EFFECTIVE DATA AUGMENTATION WITH DIFFUSION MODELS

Brandon Trabucco
Carnegie Mellon University
brandon@btrabucco.com

Kyle Doherty
MPG Ranch & School of Forestry, Northern Arizona University

Max Gurinas
University of Chicago Laboratory Schools

Ruslan Salakhutdinov
Carnegie Mellon University
rsalakhu@cs.cmu.edu

ABSTRACT

Data augmentation is one of the most prevalent tools in deep learning, underpinning many recent advances, including those from classification, generative models, and representation learning. The standard approach to data augmentation combines simple transformations like rotations and flips to generate new images from existing ones. However, these new images lack diversity along key semantic axes present in the data. Consider the task of recognizing different animals. Current augmentations fail to produce diversity in task-relevant high-level semantic attributes like the species of the animal. We address the lack of diversity in data augmentation with image-to-image transformations parameterized by pre-trained text-to-image diffusion models. Our method edits images to change their semantics using an off-the-shelf diffusion model, and generalizes to novel visual concepts from a few labelled examples. We evaluate our approach on image classification tasks in a few-shot setting, and on a real-world weed recognition task, and observe an improvement in accuracy in tested domains.

1 INTRODUCTION

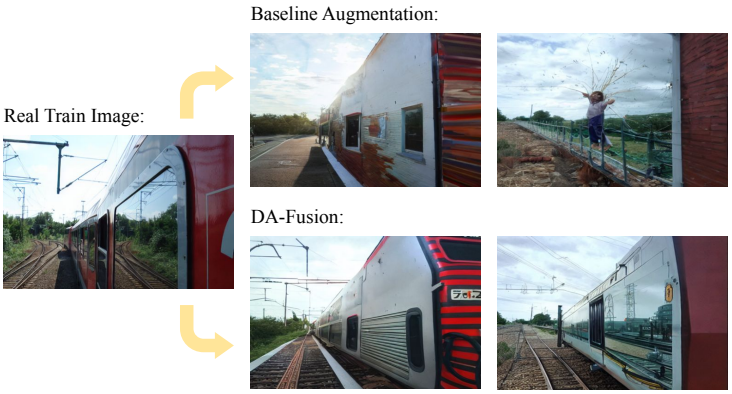


Figure 1: DA-Fusion improves generation quality. Given an image of a train, we generate augmentations using Real Guidance He et al. (2022) (top row), and compare these to our method (bottom row).

An omnipresent lesson in deep learning is the importance of internet-scale data, such as ImageNet Deng et al. (2009), JFT Sun et al. (2017), OpenImages Kuznetsova et al. (2018), and LAION-5B Schuhmann et al. (2022), which are driving advances in Foundation Models Bommasani et al. (2021) for image generation. These models use large deep neural networks Rombach et al. (2022) to synthesize photo-realistic images for a diverse landscape of prompts. Indeed, the recent success of large generative models prompts a question: can we augment visual recognition datasets with synthetic images from generative models? Answering this question promises to improve image recognition by generating large-scale image datasets from a handful of real images without human labelling effort. Generative models capture natural variations in appearance that standard data augmentation methods cannot.

Standard data augmentation aims to mitigate data scarcity by composing randomly parameterized image transformations Antoniou et al. (2017); Perez & Wang (2017); Shorten & Khoshgoftaar (2019); Zhao et al. (2020). Transformations including flips and rotations are chosen that respect

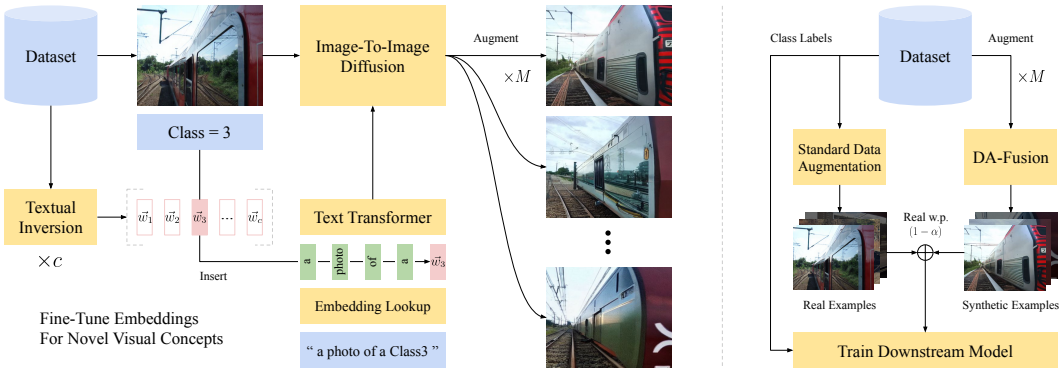


Figure 2: System architecture for DA-Fusion. Given a dataset of images, we use multi-class Textual Inversion Gal et al. (2022) to adapt the diffusion model to novel concepts.

basic invariances present in the data, such as horizontal reflection symmetry for a coffee mug. Robustness to this type of image transformation is captured well by existing methods, but models for recognizing coffee mugs should also be robust to subtle details of visual appearance like the brand of mug. Humans are exceptional at noticing these subtle details, able to distinguish varying brands of mugs from a single example. We aim to reproduce this efficiency when training deep neural networks by equipping data augmentation with large text-to-image diffusion models.

In this work, we propose a flexible data augmentation strategy that generates variations of real images using text-to-image diffusion models (DA-Fusion). Our method adapts the diffusion model to new domains by inserting and fine-tuning new tokens in the text encoder representing novel visual concepts. DA-Fusion modifies the appearance of foreground objects and backgrounds in a manner that respects object-level visual invariances, such as the design of the colors on the train in Figure 1. We test our method on three few-shot image classification tasks, including a real-world weed recognition task that lies outside the vocabulary of the diffusion model. Using the same hyper-parameters in all domains, our method outperforms prior work. DA-Fusion improves data augmentation by up to +10 percentage points, and ablations confirm that our method is robust to hyper-parameter assignment. Code for DA-Fusion will be released on acceptance.

2 Data Augmentation With Diffusion

Standard data augmentations apply to all images regardless of class and content Perez & Wang (2017). We aim to capture this flexibility with our diffusion-based augmentation. This is challenging because real images may contain elements the diffusion model is not able to generate out-of-the-box. How do we generate plausible augmentations for such images? We propose to address this shortcoming by fine-tuning new tokens in the text encoder of the diffusion model for each concept.

