TTTFlow: Unsupervised Test-Time Training with Normalizing Flow
Supplementary Material

David Osowiechi* , Gustavo A. Vargas Hakim*,
Mehrdad Noori , Milad Cheraghalikhani , Ismail Ben Ayed , and Christian Desrosiers

École de Technologie Supérieure, Montréal, Canada

david.osowiechi.1@ens.etsmtl.ca, gustavo-adolfo.vargas-hakim.1@ens.etsmtl.ca,
mehrdad.noori.1@ens.etsmtl.ca,milad.cheraghalikhani.1@ens.etsmtl.ca,
ismail.benayed@etsmtl.ca, christian.desrosiers@etsmtl.ca

1. Implementation details

RealNVP is a Normalizing Flow model based on coupling layers (see Fig. 1). The *checkerboard* masking is used to separate the inputs into two equal parts. The first part is used to compute the affine parameters to scale and translate the second part. Coupling layers alternate the masking pattern when they are stacked. In this work, RealNVP [1] consists of three coupling layers, each followed by Activation Normalization. The computation of the affine parameters is done through a small ResNet model [2] consisting of two standard bottleneck resblocks.

2. Ablation Studies

In this section, ablation studies for TTTFlow are performed. Particularly, we focus on varying the complexity of the Normalizing Flow, as it is the main component in storing the source distribution for its utilization at test-time.

The Normalizing Flow is based on RealNVP [1], and has three coupling layers (TTTFlow-3C), each followed by Activation Normalization. The architecture choice of the Normalizing Flow affects the way in which the normal distribution is modeled from input data, for which we evaluate two variations with only two and one coupling layers. Accuracy is measured on CIFAR-10-C with 20 iterations. As shown in Table 1, smaller variations of the Normalizing Flow used in TTTFlow cause important degradation in performance. A larger model than the original version (i.e. TTTFlow-3C) causes a memory overflow, thus this module needs to remain inside a feasible range of complexity.

*Equal contribution

Table 1: Accuracy values after 20 iterations of adaptation on all the corruptions of CIFAR-10-C (level 5) with different versions of the RealNVP-based Normalizing Flow in TTTFlow. Variants are referred as *TTTFlow-nC*, being n the number of coupling layers.

	Accuracy
TTTFlow-1C	0.1291
TTTFlow-2C	0.1188
TTTFlow-3C	**0.7311**

3. Additional results

This section provides additional results validating our TTTFlow method on the CIFAR-10-C dataset.

3.1. Detailed predictions of TTTFlow

In Table 2, we present a more detailed analysis on the predictions of TTTFlow before and after adaptation across executions. We use 20 iterations, as it yields the best trade-off between accuracy and execution costs. It can be observed that, generally, the number of originally misclassified samples that become correctly classified (Bad/Good) is larger than the number of correctly classified samples which become misclassified (Good/Bad). This is of paramount importance as the domain shift detector should not decrease the formerly-obtained classification efficiency.

3.2. Results for corruption levels 1 to 4

Next, we report the results for the corruption levels of severity 1 to 4.

Comparison to methods using a classifier trained with only $\mathcal{L}_{\text{cls}}$ As shown in Tables 3, 4, 5 and 6, for all corrup-

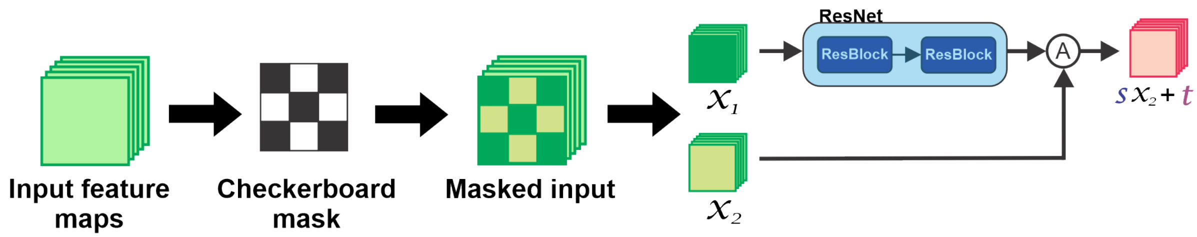

Figure 1: Coupling layer of RealNVP. The *checkerboard* masking is used to separate the input into two equally shaped parts. The first part $\mathbf{x}_1$ is used to compute affine parameters s and t through a compact ResNet-based model, and the second part $\mathbf{x}_2$ is undergone through an affine transformation using the aforementioned parameters.

