A Appendix

A.1 Proof of Theorem 1

Proof. Due to rotational symmetry, it suffices to consider the case where $U = I$, thus $\Sigma = D$.

We start by computing the derivatives of the corresponding decomposition:

$$\tilde{x}\tilde{x}^\top - \Sigma =: \dot{M}(\alpha) = \dot{V}(\alpha) \cdot S(\alpha) \cdot V^\top(\alpha) + V(\alpha) \cdot \dot{S}(\alpha) \cdot V^\top(\alpha) + V(\alpha) \cdot S(\alpha) \cdot \dot{V}^\top(\alpha).$$

From the fact that $V^\top(\alpha) \cdot V(\alpha) = I$ and the above display, we get:

$$\begin{gathered} \dot{V}^\top(\alpha)V(\alpha) + V^\top(\alpha)\dot{V}(\alpha) = 0 \\ \dot{S}(\alpha) + V^\top(\alpha) \cdot \dot{V}(\alpha) \cdot S(\alpha) + S(\alpha) \cdot \dot{V}^\top(\alpha) \cdot V(\alpha) = V^\top(\alpha) \cdot \dot{M}(\alpha) \cdot V(\alpha). \end{gathered}$$

The previous display now implies that $V^\top(\alpha) \cdot \dot{V}(\alpha)$ is skew-symmetric and hence, we get:

$$\begin{aligned} \dot{S}(\alpha) &= I \odot (V^\top(\alpha) \cdot \dot{M}(\alpha) \cdot V(\alpha)) \\ \forall i \neq j : (V^\top(\alpha) \cdot \dot{V}(\alpha))_{ij} &= \frac{(V^\top(\alpha) \cdot \dot{M}(\alpha) \cdot V(\alpha))_{ij}}{\sigma_j(\alpha) - \sigma_i(\alpha)}. \end{aligned}$$

Now, we restrict to the setting where $\alpha = 0$ where $V(\alpha) = I$ and we get:

$$\dot{V}_{i1} = \frac{(\tilde{x}\tilde{x}^\top - \Sigma)_{i1}}{\sigma_1 - \sigma} = \begin{cases} 0 & \text{if } i = 1 \\ \frac{\tilde{x}_1 \cdot \tilde{x}_i}{\sigma_1 - \sigma} & \text{o.w} \end{cases}.$$

Finally, we compute:

$$\begin{aligned} \mathbb{E}[\langle \dot{V}(0), r_1 \rangle] = \langle \mathbb{E}[\dot{V}(0)], r_1 \rangle &= \frac{1}{\sigma_1 - \sigma} \cdot \left(r_1^\top \cdot R \cdot \Sigma \cdot R^\top \cdot e_1 - r_{11} \cdot e_1^\top \cdot R \cdot \Sigma \cdot R^\top \cdot e_1 \right) \\ &= \frac{1}{\sigma_1 - \sigma} \cdot \left(\sigma_1 \cdot r_{11} - r_{11}(\sigma + (\sigma_1 - \sigma) \cdot r_{11}^2) \right) \\ &= r_{11} \cdot (1 - r_{11}^2). \end{aligned}$$

The above is greater than 0 when $0 < r_{11} < 1$. ☐

Remark on the proof. Since $\|r_1\| = 1$, the assumption that $0 < r_{11} < 1$ is very mild: it is satisfied when there exists any other nonzero entry in r_1, that is, when there exists any transformation on the first principle component between $\tilde{x}$ and x.

A.2 Additional Experiments on ImageNet-C

In all experiments, unless stated otherwise, we apply ViT-probing with the same hyper-parameters as described in the main paper, and report the accuracy (%) for ImageNet-C level-5 corruptions *after 10 steps* of TTT (instead of 20, for faster experiments).

Results on all levels. Tables 5-9 present the results before and after TTT on all 5 corruption levels.

Normalized pixels. Following [19], we experiment with two kinds of reconstruction loss: 1) MSE between the reconstructed pixels and the original pixels; (2) MSE where the target is the pixel values normalized across each masked patch. As shown in Table 10, using normalized pixels as the reconstruction target improves representation quality on most of the corruptions. The two pre-trained models for the two types of loss functions are taken from the official git repository of [19].

Training encoder only vs. encoder and decoder. We experiment with two optimization procedures for the TTT: 1) optimize both the encoder and decoder weights; 2) optimize only the encoder, while freezing the decoder weights. In both cases, the mask token and the class token are optimized as together. As shown in Table 11, the differences are negligible.

Masking ratio for TTT. Table 12 presents the results for training with different masking ratios on the test images. For pre-training, all results use a fixed masking ratio of 75%.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Baseline	78.5	74.5	68.1	73.9	70.5	70.6	74.8	68.6	72.3	73.0	75.2	75.9	73.6	69.3	63.7
TTT-MAE	78.9	74.7	72.5	74.7	72.9	72.2	76.8	72.2	75.5	74.5	75.8	77.0	75.9	71.9	69.3

Table 5: Accuracy (%) on ImageNet-C, level 1, after 10 steps of TTT.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Baseline	77.4	71.2	62.3	51.0	66.3	58.4	68.6	59.2	64.9	70.4	70.6	74.7	66.2	54.2	55.2
TTT-MAE	77.8	71.5	69.4	49.7	69.8	62.7	72.5	66.4	70.0	72.7	72.3	76.2	70.6	58.7	63.6