Adapting The Generative Model When generating synthetic images, previous work uses a prompt with the specified class name He et al. (2022). However, this is not possible for novel visual concepts that lie outside the vocabulary of the generative model. This is especially true in applied domains. We discuss this in Appendix B.1 with our contributed weed-recognition task, which our pretrained diffusion model is unable to initially generate a plausible image even when the class name is provided. We propose to address this shortcoming by inserting c new tokens into the vocabulary of the model—one for each class. We then perform Textual Inversion Gal et al. (2022) for each new token, initializing their embeddings $\vec{w}_i$ to a class-agnostic value (see Appendix H), and fine-tuning only these embeddings using the standard diffusion loss from Ho et al. (2020).

$$L_{\text{simple}}(\vec{w}_0, \vec{w}_1, \ldots, \vec{w}_c) = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\sqrt{\tilde{\alpha}_t}x_0 + \sqrt{1-\tilde{\alpha}_t}\epsilon, t)\|^2\right] \tag{1}$$

Generating Synthetic Images Many of the existing approaches generate synthetic images from scratch Antoniou et al. (2017); Tanaka & Aranha (2019); Besnier et al. (2020); Zhang et al. (2021b;a). However, novel concepts can be particularly challenging to generate from scratch when

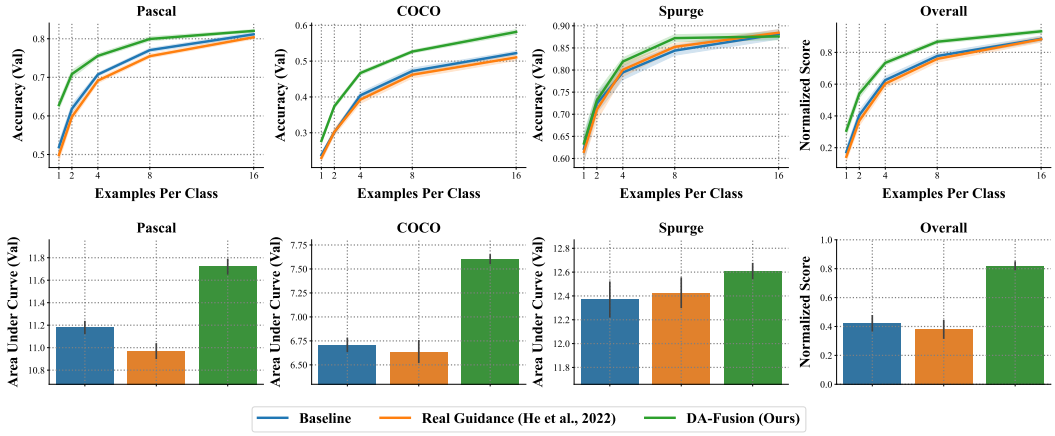


Figure 3: Few-shot performance. We evaluate DA-Fusion on three classification datasets and unilaterally outperform standard data augmentation in terms of accuracy and area under the curve.

only a handful of labelled examples are observed Gal et al. (2022). Rather than generate synthetic images from scratch, we use image-to-image transformations that splice real images into the reverse diffusion process following prior work in SDEdit Meng et al. (2022a). Given a reverse diffusion process with S steps, we insert a real image x_0^{ref} with noise $\epsilon \sim \mathcal{N}(0, I)$ at timestep $\lfloor St_0 \rfloor$, where $t_0 \in [0, 1]$ is a hyperparameter controlling the insertion position of the image.

$$x_{\lfloor St_0 \rfloor} = \sqrt{\tilde{\alpha}_{\lfloor St_0 \rfloor}} x_0^{\text{ref}} + \sqrt{1 - \tilde{\alpha}_{\lfloor St_0 \rfloor}} \epsilon \tag{2}$$

We proceed with reverse diffusion starting from the spliced image at timestep $\lfloor St_0 \rfloor$ and iterating Equation 5 to step 0. We discuss the interpretation and selection of the new hyperparameter t_0 in Section C. Generation is guided with a prompt that includes the fine-tuned token corresponding to the class of the spliced image (see Appendix H for details of the prompts used in this work).

3 Balancing Real & Synthetic Data

Training models on images from generative models often presents a trade-off between the diversity and size of the synthetic dataset, and the risk of over-emphasizing spurious qualities present in the synthetic data Antoniou et al. (2017). This is especially important considering that several recent papers have observed the benefit of curating orders of magnitude more synthetic data compared to the real data Tanaka & Aranha (2019); Besnier et al. (2020); Zhang et al. (2021b;a); He et al. (2022). The common solution assigns different probabilities to real and synthetic images He et al. (2022). We adopt a similar method for balancing real and synthetic data in Equation 3, where α denotes the probability that a synthetic image is present at the l-th location in the minibatch of images B.

$$B_{l+1} \leftarrow B_l \cup \left\{ X_i \;\; \text{w.p.} \;\; (1 - \alpha) \;\; \text{else} \;\; \tilde{X}_{ij} \right\} \tag{3}$$

Here $X \in \mathcal{R}^{N \times H \times W \times 3}$ denotes a dataset of N real images, and $i \in \mathbb{Z}$ specifies the index of a particular image X_i. For each image, we generate M augmentations, resulting in a synthetic dataset $\tilde{X} \in \mathcal{R}^{N \times M \times H \times W \times 3}$ with $N \times M$ image augmentations, where $\tilde{X}_{ij} \in \mathcal{R}^{H \times W \times 3}$ enumerates the jth augmentation for the ith image in the dataset. Indices i and j are sampled uniformly from the available N real images and their M augmented versions respectively. Given indices ij, with probability $(1 - \alpha)$ a real image image X_i is added to the batch B, otherwise its augmented image $\tilde{X}_{ij}$ is added. Hyper-parameter details are presented in Appendix H, and we find $\alpha = 0.5$ to work effectively in all domains tested, which equally balances real and synthetic images.

4 Evaluating Few-Shot Learning

While leafy spurge is confirmed to lie outside the vocabulary of the pretrained diffusion model, this is not true for Pascal Everingham et al. (2009) and COCO Lin et al. (2014), which have common

objects like boats and airplanes. To properly evaluate few-shot classification performance with these datasets, we emphasize the need to *delete* knowledge of these objects from the weights of the generative model. This remains an active area of research Meng et al. (2022b), so we hold concept deletion out-of-scope for this paper. Instead, we require that all generative models use a class agnostic prompt that does not contain the class name. This is important because generative models pre-trained at scale may have seen thousands of examples of common objects, and using a class name with an embedding trained on this large pool will not result in a proper few-shot setting. This evaluation protocol treats all classes as novel visual concepts that lie outside the vocabulary of the generative model, and we follow Section 2 for adapting the model to each class.

Experimental Details In this experiment, we test few-shot classification with three data augmentation strategies. The first, referred to as "Baseline", employs no synthetic images. This baseline implements a standard data augmentation strategy that uses random horizontal flips for COCO and Pascal, with additional random vertical flips for Spurge, with flip probabilities 0.5. The Real Guidance is based on He et al. (2022), and uses SDEdit on real images with $t_0 = 0.5$. Hyper-parameters shared between Real Guidance and our method have equal values to ensure fairness. Real Guidance is given the class agnostic prompt "a photo" whereas our method is prompted with "a photo of a ClassX" where ClassX represents a new token fine-tuned according to Section 2. Each real image is augmented M times, and a ResNet50 classifier pre-trained on ImageNet is fine-tuned on a mixture of real and synthetic images sampled as discussed in Section 3. Solid lines in plots represent the mean, and error bars denote the 68% confidence interval over 8 independent random trials.

