AI scratching your car: Using diffusion models for training data generation in automotive damage detection

Julian Strietzel[1], M. Saquib Sarfraz[1,2], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology
[2] Mercedes-Benz Tech Innovation

Abstract Demand for reliable data remains a major issue in training machine learning models in computer vision. Frequently, datasets are of insufficient scale, imbalanced, not diverse, and of poor quality, potentially resulting in biased, inaccurate, non-robust, and badly generalizing models. Moreover, real-world training data can raise privacy concerns or be extremely expensive to gather, necessitating alternative solutions.
This paper investigates the use of diffusion models for generative data augmentation in semantic image segmentation, specifically in the domain of vehicle damage detection. We propose a new approach that utilizes an existing diffusion model ControlNet to generate useful synthetic data depicting realistic vehicles with damages such as scratches, rim damages, dents and etc. Based on this we provide an analysis and show how such a generative data augmentation may help in scenarios where training data is scarce and of low quality.

Keywords Generative data augmentation, diffusion models, ControlNet, damage detection

1 Introduction

A major challenge in Deep Learning for Computer Vision persisting is the scarcity and quality of training data, which is crucial for training robust and generalizing models. Acquiring a large quantity of detailed and balanced images for training is often time-consuming, expensive,

DOI: 10.58895/ksp/1000174496-18 erschienen in:
Forum Bildverarbeitung 2024
DOI: 10.58895/ksp/1000174496 | https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000174496/

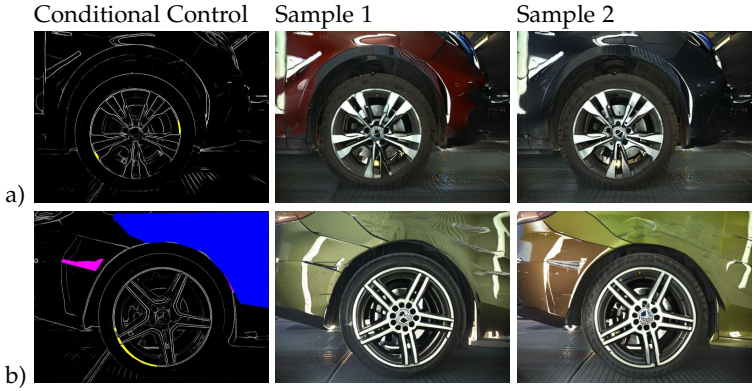

Figure 1: Edge detection maps with color patches as labels used as conditional control input and respective generated images from our trained ControlNet. With a) rim damages (yellow), b) deformations (blue), scratches (pink), and rim damages (yellow).

and sometimes impossible. This paper aims to address the challenge of limited training data in the domain of vehicle damage detection by investigating the use of ControlNet [1] for Generative Data Augmentation (GDA) [2] in scarce and low-quality data scenarios. We trained ControlNet based on StableDiffusion (SD) to generate synthetic images of damaged cars from labeled edge-detection maps and use these generated images to train segmentation models for damage detection (see fig. 1).

In automotive damage detection, semantic segmentation models may be used to recognize various types of exterior damages on images for efficient vehicle inspection. In practical industrial setting usually images of passing cars are taken autonomously from multiple angles by vehicle scanners and sent to cloud processing, to recognize several damage classes. Due to the nature of the damages it is quite impractical, even impossible in some cases to collect such data manually.

Guiding this are the questions about (1) how synthetic training data transfers to real-world evaluation, (2) its capacity to tackle challenges of scarcity, quality, and bias in training data, and (3) the effect on model generalizability. In the process, we evaluate parameters, design decisions, and training data compositions in extensive ablation studies and

experiments. The use of GDA has not been explored yet on this problem and may increase the potential of synthetic data in vehicle damage detection scenarios.

2 Related Work

Data augmentation addresses challenges like data scarcity, lack of diversity, and overfitting by creating new label-preserving examples from existing datasets [3]. Common augmentation approaches include image manipulation, image erasing, image mixing, auto-augment, feature augmentation, and neural style transfer [3]. GDA involves supplementing training data with synthetic examples to improve model performance, especially when only little training data is available and overfitting is of concern [2]. Classic methods in computer vision include CGI placement [4], model renderings [5–7], and degrading techniques [8]. Training data generation may help increase diversity, generalization capabilities, and robustness including to adversarial attacks.