tions levels, TTTFlow achieves a higher average accuracy compared to the pretrained ResNet50 baseline as well as to other Test-Time architectures. The corruption *JPEG Compression* is still challenging, since the pretrained ResNet50 Baseline outperforms Test-Time architectures at every level for this corruption type.

Comparison with TTT++ on baseline trained with $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ssl}}$ As in the main paper, we also applied our NF model on of the second layer of the same encoder as TTT++, which was trained using an auxiliary contrastive learning loss in addition to the cross-entropy loss. As seen in Tables 3, 4, 5 and 6, higher accuracy scores are obtained using TTTFlow with the same feature extractor as TTT++. The accuracy of TTT++ largely decreases across levels, whilst TTTFlow has a more stable behavior regardless of the corruption severity. These results further demonstrate that a Normalizing Flow can be used on top of any feature extractor to detect domain shift in an unsupervised way, but also that its capacity increases considerably when a strongly pretrained encoder is available.

References

[1] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

[5] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: fully test-time adaptation by entropy minimization. *arXiv:2006.10726 [cs, stat]*, Mar. 2021. arXiv: 2006.10726.

	Good/Good	Good/Bad	Bad/Good	Bad/Bad	Accuracy (%)
Gaussian Noise	1450.0 ±15.7	625.0 ±15.7	4723.2 ±20.1	3201.8 ±20.1	61.73 ±0.35
Shot Noise	1976.8 ±11.2	653.2 ±11.2	4531.0 ±15.1	2839.0 ±15.1	65.08 ±0.14
Impulse Noise	2001.6 ±3.4	646.4 ±3.4	3846.2 ±9.0	3505.8 ±9.2	58.48 ±0.12
Defocus Blur	2622.4 ±6.41	286.6 ±6.4	5852.4 ±14.4	1238.6 ±14.4	84.75 ±0.17
Glass Blur	3589.0 ±6.7	685.4 ±6.7	2603.4 ±13.0	3121.6 ±13.0	61.93 ±0.12
Motion Blur	4907.2 ±7.9	430.8 ±7.9	3324.2 ±6.3	1337.8 ±6.3	82.31 ±0.10
Zoom Blur	3137.2 ±8.0	294.8 ±8.0	5445.2 ±13.0	1122.8 ±13.0	85.82 ±0.17
Snow	5916.2 ±16.3	588.8 ±16.3	1794.8 ±10.7	1700.2 ±10.7	77.11 ±0.24
Frost	4442.6 ±11.0	581.4 ±11.0	3193.4 ±9.1	1782.6 ±9.1	76.36 ±0.16
Fog	5098.0 ±14.7	467.0 ±14.7	2960.0 ±6.3	1475.0 ±6.3	80.58 ±0.15
Brightness	8169.8 ±17.2	434.2 ±17.2	653.8 ±12.0	742.2 ±12.0	88.24 ±0.12
Contrast	2805.0 ±8.2	334.0 ±8.2	5615.0 ±11.5	1246.0 ±11.5	84.20 ±0.18
Elastic Transform	5486.2 ±19.4	667.8 ±19.4	1723.0 ±11.0	2123.0 ±11.0	72.09 ±0.20
Pixelate	3233.2 ±9.5	333.8 ±9.5	4416.4 ±14.8	2016.6 ±14.8	76.50 ±0.13
Jpeg Compression	5812.8 ±23.8	664.2 ±23.8	1126.6 ±17.0	2396.4 ±17.0	69.39 ±0.25

Table 2: Detailed breakdown of number of samples well predicted before adaptation and after 20 iterations (Good/Good), well predicted before and badly predicted after 20 iterations (Good/Bad), badly predicted before adaptation and well predicted after 20 iterations (Bad/Good), badly predicted before and after 20 iterations (Bad/Bad).

