Test-time Adaptation with Diffusion Models

Mihir Prabhudesai [1 *] **Tsung-Wei Ke** [1 *] **Alexander C. Li** [1] **Deepak Pathak** [1] **Katerina Fragkiadaki** [1]

Abstract

We find that generative models can be great test-time adapters for discriminative models. We propose a method to adapt pre-trained classifiers and large-scale CLIP models to individual unlabelled images by modulating the text conditioning of a text-conditional pretrained image diffusion model and maximizing the image likelihood using end-to-end backpropagation to the classifier parameters. We improve the classification accuracy of various pretrained classifiers on various datasets, including ImageNet and its variants. Further we show that our approach significantly outperforms previous test-time adaptation methods. To the best of our knowledge, this is the first work that adapts pre-trained large-scale discriminative models to individual images; all previous works require co-training under joint discriminative and self-supervised objectives, to apply at test time, which prevents them from adapting readily available models.

1. Introduction

Building predictive models, whether classifiers or regressors, is arguably the most fundamental problem in machine learning, and yet, whether such models be built in a generative or discriminative way remains an ever-lasting debate. Empirically, discriminative approaches have always had an upper hand in terms of generalization accuracy (He et al., 2016; Dosovitskiy et al., 2020). In the past few years, there has been a dramatic resurgence in the field of generative modeling, primarily fueled by the advent of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022) and the proliferation of large-scale datasets (Schuhmann et al., 2022; Sun et al., 2017). This has rekindled the interest in repurposing generative models for discriminative tasks. To date, current attempts to leverage visual generative models for discriminative tasks can be grouped into two kinds. First, methods that use Bayes rule to directly optimize the conditional probability from the joint distribution learned by the generative model (Grathwohl et al., 2019; Li et al., 2023a; Clark & Jaini, 2023). Such methods do not need finetuning and operate in a zero-shot manner. Second, methods that generate labeled images to augment the training or finetuning of discriminative models (Azizi et al., 2023; Yu et al., 2023; Li et al., 2022; Zhang et al., 2021). However, despite these attempts, pure discriminative methods still outperform pure generative methods across almost all popular benchmarks but it's not too surprising since generative models are aiming to solve a much more difficult problem.

In this paper, we take an alternative perspective. Instead of considering generative and discriminative models as competitive, we argue that they should be coupled in a way that leverages the best of both worlds: discriminative models are good at building powerful conditional density models but overfit to training distribution, and generative models generalize better but struggle to learn discriminative features. Instead of re-purposing generative models, we leverage them to adapt discriminative models at test time to data samples that are far from the training distribution.

We find that pre-trained generative diffusion models are great test-time adapters for pre-trained discriminative models. We propose to adapt open and closed vocabulary classifiers to individual unlabelled images by using their output to modulate the text conditioning of an image diffusion model and maximize the image diffusion likelihood. Our model, Diffusion-based Test Time Adaptation , is reminiscent of an encoder-decoder architecture, with a state-of-the-art pretrained classifier as the encoder and a state-of-the-art pretrained generative model in the decoder. At test time, the pretrained diffusion model provides guidance on how to update the pretrained classifier. We show that Diffusion-TTA effectively adapts image classifiers for both in- and out-of-distribution examples across established benchmarks, including ImageNet and its variants.

Generative models have previously been used for test time adaptation (TTA) of classifiers or segmentors, e.g., TTT-MAE (Gandelsman et al., 2022), Slot-TTA (Prabhudesai et al.), etc. However, these methods require the discrimina-

[*]Equal contribution [1]Carnegie Mellon University. Correspondence to: Mihir Prabhudesai <mprabhud@andrew.cmu.edu>.

Accepted to ICML workshop on Structured Probabilistic Inference & Generative ModelProceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