Synthetic data from Denoising Diffusion Probabilistic Models (DDPM) [9] prove to be effective for GDA in ImageNet [10] classification [11], even achieving new state-of-the-art scores using supplemented real training data [12]. Synthetic data is also employed to fight representation bias [13, 14] and privacy concerns [15] in medical image data, with curation of synthetic data proving important [14]. For segmentation model training, mostly Generative Adversarial Networks (GANs) [16] have been employed to generate samples, using a decoder to extract pixel-wise annotations from latent space [17, 18], and showing performance gains mainly in out-of-domain data. Only recently DDPMs have been explored for GDA, mostly following a similar approach, extracting labels from attention-maps [19] or training a grounding model to align pixels with textual representations [20]. Apart from their superior sample quality [21], DDPMs for GDA face challenges including dataset memorization [22], diversity [23] and do not generally outperform GANs from scratch [24].

ControlNet is a neural network structure aimed to further condition the output of DDPMs [1]. ControlNet is a copy of an arbitrary neural network block, running in parallel to the original network, incorpo-

rating an encoding of additional control input and feeding its guided output back to the main structure. During training of ControlNet, the original model is locked to preserve its distilled knowledge, only the parallel, duplicated blocks are trained for guidance. Using this architecture we can control DDPMs to exactly match input conditions, like edge detection maps, human poses, or drawings, even with comparably low training data available.

3 Methodology

The core part of this paper is the implementation of ControlNet generating pre-labeled data as GDA for segmentation model training for damage detection. We consider four commonly occurring damage classes: deformation, dent, rim damage and scratch.

In this section, we discuss how to guide and train ControlNet to generate pre-labeled samples for semantic segmentation training.

Conditional Control We utilize ControlNet's conditional control feature to generate precise images representing specific views of cars and damage positions. Edge detection maps, identifying image boundaries, serve as the control input. This approach offers a balance between detailed output descriptions and the freedom to generate varied results and has proven to work well with ControlNet [1].

For pre-labeled sampling, we need to include detailed label information in the conditional control. We propose to include the labels using color patches on the black-and-white edge detection maps (see fig.1). This approach of labeling the conditional control to generate pre-labeled training data has, to our knowledge, never been evaluated before.

It has the following advantages: (1) Efficient placement and generation from existing labels, (2) effective guidance for ControlNet, (3) easily differentiable by eye, (4) covering edge detection maps on relevant positions, and (5) referenceable in text prompts by naming the respective color.

Text prompts play a crucial role in text-to-image generation, functioning as fundamental guidance, and imparting context and semantic information to the generative process. When using ControlNet, the

textual prompt introduces background information guiding the interpretation of the conditional control. The integration of semantically relevant textual information to image generation results in more precise and sophisticated outcomes, potentially improving its capability for GDA.

Text prompts are generated following a specific prompting schema: (1) The applying short description of the relevant damage: *Rim damage at the yellow marking, Scratch at the pink marking, Dent at the green marking, Deformation at the blue marking*, (2) a background prompt to define the image, context, and style: *side of a car in a workshop, high quality, detailed, and professional image*.

Training Generative Model We train the generative model on the available real-world training data, to generate damaged cars matching the conditioning. Apart from its comparably low requirements on training data, ControlNet training is exhibiting a sudden convergence phenomenon [25], which we take into account asjusting the virtual batch size using gradient accumulation to reach around 10k steps during training.

Fine-Tuning In visual evaluation, samples of different damage classes showed significant deviations in image quality, suggesting that the generative model might be improved by fine-tuning it on specific damage classes. We filtered the 17k dataset to include only images containing instances of the respective class and trained a generative model for each one. We name these fine-tuned models (damage-class-) *specific*.

4 Experiments

Experiments have been split into (1) tuning and evaluating the image generation process and (2) optimizing training of segmentation models from (partly) synthetic data. The used dataset includes 17k hand-labeled images from a vehicle scanner containing from 1k to 9k instances per damage class (see appendix).

