LOOKSHARP: ATTENTION ENTROPY MINIMIZATION FOR TEST-TIME ADAPTATION

Yash Mali[1], **Evan Shelhamer**[1,2]
University of British Columbia, Department of Computer Science[1]
Vector Institute[2]
ymali@mail.ubc.ca

ABSTRACT

Test-time adaptation (TTA) updates models during inference to reduce error on distribution shifts. Given the established test-time loss of entropy minimization over model predictions, we propose a new test-time loss of attention entropy minimization, over the distributions computed by self-attention in the model. We propose *LookSharp* to minimize the entropy of the CLS-to-patch attention in the final layer of a ViT model (Dosovitskiy et al., 2021) and maintain focused attention on shifted data. We show that our attention entropy minimization improves robustness and is complementary to output entropy minimization on ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-R (Hendrycks et al., 2021).

1 INTRODUCTION AND RELATED WORK

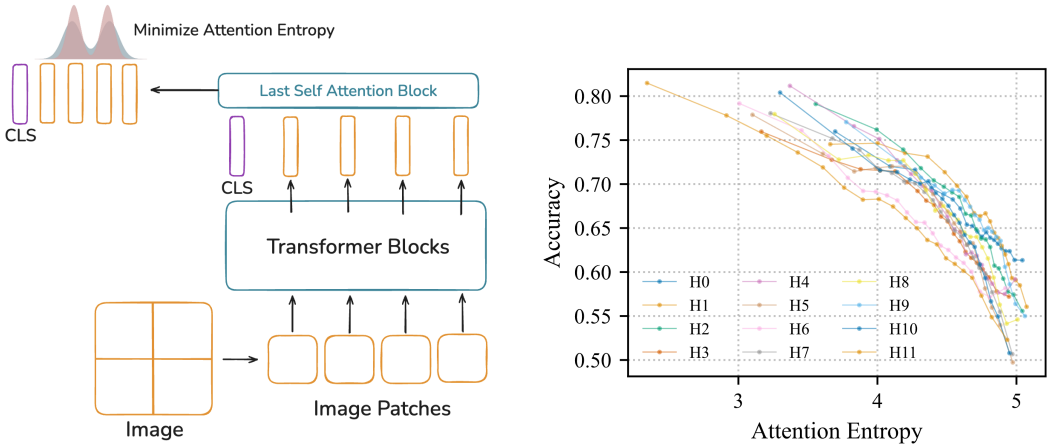
Deep networks achieve impressive performance on in-distribution data but can fail catastrophically when deployed on data from shifted distributions (Hendrycks & Dietterich, 2019). To mitigate shift, TTA by entropy minimization optimizes over the output distribution, which encourages confident predictions, and can improve generalization at test time (Wang et al., 2021). While effective, this output entropy minimization loss treats the model as a black box, and does not make use of the feature extractor representations that could potentially help adaptation. Vision Transformers (ViTs) (Dosovitskiy et al., 2021), which are now the dominant architecture for visual recognition due to their scalability, offer attention distributions that can measure the importance of spatial locations for model representations (Fuller et al., 2025).

We harness these attention distributions for TTA, minimizing the entropy of the attention distributions in vision transformers as an unsupervised loss to update the model parameters. As this sharpens the distribution to focus more on fewer tokens, we call our method *LookSharp*. Specifically, we minimize the entropy of the distribution defined by the attention scores from the CLS token to the patch tokens of the last layer across attention heads. Our approach is motivated by two key observations. First, Figure 1 (b) shows that accuracy drops sharply when the attention entropy is too diffuse. Second, self-supervised ViTs like DINOv3 can learn spatially-localized, object-centric, and interpretable attention maps through large-scale training on internet data (Siméoni et al., 2025).

We evaluate LookSharp on adaptation to corruptions on ImageNet-C in the batch episodic setting. That is, the model updates and then resets on each batch. We also show that combining attention entropy and output entropy leads to further improvement.

Entropy Minimization for Adaptation. Test-time adaptation often relies on entropy minimization. Tent (Wang et al., 2021) updates normalization layer statistics and parameters to minimize output entropy. MEMO (Zhang et al., 2022) extends this by using test-time augmentation to create a batch from a single sample and updates all parameters episodically using the same loss as Tent. Other works like SAR (Niu et al., 2023) and EATA (ETA) (Niu et al., 2022) use output entropy combined with sharpness-aware minimization, data filtering, and anchoring to the source model using regularization of the parameters.

