Data Augmentation for Road Segmentation using Diffusion Models

Gian Hess, Patrick Eppensteiner, William Wong
{gihess, eppatric, wiwong}@student.ethz.ch
Group: eppatric
Department of Computer Science, ETH Zurich, Switzerland

***Abstract*—Data annotation is a labor-intensive and expensive process. One way of making more efficient use of data is data augmentation. In this paper we investigate the application of diffusion models for data augmentation in the task of road segmentation. We aim to improve the generalization performance of a segmentation model by increasing its robustness to occlusions. To this end, we propose a method to augment the data set by masking and inpainting images with diffusion models while retaining ground truth information. We train a U-Net architecture using the inpainted data samples and compare it with other techniques such as basic data augmentation or using additional real data. The inpainting method achieves an improvement of 2% over using the original data set and 1% over using an external data set of bigger size, but only 0.02% over basic data augmentation. Although the improvement over basic data augmentation is not substantial, the method is promising because it can potentially introduce more diverse and realistic data than simpler augmentation methods.**

I. Introduction

Labelled data is often scarce, which poses a challenge for deep learning-based techniques [1]. One way to alleviate this is through data augmentation, a set of techniques aiming to increase the amount of available data [2]. These techniques create new data samples by applying certain transformations that do not change the semantic meaning of the original data. By training models on these augmented data samples, one can enforce the models to be invariant to the transformations and improve their generalization performance [3]. In image classification, typical transformations are random translations, rotations or crops [2]. These transformations are easy to compute, however, the diversity in new data they can generate is limited [4]. For example, in road segmentation, it could be beneficial to ensure that models are robust to occlusions of the road surface, such as those caused by trees or other objects. Standard data augmentation techniques may not be sufficient to simulate such scenarios. Random erasing [5] is a possible alternative, but it may compromise the realism of the augmented samples. In the last few years, GANs [6] were employed for data augmentation, e.g., to transfer the style of one image to another to create new training data [7], [8], [9]. More recently, diffusion models [10] have been used to generate new, diverse training data [4]. Most previous work using generative models for data augmentation focuses on classification, where each sample is assigned a label from a finite set of possible classes. While semantic segmentation can be formulated as a classification task in which each pixel is assigned a class, it introduces an additional difficulty, namely that the position of the objects matters. Therefore, care must be taken to ensure that the ground truth of newly generated images is still informative. In this work, we explore a data augmentation strategy that uses diffusion models for the task of road segmentation. More concretely, we base our approach on the recent work of Lugmayr et al. [11] which created a method to add semantically consistent content to regions in images specified by a binary mask. We also explore the feasibility of using diffusion models to generate completely new annotated samples or directly extracting segmentation masks for test images and discuss the challenges and limitations of these approaches.

II. Preliminaries

A. Inpainting using Diffusion Models

In this paper, we utilize the free-form inpainting capabilities of the RePaint model, as introduced in the paper "RePaint: Inpainting using Denoising Diffusion Probabilistic Models" [11]. This model is based on Denoising Diffusion Probabilistic Models (DDPM) and allows inpainting missing regions with semantically meaningful content. This is achieved by conditioning the generation process by sampling from the known pixels during the reverse diffusion iterations.

III. Models and Methods

A. Baseline

The U-Net, introduced by Ronneberger et al., has been proven to work effectively on the task of Image Segmentation [12]. For our baseline model, we trained a U-Net architecture with six blocks in the encoder and five blocks in the decoder. The encoder has (3, 64, 128, 256, 512, 1024) channels while the decoder's channel configuration is reversed, starting with 1024 and decreasing down to 64. For our two baselines, we train the U-Net using two approaches:

1) Standard data set: The first baseline is solely trained with the provided data set, which is split randomly into training and validation set.

2) Basic augmentation: The second baseline involves basic data augmentation using the Albumentations library [13]. A sequence of transformations is applied to each training sample α times (α denoting the augmentation factor). The

sequence consists of horizontal flipping, random rotation, brightness/contrast adjustment and elastic deformation. Each transformation in the sequence is applied with a specified probability to enhance diversity. One augmented example is illustrated in Figure 1.