4.1 Image Sampling

Firstly, we conducted experiments to evaluate the quality of synthetic images for different parameters and ablations. To measure improvements, we employ an existing segmentation model, trained on the existing real17k dataset and evaluate it on our generated datasets. We expect correlation between sample quality and the model's ability to recognize synthetic damages, measured by the evaluation F-Scores.

We evaluated the prompting schema defined in section 3: As the models had been trained with the fixed schema, additional background and negative text prompts, as well as no prompts at all resulted in worse samples. This suggests that our prompts need to stick to the scenario or have to be trained on a more diverse prompt landscape. We also evaluated a pre-trained ControlNet from a large edge detection dataset on damage generation, which was not able to extract meaningful results from the guidance input, though.

4.2 Training Data Compositions

Evaluating a model trained from synthetic data on the real evaluation dataset, a first approach showed a significant domain gap between real and synthetic images, with F-Scores of less than 0.1. In the second part of our experiments, we therefore evaluate (only partly) synthetic training data compositions to train segmentation models based on their downstream performance on real evaluation data.

Damage-Specific Training Data To evaluate samples from damage specific generative models, we combined 2k samples for each class with 2k samples from a general model to a new dataset - *specific10k* - for segmentation model training. Compared to 10k samples from the general model, we seem to slightly improve F-Scores on average, reflecting the the results from class-specific sample evaluation. Furthermore, we are able to improve over *real17k* training data in the deformation class, representing the most scarce and low-quality training data, increasing the F-Score by $\sim$ .03 when using specific training data.

We also supplemented fake training data to real17k, instead of using it isolated, and employed a quality filter to curate the samples for training. We used an IOU threshold of 25% per image from evaluating

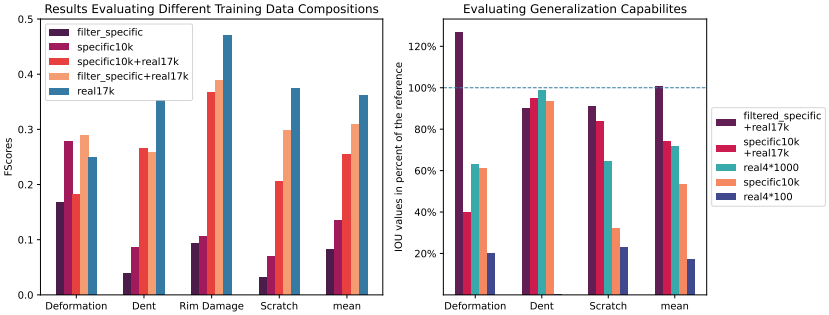

Figure 2: Left: F-Scores of segmentation models trained using different (synthetic) datasets evaluated on real test data. *Specific10k* refers to samples from damage-specific generative models, with *real17k* referring to the original training data as reference. *Filtered* referencing to a quality threshold. **Right**: Generalization IOU values of trained segmentation models (synthetic and limited training data) on out-of-domain data in proportion to the same reference model.

the images by a pre-trained segmentation model as in image sampling experiments (section 4.1). Using both supplementation and filtering greatly reduced the synthetic-to-real-data gap, resulting in only .04 difference in macro average scores. This was primarily due to a further improvement in deformation accuracy (see fig. 2), where we increased the existing lead over real training data. Notably, we still decrease overall performance by supplementing fake data to our training, especially in well-represented classes. This suggests a key difference in data distribution regardless of the visual quality of samples, but we show potential application and benefit of GDA for very scarce classes.

Comparison on limited data To find a threshold of data availability where synthetic samples outperform real training data, we limit available real training data (to 25, 100, 250, and 1000 examples per class). As expected, limiting the availability of real training data negatively impacts overall segmentation performance. Decrease differs from class to class, with scarce classes (dent and deformation) benefiting from more balanced training data. Synthetic data outperforms very limited real datasets (25 samples per class) across all damage classes and enhanced datasets are competitive to larger real datasets (up to 1k), especially in

rim damage.

