Fast Inference in Denoising Diffusion Models via MMD Finetuning

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 July 2024, accepted 29 July 2024, date of publication 1 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436698

RESEARCH ARTICLE

Fast Inference in Denoising Diffusion Models via MMD Finetuning

EMANUELE AIELLO, DIEGO VALSESIA, (Member, IEEE), AND ENRICO MAGLI, (Fellow, IEEE)
Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
Corresponding author: Emanuele Aiello (emanuele.aiello@polito.it)

ABSTRACT Denoising Diffusion Models (DDMs) have become a popular tool for generating high-quality samples from complex data distributions. These models are able to capture sophisticated patterns and structures in the data, and can generate samples that are highly diverse and representative of the underlying distribution. However, one of the main limitations of diffusion models is the complexity of sample generation, since a large number of inference timesteps is required to faithfully capture the data distribution. In this paper, we present MMD-DDM, a novel method for fast sampling of diffusion models. Our approach is based on the idea of using the Maximum Mean Discrepancy (MMD) to finetune the learned distribution with a given budget of timesteps. This allows the finetuned model to significantly improve the speed-quality trade-off, by substantially increasing fidelity in inference regimes with few steps or, equivalently, by reducing the required number of steps to reach a target fidelity, thus paving the way for a more practical adoption of diffusion models in a wide range of applications. We evaluate our approach on unconditional image generation with extensive experiments across the CIFAR-10, CelebA, ImageNet and LSUN-Church datasets. Our findings show that the proposed method is able to produce high-quality samples in a fraction of the time required by widely-used diffusion models, and outperforms state-of-the-art techniques for accelerated sampling. Code will be available at: https://github.com/diegovalsesia/MMD-DDM.

INDEX TERMS Denoising diffusion models, fast inference, image generation, MMD.

I. INTRODUCTION

Denoising Diffusion Models (DDMs) [1], [2], [3] have emerged as a powerful class of generative models. DDMs learn to reverse a gradual multi-step noising process to match a data distribution. Samples are then produced by a Markov Chain that starts from white noise and progressively denoises it into an image. This class of models has shown excellent capabilities in synthesising high-quality images [4], [5], audio [6], and 3D shapes [7], [8], recently outperforming Generative Adversarial Networks (GANs) [9], [10], [11], [12] on image synthesis.

However, GANs require a single forward pass to generate samples, while the iterative DDM design requires hundreds or thousands of inference timesteps and, consequently, forward passes through a denoising neural network. The slow sampling process thus represents one of the most significant limitations of DDMs. It is well-known that there is a trade-off between sample quality and speed, measured in the number of timesteps [4], [13]. However, it is currently unclear how low the number of timesteps can be pushed while retaining high quality for a given data distribution [6]. This issue is the focus of a lot of current research in the field, with recent works proposing acceleration solutions which can be divided into two categories: learning-free sampling and learning-based sampling. The learning-free approach focuses on modifying the sampling process without the need for training [13], [14], [15], [16]. These methods are beneficial as they do not require additional computational resources for training, making them straightforward to implement. However, they often achieve limited improvements in sampling speed and may not fully leverage the capabilities of the model. On the other hand, the learning-based approach uses techniques such as truncation [17], [18], knowledge distillation [19],

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

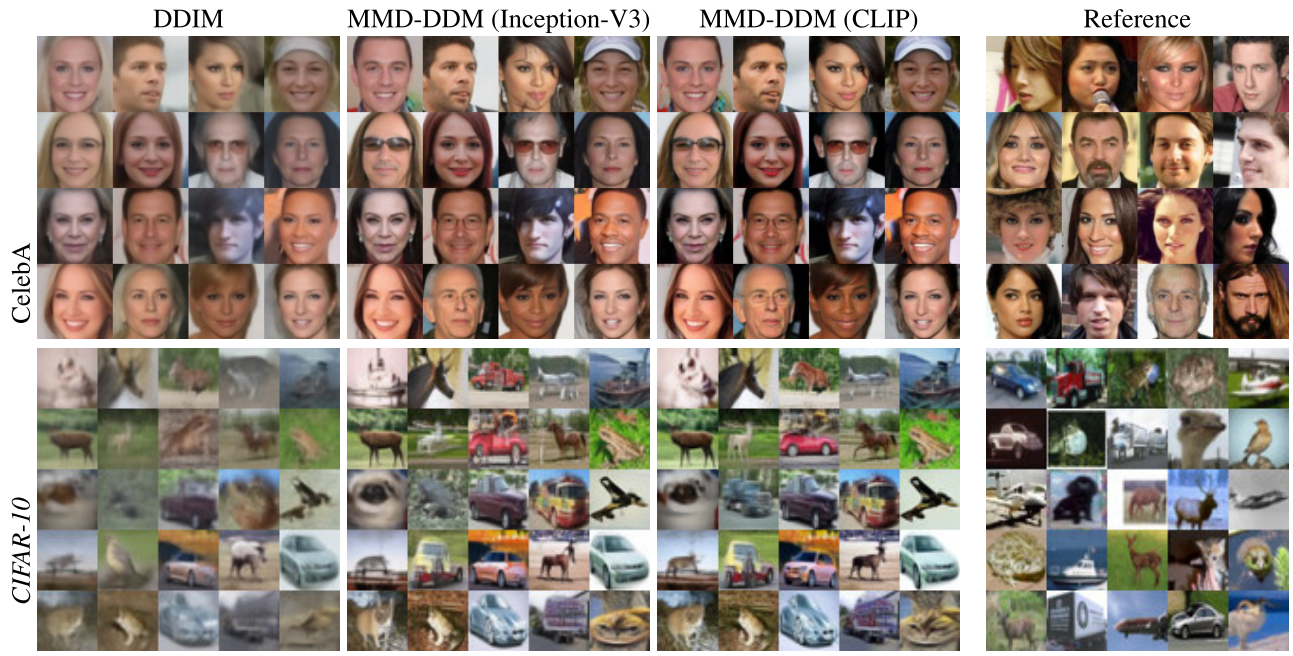

FIGURE 1. Generated samples for CelebA (top) and CIFAR-10 (bottom). The samples are obtained using 5 timesteps with the DDIM sampling procedure. Results from standard DDIM (left), the same model finetuned using MMD with Inception-V3 features (center-left) and CLIP features (center-right), reference images from the dataset (right). Samples are not cherry-picked. Finetuning improves details clarity and sharpness, occasionally introducing semantic changes.

