

# ACCELERATING DIFFUSION SAMPLING WITH CLASSIFIER-BASED FEATURE DISTILLATION

Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, Chun Chen

Zhejiang University, Hangzhou, China  
sunwujie@zju.edu.cn

## ABSTRACT

Although diffusion model has shown great potential for generating higher quality images than GANs, slow sampling speed hinders its wide application in practice. Progressive distillation is thus proposed for fast sampling by progressively aligning output images of  $N$ -step teacher sampler with  $N/2$ -step student sampler. In this paper, we argue that this distillation-based accelerating method can be further improved, especially for few-step samplers, with our proposed **Classifier-based Feature Distillation (CFD)**. Instead of aligning output images, we distill teacher’s sharpened feature distribution into the student with a dataset-independent classifier, making the student focus on those important features to improve performance. We also introduce a dataset-oriented loss to further optimize the model. Experiments on CIFAR-10 show the superiority of our method in achieving high quality and fast sampling. Code is provided at <https://github.com/zju-SWJ/RCFD>.

**Index Terms**— Diffusion model, knowledge distillation, image generation, fast sampling

## 1. INTRODUCTION

Image generation is an important research field in computer vision and various models have been invented, such as generative adversarial networks (GANs) [1] and diffusion models [2]. The adversarial nature of GANs requires careful architecture and hyper-parameter selection to stabilize the model training, while the recent diffusion models can overcome these weaknesses and achieve better performance [3]. However, diffusion models require a greatly slower iterative sampling to get the final denoised images. How to accelerate the sampling efficiency becomes a critical issue.

Two main acceleration directions are training-free sampling and training scheme [2]. Training-free sampling [4, 5, 6] aims to propose efficient sampling methods to boost sampling speed for the pre-trained diffusion models, while training scheme methods [7, 8, 9] require additional training, but it gives model the potential for more powerful performance.

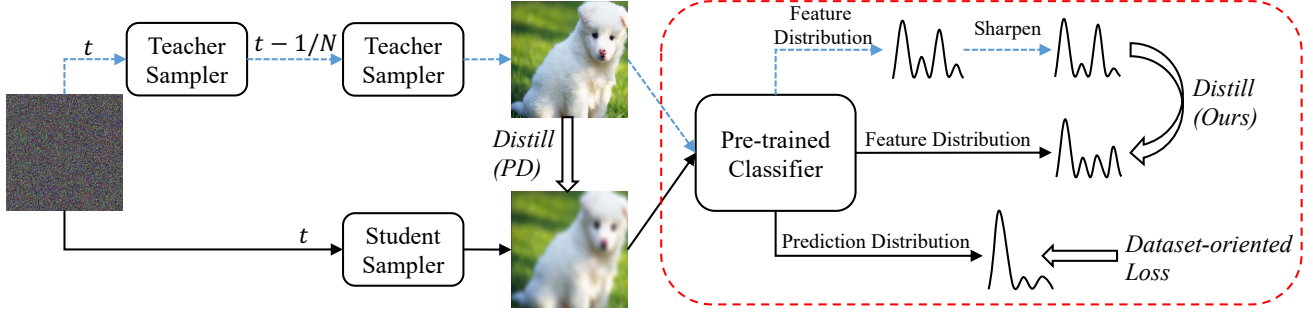
Recently, knowledge distillation-based training scheme methods [7, 10] have exhibited strong capabilities in fast sam-

pling and high performance, surpass other methods [4, 5, 6, 8, 9] with large margins. Inspired by the idea of distilling the knowledge in a powerful teacher model into a compact student model [11, 12], Progressive Distillation (PD) [7] lets the student sampler mimic the teacher sampler’s two-step output with a single step. In this way, the sampler maintains a decent performance when progressively halving its sampling steps. However, little work has been done upon this.

In this paper, by using an additional classifier, we further demonstrate the power of knowledge distillation in speeding up diffusion sampling. We argue that strictly aligning the individual pixels in output images of the student and teacher samplers is difficult, especially for student samplers with few sampling steps. With the help of a classifier, we can get the high-level feature distributions based on the images output by teacher and student. By calculating the KL-divergence of these two distributions, student is able to focus on those important features (which are closely related to image composition), thus reducing the learning burden and ensuring the consistency of the image. We name it **Classifier-based Feature Distillation (CFD)**. Notice that at this point, our classifier is NOT necessarily trained on the target dataset, since it is only used for feature extraction and does not involve category information. This allows our method to be applied to datasets that are not used for classification. Such classifier, which does not require adversarial training and pre-training on the target dataset, makes our work very different from previous works with classifiers for image generation and refinement [13, 14]. For classifiers trained on the target dataset, we further propose **Regularized CFD (RCFD)** which combines CFD with entropy and diversity losses to further optimize model performance. We provide an overview of our method in Figure 1.

