DataEdit: Dataset editor using diffusion inpainting

Uliana Izmesteva, Dmitry Yudin



Рис. 1: An overview of a multi-model diffusion pipeline of data augmentation with visual and segmentation semantics. First, the synthetic depth is computed from the input image. Based on this depth, a scene image is generated. The object size in the input target generation mask is calculated using the depth as a resizing coefficient. Finally, the generated object is then added to the obtained scene.

Аннотация—With the development of diffusion models, there are unique opportunities for the application of innovative methods, in particular for the generation of synthetic training datasets. However, there are a range of difficulties associated with computational resources and time when automating the augmentation of large amounts of data. We propose a new approach to generate training data in a similar domain as existing datasets. The proposed method is based on the Stable Diffusion v2, which has been further trained using a new loss function on a specially assembled TAOMR dataset in the street scene domain. Also we implement the automatic pipeline for generating fully synthetic datasets including the Depth Anything v2 synthetic depth model. The code for our method is publicly available: https://github.com/Areson251/diff_depth_new

Index Terms—inpainting diffusion models, object adding, synthesis data, multimodal data, neural network

I. Introduction

(see Figure 1 for details).

Modern machine learning methods require significant amounts of high-quality data to effectively train models. However, the collection, cleaning and partitioning of real data is often costly, both in time and money, and is also subject to human error. In this context, the task of automated expansion of training samples at the expense of synthetic data is of particular relevance.

The key reasons for the development of an automatic data augmentation tool are: the need to increase the robustness of models through diverse and balanced training samples; solving problems under conditions of limited data volume typical of highly specialised scenarios; and reducing the cost of data collection and partitioning by using generative models, which also reduces the influence of the human factor.

Our contribution may be summarized as follow:

- We introduce new loss for fine-tuning diffusion models via LoRA [1] to generate objects.
- We develop a modular neural network method (see Figure 1 for details) that generates a new dataset based on reference data.
- We benchmark synthesys depth models.
- We extend the InstSnowRoadDetection [2] dataset by our pipeline.

II. Related Work

In this section, we provide a brief overview of the related work in generative and depth estimation models, using image data.

A. Augmention methods

PowerPaint

Affordance-Aware Object Insertion via Mask-Aware Dual Diffusion

Add-it: Training-Free Object Insertion in Images With Pretrained Diffusion Models

Magic Insert: Style-Aware Drag-and-Drop

Flux1.1

InstructPix2Pix

B. Diffusion models

backbones SDXL, Kandinsky [3], Dall-E 2 [4] хорошие, но слишком тяжелые.

Stable Diffusion v1.5-v2 [5] больше всего подходят.

tuning (???) Textual Inversion, DreamBoth, LoRA

C. Depth estimation models

DepthAnything v1, v2 MaryGold DPT DepthLab !!!

D. Segmentation models

YOLO v8 as baseline. SAM

III. Methodology

In this section we introduce methodology of each part of our pipeline.

.

Uliana Izmesteva are with the Intelligent Transport Laboratory, Moscow Institute of Physics and Technology, Moscow Region, Dolgoprudny, 141700, Russia.

Dmitry Yudin is with Artificial Intelligence Research Institute (AIRI), Moscow, Russia and the Intelligent Transport Laboratory, Moscow Institute of Physics and Technology, Moscow Region, Dolgoprudny, 141700, Russia, e-mail: yudin@airi.net.

A. Loss function

In this paper, the Stable Diffusion v2 diffusion model has been further trained with a new loss. The idea is as follows: we duplicate and freeze the original Unet weights, whose output will be ϵ_{empt} noise. Then, using a special prompt, we ask the model to remove the object from the image, thereby obtaining an "empty"image and counting the inverse difference between the ϵ_θ noise predicted by the trained LoRA weights and the ϵ_{empt} noise predicted by the frozen weights. This is the second term in the total error. By this way, we not only encourage the model to generate objects, but also penalise it when it tries to generate a "emptiness".

$$L_{new} := \alpha \|\epsilon_{obj} - \epsilon_\theta(z_t, t, F_\theta(\hat{c}))\|^2 + \frac{\beta}{\|\epsilon_{empt} - \epsilon_\theta(z_t, t, F_\theta(\hat{c}))\|^2 + \gamma} \tag{3}$$

where:

- ϵ_{obj} - GT noise;
- ϵ_{empt} - noise of a synthetically generated emptiness;
- $F_\theta(\hat{c})$ - model output;
- α and β - weighting coefficients;
- γ - division by zero elimination processing.