[20], dynamic programming [21] and differentiable sampler search [22] to improve sampling speed. These methods can provide more substantial acceleration but at the cost of additional training complexity and computational overhead. Despite these advancements, significant technical gaps remain, particularly in balancing the trade-off between speed and sample quality without incurring high computational costs.

In this paper, we propose MMD-DDM, a technique to finetune a pretrained DDM with a large number of timesteps in order to optimize the features of the generated data under the constraint of a reduced number of timesteps. This is done by directly optimizing the weights of the denoising neural network via backpropagation through the sampling chain. The minimization objective is the Maximum Mean Discrepancy (MMD) [23] between real and generated samples in a perceptually-relevant feature space. This allows to specialize the model for a fixed and reduced computational budget with respect to the original training; the use of MMD represents a different and, possibly complementary, objective to the original denoising loss. Our proposed approach is extremely fast, requiring only a small number of finetuning iterations. Indeed, the finetuning procedure can be performed in minutes, or at most few hours for more complex datasets, on standard hardware. Moreover, it is agnostic to the sampling procedure, making it appealing even for future models employing new and improved procedures.

Extensive experimental evaluation suggests that the proposed solution significantly outperforms state-of-the-art approaches for fast DDM inference. MMD-DDM is able to substantially reduce the number of timesteps required to reach a target fidelity. We also need to remark that the choice of feature space for the MMD objective may artificially skew results, if the evaluation metric is based on the same feature space being optimized, such as Inception features and the FID score. We discuss the importance of this point for fair evaluation and present results on different metrics and features spaces, such as CLIP features [24], in order to present a fair assessment of the method. The proposed MMD-DDM addresses several gaps left by existing acceleration techniques. By optimizing directly in a perceptually-relevant feature space, it ensures high sample quality even with a reduced number of timesteps, bridging the gap between speed and fidelity without the extensive computational demands of additional training. This makes our method a versatile and efficient solution for accelerating DDMs while maintaining or even enhancing the quality of the generated samples.

To summarize our contributions are:

- We Introduce MMD-DDM, a novel technique that significantly improves the speed-quality trade-off in denoising diffusion models by finetuning with Maximum Mean Discrepancy (MMD) in a perceptually-relevant feature space, allowing for high-quality samples with fewer timesteps.
- We demonstrate that the proposed finetuning process is extremely fast, requiring only a small number of iterations and minimal computational resources, making it practical for a wide range of applications without the need for extensive additional training.

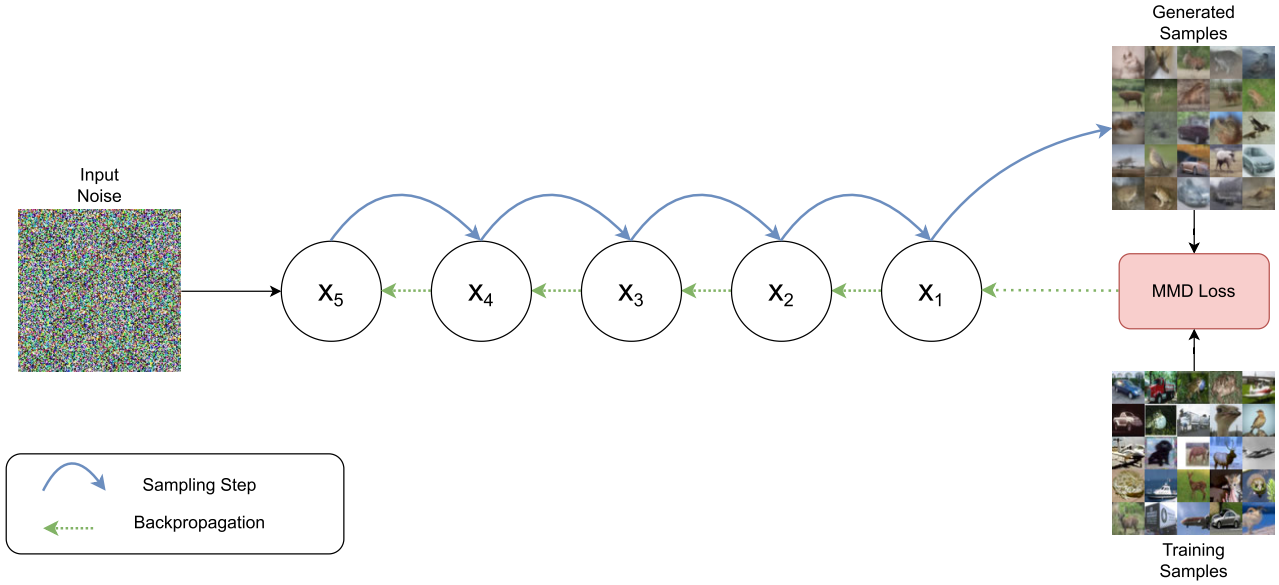

FIGURE 2. Overview of finetuning scheme in the proposed MMD-DDM. During the forward pass samples are generated with a fixed number of sampling steps (e.g 5 in the example). During the optimization process the gradient flows through the sampling chain, directly optimizing the sampling procedure with a limited number of steps.

- We showcase that the method is decoupled from the sampling procedure, ensuring compatibility with future models and samplers.
- We focus on the importance of fair evaluation by presenting results across different metrics and feature spaces, addressing potential biases in standard evaluation methods.

II. RELATED WORK

DDMs ([1], [2]) leverage the diffusion process to model a specific distribution starting from random noise. They are based on a predefined Markovian forward process, by which data are progressively noised in T steps. T is set to be sufficiently large such that $\boldsymbol{x}_T$ is close to white Gaussian noise (in practice, $T \geq 1000$ is often used). The forward process can be written as:

$$q(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T) = q(\boldsymbol{x}_0) \prod_{t=1}^{T} q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) \tag{1}$$

$$q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t | \sqrt{1-\beta_t}\boldsymbol{x}_s, \beta_t \boldsymbol{I}) \tag{2}$$

where $q(\boldsymbol{x}_0)$ denotes the real data distribution and β_t the variance of the Gaussian noise at timestep t. The reverse process traverses the Markov Chain backwards and can be written as:

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \tag{3}$$

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1} \mid \mu_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I}) \,. \tag{4}$$

The parameters of the learned reverse process p_θ can be optimized by maximizing an evidence lower bound (ELBO) on the training set. Under a specific parametrization choice [2], the training objective can be simplified to that of a noise conditional score network [3], [25]:

$$\min_\theta \mathcal{L}(\theta) = \mathbb{E}_{x_0, \epsilon, t} ||\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)||_2^2 \tag{5}$$

where $\boldsymbol{x}_0 \sim q_{\text{data}}$, $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ and t is uniformly sampled from $\{1, \ldots, T\}$.