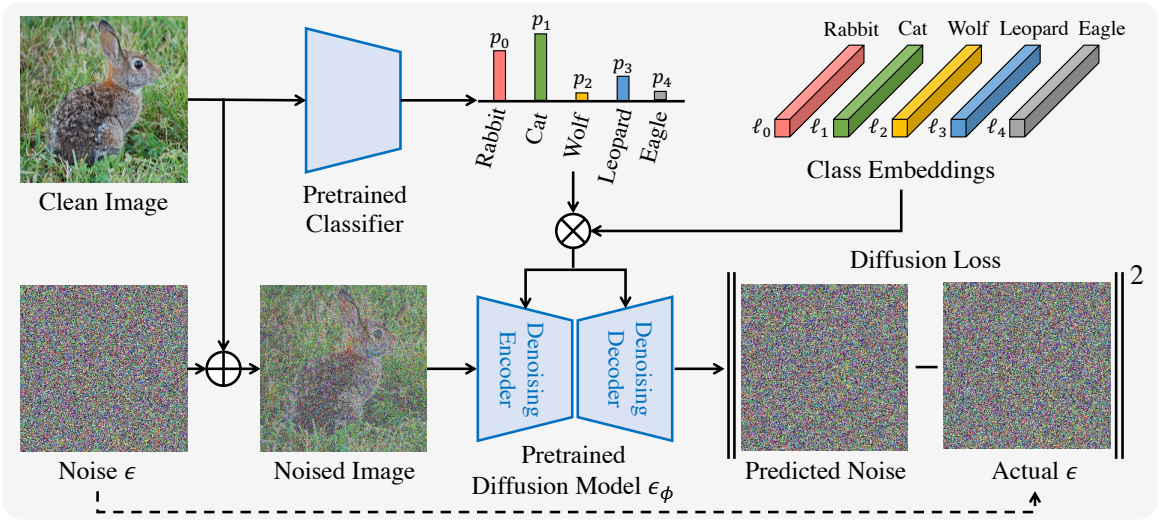

Figure 1: **Architecture of Diffusion-TTA.** Our method consists of discriminative and generative modules. The discriminative classifier uses the clean input image to predict a distribution over labels, which we use for a weighted sum of learnt class embeddings into a prompt. The generative diffusion model takes as input the noisy image and the prompt, and predicts the added noise. We update the classifier's weights to maximize the image likelihood using the diffusion loss.

tive models to be trained jointly with a self-supervised image generation loss. At test time, the discriminative model is adapted by finetuning with the generation loss, e.g. the masked autoencoder loss in TTT-MAE (Gandelsman et al., 2022), thus enhancing its task performance. While these models demonstrate substantial performance improvements upon adaptation, they often stem from the subpar performance of their initial feed-forward discriminative model. This trend can be clearly seen in our TTT-MAE baseline (Gandelsman et al., 2022), where the before-adaptation results are significantly lower than the results of a pre-trained feedforward classifier. In contrast, our approach refrains from any initial joint training, opting instead to directly adapt usual pre-trained discriminative models at test time utilizing pre-trained generative diffusion models.

We test our approach on multiple datasets including ImageNet (Deng et al., 2009), its out-of-distribution variants (C, R, A, v2), CIFAR100, Food101, FGVC, Oxford Pets, and Flowers102 and on the adaptation of numerous ImageNet-trained and CLIP-based classifiers, and under various text conditioned diffusion models, such as Stable Diffusion (Rombach et al., 2022) and DiT (Peebles & Xie, 2022). We show consistent improvements over the initially employed classifier. We hope our work to stimulate research on combining feed-forward and generative models, for more deliberate slow perception to handle images outside of the training distribution. Our code and trained models will be publicly available upon publication.

2. Method

In this section, we present Diffusion-TTA. We discuss relevant diffusion model preliminaries in Section 2.1 and describe our method in Section 2.2. Our Diffusion-TTA model is shown in Figure 1.

2.1. Diffusion Model Preliminaries

A diffusion model learns to model a probability distribution $p(x)$ by inverting a process that gradually adds noise to the image x. The diffusion process is associated with a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$, which defines how much noise is added at each time step. The noisy version of sample x at time t can then be written $x_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, is a sample from a Gaussian distribution (with the same dimensionality as x), $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. One then learns a denoising neural network $\hat{\epsilon} = \epsilon_\phi(x_t; t)$ that takes as input the noisy image x_t and the noise level t and tries to predict the noise component ϵ.

Diffusion models can be easily extended to draw samples from a distribution $p(x|\mathbf{c})$ conditioned on a prompt $\mathbf{c}$, where $\mathbf{c}$ can be an image category, class index, image caption, image semantic map, *etc* (Rombach et al., 2022; Li et al., 2023b; Zhang & Agrawala, 2023). Conditioning on the prompt can be done by adding $\mathbf{c}$ as an additional input of the network ϵ_ϕ. In this work, we focus on text-conditioned diffusion models. Modern text-conditioned image diffusion models are trained on large collections $\mathcal{D}' = \{(x^i, \mathbf{c}^i)\}_{i=1}^N$ of images paired with text prompts by minimizing the loss: $\mathcal{L}_{\text{diff}}(\phi; \mathcal{D}') =$

$\frac{1}{|\mathcal{D}'|}\sum_{x^i,\mathbf{c}^i\in\mathcal{D}'}||\epsilon_\phi(\sqrt{\bar{\alpha}_t}x^i+\sqrt{1-\bar{\alpha}_t}\epsilon,\mathbf{c}^i,t)-\epsilon||^2$.

This loss, which trains the network ϵ_ϕ to predict the noise ϵ added to an image, corresponds to a reweighted form of the variational lower bound for $\log p(x|\mathbf{c})$ (Ho et al., 2020).