B. Scene generation pipeline

Controlnet - Depth Version [6] is used to generate scenes. Synthetic depth model input receives the original image. Then the calculated synthetic depth is fed to the Controlnet input and a new scene image is generated. The special feature is that the original markup of the dataset is preserved, thus no additional time is required.

A comparative analysis of the depth models is presented in Table Section IV-D. Metrics Equation (1) and Equation (2) are used here. According to its results, the Depth Anything 2 model was selected.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2} \tag{1}$$

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|z_i - \hat{z}_i|}{z_i} \tag{2}$$

where z_i — grounded true depth, $\hat{z}_i$ — predicted depth.

C. Object generation pipeline

Often it is necessary to generate rather small objects and the usual weights of diffusion models do not handle this task well. We propose the following pipeline for generating objects Figure 2: randomly select the coordinates of the object centre from the initial mask, where the object is supposed to be located (in our case, a road). Choose one of the template masks and resize it taking into account the depth factor. We will use the same synthetic depth that we used to generate the synthetic scene.

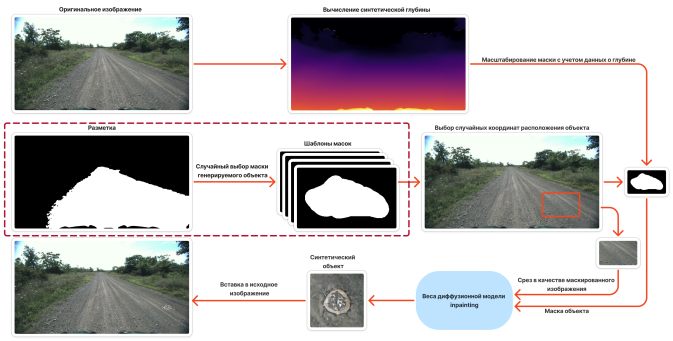

Рис. 2: An overview of object generation pipeline. Feature work (the highlighted in red area): replacing the selection of a random mask and object coordinates with a trained element

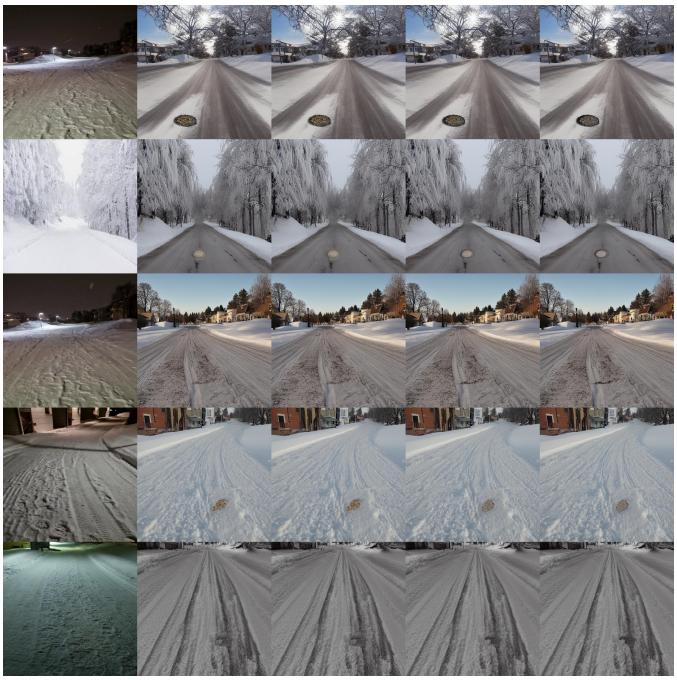

Рис. 3: An overview of full synthesis dataset and generated objects.

Figure Figure 5 shows examples of pothole object generation on trained Stable Diffusion weights using the proposed loss and pothole object generation pipelines.

D. Common pipeline

Examples of generating a fully synthetic dataset are shown in Figure Figure 7. The first column contains the original image. The next columns contain images for $guidencescale = [0.7, 1, 2, 3]$ respectively. For each original image from the dataset, one synthetic image with one object (pothole) is generated. A total of 842 images are obtained from train sampling, 38 images in validation and 24 images in test. All experiments were performed on an two NVIDIA Tesla V100 32GB.

