# FMBoost: Boosting Latent Diffusion with Flow Matching

Johannes S. Fischer*, Ming Gui*, Pingchuan Ma*, Nick Stracke, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer

CompVis @ LMU Munich & MCML
https://compvis.github.io/fm-boosting

**Abstract.** Visual synthesis has recently seen significant leaps in performance, largely due to breakthroughs in generative models. Diffusion models have been a key enabler, as they excel in image diversity. However, this comes at the cost of slow training and synthesis, which is only partially alleviated by latent diffusion. To this end, flow matching is an appealing approach due to its complementary characteristics of faster training and inference but less diverse synthesis. We demonstrate our *FMBoost* approach, which introduces flow matching between a frozen diffusion model and a convolutional decoder that enables high-resolution image synthesis at reduced computational cost and model size. A small diffusion model can then effectively provide the necessary visual diversity, while flow matching efficiently enhances resolution and detail by mapping the small to a high-dimensional latent space, producing high-resolution images. Combining the diversity of diffusion models, the efficiency of flow matching, and the effectiveness of convolutional decoders, state-of-the-art high-resolution image synthesis is achieved at $1024^2$ pixels with minimal computational cost. Cascading FMBoost optionally boosts this further to $2048^2$ pixels. Importantly, this approach is orthogonal to recent approximation and speed-up strategies for the underlying model, making it easily integrable into the various diffusion model frameworks.

**Keywords:** Generative Models · Diffusion Models · Flow Matching

## 1 Introduction

Visual synthesis has recently witnessed unprecedented progress and popularity in computer vision and beyond. Various generative models have been proposed to address the diverse challenges in this field [78], including sample diversity, quality, resolution, training, and test speed. Among these approaches, diffusion models (DMs) [58, 59, 61] currently rank among the most popular and highest quality, defining the state of the art in numerous synthesis applications. While DMs excel in sample quality and diversity, they face challenges in high-resolution synthesis, slow sampling speed, and a substantial memory footprint.

* Equal contribution

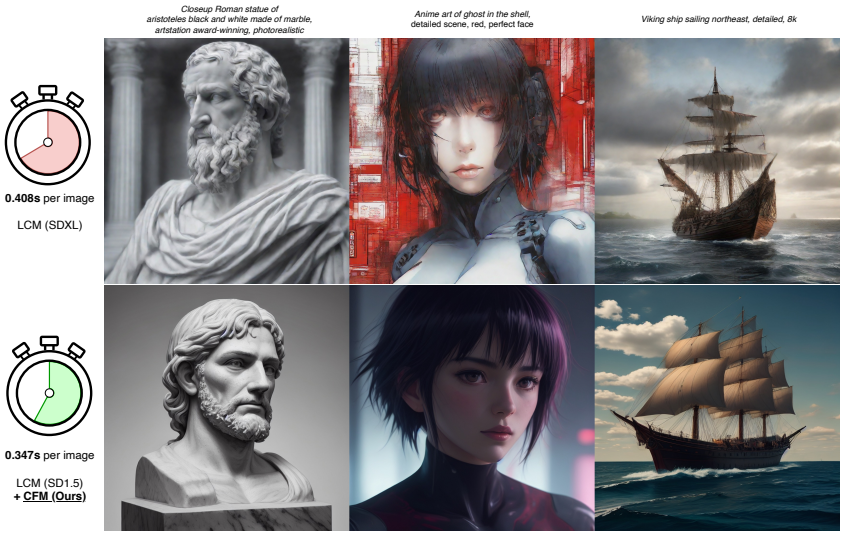



**Fig. 1: Samples synthesized in $1024^2$ px.** FMBoost elevates Diffusion Models (DMs) and similar architectures to a higher-resolution domain, achieving exceptionally rapid processing speeds. We use Latent Consistency Models (LCM) [48], distilled from SD1.5 [59] and SDXL [55], respectively. To achieve the same resolution as LCM-SDXL, we boost LCM-SD1.5 with our approach. The LCM-SDXL model fails to produce competitive results within this shortened timeframe, highlighting the effectiveness of our approach in achieving both speed and quality in image synthesis.

Lately, numerous efficiency improvements to DMs have been proposed [12, 57, 68], but the most popular remedy has been the introduction of Latent Diffusion Models (LDMs) [59]. Operating only in a compact latent space, LDMs combine the strengths of DMs with the efficiency of a convolutional encoder-decoder that translates the latents back into pixel space. However, Rombach et al. [59] also showed that an excessively strong first-stage compression leads to information loss, limiting generation quality. Efforts have been made to expand the latent space [55] or stack a series of different DMs, each specializing in different resolutions [20, 61]. Nevertheless, these approaches are still computationally costly, especially when synthesizing high-resolution images.

The inherent stochasticity of DMs is key to their proficiency in generating diverse images. In the later stages of DM inference, as the global structure of the image has already been generated, the advantages of stochasticity diminish. Instead, the computational overhead due to the less efficient stochastic diffusion trajectories becomes a burden rather than helping in up-sampling to and improving higher resolution images [8]. At this stage, converse characteristics become beneficial: reduced diversity and a short and straight trajectory toward

the high-resolution latent space of the decoder. These goals align precisely with the strengths of Flow Matching (FM) [4, 41, 44], another emerging family of generative models currently gaining significant attention. In contrast to DMs, Flow Matching enables the modeling of an optimal transport conditional probability path between two distributions that is significantly straighter than those achieved by DMs, making it more robust, and efficient to train. The deterministic nature of Flow Matching models also allows the utilization of off-the-shelf Ordinary Differential Equation (ODE) solvers, which are more efficient to sample from and can further accelerate inference.