In this damage detection scenario, the targeted threshold seems to be between 25 and 100 images per class.

Generalizability To assess generalizability, we evaluate models trained with GDA on a new dataset of 250 labeled samples from unseen locations and vehicle scanners (out-of-domain dataset) and compare them to limited real datasets. Especially in the deformation class, the GDA-trained model (filtered, specific, and supplemented samples) significantly outperforms the reference model in the out-of-domain setting by 25 % (see fig. 2). Even on average, filtered supplemented data outperforms real training data, with even non-filtered outperforming up to 1k real samples per class, and generated samples only still significantly dominating over 100 real samples. Synthetic data improves generalization performance compared to limited datasets, particularly outperforming the original training data in scarce classes. This underlines the potential of synthetic data especially when it comes to generalization, where GDA shows more competitiveness than during in-domain evaluation.

5 Discussion

We show how ControlNet with StableDiffusion can be effectively used to generate pre-labeled, high-fidelity images for GDA in image segmentation tasks. In the context of vehicle damage detection the model demonstrates the ability to accurately place damages on vehicles.

Our experiments and ablation studies have revealed several key factors that can contribute to optimizing ControlNet generative performance. Parameter tuning and input specifications significantly improved sample fidelity. The ablation studies further provided valuable insights into the role of various techniques guiding and training ControlNet: We discovered the necessity of fine-tuning and additional text prompts, incorporating quality guidance. Finally, tuning damage-class-specific generative models for specific damage classes is beneficial, compared to a general multi-class generative-model. We showed how our synthetic data can be effectively used to train segmentation models for damage detection: Synthetic data alone can enhance seg-

mentation performance for very scarce classes and generally outperform limited real data when only a few samples (less than 50) are available. Especially scarce classes can benefit from *additional* synthetic training data. Furthermore, GDA can, with some limitations, be used to increase the generalization capabilities of our segmentation models, where supplemented fake data is outperforming the real dataset, especially limited to a few hundred examples only. Key findings from this study align with prior research GDA.

As a key takeaway we note that filtered, specific generated datasets supplementing real data can increase in-domain and especially out-of-domain performance significantly for scarce data classes. However, synthetic-only data remains no match to real large scale datasets, due to a significant distribution shift between real and fake samples. This is emphasized by a significant performance drop when GDA-trained models are evaluated on real-world test data, indicating how synthetic data may not fully capture the nuances and variations present in real-world data. We show that GDA in this use-case is mostly not effective for well represented data classes.

Future Work To guide effective utilization of GDA we suggest employing synthetic data when real data is very scarce. When using ControlNet for GDA, stronger guidance and increased steps benefit sample fidelity. Furthermore, analyzing the characteristics of synthetic compared to real samples and employing inpainting for GDA might be promising directions.

Acknowledgment We would like to express our gratitude towards Mercedes-Benz Tech Innovation GmbH for providing valuable compute, data, and hardware that significantly contributed to the research presented in this paper. Their cooperation and support for this research played a crucial role in enabling us to carry out the experiments and analysis necessary.

We would also like to thank our colleagues at the Autonomous Systems Karlsruhe Team, for their valuable insights, constructive feedback, and collaboration throughout the research process. Their expertise and dedication have significantly contributed to the quality of this work.