2.2. Test-time Adaptation with Diffusion

Test-time adaptation (TTA) methods (Sun et al., 2020) aim to adapt a model at test time to account for unforeseen distribution shifts. We consider image classification as our task for test-time adaptation. Our core idea is to update the image classifier using a self-supervised objective–image likelihood, under the guidance of general representation encoded in powerful diffusion models.

To do so, we reformulate conditioning prompt $\mathbf{c}$ in the diffusion model to be dependent on the classifier's output. Let f_θ denote the classifier parameterized by θ, which takes input as image x and outputs classification logits $z=f_\theta(x)$, where $z=(z_1,z_2,...,z_L)$ be the logits predicted by the classifier, over all the L categories in the dataset. We observe that most classifiers maintain a high top-K accuracy, while their top-1 accuracy is significantly lower, this motivates to only adapt the top-K predictions of the classifier using the diffusion objective. We subselect the top-K categories $\mathcal{T}$ and do a softmax normalization of the logits z only over them. The probability of category $i\in\mathcal{T}$ is:

$p_i=\text{Softmax}(z_i)=\frac{e^{\mathbf{z}_i}}{\sum_{j\in\mathcal{T}}e^{\mathbf{z}_j}}$

Given the learnt class or text embeddings of the diffusion model for the top-K categories $\{\ell_i\}_{i\in T}$, we write the classifier conditioned text prompt as $\hat{c}=\sum_{i\in\mathcal{T}}p_i\ell_i$. As p is differentiable with respect to the classifier weights, we can now update the classifier weights θ via gradient descent by minimizing the following diffusion loss, after replacing $\mathbf{c}$ with $\hat{c}$.

$L(\theta,\phi)=\mathbb{E}_{t,\epsilon}\|\epsilon_\phi(\sqrt{\bar{\alpha}_t}x+\sqrt{1-\bar{\alpha}_t}\epsilon,\hat{c},t)-\epsilon\|^2$.

We minimize this loss over each example in the test-set independently, while sampling random timesteps t coupled with random noise variables ϵ.

For further implementation details, please refer to our Appendix Sections A.2 and A.3.

3. Experiments

We test Diffusion-TTA in adapting CLIP models (Radford et al., 2021) and ImageNet classifiers (He et al., 2016; Dosovitskiy et al., 2020; Liu et al., 2022) across multiple image classification datasets, at both in-distribution and out-of-distribution test images. CLIP models are trained across millions of image-caption pairs using contrastive matching of the language and image feature embeddings. ImageNet classifiers are all trained on the ImageNet classification task. We report the top-1 classification in our experiments. For further results and ablations of our design choices, please refer to Section A.1 in Appendix.

3.1. Test-time Adaptation of on Open-Vocabulary CLIP Classifiers

We test Diffusion-TTA for adapting three different CLIP models with different backbone sizes: ViT-B/32, ViT-B/16, and ViT-L/14. We follow (Radford et al., 2021) to perform zero-shot classification. We consider a variety of datasets for TTA of the CLIP classifiers, such as CIFAR-100 (Krizhevsky et al., 2009), Food101 (Bossard et al., 2014), Flowers102 (Nilsback & Zisserman, 2008), FGVC Airplane (Maji et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012) and ImageNet (Deng et al., 2009).

Baselines. We compare Diffusion-TTA against the following state-of-the-art TTA approaches:

- Diffusion Classifier (Li et al., 2023a) is a generative image classifier building atop diffusion models. This method searches over the set of discrete classes and predicts the class that maximizes the conditional image likelihood. We directly report the numbers from their paper in our table.
- Synthetic SD Data (Ravuri & Vinyals, 2019; Azizi et al., 2023) is a baseline reported in Diffusion Classifier. Using class names as prompts, they generate synthetic class-image data with the text-to-image capabilities of Stable Diffusion. Afterwards, they train a ResNet-50 classifier on the synthetic dataset.
- SD Features is a baseline inspired by Label-DDPM (Baranchuk et al., 2021) and reported in Diffusion Classifier, where they use the Stable Diffusion features and then train a ResNet-50 classifier supervised on top of the extracted features and ground truth labels.

We show quantitative zero-shot classification results in Table 1. We find that our method improves CLIP of different sizes, including the strong backbone of ViT-L/14. We get consistent improvement over all datasets, including small-scale (CIFAR-100), large-scale (ImageNet), and fine-grained (Food101, FGVC, Pets, and Flowers102) datasets.

3.2. Test-time Adaptation on Pre-trained ImageNet Classifiers

We use Diffusion-TTA to adapt ImageNet classifiers with different backbones: ResNet18 (He et al., 2016), ViT-B/32 (Dosovitskiy et al., 2020), and ConvNext-Tiny (Liu et al., 2022). For our class-conditional generative model, we use Diffusion Transformer (DiT) (Peebles & Xie, 2022), which is trained from scratch on ImageNet. This enables a fair comparison as we do not use extra data.