Attention for Adaptation. There has been less use of attention for updates. **Attent** (Kojima et al., 2023) aligns test-time attention statistics with stored source statistics. Unlike Attent, our method is



(a) Summary of the Method

(b) Attention Entropy vs Accuracy across Heads

Figure 1: **Left** (a) Our method: attention entropy minimization. We minimize the entropy of the attention distribution from CLS to patch tokens at test time. We combine this with output entropy minimization for best results. **Right** (b) Visualization of attention entropy and accuracy. The entropy of final-layer CLS to patch token attention and the accuracy on shifted data are shown across all heads on a 10% sample of ImageNet-C for the unadapted DINOv3-Base model. Higher entropy (x-axis, right) tends toward lower accuracy (y-axis, bottom).

purely test-time and does not require storing source statistics. Instead, it relies on the confidence of attention during inference alone. We therefore only compare with other fully test-time updates.

2 Method: Attention Entropy Minimization

Given a model f_θ trained on a source distribution $\mathcal{D}_{source}$, we encounter a batch from a shifted distribution $\mathcal{D}_{shift}$ at test time. For each test batch, $\mathcal{B} = \{x_i\}_{i=1}^{B}$ where $x_i \sim \mathcal{D}_{shift}$, we aim to:

1. Adapt model parameters θ using an unsupervised loss $\mathcal{L}(x_i; \theta)$.
2. Generate prediction $\hat{y}_i = f_{\theta'}(x_i)$ using the adapted model.
3. Reset the model parameters to the original pretrained state (for episodic adaptation).

Loss: Attention Entropy Minimization. Let $\mathbf{A}(x_i) \in \mathbb{R}^{H \times T \times T}$ denote the post-softmax attention tensor from the final transformer layer for input image x_i, where H is the number of attention heads, T is the sequence length (CLS, register, and patch tokens), and t_{cls} is the index of the CLS token. Let $\mathbb{P}$ denote the set of patch-token indices (excluding the CLS and register tokens), with $P = |\mathbb{P}|$. We extract scores for the CLS token attending to the patch tokens and renormalize them to form a distribution $a^{(h)}(x_i)$:

$$a_j^{(h)}(x_i) = \frac{\mathbf{A}_{h,t_{\text{cls}},j}(x_i)}{\sum_{k \in \mathbb{P}} \mathbf{A}_{h,t_{\text{cls}},k}(x_i)}, \quad \mathcal{L}_{Attention}(x_i) = -\frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathbb{P}} a_j^{(h)}(x_i) \log a_j^{(h)}(x_i) \tag{1}$$

We **exclude** the attention to the CLS token itself and to register tokens as we want to focus on the spatial patches of the image as opposed to global information. Minimizing this loss encourages each attention head to place concentrated (low-entropy) focus on a smaller subset of patch tokens, rather than distributing attention more diffusely. Averaging the distributions first, then taking their entropy, was tried but performed worse. This is reasonable as heads tend to specialize (Raghu et al., 2021). We update the attention of the last layer as this is closest to the model output.

Combining the standard output entropy minimization, as in (Wang et al., 2021), with attention entropy minimization further improves performance. We combine the attention entropy loss and the output entropy loss using a convex weighting parameter $\alpha \in [0, 1]$. We get best results by weighting

Corruption	Source	Tent (Online)	Tent (Episodic)	Output	Attention (*Ours*)	**Combined (*Ours*)**
Brightness	76.06	76.46	76.59	77.26	77.37	77.60
Contrast	51.96	55.77	54.88	60.49	57.52	60.36
Defocus Blur	42.87	44.63	45.30	48.17	49.53	50.59
Elastic Transform	33.94	40.83	37.60	45.74	47.29	48.70
Fog	57.20	59.94	59.20	62.69	63.07	64.56
Frost	49.72	51.82	51.23	54.66	55.42	56.42
Gaussian Noise	33.15	36.57	36.40	41.67	32.74	41.17
Glass Blur	21.68	31.26	26.59	39.40	38.36	40.47
Impulse Noise	37.05	37.13	39.58	44.44	40.68	44.97
JPEG Compression	59.27	61.72	60.93	63.14	61.89	63.21
Motion Blur	48.49	51.03	50.28	54.20	53.48	54.98
Pixelate	63.27	65.44	64.94	67.92	67.99	68.73
Shot Noise	35.33	39.65	38.78	45.93	39.26	45.92
Snow	56.94	59.58	58.87	62.52	60.19	62.55
Zoom Blur	46.11	49.72	48.76	53.41	53.21	54.47
Mean	47.54	50.77 (+3.23)	50.00 (+2.46)	54.78 (+7.24)	53.20 (+5.66)	**55.65 (+8.11)**

Table 1: Top-1 Accuracy (%) on ImageNet-C level 5 corruptions. We report the source model and test-time adaptation variants: Tent (online/episodic), output entropy, attention entropy, and their combination (*LookSharp*). All results use batch size 128. Attention, output and combined reset all the parameters. $\alpha = 0.8$ for the combined loss.