	Encoder trained with $\mathcal{L}_{\text{cls}}$ only				Encoder trained with $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ssl}}$	
	Baseline	TENT [5]	TTT [4]	TTTFlow	TTT++ [3]	TTTFlow
Gaussian Noise	59.05	47.67 ±0.31	65.15 ±0.12	**65.71 ±0.17**	78.70 ±4.28	**81.12 ±0.06**
Shot Noise	66.78	47.90 ±0.11	69.84 ±0.23	**70.56 ±0.11**	80.12 ±0.12	**83.92 ±0.10**
Impulse Noise	57.17	40.60 ±0.11	66.02 ±0.24	**66.55 ±0.08**	70.64 ±0.53	**74.50 ±0.08**
Defocus Blur	73.17	61.62 ±0.08	88.61 ±0.17	**88.85 ±0.12**	81.75 ±0.43	**91.98 ±0.03**
Glass Blur	52.23	41.10 ±0.14	60.44 ±0.34	**61.42 ±0.11**	62.85 ±0.50	**71.25 ±0.14**
Motion Blur	71.14	57.99 ±0.19	80.75 ±0.12	**84.94 ±0.10**	68.42 ±1.08	**87.66 ±0.06**
Zoom Blur	70.63	60.06 ±0.12	86.87 ±0.17	**87.69 ±0.16**	70.74 ±2.05	**93.09 ±0.05**
Snow	78.35	55.50 ±0.17	77.67 ±0.12	**79.66 ±0.18**	52.43 ±0.56	**83.99 ±0.08**
Frost	77.19	53.5 ±0.11	81.05 ±0.12	**81.28 ±0.11**	52.80 ±2.67	**88.16 ±0.09**
Fog	81.18	58.58 ±0.24	88.54 ±0.11	**88.66 ±0.10**	41.75 ±0.09	**84.87 ±0.05**
Brightness	90.26	66.34 ±0.16	90.96 ±0.04	**91.48 ±0.10**	50.95 ±2.19	**93.42 ±0.06**
Contrast	61.11	57.20 ±0.10	**88.55 ±0.10**	88.02 ±0.12	45.28 ±0.55	**93.22 ±0.05**
Elastic Transform	78.57	55.38 ±0.05	80.58 ±0.17	**80.64 ±0.16**	35.53 ±1.51	**86.81 ±0.09**
Pixelate	66.65	56.23 ±0.24	81.08 ±0.13	**82.63 ±0.20**	33.64 ±0.83	**91.08 ±0.07**
Jpeg Compression	**83.82**	56.96 ±0.20	72.96 ±0.16	73.32 ±0.14	28.01 ±1.75	**87.81 ±0.10**
Average	71.15	54.44	78.60	**79.43**	56.91	**86.19**

Table 3: Accuracy (%) on CIFAR-10-C dataset with Level 4 corruption for TTTFlow compared to ResNet50, TENT, TTT, and TTT++ with different encoders. Mean and standard deviation are reported over 5 runs.

