How Good Are Deep Generative Models for Solving Inverse Problems?

Shichong Peng, Alireza Moazeni, Ke Li
APEX Lab
School of Computing Science
Simon Fraser University
{shichong_peng,seyed_alireza_moazenipourasil,keli}@sfu.ca

Abstract

Deep generative models, such as diffusion models, GANs, and IMLE, have shown impressive capability in tackling inverse problems. However, the validity of model-generated solutions w.r.t. the forward process and the reliability of associated uncertainty estimates remain understudied. This study evaluates recent diffusion-based, GAN-based, and IMLE-based methods on three inverse problems, i.e., $16\times$ super-resolution, colourization, and image decompression. We assess the validity of these models' outputs as solutions to the inverse problems and conduct a thorough analysis of the reliability of the models' estimates of uncertainty over the solution. Overall, we find that the IMLE-based CHIMLE method outperforms other methods in terms of producing valid solutions and reliable uncertainty estimates.

1 Introduction

Deep generative models have seen remarkable growth in adoption across various applications. These models, exemplified by methods such as diffusion models [22, 9], Generative Adversarial Networks (GANs) [8], and Implicit Maximum Likelihood Estimation (IMLE) [16], have demonstrated remarkable capabilities in capturing complex data distributions.

One particularly intriguing application of deep generative models is solving inverse problems, a class of challenges that involves reconstructing clean data with complex structure, often from incomplete or noisy observations. Instead of producing a single deterministic solution, these models can learn entire distributions over the possible solutions. This capability is pivotal when the forward process is not injective, in which case there is no unique solution, but rather many solutions. In such cases, having both a clean solution and a quantification of the uncertainty in the solution are important.

However, a critical aspect often overlooked in the evaluation of deep generative models is the validity of their outputs as solutions to the underlying inverse problem. More concretely, a valid solution is one where if the forward process were applied to the solution, the result would be similar with the provided input. In other words, while deep generative models excel at generating outputs that appear plausible, they may not actually be valid solutions to the inverse problem. This validation is of great importance, since otherwise the generated solutions would not truly invert the forward process.

Since the forward process is often not injective, another dimension of evaluation that deserves more attention is the accuracy of uncertainty estimates. Accurate pixel-level modelling of uncertainty in the generated solution is crucial, particularly in domains where decisions carry significant consequences.

In this study, we conduct an evaluation of recent deep generative models based on a variety of methods, including diffusion, GAN, and IMLE, in the context of three challenging inverse problems: $16\times$ single image super-resolution, image colourization and image decompression. Our evaluation focuses on two critical dimensions that are traditionally underemphasized: output validity and model

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

uncertainty. We aim to assess the extent to which these models' solutions align with the input data when passed through the forward process, and to quantify and analyze the inherent uncertainty in their solutions.

2 Background

2.1 Deep Generative Models for Inverse Problems

Given the input $\mathbf{x}$ and a forward process $\mathcal{F}$, the goal of inverse problems is to find solutions $\mathbf{y}$ such that $\mathcal{F}(\mathbf{y}) = \mathbf{x}$. When $\mathcal{F}$ is not injective, there can be multiple plausible solutions $\mathbf{y}$ for the same observed data $\mathbf{x}$, i.e., $\mathbf{y} \in \{\mathbf{y}' | \mathcal{F}(\mathbf{y}') = \mathbf{x}\}$. In light of this, deep generative models estimate $p(\mathbf{y}|\mathbf{x})$ to capture the distribution of all plausible solutions.

When $\mathcal{F}$ is additive Gaussian, one way to model it is with diffusion models [22, 9]. Specifically, diffusion models break down $\mathcal{F}$ into T additive Gaussian steps that progressively add noise to the data according to a predefined variance schedule. If we denote the original data as $\mathbf{x}_0$, then each step of the forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ can be written as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \tag{1}$$

Here, $\beta_1, \ldots, \beta_T$ represent the variance schedule. To model the reverse process $p_\theta(\mathbf{x}_{0:T})$, diffusion models make an additional assumption, namely that the reverse process is a Markov chain with Gaussian transition kernels, starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; 0, \mathbf{I})$:

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)) \tag{2}$$

where $\mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$ are the predicted mean and covariance at time step t. It's worth noting that the reverse process is only truly the inverse of the forward process when there are an infinite number of time steps T; otherwise the Gaussianity assumption in the transition kernel of the reverse process would not hold. In practice, a finite T is used, in which case the Gaussianity assumption would introduce approximation errors. Therefore, the assumptions of diffusion models are technically not met when the forward process is not additive Gaussian or when the number of time steps is small.