Our contributions can be summarized as follows:

- We propose a novel classifier-based distillation method for speeding up diffusion sampling speed.
- Our method does not involve adversarial training, and does not require the classifier to be pre-trained on the target dataset, making our method easy to use and widely applicable.
- Experiments on CIFAR-10 show that our method out-



**Fig. 1.** Overall framework of our method. Instead of directly aligning images as PD, we use the pre-trained classifier to get feature and prediction distributions. The teacher’s feature distribution is sharpened and distilled to the student, which involves no category information, giving our method wide applicability. The prediction distribution is guided by dataset-oriented loss to further improve performance, which can be used when a pre-trained classifier on the target dataset is available.

performs SOTA methods with large margins.

## 2. RELATED WORK

### 2.1. Diffusion model

Diffusion models aim to sample high-quality images from random noises, which contains two processes: training and sampling. A standard training process is proposed by DDPM [15]. The well-trained network with parameter  $\theta$  could take noisy image  $\mathbf{z}_t$  and time  $0 \leq t \leq 1$  as inputs, and outputs the predicted denoised image  $\mathbf{x}_t = \theta(\mathbf{z}_t, t) = \theta(\mathbf{z}_t)$ . Starting from  $t = 1$ , the sampling process is then repeated  $N$  times to get the final generated image. Since such sampling process is very time-consuming, DDIM [4] proposes an implicit sampling to speed up, which can be represented as

$$\mathbf{z}_s = \alpha_s \underbrace{\theta(\mathbf{z}_t)}_{\text{predicted denoised image } \mathbf{x}_t} + \sigma_s \underbrace{\frac{\mathbf{z}_t - \alpha_t \theta(\mathbf{z}_t)}{\sigma_t}}_{\text{direction pointing to } \mathbf{z}_t}, \quad (1)$$

where  $\alpha$  and  $\sigma$  are pre-defined time-related functions,  $\mathbf{z}_0$  is the final denoised image, and  $0 \leq s < t \leq 1$ . We provide a more detailed explanation in Appendix A. Based on DDIM, Progressive Distillation (PD) [7] uses knowledge distillation to improve sampling speed. Other methods such as PNDMs [5] and DPM-Solver [6] also manage to speed up sampling, but fail to outperform PD with huge margins.

### 2.2. Knowledge distillation

Knowledge distillation [11] is an efficient method for model compression. Diverse knowledge such as logits [11] and intermediate features [16], can be transferred from a superior teacher model to a compact student model. In addition, online knowledge distillation [17] introduces multiple training models, while self-distillation [18] contains only a single model

architecture. Although knowledge distillation has a wide applications such as image classification [11] and semantic segmentation [19], distillation for fast diffusion sampling [7] has rarely been explored yet. We believe this field holds great promise.

### 2.3. Classifier for image generation

Classifier is important for image classification. Recent works show that it can also be applied to image generation [13, 14]. However, these methods need a robust classifier with adversarial training, which increases training difficulty. Classifier is also used in diffusion models to provide class-related guidance and improve performance [3]. Different from the above works, in this paper, we use the classifier to extract the feature/prediction distribution of images and transfer it to the student model as knowledge. Such classifier does not require adversarial training and can be pre-trained on a different dataset.

## 3. METHODOLOGY

### 3.1. Progressive distillation

Progressive Distillation (PD) [7] introduces knowledge distillation to speed up sampling. Once teacher sampler with  $N$  steps is given, student sampler with  $N/2$  steps is trained to speed up sampling. Assuming that the sampling time is now  $t$ , we can get the predicted denoised image  $\mathbf{x}^T$  at time  $t - 2/N$  by sampling the teacher model for two steps. The detailed derivation for  $\mathbf{x}^T$  is provided in Appendix B. The training loss for PD is represented as

$$L_{PD} = w_t \|\mathbf{x}^T - \theta(\mathbf{z}_t)\|_2^2, \quad (2)$$

where  $w_t = \max(\alpha_t^2/\sigma_t^2, 1)$  is used for better distillation.