	Food101	CIFAR-100	FGVC	Oxford Pets	Flowers102	ImageNet
Synthetic SD Data (Ravuri & Vinyals, 2019; Azizi et al., 2023)	12.6	-	9.4	31.3	22.1	18.9
SD Features	73.0	-	35.2	75.9	70.0	56.6
Diffusion Classifier (Li et al., 2023a)	77.9	-	24.3	85.7	56.8	58.4
CLIP-ViT-B/32	78.4	60.0	18.8	77.8	64.1	56.3
+ Diffusion-TTA (Ours)	**80.2** (+1.8)	**61.8** (+1.6)	**22.2** (+3.4)	**81.1** (+3.3)	**64.3** (+0.2)	**58.1** (+1.8)
CLIP-ViT-B/16	84.7	**68.8**	21.4	80.5	67.6	60.6
+ Diffusion-TTA (Ours)	**85.5** (+0.8)	67.6	**22.6** (+1.2)	**80.8** (+0.3)	**69.2** (+1.6)	**61.5** (+0.9)
CLIP-ViT-L/14	**91.2**	79.6	29.0	89.2	75.2	68.9
+ Diffusion-TTA (Ours)	**91.2** (+0.0)	**80.6** (+1.0)	**30.6** (+1.6)	**89.8** (+0.6)	**76.1** (+0.9)	**69.9** (+1.0)

Table 1: Our method consistently improves open-vocabulary classifiers for zero-shot classification. Our evaluation is performed across multiple model sizes and a variety of zero-shot datasets.

Datasets We consider the following datasets for TTA of ImageNet classifiers: ImageNet (Deng et al., 2009) (in-distribution) and its out-of-distribution variants, whic indclude ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), and ImageNetV2 (Recht et al., 2019). We report the average accuracy over all these datasets in ImageNet-OOD column. In Table 4 in Appendix, we detail and provide results for each dataset.

Baselines. TTT-MAE (Gandelsman et al., 2022) is a state-of-the-art TTA approach. Specifically it is a per-example test-time adaptaion model that uses masked-autoencoding as self-supervised loss for test-time adaptation. Before training for adaptation, TTT-MAE trains a classification head supervised on top of a frozen pre-trained MAE model. In contrast, our method doesn't require any form of inital supervised training but directly uses pre-trained classifiers for TTA. TTT-MAE already showed results on most of the datasets in their paper. For the remaining ones we use their official pre-trained weights and code to train for TTA.

In Table 2 we present quantitative classification results on in-distribution (ImageNet) and out-of-distribution (ImageNet-OOD) which includes ImageNet-C, R, A, and V2. We conclude that that our method:

(i) Our method consistently improves classifiers on the in-distribution (ImageNet) testing images. Without any effort of introducing hand-crafted augmentations, our method increases classification accuracy of strong classifiers (81.9% to 83.1%).

(ii) Our method works well across different types of image classifiers. For all ResNet, ViT, and ConvNext-Tiny, we observe significant performance gains.

(iii) Our method is robust to different types of out-of-distribution testing images.

	ImageNet	ImageNet-OOD
TTT-MAE (Gandelsman et al., 2022): before	82.1	34.4
TTT-MAE: after TTA	82.0	40.1
ResNet18	69.5	23.9
+ Diffusion-TTA (Ours)	**77.2** (+7.7)	**27.8** (+3.9)
ViT-B/32	75.7	38.7
+ Diffusion-TTA (Ours)	**77.6** (+1.9)	**40.6** (+1.9)
ConvNext-Tiny	81.9	39.4
+ Diffusion-TTA (Ours)	**83.1** (+1.2)	**41.9** (+2.5)

Table 2: Our method improves pre-trained image classifiers on both in-distribution and out-of-distribution images. ResNet18, ViT-B/32, and ConvNext are pre-trained on ImageNet. We also test on ImageNet out-of-distribution variants. We observe consistent and significant performance gain across all types of classifiers and distribution drifts.

(iv) TTT-MAE is not as general as our method to different types of data distribution. TTT-MAE excels at certain types of distribution drifts. However, it decreases the performance on in-distribution testing images (-0.1% on ImageNet).

4. Conclusion

We introduced Diffusion-TTA, an effective plug-and-play method for test time adaptation that uses generative feedback from a pre-trained diffusion model to adapt large scale image classifiers. Test time adaptation is carried out for each example in the test-set independently by backpropagating diffusion likelihood gradients to the discriminative model weights. We have shown that our model outperforms previous state-of-the-art TTA methods, and that it is effective across multiple classifier and diffusion model variants.