A. ACCELERATED SAMPLING FOR DDMS

Accelerated DDM sampling is currently a hot research topic. At a high level, the different approaches can be divided into two categories: learning-free sampling and learning-based sampling [26]. The learning-free approaches do not require training and instead focus on modifying the sampling process to make it more efficient. One example is the work of Song et al. [13] (DDIM), in which they define a new family of non-Markovian diffusion processes that maintains the same training objectives as a traditional DDPM. They demonstrate that alternative ELBOs may be built using only a subsequence of the original timesteps $\tau \in \{1, \ldots, T\}$, obtaining faster samplers compatible with a pre-trained DDPM. Other works focus on using the Score SDE formulation [14] of continuous-time DDMs to develop faster sampling methods. For example, Song et al. [14] propose the use of higher-order solvers such as Runge-Kutta methods, while Jolicoeur et al. [15] propose the use of SDE solvers with adaptive timestep sizes. Another approach is to solve the probability flow ODE, which has been shown by Karras et al. [16] to provide a good balance between sample quality and sampling speed when using Heun's second-order method. Additionally, customized ODE solvers such as the DPM-solver [27] and the Diffusion Exponential Integrator sampler [28] have been developed specifically for DDMs and have been shown to be

more efficient than general solvers. These methods provide efficient and effective ways to speed up the sampling process in continuous-time DDMs.

The other main line of approaches for efficient sampling is the learning-based one. Among these approaches we can distinguish between approach that incorporate the acceleration component of the loss at pre-training time and those who have a separate finetuning stage focused on sampling acceleration. The first line of works is represented by methods that use hybrid adversarial and diffusion objectives [29], [30], these methods achieve several order of acceleration without compromising the sample quality. On the same line of work some recent advancements [31] propose to use a fractional multi-phase distilled diffusion prior to improve sampling efficiency, based on the observation that a single denoiser may be insufficient to capture the reverse diffusion process [32]. On the other hand, finetuning based approach does not require to retrain the model and can be adopted on top of large scale pretrained models. Recently proposed Consistency Models [33] share some similarities with MMD-DDMs: both aim to achieve faster sampling through a form of temporal distillation, but they differ significantly in their objectives and theoretical approaches. The primary distinction lies in how the models handle the optimization process. Consistency models do not backpropagate through the sampling chain; instead, they focus on ensuring consistency in the outputs across different timesteps through a fixed procedure. Some of these approaches [17], [18] involve truncating the forward and reverse diffusion processes to improve sampling speed, while others [19], [20] use knowledge distillation to create a faster model that requires fewer steps. Another approach (GENIE) [34], based on truncated Taylor methods, trains an additional model on top of a first-order score network to create a second-order solver that produces better samples with fewer steps. Dynamic programming techniques [21] have also been used to find the optimal discretization scheme for DDMs by selecting the best time steps to maximize the training objective, although the variational lower bound does not correlate well with sample quality, limiting the performance of the method. In a successive work [22], the sampling procedure was directly optimized using a common perceptual evaluation metric (KID) [35], but this required a long training time (30k training iterations). In this work. the authors backpropagate through the sampling chain using reparametrization and gradient rematerialization in order to make the optimization feasible. Our work is closely related to [22], since we similarly backpropagate through the sampling chain. However, we use the MMD [23], [36] to finetune the weights of a pretrained DDM without optimizing the sampling strategy. In essence, the proposed method is complementary to [22]: instead of optimizing the sampling procedure, keeping the model fixed, we directly optimize the model leaving the sampling procedure unchanged. This leads to better results with as few as 500 finetuning iterations. We also remark that our approach is decoupled from the sampling strategy and can be used in conjunction with other training-free acceleration methods such as DDIM.

B. MMD IN GENERATIVE MODELS

The MMD [23], [36] is a distance on the space of probability measures. It is a non-parametric approach that does not make any assumptions about the underlying distributions, and can be used to compare a wide range of distributions. Generative models trained by minimizing the MMD were first considered in [37] and [38]. These works optimized a generator to minimize the MMD with a fixed kernel, but struggled with the complex distribution of natural images where pixel distances are of little value. Successive works [35], [39] addressed this problem by adversarially learning the kernel for the MMD loss, reaching results comparable to GANs trained with a Wasserstein critic. In this work we apply the MMD in the context of diffusion models, demonstrating its effectiveness in finetuning a pretrained DDM under a more restrictive timesteps constraint.

III. METHOD

A. OVERVIEW

We propose MMD-DDM, a technique to accelerate inference in DDMs while maintaining high sample quality, based on finetuning a pretrained diffusion model. The finetuning process minimizes an unbiased estimator of the MMD between real and generated samples, evaluated over a perceptually-relevant feature space. We backpropagate through the sampling process with the aid of the reparametrization trick and gradient checkpointing. This is done only for a small subset of the original timesteps and can be combined with existing techniques for timestep selection or acceleration of the sampling process. The overview of our approach is showed in Figure 2. The reduction in timesteps with respect to the original model degrades the distribution of the generated data. However, the main idea behind the proposed approach is that it is possible to recover part of this degradation by analyzing the generated data in a perceptual feature space and imposing that the reduced DDM produces perceptual features similar to those of real data via MMD minimization. By utilizing this approach, we are thus able to maximize the model performance under a fixed computational budget. It is interesting to notice that older approaches that utilized MMD as sole objective for image generation failed to capture their complex data distribution. On the other hand, our approach avoids that as it leverages the strong baseline provided by the pretrained DDM, albeit degraded by the timesteps constraint.