Directly aligning images is very effective when the sampler has many steps, but it degrades rapidly when there are

few steps. We believe that when the sampling steps are small, it becomes difficult for the student to strictly align the pixels on the image, which hinders the model learning. Therefore, we argue that in this situation, the student model should pay more attention to learning the key features associated with images, so as to improve the learning efficiency and quality.

### 3.2. Classifier-based feature distillation

A classifier  $cls$  is usually composed of two parts, feature extractor  $extr$  and fully connected layers.

Instead of aligning  $\mathbf{x}^T$  and  $\theta(\mathbf{z}_t)$  as PD [7], we use a classifier to extract features and use them as transferred knowledge. To be more specific, student’s output image  $\mathbf{x}^S = \theta(\mathbf{z}_t)$  and teacher’s output image  $\mathbf{x}^T$  are input to the same extractor  $extr$ , and output the last features before the fully connected layers, which can be represented as

$$\mathbf{F}^S = extr(\mathbf{x}^S), \quad \mathbf{F}^T = extr(\mathbf{x}^T). \quad (3)$$

After that, we convert feature into distribution using softmax function  $\sigma(\cdot)$ , and calculate the KL-divergence between teacher and student feature distributions

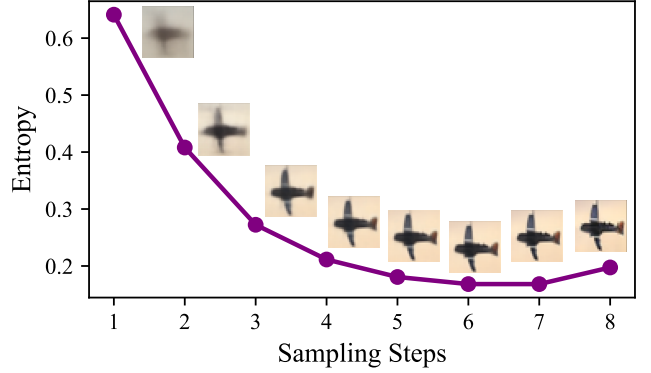
$$L_{CFD} = \text{KL} \left( \sigma(\mathbf{F}^S), \sigma^\tau(\mathbf{F}^T) \right), \quad (4)$$

where temperature  $0 < \tau \leq 1$  is used to sharpen the distribution. Note that  $\tau$  is only applied to teacher feature distribution, which we find to be more effective than applying to both distributions. In this case, the upper limit of student performance is no longer the teacher, so in some cases (see section 4.2), students can even surpass the teacher model!

Different from the L2 distance used in PD, KL divergence can give large feature values greater weight in gradient descent, thus helping the model focus more on aligning these features. After the image is input into the feature extractor, the features change from the shallow fine-grained features to the deep coarse-grained features as the layer increases. Deep features contain more semantic information related to categories, which is crucial for image composition. By aligning important teacher features, and reducing the interference of irrelevant features on model training, students with poor ability can learn more useful knowledge to generate high-quality images and improve performance.

Note that the loss in Equ. 4 is NOT oriented to a specified dataset, since we only use the feature extractor and do not include the subsequent fully connected layers for classification. This advantage makes our proposed distillation method can be extended to more datasets, such as CelebA and LSUN bedrooms. Next, we further introduce dataset-oriented loss to help the model better improve performance.

**Dataset-oriented loss.** For a  $N$ -step sampler, as the sampling step increases, the image obtained by  $\theta(\mathbf{z}_t)$  tends to be clearer. A clearer denoised image in the early steps will benefit the subsequent sampling steps.



**Fig. 2.** Entropy evaluation on CIFAR-10 using 8-step sampler with different sampling steps. The classifier is ResNet18 and the entropy is calculated by averaging 4096 generated images. We give an illustration of the denoised image from the same noise initialization.

By feeding the images obtained from each sampling step into a classifier, we can calculate the entropy as follows

$$L_{\text{entropy}} = - \sum_{c=1}^C \mathbf{p}_c \log \mathbf{p}_c, \quad \mathbf{p} = \sigma(cls(\mathbf{x}^S)). \quad (5)$$

where  $C$  is the total number of classes,  $\mathbf{p}$  denotes prediction results. Figure 2 shows that sampling with fewer steps yield a larger entropy and generate more blurred images. This means if we minimize the entropy of prediction results, we could get relatively clearer images, especially for early sampling steps.