Interpreting The Results Figure 3 shows results. We observe a consistent improvement by as much as +10 validation accuracy on Pascal and COCO when compared to standard data augmentation. DA-Fusion exceeds Real Guidance He et al. (2022) in all domains while utilizing the same hyperparameters, without any prior information about class names given. In this setting, Real Guidance performs comparably to the baseline, which suggests that gains in Real Guidance may stem from information provided by the class name. This experiment shows DA-Fusion improves few-shot learning and generalizes to out-of-vocabulary concepts. In the following sections, we ablate these results to understand how important each part of the method is to these gains in performance.

4.1 Robustness To The Mixing Ratio

We next conduct an ablation to understand the sensitivity of our method to the mixing parameter α that controls the sampling ratio of real and synthetic images. We chose this hyper-parameter as $\alpha = 0.5$ throughout this paper for simplicity as this value is at the center of the range for this hyper-parameter ($\alpha \in [0, 1]$). Insensitivity to the particular value of α is a desireable trait for any data augmentation method using generative models, as it simplifies hyper-parameter tuning. We test sensitivity to α by comparing runs of DA-Fusion with different alpha values. We report the gained accuracy over Real Guidance with the same α in Figure 4. These results show stability as α varies, and that $\alpha = 0.7$ performs marginally better than $\alpha = 0.5$, which suggests our method improves synthetic images quality because sampling them more often improves accuracy.

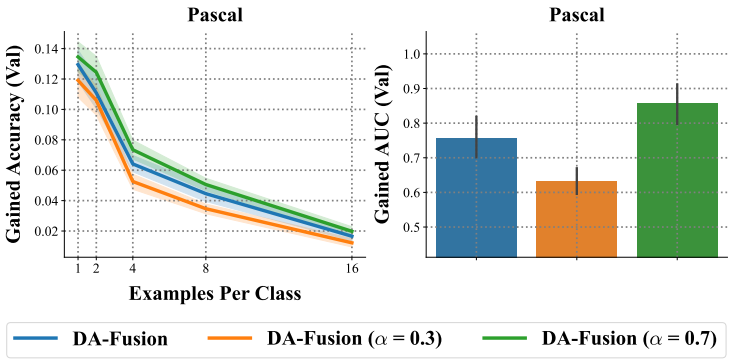


Figure 4: Ablation for mixing ratio. We run our method with three different mixing ratios $\alpha \in \{0.3, 0.5, 0.7\}$ and report the improvement in few-shot classification accuracy over a Real Guidance baseline using the same α. The plot shows our method is robust to this hyper-parameter and outperforms prior work for every α.

5 DISCUSSION

We proposed a flexible method for semantic data augmentation with diffusion models, DA-Fusion. Our method adapts a pretrained diffusion model to generate high quality augmentations for all images regardless of their content. Our method improves few-shot classification accuracy in all domains tested, and by up to +10 percentage points on the Pascal and COCO datasets. Similarly, our method produces gains on a contributed weed-recognition dataset that lies outside the vocabulary of the diffusion model. To understand these gains, we performed ablations that test the flexibility and robustness of our method. We first observe that few-shot performance is boosted when using multiple augmentations in parallel in Appendix D. Next, we test flexibility by selectively applying DA-Fusion to either foreground objects or backgrounds in Appendix F. We obtain higher robustness in the masked setting, suggesting our method remains effective when controlled with masks.

6 ACKNOWLEDGEMENTS

We thank MPG Ranch for supporting the leafy spurge component of this work. MPG Ranch staff, including Charles Casper, Erik Samsoe, Beau Larkin, and Philip Ramsey coordinated planning and acquisition of the leafy spurge imagery. In addition, we thank the effort of reviewers for helping to improve the paper, and the feedback from peers on intermediate drafts. We specifically thank Jing Yu Koh, Yutong He, Murtaza Dalal, So Yeon Min, and Martin Ma for their feedback. We thank Stability AI and Huggingface for providing open source models. Brandon Trabucco is supported by Amazon, and Ruslan Salakhutdinov is supported by ONR award N000141812861 and DSTA.

REFERENCES

Giuseppe Amatulli, Sami Domisch, Mao-Ning Tuanmu, Benoit Parmentier, Ajay Ranipeta, Jeremy Malczyk, and Walter Jetz. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data*, 5:180040, March 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.40. URL https://europepmc.org/articles/PMC5859920.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2017. URL https://arxiv.org/abs/1711.04340.

Romain Beaumont. Clip retrieval. https://github.com/rom1504/clip-retrieval, 2022.

Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: Training models from generated images. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 1–5. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053146. URL https://doi.org/10.1109/ICASSP40776.2020.9053146.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Pham Thanh Dat, Anuvabh Dutt, Denis Pellerin, and Georges Quénot. Classifier training from a generative model. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, 2019. doi: 10.1109/CBMI.2019.8877479.

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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Kyle D. Doherty, Marirose P. Kuhlman, Rebecca A. Durham, Philip W. Ramsey, and Daniel L. Mummey. Fine-grained topographic diversity data improve site prioritization outcomes for bees. *Ecological Indicators*, 132:108315, 2021. ISSN 1470-160X. doi: https://doi.org/10.1016/j.ecolind.2021.108315. URL https://www.sciencedirect.com/science/article/pii/S1470160X21009808.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–308, September 2009. URL https://www.microsoft.com/en-us/research/publication/the-pascal-visual-object-classes-voc-challenge/. Printed version publication date: June 2010.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Ayaan Haque. EC-GAN: low-sample classification using semi-supervised algorithms and gans (student abstract). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 15797–15798. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17895.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition?, 2022. URL https://arxiv.org/abs/2210.07574.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12,*

2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=qhAeZjs7dCL.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. URL http://arxiv.org/abs/1811.00982.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11451–11461. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01117. URL https://doi.org/10.1109/CVPR52688.2022.01117.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=-h6WAS6eE4.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. URL https://arxiv.org/abs/2211.09794.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017. URL https://arxiv.org/abs/1712.04621.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14837–14847, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Abstract.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In Munkhtsetseg Nandigjav, Niloy J. Mitra, and Aaron Hertzmann (eds.), *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pp. 15:1–15:10. ACM, 2022a. doi: 10.1145/3528233.3530757. URL https://doi.org/10.1145/3528233.3530757.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022b. URL https://arxiv.org/abs/2205.11487.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2256–2265. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 843–852. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.97. URL https://doi.org/10.1109/ICCV.2017.97.

Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans, 2019. URL https://arxiv.org/abs/1904.09135.

Toan Tran, Trung Pham, Gustavo Carneiro, Lyle J. Palmer, and Ian D. Reid. A bayesian data augmentation approach for learning deep models. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2797–2806, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/076023edc9187cf1ac1f1163470e479a-Abstract.html.