6 Appendix

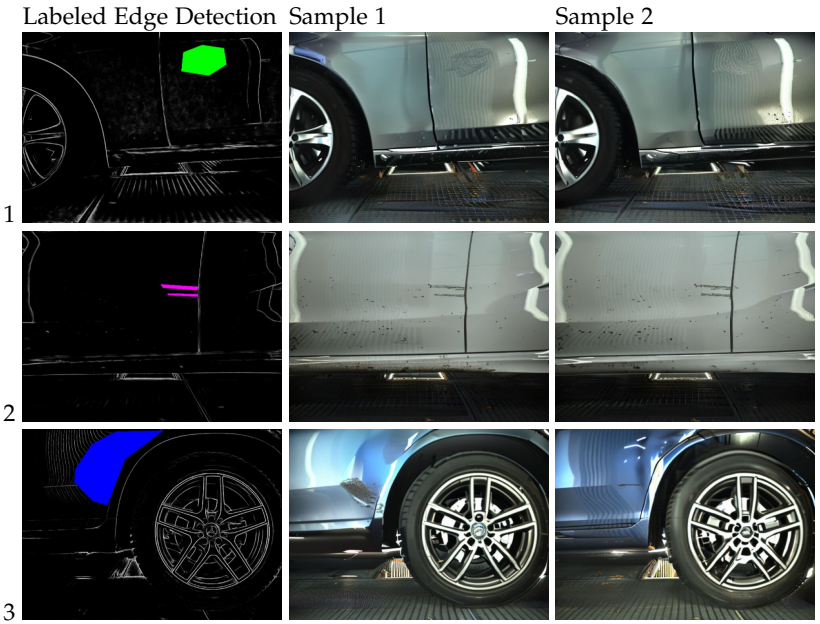

Figure 3: Sampled images of different damage classes from shown labeled edge detection maps as conditional control input for ControlNet.

Segmentation Model We use a U-Net [26] from the segmentation model library [27]. The reference model is trained on the real17k training data and does not represent the performance of similar models in production. **ControlNet** version 1.0 with StableDiffusion 2.1 is trained on the same T4 for 15 epochs on a virtual batch size of 32 for about 8,000 steps, taking around 100 hours per model. Virtual batch size is reached by using gradient accumulation of 32.

6.1 Datasets

The used dataset - real17k - contains 17467 manually labeled images of cars with different damages (9234 rim damages, 8685 scratches, 1803 dents, 972 deformations). The test dataset contains 739 images of which 177 include rim damages, 168 scratches, 104 dents, and 11 deformations. The in-domain images are taken from an automatic vehicle scanner at entry points to a workshop. They all come from the same location, taken with the same equipment, lighting conditions, background, and surroundings. The weather is similar with only a few images containing rainy or snowy conditions.

The **limited datasets** real4*x with $x \in 25, 100, 250, 1000$ contain a random sample from the real17k dataset. They are used to simulate a scenario, where only a limited but balanced amount of data is available.

The **out-of-domain** dataset to test generalization contains 170 images, with some being from different locations and types of vehicle scanners. The dataset contains 169 images of which 31 deformations, 32 dents, 0 include rim damages, and 128 scratches.

Synthetic Datasets The **specific10k** dataset of generated synthetic data contains 2k images per class generated from class-specific generative models and 2k images from a general model. The **filtered** dataset contains all images from specific10k, that passed a quality threshold established as an IOU greater than .25 in evaluation using a real-world pre-trained segmentation model: Deformation 339 (from 2940 samples containing instances in total), Dent 392 (3095), Rim Damage 876 (3804) & Scratch 490 (5164).

References

1. L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," Feb. 2023, arXiv:2302.05543 [cs]. [Online]. Available: http://arxiv.org/abs/2302.05543
2. C. Zheng, G. Wu, and C. Li, "Toward Understanding Generative Data Augmentation," May 2023, arXiv:2305.17476 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2305.17476
3. T. Kumar, A. Mileo, R. Brennan, and M. Bendechache, "Image Data Augmentation Approaches: A Comprehensive Survey and

Future directions," Mar. 2023, arXiv:2301.02830 [cs]. [Online]. Available: http://arxiv.org/abs/2301.02830