B. FINETUNING WITH MMD

We are interested in learning a model distribution $p_\theta(\boldsymbol{x}_0)$ that approximates the real data distribution $q(\boldsymbol{x}_0)$. Starting from a pretrained diffusion model, we know from previous work (DDIM [13]) that it is possible to sample from $p_\theta^{(\mathcal{T})}(\boldsymbol{x}_0)$, i.e., the learned distribution using a subset of the original timesteps $\mathcal{T} \subset \{1, \dots, T\}$, accepting a complexity-quality

tradeoff. The MMD [23] is an integral probability metric that we use to measure the discrepancy between the real data distribution $q(\boldsymbol{x}_0)$ and the generated data distribution with the given budget of timesteps $p_\theta^{(\mathcal{T})}(\boldsymbol{x}_0)$. Mathematically, it is defined as:

$$\mathrm{MMD}(p_\theta^{(\mathcal{T})}, q) = \|\mathbb{E}_{\boldsymbol{x}\sim p_\theta^{(\mathcal{T})}}\varphi(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y}\sim q}\varphi(\boldsymbol{y})\| \tag{6}$$

where φ represents a function mapping raw images to a perceptually-meaningful feature space. This is needed as MMD would not perform well on the pixel space, since it is well known that images live on a low-dimensional manifold within the high-dimensional pixel space. However, once the images are mapped into an appropriate feature space, MMD is proven to have strong discriminative performances, as proved by the success of the KID [35] as evaluation metric for perceptual quality. The choice of feature space is critical for the performance of the proposed method and for the fair assessment of methods optimizing quality metrics, which will be presented in Secs. III-C and IV-C.

In order to use the MMD as our loss function, given a batch of generated samples $\{\boldsymbol{x}_i\}_{i=1}^N \sim p_\theta^{(\mathcal{T})}(\boldsymbol{x}_0)$ and a batch of real samples $\{\boldsymbol{y}_i\}_{i=1}^N \sim q(\boldsymbol{x}_0)$, we use the unbiased estimator proposed by Gretton et al. [36]:

$$\begin{aligned}\mathcal{L}_{\mathrm{MMD}^2}^{\mathrm{unbiased}} =& \frac{1}{N(N-1)}\sum_{i\neq j}^{n} k(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)) \\ &- \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{y}_j)) + c\end{aligned} \tag{7}$$

where N is the batch size, c is a constant, and k is a generic positive definite kernel (in our experiments we consider linear, cubic and Gaussian kernels, see Sec. IV-G). The loss function is minimized in order to finetune the values of the parameters θ of a pretrained denoising neural network composing the diffusion model. Next, we are going to discuss the choice of the feature extraction function ϕ.

C. PERCEPTUALLY-RELEVANT FEATURE SPACES

As we previously mentioned, it is necessary to embed real and generated images in some perceptually-relevant feature space, so that the MMD objective could be effective. The feature mapping network ϕ plays a crucial role in the performance of the method. However, this is not a trivial choice. The most popular choice could be to use the feature space of the penultimate layer of an ImageNet-pretrained Inception-V3 classifier [44]. This choice is widely used to evaluate performance of generative models, with Inception Score (IS) [45], FID [46] and KID [35] all using it.

However, a recent study [47] has examined the effectiveness of using ImageNet-pretrained representations to evaluate generative models, and found that the presence of ImageNet classes has a significant impact on the evaluation. The study highlights some potential pitfalls in using these metrics, and how they can be manipulated by the use of ImageNet pretraining. This suggests that care should be taken when using ImageNet features to optimize generative models as this can potentially distort the FID quality metric and make it unreliable. Indeed, for any image generation method, part of the improvement might lie in the *perceptual null space* [47] of FID, which encompasses all the operations that change the FID without affecting the generated images in a perceptible way. For our finetuning procedure, we have experimentally observed a better overall visual quality of generated images and a consistent gain in FID, when optimizing MMD with Inception features. However, it is hard to quantitatively assess how much of this improvement is due to actual perceptual improvements versus optimizations in the perceptual null space. These considerations apply also to the work of Watson et al. [22].

One solution to this problem is to use a different feature space for the feature mapping network, such as one that has not been pretrained on ImageNet. Thus, we propose to optimize the MMD using the feature space of the CLIP image encoder [24], which has been trained in a self-supervised way and is supposed to have richer representations without exposure to ImageNet classes. Moreover, we also consider the case in which we optimize MMD with Inception features and measure performance with a variant of FID using CLIP features. More comments, details, and a discussion of the various results can be found in Sec. IV-C.

IV. EXPERIMENTS

A. SETTING

a: DATASETS

In order to demonstrate the effectiveness of the proposed solution, we validate it on several datasets with different resolutions. We use CIFAR-10 [48] at resolution 32×32, CelebA [49] at resolution 64×64, Image-Net [50] at resolution 64×64, and LSUN-Church [51] at resolution 256×256.

b: MODELS AND SAMPLING

We use the models pretrained by Ho et al. [2] for the CIFAR-10 and LSUN experiments, the model pretrained by Song et al. [13] for CelebA, and the model pretrained by Nichol and Dhariwal [4] with the L_{hybrid} objective for ImageNet. All the architectures are based on the modified UNet [52] that incorporates self-attention layers [53]. We perform our experiments using the efficient sampling strategy of DDIM [13], as it already has good performance in few-timesteps regime. We fix the timestep schedule in the main experiments to be linear. The MMD kernel is polynomial cubic in all experiments, except the kernel ablation one. We also test the proposed solution with the DDPM [2] sampling strategy in Sec. IV-G.

c: EVALUATION

We use the FID [46] to evaluate sample quality. All the values are evaluated by comparing 50k real and generated samples as this is the literature's standard. We also use

TABLE 1. Unconditional CIFAR-10 generative performance (Inception FID).

Method	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 15$	$\lvert\mathcal{T}\rvert = 20$
DDPM [2]	76.3	42.1	31.4	25.9
DDIM [13]	32.7	13.6	9.31	7.50
DDIM + MMD-DDM (Inception-V3)	**5.48**	**3.80**	**4.11**	**3.55**
DDIM + MMD-DDM (CLIP)	6.79	4.87	4.79	4.52
GENIE [34]	13.9	5.97	4.49	3.94
PNDM [40]	35.9	10.3	6.61	5.20
FastDPM [41]	-	9.90	-	5.05
Learned Sampler [22]	13.8	8.22	6.12	4.72
Analytic DDIM [42]	-	14.7	9.16	7.20
DPM-Solver(Type-1) [27]	-	6.37	3.78	4.28
DPM-Solver(Type-2) [27]	-	10.2	4.17	3.72

TABLE 2. Unconditional CelebA generative performance (Inception FID).