	Encoder trained with $\mathcal{L}_{\text{cls}}$ only				Encoder trained with $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ssl}}$	
	Baseline	TENT [5]	TTT [4]	TTTFlow	TTT++ [3]	TTTFlow
Gaussian Noise	64.92	50.39 ±0.25	68.96 ±0.11	**69.14 ±0.28**	80.29 ±0.81	**83.78 ±0.04**
Shot Noise	71.98	52.41 ±0.19	73.40 ±0.25	**73.77 ±0.29**	82.46 ±0.37	**86.31 ±0.04**
Impulse Noise	72.58	48.32 ±0.05	74.10 ±0.15	**74.58 ±0.18**	79.20 ±0.38	**82.62 ±0.05**
Defocus Blur	84.68	64.42 ±0.14	90.55 ±0.11	**90.89 ±0.11**	87.68 ±0.38	**93.36 ±0.11**
Glass Blur	66.68	46.66 ±0.08	70.96 ±0.10	**71.42 ±0.13**	72.52 ±0.56	**82.75 ±0.04**
Motion Blur	71.61	57.14 ±0.13	81.48 ±0.14	**85.11 ±0.09**	69.59 ±1.38	**87.99 ±0.06**
Zoom Blur	75.77	61.71 ±0.09	87.64 ±0.12	**88.41 ±0.14**	73.23 ±2.33	**93.14 ±0.02**
Snow	80.80	57.09 ±0.13	79.47 ±0.12	**81.70 ±0.16**	57.96 ±1.02	**85.75 ±0.06**
Frost	77.54	53.60 ±0.16	**80.98 ±0.11**	80.94 ±0.12	49.94 ±3.53	**88.87 ±0.05**
Fog	86.37	61.35 ±0.14	90.33 ±0.06	**90.62 ±0.09**	52.89 ±4.13	**89.55 ±0.03**
Brightness	91.41	66.44 ±0.23	91.34 ±0.16	**91.96 ±0.07**	57.96 ±1.32	**93.75 ±0.03**
Contrast	77.92	59.37 ±0.25	**89.89 ±0.14**	89.84 ±0.06	53.44 ±2.37	**93.50 ±0.06**
Elastic Transform	82.84	60.20 ±0.22	86.70 ±0.11	**86.77 ±0.17**	36.49 ±3.72	**91.16 ±0.06**
Pixelate	82.16	60.12 ±0.14	85.65 ±0.15	**86.86 ±0.07**	33.41 ±3.02	**92.31 ±0.02**
Jpeg Compression	**85.43**	59.04 ±0.20	75.00 ±0.13	75.66 ±0.06	28.82 ±2.74	**89.14 ±0.04**
Average	78.18	57.22	81.76	**82.51**	61.06	**88.93**

Table 4: Accuracy (%) on CIFAR-10-C dataset with Level 3 corruption for TTTFlow compared to ResNet50, TENT, TTT, and TTT++ with different encoders. Mean and standard deviation are reported over 5 runs.

	Encoder trained with $\mathcal{L}_{\text{cls}}$ only				Encoder trained with $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ssl}}$	
	Baseline	TENT [5]	TTT [4]	TTTFlow	TTT++ [3]	TTTFlow
Gaussian Noise	**76.26**	55.07 ±0.38	75.65 ±0.14	75.96 ±0.27	85.41 ±2.26	**88.29 ±0.12**
Shot Noise	**84.73**	57.23 ±0.13	81.74 ±0.15	82.15 ±0.11	88.79 ±0.44	**90.74 ±0.07**
Impulse Noise	79.87	51.75 ±0.15	79.69 ±0.15	**80.16 ±0.08**	84.27 ±0.29	**86.99 ±0.07**
Defocus Blur	90.41	65.09 ±0.15	91.62 ±0.14	**91.66 ±0.07**	90.85 ±0.42	**93.68 ±0.06**
Glass Blur	63.14	46.47 ±0.19	70.82 ±0.26	**71.70 ±0.09**	71.60 ±1.95	**83.14 ±0.11**
Motion Blur	79.25	59.73 ±0.13	85.35 ±0.07	**87.60 ±0.20**	77.38 ±1.12	**90.55 ±0.04**
Zoom Blur	81.12	62.69 ±0.18	88.82 ±0.09	**89.57 ±0.08**	80.30 ±1.45	**93.43 ±0.08**
Snow	78.63	55.49 ±0.20	78.33 ±0.14	**81.32 ±0.17**	68.56 ±1.36	**88.22 ±0.06**
Frost	84.65	59.73 ±0.24	85.56 ±0.13	**85.88 ±0.04**	63.66 ±3.39	**91.49 ±0.03**
Fog	90.08	63.90 ±0.25	91.72 ±0.10	**91.80 ±0.19**	64.26 ±3.37	**91.92 ±0.04**
Brightness	92.05	67.01 ±0.16	91.76 ±0.07	**92.61 ±0.04**	67.19 ±1.23	**93.88 ±0.05**
Contrast	85.35	63.05 ±0.14	**90.75 ±0.09**	90.66 ±0.12	62.90 ±1.93	**93.68 ±0.05**
Elastic Transform	87.56	63.19 ±0.18	87.84 ±0.08	**87.85 ±0.04**	50.06 ±2.37	**91.48 ±0.05**
Pixelate	86.89	62.50 ±0.1	86.56 ±0.15	**87.77 ±0.08**	43.33 ±3.31	**92.69 ±0.05**
Jpeg Compression	**86.76**	61.28 ±0.14	77.51 ±0.22	78.15 ±0.17	28.26 ±2.78	**90.07 ±0.06**
Average	83.12	59.61	84.25	**84.99**	68.45	**90.68**

Table 5: Accuracy (%) on CIFAR-10-C dataset with Level 2 corruption for TTTFlow compared to ResNet50, TENT, TTT, and TTT++ with different encoders. Mean and standard deviation are reported over 5 runs.