1. Добавить глубину в датасет по итоговой картинке с ямой
2. Добавить глубину в датасет по исходной картинке + кроп глубины на генерации + poison blend (посчитать метрики с гт, мб на Чехове - нужна реальная глубина с камеры)
3. Добавить метрики генерации
4. Сравнить с другими loss функциями
5. Проверить качество датасета
6. Протестировать на моделях сегментации

Варианты научной новизны:

- Учить лора с глубиной (???) подумать как реализовать

IV. Experiments

A. Training details

Image Figure 4 shows a plot of the loss function for training the Stable Diffusion v2 model using LoRA. The model was trained for 10 thousand iterations with Learning rate equal to $1e-06$, train batch size equal to 4 and gradient accumulation steps equal to 32.

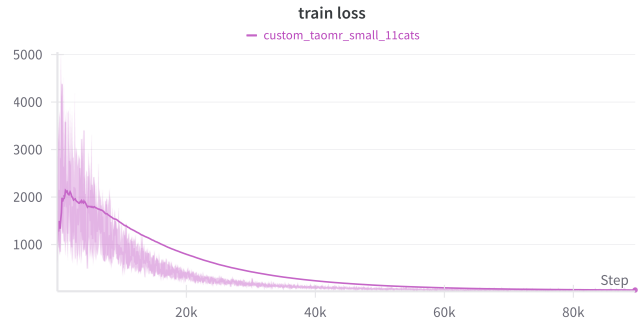

Рис. 4: Training new custom loss Stable Diffusion v2

All experiments were performed on an two NVIDIA Tesla V100 32GB.

B. Evaluation metrics

C. Object generation

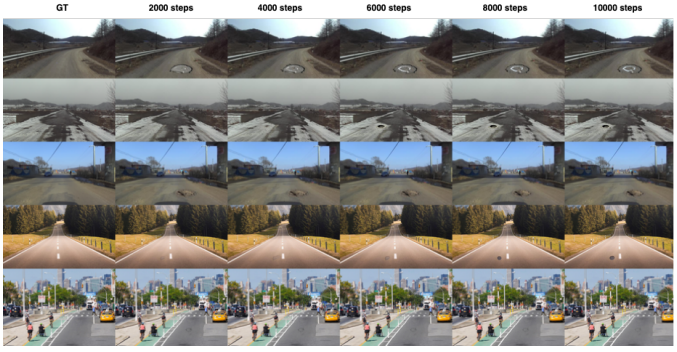

Рис. 5: Experimental results using Stable Diffusion 2 and image crops including object masks

Image Figure 6 shows qualitative results for generating several different objects on the weights trained by New Loss.

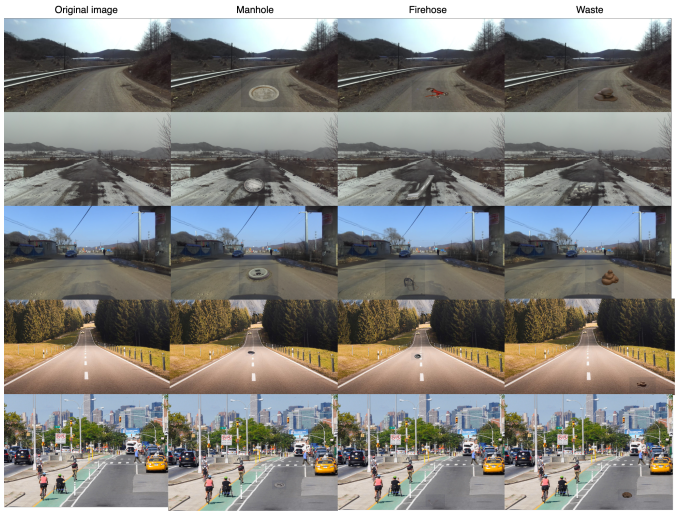

Рис. 6: Experimental results using Stable Diffusion 2 for several objects of TAOMR dataset

D. Depth estimation

Model	T↓	R↓	A↓
Depth Anything	0.04	9.11	7.15
Depth Anything 2	0.04	8.96	7.14
MariGold	0.43	10.49	9.78
DPT	0.07	10.38	6.94

E. Synthesis dataset

Examples of generating a fully synthetic dataset are shown in Figure Figure 7. The first column contains the original image. The next columns contain images for $guidencescale = [0.7, 1, 2, 3]$ respectively.

For each original image from the dataset, one synthetic image with one object (pit) is generated. A total of 842 images are obtained from train sampling, 38 images in validation and 24 images in test.

V. Conclusion

Our research has contributed to the field of object adding by introducing a modular pipeline named ***. The proposed method showed good qualitative results.