Method	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 15$	$\lvert\mathcal{T}\rvert = 20$
DDIM [13]	22.4	17.3	16.0	13.7
DDIM + MMD-DDM (Inception-V3)	**3.04**	**2.58**	**2.13**	**2.24**
DDIM + MMD-DDM (CLIP)	4.65	3.90	3.17	3.27
ES+StyleGAN2+DDIM [17]	9.15	6.44	-	4.90
PNDM [40]	11.3	7.71	-	5.51
FastDPM [41]	-	15.3	-	10.7
Diffusion Autoencoder [43]	-	12.9	-	10.2
Analytic DDPM [42]	-	29.0	21.8	18.1
Analytic DDIM [42]	-	15.6	12.3	10.45
DPM-Solver(Type-1) [27]	-	6.92	3.05	2.82
DPM-Solver(Type-2) [27]	-	5.83	3.11	3.13

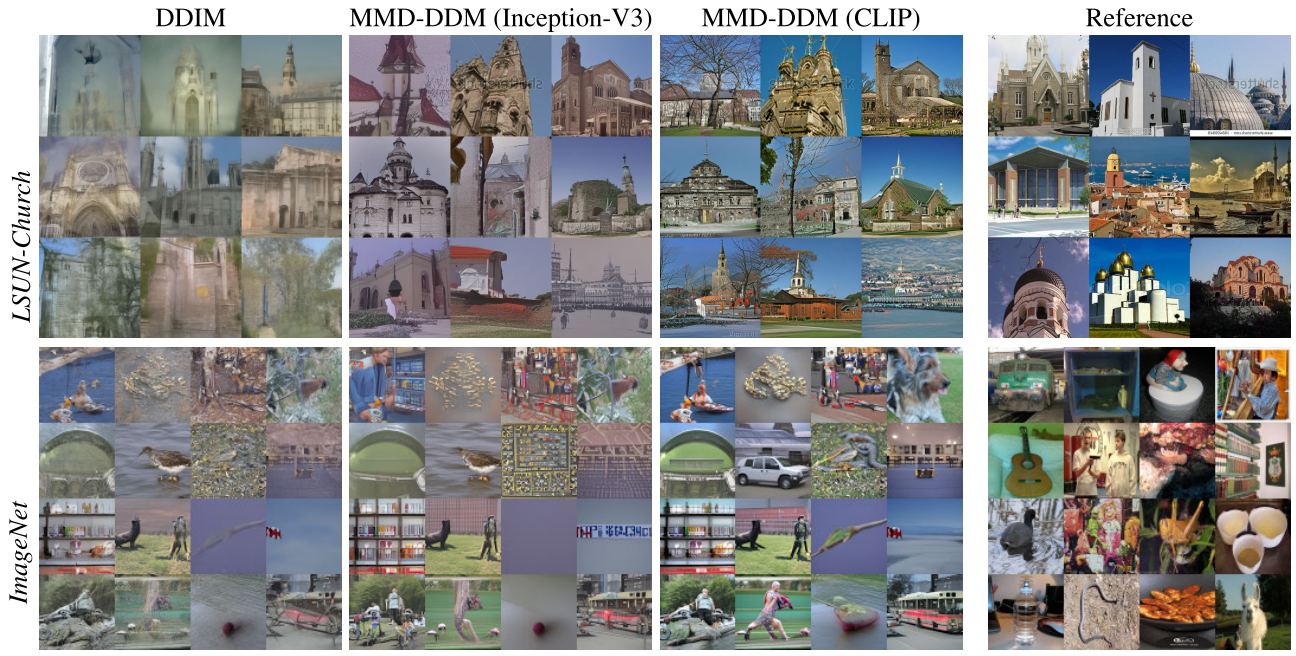

FIGURE 3. Generated samples for LSUN-Church (top) and ImageNet (bottom). The samples are obtained using 5 timesteps for LSUN-Church and 10 timesteps for ImageNet, with the DDIM sampling procedure. Results from Standard DDIM (left), the same model finetuned using Inception-V3 features (center-left) and CLIP features (center-right), reference images from the dataset (right). Samples are not cherry-picked.

FID_{CLIP} [47] in some experiments to remove the effect of Image-Net classes in the evaluation. Additional evaluation metrics such as Inception Score [45], Spatial FID [54], and Precision and Recall [55] can be found in the Supplementary Material.

d: IMPLEMENTATION DETAILS

For all the experiments we set the batch size equal to 128. We use Adam as optimizer [56] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$ and learning rate equal to 5×10^{-6}. When DDIM is used, we set $\sigma_t = 0$. As feature extractors, we use

TABLE 3. Unconditional ImageNet generative performance (Inception FID).

Method	$\|\mathcal{T}\| = 5$	$\|\mathcal{T}\| = 10$	$\|\mathcal{T}\| = 20$
DDIM [13]	131.5	35.2	20.7
DDIM + MMD-DDM (Inc-V3)	33.1	21.1	**12.4**
DDIM + MMD-DDM (CLIP)	**27.5**	**16.4**	14.5
Learned Sampler [22]	55.1	37.2	24.6
Analytic-DDIM [42]	-	70.6	30.9
Analytic-DDPM [42]	-	60.6	37.7
DPM-Solver(T2) [27]	-	24.4	18.53

TABLE 4. Unconditional LSUN-Church Outdoor generative performance (Inception FID).

Method	$\|\mathcal{T}\| = 5$	$\|\mathcal{T}\| = 10$	$\|\mathcal{T}\| = 20$
DDIM [13]	49.6	19.4	12.5
DDIM + MMD-DDM (Inc-V3)	**4.75**	**7.55**	**6.21**
DDIM + MMD-DDM (CLIP)	14.2	10.7	8.82
S-PNDM [40]	20.5	11.8	9.20
F-PNDM [40]	14.8	8.69	9.13

TABLE 5. Comparison of relative improvements evaluating FID in Inception-V3 feature space versus CLIP feature space.

	$\|\mathcal{T}\| = 5$		$\|\mathcal{T}\| = 10$	
CIFAR-10				
	FID	FID_{CLIP}	FID	FID_{CLIP}
DDIM	32.7	13.7	13.6	6.87
DDIM + MMD-DDM (Inception-V3)	5.48	2.11	3.80	2.01
Improvement	-83.2%	-84.4%	-72.0%	-70.7%
CelebA				
	FID	FID_{CLIP}	FID	FID_{CLIP}
DDIM	22.4	12.2	17.3	9.48
DDIM + MMD-DDM (Inception-V3)	3.04	4.94	2.58	4.26
Improvement	-86.4%	-59.3%	-85.0%	-55.0%
ImageNet				
	FID	FID_{CLIP}	FID	FID_{CLIP}
DDIM	131.5	29.5	35.2	11.9
DDIM + MMD-DDM (Inception-V3)	33.1	15.2	21.1	9.21
Improvement	-74.8%	-56.8%	-40.0%	-22.6%

the standard Inception-V3[1] pretrained on Image-Net, and the ViT-B/32[2] model from CLIP [24]. We use *torch-fidelity* [57] for the FID evaluation. We train all the models for about 500 iterations. Finetuning with a budget of 5 timesteps required about 10 minutes for CIFAR-10, about 45 minutes for CelebA, and about one hour for ImageNet on a single Nvidia RTX A6000. For LSUN-Church and for the other timesteps budgets, finetuning has been performed on four Nvidia RTX A6000. Finetuning for 5 timesteps of LSUN-Church required about two hours on the mentioned hardware.