Figure 1: Example of augmented training sample with $\alpha = 3$

B. Equipping U-Net with Attention

As an additional neural network used for prediction, we modified the U-Net by implementing attention gates [14] on the skip connections, hoping to improve the predictive power of the network, following the approach of [15] and [16].

C. Data Augmentation by Inpainting

In the following we describe our strategy to augment our data set using diffusion models. Our original images have a resolution of 400x400 pixels, which is incompatible with the RePaint model that operates on 256x256 pixels. Note that this restriction is not due to RePaint itself, but due to the underlying pre-trained diffusion model used. Therefore, we first crop the images to fit the RePaint input size. Next, we generate a random mask with three fixed-size boxes that cover some parts of the image. We use RePaint to inpaint the masked regions with realistic textures. Finally, we combine the inpainted image with the original image to restore the full resolution. An example illustrating the masking and inpainting of an image is shown in Figure 2.

(a) (b) (c)

Figure 2: a) A cropped image from our data set. b) The same image with a random mask consisting of three boxes applied to it. c) The inpainted image obtained by filling in the masked regions using RePaint. Note that a part of the road is now occluded by a newly added tree.

The form of the masks was motivated by the following criteria: The occluded regions should not be too small, because RePaint would restore most of the image details. This is also the case if a large part of the image is masked in total, but there is no large contiguous masked region, see Figure 5 in the appendix for an example. Nevertheless, they should not cover too much of the image to maintain consistency with the ground truth. If the area to be painted over is too large, not enough context information is available to keep the road network unchanged. The problem is exacerbated by the fact that the underlying diffusion model was not trained on satellite images, but on the Places2 [17] data set. Therefore, if large regions have to be inpainted, the images completely lose their resemblance to satellite images.

D. Using additional real data

To compare the performance of our augmented data set with the performance of a similar sized data set consisting of real samples, we conducted experiments with data obtained from the DeepGlobe Road Extraction Challenge [18]. The DeepGlobe data set contains 6'226 satellite images, each with a size of 1024x1024 pixels, along with their road labels as a grey scale mask. To align with our data set's dimensions, we cut each image in four smaller images and resize them to 400x400 pixels.

However, we observed that the DeepGlobe data set contains a substantial amount of images of rural areas with only a few small roads visible. As a result, when we divided the image into four pieces, some parts did not contain any roads or only contained a small fraction of the road. To make the DeepGlobe data set comparable to ours, we decided to discard images with segmentation masks that had very few road pixels.

To determine a suitable threshold, we analyzed the distribution of the number of road pixels per training sample. We found that approximately 98.5% of the samples have a segmentation mask with more than 7.5% of the pixels annotated as roads. We utilized this threshold and removed images accordingly from the DeepGlobe data set, resulting in a final data set of 4'025 images from DeepGlobe.

E. Generating new annotated samples with Diffusion Models

(a) (b) (c)

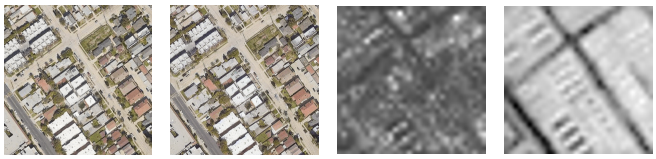
Figure 3: a) Image of simple roads generated by the diffusion model using the prompt "a image of a road from above". b) The aggregated cross-attention map for the token "road". c) Simple threshold applied to the cross-attention map. The full attention maps for all tokens can be found in the appendix in Figure 7.

When generating novel samples from scratch, rather than modifying existing ones, we require a method to annotate them automatically. Hertz et al. [19] observed that the underlying attention maps respect the spatial and geometric layout of the generated image. Therefore, we can use the spatial attention maps associated with the words ”road” or ”street” as ground truth.

However, the images generated are very simple, usually depicting only a single road and as such do not fit quite as well into the existing data set. Therefore, we did not consider this method in our experiments. An example can be found in Figure 3.

F. Segmenting with Diffusion Models

The above approaches involve creating or modifying data with a diffusion model and then training another model on that data. The question may arise whether the information contained in the diffusion model can be extracted in a more direct way. One possible approach we can imagine is as follows: Using null-text inversion [20] with a base prompt allows inverting images to their latent representation and reconstructing them, therefore enabling us to retrieve attention maps. Depending on the quality of the attention maps, they could then be used directly as prediction or as further aid, e.g., as an extra input or channel of the image.