In addition, with the progressive distillation, it inevitably makes the current sampler’s output image distribution deviate more and more from the original optimal one. Since the dataset we used is balanced, we expect the predicted probability to remain equal for each class within each batch:

$$L_{\text{diversity}} = \sum_{c=1}^C \hat{\mathbf{p}}_c \log \hat{\mathbf{p}}_c, \quad \hat{\mathbf{p}} = \frac{\sum_{b=1}^B \mathbf{p}}{B}, \quad (6)$$

where  $B$  is the batch size. These two losses do not involve teacher guidance and are thus less effective when used alone. But better results can be achieved by combining them with  $L_{CFD}$ .

**Overall loss.** The overall loss function can be represented as

$$L_{RCFD} = L_{CFD} + \beta[\gamma L_{\text{entropy}} + (1 - \gamma)L_{\text{diversity}}], \quad (7)$$

where  $\beta$  and  $\gamma$  are hyper-parameters, and RCFD stands for **Regularized Classifier-based Feature Distillation**.

## 4. EXPERIMENT

### 4.1. Setting

In this section, we use CIFAR-10 to validate the superiority of our method. We use the cosine schedule introduced in [8]

Sampling Steps	Method	IS $\uparrow$	FID $\downarrow$
1	<b>RCFD-DenseNet201</b>	<b>8.87</b>	<b>8.92</b>
	RCFD-ResNet18	8.56	12.03
	PD (ICLR 2022)	7.88	15.06
2	<b>RCFD-DenseNet201</b>	<b>9.19</b>	<b>5.07</b>
	RCFD-ResNet18	9.09	6.12
	PD (ICLR 2022)	8.70	7.42
4	<b>RCFD-DenseNet201</b>	<b>9.34</b>	<b>3.80</b>
	RCFD-ResNet18	9.24	4.24
	PD (ICLR 2022)	9.04	4.83
8	PD (ICLR 2022)	9.14	4.14
	DDIM (ICLR 2021)	8.14	20.97
10	PNDMs (ICLR 2022)	-	7.05
12	DPM-Solver (NIPS 2022)	-	4.65
1024	DDIM (ICLR 2021)	9.21	3.78

**Table 1.** Performance comparison with state-of-the-art methods on CIFAR-10. Higher IS and lower FID are better. Results are the average of 3 runs.

to calculate  $\alpha_t$  and  $\sigma_t$ . We use the U-Net [20] as the diffusion model. ResNet18 [21] and DenseNet201 [22] are used as the classifiers. The base diffusion model is trained with 1024 steps. More details are provided in Appendix C.

We compare our method with DDIM [4], Progressive Distillation (PD) [7], PNDMs [5], and DPM-Solver [6]. The distillation-based acceleration method requires iterative training to halve sampling steps. Based on results in [7] and our own experiments, we find that performance changes rapidly in distillation from 8-step to 1-step. So we focus on distillation process starting from 8-step and train the teacher model as PD [7] from 1024 to 8 steps without the classifier. We reimplemented DDIM and PD for better comparison. Note that the PD performance we reported in Table 1 is different from the original paper [7] because we failed to train a good base model on the U-Net architecture used by PD. Therefore, we chose the architecture introduced in DDPM [15], and used a smaller distillation iterations to reduce training overhead.

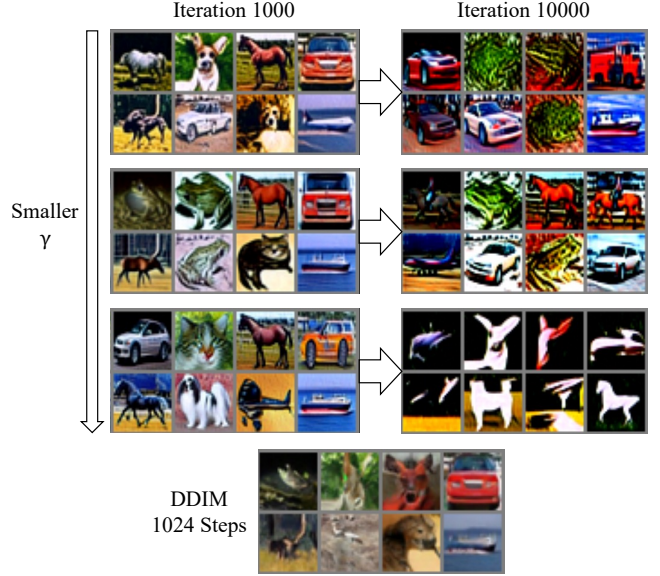
## 4.2. Result

The result is shown in Table 1. As we can see, distillation-based methods (RCFD and PD) surpass other methods with large margin (4-step distillation-based samplers can achieve the performance of other samplers with 10+ steps). Also, the difference between the 8-step sampler obtained by PD and the 1024-step DDIM sampler (base diffusion model) is small, indicating the effectiveness of distillation.