B. IMAGE GENERATION RESULTS

We evaluate MMD-DDM using the following timesteps budgets: $|\mathcal{T}| \in \{5, 10, 15, 20\}$. We report the values of FID on unconditional generation experiments for CIFAR-10 in Table 1, for CelebA in Table 2, for ImageNet in Table 3 and for LSUN-Church in Table 4.

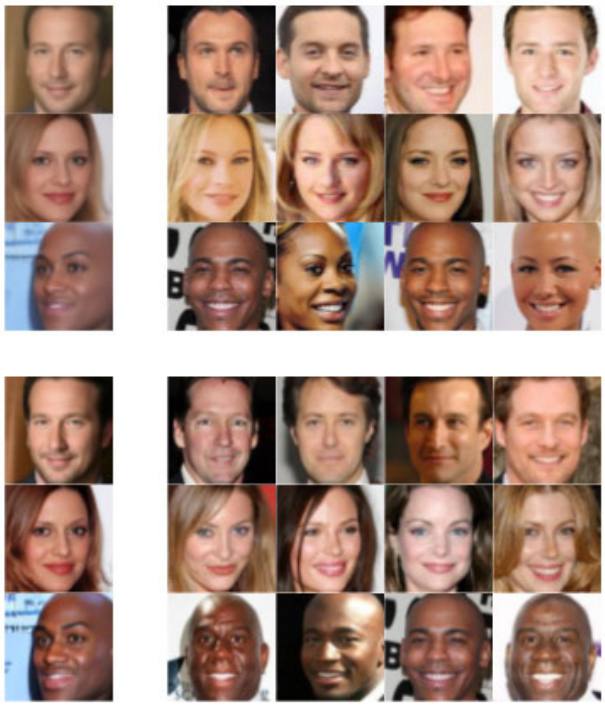

FIGURE 4. Generated samples by the DDIM model (top) and the finetuned model (bottom) for CelebA. For each generated samples we visualize the top-4 nearest neighbours.

We compare against several state-of-the-art methods for accelerating DDMs. The tables report the results for MMD-DDM trained with Inception features and we also report results taken from literature for other methods. For all datasets and timesteps budgets, MMD-DDM provides superior or, occasionally, comparable quality to state-of-the art approaches. For the ImageNet experiment, we remark that we report the result of the Learned Sampler approach [22], which uses an improved version of the model from [4] trained for 3M iterations, instead of the 1.5M iterations used by our checkpoint, thus making the comparison slightly unfavourable for our method. We do not compare with the progressive distillation method [19], as it cannot be considered a post-training acceleration technique but rather a very computationally-demanding modification of the DDM training procedure.

Qualitative comparisons for CIFAR-10 and CelebA are shown in Fig. 1 and for LSUN-Church and ImageNet in Fig. 3. More generated samples, for different numbers of timesteps, can be found in the Supplementary Material. It can be noticed that MMD-DDM provides substantial improvements in visual quality when the number of timesteps is highly constrained. As the available timesteps budget is relaxed to 20 or more, the improvement provided MMD-DDM diminishes, although all approaches start providing high quality samples.

[1] http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz

[2] https://github.com/openai/CLIP

C. FEATURE SPACE DISCUSSION

Results in the previous section were presented with the commonly-used FID metric exploiting Inception features. However, as detailed in Sec. IV-C, our optimization of Inception features via the MMD loss could raise concerns about the reliability of the FID metric. In this section, we present results using the CLIP feature space in either the MMD loss or the FID metric.

Figs. 1 and 3 already show a visual comparison between using the MMD with Inception features and CLIP features and more results are present in the Supplementary Material. It can be noticed that optimizing over CLIP features leads to higher visual quality, including sharper details and clarity, confirming that the CLIP space is a superior embedding of perceptually-relevant features. As a reference, we also report the FID scores obtained by MMD-DDM with CLIP features in Tables 1,2,3,4. Notice that lower values are observed, possibly due to the reliance of the FID on flawed Inception features and the metric not accurately tracking a genuine improvement in visual quality.

We further expand the set of results in Table 5, in which we optimize Inception features with the MMD but then measure quality using the FID computed on CLIP features (FID_{CLIP}), as proposed in [47]. Since the feature space used for evaluation is different from the one used in optimization, the observed gains in FID_{CLIP} suggest us that the quality improvement in quantitatively meaningful and not just an artifact of the metric. Percentage improvements in FID_{CLIP} mostly track those of regular FID, albeit being lower in some cases, suggesting that some overfitting of the perceptual null space does indeed happen when Inception features are used for both MMD and FID.

D. ANALYSIS OF OVERFITTING

One might wonder whether finetuning via the MMD loss leads to images overfitting the features of the training set. This section presents a experiment to dispel this concern. To do so, we looked at the top-K nearest neighbors of generated samples when the CLIP feature space is used for both the optimization with the MMD and the space for nearest neighbor search (Euclidean distance betweeen CLIP features). Fig. 4 provides the results of this experiment for samples generated with both the pretrained model and the finetuned model. We can see that the nearest neighbors of the samples generated after finetuning are not more significantly similar to the generated image than those for the pretrained model. More samples can be found in the Supplementary material.

E. ADDITIONAL EVALUATION METRICS

We report additional evaluation metrics computed on CIFAR-10. The Inception Score [45], the Spatial FID [54], i.e. the FID evaluated using the first 7 channels from the intermediate *mixed_6/conv* feature maps, and the Precision and Recall metrics [55]. In particular Recall is used to evaluate the diversity of generated samples. We denote an overall improvement on all the metrics for both our proposed solutions. The results are showed in Table 6. Interestingly, our proposed finetuning increases the diversity of the generated samples leading to higher recall scores. [3]

TABLE 6. Unconditional CIFAR-10 generative performance over different metrics.