(a) (b) (c) (d)

Figure 4: a) Original image of the base data set. b) The reconstructed image using the prompt ”satellite image of city streets”. c) Aggregated cross-attention map of the token ”streets”. d) Suitable aggregated self-attention map. The full attention maps can be found in the appendix in Figure 8.

IV. Results

Each method is trained for 35 epochs with an early stopping strategy based on the patch accuracy on the validation set. We checkpoint the model each time a new highest patch accuracy is achieved. We perform k-fold cross validation with $k = 6$ for each method and report the average validation patch accuracy and the standard error in Table I.

As the provided data set contains 144 images, the validation set consists of 24 and the training set of 120 images. To compare the proposed augmentation method with the other methods, we ensured that the number of training samples are the same across all methods, except for the vanilla U-Net with the Standard data set. For basic augmentation we used the augmentation factor of $\alpha = 7$. For DeepGlobe, we took $7 \cdot 120 = 840$ additional samples from the DeepGlobe data set. For Masking and Inpaiting, we augment the training data set with $\alpha = 3$ and take 4 inpainted/masked versions for each training sample and perform random transformation on them similar to the augmentation. Therefore, each method is trained with 960 samples.

Additionally, for each method, we select the model with the highest validation patch accuracy to submit on the ETHZ CIL Road Segmentation 2023 Kaggle competition, and present the public score also in Table I.

Method	Mean	SE	Kaggle
U-Net + Standard data set	0.8771	0.00461	0.88364
U-Net + Basic augmentation	0.9041	0.00356	0.90671
U-Net + DeepGlobe	0.8957	0.00372	0.89657
U-Net + Masking	0.9046	0.00435	0.90708
U-Net + Inpainting	0.9060	0.00438	0.90746
U-TNet + Inpainting	0.9049	0.00375	0.90311

Table I: Mean accuracy of the different methods

V. Discussion

The cross validation results of the experiments closely align with the competition’s public score. In the following, we discuss the patch accuracies achieved during cross validation. The inpainting technique improved patch accuracy by about 2.9% compared to the model solely trained with the standard data set and about 1% improvement over the DeepGlobe data set. However, inpaiting only achieves a marginal 0.02% improvement over basic augmentation. One possible reason for the marginal improvement is that the inpainting might not introduce enough variation for the U-Net to learn additional information, e.g., when the original details are almost completely restored. Additionally, inpainting may introduce roads in areas which are not defined as one in the segmentation mask (as presented in Figure 6), possibly negatively affecting the training. However, manual removal of such samples did not show any significant increase in performance in preliminary experiments. These limitations could potentially be addressed by fine-tuning a diffusion model to satellite imagery and constraining the generation process to create realistic occlusions only in road-containing areas and ensuring greater variation outside of roads.

The experiments further show that the inpainting method achieves similar results to plain masking, both achieving around 90% patch accuracy. Given the computational efficiency of plain masking, it presents a practical alternative.

The U-Net with attention gates, called UT-Net in Table I, did not outperform the unmodified U-Net.

The diffusion model used for both generation and prediction tasks is sufficient to generate simple images of roads with attention maps very close to what could be considered ground truth. As for the null-text inversion for prediction, the image can be reconstructed almost perfectly, and while the structure of the road can still be faintly recognized by a human (see Figure 4c), this would hardly suffice for the

segmentation process. On the other hand, the self-attention maps yielded some interesting results as seen in Figure 4d. The issue is that the self-attention maps are not necessarily associated to any tokens in the given prompt and the order of the self-attention maps cannot be compared between different inversions, as such they cannot be picked statically (see Figure 8b).

The issue of currently only being able to generate very simple road systems could potentially be resolved by using a diffusion model specialized towards the task of generating satellite images of roads. In addition this could improve the quality of the attention maps produced by using null-text inversion.

VI. Summary

We explored various methods to acquire more data to train a standard U-Net for road segmentation. This included a novel approach utilizing an inpainting model. Moreover, we experimented with a modified U-Net network which incorporates attention gates on the skip connections. The proposed inpainting method demonstrated a 2.9% improvement over the model trained with standard data set and a 1% improvement over the use of an additional data set not associated with the competition. However, the inpainting technique does not perform significantly better than basic data augmentation. We also showed that the inpainting and plain masking achieve similar results, both reaching a patch accuracy of 90%, meaning plain masking provides an efficient alternative to inpainting.