4. V. Shakhuro, B. Faizov, and A. Konushin, "Rare Traffic Sign Recognition using Synthetic Training Data," in *Proceedings of the 3rd International Conference on Video and Image Processing*, ser. ICVIP '19. New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 23–26. [Online]. Available: https://dl.acm.org/doi/10.1145/3376067.3376105
5. W. Armstrong, S. Drakontaidis, and N. Lui, "Synthetic Data for Semantic Image Segmentation of Imagery of Unmanned Spacecraft," in *2023 IEEE Aerospace Conference*. Big Sky, MT, USA: IEEE, Mar. 2023, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10115564/
6. J. Adams, J. Sutor, A. Dodd, and E. Murphy, "Evaluating the Performance of Synthetic Visual Data for Real-Time Object Detection," in *2021 6th International Conference on Communication, Image and Signal Processing (CCISP)*, Nov. 2021, pp. 167–171.
7. M. Pergeorelis, M. Bazik, P. Saponaro, J. Kim, and C. Kambhamettu, "Synthetic Data for Semantic Segmentation in Underwater Imagery," in *OCEANS 2022, Hampton Roads*, Oct. 2022, pp. 1–6, iSSN: 0197-7385.
8. X. Nie, M. Yang, and R. W. Liu, "Deep Neural Network-Based Robust Ship Detection Under Different Weather Conditions," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Oct. 2019, pp. 47–52.
9. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020, event-place: Vancouver, BC, Canada.
10. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y
11. M. B. Sariyildiz, K. Alahari, D. Larlus, and Y. Kalantidis, "Fake it till you make it: Learning transferable representations from synthetic ImageNet clones," Mar. 2023, arXiv:2212.08420 [cs]. [Online]. Available: http://arxiv.org/abs/2212.08420
12. S. Azizi, S. Kornblith, C. Saharia, M. Norouzi, and D. J. Fleet, "Synthetic Data from Diffusion Models Improves ImageNet Classification," Apr. 2023, arXiv:2304.08466 [cs]. [Online]. Available: http://arxiv.org/abs/2304.08466

13. L. W. Sagers, J. A. Diao, M. Groh, P. Rajpurkar, A. Adamson, and A. K. Manrai, "Improving dermatology classifiers across populations using images generated by large diffusion models," in *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022. [Online]. Available: https://openreview.net/forum?id=Vzdbjtz6Tys

14. M. Akrout, B. Gyepesi, P. Holló, A. Poór, B. Kincső, S. Solis, K. Cirone, J. Kawahara, D. Slade, L. Abid, M. Kovács, and I. Fazekas, "Diffusion-based Data Augmentation for Skin Disease Classification: Impact Across Original Medical Datasets to Fully Synthetic Images," Jan. 2023, publication Title: arXiv e-prints ADS Bibcode: 2023arXiv230104802A. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2023arXiv230104802A

15. S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle, "Differentially Private Diffusion Models Generate Useful Synthetic Images," Feb. 2023, arXiv:2302.13861 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2302.13861

16. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, Dec. 2014, pp. 2672–2680.

17. Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 10 140–10 150, iSSN: 2575-7075.

18. D. Li, H. Ling, S. W. Kim, K. Kreis, S. Fidler, and A. Torralba, "BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 21 298–21 308. [Online]. Available: https://ieeexplore.ieee.org/document/9878775/

19. W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models," Mar. 2023, arXiv:2303.11681 [cs]. [Online]. Available: http://arxiv.org/abs/2303.11681

20. Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Guiding Text-to-Image Diffusion Model Towards Grounded Generation," Jan. 2023, arXiv:2301.05221 [cs]. [Online]. Available: http://arxiv.org/abs/2301.05221

21. P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*,

M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

22. M. U. Akbar, W. Wang, and A. Eklund, "Beware of diffusion models for synthesizing medical images – A comparison with GANs in terms of memorizing brain tumor images," May 2023, arXiv:2305.07644 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2305.07644
23. M. F. Burg, F. Wenzel, D. Zietlow, M. Horn, O. Makansi, F. Locatello, and C. Russell, "A data augmentation perspective on diffusion models and retrieval," Apr. 2023, arXiv:2304.10253 [cs]. [Online]. Available: http://arxiv.org/abs/2304.10253
24. M. U. Akbar, M. Larsson, and A. Eklund, "Brain tumor segmentation using synthetic MR images – A comparison of GANs and diffusion models," Jun. 2023, arXiv:2306.02986 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2306.02986
25. L. Zhang and M. Agrawala, "ControlNet/docs/train.md at main · lllyasviel/ControlNet." [Online]. Available: https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md
26. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
27. P. Iakubovskii, "Segmentation Models," 2019, publication Title: GitHub repository. [Online]. Available: https://github.com/qubvel/segmentation_models