Sajila Wickramaratne and Md Shaad Mahmud. Conditional-gan based data augmentation for deep learning task classifier improvement using fnirs data. *Frontiers in Big Data*, 4:659146, 07 2021. doi: 10.3389/fdata.2021.659146.

Wei Xiong, Yutong He, Yixuan Zhang, Wenhan Luo, Lin Ma, and Jiebo Luo. Fine-grained image-to-image transformation towards visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Shin'ya Yamaguchi, Sekitoshi Kanai, and Takeharu Eda. Effective data augmentation with multi-domain learning gans. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 6566–6574. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6131.

Xiaohui Yang, Anne M. Smith, Robert S. Bourchier, Kim Hodge, and Dustin Ostrander. Flowering leafy spurge (euphorbia esula) detection using unmanned aerial vehicle imagery in biological control sites: Impacts of flight height, flight time and detection method. *Weed Technology*, 34(4): 575–588, 2020. doi: 10.1017/wet.2020.8.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=_SJ-_yyes8.

Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=yWkP7JuHX1.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 10145–10155. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.01001. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_DatasetGAN_Efficient_Labeled_Data_Factory_With_Minimal_Human_Effort_CVPR_2021_paper.html.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient GAN training. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55479c55ebd1efd3ff125f1337100388-Abstract.html.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 3774–3782. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.405. URL https://doi.org/10.1109/ICCV.2017.405.

A Related Work

Generative models have been the subject of growing interest and rapid advancement. Earlier methods, including VAEs Kingma & Welling (2014) and GANs Goodfellow et al. (2014), showed initial promise generating realistic images, and were scaled up in terms of resolution and sample quality Brock et al. (2019); Razavi et al. (2019). Despite the power of these methods, many recent successes in photorealistic image generation were the result of diffusion models Ho et al. (2020); Nichol & Dhariwal (2021); Saharia et al. (2022b); Nichol et al. (2022); Ramesh et al. (2022). Diffusion models have been shown to generate higher-quality samples compared to their GAN counterparts Dhariwal & Nichol (2021), and developments like classifier free guidance Ho & Salimans (2022) have made text-to-image generation possible. Recent emphasis has been on training these models with internet-scale datasets like LAION-5B Schuhmann et al. (2022). Resulting models Rombach et al. (2022); Saharia et al. (2022b); Nichol et al. (2022); Ramesh et al. (2022) have unlocked many application areas for generative models.

Image Editing One application area that diffusion has popularized makes edits to existing real images. Inpainting with diffusion is one such approach that allows the user to specify what to edit as a mask Saharia et al. (2022a); Lugmayr et al. (2022). Other works avoid masks and modify the attention weights of the diffusion process that generated the image instead Hertz et al. (2022); Mokady et al. (2022). Perhaps the most relevant technique to our work is SDEdit Meng et al. (2022a), where real images are inserted partway through the reverse diffusion process. SDEdit is applied by He et al. (2022) to generate synthetic data for training classifiers, but differs from our method in two key ways. Our method adapts the generative model to novel visual concepts and we consider augmentations to individual objects.

Synthetic Data Training neural networks on synthetic data from generative models was popularized using GANs Antoniou et al. (2017); Tran et al. (2017); Zheng et al. (2017). Various applications for synthetic data generated from GANs have been studied, including representation learning Jahanian et al. (2022), inverse graphics Zhang et al. (2021a), semantic segmentation Zhang et al. (2021b), and training classifiers Tanaka & Aranha (2019); Dat et al. (2019); Yamaguchi et al. (2020); Besnier et al. (2020); Xiong et al. (2020); Wickramaratne & Mahmud (2021); Haque (2021). More recently, synthetic data from diffusion models has also been studied in a few-shot setting He et al. (2022). These works position synthetic data from generative models as an additional dataset, while we develop a framework for generative models as a stackable transformation.

B Background

Diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020); Nichol & Dhariwal (2021); Song et al. (2021); Rombach et al. (2022) are sequential latent variable models inspired by thermodynamic diffusion Sohl-Dickstein et al. (2015). They generate samples via a Markov chain with learned Gaussian transitions, called the *reverse process*, starting from an initial noise distribution $p(x_T) = \mathcal{N}(x_T; 0, I)$.

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{4}$$

Transitions $p_\theta(x_{t-1}|x_t)$ are designed to gradually reduce variance according to a schedule $\beta_1, \ldots, \beta_T$ so the final sample x_0 represents a sample from the true distribution. Transitions are often parameterized by a fixed covariance $\Sigma_t = \beta_t I$ and a learned mean $\mu_\theta(x_t, t)$ defined below.

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \tilde{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \tag{5}$$

This parameterization choice results from deriving the optimal reverse process Ho et al. (2020), where $\epsilon_\theta(\cdot)$ is a neural network that is trained to process a noisy sample x_t to predict the noise added to real images by the *forward process*. Given real samples x_0 and noise $\epsilon \sim \mathcal{N}(0, I)$, the forward process can be sampled at an arbitrary timestep below.

$$x_t(x_0, \epsilon) = \sqrt{\tilde{\alpha}_t} x_0 + \sqrt{1 - \tilde{\alpha}_t} \epsilon \tag{6}$$

We borrow the notation introduced by Ho et al. (2020) to define $\alpha_t = 1 - \beta_t$ and $\tilde{\alpha}_t = \prod_{s=1}^{t} \alpha_t$. These components allow training and sampling from a diffusion model, which is used as the generative backbone in this work. In this work, we use a pretrained Stable Diffusion model trained by Rombach et al. (2022). Among other differences, this model includes a text encoder that enables text-to-image generation (refer to Appendix H for model details).

B.1 Data Preparation

Leafy Spurge We contribute a dataset of top-down drone images of semi-natural areas in the western United States. These data were gathered in an effort to better map the extent of a problematic invasive plant, leafy spurge (*Euphorbia esula*), that is a detriment to natural and agricultural ecosystems in temperate regions of North America. Prior drone-based work to detect leafy spurge achieved an accuracy of 0.75 Yang et al. (2020). To our knowledge, top-down aerial imagery of leafy spurge was not present in the Stable Diffusion training data. Results of CLIP-retrieval Beaumont (2022) returned close-up, side-on images of members of the same genus (Figure 5) in the top 20 results. We observed the first instance of our target species, *Euphorbia esula*, as a 35th result. Thus, the Spurge dataset represents a unique opportunity to explore few-shot learning setting, and state-of-the art classification outcomes would directly benefit efforts to restore natural ecosystems. Additional details about the Spurge dataset are in Appendix I.

Figure 5: A sample from the Spurge dataset (the first on the left), compared with top results of CLIP-retrieval queried on the prompt: "a drone image of leafy spurge". We note closeup images from members of the same genus (second, and third) in the top 20 results and a closeup of the same species for the 35th result (fourth).