	Encoder trained with $\mathcal{L}_{\text{cls}}$ only				Encoder trained with $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{ssl}}$	
	Baseline	TENT [5]	TTT [4]	TTTFlow	TTT++ [3]	TTTFlow
Gaussian Noise	**85.59**	60.67 $\pm$0.14	83.38 $\pm$0.08	83.71 $\pm$0.21	90.14 $\pm$1.05	**91.60 $\pm$0.05**
Shot Noise	**89.01**	61.03 $\pm$0.21	86.02 $\pm$0.08	86.44 $\pm$0.10	90.89 $\pm$0.29	**92.46 $\pm$0.02**
Impulse Noise	**87.50**	58.01 $\pm$0.15	84.75 $\pm$0.08	85.55 $\pm$0.07	87.76 $\pm$0.06	**90.03 $\pm$0.06**
Defocus Blur	92.31	66.77 $\pm$0.18	92.38 $\pm$0.03	**92.80 $\pm$0.04**	91.51 $\pm$0.48	**93.87 $\pm$0.04**
Glass Blur	62.79	47.40 $\pm$0.13	70.72 $\pm$0.15	**71.50 $\pm$0.14**	72.12 $\pm$2.13	**83.04 $\pm$0.05**
Motion Blur	87.08	64.63 $\pm$0.02	89.05 $\pm$0.10	**89.60 $\pm$0.12**	84.11 $\pm$0.91	**92.25 $\pm$0.01**
Zoom Blur	84.12	64.12 $\pm$0.37	88.65 $\pm$0.08	**89.75 $\pm$0.13**	81.76 $\pm$1.38	**92.97 $\pm$0.06**
Snow	87.87	62.40 $\pm$0.14	86.10 $\pm$0.08	**88.05 $\pm$0.07**	75.89 $\pm$0.75	**91.53 $\pm$0.04**
Frost	**89.52**	64.40 $\pm$0.08	88.54 $\pm$0.10	89.08 $\pm$0.09	71.54 $\pm$3.13	**93.07 $\pm$0.06**
Fog	92.17	67.52 $\pm$0.17	92.36 $\pm$0.05	**92.74 $\pm$0.10**	70.58 $\pm$1.29	**93.54 $\pm$0.04**
Brightness	92.50	68.60 $\pm$0.27	92.27 $\pm$0.02	**93.00 $\pm$0.03**	64.40 $\pm$2.69	**93.99 $\pm$0.04**
Contrast	91.63	67.14 $\pm$0.13	92.05 $\pm$0.12	**92.44 $\pm$0.07**	53.60 $\pm$3.80	**93.96 $\pm$0.05**
Elastic Transform	87.25	63.51 $\pm$0.11	87.34 $\pm$0.04	**87.71 $\pm$0.09**	39.92 $\pm$1.52	**91.21 $\pm$0.03**
Pixelate	**90.80**	65.75 $\pm$0.16	89.16 $\pm$0.15	90.42 $\pm$0.05	36.04 $\pm$3.47	**93.16 $\pm$0.08**
Jpeg Compression	**89.73**	64.63 $\pm$0.11	82.64 $\pm$0.12	83.38 $\pm$0.13	30.90 $\pm$1.18	**92.34 $\pm$0.06**
Average	87.32	63.11	87.02	**87.74**	69.41	**91.93**

Table 6: Accuracy (%) on CIFAR-10-C dataset with Level 1 corruption for TTTFlow compared to ResNet50, TENT, TTT, and TTT++ with different encoders. Mean and standard deviation are reported over 5 runs.