References

[1] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. Albahri, B. Al-dabbagh, M. Fadhel, M. Manoufali, J. Zhang, A. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, and Y. Gu, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, 04 2023.

[2] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, p. 60, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[3] Y. Zou, J. Choi, Q. Wang, and J. Huang, "Learning representational invariances for data-efficient action recognition," *CoRR*, vol. abs/2103.16565, 2021. [Online]. Available: https://arxiv.org/abs/2103.16565

[4] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," 2023.

[5] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *CoRR*, vol. abs/1708.04896, 2017. [Online]. Available: http://arxiv.org/abs/1708.04896

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[7] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *CoRR*, vol. abs/1712.04621, 2017. [Online]. Available: http://arxiv.org/abs/1712.04621

[8] S. D. Wickramaratne and M. Mahmud, "Conditional-gan based data augmentation for deep learning task classifier improvement using fnirs data," *Frontiers in Big Data*, vol. 4, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2021.659146

[9] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518308430

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[11] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. V. Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," *CoRR*, vol. abs/2201.09865, 2022. [Online]. Available: https://arxiv.org/abs/2201.09865

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[13] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[15] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[16] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation," in *MICCAI workshop MLMI*, Strasbourg (virtuel), France, Sep. 2021. [Online]. Available: https://hal.science/hal-03337089

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[18] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

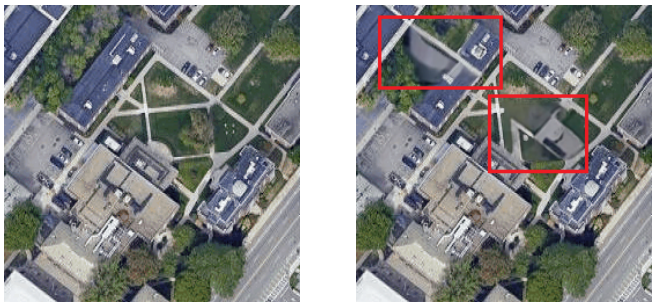
[19] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022.

[20] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” 2022.

APPENDIX A.
ADDITIONAL FIGURES

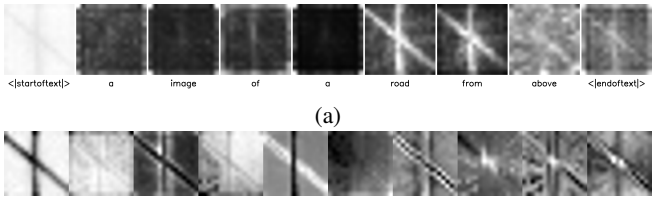
(a) (b) (c)

Figure 5: a) A cropped image from our data set. b) The same image with a mask consisting of stripes that have a width of four pixels each applied to it. The stripes are also placed four pixels apart from each other so that in total one half of the image is masked. c) The inpainted image obtained by filling in the masked regions using RePaint, which closely resembles the original image.



(a) (b)

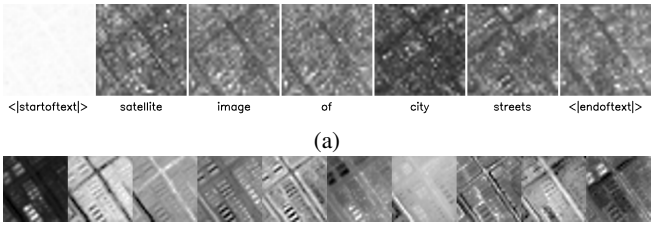
Figure 6: a) The original image. b) The inpainted image. The red areas highlight the regions that underwent inpainting, resulting in an addition of a road and a removal of another.



(a)

(b)

Figure 7: a) Cross-attention maps, each map is associated with a token of the prompt ”a image of a road from above”. b) Self-attention maps of the sentence ”a image of a road from above”.



(a)

(b)

Figure 8: a) Cross-attention maps of the null-text inversion with prompt ”satellite image of city streets”. b) Self-attention maps of the null-text inversion with prompt ”satellite image of city streets”.