PASCAL We leverage the 2012 version of the PASCAL Visual Object Classes challenge Everingham et al. (2009). This dataset contains 11,530 images and 6,929 object segmentation masks. We adapt this dataset into an object classification task by filtering images that have at least one object segmentation mask. We assign these images labels corresponding to the class of object with largest area in the image, as measured by the pixels contained in the mask. There are 20 classes in total using this methodology. We utilize the official training and validation sets for the 2012 challenge, and randomly select q images per class from the training set, which are used to measure few-shot classification accuracy.

COCO We process the 2017 version of the COCO dataset Lin et al. (2014) in a similar manner to PASCAL. This dataset contains 330K images with 1.5M object segmentation masks. As before, we filter images that have at least one object segmentation mask. We assign these images labels corresponding to the class of the largest object, measured by segmentation mask area. This dataset has 80 classes. We use the official training and validation sets for the 2017 dataset, and measure few-shot classification accuracy using the same methodology described for the PASCAL dataset.

C Stackable Augmentations

Having appropriately balanced the real and synthetic images, our goal becomes to maximize the augmentation diversity. This goal is shared with the standard data augmentation Perez & Wang (2017); Shorten & Khoshgoftaar (2019), where simple transformations are typically composed, yielding more sophisticated and diverse data. Despite the prevalence of composition for the data augmentation, methods using generative models treat their "augmentations" like a secondary dataset, not stackable transformations Antoniou et al. (2017); Tanaka & Aranha (2019); Yamaguchi et al. (2020); Zhang et al. (2021b;a); He et al. (2022). Inspired by the success of standard data augmentation, we propose stackable data augmentations based on generative models.

We define a sequence of augmentations $\mathcal{A}$ that consists of tuples of image transformations D_i and respective activation probabilities $p_i \in [0, 1]$ that collectively sum to one. This definition supports multiple options for selecting which augmentations to use, and a complete empirical study of them is an exciting future direction. In this work we opt for a simple approach: randomly sampling one augmentation D_a from $\mathcal{A}$ weighted by the probability p_a. Refer to Appendix H for hyperparameters associated with this stacking approach.

$$D_i : \mathcal{R}^{H \times W \times 3} \rightarrow \mathcal{R}^{H \times W \times 3} \tag{7}$$

$$\mathcal{A} = \big[(D_1, p_1), (D_2, p_2), \ldots, (D_k, p_k)\big] \tag{8}$$

To ensure stacking augmentations creates diversity in the samples, the image transformations D_i should be sufficiently unique. Several options are possible here, and we consider a first-principles heuristic that transforms increasingly higher-level features of the image as i increases. This heuristic ensures each D_i operates with a different granularity on the image. We accomplish this with SDEdit Meng et al. (2022a), and insert a noisy source image at a fraction $t_0 = i/k$ through the diffusion process to guide generation. Source images are taken from X, observed training images. Previous work in He et al. (2022) shows the effectiveness of SDEdit for creating synthetic images, and our work turns this methodology into a stackable data augmentation.

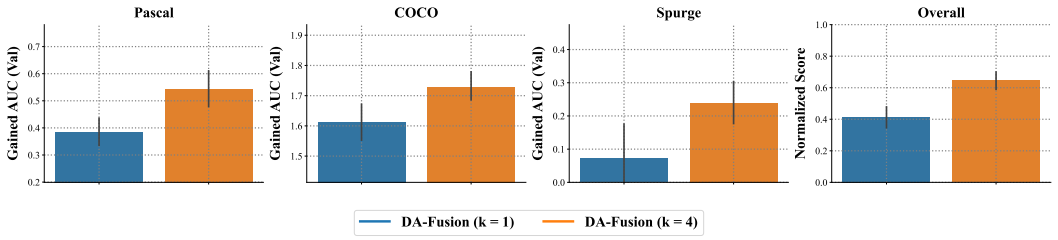


Figure 6: Ablation for stacking parameters. We evaluate DA-Fusion on three classification datasets and vary the number of augmentations in $\mathcal{A}$, from the setting used in the main experiments ($k = 4$), to only one ($k = 1$). We report the improvement in area under the curve (AUC) for few-shot classification accuracy, and observe a consistent improvement due to stacking in all domains.

D How Important Is Stacking?

Our goal in this section is to understand what fraction of gains are due to the stacking methodology of Section C. We employ the same experimental settings as in Section 4, and run an additional version of our method without stacking ($k = 1$) and with $t_0 = 0.5$, following the settings previously used with Real Guidance. In Figure 6 we report the improvement in area under the curve (AUC) versus standard data augmentation for our method. These results show that both versions of our method outperform the baseline, and stacking improves our method in all domains, leading to an overall improvement of 51%.

E Object-Centric Augmentations

Our approach thus far applies to all images, regardless of their class and content. However, real images often contain multiple classes with different visual invariances. Considering these visual differences, applying transformations at the object level has an appealing property. While traditional data augmentations violate this principle Perez & Wang (2017); Shorten & Khoshgoftaar (2019), ours permits independent transformations that separate objects and backgrounds. To accomplish this, we leverage inpainting Lugmayr et al. (2022); Saharia et al. (2022a). Given a pixelwise mask $v \in [0, 1]^{H \times W}$ specifying which image content to modify, we insert content into the diffusion process at locations where v_{ij} is close to one. In particular, after every timestep t in the reverse diffusion process, we assign the current sample x_t to a new value, where $\eta \sim \mathcal{N}(0, I)$ is unit Gaussian noise with the same cardinality as x_t.

$$x_t \leftarrow \big(1 - v\big) \circ x_t + v \circ \big(\sqrt{\tilde{\alpha}_t} x_0^{\text{ref}} + \sqrt{1 - \tilde{\alpha}_t}\eta\big) \tag{9}$$

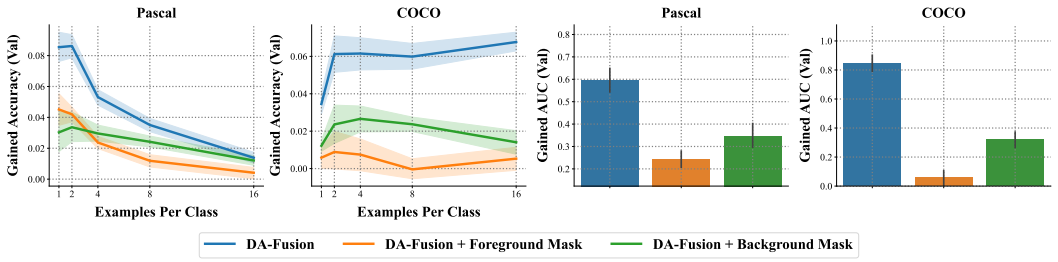


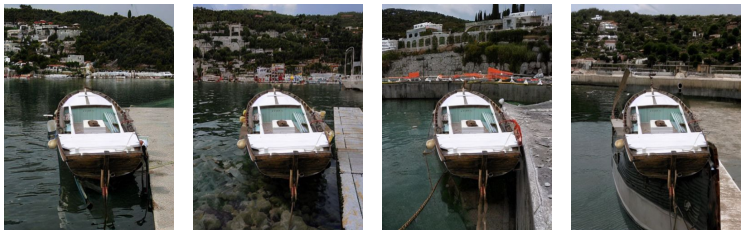
Figure 8: Ablation for masking. We evaluate our method when applied to foregrounds and backgrounds separately. We report the improvement in few-shot classification accuracy on a validation set for our method compared to a Real Guidance baseline using object segmentation masks from the Pascal and COCO datasets. Results show a consistent improvement when masks are used.