In addition, RCFD with DenseNet201 achieved 6.14 ( $\downarrow 40.7\%$ ), 2.35 ( $\downarrow 31.6\%$ ), and 1.03 ( $\downarrow 21.3\%$ ) FID improvement compared to PD in the 1, 2, and 4-step samplers, respectively, demonstrating its superiority. Also, with the help of the classifier, we offer the possibility for the student sampler (4-

Method	$L_{CFD}$	$L_{entropy}$	$L_{diversity}$	IS $\uparrow$	FID $\downarrow$
RCFD	$\checkmark$			9.14	4.42
		$\checkmark$		2.18	330.27
			$\checkmark$	1.22	308.61
		$\checkmark$	$\checkmark$	5.87	92.07
PD	$\checkmark$	$\checkmark$	$\checkmark$	<b>9.24</b>	<b>4.24</b>
				9.04	4.83

**Table 2.** Impact of each loss on performance.



**Fig. 3.** The first three rows indicate the images of different training stages obtained by only using different scales of  $L_{entropy}$  and  $L_{diversity}$ . It can be seen that in the absence of  $L_{CFD}$ , the visual quality of the images at early training stage surpasses even the sampler with 1024 steps of DDIM, but those images are easily collapsed latter.

step of RCFD-DenseNet201) to significantly outperform its teacher (8-step sampler obtained from PD).

## 4.3. Ablation study

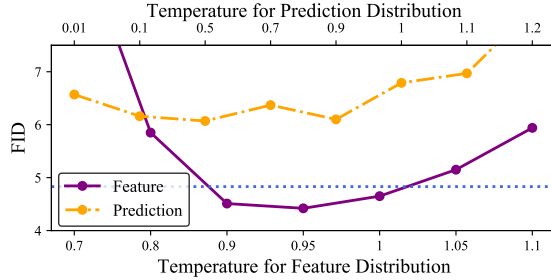
In this section, we perform some ablation studies to verify the importance of each component in our method. If not specified, we use ResNet18 as the classifier and use the 8-step sampler trained by PD as the teacher to train a 4-step student. **Ablation study on each loss.** Three losses are included in our method,  $L_{CFD}$ ,  $L_{entropy}$ , and  $L_{diversity}$ .  $L_{CFD}$  is a dataset-independent loss, which introduces classifier-based distillation to align student’s feature distribution with teacher’s sharpened feature distribution. The latter two are dataset-oriented losses, where  $L_{entropy}$  is used to generate clearer images and  $L_{diversity}$  maintains the class balance.

As we can see from Table 2, with only  $L_{CFD}$ , we can already achieve better performance than PD. Although good



Pre-trained Dataset	Classifier	IS $\uparrow$	FID $\downarrow$
CIFAR-10	ResNet18	9.14	4.42
	ResNet50	9.16	4.24
	DenseNet201	<b>9.34</b>	<b>3.80</b>
ImageNet	ResNet50	9.05	4.62

**Table 3.** Impact of different pre-trained classifiers on CIFAR-10 image generation performance. For better comparison, we use  $L_{CFD}$  only.



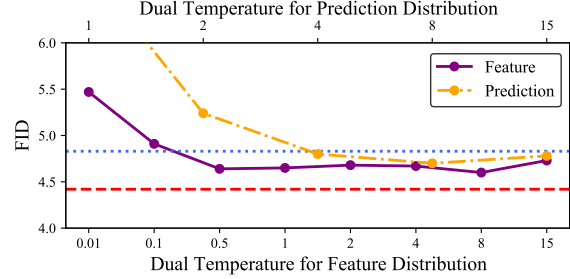
**Fig. 4.** Impact of softmax temperature on performance. The FID of PD is shown by the blue dotted line. For better comparison, we use  $L_{CFD}$  only.

results cannot be achieved using  $L_{entropy}$  and  $L_{diversity}$  when  $L_{CFD}$  is not available, optimal performance can be achieved by combining all three terms. The reason is that the teacher constraint ( $L_{CFD}$ ) will prevent the generated images from being too abstract and meaningless during training (as shown in Figure 3), which improves the model performance.

