Data Augmentation in Diffusion Inversion Space

Junfeng Wei[1] Rongsen Luo[1] Ziming Cheng[1] An Mo[1] Chao Ji[1,2]
[1]LindenBot [2]University of Science and Technology of China

Abstract

Visual imitation learning methods have demonstrated strong performance and potential, but their generalization ability to unseen environments remains limited. Although data augmentation offers an effective solution to this problem, current approaches depend on complex preprocessing procedures, require substantial hardware resources, are time-consuming, and struggle to comprehensively account for all possible environments. Our goal is to develop a data augmentation method that is simple, efficient, plug-and-play, and incurs no additional computational overhead. Our core idea is that, instead of performing data augmentation in the raw image space, conducting it in the diffusion inversion space can significantly simplify the augmentation process — to the extent that inserting simple geometric shapes is sufficient to achieve broader coverage of environmental variations. We designed a simple industrial-style scenario experiment to preliminarily validate our idea.

1. Introduction

Visual imitation learning has demonstrated strong capabilities and potential in the field of embodied intelligence [12][2]. However, its performance is highly sensitive to environmental variations, often showing good results only in specific training environments and struggling to generalize to unseen scenarios [5].

Data augmentation methods offer an effective solution to the issue of overfitting to a single environment [11]. However, existing augmentation techniques typically rely on complex image preprocessing pipelines [10][1], which place high demands on hardware resources, require considerable processing time, and still fail to comprehensively cover the full spectrum of possible environmental changes.

In contrast to the complex data augmentation methods used in visual imitation learning, augmentation techniques in purely visual tasks such as image classification [3] are significantly simpler. Effective augmented data can be obtained through basic operations like rotation, cropping, deformation, and even masking [13][4], without the need for elaborate preprocessing procedures.

We believe that the ideal data augmentation approach for visual imitation learning should mirror the simplicity and efficiency of augmentation techniques used in visual tasks.

The emergence of diffusion models [2][9] has provided significant inspiration for our work. In applications such as AI-generated art [8], users can simply sketch basic geometric shapes, and the diffusion model can generate multiple high-quality images related to those shapes.

This implies that, due to the generalization capability of diffusion models, simple geometric shapes in the diffusion inversion space can serve as a shared representation for the diverse set of images that may be generated through the reverse process. Based on this insight, we propose the following hypothesis: performing data augmentation directly in the raw image space requires extensive and complex preprocessing to generate realistic variations, yet still struggles to capture the full spectrum of environmental changes [7]. In contrast, data augmentation in the diffusion inversion space significantly reduces procedural complexity. Simple operations, such as inserting geometric shapes, can exploit the diffusion model's inherent generalization to account for a wide range of environmental variations.

Based on this hypothesis, we propose data augmentation in diffusion inversion space (DADIS), a simple yet effective data augmentation strategy for visual imitation learning. We conducted preliminary experiments by inserting ellipses of different colors, shapes, sizes and opacities in the inversion space for data augmentation, and designed a simple part-sorting task to initially validate our idea.

2. Method

2.1. The inversion of diffusion models

Similar to Stem-OB [5], we adopt the DDPM inversion framework introduced in [6].

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\tilde{\boldsymbol{\epsilon}}_t \tag{1}$$

Here, $\boldsymbol{x}_0$ denotes the original image, $\boldsymbol{x}_t$ represents the image after t steps of inversion, $\tilde{\boldsymbol{\epsilon}}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is independently sampled Gaussian noise, $\bar{\alpha}_t$ represents the proportion of the original image that is retained. In DDPM [2], $\bar{\alpha}_t$ is computed from each step of the forward diffusion process