The source image to inpaint is given by x_0^{ref}, and the multiplication $\circ$ represents the elementwise product of tensors. Before splicing, x_t is taken from Equation 5 and represents standard reverse diffusion applied to the entire image. Equipped with our flexible data augmentation strategy that respects objects, stacks with other augmentations, and applies to all images regardless of class and content, we proceed with evaluation.

F Robustness To Object Masks

Our previous results show DA-Fusion is an effective data augmentation strategy when applied to whole images. However, real images often contain several objects with different visual invariances. For example, in a photo of a cat sitting on a table, we may aim to change only the breed of the cat or the brand of the table, which are defined independently. Augmentations based on generative models can achieve this behavior using inpainting Saharia et al. (2022a); Lugmayr et al. (2022), and stable performance when various masks are given is a desireable trait for any such method. In this section we test the robustness of DA-Fusion when various masks are used to constrain where image changes occur.

Augmenting The Background:



Augmenting The Foreground:

Figure 7: Foreground and background augmentations. Our method can be applied independently to each region of the image and maintains an improvement over prior work. In this example of the boat class from the PASCAL dataset, we augment the boat (foreground) separately from the water and hills (background).

Experimental Details We test robustness with two mask types: foreground object masks, and background masks. In the foreground type, a mask around the focal object in each image is used for inpainting with our pre-trained diffusion model. The mask is dilated by 16 pixels before use for inpainting to ensure the focal object is fully contained within. Background type masks are generated by inverting the foreground masks for each image. We test three versions of our method that augment the whole image, only the foreground, or just the background. Performance represents

the gain in few-shot classification accuracy over Real Guidance when training a classifier on samples from our masked augmentation. To ensure fairness, we omit stacking in this experiment and share all hyper-parameters with Real Guidance, including the mask used for inpainting, and the strength value $t_0 = 0.5$ from Section 4. As before, solid lines in plots represent the mean, and error bars denote the 68% confidence interval over 8 independent random trials.

Interpreting The Results Figure 8 shows that our method consistently outperforms prior work given masks for objects and backgrounds. Plots are shifted so that Real Guidance performance corresponds to a gained accuracy of 0, and positive gains represent an improvement over the corresponding masked Real Guidance baseline. Interestingly, performance gains are lower for masked augmentations than for whole images, suggesting there is room for improving DA-Fusion in the masked case. Additionally, the common ordering in the plots suggests that DA-Fusion is more effective at modifying backgrounds than objects. These results suggest our method has greater flexibility than prior work and produces effective synthetic images when directed with masks.

G Future Work

There are several directions to further improve the flexibility and performance of our method as future work. First, our method does not explicitly control *how* an image is augmented by the diffusion model. Extending the method with a mechanism to better control how objects in an image are modified, e.g. changing the breed of a cat, could improve the results. We explored this notion in Section F when masks are present; though, recent work in prompt-based image editing Hertz et al. (2022) suggests diffusion models can make localized edits without such masks. Second, data augmentation is becoming increasingly important in the decision-making setting Yarats et al. (2022). Maintaining temporal consistency is an important challenge faced when using our method in this setting. Solving this challenge could improve the few-shot generalization of policies in complex visual environments. Finally, improvements to diffusion models that enhance photo-realism when adapting the model to novel visual concepts are likely to improve sample quality and task performance.

H Hyperparameters

Our method inherits the hyperparameters of text-to-image diffusion models and SDEdit Meng et al. (2022a). In addition, we introduce several other hyperparameters in this work that control the diversity of the synthetic images. Specific values for these hyperparameters are given in Table 1.

I Leafy Spurge Dataset Acquisition and Pre-processing

In June 2022 botanists visited areas in western Montana, United States known to harbor leafy spurge and verified the presence or absence of the target plant at 39 sites. We selected sites that represented a range of elevation and solar input values as influenced by terrain. These environmental axes strongly drive variation in the structure and composition of vegetation Amatulli et al. (2018); Doherty et al. (2021). Thus, stratifying by these aspects of the environment allowed us to test the performance of classifiers when presented with a diversity of plants which could be confused with our target.

During surveys, each site was divided into a 3 x 3 grid of plots that were 10m on side (**Fig. 9**), and then botanists confirmed the presence or absence of leafy spurge within each grid cell. After surveying we flew a DJI Phantom 4 Pro at 50m above the center of each site and gathered still RGB images. All images were gathered on the same day in the afternoon with sunny lighting conditions.

We then cropped the the raw images to match the bounds of plots using visual markers installed during surveys as guides (**Fig. 10**). Resulting crops varied in size because of the complexity of terrain. E.G., ridges were closer to the drone sensor than valleys. Thus, image side lengths ranged from 533 to 1059 pixels. The mean side length was 717 and the mean spatial resolution, or ground sampling distance, of pixels was 1.4 cm.

In our initial hyperparameter search we found that the classification accuracy of plot-scale images was less than that of a classifier trained on smaller crops of the plots. Therefore, we generated four 250x250 pixel crops sharing a corner at plot centers for further experimentation (**Fig. 11**). Because

Hyperparameter Name	Value
Synthetic Probability α	0.5
Stacked Augmentations k	4
Activation Probabilities p_i	$1/k$
Synthetic Images Per Real M	10
Synthetic Images Per Real M (spurge)	50
Textual Inversion Token Initialization	"the"
Textual Inversion Batch Size	4
Textual Inversion Learning Rate	0.0005
Textual Inversion Training Steps	1000
Class Agnostic Prompt	"a photo"
Textual Inversion Prompt	"a photo of a ClassX"
Real Guidance Strength t_0	0.5
Stable Diffusion Checkpoint	`CompVis/stable-diffusion-v1-4`
Stable Diffusion Guidance Scale	7.5
Stable Diffusion Resolution	512
Stable Diffusion Denoising Steps	1000
Classifier Architecture	ResNet50
Classifier Learning Rate	0.0001
Classifier Batch Size	32
Classifier Training Steps	10000
Classifier Early Stopping Interval	200

Table 1: Hyperparameters and their values.

spurge plants were patchily distributed within a plot, a botanist reviewed each crop in the present class and removed cases in which cropping resulted in samples where target plants were not visually apparent.

J Benchmarking the Leafy Spurge Dataset

We benchmark classifier performance here on the full leafy spurge dataset, comparing a baseline approach incorporating legacy augmentations with our novel DA-fusion method. For 15 trials we generated random validation sets with 20 percent of the data, and fine-tuned a pretrained ResNet50 on the remaining 80 percent using the training hyperparameters reported in section **??** for 500 epochs. From these trials we compute cross-validated mean accuracy and 68 percent confidence intervals.