Therefore, we propose *FMBoost* to leverage the complementary strengths of DMs, FMs, and VAEs: the diversity of stochastic DMs, the speed of Flow Matching in training and inference stages, and the efficiency of a convolutional decoder when mapping latents into pixel space. This synergy results in a small diffusion model that excels in generating diverse samples at a low resolution. Flow Matching then takes a direct path from this lower-resolution representation to a higher-resolution latent, which is subsequently translated into a high-resolution image by a convolutional decoder. Moreover, the Flow Matching model can establish data-dependent couplings with the synthesized information from the DM, which automatically and inherently forms optimal transport paths from the noise to the data samples in the Flow Matching model [5, 72].

Note that our work is complementary to recent work on sampling acceleration of diffusion models like DDIM [68], DPM-Solver [46], and LCM-LoRA [47, 48]. FMBoost can be directly integrated into any existing DM architecture to increase the final output resolution efficiently.

## 2 Related Work

**Diffusion Models** Diffusion models [19, 67, 70] have shown broad applications in computer vision, spanning image [59], audio [42], and video [9, 22]. Albeit with high fidelity in generation, they do so at the cost of sampling speed compared to alternatives like Generative Adversarial Networks [16, 28, 31]. Hence, several works propose more efficient sampling techniques for diffusion models, including distillation [49, 63, 69], noise schedule design [33, 52, 56], and training-free sampling [30, 43, 45, 68]. Nonetheless, it is important to highlight that existing methods have not fully addressed the challenge imposed by the strong curvature in the sampling trajectory [36], which limits sampling step sizes and necessitates the utilization of intricately tuned solvers, making sampling costly.

**Flow Matching-based Generative Models** A recent competitor, known as Flow Matching [6, 41, 44, 50], has gained prominence for its ability to maintain straight trajectories during generation by modeling the synthesis process using an optimal transport conditional probability path with Ordinary Differential Equations (ODE), positioning it as an apt alternative for addressing trajectory straightness-related issues encountered in diffusion models. The versatility of Flow Matching has been showcased across various domains, including im-

age [13, 23, 24, 41], video [7], audio [34], depth [17], and even text [25]. This underscores its capacity to address the inherent trajectory challenges associated with diffusion models, mitigating the limitations of slow sampling in the current generation based on diffusion models. Considerable effort has been directed towards optimizing transport within Flow Matching models [44, 72], which contributes to enhanced training stability and accelerated inference speed by making the trajectories even straighter and thus enabling larger sampling step sizes. However, the generation capabilities of Flow Matching presently do not parallel those of diffusion models [13, 41]. We remedy this limitation by using a small diffusion model for synthesis quality.

**Image Super-Resolution** Image super-resolution (SR) is a fundamental problem in computer vision. Prominent methodologies include GANs [28, 35, 75, 81], diffusion models [37, 62, 80] and Flow Matching methods [5, 41].

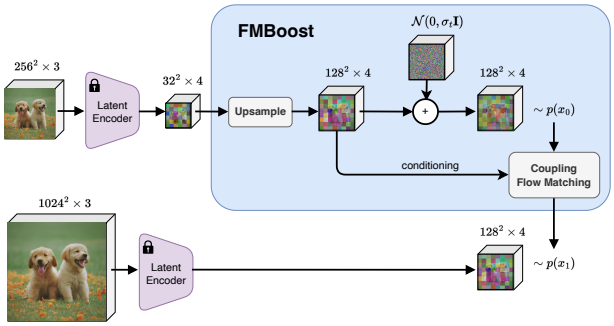
FMBoost adopts the Flow Matching approach, leveraging its objective to achieve faster training and inference compared to diffusion models. We take inspiration from latent diffusion models [59] and transition the training to the latent space, which further enhances computational efficiency. This enables the synthesis of images with significantly higher resolution, thereby advancing the capacity for image generation in terms of both speed and output resolution.

## 3 Method

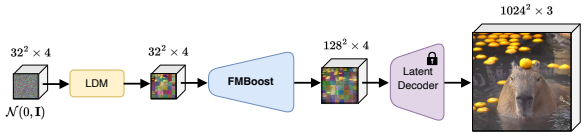
We speed up and increase the resolution of existing LDMs by integrating Flow Matching in the latent space. The proposed architecture should not be limited to unconditional image synthesis but also be applicable to text-to-image synthesis [51, 58, 59, 61] and Diffusion models with other conditioning including depth maps, canny edges, etc. [14, 38, 82]. The main challenge is not a deficiency in diversity within the Diffusion model; rather it is the slow convergence of the training procedure, the huge memory demand, and the slow inference [55, 61, 79]. While there are substantial efforts to accelerate inference speed of DMs either by distillation techniques [49], or by an ODE approximation at inference [45, 46, 68], we argue that we can achieve faster training and inference speed by training with an ODE assumption [41]. Flows characterized by straight paths without Wiener process inherently incur minimal time-discretization errors during numerical simulation [44] and can be simulated with only a few ODE solver steps. FMBoost employs a compact Diffusion model and a Flow Matching model aimed at high-resolution image generation (Sec. 3.1, Sec. 3.2). The combination of both models (Sec. 3.3) ensures efficient and detailed image generation.

