table 1 shows the results for the two evaluated metrics for each pipeline. In terms of image quality (measured by the FID score), the diffusion-based methods achieve significantly lower results than the GAN-based methods. The *Kolors_steps* method achieves the best result, followed by the other diffusion-based methods. In terms of anonymization effectiveness (measured by the LPIPS score), there are only minor differences between the two architecture types and between the implementation methods “whole” and “steps” for the diffusion models. GAN-based approaches achieve only slightly higher LPIPS values (which represent stronger anonymization). We can conclude from this quantitative evaluation that diffusion-based methods have a more balanced relationship between image quality and anonymization effectiveness than GAN-based approaches.

Qualitative Evaluation. As we were concerned that the metrics from the quantitative analysis might not correspond to the assessment of real people, we also conducted a survey with ten test subjects. This qualitative analysis is divided into two parts, in which four questions are answered for different images.
In the first part, the test subjects receive both the original image and the anonymized images created with the six anonymization pipelines (see an example in Figure 7). A total of 80 original images are evaluated. Each image is shown to three test subjects to avoid bias. The test subjects answer the following questions: *Q1: Which image achieves the best balance between anonymization and realism?* and *Q2: Which image looks the most realistic?*. The test subjects were instructed to pay attention to

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W7-2025-9-2025 | © Author(s) 2025. CC BY 4.0 License.

the visual integrity of the images (e.g., shading, lighting, sharpness) and only then to include logical aspects of the scene (e.g., correct direction of travel of the cars, realisitc size of the cars). The aim of this approach is primarily to determine whether the anonymization is recognizable at first glance.
In the second part, the test subjects receive pairs of images, consisting of the original image and an anonymized version. Each method is evaluated 40 times in total. Figure 8 shows an example. The test subjects answer the following questions on a five-point Likert scale (1 = very poor, 2 = poor, 3 = satisfactory, 4 = good, 5 = very good): *Q3: How anonymized are the cars compared to the original?*, and *Q4: How well do the generated cars fit into the scene?*. As a guideline, the test subjects were encouraged to consider features such as the color or model of the cars. In case of doubt, they should rely on their personal perception. The goal is to identify those anonymization methods that deliver convincing results, especially for Q1, as it directly reflects the trade-off between anonymization and image integrity. Q2 helps us understand which methods produce particularly realistic images with high visual integrity. Q3 and Q4 provide additional insights into the performance of individual anonymization methods.

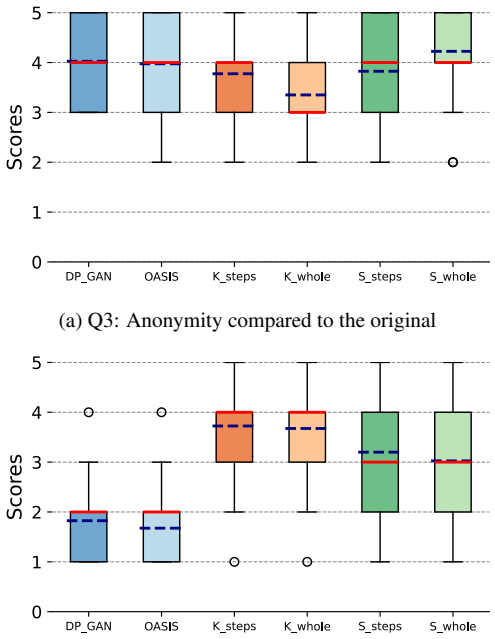


(a) Q1: Best balance between anonymization and realism

(b) Q2: Most realistic

Figure 9. Models selected in Q1 and Q2.

Regarding the balance between anonymization and realism (Q1, see Figure 9), the test subjects showed a clear preference for diffusion-based methods. *Kolors_steps* was chosen most frequently, followed by the other diffusion approaches *Kolors_whole*, *SDXL_whole*, and *SDXL_steps*. To our surprise *SDXL_steps* was chosen less frequently than both "whole" variants. The superiority trend of the "steps" methods is less evident here than in the quantitative analysis. The GAN methods were chosen much less frequently, indicating difficulties in presenting vehicles anonymously and realistically at the same time. *Kolors_steps* also performed by far the best in terms of perceived realism (Q2, see Figure 9), again followed by other diffusion-based methods. GAN-based methods seem to produce images with lower visual integrity of the cars.
As all methods were regarded separately, we present the results for Q3 and Q4 as a distribution of votes (see Figure 10). For Q3, the *SDXL_whole* method was rated highest, while this method also had the lowest standard deviation compared to the other approaches. The GAN-based methods were rated better than the other three diffusion-based approaches, which corresponds to the measures of LPIPS values from the quantitative analysis. The *SDXL_steps* method scored the lowest mean value in the comparison. Since all methods are in a similar range, no clear superiority trend for any one method can be inferred in terms of anonymization performance. Regarding the visual integration of the generated vehicles into the scenes (Q4, see Figure 10), there are clear tendencies in favor of the diffusion-based approaches. Both models using the "steps"



(a) Q3: Anonymity compared to the original

(b) Q4: Integration of cars

Figure 10. Answer distribution for Q3 and Q4. Mean: Dashed blue line. Median: red line.

method achieved high mean values. The test subjects predominantly agree that the GAN-based methods integrate the vehicles poorly into the scene, which is consistent with our findings in Q2. The diffusion-based methods have a higher standard deviation, indicating greater disagreement among testers about the integration quality. We also noted that diffusion models implemented with the "steps" method have slightly lower standard deviations than those using the "whole" method.