In the case of baseline experiments, we augment data by flipping vertically and horizontally, as well as randomly rotating by as much as 45 degrees with a probability of 0.5. For DA-Fusion augmentations we take two approaches(**Fig. 12**) The first we refer to as DA-Fusion Pooled, and we apply the methods of Textual Inversion Gal et al. (2022), but include all instances of a class in a single session of fine-tuning, generating one token per class. In the second approach we refer to as DA-Fusion Specific, we fine-tune and generate unique tokens for each image in the training set. In the specific case, we generated 90, 180, and 270 rotations as well as horizontal and vertical flips and contribute these along with original image for Stable Diffusion fine-tuning to achieve the target number of images suggested to maximize performanceGal et al. (2022). In both DA-Fusion approaches we generated ten synthetic images per real image for model training. We maintain $\alpha = 0.5$, evenly mixing real and synthetic data during training. We also maximize synthetic diversity by randomly selecting 0.25, 0.5, 0.75, and 1.0 t_0 values.

Both approaches to DA-Fusion offer slight performance enhancements over baseline augmentation methods for the full leafy spurge dataset. We observe a 1.0% gain when applying DA-Fusion Pooled and a 1.2% gain when applying DA-Fusion Specific(**Fig. 13**). It is important to note that, as implemented currently, compute time for DA-Fusion Specific is linearly related to data amount, but DA-Fusion Pooled compute is the same regardless of data size.

While pooling was not the most beneficial in this experiment, we support investigating it further. This is because fine-tuning a leafy spurge token in a pooled approach might help to orient our target

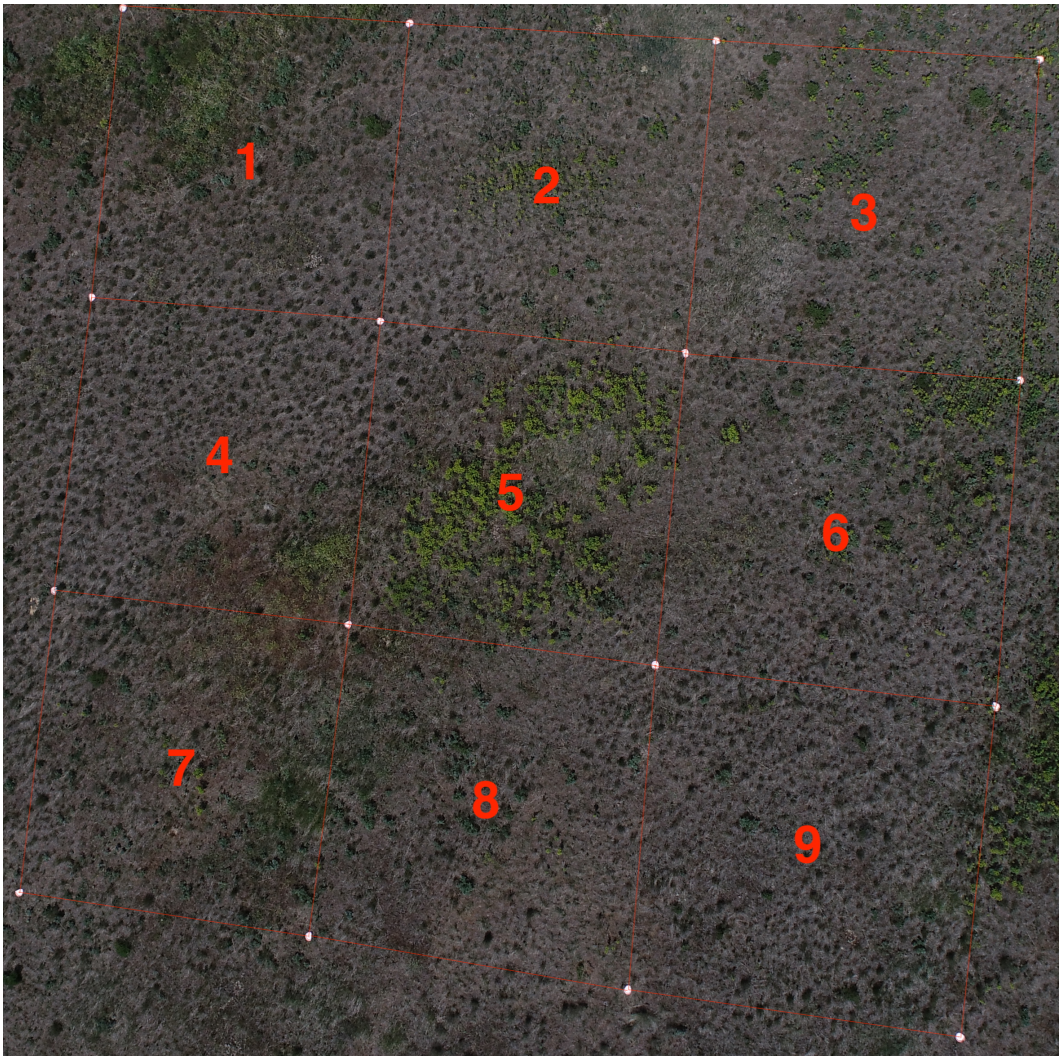


Figure 9: A drone image of surveyed areas containing leafy spurge. At each site botanists verified spurge presence or absence in a grid of nine spatially distinct plots. Note that cell five is rich in leafy spurge.