$\mathcal{L}_{Attention}$ more (Appendix B). Both loss terms are normalized to lie in the range $[0, 1]$ by dividing by the theoretical maximum to ensure that the two components are on a comparable scale.

Thus, our loss is:

$$\mathcal{L}_{Combined}(x_i) = \alpha \times \frac{\mathcal{L}_{Attention}(x_i)}{\log P} + (1 - \alpha) \times \frac{\mathcal{L}_{Output}(x_i)}{\log C} \tag{2}$$

where C is the number of classes.

3 Experiments and Results

We experiment with the standard benchmark for test-time adaptation applied to image classification, using a common architecture and a recent self-supervised backbone. We consider the batch-wise episodic test-time adaptation setting where the parameters are reset after each batch Zhang et al. (2022), and also compare to an online (no resetting) method (Wang et al., 2021).

Dataset: We evaluate on ImageNet-C (Hendrycks & Dietterich, 2019) and on ImageNet-R (Hendrycks et al., 2021) in Appendix C. We only evaluate on level 5, which is the most severe level of shift. We also perform TTA on clean data to ensure our method maintains performance without distribution shift.

Model: We use DINOv3-Base (Siméoni et al., 2025), pretrained on an internet-scale image dataset. We train a linear classification head with this representation on the source data (ImageNet training split) using the standard cross-entropy loss (a.k.a. linear probing). This yields 83.57% top-1 accuracy on the validation set. The images are preprocessed to the standard ImageNet size (224×224) as in Krizhevsky et al. (2012).

Evaluation Protocol: For each corruption type, we report per-corruption accuracy and mean corruption accuracy at level 5. We use a batch size of 128. The data is loaded in a randomized order for each shift, and as a result, each batch contains a mix of classes. We optimize by Adam (Kingma & Ba, 2015) with learning rate 5×10^{-5} for all methods except Tent. For Tent, we use 10^{-3} in the

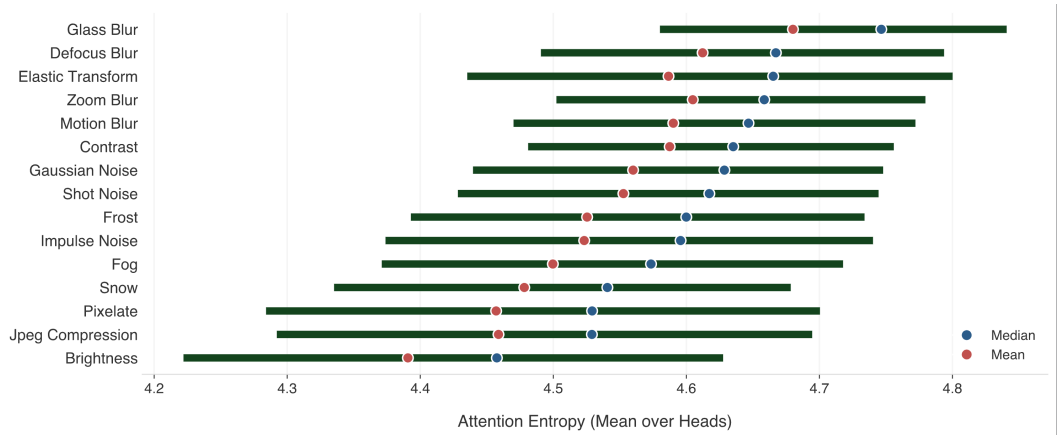


Figure 2: Median and interquartile range (IQR) of per-image mean attention entropy across a 10% sample of ImageNet-C at level 5. For each corruption, attention entropy is first averaged over heads for each image. Blurs and blur-like corruptions tend to have higher attention entropies.

episodic setting and 10^{-5} in the online setting. These values are selected by a learning-rate sweep on the level 5 test set with mean accuracy as the metric. We perform 1 gradient update per batch and update all parameters.

Baselines: We evaluate the source model as a baseline for robustness without any test-time updates. We also compare with Tent (Wang et al., 2021), where only the normalization layer parameters are updated, in the episodic and in the online settings.