Performance. As a first performance indicator, we use the inference time of the anonymization methods for full images. Additional overheads such as combining car sections with the original or creating the input masks are not considered. Second, we record the VRAM used during inference to compare the memory requirements of the models and to better evaluate their practical applicability in resource-constrained environments.
Table 2 shows the measured times on a Tesla V100 GPU and the occupied VRAM. While GAN-based methods require only milliseconds to process an image, the processing time for diffusion-based methods ranges from several seconds to almost three minutes. In particular, models implemented with the "steps" method have long image processing times. On average, the "steps" method takes 9.32 times longer than the "whole" for the same model. The analysis of GPU memory utilization (VRAM) shows that diffusion-based models place significantly higher demands on the hardware than the GAN-based approaches consuming more than twice as much memory.

Method	Time per Image [s]	VRAM [MiB]
DP-GAN	0,0036	5.684
OASIS	0,0043	5.622
Kolors_steps	172,52	12.680
SDXL_steps	156,99	14.002
Kolors_whole	18,72	12.680
SDXL_whole	16,67	14.002

Table 2. Inference time and occupied VRAM.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W7-2025-9-2025 | © Author(s) 2025. CC BY 4.0 License.

Findings. The results clearly show that diffusion-based approaches generate cars of higher image quality, while GAN-based methods are particularly convincing due to their short inference time and lower resource load. This makes GANs particularly interesting for cases where high-performance GPUs cannot be used, such as with edge devices. All methods perform similarly well in terms of anonymization effectiveness. Thus, the primary trade-off is between image integrity and performance, while anonymization effectiveness and image integrity can be achieved together. This described trade-off also exists between our implementation methods "whole" and "steps"for the diffusion models. The integration into the image is rated better with the "steps" method but takes longer.
We have also seen that the Kolors model generally achieves better results in terms of image integrity than the SDXL model, albeit with a higher inference time. Regarding GAN-based methods, DP-GAN achieves better results compared to OASIS in basically all criteria examined, with the exception of the LPIPS score. While *Kolors_steps* achieves the best ratings for Q1 (balance), Q2 (realism), Q4 (integration), and the FID score, indicating high image quality and visual integration, the method performs weaker for Q3 (anonymization effectiveness) and the LPIPS score. The *SDXL_whole* method achieves the best result for Q3, and second best for Q2, Q4, and the FID score.
The evaluation of anonymization effectiveness shows a smaller range of standard deviations compared to image integrity, which might indicate a more uniform understanding of the term "anonymization" than "image integrity" among the testers. We suspect that the requirements for an anonymized car, such as changing the model or color, are relatively well-defined and intuitively understandable. The evaluation of image integrity may be more subjective, as it depends on aesthetic preferences or expectations of visual coherence.

5. DISCUSSION

In future work, it would be interesting to test lightweight diffusion models, which may provide a better trade-off between image integrity and performance. In addition, a GAN generator developed specifically for vehicle inpainting could be promising. Another option would be to fine-tune diffusion models to have a bigger diversity of of generated cars. A cost-benefit analysis should be considered for practical use cases. So far, we only tested the models for the case of modifying cars. And although we would expect similar results for other objects such as trucks or windows, it would be interesting to extend the study. As for practically all vision models that are used in outdoor scenes, experiments with different light and weather conditions would help evaluate the robustness of the generative models.
Future work may also include an analysis of different downstream models (e.g., detection of traffic signs) to evaluate the practical use of the anonymized images. To increase the reality of the generated images, the driving or parking direction of the cars could also be taken into account in further studies.

6. CONCLUSION

In this work, we investigated the suitability of GANs and diffusion models for modifying cars in urban image datasets by replacing them with synthetically generated image patches. We compared different models in terms of their anonymization effectiveness and the degree to which differences occur in image quality and performance. We identified two GAN-based I2IT models (DP-GAN, OASIS) and two diffusion-based inpainting models (Kolors, SDXL) as suitable and implemented them for the task. DP-GAN was found to perform better than OASIS in almost all evaluations. Among the diffusion models, Kolors was particularly convincing because of its high image quality and integrity, which can be attributed to the additional diffusion steps in the SDXL base architecture. GANs scored with fast inference times and lower resource consumption, while diffusion models offered great flexibility and visual quality.
To understand the trade-off between anonymization effectiveness, image integrity, and performance, we conducted a quantitative and a qualitative analysis. All methods achieved similar anonymization performance, while diffusion-based approaches delivered higher image integrity. Our developed "steps" method achieved the highest visual integrity but is computationally intensive. The "whole" method offers a pragmatic compromise between image integrity and performance. Thus, the central trade-off is between image integrity and technical performance. This finding is particularly relevant for practical use as it provides a basis for deciding whether a higher image quality can be aimed for in a specific application scenario, or whether limited resources require the use of more efficient methods.