References

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., and Fleet, D. J. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., and Babenko, A. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Clark, K. and Jaini, P. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gandelsman, Y., Sun, Y., Chen, X., and Efros, A. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023a.

Li, D., Ling, H., Kim, S. W., Kreis, K., Fidler, S., and Torralba, A. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21330–21340, 2022.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023b.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. Technical report, 2013.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

Prabhudesai, M., Paul, S., van Steenkiste, S., Sajjadi, M. S., Goyal, A., Pathak, D., Fragkiadaki, K., Aggarwal, G., and Kipf, T. Test-time adaptation with slot-centric models. In *Sixth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ravuri, S. and Vinyals, O. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Wang, Q., Fink, O., Van Gool, L., and Dai, D. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Yu, T., Xiao, T., Stone, A., Tompson, J., Brohan, A., Wang, S., Singh, J., Tan, C., Peralta, J., Ichter, B., et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.-F., Barriuso, A., Torralba, A., and Fidler, S. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10145–10155, 2021.

A. Appendix

We present Diffusion-TTA, a test-time adaptation approach that modulates the text conditioning of a text conditional pre-trained image diffusion model to adapt pre-trained image classifiers and large-scale CLIP models to individual unlabelled images. We show improvements on multiple datasets including ImageNet and its out-of-distribution variants (C, R, A, and V2), CIFAR-100, Food101, FGVC, Oxford Pets, and Flowers102 over the initially employed classifiers. In the Supplementary, we include further details on our work:

1. In Section A.1, we show qualitative results and expand Table 2 in the main paper by reporting scores for each dataset seperately. Further we include additional pseudo-labelling baselines.

2. We provide a detailed analysis of the computational speed of Diffusion-TTA in Section A.2.

3. We detail the hyperparameters and input pre-processing in Section A.3.

4. We provide details of the datasets used for our experiments in Section A.4.

5. We provide details on the baseline methods used in our experiments in Section A.5.

A.1. Additional Results

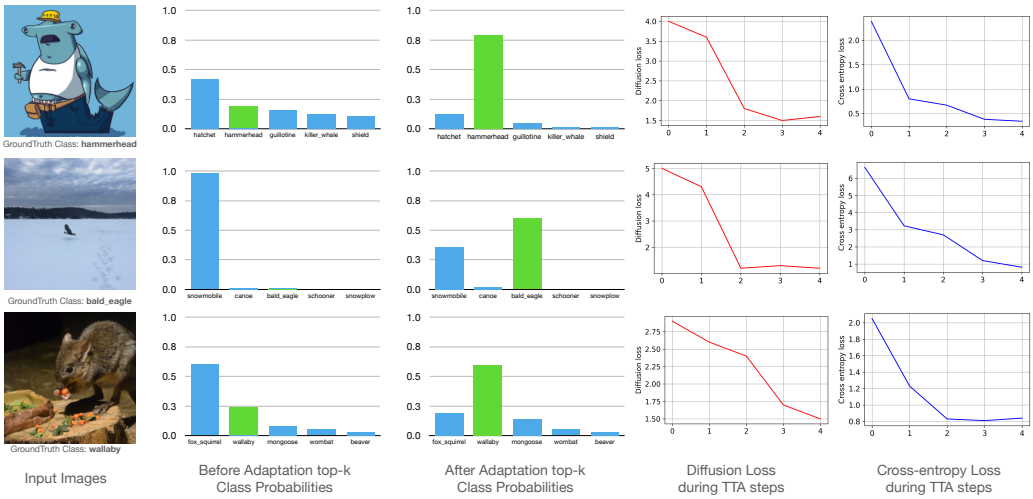

Figure 2: **Visualizing Diffusion-TTA improvement across adaptation steps.** From left to right: We show input image, predicted class probabilities **before** and **after** adaptation (green bars indicate the ground truth category), the diffusion loss as we optimize over TTA steps, and the classification loss across TTA steps. The cross-entropy loss decreases as we minimize the diffusion loss, indicating that there is strong correlation between the diffusion and the classification objective. Note that we only show the cross-entropy loss for this analysis, as we cannot compute it on an unlabeled test image.

A.1.1. Ablative analysis.

We ablate different choices of our model on ImageNet-A dataset with ConvNext as our pre-trained classifier and DiT as our diffusion model in Table 3 and Figure 3. If we do test-time adaptation using a single randomly sampled timestep and noise latent (+*diffusion TTA*), we find a significant reduction in the classification accuracy (-2.9%). Increasing the batch size from 1 to 180 by sampling random timesteps from an uniform distribution (+ *timestep aug BS=180*) gives a significant boost in accuracy ($+4.9\%$). Further sampling random noise latents per timestep (+ *noise aug*) gives an added boost of ($+0.2\%$). Adapting only the top-K predicted classes of the classifier (+ *top-K selection*) gives further ($+0.2\%$) boost (Figure 3). Finally adapting the diffusion weights of DiT (+ *adapting diffusion weights*) in Section 3.2, gives further ($+0.7\%$) boost.