### 3.1 From LDM to FM-LDM

Diffusion Models (DMs) [19] are generative models that learn a data distribution $p(x)$ by learning to denoise noisy samples. During inference, they generate samples in a multi-step denoising process starting from Gaussian noise. Their




**(a)** Training.




**(b)** Inference.

**Fig. 2:** Approach overview. **a)** During training we feed both a low- and a high-res image through the pre-trained encoder to their respective latent code. Based on the concatenated low-res latent code and a noisy version of it, we model a vector field within $t \in [0, 1]$. **b)** During inference we can take any LDM, generate the low-res latent, and then use our coupling flow matching model to synthesize the higher dimensional latent code. Finally, the pre-trained decoder projects the latent code back to pixel space.

inherent stochasticity allows them to effectively approximate the data manifold with high diversity, even in high-dimensional complex data domains such as images [52, 61] or videos [9, 22, 65], but makes generation inefficient, requiring many denoising steps at the data resolution. This problem has previously partially been addressed by *Latent Diffusion Models* (LDMs), which move the diffusion process to an autoencoder latent space, but efficiency is still a problem. While diffusion models' stochasticity helps them generate high-quality samples, we propose that this stochasticity is not needed for later stages of generation and that the diffusion generation process can be separated into two parts without substantial loss in quality: one diffusion-based low-resolution stage for generating image semantics with high variation and a light-weight high-resolution stage with reduced stochasticity. Recently, the formulation of generative processes as optimal transport conditional probability paths has gained much attraction [6, 41, 72], perfectly suiting this task of modeling straight trajectories between two distributions.

### 3.2 Flow Matching

Flow Matching models are generative models that regress vector fields based on fixed conditional probability paths. Let $\mathbb{R}^d$ be the data space with data points $x$. Let $u_t(x) : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ be the time-dependent vector field, which defines

the ODE in the form of $dx = u_t(x)dt$, and let $\phi_t(x)$ denote the solution to this ODE with the initial condition $\phi_0(x) = x$.

The probability density path $p_t : [0, 1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$ depicts the probability distribution of $x$ at timestep $t$ with $\int p_t(x)dx = 1$. The pushforward function $p_t = [\phi_t]_{\#}(p_0)$ then transports the probability density path $p$ along $u$ from timestep 0 to $t$. Assuming that $p_t(x)$ and $u_t(x)$ are known, and the vector field $u_t(x)$ generates $p_t(x)$, we can regress a vector field $v_\theta(t, x)$ parameterized by a neural network with learnable parameters $\theta$ using the Flow Matching objective

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t(x)}||v_\theta(t, x) - u_t(x)||. \tag{1}$$

While we generally do not have access to a closed form of $u_t$ because this objective is intractable, Lipman et al. [41] showed that we can acquire the same gradients and therefore efficiently regress the neural network using the coupling Flow Matching (CFM) objective, where we can compute $u_t(x|z)$ by efficiently sampling $p_t(x|z)$,

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,q(z),p_t(x|z)}||v_\theta(t, x) - u_t(x|z)||, \tag{2}$$

with $z$ as a conditioning variable and $q(z)$ the distribution of that variable. We parameterize $v_\theta$ as a U-Net [60], which takes the data sample $x$ as input and $z$ as conditioning information.

**Naïve Flow Matching** We first assume that the probability density path bridges the a random Gaussian distribution and the data distribution $p_1$, and let the corresponding sample from that distribution be $x_1$. We smooth the data sample with a Gaussian with minimal variance $\mathcal{N}(x_1, \sigma^2_{\min})$ following [41]. Given the conditioning $z := x_1$, the transportation path can then be formulated as follows:

$$p_t(x|z) = \mathcal{N}(x|tx_1, (t\sigma_{\min} - t + 1)^2\mathbf{I}), \tag{3}$$

$$u_t(x|z) = \frac{x_1 - (1 - \sigma_{\min})x}{1 - (1 - \sigma_{\min})t}; \quad \phi_t(x|z) = (1 - (1 - \sigma_{\min})t)x + tx_1. \tag{4}$$

The resulting FM loss takes the form of

$$\begin{aligned} \mathcal{L}_{FM}(\theta) &= \mathbb{E}_{t,z,p_t(x|z)}||v_\theta(t, \phi_t(x_0)) - \frac{d}{dt}\phi_t(x_0)|| \\ &= \mathbb{E}_{t,z,p(x_0)}||v_\theta(t, \phi_t(x_0)) - (x_1 - (1 - \sigma_{\min})x_0)||. \end{aligned} \tag{5}$$

**Data-Dependent Couplings** In our case, we also have access to the representation of a low-resolution image generated by a DM at inference time. It seems intuitive to incorporate the inherent relationship between the conditioning signal and our target within the Flow Matching objective, as is also stated in [5]. Let $x_1$ denote a high-resolution image. Instead of randomly sampling from a Gaussian distribution in the naïve Flow Matching method, the starting point $x_0 = \mathcal{E}(x_1)$ corresponds to an encoded representation of the image, with $\mathcal{E}$ being

a fixed encoder. The conditioning $z$ consequently encompasses information from both $x_0$ and $x_1$.