ACKNOWLEDGEMENT

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI
We also thank the German Federal Ministry for Digital and Transport for funding the DiGuRaL project as part of the mFund program.

REFERENCES

Adeboye, O., Dargahi, T., Babaie, M., Saraee, M., Yu, C.-M., 2022. DeepClean: A robust deep learning technique for autonomous vehicle camera data privacy. *IEEE Access*.

AI, S., 2024. Stable diffusion 3 medium - diffusers. `https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers`. Accessed March 28, 2025.

alimama-creative, 2024. SD3-Controlnet-Inpainting. `https://huggingface.co/alimama-creative/SD3-Controlnet-Inpainting`. Accessed April 10, 2025.

Brkic, K., Sikiric, I., Hrkac, T., Kalafatic, Z., 2017. I know that person: Generative full body and face de-identification of people in images. *CVPRW*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *CVPR*.

Croitoru, F.-A., Hondru, V., Ionescu, R. T., Shah, M., 2023. Diffusion models in vision: A survey. *TPAMI*.

Datenschutz-Grundverordnung (DSGVO), 2016. Regulation 2016/679 of the European Parliament and of the Council.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. *NeurIPS*.

diffusers, 2023. Stable diffusion xl inpainting. `https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1`. Accessed March 26, 2025.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. *Communications of the ACM*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Hukkelås, H., Lindseth, F., 2023. Deepprivacy2: Towards realistic full-body anonymization. *WACV*.

Hukkelås, H., Smebye, M., Mester, R., Lindseth, F., 2023. Realistic full-body anonymization with surface-guided gans. *WACV*.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *CVPR*.

Klemp, M., Rösch, K., Wagner, R., Quehl, J., Lauer, M., 2023. Ldfa: Latent diffusion face anonymization for self-driving applications. *CVPRW*.

Kolors Team, 2024. Kolors: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis. *arXiv preprint*.

Kwai-Kolors, 2024. Kolors inpainting. `https://huggingface.co/Kwai-Kolors/Kolors-Inpainting`. Accessed March 28, 2025.

Lee, J. H., You, S. J., 2024. Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision. *IEEE access*.

Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., Doshi, S., 2024. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint:2402.17245*.

Li, S., Cheng, M.-M., Gall, J., 2022. Dual pyramid generative adversarial networks for semantic image synthesis. *arXiv preprint: 2210.04085*.

Liu, D., Wang, X., Chen, C., Wang, Y., Yao, S., Lin, Y., 2024. Svia: A street view image anonymization framework for self-driving applications. *ITSC*.

Malm, S., Rönnbäck, V., Håkansson, A., Le, M.-h., Wojtulewicz, K., Carlsson, N., 2023. Rad: Realistic anonymization of images using stable diffusion. *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*.

Piano, L., Basci, P., Lamberti, F., Morra, L., 2024. Latent diffusion models for attribute-preserving image anonymization. *arXiv preprint: 2403.14790*.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R., 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint: 2307.01952*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis. *ICML*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. *CVPR*.

Shakhmatov, A., 2023. kandinsky 2.1. `https://huggingface.co/kandinsky-community/kandinsky-2-1-inpaint`. Accessed April 10, 2025.

Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A., 2020. You only need adversarial supervision for semantic image synthesis. *arXiv preprint: 2012.04781*.

Szeliski, R., 2022. *Computer vision: algorithms and applications*. Springer Nature.

Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., Gavrila, D. M. et al., 2019. Privacy protection in street-view panoramas using depth and multi-view imagery. *CVPR*.

Wang, H., 2024. Comparative Analysis of GANs and Diffusion Models in Image Generation. *Highlights in Science, Engineering and Technology*.

Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D. J., Soricut, R. et al., 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *CVPR*.

Wu, Y., Yang, F., Ling, H., 2018. Privacy-protective-gan for face de-identification. *arXiv preprint: 1806.08906*.

Xiong, Z., Cai, Z., Han, Q., Alrawais, A., Li, W., 2020. ADGAN: Protect your location privacy in camera data of auto-driving vehicles. *IEEE Transactions on Industrial Informatics*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*.

Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I., Xu, Y., 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint:2103.10428*.

Zheng, H., Lin, Z., Lu, J., Cohen, S., Shechtman, E., Barnes, C., Zhang, J., Xu, N., Amirghodsi, S., Luo, J., 2022. CM-GAN: Image Inpainting with Cascaded Modulation GAN and Object-Aware Training. *arXiv preprint:2203.11947*.

Zulfiqar, A., Muhammad Daudpota, S., Shariq Imran, A., Kastrati, Z., Ullah, M., Sadhwani, S., 2024. Synthetic Image Generation Using Deep Learning: A Systematic Literature Review. *Computational Intelligence*.