**Ablation study on classifier.** In this section, we try different classifiers and see how the performance changes. As shown in Table 3, no matter what classifier we use, we can achieve better results than PD. However, if possible, it is better to train the classifier on the target image generation dataset. In addition, as the classifiers become more and more powerful, they also help the student samplers produce higher quality images, which even achieve significantly better FID than the teacher. We believe that for a more powerful classifier, it will extract more accurate and meaningful features, therefore, it provides students with more effective knowledge for distillation, thus helping students produce better images.

**Ablation study on  $L_{CFD}$ .** In our method, we align feature distributions rather than prediction distributions since the former is dataset-independent and achieve better results.

In this section, we provide the performance comparison of these two approaches under different softmax temperatures. Figure 4 shows that aligning student’s feature distribution with teacher’s slightly sharpened feature distribution (temperature  $0.9 \leq \tau \leq 1$ ) obtain better results than PD and almost always outperform distilling the prediction distributions. For the feature distribution, over large temperature will make the teacher’s feature distribution tend to be uniformly distributed, hindering the learning of important features and



**Fig. 5.** Impact of dual softmax temperature. The FIDs of PD and CFD are shown by the blue dotted line and red dashed line, respectively. For better comparison, we use  $L_{CFD\_dual}$  only.

making the image meaningless, while over small temperature makes few features to be highlighted, making the image too abstract and causing performance degradation. For the prediction distribution, since it has smaller constraints compared to the feature distribution (i.e., different feature distributions may yield the same prediction results), the learning of image details can be weakened, which leads to bad performance.

**Ablation study on dual softmax temperature.** In our method, softmax temperature  $\tau$  is only used for the teacher, as shown in Equ. 4. We now apply the same temperatures  $\tau$  to both the student and the teacher, and change the loss as

$$L_{CFD\_dual} = \tau^2 \text{KL} \left( \sigma^\tau(\mathbf{F}^S), \sigma^\tau(\mathbf{F}^T) \right). \quad (8)$$

Figure 5 show that, for a wide range of temperatures, aligning feature distributions and prediction distributions achieve better performance than PD, but fails to outperform the original  $L_{CFD}$  which only uses temperature for the teacher. Although large dual temperature helps to improve performance when prediction distributions are aligned, we believe that such aligning (no matter it is feature or prediction distribution) determines that the upper limit of the student is the teacher (unlike traditional knowledge distillation for image classification, there is no additional guidance such as labels during distillation), which limits performance improvement.