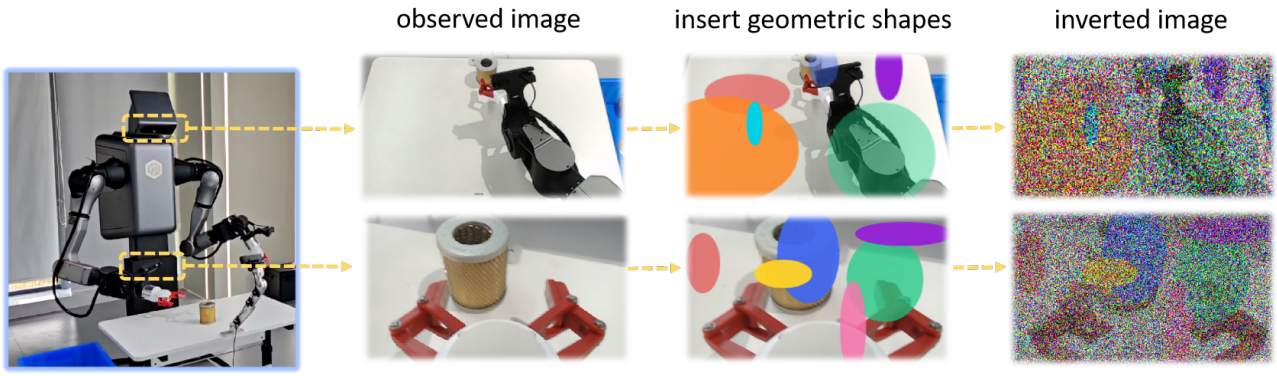

Figure 1. The process of performing data augmentation in the diffusion inversion space.

$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, where α_i denotes the scheduler parameter at step i in the forward diffusion process. However, in our method, since all images are inverted to the same number of steps, $\bar{\alpha}_t$ becomes a fixed constant.

2.2. Implementation

Training Stage: As shown in Figure 1, during the training stage, we first insert ellipses into the robot's observation images, with randomly sampled quantities, colors, brightness levels, sizes, and shapes.

$$\boldsymbol{o}_i^d = \boldsymbol{f}(\boldsymbol{o}_i) \tag{2}$$

$\boldsymbol{o}_i$ denote the original observation image of the robot, $\boldsymbol{f}(\cdot)$ denote the ellipse insertion function. Then $\boldsymbol{o}_i^d$ represents the image augmented by inserting ellipses. We then invert the image into the diffusion inversion space with equation1.

$$\hat{\boldsymbol{o}_i}^d = \sqrt{\alpha}\boldsymbol{o}_i^d + \sqrt{1-\alpha}\tilde{\boldsymbol{\epsilon}}_t \tag{3}$$

$\hat{\boldsymbol{o}_i}^d$ denote the inverted image, α denotes the specific value of $\bar{\alpha}_t$.

Testing Stage: During the testing stage, we only invert the image into the diffusion inversion space.

3. Experiments

We designed a simple part-sorting task, where the robot is required to pick up parts from the table and place them into a nearby blue box. Under identical environmental conditions, we collected 200 demonstration trajectories for training. For imitation learning, we adopted the ACT (Action Chunking with Transformers) model[12]. We evaluated the model's ability to generalize across different objects and background settings. During testing, we fixed several object placement positions, conducted 20 trials, and recorded the number of successful attempts. The results of the object generalization test are shown in Table 1. ori refers to the original part that is identical to the one used in the training data. $obj1$ is the same part with colorful stickers placed on some of its key features. $obj2$ is a water bottle with red packaging. $obj3$ is a blue cookie box.

	ori	$obj1$	$obj2$	$obj3$
w/o DADIS	19	8	2	0
DADIS	18	16	11	6

Table 1. Results on object generalization.

It is worth noting that the failures of the model using DADIS augmentation were all due to misaligned grasps after attempting to pick up the objects, whereas the ACT model without DADIS augmentation failed on $obj2$ and $obj3$ always due to a complete lack of grasping intention. The results of the background generalization test are shown in Table 2. $bg1$ refers to a tabletop covered with a tablecloth, while $bg2$ introduces additional clutter on the tablecloth-covered tabletop. The results demonstrate that DADIS exhibits no significant improvement in background generalization, particularly in $bg2$, where both methods are only capable of completing the task in regions distant from the interfering objects.

	ori	$bg1$	$bg2$
w/o DADIS	19	13	4
DADIS	18	15	4

Table 2. Results on background generalization.

This implies that even within the diffusion inversion space, simple graphical elements such as ellipses cannot adequately represent three-dimensional objects. However, they can effectively augment the color and texture of the target object, while enabling the model to better focus on the geometric structures of three-dimensional shapes.

References

[1] Ezra Ameperosa, Jeremy A Collins, Mrinal Jain, and Animesh Garg. Rocoda: Counterfactual data augmentation for data-efficient robot learning from demonstrations. *arXiv preprint arXiv:2411.16959*, 2024. 1

[2] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[5] Kaizhe Hu, Zihang Rui, Yao He, Yuyao Liu, Pu Hua, and Huazhe Xu. Stem-ob: Generalizable visual imitation learning with stem-like convergent observation through diffusion inversion. *arXiv preprint arXiv:2411.04919*, 2024. 1

[6] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024. 1

[7] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[10] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025. 1

[11] Chengbo Yuan, Suraj Joshi, Shaoting Zhu, Hang Su, Hang Zhao, and Yang Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025. 1

[12] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 1, 2

[13] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1