Denoising diffusion models for out-of-distribution detection
Supplemental Material

Mark S. Graham
King's College London
mark.graham@kcl.ac.uk

Walter H.L. Pinaya
King's College London
walter.diaz_sanz@kcl.ac.uk

Petru-Daniel Tudosiu
King's College London
petru.tudosiu@kcl.ac.uk

Parashkev Nachev
University College London
p.nachev@ucl.ac.uk

Sebastien Ourselin
King's College London
sebastien.ourselin@kcl.ac.uk

M. Jorge Cardoso
King's College London
m.jorge.cardoso@kcl.ac.uk

A. Sample reconstructions on computer vision datasets.

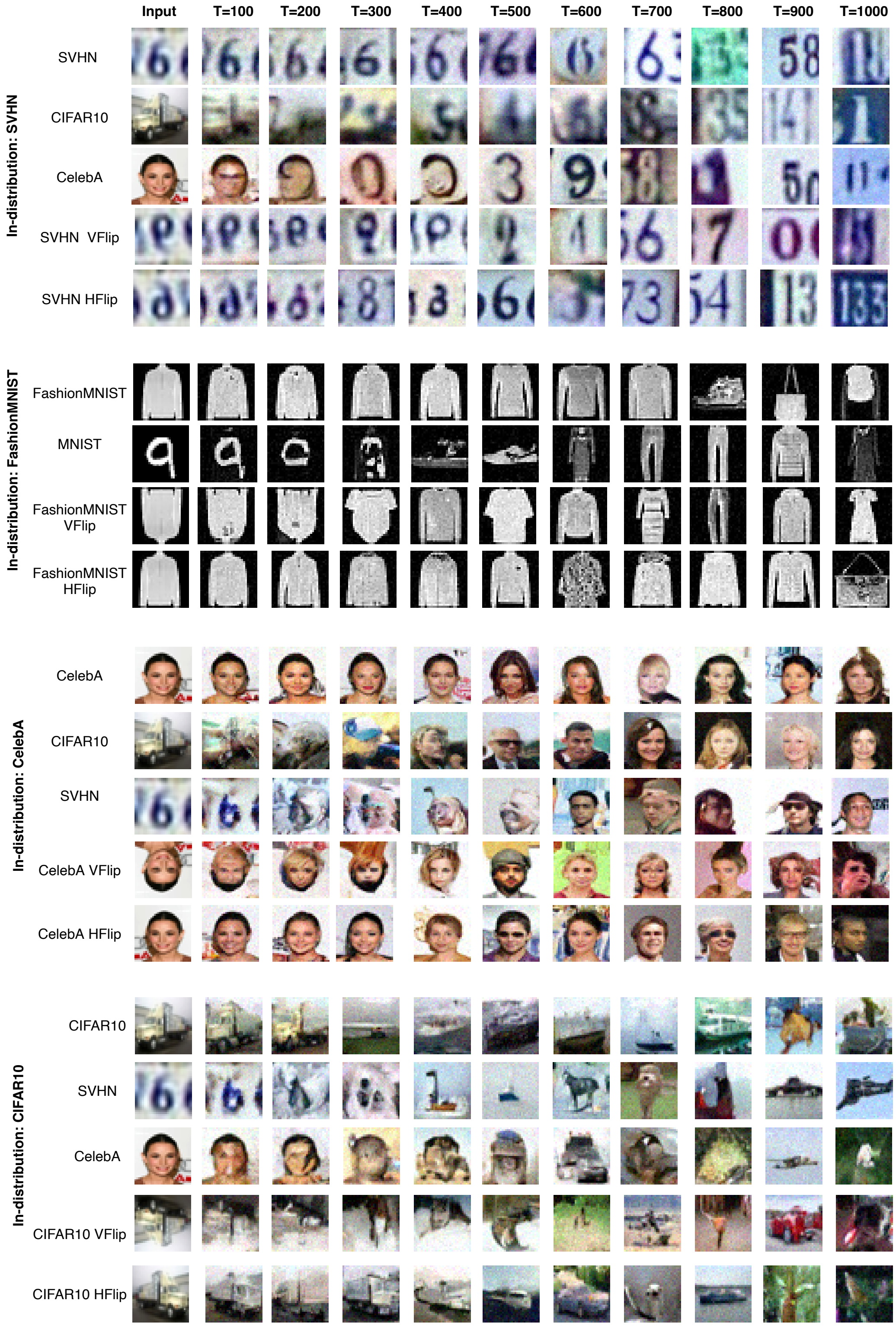

Figure 1. Example reconstructions from all models from ten different starting points spanning the full T-chain.