Similar to the previously described case, we smooth around the data samples within a minimal variance to acquire the corresponding data distribution $\mathcal{N}(x_0, \sigma^2_{\text{min}})$ and $\mathcal{N}(x_1, \sigma^2_{\text{min}})$. The Gaussian flows can be defined by the equations

$$p_t(x|z) = \mathcal{N}(x|tx_1 + (1-t)x_0, \sigma^2_{\text{min}}\mathbf{I}), \tag{6}$$

$$u_t(x|z) = x_1 - x_0; \quad \phi_t(x|z) = tx_1 + (1-t)x_0. \tag{7}$$

Notably, the optimal transport condition between the probability distributions $p_0(x|z)$ and $p_1(x|z)$ is inherently satisfied due to the data coupling. This automatically solves the dynamic optimal transport problem in the transition from low to high resolution within the Flow Matching paradigm and enables more stable and faster training [72]. We name these Flow Matching models with data-dependent couplings *Coupling Flow Matching* (CFM) models, and the CFM loss then takes the form of

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,z,p(x_0)}||v_\theta(t, \phi_t(x_0), x_0) - (x_1 - x_0)||. \tag{8}$$

**Noise Augmentation** Noise Augmentation is a technique for boosting generative models' performance introduced for cascaded Diffusion models [20]. The authors found that applying random Gaussian noise or Gaussian blur to the conditioning signal in super-resolution Diffusion models results in higher-quality results during inference. Drawing inspiration from this, we also implement Gaussian noise augmentation on $x_0$. Following variance-preserving DMs, we noise $x_0$ according to the cosine schedule first proposed in [52]. In line with [20], we empirically discover that incorporating a specific amount of Gaussian noise enhances performance. We hypothesize that including a small amount of Gaussian noise smoothes the base probability density $p_0$ so that it remains well-defined over the higher-dimensional space. Note that this noise augmentation is only applied to $x_0$ but not to the conditioning information $z$, since the model relies on the precise conditioning information to construct the straight path.

**Latent Flow Matching** In order to reduce the computational demands associated with training FM models for high-resolution image synthesis, we take inspiration from [59] and utilize an autoencoder model that provides a compressed latent space that aligns perceptually with the image pixel space similar to LDMs. By training in the latent space, we get a two-fold advantage: i) The computational cost associated with the training of flow-matching models is reduced substantially, thereby enhancing the overall training efficiency. ii) Leveraging the latent space unlocks the potential to synthesize images with significantly increased resolution efficiently and with a faster inference speed. On the other hand, our method learns a path from low-resolution to high-resolution images in comparison to Dao et al. [13], which starts from noise. This optimization accelerates training and enables sampling at a significantly higher efficiency, as discussed in Section 4.5

### 3.3 High-Resolution Image Synthesis

Our *FMBoost* approach integrates all aforementioned components into a cohesive pipeline, as depicted in detail in Fig. 2. We start from a DM for content synthesis and move the generation to a latent space with a pretrained VAE encoder, which optimizes memory usage and enhances inference speed. To further alleviate the computational load of the DM and achieve additional acceleration, we adopt a compact DM to produce compressed information. Subsequently, the FM model maps the compressed information to a high-resolution latent image with a straight conditional probability path. Finally, we decompress the latent space using a pre-trained VAE decoder. Note that the VAE decoder performs well across various resolutions, we show addtional proof in the appendix.

The integration of FM with DMs in the latent space presents a promising approach to address the trade-off between flexibility and efficiency in modeling the dynamic image synthesis process. The inherent stochasticity within a DM's sampling process allows for a more nuanced representation of complex phenomena, while the FM model exhibits greater computational efficiency, which is useful when handling high-resolution images, but lower flexibility and image fidelity as of yet when it comes to image synthesis [41]. By combining them in the pipeline, we benefit from the flexibility of the DM while capitalizing on the efficiency of FM as well as a VAE.

## 4 Experiments

### 4.1 Metrics and datasets

For quantitative evaluation, we use the standard Fréchet Inception Distance (FID) [18], SSIM [77], and PSNR to measure the realism of the output distribution and the quality of the image. The general dataset we use for initial experiments and ablations is FacesHQ, a compilation of CelebA-HQ [29] and FFHQ [31], as used in previous work [15,62] for high-resolution synthesis tasks. However, as highlighted in [11], standard FID struggles to capture detail and measure fidelity at higher resolutions due to the downsampling inherited from Inception network [71]. To remedy this, we also report patch FID (p-FID) [11] for a more comprehensive evaluation, especially when images contain objects at different scales, such as LHQ [66], which contains 90k high-resolution landscape images and offers a more diverse scale of scenes/objects presented in the image compared to FacesHQ. These two datasets serve as the basis for the evaluation.

For general T2I image synthesis, we train on the Unsplash dataset [2], which provides diverse and high-quality images for training our model. To show our generalization ability, we evaluate on a high-resolution subset of LAION-5B [64].

### 4.2 Boosting LDM with CFM

FMBoost combines LDM with CFM to achieve an optimal trade-off between computational efficiency and visual fidelity. We demonstrate the time taken by


**Fig. 3:** Uncurated samples from the Coupling Flow Matching model on top of SD 1.5 [59] using a classifier-free guidance scale of 7.5. Samples are generated in $64^2$ latent space and up-sampled with CFM from $64^2$ to $128^2$. The resulting images have a resolution of $1024 \times 1024$ pixels. Best viewed via zoomed-in.

LDM and FM, respectively, to synthesize 1k resolution images in Fig. 4, where LDM's inference time scales quadratically with increasing resolution, and inference is nearly impractical for real-time inference for a latent space of $128^2$. To ensure a fair comparison within the limited time frame, we compare our combination to the LCM-LoRA SDXL model [48, 69], which is known for its significantly faster inference than the original SDXL model. Tab. 1 shows that our approach with a standard SD baseline model yields superior performance in terms of FID and inference speed. Note that we apply attention scaling [27] on SD to synthesize images for varying resolutions and finetune the models, with more details in the Appendix. We present a selection of image samples from the baseline SD1.5 model and CFM $64^2 \rightarrow 128^2$ in Fig. 3. We can equally upscale the LCM-LoRA SD1.5 model from 512 to 1k resolution images with our CFM model. We present our synthesized results in Fig. 1. The inference time for a batch of four samples is 1.388 seconds on an NVIDIA A100 GPU. The LCM-LoRA SDXL model fails to produce images with similar fidelity at the same resolution within the same time.

We further demonstrate the effectiveness of our approach by comparing it to state-of-the-art models [54, 55, 83] in image synthesis on COCO $1024 \times 1024$, including CogView3 [83]. We reduce the computational cost of the diffusion component by using a lower resolution and fewer steps while offloading the remaining steps to our CFM module. This approach significantly reduces the inference time and maintains a good trade-off between speed and accuracy, as shown by the FID in Tab. 5 in the Appendix. In summary, we achieve a competitive FID at a faster inference speed than the counterpart diffusion models.

**Table 1:** Quantitative comparison for $1024^2$ image synthesis using SD v1.5 [59] plus our Coupling Flow Matching (CFM) method against a state-of-the-art diffusion speed-up method. The numbers after CFM symbolize the starting and ending resolutions in pixel space. FID and p-FID are computed for 5k samples. We use the fixed step-size Euler ODE solver with 40 NFE for CFM. For LCM-LoRA SDXL [48] we use 4 sampling steps.

| **Zero-shot LAION-5k 1024×1024** | | | | |
|---|---|---|---|---|
| Model | CLIP ↑ | FID ↓ | p-FID ↓ | time ($s$/im) ↓ |
| SD1.5 + CFM $\mathbf{256^2 \rightarrow 1024^2}$ | 23.75 | 25.47 | 23.31 | **0.62** |
| SD1.5 + CFM $\mathbf{512^2 \rightarrow 1024^2}$ | **26.14** | **21.67** | **15.96** | 3.16 |
| LCM-LoRA SDXL [48] | 24.51 | 28.98 | 24.00 | 1.83 |

### 4.3 Baseline Comparison

We compare our CFM model to three baseline methods on the FacesHQ and LHQ datasets. For a fair comparison, we fix the UNet architecture and hyper-parameters so that the models only differ in their respective training objectives.

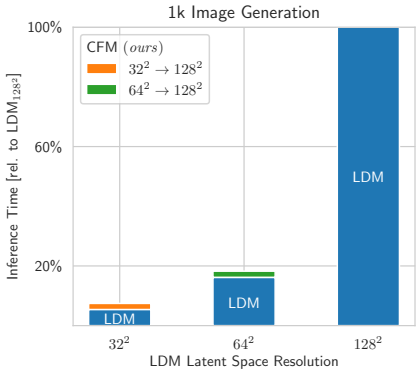



**Fig. 4:** Comparison of 1k image synthesis performance using different architectures. We utilize SD v1.5 as our base model for LDM and adapt its resolution based on [27]. LDM's inference time grows quadratically with higher resolutions, making real-time inference nearly impractical at a $128^2$ resolution latent space. In contrast, the integration of Coupling Flow Matching (CFM) with 50 function evaluations exhibits consistently faster inference, highlighting its efficiency in high-resolution image synthesis.

**Regression Baseline**. Similar to [62], we compare simple one-step regression models with $L1$ and $L2$ loss, respectively. The input is the low-resolution latent code and the target is the corresponding high-resolution latent code of the pre-

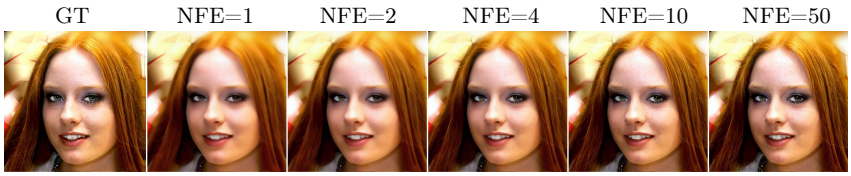



**Fig. 5:** Sample quality for different number of function evaluations (NFE). From left to right, 1st column represents the ground truth, high-resolution image. From 2nd column on, we show the results for NFE $= 1, 2, 4, 10, 50$ with the Euler ODE solver.

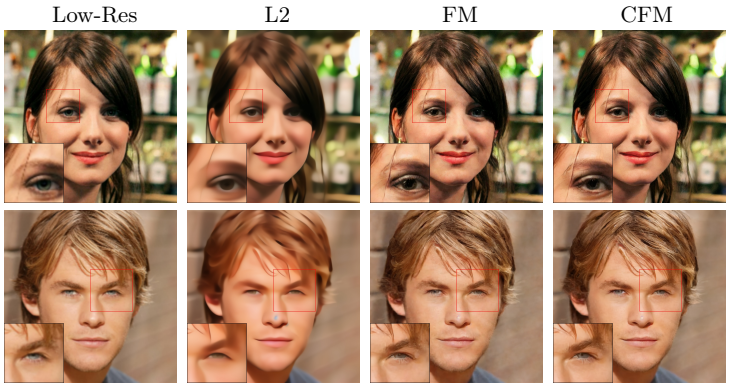



**Fig. 6:** Results for different baseline methods, increasing resolution from $256^2$ px to $1024^2$ px. *Low-Res* corresponds to bi-linear upsampling of the low-resolution image, *L2* refers to the L2 regression baseline. *FM* and *CFM* correspond to Flow Matching and Coupling Flow Matching, respectively. Best viewed when zoomed in.

trained KL autoencoder. In contrast, our method is trained with $L2$ loss on intermediate vector fields. Tab. 2 shows that CFM yields superior metric results. This is also reflected qualitatively, as visualized in Fig. 6, where the images from the regression baseline are visually blurry due to the mode-averaging behavior of the MSE regression. CFM excels at adding fine-grained, high-resolution detail to the image. We conclude that simple regression models trained with $L1$ or $L2$ loss are not sufficient to increase resolution in latent space.

**Diffusion Models.** Based on optimal transport theory, the training of a constant velocity field presents a more straightforward training objective when contrasted with the intricate high-curvature probability paths found in DMs [13, 41]. This distinction often translates to slower training convergence and potentially sub-optimal trajectories for DMs, which could detrimentally impact both training duration and overall model performance. Fig. 7 shows that within 100k iterations and for different numbers of function evaluations (NFE) after convergence, we consistently achieve a lower FID compared to the DM. In particular,

**Table 2:** Metric results for $L1$ and $L2$ regression, diffusion-based ($DM$) similar to [62], Flow Matching ($FM$) [41], and our Coupling Flow Matching ($CFM$) on 5k samples from FacesHQ and LHQ high-resolution datasets, respectively.

| Model | FacesHQ | | | | LHQ | | | |
|---|---|---|---|---|---|---|---|---|
| | SSIM↑ | PSNR↑ | FID↓ | p-FID↓ | SSIM↑ | PSNR↑ | FID↓ | p-FID↓ |
| $L1$ | **0.86** | **31.78** | 4.52 | 6.51 | **0.72** | **26.99** | 4.88 | 6.54 |
| $L2$ | 0.85 | 31.48 | 5.73 | 9.07 | **0.72** | 26.87 | 6.02 | 8.59 |
| DM | 0.73 | 23.68 | 2.72 | 4.71 | 0.61 | 19.94 | 4.29 | 4.55 |
| FM | 0.82 | 30.46 | 1.37 | 2.10 | 0.68 | 25.50 | 2.31 | 2.61 |
| CFM (*ours*) | 0.82 | 30.40 | **1.36** | **1.62** | 0.69 | 25.69 | **2.27** | **2.38** |

the CFM model shows a faster reduction of the FID and provides better results. Tab. 3 shows that the combination of DM and CFM outperforms the cascaded DMs across the board.

Taken together, these results underscore the training efficiency of our CFM model over DMs and its superior performance on the up-sampling task after fewer training steps.

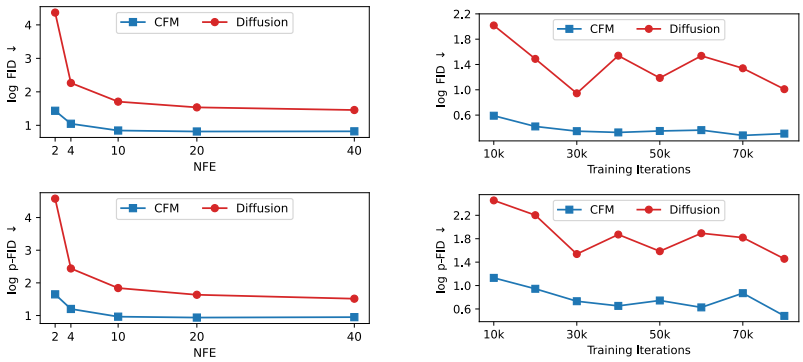



**Fig. 7:** Comparison of a diffusion-based [62] and our Coupling Flow Matching (CFM) module over the training for 4× up-sampling of the latent codes from $32^2 \rightarrow 128^2$. The decoded output resolution is $1024^2$. We report FID and p-FID for **a)** different numbers of function evaluations (NFE) and **b)** throughout training. Architecture and hyperparameters are kept fixed. FID evaluated on 5k samples from the LHQ validation set. We use 50 steps for both DDIM [68] sampling and the Euler ODE solver.

**Naïve Flow Matching**. Lastly, we compare to Naïve Flow Matching (FM). Similar to the DM, FM is conditioned on the low-resolution latent code and starts with Gaussian noise, but uses an optimal transport-based objective to regress the vector fields. In contrast, our CFM method directly starts from the low-

**Table 3:** Quantitative comparison for $1024^2$ px image synthesis using SD1.5 [59] for sampling and either our Coupling Flow Matching (CFM) method or a diffusion-based latent space up-sampling model (DM) [62]. FID and p-FID are computed for 5k samples. We use the Euler ODE solver for CFM and DDIM sampling for the DM.

| **Zero-shot LAION-5k 1024×1024** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **$256^2 \rightarrow 1024^2$** | | | | **$512^2 \rightarrow 1024^2$** | | | |
| Model | Steps | CLIP ↑ | FID ↓ | p-FID ↓ | time ($s$/im) ↓ | CLIP ↑ | FID ↓ | p-FID ↓ | time ($s$/im) ↓ |
| Diffusion | 4 | 22.92 | 35.22 | 50.53 | 0.19 | 25.68 | 26.53 | 25.20 | 2.72 |
| CFM (*ours*) | 4 | 23.77 | 27.54 | 24.02 | 0.19 | 26.16 | 21.61 | 15.83 | 2.72 |
| Diffusion | 40 | 23.55 | 26.67 | 24.16 | 0.62 | 26.05 | 22.29 | 16.36 | 3.16 |
| CFM (*ours*) | 40 | 23.75 | 25.47 | 23.31 | 0.62 | 26.14 | 21.67 | 15.96 | 3.16 |

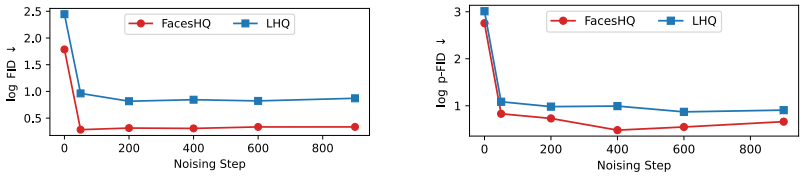



**Fig. 8:** FID (*left*) and p-FID (*right*) for our model when applying different degrees of noise augmentation. Evaluated on 5k samples.

resolution latent code and regresses the vector field towards the high-resolution counterpart. Due to data-dependent coupling, we have optimal transport guaranteed during training. All methods have the low-resolution latent code available as conditioning throughout the full generation trajectory. We evaluated the aforementioned two variants quantitatively (Tab. 2) and qualitatively (Fig. 6), where we observed that the CFM model with data-dependent coupling readily outperforms the ones without. We provide more information about the noise augmentation in Fig. 8. Notably, in the specific upsampling scenario from $256^2$ to $1024^2$, we observe an optimal configuration with a noising timestep of 400. The introduction of Gaussian noise proves beneficial as it imparts a smoothing effect on the input probability path, resulting in improved performance. However, excessive Gaussian noise can lead to the loss of valuable information, subsequently deteriorating the data-dependent coupling and reverting the model's behavior to FM's Gaussian assumption of $p(x_0)$. This finding underscores the delicate balance required in incorporating noise for optimal model performance.

### 4.4 CFM for Degraded Image Super-Resolution

Our model is originally intended to render image synthesis with existing diffusion models more effectively by enabling them to operate on a lower resolution while increasing pixel-level resolution. However, our method can also be generalized to work on super-resolution tasks which usually include image degradations [75] for low-resolution images. By fine-tuning our method, we can achieve state-of-the-art results on two common benchmark datasets on a 4× upsampling task from

$128^2$ to $512^2$ pixels. We provide quantitative (Tab. 4) and qualitative (Fig. 12) results in the Appendix.

### 4.5 CFM Model Ablations

**Upsampling Methods** Since the dimensionality of the samples from both terminal distributions must be consistent for CFM, we need to upsample the low-resolution latent code $x_0$ to match the resolution at $x_1$. In this context, we perform an ablation study comparing three different upsampling methods: nearest neighbor upsampling, bilinear upsampling, and pixel space upsampling (PSU). The first two methods operate in latent space, while PSU requires the use of the KL autoencoder to upsample in pixel space. Denoting the latent encoder as $\mathcal{E}$, the decoder as $\mathcal{D}$, and the bilinear upsampling operation as $UP$, the upsampling operation PSU can be represented as $\mathcal{E}(UP(\mathcal{D}(\cdot)))$. We empirically find that upsampling in latent space works well, but introduces artifacts that make distribution matching with CFM more difficult. In contrast, PSU yields faster and better model convergence at minimal additional cost (*cf*. Fig. 19 in the Appendix) and also makes our approach invariant to the autoencoder used. Therefore, we use PSU unless otherwise stated.
**Noise Augmentation** We systematically investigate the impact of varying levels of noise augmentation. Fig. 8 shows the FID and p-FID for different noise augmentation steps, with higher values corresponding to more noise. The highest amount of noise eradicates all information at $x_0$ and approximates [13]. Our findings suggest that noise augmentation is crucial for model performance, albeit being robust to the amount of noise. Empirically, we discovered that $t = 400$ yields the best results overall.
**Intermediate Results along the ODE Trajectory** In Fig. 16 we show intermediate results along the ODE trajectory. It can be seen that the CFM model gradually transforms the noise-augmented image representation to its high-resolution image counterpart.