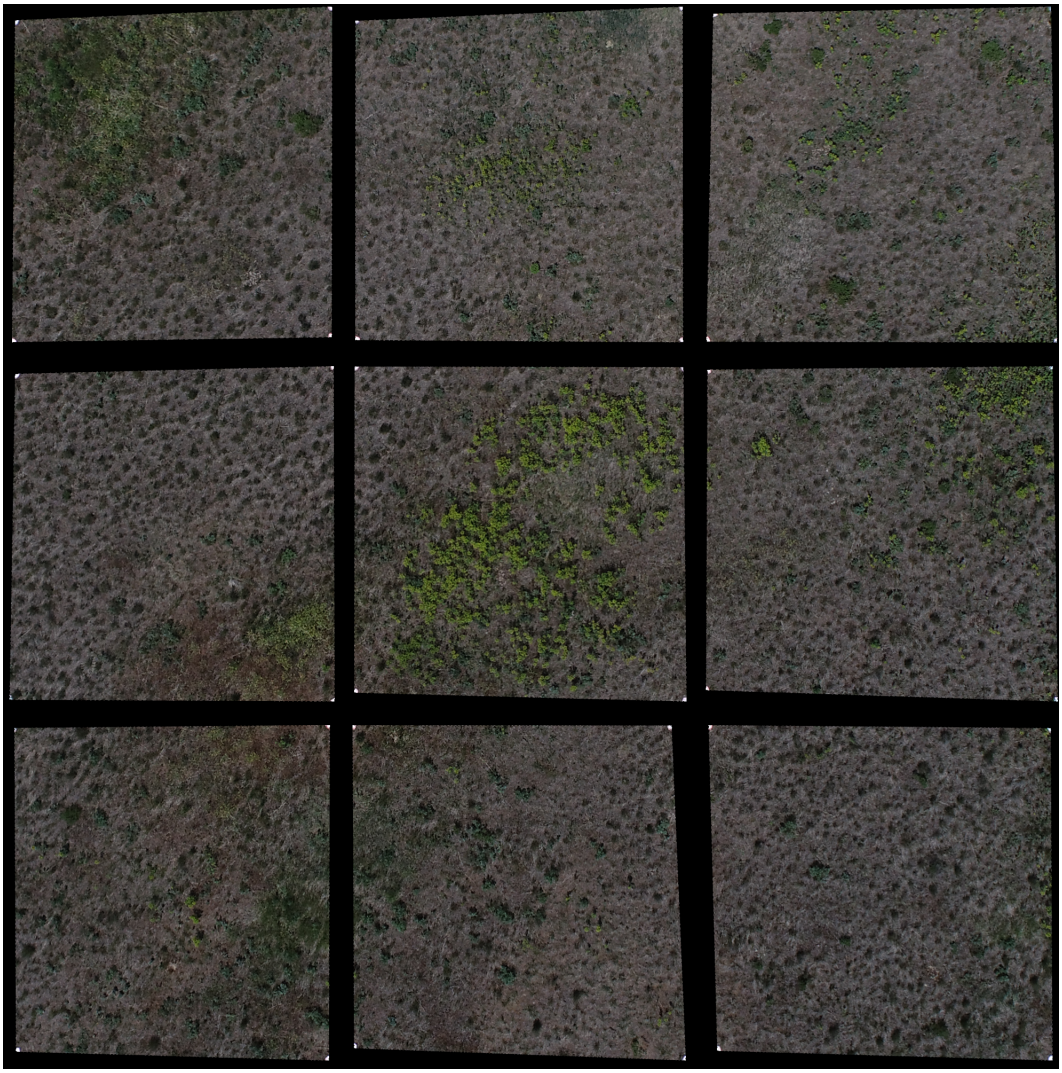


Figure 10: Markers installed at the corners of plots were used to crop plots from source images.

Figure 11: At each plot image center we cropped four 250x250 pixel sub-plots. We did this to amplify our data and improve classifier performance. The crops of plots with spurge present labels were inspected by a botanist to filter out examples where cropping excluded the target plant or the plants were not apparent.

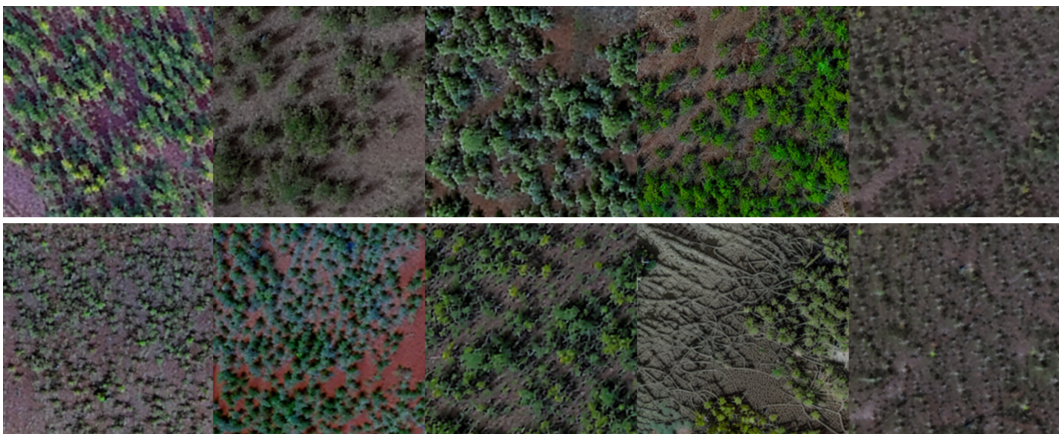


Figure 12: Here we show examples of synthetic images generated from the leafy spurge dataset with DA-Fusion methods. The top row shows output where images are pooled to fine-tune a single token per class. The bottom row shows examples where tokens are generated specifically for each image. Source images, inference hyperparameters, and seed are otherwise identical in each column.

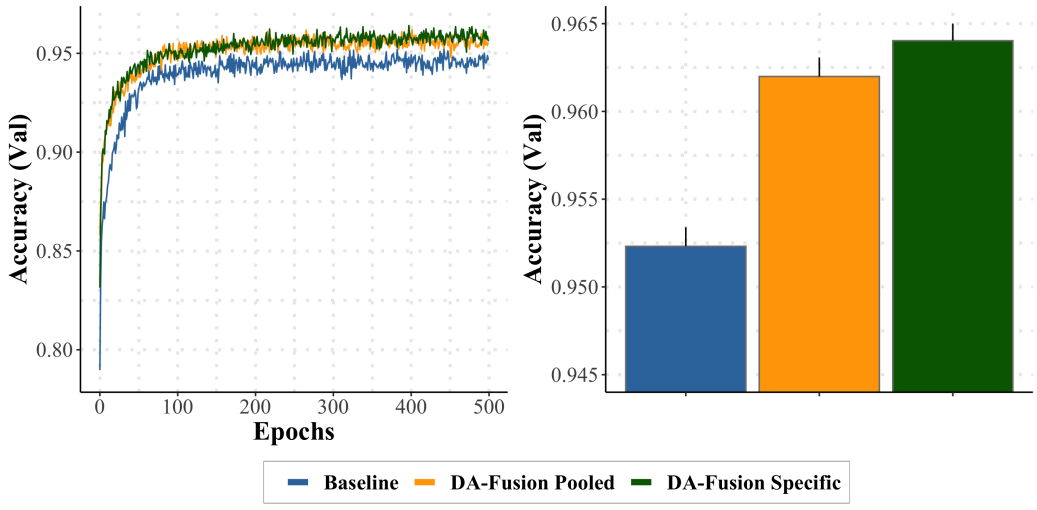


Figure 13: Cross-validated accuracy of leafy spurge classifiers when trained with baseline augmentations versus DA-Fusion methods on the full dataset. In addition to the benefits of DA-Fusion in few-shot contexts, we also find our method improves performance on larger datasets. Generating image-specific tokens (green line and bar) offers the most gains over baseline, though at the cost of greater compute.

in the embedding space where plants with similar diagnostic properties, such as flower shape and color from the same genus, may be well represented. However, the leafy-spurge negative cases do not correspond to a single semantic concept, but a plurality, such as green fields, brown fields, and wooded areas. It is unclear if fine-tuning a single token for negative cases by a pooled method would remove diversity from synthetic samples of spurge-free background landscapes, relative to an image-specific approach. For this reason, we suspect a hybrid approach of pooled token for the positive case and specific tokens for the negative cases could offer further gains, and support the application of detecting weed invasions into new areas.