	ImageNet	ImageNet-A	ImageNet-R	ImageNet-C	ImageNet-V2
COTTA (Wang et al., 2022)	71.2	0.0	24.2	11	58.0
TENT (Wang et al., 2020)	71.1	0.0	24.3	12	58.1
TTT-MAE (Gandelsman et al., 2022): before	82.1	14.4	33.0	17.5	**72.5**
TTT-MAE: after TTA	82.0	21.3	39.2	27.5	72.3
ResNet18	69.5	1.4	34.6	2.6	57.1
+ Diffusion-TTA (Ours)	**77.2** (+7.7)	**3.0** (+1.6)	**39.7** (+5.1)	**4.5** (+1.9)	**63.8** (+6.7)
ViT-B/32	75.7	9.0	45.2	39.5	61.0
+ Diffusion-TTA (Ours)	**77.6** (+1.9)	**10.0** (+1.0)	**46.5** (+1.3)	**41.4** (+7.7)	**64.4** (+3.4)
ConvNext-Tiny	81.9	22.3	47.8	16.4	70.9
+ Diffusion-TTA (Ours)	**83.1** (+1.2)	**25.2** (+2.9)	**49.7** (+1.9)	**21.0** (+4.6)	**71.5** (+0.6)

Table 4: Our method improves pre-trained image classifiers on both in-distribution and out-of-distribution images. ResNet18, ViT-B/32, and ConvNext are pre-trained on ImageNet, where images are in-distribution to the classifiers. We also test on ImageNet variants where images are out-of-distribution to the classifiers. We report top-1 accuracy of image classification before and after adaptation. We observe consistent and significant performance gain across all types of classifiers and distribution drifts.

	ImageNet-A
ConvNext-Tiny (Liu et al., 2022)	22.3
+ diffusion loss TTA	19.4 **(-2.9)**
+ timestep aug	24.1 **(+4.9)**
+ noise aug	24.3 **(+0.2)**
+ top-K selection	24.5 **(+0.2)**
+ adapting diffusion weights	**25.2** **(+0.7)**

Table 3: **Ablative analysis** of Diffusion-TTA components for ConvNext classifier and DiT diffusion model.

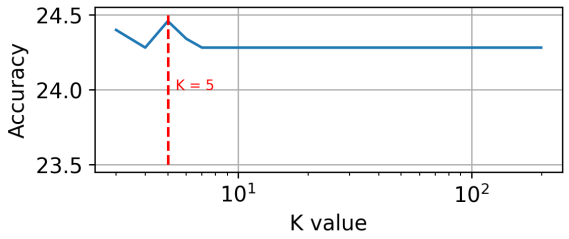

Figure 3: We report how the number of K predictions we adapt affects top-1 accuracy on ImageNet-A with the ConvNext classifier.

A.1.2. Detailed Results.

In Table 4, we expand Table 2 by reporting results on each dataset in ImageNet-OOD column and including additional baselines. Further In Figure 2, we visualize intermediate test-time adpatation results following the setting in Section 3.2.

A.2. Analysis of Computation Speed

In this section we detail the computation requirements of Diffusion-TTA. We conduct all our experiments on a single A100 GPU (40 GB RAM), we do 5 test-time adaptation per example, each adaptation step requiring about 1 to 1.38 seconds dependent on the setting. For instance, the setting with the largest models i.e CLIP L/14 + Stable diffusion takes about 1.45 seconds per TTA step, while our smallest model setting of ResNet-18 + DiT takes about 1 second per TTA step. We maintain a batch size of 180 for all our experiments. As we can only fit a batch size of 16 and 24 in our GPU memory for Stable Diffusion and DiT experiments, we use gradient accumulation for 11 and 9 steps respectively. This increases our per-step adaptation time from 1 seconds to 11 seconds. Therefore given 5 adaptation steps, our total adaptation time per

example ranges from 55 to 66 seconds dependent on the setting.