Limitations:

Acknowledgment

Список литературы

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen et al., "Lora: Low-rank adaptation of large language models." ICLR, vol. 1, no. 2, p. 3, 2022. 1

[2] DEBIGP, "Instsnowroaddetection dataset," , dec 2023, visited on 2025-04-07. 1

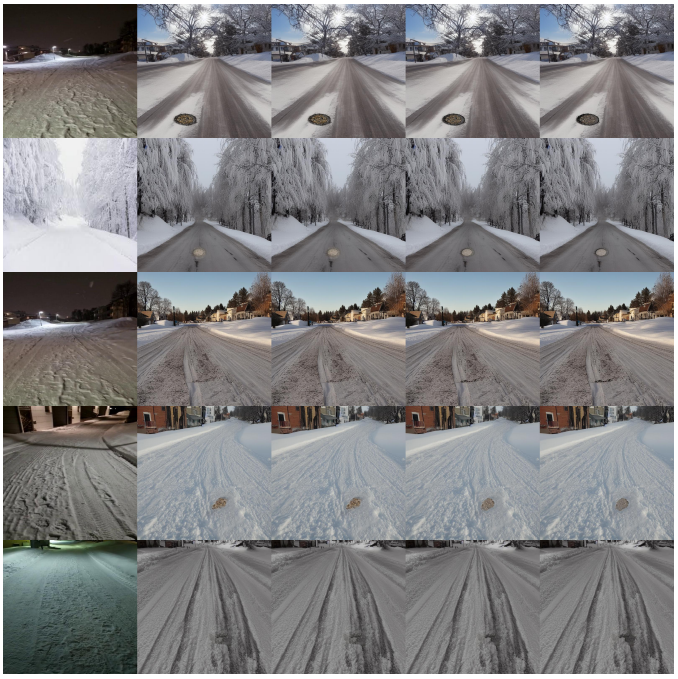

Рис. 7: An overview of full synthesis dataset and generated objects.

[3] A. Razzhigaev, A. Shakhmatov, A. Maltseva, V. Arkhipkin, I. Pavlov, I. Ryabov, A. Kuts, A. Panchenko, A. Kuznetsov, and D. Dimitrov, "Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion," arXiv preprint arXiv:2310.03502, 2023. 1

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, vol. 1, no. 2, p. 3, 2022. 1

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10 684–10 695. 1

[6] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. 2



Uliana A. Izmesteva received the bachelor's degree with a specialty "software engineering"with Kalashnikov Izhevsk State Technical Univer- sity, Izhevsk, Russia, in 2023. Since August 2023, she has been an research engineer with the Center of Cognitive Modeling, Moscow Institute of Physics and Technology, Moscow, Russia. Her research interests include generative models,computer vision adn deep learning.



Dmitry A. Yudin received the engineering diploma in automation of technological processes and production in 2010 and the Ph.D. degree in computer science from the Belgorod State Technological University (BSTU) named after V.G. Shukhov, Belgorod, Russia in 2014.

From 2009 to 2019, he was a Researcher and Assistant Professor with Technical Cybernetics Department at BSTU n.a. V.G. Shukhov. Since 2019, he has been the head of the Intelligent Transport Laboratory at the Moscow Institute of Physics and Technology, Moscow, Russia. Since 2021, he has been a Senior and Leading Researcher at AIRI (Artificial Intelligence Research Institute), Moscow, Russia. His research interests include computer vision, deep learning, and robotics.

VI. Ablation Study

Some details about our work.

A. Scene generation



Рис. 8: Результаты экспериментов с использованием весов Depth Anything V2 на датасете INSTSnowRoadDetection

B. Object generation

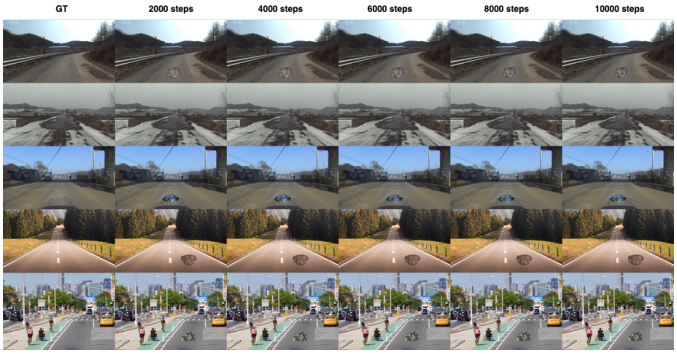

Рис. 9: Результаты экспериментов с использованием весов PowerPaint и кропов изображений, включающих маски объекта

На рисунках Figure 9, Figure 9, Figure 9, Figure 9 представлены результаты работы пайпллайна с генерацией одного объекта (pothole) на тестовой выборке датасета INSTSnowRoadDetection.