Method	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 15$	$\lvert\mathcal{T}\rvert = 20$
Inception Score				
DDIM [13]	6.95	8.17	8.43	8.54
MMD-DDM (Inception-V3)	**9.39**	**9.91**	**9.85**	**9.94**
MMD-DDM (CLIP)	9.10	9.38	9.52	9.19
sFID				
DDIM [13]	22.2	10.73	9.35	7.36
MMD-DDM (Inception-V3)	**7.20**	6.09	**5.75**	5.46
MMD-DDM (CLIP)	8.25	**6.07**	6.04	**4.90**
Precision				
DDIM [13]	0.51	0.60	0.61	0.63
MMD-DDM (Inception-V3)	0.63	0.65	0.65	0.67
MMD-DDM (CLIP)	**0.65**	**0.67**	**0.67**	**0.68**
Recall				
DDIM [13]	0.31	0.46	0.50	0.52
MMD-DDM (Inception-V3)	**0.56**	**0.59**	**0.60**	**0.58**
MMD-DDM (CLIP)	0.53	0.56	0.57	0.57

F. LATENT SPACE INTERPOLATION

We create interpolations (Fig. 5) on a line by selecting two random x_T values from a standard Gaussian distribution, using them in a spherical linear interpolation method and then applying the DDIM sampling [2] to generate x_0 samples.

$$\boldsymbol{x}_T^{(\alpha)} = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)}\boldsymbol{x}_T^{(0)} + \frac{\sin(\alpha\theta)}{\sin(\theta)}\boldsymbol{x}_T^{(1)} \tag{8}$$

where $\theta = \arccos\left(\frac{(\boldsymbol{x}_T^{(0)})^\top \boldsymbol{x}_T^{(1)}}{\|\boldsymbol{x}_T^{(0)}\|\|\boldsymbol{x}_T^{(1)}\|}\right)$.

G. ABLATION STUDIES

In this section we consider how the choice of MMD kernel, timestep scheduling and sampling process affect the performance of the proposed method. In all the experiments, unless otherwise specified, we use the DDIM sampling procedure and the Inception-V3 feature space. First, we ablate the choice of the kernel for the MMD loss by comparing three different kernels: the linear kernel $k^{\text{lin}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\top\boldsymbol{y}$, the polynomial cubic kernel $k^{\text{cub}}(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{1}{d}\boldsymbol{x}^\top\boldsymbol{y}+1\right)^3$ [35] and the Gaussian RBF kernel $k^{\text{rbf}}(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$. Table 7 reports the results for different kernels in terms of FID, showing a marginal preference for the cubic kernel and overall robustness of MMD-DDM to kernel choice.

Next, we ablate the influence of timesteps selection in Table 8. We consider the two commonly-used alternatives to select $\mathcal{T}$: *linear* $\tau_i = \lfloor ci \rfloor$, and *quadratic* $\tau_i = \lfloor ci^2 \rfloor$, where c is selected to make $\tau_1 \approx T$. This experiment does not show a preference for either selection method. However, it is possible

[3]The values are obtained using OpenAI evaluator that can be found at: https://github.com/openai/guided-diffusion/tree/main/evaluations

FIGURE 5. Generated interpolations by the DDIM model (first row) and the Inception finetuned model (second row) and the CLIP finetuned model (third row) for LSUN-Church The samples are generated using 5 timesteps.

TABLE 7. Ablation study for the kernel choice - CIFAR-10.

Kernel	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 20$
Linear	5.61	4.69	4.06
Gaussian RBF	5.89	3.88	3.61
Cubic	**5.48**	**3.80**	**3.55**

TABLE 8. Ablation study for the timestep schedule - CIFAR10.

Selection Method	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 20$
Linear	5.48	3.80	**3.55**
Quadratic	**5.19**	3.80	3.67

TABLE 9. Ablation study for the sampling procedure - CIFAR10.

Sampling	$\lvert\mathcal{T}\rvert = 5$	$\lvert\mathcal{T}\rvert = 10$	$\lvert\mathcal{T}\rvert = 20$
DDPM [2]	76.3	42.1	25.9
DDIM [13]	32.7	13.6	7.50
DDPM+MMD-DDM	6.65	5.19	4.48
DDIM+MMD-DDM	**5.48**	**3.80**	**3.55**

that other subset selection strategies such as grid search [34] or learning the optimal timesteps [22] could further improve results. We remark that MMD-DDM is decoupled from the specific timesteps selection technique.

Finally, we also test MMD-DDM with the DDPM sampling procedure, instead of DDIM. Results are reported in Table 9. As expected, the DDIM sampling procedure is more powerful and produces better results with a low number of timesteps. However, we notice that MMD-DDM produces significant improvements even when applied to DDPM.

V. CONCLUSION AND DISCUSSION

This paper addressed the problem of inference speed of DDMs. We showed that finetuning a DDM with a constraint on the number of timesteps using the MMD loss provides substantial improvements in visual quality. The limited computational complexity of the finetuning procedure offers a way to quickly obtain an improved tradeoff between inference speed and visual quality for a wide range of DDM designs. A limitation of the current technique lies in the memory requirements when the finetuning needs to be performed over a larger number of timesteps, although gradient checkpoint partially addresses this issue in most practical settings. Furthermore, coupling MMD-DDM with more advanced timestep selection and optimization techniques, possibly via joint optimization, could represent an interesting avenue to further improve speed-quality tradeoffs. Integration with conditional DDMs could also represent a direction for future work.

VI. LIMITATIONS AND FUTURE WORK

MMD-DDM significantly enhances the speed-quality trade-off in denoising diffusion models. However, it still faces certain limitations. One major challenge is maintaining the highest sample quality at extremely low timestep counts. Our experiments indicate that while our method performs exceptionally well with fewer timesteps, the quality improvements diminish as the number of timesteps increases. Another limitation is that our approach relies on pretrained diffusion models, meaning its performance is contingent on the quality and characteristics of the base model. Variations in the base model's architecture or training data can influence the effectiveness of MMD-DDM's finetuning process. For future work, several areas could be explored to overcome these limitations. Advanced techniques for timestep selection and optimization, possibly through joint optimization methods, could further enhance the speed-quality trade-offs. Techniques like grid search or learning optimal timesteps could be beneficial. Extending MMD-DDM to conditional diffusion models could broaden its applicability and improve its performance in specific tasks such as image-to-image translation and super-resolution. Investigating alternative or complementary loss functions to MMD could potentially enhance the model's performance. For instance, incorporating perceptual losses might yield better sample quality and diversity. Additionally, testing and optimizing MMD-DDM on larger and more diverse datasets

FIGURE 6. Generated samples for ImageNet. The samples are obtained using 5, 10 and 20 timesteps with the DDIM sampling procedure. Results from Standard DDIM (left), the same model finetuned using Inception-V3 features (center) and CLIP features (right).

could provide insights into its scalability and robustness. By addressing these limitations and exploring the suggested future directions, we aim to enhance the practicality and performance of MMD-DDM in various applications of denoising diffusion models.