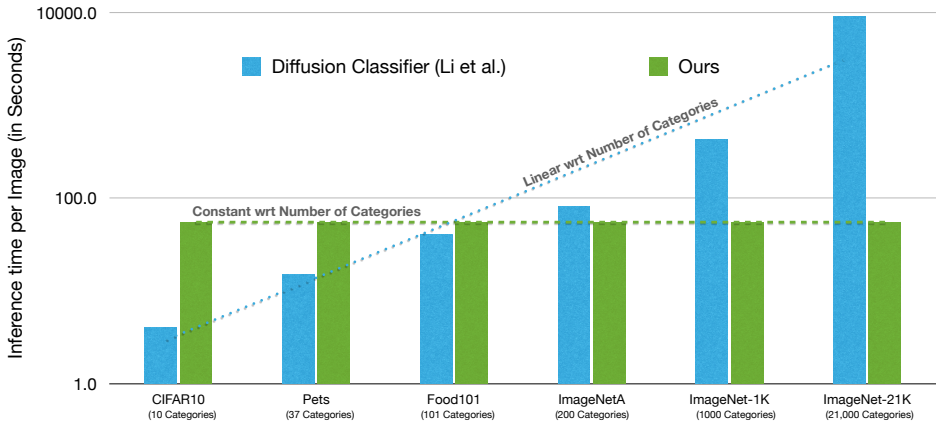

Figure 4: Our method becomes more computationally efficient than Diffusion Classifier (Li et al., 2023a) as the number of categories in the dataset increases. Diffusion Classifier requires to do a forward pass through the diffusion model for each category in the dataset and the computation increases linearly with the number of categories. On the other hand, our method adapts pre-trained classifiers and thus the computation is instead dependent on the throughput of the classifier.

Computation compared to Diffusion Classifier. Diffusion Classifier (Li et al., 2023a) inverts a diffusion model by performing discrete optimization over categorical text prompts, instead we obtain continuous gradients to search much more effectively over a pre-trained classifier's parameter space. This design choice makes significant trade-offs in-terms of computation costs. For instance, Diffusion Classifier requires to do a forward pass for each category seperately and thus the computation increases linearly with the number of categories, however for us it's independent of the number of categories. We compare the per-example inference speed in Figure 4, across various datasets.

A.3. Hyper-parameters and Input Pre-Processing

For test-time-adaptation of individual images, we randomly sample 180 different pairs of noise ϵ and timestep t for each adaptation step, composing a mini-batch of size 180. Timestep t is sampled over an uniform distribution from the range 1 to 1000 and epsilon ϵ is sampled from an unit gaussian. We apply 5 test-time adaptation steps for each input image. We adopt Stochastic Gradient Descent (SGD), and set learning rate, weight decay and momentum to 0.005, 0, and 0.9, respectively. To isolate the contribution of the improvement due to the diffusion model we do not use any form of data augmentation during test-time adaptation.

We use Stable Diffusion v1.4 (Rombach et al., 2022) to adapt CLIP models. For the CLIP classifier, we follow their standard preprocessing step where we resize and center-crop to a resolution of 224×224 pixels. For Stable Diffusion, we process the images by resizing them to a resolution of 512×512 pixels which is the standard image size in Stable Diffusion models. For the CLIP text encoder in Stable Diffusion and the Classifier we use the text prompt of "a photo of a <class_name>", where <class_name> is the name of the class label as mentioned in the dataset.

For the adaptation of ImageNet classifiers, we use pre-trained Diffusion Transformers (DiT) (Peebles & Xie, 2022) specifically their XL/2 model of resolution 256×256, which is trained on ImageNet1K. For ImageNet classifiers we follow the standard data pre-processing pipeline where we first resize the image to 232×232 and then do center crop to a resolution of 224×224 pixels. For DiT we resize the image to 256×256, before passing it as input to their model.

For top-K filtering we set the value of K as 5, for all the experiments in our paper. For all experiments, we adjust all the parameters of the classifier. We freeze the diffusion model parameters for the open-vocabulary experiments in Section 3.1 with CLIP and Stable Diffusion. For our experiments on adapting to ImageNet distribution shift in Section 3.2, we observe a slight performance boost when we adapt the parameters of both the diffusion model and the classifier. This is likely because the Diffusion Transformer (Peebles & Xie, 2022) class-conditional generative model was trained only on ImageNet. Thus, optimizing its parameters at test time helps it adapt to each distribution shift and give better gradient signal to the classifier.

A.4. Datasets

In the following section, we provide further details on the datasets used in our experiments. We adapt CLIP classifiers to the following datasets incuding ImageNet: **1) CIFAR-100** (Krizhevsky et al., 2009) is a compact, generic image classification dataset featuring 100 unique object classes, each with 100 corresponding test images. The resolution of these images is relatively low, standing at a size of 32×32 pixels. **2) Food101** (Bossard et al., 2014) is an image dataset for fine-grained food recognition. The dataset annotates 101 food categories and includes 250 test images for each category. **3) FGVC Airplane** (Maji et al., 2013) is an image dataset for fine-grained categorization containing 100 different aircraft model variants, each with 33 or 34 test images. **4) Oxford Pets** (Parkhi et al., 2012) includes 37 pet categories, each with roughly 100 images for testing. **5) Flower102** (Nilsback & Zisserman, 2008) hosts 6149 test images, annotated in 102 flower categories commonly found in the United Kingdom. For all of these datasets, we sample 5 images per category for testing. In Figure 5, we visualize an example from each dataset.