One alternative is to use a conditional GAN, comprising a generator and a discriminator. The generator G_θ takes in the observed input $\mathbf{x}$ and a latent code $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and generates an output $\mathbf{y} \coloneqq G_\theta(\mathbf{x}, \mathbf{z})$. The discriminator's role is to distinguish between the generated sample and real data example. However, due to mode collapse, the generator tends to generate identical samples for the same input and ignores the latent noise [12]. To mitigate this problem, various approaches have been proposed, including introducing additional losses in the latent space [28], adding mode-seeking terms [19], or using contrastive losses [18, 25, 11]. While these methods succeed in enhancing sample diversity, they often come at the expense of reduced sample fidelity.

Another alternative approach is Implicit Maximum Likelihood Estimation (IMLE) [16]. IMLE differs from GANs in two key ways: it explicitly aims to cover all modes, and optimizes a non-adversarial objective. To achieve the former, IMLE reverses the direction in which generated samples are matched to real data: rather than making each generated sample similar to some real data point, it makes sure each real data point has a similar generated sample. To achieve the latter, it removes the discriminator (which matches generated samples to real data implicitly) and instead explicitly performs matching using nearest neighbour search. The latter can be done efficiently using DCI [14, 15], which avoids the curse of dimensionality. Recent advancements to IMLE include cIMLE [17], which extends IMLE to the conditional setting, and CHIMLE [20], which further enhances generated image quality by adopting a hierarchical approach to constructing latent codes.

2.2 Model Uncertainty Quantification

A popular and statistically rigorous approach to perform uncertainty quantification is conformal prediction [4, 1, 7, 24]. Conformal prediction stands out for its ability to construct statistically valid uncertainty guarantees for general predictors without relying on distributional or model assumptions.

	Super-Resolution (SR)			Colourization (Col)			Image Decompression (DC)		
	LPIPS ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑	*LPIPS* ↓	*PSNR* ↑	*SSIM* ↑
GAN-based:									
BicyleGAN [27]	0.105	22.31	0.832	0.322	20.28	0.716	0.349	19.75	0.781
MSGAN [19]	0.135	20.83	0.810	0.370	18.98	0.665	0.416	17.24	0.712
DivCo [18]	0.136	20.60	0.763	0.336	18.84	0.687	0.356	18.94	0.731
MoNCE [25]	0.091	27.41	0.943	0.030	<u>39.52</u>	**0.990**	<u>0.213</u>	<u>25.78</u>	<u>0.854</u>
Diffusion-based:									
DDRM [13]	0.045	<u>29.57</u>	0.948	0.127	28.21	0.922	0.539	19.07	0.459
NDM [5]	0.082	23.50	0.854	0.057	37.34	0.951	0.525	14.15	0.575
IMLE-based:									
cIMLE [17]	<u>0.040</u>	28.57	<u>0.949</u>	<u>0.022</u>	36.12	<u>0.981</u>	0.261	22.39	0.790
CHIMLE [20]	**0.009**	**33.26**	**0.988**	**0.011**	**41.25**	**0.990**	**0.191**	**26.24**	**0.877**

Table 1: Comparison of model output validity among diffusion-based, GAN-based and IMLE-based methods. The validity is measured by computing the LPIPS [26], PSNR and SSIM between the model input image and the output obtained by running the forward process on the model's generated samples, averaged over 50 runs. Lower values of LPIPS are better and higher values of PSNR and SSIM are better. Notably, CHIMLE [20] consistently achieves the best output validity in our evaluation.

Recent studies [2, 23, 10] have extended the application of conformal prediction to address image-to-image translation problems. In this study, we adopt a sampling-based approach, as outlined in [10], to evaluate model uncertainty across different methods.

3 Experiments

We evaluate recent deep generative models for inverse problems that are based on diffusion, GAN and IMLE. For diffusion-based methods, we choose DDRM [13] and NDM [5]. For GAN-based methods, we choose BicyleGAN [27], MSGAN [19], DivCo [18] and MoNCE [25]. For IMLE-based methods, we choose cIMLE [17] and CHIMLE [20].

Our evaluation includes three challenging inverse problems: $16\times$ single image super-resolution, image colourization and image decompression. For image super-resolution, we choose three categories from ILSVRC-2012 [21] with an input resolution of 32×32, and output resolution of 512×512. For image colourization, we choose two categories from ILSVRC-2012 [21] and the Natural Color Dataset (NCD) [3]. For image decompression, we choose the RAISE1K [6] dataset and compressed each image using JPEG with a quality of 1%.

3.1 Output Validity Assessment

The first evaluation focuses on assessing the validity of the model output as a solution to the inverse problem. Recall that in an inverse problem, it is essential that the result of applying the forward process as defined by the task to the output is similar to the provided input. To assess this, we generate 50 random samples for each input image in the test set. These samples are then processed through the corresponding forward process for each task, namely bicubic downsampling by a factor of $16\times$ for super-resolution, conversion of RGB images to grayscale for colourization, and compression of images using JPEG at a quality of 1% for image decompression. Subsequently, we compare the processed outputs to the original input.

Table 1 shows the average LPIPS [26], PSNR and SSIM metrics between the model input and the result of applying the forward process to the model output. Lower LPIPS distances are better, whereas higher PSNR and SSIM are better. We found that CHIMLE [20], an IMLE-based method, performs the best across tasks and metrics. In second place, we have cIMLE [17] according to LPIPS and SSIM on super-resolution and colourization, DDRM [13] according to PSNR on super-resolution, and MoNCE [25] according to PSNR on colourization and according to all metrics on decompression.

3.2 Model Uncertainty Evaluation

To evaluate model uncertainty for each method, we adopt a sampling-based approach mentioned in [10]. For each input, we randomly generate 50 samples from each model and calculate the confidence interval at each pixel. In our visualization, the confidence interval is superimposed as highlights on top of the dimmed mean generated sample. A wider confidence interval corresponds to greater model

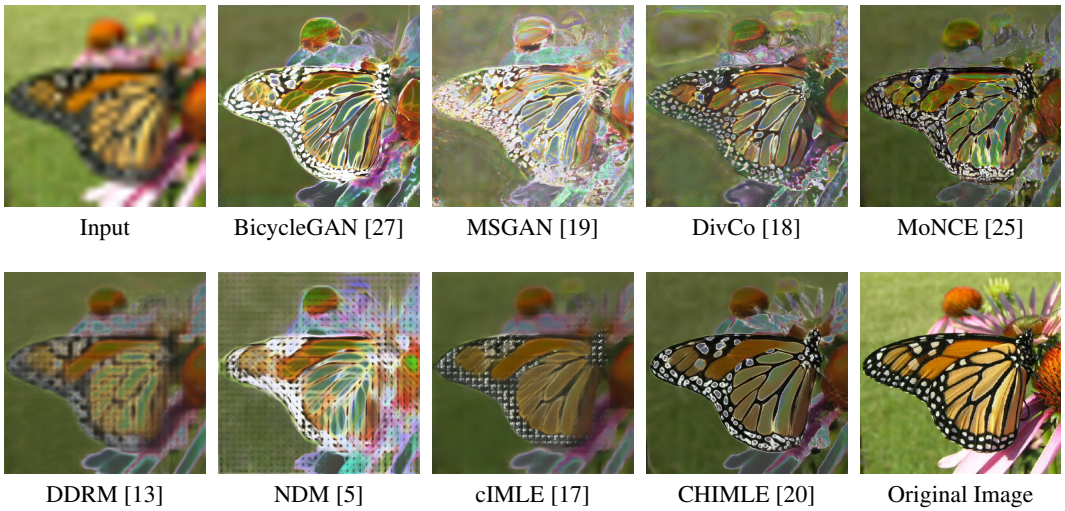

Figure 1: Comparison of model uncertainty visualized as confidence intervals extracted from 50 randomly generated samples from each method using the same input. As shown, CHIMLE [20] produces high-fidelity outputs and uncertainty estimates that closely align with ambiguous regions in the input (e.g., around the black rim of the butterfly's wing). In contrast, other methods either produce low fidelity results (DivCo [18], MoNCE [25], DDRM [13] and cIMLE [17]) or contain excessive uncertainty in less ambiguous input regions such as the thick black stripes in the wing's center (BicycleGAN [27], MSGAN [19] and NDM [5]).

uncertainty, visually appeared as a brighter colour. To determine the size of the confidence interval, we subtract the lower quantile from the upper quantile for each pixel within the set of generated samples for the same input. In our evaluation, we set the lower quantile at the 5^{th} percentile and the upper quantile at the 95^{th} percentile. Additionally, when visualizing the confidence interval for a specific input, we calibrate the lower and upper bounds using the remaining images in the test set, with more details available in [10].

Figure 1 shows the model uncertainty comparison in the task of $16\times$ image super-resolution. As shown, BicycleGAN [27], MSGAN [19] and NDM [5] generate samples with significant variations. However, these methods also show substantial uncertainty in regions that are constrained significantly by the input and therefore should not be uncertain, e.g., the thick black strips in the wing's center should be just solid black with no other possibilities. On the other hand, DDRM [13] generates blurry outputs and shows low uncertainty in regions where the input is ambiguous, where one would expect a higher degree of uncertainty, e.g., along the rim of the butterfly's wing. DivCo [18], MoNCE [25], and cIMLE [17] show model uncertainties that align with the degree of ambiguity present in the input, but their outputs are low in visual fidelity. CHIMLE [20] achieves the overall best performance and produces realistic outputs and uncertainty estimates that accurately capture input ambiguity. Visual comparisons of model uncertainty for the other tasks, namely image colourization and image decompression, can be found in the appendix.

4 Conclusion

In this study, we thoroughly evaluated diffusion-based, GAN-based, IMLE-based methods on three challenging inverse problems, namely $16\times$ super-resolution, colourization, and image decompression. Our assessment focused on two critical dimensions: output validity and model uncertainty. Surprisingly, despite their popularity, diffusion-based methods do not perform well on either dimension, whereas GAN-based and IMLE-based methods perform better. Among GAN-based and IMLE-based methods, CHIMLE [20] seems to perform the best along both dimensions.

References

[1] Anastasios Nikolas Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *ArXiv*, abs/2107.07511, 2021.

[2] Anastasios Nikolas Angelopoulos, Amit Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. *ArXiv*, abs/2202.05265, 2022.

[3] Saeed Anwar, Muhammad Tahir, Chongyi Li, A. Mian, F. Khan, and A. W. Muzaffar. Image colorization: A survey and dataset. *ArXiv*, abs/2008.10774, 2020.

[4] Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68:43:1–43:34, 2021.

[5] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Non-uniform diffusion models. *arXiv preprint arXiv:2207.09786*, 2022.

[6] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224, 2015.

[7] Bat-Sheva Einbinder, Stephen Bates, Anastasios Nikolas Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to label noise. *ArXiv*, abs/2209.14295, 2022.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[10] Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *ArXiv*, abs/2211.09795, 2022.

[11] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. *arXiv preprint arXiv:2203.08483*, 2022.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[13] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

[14] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via Dynamic Continuous Indexing. In *International Conference on Machine Learning*, pages 671–679, 2016.

[15] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via Prioritized DCI. In *International Conference on Machine Learning*, pages 2081–2090, 2017.

[16] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018.

[17] Ke Li*, Shichong Peng*, Tianhao Zhang*, and Jitendra Malik. Multimodal image synthesis with conditional implicit maximum likelihood estimation. *International Journal of Computer Vision*, May 2020.

[18] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. DivCo : Diverse conditional image synthesis via contrastive generative adversarial network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16372–16381, 2021.

[19] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1429–1437, 2019.

[20] Shichong Peng, Seyed Alireza Moazenipourasil, and Ke Li. Chimle: Conditional hierarchical imle for multimodal conditional image synthesis. *Advances in Neural Information Processing Systems*, 35:280–296, 2022.

[21] Olga Russakovsky, J. Deng, Hao Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[23] Jacopo Teneggi, Matthew Tivnan, J. Webster Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, 2023.

[24] Jiaye Teng, Chuan Wen, Dinghuai Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. Predictive inference with feature conformal prediction. *ArXiv*, abs/2210.00173, 2022.

[25] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[27] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017.

[28] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.

A Additional Results on Model Uncertainty

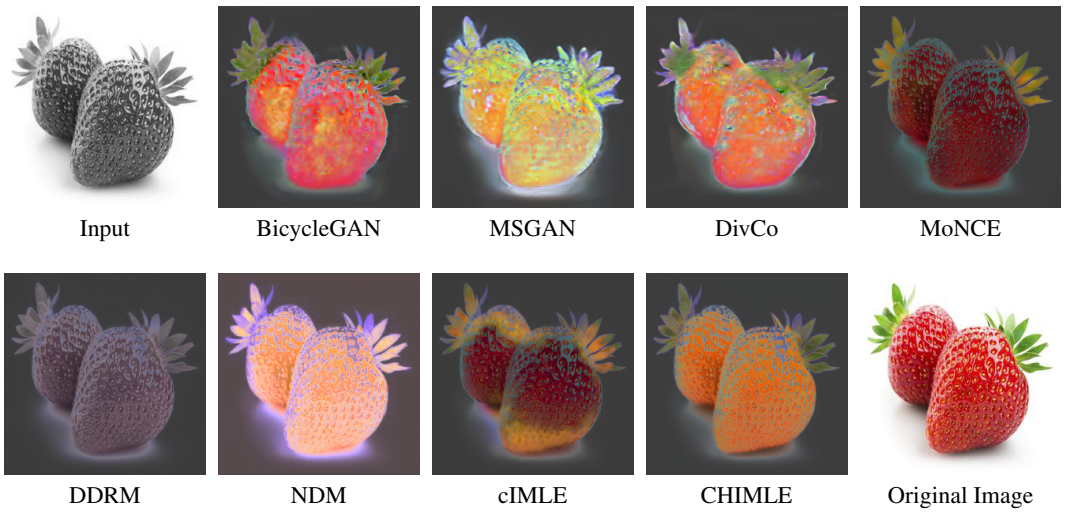

Figure 2: Model uncertainty comparison on image colourization.

Figure 2 shows the model uncertainty comparison of different methods on the task of image colourization. In our visualization, the confidence interval is superimposed as highlights on top of the dimmed mean generated sample. A wider confidence interval corresponds to greater model uncertainty, visually appeared as a brighter colour. As shown, BicycleGAN, MSGAN and DivCo produce diverse results but their samples fail to preserve the details from the input and lack visual fidelity (i.e., look unrealistic). Conversely, MoNCE generates samples with limited diversity, with uncertainty primarily concentrated in less ambiguous areas, such as the strawberry's leaf (where the colour of the strawberry leaf should be green). DDRM produces samples that show desaturation and minimal diversity, and fails to capture different plausible solutions (e.g., the strawberry's colour could be either red or green). NDM provides diverse samples but shows high uncertainty in regions that should not contain much diversity (e.g., the leaf on the strawberry). cIMLE's model uncertainty does not fully encompass the entire strawberry. In contrast, CHIMLE generates diverse samples while assigning low model uncertainty to regions in the input that are less ambiguous and high uncertainty to regions that should contain more diversity, capturing a comprehensive and precise representation of uncertainty in the input data.

Figure 3 shows the model uncertainty comparison of different methods on the task of image decompression. As shown, BicycleGAN, MSGAN and DivCo produce excessively diverse samples, resulting in model uncertainty that spans across the entire image. Conversely, MoNCE lacks sample diversity and fails to capture the inherent ambiguity present in the input. DDRM struggles to generate reasonable samples in this example. NDM produces diverse samples but the generated samples are blurry. The IMLE-based methods, cIMLE and CHIMLE, show model uncertainties that effectively capture the ambiguous regions in the input, such as the edges of the human and the trees in the top right corner.

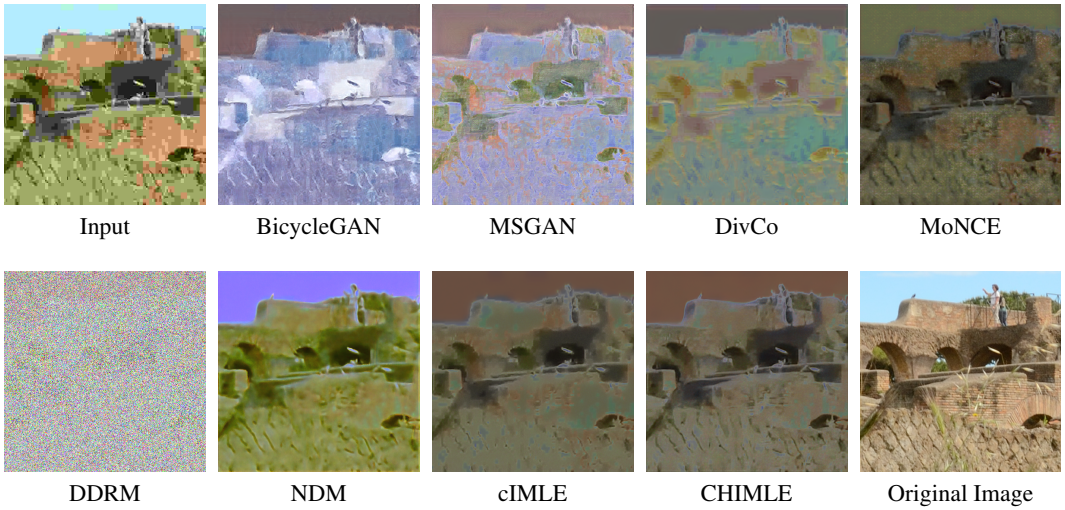

Figure 3: Model uncertainty comparison on image decompression.