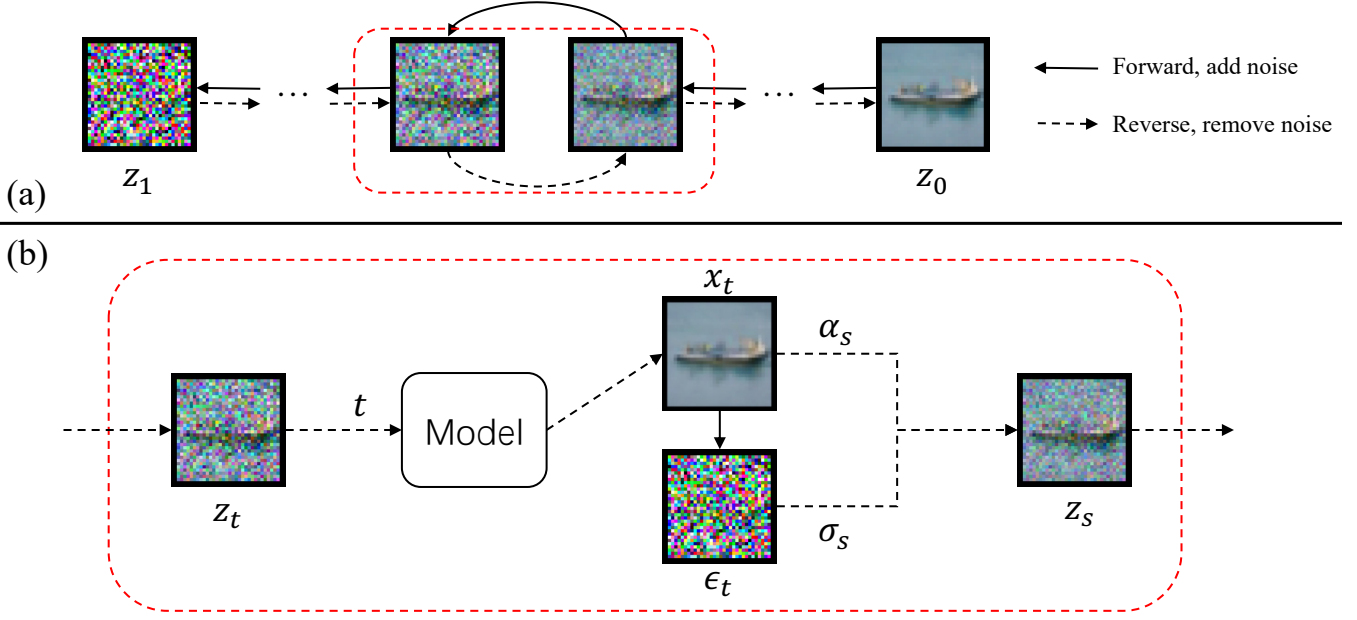
## 5. CONCLUSION

In this paper, we propose a novel classifier-based distillation method to speed up the sampling of the diffusion models. We let student align its feature distribution with teacher’s sharpened feature distribution, rather than aligning the generated images. In this way, student can focus on learning important features that make up an image, resulting in even better performance than the teacher. This distillation method is also applicable when the classifier is pre-trained on other datasets. When the classifier pre-trained on the target dataset is available, we propose a dataset-oriented loss to further improve

performance. Experiments on CIFAR-10 show the superiority of our method.

## 6. REFERENCES

- [1] Divya Saxena and Jiannong Cao, “Generative adversarial networks (gans) challenges, solutions, and future directions,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–42, 2021. [1](#)
- [2] Hanqun Cao, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li, “A survey on generative diffusion model,” *arXiv preprint arXiv:2209.02646*, 2022. [1](#)
- [3] Prafulla Dhariwal and Alexander Nichol, “Diffusion models beat gans on image synthesis,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021. [1](#), [2](#)
- [4] Jiaming Song, Chenlin Meng, and Stefano Ermon, “Denoising diffusion implicit models,” in *International Conference on Learning Representations*, 2020. [1](#), [2](#), [4](#)
- [5] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao, “Pseudo numerical methods for diffusion models on manifolds,” in *International Conference on Learning Representations*, 2021. [1](#), [2](#), [4](#)
- [6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu, “Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps,” *arXiv preprint arXiv:2206.00927*, 2022. [1](#), [2](#), [4](#)
- [7] Tim Salimans and Jonathan Ho, “Progressive distillation for fast sampling of diffusion models,” in *International Conference on Learning Representations*, 2021. [1](#), [2](#), [3](#), [4](#)
- [8] Alexander Quinn Nichol and Prafulla Dhariwal, “Improved denoising diffusion probabilistic models,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171. [1](#), [3](#)
- [9] Zhifeng Kong and Wei Ping, “On fast sampling of diffusion probabilistic models,” in *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021. [1](#)
- [10] Eric Luhman and Troy Luhman, “Knowledge distillation in iterative generative models for improved sampling speed,” *arXiv preprint arXiv:2101.02388*, 2021. [1](#)
- [11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al., “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015. [1](#), [2](#)
- [12] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen, “Knowledge distillation with the reused teacher classifier,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11933–11942. [1](#)
- [13] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, “Image synthesis with a single (robust) classifier,” *Advances in Neural Information Processing Systems*, vol. 32, 2019. [1](#), [2](#)
- [14] Roy Ganz and Michael Elad, “Bigroc: Boosting image generation via a robust classifier,” *arXiv preprint arXiv:2108.03702*, 2021. [1](#), [2](#)
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. [2](#), [4](#), [7](#)
- [16] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen, “Cross-layer distillation with semantic calibration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 7028–7036. [2](#)
- [17] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen, “Online knowledge distillation with diverse peers,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 3430–3437. [2](#)
- [18] Tommaso Furlanello, Zachary Lipton, Michael Tschanen, Laurent Itti, and Anima Anandkumar, “Born again neural networks,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 1607–1616. [2](#)
- [19] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen, “Channel-wise knowledge distillation for dense prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320. [2](#)
- [20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. [4](#)
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [4](#)
- [22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. [4](#)
- [23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016. [7](#)
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017. [7](#)