Figure 5: Image datasets for zero-shot classification. **From left to right:** we show an image example obtained from Food101, CIFAR-100, FGVC, Oxford Pets, Flowers102, and ImageNet dataset. CLIP classifiers are not trained but tested on these datasets.

We consider ImageNet and its out-of-distribution variants for test-time adaptation as shown in Figure 6. **1) ImageNet** (Deng et al., 2009) is a large-scale generic object classification dataset, featuring 1000 object categories. We test on the validation set that consists of 50 images for each category. **2) ImageNet-C** (Hendrycks & Dietterich, 2019) applies various visual corruptions to the same validation set as ImageNet. We consider the shift due to gaussian noise (level-5) in our experiments. **3) ImageNet-A**(Hendrycks et al., 2021b) consists of real-world image examples that are misclassified by existing classifiers. The dataset convers 200 categories in ImageNet and 7450 images for testing. **4) ImageNet-R** (Hendrycks et al., 2021a) is a synthetic dataset that renders 200 ImageNet classes in *art*, *cartoons*, *deviantart*, *graffiti*, *embroidery*, *graphics*, *origami*, *paintings*, *patterns*, *plastic objects*, *plush objects*, *sculptures*, *sketches*, *tattoos*, *toys*, and *video game* styles. The dataset consists of 30K test images. **5) ImageNetV2** collects images from the same distribution as ImageNet, composed of 1000 categories, each with 10 test images. For ImageNet and its C, R, and V2 variants except A, we sample 3 images per category, as these datasets contains a lot more classes than the other datasets. For ImageNet-A we evaluate on it's whole test-set.

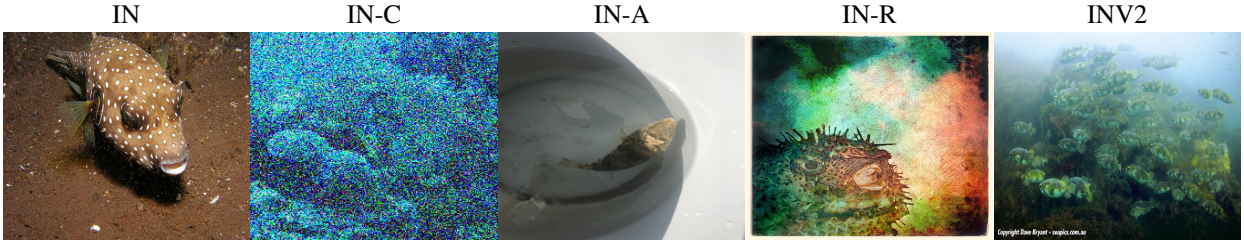

Figure 6: Image distribution drifts cast great challenges for visual recognition. **From left to right:** we show images of the *puffer* category in the ImageNet (IN), ImageNet-C, ImageNet-A, ImageNet-R, and ImageNetV2 dataset. ImageNet-C applies different corruptions to images (we use *Gaussian Noise* in this paper). ImageNet-A consists of real-world image examples that are misclassified by existing classifiers. ImageNet-R renders images in artistic styles, e.g. cartoon, sketch, graphic, *etc*. ImageNetV2 attempts to collect images from the same distribution as ImageNet, but still suffers from minor distribution shift. Though these images all correspond to the same category, they are visually dissimilar and can easily confuse ImageNet-trained classifiers.

A.5. Baselines

We describe the baselines used in our experiments. We *exactly* follow the set up employed in their official codebase.

TTT-MAE consists of a ViT-L/16 as the feature extractor and a ViT-B/16 as the classifier, where each model is pre-trained for masked image reconstruction and classification on ImageNet, respectively. We follow the default hyper-parameters used in the released code base[1].

TENT (Wang et al., 2020) is a TTA method that adapt only the normalization statistics and model parameters in the normalization of the classifier. The optimization objective is to minimize the entropy (maximize the confidence) of the classifier's predictions. Our method instead optimizes the diffusion loss. We train TENT for test-time adaptation on our benchmarks using their official codebase.

COTTA (Wang et al., 2022) adopts a teacher-student model distillation strategy to adapt the classifier. COTTA supervises the student classifier using *pseudo labels* that are the weight averaged classifications of multiple augmentations predicted by the teacher model. The teacher classifier is updated via moving average with the student's weights. In contrast, our approach updates the classifier simply with gradient descent. We evaluate COTTA on our benchmarks using their official codebase.

[1] https://github.com/yossigandelsman/test_time_training_mae