B. Sample reconstructions on medical datasets.

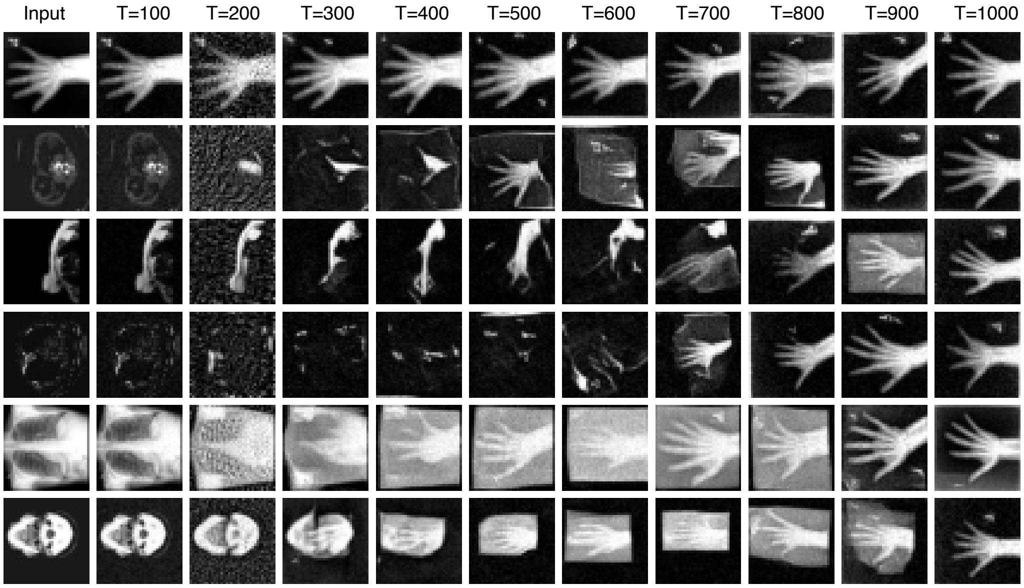

Figure 2. Example reconstructions from a model trained on MedNIST hand at $6x \times 64$ for ten different t-values spaced equally across the chain. Plot shows an in-distribution input (top row) and OOD inputs from Abdomen CT, Breast MRI, Chest CT, Chest X-ray, Head CT (rows 2-6).

	FashionMNIST			CIFAR-10				CelebA				SVHN			
	MNIST	VFlip	HFlip	SVHN	CelebA	VFlip	HFlip	CIFAR10	SVHN	VFlip	HFlip	CIFAR10	CelebA	VFlip	HFlip
AutoEncoder	75.0	59.0	**50.4**	**3.2**	**75.3**	50.1	**49.9**	42.1	2.7	53.2	**49.9**	**99.4**	**99.9**	49.9	**49.9**
AutoEncoder (4layer)	**95.1**	**71.9**	49.5	2.7	66.4	**50.4**	**49.9**	**65.2**	**6.7**	**68.5**	**49.9**	98.8	99.5	**50.3**	**49.9**
AutoEncoder Mahlabonis	94.9	**79.5**	63.0	4.5	**71.8**	50.9	**50.0**	64.2	9.6	71.6	**50.1**	**99.3**	**99.8**	50.6	**50.6**
AutoEncoder Mahlabonis (4layer)	**98.3**	77.7	**64.1**	**6.8**	69.0	**51.4**	**50.0**	**73.3**	**19.5**	**78.7**	**50.1**	96.9	98.6	**51.1**	**50.6**
MemAE (mem size rec)	56.9	59.0	48.7	4.21	**69.4**	50.3	49.9	51.5	5.8	56.4	49.9	98.6	99.5	49.8	49.7
MemAE (mem size 2000)	88.4	58.2	49.1	26.8	61.2	49.3	50.0	67.2	36.0	63.1	**50.0**	95.2	98.5	50.1	**50.7**
MemAE (mem size 50)	28.6	53.0	49.9	7.2	67.1	**50.7**	49.8	63.5	33.3	53.5	49.8	**98.7**	**99.6**	50.1	50.0
MemAE (4layer, mem size rec)	77.3	61.0	51.1	7.1	66.1	50.6	**50.1**	75.1	25.4	73.3	49.9	97.1	98.6	50.6	50.2
MemAE (4layer, mem size 2000)	**91.5**	**71.4**	**52.5**	**33.1**	66.4	**50.7**	**50.1**	**78.6**	**55.8**	**80.3**	**50.0**	89.3	94.5	**52.5**	50.4
MemAE (4layer, mem size 50)	87.6	65.2	51.3	7.0	62.3	50.6	50.0	71.7	23.8	64.4	**50.0**	97.2	98.7	50.3	50.2
AnoDDPM-Mod $t = 100$	76.6	79.5	62.2	23.6	55.8	53.2	50.3	79.2	52.7	77.3	49.7	**95.3**	**98.1**	46.8	48.1
AnoDDPM-Mod $t = 250$ (rec)	**91.8**	**81.0**	**64.2**	37.8	60.2	54.2	50.5	**80.2**	**67.3**	**78.1**	49.4	90.4	94.2	**50.2**	**52.7**
AnoDDPM-Mod $t = 500$	81.5	68.8	58.4	**51.0**	**64.1**	**54.6**	**50.8**	71.8	68.4	72.6	**50.9**	74.8	79.9	50.0	51.6