APPENDIX

A. IMAGE GENERATION RESULTS ON IMAGENET

In Fig. 6, we report visualizations of generated images for the ImageNet dataset with several timesteps choices.

REFERENCES

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2020, pp. 6840–6851.

[3] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–20.

[4] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 8162–8171.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[6] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," 2020, *arXiv:2009.00713*.

[7] L. Zhou, Y. Du, and J. Wu, "3D shape generation and completion through point-voxel diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5806–5815.

[8] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "LION: Latent point diffusion models for 3D shape generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–12.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–15.

[10] W. Zhang et al., "GACNet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5503314, doi: 10.1109/TGRS.2023.3347745.

[11] W. Zhang, W. Zhao, J. Li, P. Zhuang, H. Sun, Y. Xu, and C. Li, "CVANet: Cascaded visual attention network for single image super-resolution," *Neural Netw.*, vol. 170, pp. 622–634, Feb. 2024.

[12] W. Zhao, C. Li, W. Zhang, L. Yang, P. Zhuang, L. Li, K. Fan, and H. Yang, "Embedding global contrastive and local location in self-supervised learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2275–2289, May 2023.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–30.

[15] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," 2021, *arXiv:2105.14080*.

[16] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," 2022, *arXiv:2206.00364*.

[17] Z. Lyu, X. Xu, C. Yang, D. Lin, and B. Dai, "Accelerating diffusion models via early stop of the diffusion process," 2022, *arXiv:2205.12524*.

[18] H. Zheng, P. He, W. Chen, and M. Zhou, "Truncated diffusion probabilistic models," *Stat*, vol. 1050, p. 7, Jun. 2022.

[19] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–23.

[20] E. Luhman and T. Luhman, "Knowledge distillation in iterative generative models for improved sampling speed," 2021, *arXiv:2101.02388*.

[21] D. Watson, J. Ho, M. Norouzi, and W. Chan, "Learning to efficiently sample from diffusion probabilistic models," 2021, *arXiv:2106.03802*.

[22] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–15.

[23] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–20.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[25] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Comput.*, vol. 23, no. 7, pp. 1661–1674, Jul. 2011.

[26] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.

[27] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," 2022, *arXiv:2206.00927*.

[28] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," 2022, *arXiv:2204.13902*.

[29] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, S. Özturk, A. Gungör, and T. Çukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3524–3539, Sep. 2023.

[30] A. Gungör, S. U. Dar, S. Özturk, Y. Korkmaz, H. A. Bedel, G. Elmas, M. Ozbey, and T. Çukur, "Adaptive diffusion priors for accelerated MRI reconstruction," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102872.

[31] H. Atakan Bedel and T. Çukur, "DreaMR: Diffusion-driven counterfactual explanation for functional MRI," 2023, *arXiv:2307.09547*.

[32] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.-Y. Liu, "EDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," 2022, *arXiv:2211.01324*.

[33] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," 2023, *arXiv:2303.01469*.

[34] T. Dockhorn, A. Vahdat, and K. Kreis, "GENIE: Higher-order denoising diffusion solvers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–20.

[35] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–24.

[36] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[37] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1718–1727.

[38] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani, "Training generative neural networks via maximum mean discrepancy optimization," in *Proc. 31st Conf. Uncertainty Artif. Intell.*, 2015, pp. 258–267.

[39] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "MMD GAN: Towards deeper understanding of moment matching network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–18.

[40] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–15.

[41] Z. Kong and W. Ping, "On fast sampling of diffusion probabilistic models," in *Proc. ICML Workshop Invertible Neural Netw., Normalizing Flows, Explicit Likelihood Models*, 2021, pp. 1–17.

[42] F. Bao, C. Li, J. Zhu, and B. Zhang, "Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–33.

[43] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10619–10629.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[45] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image GANs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1–36.

[46] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–19.

[47] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of ImageNet classes in Fréchet inception distance," 2022, *arXiv:2203.06026*.

[48] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, 2009.

[49] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[51] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–6.

[54] C. Nash, J. Menick, S. Dieleman, and P. Battaglia, "Generating images with sparse representations," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 7958–7968.

[55] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–19.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–11.

[57] A. Obukhov, M. Seitzer, P.-W. Wu, S. Zhydenko, J. Kyl, and E. Y.-J. Lin, "High-fidelity performance metrics for generative models in PyTorch," 2020. [Online]. Available: https://torch-fidelity.readthedocs.io/en/latest/

EMANUELE AIELLO received the bachelor's degree (Hons.) in electronics and communications engineering and the master's degree (Hons.) in telecommunications from the Politecnico di Torino, where he is currently pursuing the Ph.D. degree, specializing in artificial intelligence. He has gained practical experience through prestigious internships, as a Research Scientist Intern with Meta. He is also a Teaching Assistant with the Politecnico di Torino. His research focusing on multimodal deep learning.

DIEGO VALSESIA (Member, IEEE) received the Ph.D. degree in electronic and communication engineering from the Politecnico di Torino, in 2016. He is currently an Assistant Professor with the Department of Electronics and Telecommunications (DET), Politecnico di Torino. His main research interests include processing of remote sensing images and deep learning for inverse problems in imaging. He is a member of the EURASIP Technical Area Committee for Signal and Data Analytics for Machine Learning and the ELLIS Society. He was a recipient of the IEEE ICIP 2019 Best Paper Award and the IEEE Multimedia 2019 Best Paper Award. He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, for which he received the 2023 Outstanding Editorial Board Member Award.

ENRICO MAGLI (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the Politecnico di Torino, Italy, in 1997 and 2001, respectively. He is currently a Full Professor with the Politecnico di Torino, Italy, where he leads the Image Processing and Learning Group, performing research in the fields of deep learning for image and video processing, image compression and image forensic for multimedia, and remote sensing applicaitons. He is a fellow of the ELLIS Society for the Advancement of Artificial Intelligence, Europe. He was a recipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award, the IEEE ICIP 2015 Best Student Paper Award (as a senior author), the IEEE ICIP 2019 Best Paper Award, the IEEE Multimedia 2019 Best Paper Award, and the 2010 and 2014 Best Associate Editor Award of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *EURASIP Journal on Image and Video Processing*. He was an IEEE Distinguished Lecturer, from 2015 to 2016.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement