University *of Ljubljana*
Faculty *of Computer and Information Science*

Efficient Face Super-Resolution with Conditional Flow Matching: Do Perceptual Losses Help?

Camile Lendering

1. Introduction

Upsampling very low-resolution face images to high resolution while preserving realism and identity remains a difficult task in computer vision. Although super-resolution (SR) methods have improved significantly, many still struggle when the input is extremely small, such as 16×16 pixels. This is particularly challenging for faces, where subtle visual features like eyes, skin texture, and hair are essential for both perceptual quality and identity consistency.

Classic SR models such as SRCNN [1], SRResNet [2], and ESRGAN [3] use convolutional architectures with pixel-wise or perceptual losses. These approaches work well for moderate scaling factors (e.g., $2\times$, $4\times$), but tend to produce smooth or unrealistic outputs at higher scales (e.g., $8\times$). Pixel losses often fail to reconstruct fine details when most of the original content is missing [4].

More recently, diffusion-based models like SR3 [5] have achieved state-of-the-art performance by modeling the conditional distribution of high-resolution images given low-resolution inputs. However, diffusion models are data-hungry, and often require hundreds or thousands of iterative denoising steps per sample, which makes them computationally expensive at inference time.

Conditional Flow Matching (CFM) [6] provides a simpler and more efficient alternative. CFM learns to generate data by predicting velocity fields that map noise to samples through continuous-time integration. Unlike diffusion models, it avoids iterative simulation during training and converges in far fewer steps at test time. Because CFM supports conditioning, it is well-suited for tasks like SR.

In this work, we apply CFM to $8\times$ face super-resolution on CelebA-HQ. The model is conditioned on bicubic-upsampled inputs and trained using a simple velocity-matching objective. We evaluate two variants: a default version trained only with the CFM loss, and a second version augmented with LPIPS [7] and FaceNet-based identity loss [8].

Surprisingly, we find that the default CFM model performs competitively across all key metrics. While LPIPS and identity losses improve targeted aspects such as perceptual similarity and embedding-based identity at short sampling depths, the default model achieves comparable, often stronger, results overall, particularly at higher sampling steps.

The complete implementation is available at: https://github.com/CLendering/conditional-flow-matching-sr.

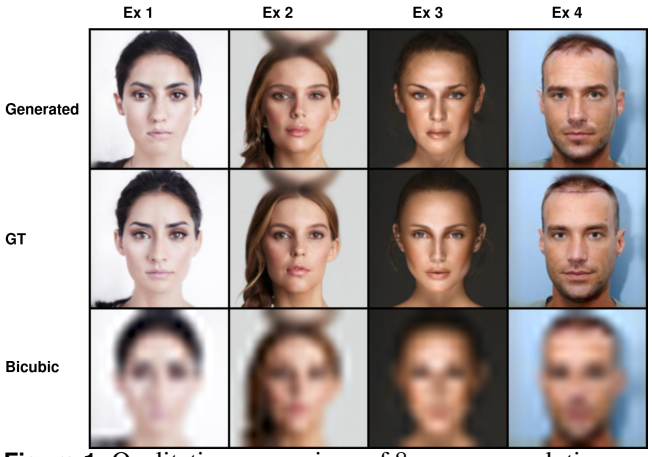

Figure 1. Qualitative comparison of $8\times$ super-resolution. Top: results from our model. Middle: ground truth high-resolution images. Bottom: bicubic upsampling. Our method preserves facial structure and fine details significantly better than the baseline.

2. Method

We apply Conditional Flow Matching (CFM) to $8\times$ face super-resolution on the CelebA-HQ dataset. The task is to reconstruct 128×128 high-resolution (HR) face images from 16×16 low-resolution (LR) inputs, conditioned on upsampled LR images. Our approach uses a simulation-free training objective and a time- and condition-aware U-Net architecture.

2.1 Conditional Flow Matching (CFM)

Conditional Flow Matching (CFM) [6] trains a neural network to model a smooth, continuous transformation from a simple noise distribution p_0 (typically a standard Gaussian $\mathcal{N}(0,I)$) to a complex data distribution p_1 (e.g., high-resolution face images). The transformation is conditioned on auxiliary input c, such as a low-resolution image.

The model learns a time-dependent velocity field $v_\theta(x_t,t,c)$, parameterized by a neural network, which defines the evolution of samples over time via the following ordinary differential equation (ODE):

$$\frac{dx_t}{dt} = v_\theta(x_t,t,c), \tag{1}$$

where x_t is the state at continuous time t, interpolating between the starting point $x_0 \sim p_0$ and the target $x_1 \sim p_1$, and $t \in [t_{\min}, t_{\max}]$.

Unlike score-based or likelihood-based methods, CFM directly regresses to the target velocity. It defines a simple linear interpolation between x_0 and x_1:

$$x_t = t \cdot x_1 + (1-t)\cdot x_0, \tag{2}$$

and sets the target velocity at each point along this path to the constant vector:

$$u_t(x_t|x_1,x_0,c) = x_1 - x_0. \tag{3}$$

The network is then trained to match the predicted velocity v_θ to this target using a mean squared error (MSE) loss on vector fields:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{x_0,x_1,t,c}\left\|v_\theta(x_t,t,c) - (x_1-x_0)\right\|^2. \tag{4}$$

During training, pairs (x_0,x_1) are sampled from the noise and data distributions, the conditioning input c is provided, and time t is sampled uniformly from the interval $[t_{\min}, t_{\max}]$. The training process is efficient because it does not require solving the ODE itself, only evaluating it at sampled times. This makes CFM a simple and tractable method for learning conditional generative models.

2.2 Architecture

We parameterize the velocity field v_θ using a time- and condition-aware U-Net [9], denoted `UNetSR`. The model adopts a symmetric encoder–decoder structure with FiLM-modulated residual blocks [10], self-attention in deeper layers, and a dedicated bottleneck module. The network is conditioned on both time and low-resolution input features. The total number of trainable parameters is approximately 35 million.

Table 1 summarizes the architecture; a schematic is shown in Figure 2.

Component	Description	Location
Time embedding	Sinusoidal positional encoding → 2-layer MLP → $t_{\text{emb}} \in \mathbb{R}^d$	Used in all residual blocks
FiLM modulation	Each residual block receives $(\gamma,\beta) = \text{Linear}(t_{\text{emb}})$ and applies $x \leftarrow x\cdot(1+\gamma)+\beta$	Before each 3×3 conv
Conditional input	The low-resolution image is upsampled to 128×128, processed by a residual block, and concatenated with the noisy HR image (6 channels total)	Encoder input
Encoder	Four resolution levels with channel sizes $[64,128,256,512]$; each level has 2 pre-activation ResBlocks, self-attention, and a stride-2 downsampling conv (except the last level)	Downsampling path
Bottleneck	ResBlock → 8-head attention → 1×1 conv → SiLU → dropout → ResBlock (scale = 0)	Centre of the network
Decoder	Three upsampling stages; each stage: transposed conv → skip connection → 2 ResBlocks → attention	Upsampling path
Output head	GroupNorm → SiLU → 3×3 conv to predict the velocity field	Final output

Table 1. Main components of `UNetSR`, the architecture used to parameterise v_θ.

2.3 Training Configuration

The model is trained to minimize a weighted loss that balances pixel accuracy, perceptual quality, and identity preservation.

2.3.1 Loss Functions

CFM Loss ($\mathcal{L}_{\text{CFM}}$). Given a pair (x_0,x_1) and a randomly sampled timestep $t \sim \mathcal{U}(t_{\min},t_{\max})$, we form an interpolated state:

$$x_t = t\,x_1 + (1-t)\,x_0.$$

The model predicts a velocity field at x_t, which is supervised using a constant target $u_t = x_1 - x_0$:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{x_0,x_1,t,c}\left\|v_\theta(x_t,t,c) - (x_1-x_0)\right\|_2^2. \tag{5}$$

While the target velocity is constant across time, the predicted velocity depends on x_t, which encodes time implicitly.

Perceptual Loss (LPIPS). To encourage perceptual similarity, we include an LPIPS loss [7] comparing deep features of denormalized predictions and targets.

$$\mathcal{L}_{\text{LPIPS}} = \text{LPIPS}(\text{denorm}(x_{\text{pred}}), \text{denorm}(x_1)). \tag{6}$$

This term is weighted by a hyperparameter λ_{LPIPS}.

Identity Loss. To preserve identity, we use a cosine distance loss between 512-dimensional FaceNet embeddings:

$$\mathcal{L}_{\text{ID}} = 1 - \cos\angle(f(x_{\text{pred}}), f(x_1)), \tag{7}$$

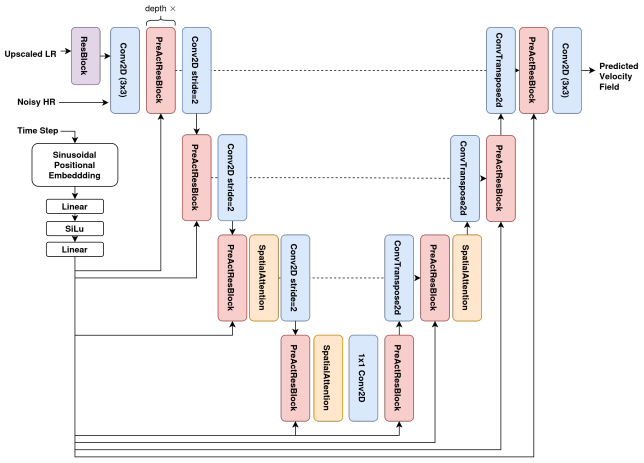

Figure 2. Diagram of the `UNetSR` architecture used to model v_θ. The network takes as input a noisy high-resolution image and an upsampled low-resolution condition, and outputs a 3-channel velocity field. Time conditioning is applied via FiLM layers using a sinusoidal embedding. Attention blocks are included at the deeper levels and in the bottleneck.

weighted by λ_{ID}, where $f(\cdot)$ is a frozen identity encoder.

Total Loss. The full objective combines all terms:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{CFM}} + \lambda_{\text{LPIPS}} \cdot \mathscr{L}_{\text{LPIPS}} + \lambda_{\text{ID}} \cdot \mathscr{L}_{\text{ID}}. \tag{8}$$

Approximating x_{pred}. The network predicts velocity, not images. To approximate a denoised image during training, we apply a single reverse Euler step:

$$x_{\text{pred}} = x_t + (1-t)\, v_\theta(x_t, t, c). \tag{9}$$

This serves as a one-step estimate of x_1. Multi-step integration is only used during inference.

Training Setup. We use the AdamW optimizer [11] with cosine learning rate annealing over T epochs. All inputs are Z-normalized using training set statistics. Denormalization is applied only when computing LPIPS and identity similarity metrics.

Validation Metrics. We evaluate the model on a held-out validation set using multiple metrics computed on denormalised outputs. Each metric captures a different aspect of reconstruction quality:

- **LPIPS** (VGG backbone): the Learned Perceptual Image Patch Similarity [7] measures perceptual similarity by comparing deep features of the predicted and ground truth images extracted from a pretrained VGG network. LPIPS aligns more closely with human visual judgment than pixel-wise metrics. Lower is better.
- **PSNR** (Peak Signal-to-Noise Ratio): A pixel-level fidelity metric defined as:

$$\text{PSNR}(x_{\text{pred}}, x_1) = 10 \cdot \log_{10}\left(\frac{1}{\text{MSE}(x_{\text{pred}}, x_1)}\right),$$

where MSE is the mean squared error between the predicted and ground truth images. PSNR is expressed in decibels (dB); higher values indicate better reconstruction quality.

- **SSIM** (Structural Similarity Index Measure): This metric evaluates perceived structural similarity between images, taking into account luminance, contrast, and structure. It is computed over local image windows and ranges from 0 (no similarity) to 1 (perfect match). The formula for a single window is:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where μ_x, μ_y are local means, σ_x^2, σ_y^2 are variances, and σ_{xy} is the covariance between x and y.

- **Identity Similarity**: Cosine similarity between FaceNet embeddings of the predicted and ground truth images. Defined as:

$$\text{cosine\_sim}(f(x_{\text{pred}}), f(x_1)) = \frac{f(x_{\text{pred}}) \cdot f(x_1)}{\|f(x_{\text{pred}})\|\, \|f(x_1)\|},$$

where $f(\cdot)$ denotes the 512-D embedding extracted by the FaceNet encoder. Higher values indicate better preservation of facial identity.

- **FID** (Fréchet Inception Distance): Measures distribution-level similarity between generated and real images in Inception feature space. Let $\mathscr{N}(\mu_r, \Sigma_r)$ and $\mathscr{N}(\mu_g, \Sigma_g)$ be the Gaussian statistics of real and generated features, respectively. Then:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}).$$

Lower FID indicates that generated images match the distribution of real images more closely in a perceptual sense. This was only evaluated on the test set.

Validation is performed every K epochs. The best model checkpoint is selected based on the lowest $\text{LPIPS}_{\text{VGG}}$ score, prioritising perceptual quality more aligned with human visual preference.

3. Results

We evaluate two variants of our Conditional Flow Matching model on the CelebA-HQ test set:

1. **CFM + LPIPS + Identity Loss:** Trained with the standard CFM objective and augmented with LPIPS and FaceNet-based identity similarity loss.
2. **Default CFM:** Trained solely with the CFM loss, without any perceptual or identity-based supervision.

We assess performance using five metrics computed on denormalized outputs: PSNR, SSIM, LPIPS (VGG backbone), identity cosine similarity (FaceNet), and Fréchet Inception

Distance (FID). We evaluate all models across multiple Euler integration step counts (1, 10, 100, 1000) to analyze the effect of sampling depth.

3.1 Quantitative Results

3.1.1 Perceptual and Identity-Aware Model

Table 2 shows the evaluation results for the model trained with LPIPS and identity loss. As the number of sampling steps increases, perceptual similarity (LPIPS) and identity consistency (ID Cos) worsen steadily, indicating a gradual loss of fine details and semantic fidelity. PSNR and SSIM also decline with more steps, suggesting that pixel-level and structural fidelity are both reduced. Contrary to expectations, FID does not consistently improve: it decreases from 1 to 10 steps, but then increases again at 100 and 1000 steps. This suggests a trade-off: shorter sampling trajectories (e.g., 10 steps) may strike a better balance between realism and fidelity, while longer integrations can introduce divergence.

Table 2. Evaluation of the full model (CFM + LPIPS + ID loss) across sampling steps.

Steps	PSNR	SSIM	$\text{LPIPS}_{\text{VGG}}$	ID Cos	FID
1	25.17	0.7384	**0.1753**	**0.7871**	23.24
10	23.78	0.6840	0.1992	0.7727	18.56
100	23.07	0.6460	0.2222	0.7630	25.67
1000	22.98	0.6398	0.2276	0.7618	26.89

3.1.2 Default Conditional Flow Model (No Perceptual or Identity Loss)

Table 3 presents the evaluation of the default CFM model, trained without LPIPS or identity loss. The model achieves the best pixel-level performance (PSNR and SSIM) at 1 sampling step, with a gradual decline as steps increase. However, perceptual metrics (LPIPS, FID) show the opposite trend: realism improves with more steps, with FID reaching its best value at 1000 steps. Identity similarity (ID Cos) decreases slightly with more steps but remains competitive. This suggests that while the default model lacks explicit perceptual supervision, longer sampling schedules enhance realism, though at the cost of structural fidelity. The model benefits from additional steps more clearly than the LPIPS+ID variant.

Table 3. Evaluation of the default CFM model (no LPIPS, no ID loss).

Steps	PSNR	SSIM	$\text{LPIPS}_{\text{VGG}}$	ID Cos	FID
1	**25.40**	**0.7570**	0.2300	0.7716	45.27
10	24.44	0.7166	0.2007	0.7419	23.40
100	23.97	0.6901	**0.2019**	0.7334	16.16
1000	23.91	0.6861	0.2029	0.7291	**15.98**

3.1.3 Model Comparison

Table 4 directly compares the two models at 1000 steps. The default model clearly outperforms the LPIPS+ID variant in all metrics except identity similarity, where the difference is notable. These results suggest that the additional losses are not needed when using a well-parameterized conditional flow model.

Table 4. Comparison of model variants at 1000 Euler steps.

Model	PSNR	SSIM	$\text{LPIPS}_{\text{VGG}}$	ID Cos	FID
CFM (full)	22.98	0.6398	0.2276	**0.7618**	26.89
CFM (default)	**23.91**	**0.6861**	**0.2029**	0.7291	**15.98**

3.2 Qualitative Results

Figure 3 shows qualitative outputs from both models. Overall differences are subtle: both maintain identity, structure, and general perceptual quality. However, the perceptual and identity-aware model produces slightly sharper textures and more high-frequency details, while the default model yields smoother, more denoised outputs. This reflects the trade-off between perceptual sharpness and pixel-level consistency.

Figure 3. Face SR comparison: Left—Ground truth. Middle—Default CFM. Right—CFM with LPIPS + ID loss.

Figure 4 visualizes how outputs evolve across sampling steps. Already at 10 steps, the model recovers detailed textures. At 1 step, outputs retain identity but lose fine-grained detail. This highlights CFM's ability to generate high-fidelity images with minimal compute. Figure 6 shows the same progression for the perceptual and identity-aware model, which maintains sharper textures even at low step counts but introduces slightly more variation across samples. The trade-off between perceptual richness and structural stability becomes more evident with increased steps.

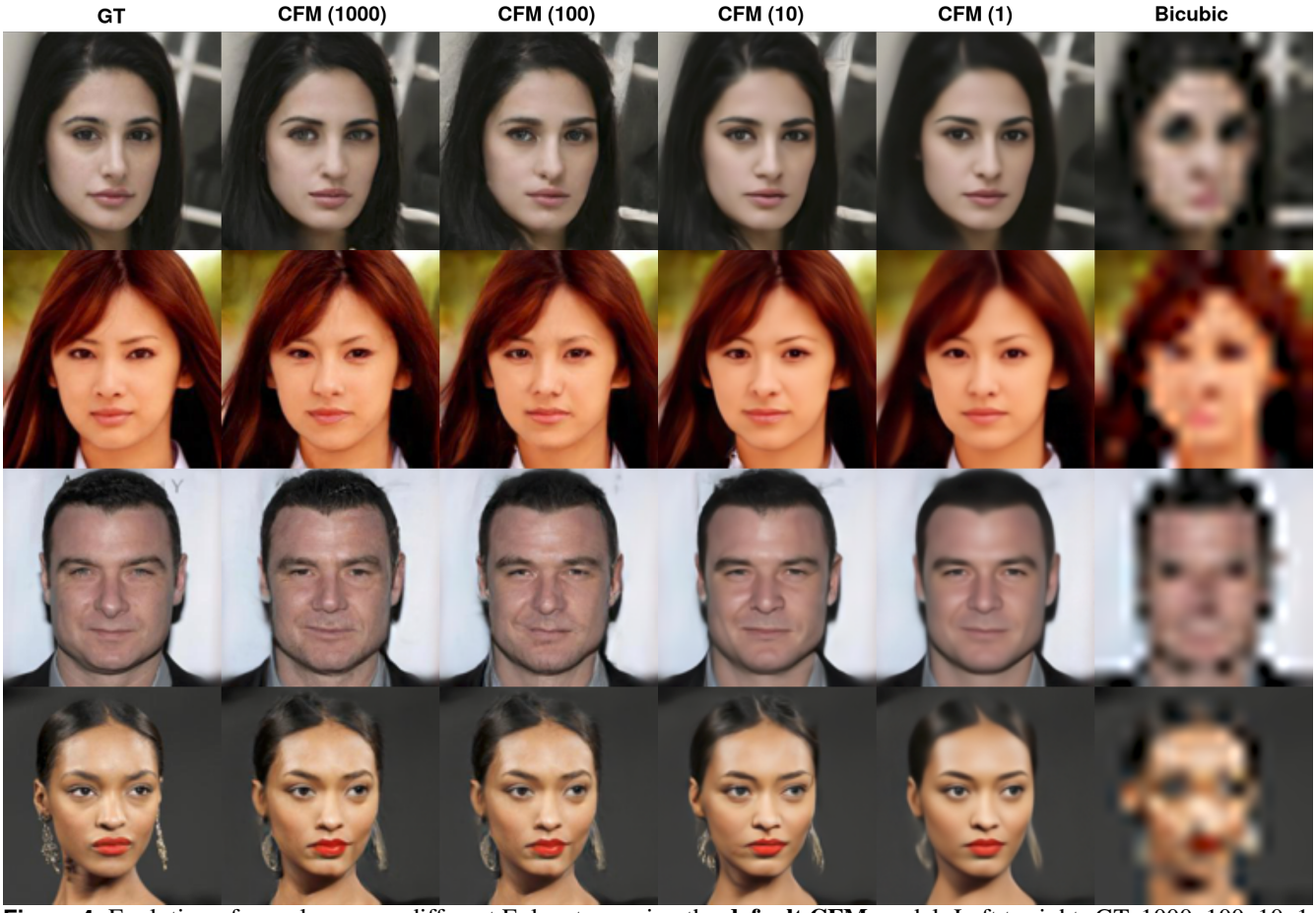

Figure 4. Evolution of samples across different Euler steps using the **default CFM** model. Left to right: GT, 1000, 100, 10, 1, bicubic input. Each row: one CelebA-HQ test image.

3.3 Comparison with Prior Work

Table 5 compares our default CFM model against published face super-resolution methods including PULSE [12], FSRGAN [13], SR3 [5], and IDM [14]. These baselines are trained on FFHQ and evaluated on full CelebA, whereas our model is trained and tested on CelebA-HQ. Although not directly comparable due to dataset and evaluation differences, CFM achieves competitive or superior results, despite being trained on a smaller dataset. Notably, at just one sampling step, our model outperforms all baselines in both PSNR and SSIM. This highlights the efficiency and effectiveness of the CFM framework even under limited training data conditions.

Table 5. Reported PSNR and SSIM values from prior work vs. our model.

Method	PSNR	SSIM
PULSE [12]	16.88	0.44
FSRGAN [13]	23.01	0.62
Regression [15]	23.96	0.69
SR3 [5]	23.04	0.65
IDM [14]	24.01	0.71
CFM (1 step)	**25.40**	**0.76**
CFM (1000 steps)	23.91	0.69

3.4 Failure Cases

While Conditional Flow Matching achieves high reconstruction quality on most faces, we observe systematic failure modes on uncommon visual features. Figure 5 highlights several such cases.

- **Rare accessories:** Objects like cigarettes, microphones, or facial jewelry are often blurred, hallucinated, or omitted entirely. This likely stems from their low frequency in the training set and weak conditioning signal.
- **Unusual headwear and eyewear:** Glasses, hats, and helmets with distinctive shapes or textures (e.g., round sunglasses, military berets) are often poorly reconstructed or replaced with more common alternatives.
- **Extreme lighting and contrast:** Images with harsh shadows, side lighting, or high contrast tend to yield unstable reconstructions, particularly in background regions and occluded facial areas.
- **Fine-grained texture:** Details such as frizzy hair, facial wrinkles, and fur (e.g., beards or costume elements) are often smoothed out or replaced with lower-frequency texture, even at 1000 sampling steps.

These failure cases suggest that while the model generalizes well to typical face structures, it struggles with underrepresented visual features. This points to potential gains from data augmentation or training on a larger and more diverse dataset like FFHQ.

Figure 5. Representative failure cases. Top row: predictions from the default CFM model. Middle: ground truth HR images. Bottom: LR bicubic inputs. Failures include poor reconstruction of rare accessories (e.g., cigarettes), unique hats and glasses, and occlusions.

4. Conclusion

This work investigated Conditional Flow Matching (CFM) as a lightweight generative approach for $8\times$ face super-resolution. We trained and evaluated two U-Net-based CFM models on the CelebA-HQ dataset: a default model using only the flow-matching objective, and an augmented variant with additional perceptual (LPIPS) and identity supervision.

Results show that the default CFM model achieves competitive performance across a range of metrics, including PSNR, SSIM, LPIPS, identity similarity, and FID. While perceptual and identity losses slightly improve LPIPS and embedding-based similarity at low sampling depths, the default model performs comparably, and often better, in pixel fidelity and distributional alignment, particularly at larger step counts.

Despite being trained on a smaller dataset than most prior work, the CFM model performs well relative to published methods such as SR3 and IDM. These results highlight CFM's strong inductive bias for structure and semantics, even without explicit feature-level constraints.

Failure cases remain on visually rare features (e.g., accessories, extreme lighting), suggesting that broader datasets may further improve robustness.

In summary, Conditional Flow Matching provides an efficient and scalable alternative to diffusion-based SR models, with promising results under realistic training and inference constraints.

References

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[2] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[3] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[5] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[6] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[10] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[12] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

[13] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2492–2501, 2018.

[14] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[15] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023.

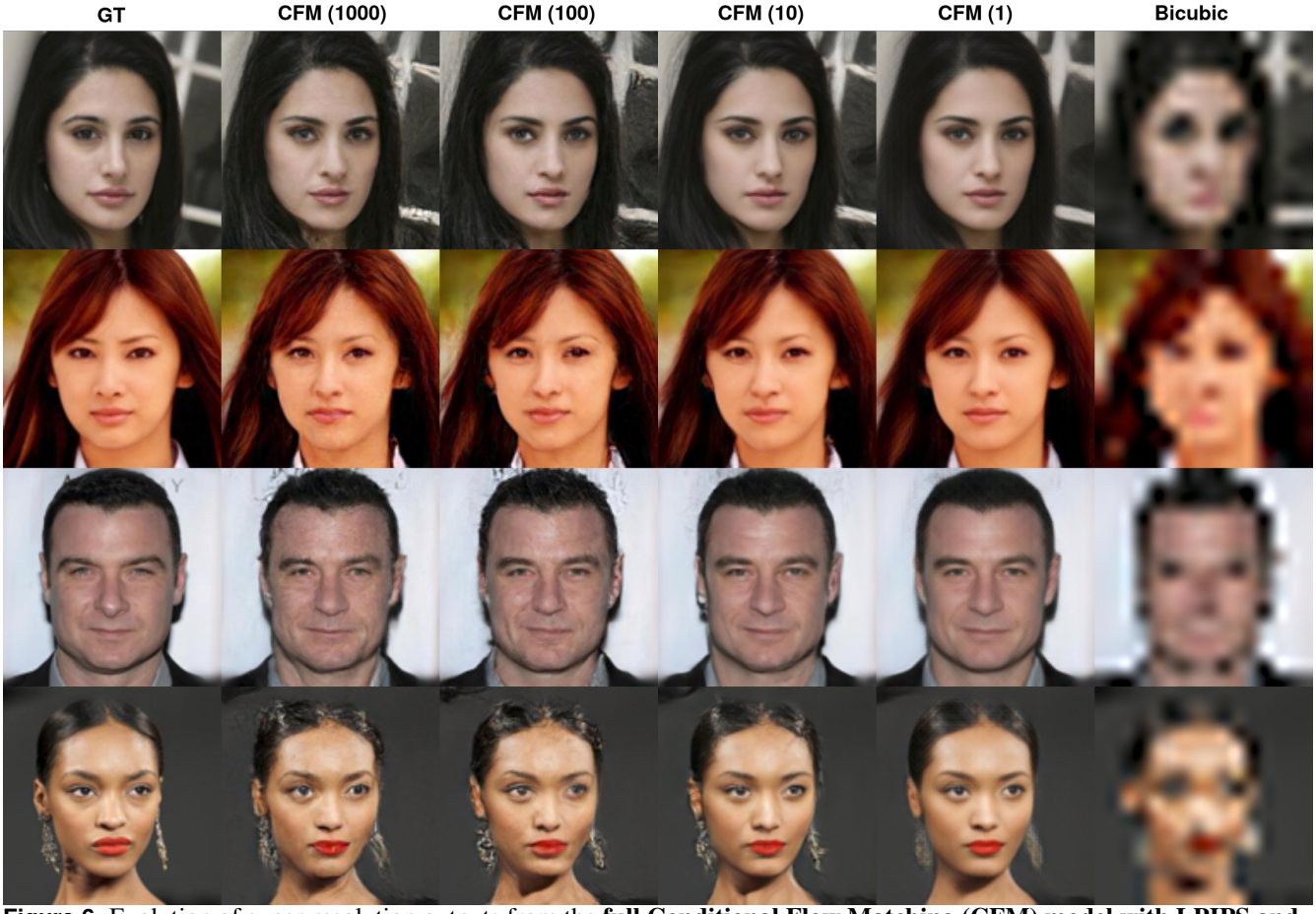

Figure 6. Evolution of super-resolution outputs from the **full Conditional Flow Matching (CFM) model with LPIPS and identity losses**. Columns show, from left to right: Ground Truth (GT), samples at 1000, 100, 10, and 1 Euler steps, and the bicubic input. Each row is a different CelebA-HQ test image. The model often introduces high-frequency textures, such as detailed hair strands or skin patterns, reflecting the effect of perceptual supervision. While visually convincing, these details may not always be faithful to the input.