Table 1. These results show how performance on datasets vary as the information bottleneck is varied. Results show AUC scores. Bold text indicates highest value per column, done separately for each model class. Recommended memory size means 100 for FashionMNIST and 500 for other datasets, according to the settings in [1].

C. Effect of bottleneck size on the performance of reconstruction-based methods

Full results are shown in Table 1. For the AE the bottleneck is the size of the latent space; made smaller by increasing the number of layers in the AE. For the MemME we change both the size of the latent space and the size of the memory bank. For AnoDDPM-Mod we vary the amount of noise added to the image before reconstruction. Different bottleneck sizes could improve performance on some datasets, but at the expense of performance on others. For example, a 4-layer MemAE with a memory size of 2000 achieves 91.5 on FashionMNIST vs MNIST, substantially higher than the 3-layer model with recommended memory size, but this architecture's performance on SVHN vs CIFAR10 drops to 89.3, making it the poorest-performing of any model (not just MemAE models) on this pairing. The results highlight that tuning the information bottleneck can improve performance on a given dataset, but at the expense of performance on other datasets.

D. Performance of DDPM vs number of reconstructions

Results in Table 2, both for reducing the number of reconstructions performed and for reducing the maximum value of t that reconstructions are performed from, $\max_T$. The results demonstrate the number of reconstructions/model evaluations can be reduced substantially with limited effect on OOD performance.

Recons	Model evals	FashionMNIST			CIFAR10				CelebA				SVHN			
		MNIST	VFlip	HFlip	SVHN	CelebA	VFlip	HFlip	CIFAR10	SVHN	VFlip	HFlip	CIFAR10	CelebA	VFlip	HFlip
$\max_t = 1000$																
100	5050	97.4	88.6	65.1	97.9	68.5	63.2	50.5	99.0	100.0	93.3	50.3	99.0	99.6	58.2	61.6
50	2500	97.1	88.1	65.1	97.8	67.8	63.0	50.5	98.9	100.0	92.9	50.3	99.0	99.5	58.0	61.4
34	1717	97.0	87.3	64.5	97.7	67.3	62.9	50.7	98.8	100.0	92.7	50.5	98.8	99.4	57.8	60.9
25	1225	96.7	86.9	64.7	97.6	66.2	62.3	50.6	98.7	100.0	92.2	50.3	98.7	99.4	57.9	61.1
20	970	96.4	86.2	64.0	97.3	65.3	62.0	50.3	98.6	100.0	91.8	50.3	98.6	99.3	57.7	60.8
13	637	95.4	84.6	63.8	97.2	63.9	61.0	50.4	98.3	100.0	90.7	50.2	98.2	98.9	57.0	60.2
7	343	91.7	80.6	62.3	96.6	61.1	59.4	50.7	97.3	99.9	87.4	50.4	96.6	97.4	55.9	58.7
4	196	82.2	73.5	60.0	95.2	56.4	56.4	50.4	94.7	99.8	80.9	50.1	92.0	93.0	54.0	55.9
2	66	39.0	60.7	54.8	91.8	46.5	52.6	50.0	87.8	98.8	69.7	50.2	83.3	80.0	51.4	52.3
$\max_t = 800$																
80	3240	97.2	88.6	65.3	96.7	62.5	62.1	50.3	98.7	100.0	92.7	50.3	99.2	99.5	60.2	64.0
40	1600	96.9	88.1	65.3	96.6	62.0	62.0	50.4	98.6	100.0	92.4	50.3	99.0	99.4	59.9	63.7
27	1080	96.7	87.5	64.7	96.5	61.4	61.8	50.4	98.5	100.0	92.1	50.5	98.9	99.2	59.7	63.3
20	780	96.4	87.1	65.0	96.3	60.7	61.4	50.5	98.4	100.0	91.7	50.3	98.8	99.1	59.6	63.3
16	616	96.0	86.5	64.4	96.1	60.0	61.3	50.3	98.3	99.9	91.3	50.3	98.7	99.0	59.5	63.0
10	370	94.8	85.1	64.1	95.7	58.0	60.2	50.2	97.9	99.9	90.0	50.2	98.2	98.5	58.9	62.5
5	165	90.3	81.7	62.9	94.7	54.0	58.7	50.4	96.4	99.9	86.6	50.4	96.6	96.3	58.1	61.2
3	99	80.4	74.9	60.3	93.6	50.5	56.2	50.3	93.7	99.7	80.3	50.1	92.2	90.6	55.5	57.6
2	66	39.0	60.7	54.8	91.8	46.5	52.6	50.0	87.8	98.8	69.7	50.2	83.3	80.0	51.4	52.3
$\max_t = 600$																
60	1830	96.6	88.4	65.2	95.0	56.3	61.4	50.3	98.2	99.9	92.1	50.3	99.0	99.3	62.0	66.0
30	900	96.2	88.0	65.2	94.8	55.9	61.2	50.4	98.1	99.9	91.7	50.3	98.9	99.1	61.6	65.6
20	590	95.9	87.5	64.8	94.7	55.2	61.1	50.4	97.9	99.9	91.3	50.5	98.7	98.9	61.5	65.4
15	435	95.4	87.0	64.8	94.6	54.9	60.8	50.4	97.8	99.9	90.9	50.4	98.6	98.8	61.2	65.0
12	342	94.9	86.5	64.4	94.4	54.4	60.9	50.2	97.6	99.9	90.3	50.4	98.4	98.7	61.2	64.8
8	232	93.7	85.2	64.0	94.3	53.8	60.0	50.3	97.3	99.9	89.4	50.3	98.0	98.1	60.1	63.6
4	100	87.9	81.7	62.7	93.2	50.1	58.6	50.5	95.6	99.8	85.4	50.4	96.2	95.6	59.0	62.1
2	34	72.4	75.0	59.9	90.5	43.1	56.0	50.3	91.0	99.2	76.6	50.1	90.3	87.1	56.7	58.6
$\max_t = 400$																
40	820	93.2	87.6	64.3	92.9	50.5	61.4	50.3	97.5	99.9	90.7	50.4	98.8	99.2	63.6	67.5
20	400	92.3	87.1	64.1	92.7	50.1	61.2	50.4	97.3	99.9	90.1	50.4	98.7	99.0	63.2	66.9
14	287	92.3	86.7	63.8	92.8	50.1	61.2	50.3	97.2	99.9	89.8	50.5	98.5	98.8	62.9	66.7
10	190	90.2	86.0	63.6	92.4	49.1	60.8	50.5	97.0	99.8	89.0	50.5	98.4	98.7	62.7	66.2
8	148	88.8	85.3	63.3	92.3	48.6	60.7	50.1	96.7	99.8	88.3	50.2	98.2	98.4	62.5	65.8
5	85	84.4	83.7	62.6	91.7	47.0	59.9	50.4	96.0	99.7	86.4	50.3	97.6	97.6	61.7	64.7
3	51	79.8	80.6	61.6	91.3	45.5	58.4	50.3	94.4	99.6	82.7	50.2	95.7	94.7	59.8	62.5
2	34	72.4	75.0	59.9	90.5	43.1	56.0	50.3	91.0	99.2	76.6	50.1	90.3	87.1	56.7	58.6
$\max_t = 200$																
20	210	69.7	82.8	61.7	90.8	45.6	60.5	50.3	96.2	99.5	87.4	50.3	98.4	99.1	62.7	65.7
10	100	66.5	81.9	61.4	90.5	45.0	60.2	50.3	95.8	99.5	86.4	50.3	98.1	98.9	62.1	64.9
7	70	64.8	81.3	61.1	90.5	44.7	60.1	50.4	95.6	99.5	85.6	50.3	97.8	98.5	61.7	64.6
5	45	58.0	79.7	60.5	90.1	43.4	59.2	50.3	95.0	99.3	84.1	50.5	97.5	98.1	61.1	63.5
4	34	53.6	78.5	60.3	90.0	42.6	59.0	50.0	94.5	99.2	82.8	50.1	97.0	97.5	60.3	62.6
3	27	53.0	77.7	59.7	89.7	42.0	58.3	50.2	93.8	99.2	81.1	50.4	96.3	96.2	60.0	62.4
2	18	46.3	74.6	58.4	89.0	40.2	57.0	50.1	91.6	99.0	76.5	50.3	93.8	91.9	58.5	60.7

Table 2. Variation in the performance of the DDPM OOD detection as the number of reconstructions used is changed. Values reported for each dataset pairing ar AUCs. The first row ($\max_t = 1000$, 100 reconstructions) are the parameters for the results reported in Table 1.

References

[1] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, pages 1705–1714, 2019. 4