## 5 Conclusion

Our work introduces a novel and effective approach, termed FMBoost, to high-resolution image synthesis, combining the generation diversity of Diffusion Models, the efficiency of Flow Matching, and the effectiveness of convolutional decoders. Strategically integrating Flow Matching models between a standard latent Diffusion model and the convolutional decoder enables a significant reduction in the computational cost of the generation process by letting the expensive Diffusion model operate at a lower resolution and up-scaling its outputs using an efficient Flow Matching model. Our Flow Matching model efficiently enhances the resolution of the latent space without compromising quality. Our approach complements DMs with their advancements and is orthogonal to their recent enhancements such as sampling acceleration and distillation techniques e.g., LCM [48]. This allows for mutual benefits between different approaches and ensures the smooth integration of our method into existing frameworks.

## Acknowledgements

This project has been supported by the German Federal Ministry for Economic Affairs and Climate Action within the project "NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung", the bidt project KLIMA-MEMES, Bayer AG, and the German Research Foundation (DFG) project 421703927. The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS at JSC and the HPC resources supplied by the Erlangen National High Performance Computing Center (NHR@FAU funded by DFG).

## References

1. LAION-Aesthetics | https://laion.ai/blog/laion-aesthetics, https://laion.ai/blog/laion-aesthetics
2. Unsplash | https://unsplash.com/data, https://unsplash.com/data
3. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR (2017)
4. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv (2023)
5. Albergo, M.S., Goldstein, M., Boffi, N.M., Ranganath, R., Vanden-Eijnden, E.: Stochastic interpolants with data-dependent couplings. arXiv (2023)
6. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: ICLR (2023)
7. Aram Davtyan, S.S., Favaro, P.: Efficient video prediction via sparsely conditioned flow matching. In: ICCV (2023)
8. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv (2022)
9. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
10. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019)
11. Chai, L., Gharbi, M., Shechtman, E., Isola, P., Zhang, R.: Any-resolution training for high-resolution image synthesis. In: ECCV (2022)
12. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv (2023)
13. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv (2023)
14. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
15. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

17. Gui, M., Fischer, J.S., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast monocular depth estimation with flow matching. arXiv preprint arXiv:2403.13788 (2024)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
20. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)
21. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
22. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: arXiv (2022)
23. Hu, T., Zhang, D.W., Mettes, P., Tang, M., Zhao, D., Snoek, C.G.: Latent space editing in transformer-based flow matching. In: AAAI (2024)
24. Hu, V.T., Baumann, S.A., Gui, M., Grebenkova, O., Ma, P., Fischer, J., Ommer, B.: Zigma: A dit-style zigzag mamba diffusion model. In: ECCV (2024)
25. Hu, V.T., Wu, D., Asano, Y.M., Mettes, P., Fernando, B., Ommer, B., Snoek, C.G.M.: Flow matching for conditional text generation in a few sampling steps. In: EACL (2024)
26. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. PMLR (2021)
27. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis (2023)
28. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: CVPR (2023)
29. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv (2017)
30. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
31. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
32. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
33. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: NeurIPS (2021)
34. Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al.: Voicebox: Text-guided multilingual universal speech generation at scale. In: arXiv (2023)
35. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
36. Lee, S., Kim, B., Ye, J.C.: Minimizing trajectory curvature of ode-based generative models. arXiv (2023)
37. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing (2022)

38. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023)
39. Liang, J., Zeng, H., Zhang, L.: Efficient and degradation-adaptive network for real-world image super-resolution. In: ECCV (2022)
40. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
41. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
42. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: ICML (2023)
43. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: ICLR (2022)
44. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)
45. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. NeurIPS (2022)
46. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv (2022)
47. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv (2023)
48. Luo, S., et al.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint (2023)
49. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: CVPR (2023)
50. Neklyudov, K., Brekelmans, R., Severo, D., Makhzani, A.: Action matching: Learning stochastic dynamics from samples. In: ICML (2023)
51. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv (2021)
52. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
53. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
54. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models (2023)
55. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv (2023)
56. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: CVPR (2022)
57. Rabe, M.N., Staats, C.: Self-attention does not need $o(n^2)$ memory. arXiv (2021)
58. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv (2022)
59. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
60. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)

61. 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. NeurIPS (2022)
62. Saharia, C., et al.: Image super-resolution via iterative refinement. TPAMI (2022)
63. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: ICLR (2022)
64. 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. NeurIPS (2022)
65. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv (2022)
66. Skorokhodov, I., Sotnikov, G., Elhoseiny, M.: Aligning latent and image spaces to connect the unconnectable. In: ICCV (2021)
67. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
68. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
69. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: ICML (2023)
70. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
71. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
72. Tong, A., et al.: Improving and generalizing flow-based generative models with minibatch optimal transport. In: ICML Worshop (2023)
73. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv (2023)
74. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV (2021)
75. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCV Workshop (2018)
76. Wang, X., et al.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: ICCV (2021)
77. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
78. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv (2021)
79. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv (2023)
80. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. NeurIPS (2024)
81. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: ICCV (2021)
82. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
83. Zheng, W., Teng, J., Yang, Z., Wang, W., Chen, J., Gu, X., Dong, Y., Ding, M., Tang, J.: Cogview3: Finer and faster text-to-image generation via relay diffusion. arXiv (2024)