**Fig. 6.** An overview of the training and sampling of diffusion models. (a) The diffusion models contain two processes, forward (turn the image into noise) and reverse (remove noise from the image). Our target is to model the reverse process using the neural network (which can be achieved by using DDPM [15]), so that we can get images from random noises. (b) Once the model is well trained, we can use it multiple times to get the denoised image. For sampling time  $t$ , we can get  $\mathbf{z}_t$  from previous step, and we input  $\mathbf{z}_t$  and  $t$  into the model, which outputs the predicted denoised image  $\mathbf{x}_t$  (or predicted noise  $\epsilon_t$ , depending on the training target of the model). Based on the  $\mathbf{x}_t$  (or  $\epsilon_t$ ), we can get the corresponding  $\epsilon_t$  (or  $\mathbf{x}_t$ ). After that, we use DDIM (Equ. 1) to calculate the noisy image  $\mathbf{z}_s$ , which is the model input for next step.

## A. TRAINING AND SAMPLING OF DIFFUSION MODELS

We provide an overview of the training and sampling of diffusion models in Figure 6.

### B. DERIVATING DENOISED IMAGE FOR DISTILLATION

Assume we have  $N$ -step teacher, and the current time is  $t$ , then we can get  $t' = t - 1/N$  and  $t'' = t - 2/N$ .  $\mathbf{z}'_t$  and  $\mathbf{z}''_t$  are calculated as

$$\mathbf{z}'_t = \alpha_{t'}\eta(\mathbf{z}_t) + \sigma_{t'}\frac{(\mathbf{z}_t - \alpha_t\eta(\mathbf{z}_t))}{\sigma_t}, \quad (9)$$

$$\mathbf{z}''_t = \alpha_{t''}\eta(\mathbf{z}'_t) + \sigma_{t''}\frac{(\mathbf{z}'_t - \alpha_{t'}\eta(\mathbf{z}'_t))}{\sigma_{t'}}, \quad (10)$$

where  $\eta$  is the teacher model.

Assume student has denoised image  $\mathbf{x}^S$  and gets noisy image  $\tilde{\mathbf{z}}_{t''}$  in one step. If well aligned, we should have

$$\mathbf{z}_{t''} = \tilde{\mathbf{z}}_{t''} = \alpha_{t''}\mathbf{x}^S + \sigma_{t''}\frac{(\mathbf{z}_t - \alpha_t\mathbf{x}^S)}{\sigma_t}. \quad (11)$$

The distillation target  $\mathbf{x}^T$  can thus be represented as

$$\mathbf{x}^T = \mathbf{x}^S = \frac{\mathbf{z}_{t''} - (\sigma_{t''}/\sigma_t)\mathbf{z}_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\alpha_t}. \quad (12)$$

## C. EXPERIMENT DETAILS

### C.1. Model Architecture

The U-Net includes four feature map resolutions ( $32 \times 32$  to  $4 \times 4$ ), and it has two convolutional residual blocks per resolution level and self-attention blocks at  $8 \times 8$  resolution. Diffusion time  $t$  is embedded into each residual block. Initial channel number is 128 and is multiplied by 2 at last three resolutions.

### C.2. Performance Evaluation

We report the Inception Score (IS) [23] and Fréchet Inception Distance (FID) [24] results of each method. IS measures the class balance and confidence of the generated images, while FID measures the difference in feature distribution between the generated and real images. Therefore, higher IS and lower FID represent better generated images.

### C.3. Training Setting

Learning rate (warmup for 5000 iterations) 0.0002, dropout 0.1, batch size 128, ema decay 0.9999, gradient clip 1, total iterations 800000.

### C.4. Distillation Setting

**Common setting.** Learning rate (cosine annealing)  $5e-5$ , batch size 128, gradient clip 1, total iterations 10000 for 1024 to 4-step and 20000 for 4 to 1-step.

**RCFD (ResNet18) setting.**

- 8 to 4-step:  $\tau = 0.95, \beta = 0.003, \gamma = 0.75$ .
- 4 to 2-step:  $\tau = 0.95, \beta = 0.003, \gamma = 0.75$ .
- 2 to 1-step:  $\tau = 0.85, \beta = 0.003, \gamma = 0.5$ .

**RCFD (DenseNet201) setting.** Since introducing dataset-oriented loss makes it more difficult to tune hyper-parameters, we only use  $L_{CFD}$  for DenseNet201.

- 8 to 4-step:  $\tau = 0.9, \beta = 0$ .
- 4 to 2-step:  $\tau = 1, \beta = 0$ .
- 2 to 1-step:  $\tau = 0.85, \beta = 0$ .