Results. Table 1 shows that our method improves mean accuracy compared to the non-adapted source model on ImageNet-C. The output entropy loss alone performs better than attention entropy alone, but combining both losses yields even better results. On clean data, the attention-only loss *slightly* hurts performance ($83.57\% \rightarrow 82.95\%$). Using the combined loss *slightly* improves accuracy ($83.57\% \rightarrow 83.80\%$).

Overall, our combined losses achieves the best mean corruption accuracy, improving mean accuracy from 47.54% (Source) to 55.65% (+8.11 %). Attention-based entropy minimization works best for blur and blur-like corruptions (elastic transform). We can see from Figure 2 that this is because blurring images makes the attention maps more diffuse, and this is what $\mathcal{L}_{Attention}$ is directly adressing. A visualization of the attention loss is shown in Appendix A.

In our experiments, we found that Tent (Online) is highly sensitive to the learning rate, consistent with Zhao et al. (2023). Larger learning rates improve performance on some corruptions but cause the model to collapse on others, resulting in mean accuracy below the source model. The learning rate we select is the one that achieves maximal mean accuracy on the shifts at severity level 5.

4 Conclusion and Future Work

We introduce *LookSharp* as a simple test-time *attention* entropy minimization method that measures CLS-to-patch attention. LookSharp shows consistent gains on ImageNet-C especially for blur-like corruptions.

Limitations. The method incurs computational overhead due to the forward-backward-forward passes needed and requires self-attention in the model architecture. Attention-based adaptation likely also depends on the quality of the learned attention maps, which vary across architectures and pretraining regimes (Darcet et al., 2024).

While this work presents focused experiments to show the effectiveness of attention entropy as an unsupervised TTA loss, future work can explore trying to extract more performance by exploring multi-layer attention losses that span the model from shallow to deep.

ACKNOWLEDGEMENTS

We thank Vivian White for her helpful review, discussion, and feedback on the exposition and experiments. We also thank the Digital Research Alliance of Canada and the Vector Institute for computational resources.

REFERENCES

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Anthony Fuller, Yousef Yassin, Junfeng Wen, Daniel G. Kyrollos, Tarek Ibrahim, James R. Green, and Evan Shelhamer. Lookwhere? efficient visual recognition by learning where to look and what to see from self-supervision, 2025. URL https://arxiv.org/abs/2505.18051.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Takuya Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Robustifying vision transformer without retraining from scratch using attention-based test-time adaptation. *New Generation Computing*, 41:5–24, 2023. doi: 10.1007/s00354-022-00197-9.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien

Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fc28053a08f59fccb48b11f2e31e81c7-Paper-Conference.pdf.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

A Appendix

The figure below shows the CLS-to-patch attention distribution of the model taken from the final layer and averaged across the heads.

Figure 3: This shows the attention map before and after adaptation using our attention entropy loss ($\mathcal{L}_{Attention}$).

B Loss Interaction

We study the loss terms interacting by varying the weighting between them. Specifically, we combine the attention entropy loss and the output entropy loss using a convex weighting parameter $\alpha \in [0, 1]$.

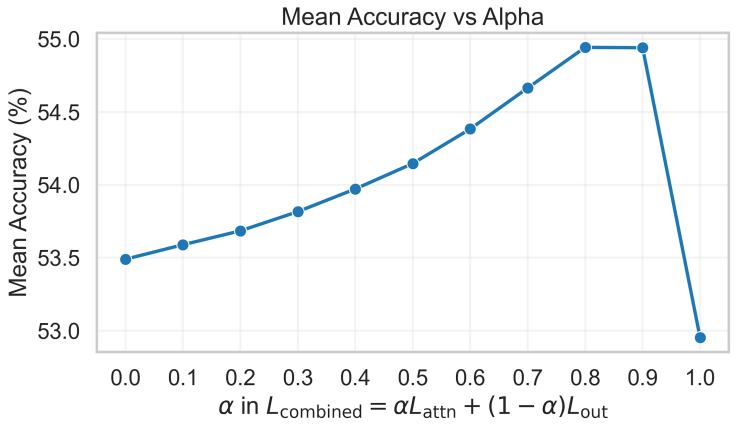


Figure 4: We get best results by weighting $\mathcal{L}_{Attention}$ more. Both loss terms are normalized to lie in the range $[0, 1]$ by dividing by the theoretical maximum to ensure that the two components are on a comparable scale. The y-axis is the mean accuracy on level 5 ImageNet-C.

C IMAGENET-R

The following results on ImageNet-R used a learning rate of 10^{-4} and batch size 128.

Table 2: Performance on ImageNet-R

Method	**Score**
Baseline	56.92
Attention	60.05
Output	60.06
Combined ($\alpha = 0.5$)	61.20