Table 6: Accuracy (%) on ImageNet-C, level 2, after 10 steps of TTT.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Baseline	75.8	62.7	49.5	67.1	59.8	47.6	57.1	35.0	57.4	68.6	60.2	70.1	54.3	54.7	48.0
TTT-MAE	75.8	64.4	59.4	71.2	64.0	54.0	63.6	50.7	64.2	71.3	64.2	73.1	61.8	58.0	57.4

Table 7: Accuracy (%) on ImageNet-C, level 3, after 10 steps of TTT.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Baseline	73.1	33.1	35.8	56.9	54.2	45.2	39.6	26.0	38.2	62.0	43.2	60.3	32.2	44.2	40.7
TTT-MAE	72.7	39.6	45.7	64.9	58.3	52.6	48.5	42.8	47.6	67.0	50.5	66.6	42.4	45.7	51.5

Table 8: Accuracy (%) on ImageNet-C, level 4, after 10 steps of TTT.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Baseline	68.3	6.4	24.2	31.6	38.6	38.4	17.4	18.4	18.2	51.2	32.2	49.7	18.2	35.9	32.2
TTT-MAE	67.7	9.2	31.7	45.5	42.8	46.3	25.1	31.8	27.0	59.9	39.5	59.9	27.0	37.9	42.9

Table 9: Accuracy (%) on ImageNet-C, level 5, after 10 steps of TTT.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Base w/o norm.	66.7	5.6	25.7	40.4	33.7	37.7	13.2	20.9	15.5	48.0	39.6	47.0	15.3	37.6	32.7
Base w/ norm.	**68.3**	6.4	24.2	31.6	38.6	38.4	17.4	18.4	18.2	51.2	32.2	49.7	18.2	35.9	32.2
Ours w/o norm.	**68.3**	5.7	31.5	**47.8**	33.3	41.5	19.2	25.3	21.5	54.5	**43.5**	54.1	22.1	**44.3**	39.5
Ours w/ norm.	67.7	**9.2**	**31.7**	45.5	**42.8**	**46.3**	**25.1**	**31.8**	**27.0**	**59.9**	39.5	**59.9**	**27.0**	37.9	**42.9**

Table 10: MSE loss with normalized pixels vs. without. Base: baseline. Ours: TTT-MAE.

	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
Encoder	67.7	9.2	31.7	45.5	42.8	46.3	25.1	31.8	27.0	59.9	39.5	59.9	27.0	37.9	42.9
Both	67.8	9.1	31.6	45.4	43.3	46.4	25.1	31.8	27.0	59.9	39.7	59.9	27.0	38.3	43.0

Table 11: Training only the encoder vs. both the encoder and decoder.

Mask ratio	brigh	cont	defoc	elast	fog	frost	gauss	glass	impul	jpeg	motn	pixel	shot	snow	zoom
50%	68.3	11.4	29.0	39.7	44.4	46.3	27.1	25.1	29.3	59.6	34.9	57.8	29.5	40.3	36.8
75%	67.7	9.2	31.7	45.5	42.8	46.3	25.1	31.8	27.0	59.9	39.5	59.9	27.0	37.9	42.9
90%	62.7	6.2	24.2	43.0	38.3	40.2	21.4	28.3	23.3	54.0	35.6	54.7	22.5	27.2	40.8

Table 12: Different masking ratios during TTT.

Dataset	License
ImageNet-C	Creative Commons Attribution 4.0 International
ImageNet-A	MIT License
ImageNet-R	MIT License
Yearbook Dataset	MIT License

Table 13: Licenses of the datasets used in our paper.

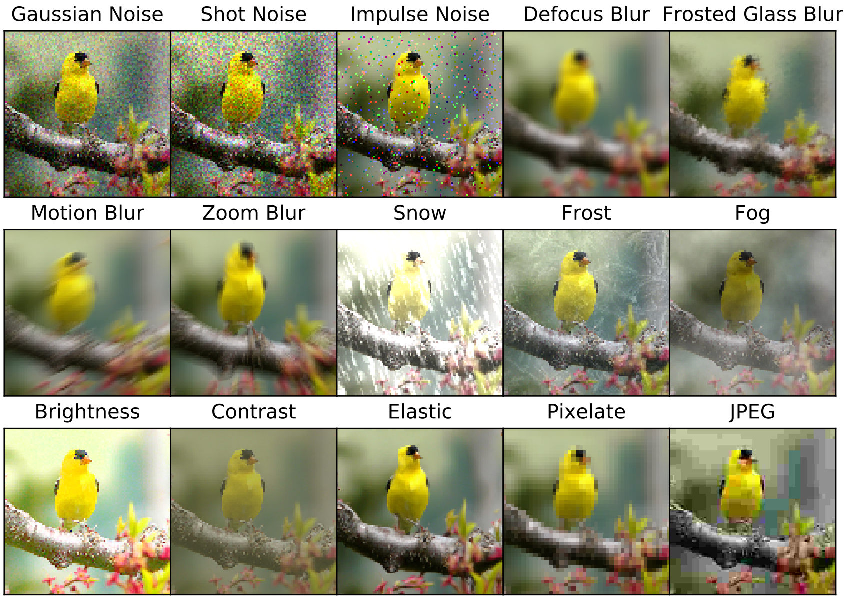

Figure 4: Sample images from the ImageNet-C benchmark, as shown in the original paper [22].