REGULARIZING ATTENTION SCORES WITH BOOTSTRAPPING

Neo Christopher Chung
University of Warsaw
Samsung AI Center, Warsaw

Maxim Laletin
University of Warsaw

Abstract

Vision transformers (ViT) rely on attention mechanism to weigh input features, and therefore attention scores have naturally been considered as explanations for its decision-making process. However, attention scores are almost always non-zero, resulting in noisy and diffused attention maps and limiting interpretability. Can we quantify uncertainty measures of attention scores and obtain regularized attention scores? To this end, we consider attention scores of ViT in a statistical framework where independent noise would lead to insignificant yet non-zero scores. Leveraging statistical learning techniques, we introduce the bootstrapping for attention scores which generates a baseline distribution of attention scores by resampling input features. Such a bootstrap distribution is then used to estimate significances and posterior probabilities of attention scores. In natural and medical images, the proposed *Attention Regularization* approach demonstrates a straightforward removal of spurious attention arising from noise, drastically improving shrinkage and sparsity. Quantitative evaluations are conducted using both simulation and real-world datasets. Our study highlights bootstrapping as a practical regularization tool when using attention scores as explanations for ViT.

Code available: `https://github.com/ncchung/AttentionRegularization`

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

The transformer is a highly influential deep learning architecture that has demonstrated remarkable performance across various domains, including natural language processing (Vaswani et al., 2017; Devlin et al., 2019), computer vision (Dosovitskiy et al., 2020; Wu et al., 2020), multi-modal tasks (Radford et al., 2021), and others. Central to the efficacy of transformers is the attention mechanism, particularly self-attention, which allows the model to weigh the relative significance of each input feature. Despite this built-in capability that could be interpreted as explanations for the model (Bahdanau et al., 2014; Mullenbach et al., 2018; Serrano and Smith, 2019; Thorne et al., 2019), attention scores generated by these mechanisms can be noisy and often lack sparsity, resulting in poor interpretability and inefficiency. This paper proposes a straightforward yet effective regularization method for attention scores, employing the bootstrapping techniques to suppress noise and enhance meaningful feature selection.

The Transformer architecture is built on self-attention, which enables each token in the input to attend to attend to all other tokens (Vaswani et al., 2017). This mechanism is expressed as a weighted sum of values, driven by the interactions between query, key, and value projections of the input data. Despite its ability to capture dependencies and relative importances regardless of their distance within the sequence, naive attention scores tend to be diffuse and overly distributed across many features, reducing the model's ability to highlight the most pertinent features for decision-making. This widespread fat-tailed distribution necessitates regularization techniques to refine the attention scores and improve the model's interpretability and performance.

Bootstrapping, a resampling technique in statistics, provides a data-driven approach to estimate the distribution of a statistic by repeatedly sampling with replacement from the data. When a statistical distri-

bution is not well characterized, the bootstrap could prove to be advantageous due to its flexibility and capacity to assess variability and uncertainty in model predictions (Efron, 1979; Efron and Tibshirani, 1994). In our proposed method, we leverage bootstrapping to generate a baseline distribution of the attention scores by resampling input features to isolate and understand spurious attention attributable to noise rather than informative features. By establishing this distribution, we can estimate the significance of attention scores through p-values and calculate local false discovery rates (lFDR), thereby enhancing the model's interpretability and reliability from an empirical Bayes perspective (Efron et al., 2001; Storey, 2001).

To validate our approach, we conduct a series of simulation studies that quantitatively assess the accuracy and effectiveness of the proposed regularization method. Our experiments demonstrate how the regularized attention mechanism consistently enhances model performance across several applications. Furthermore, qualitative evaluations reveal marked improvements in the model's ability to focus on salient features, providing clearer interpretability and richer insights into the decision-making process. These empirical findings underscore the effectiveness of bootstrapping as a tool for refining attention mechanisms within the Transformer architecture.

The remainder of this paper is organized as follows: Section 2 discusses related works in the field, detailing existing methods of attention score regulation and bootstrapping applications. In Section 3 we elaborate on the proposed, detailing the application of bootstrapping to attention scores, and describe in detail the methodology of our simulation study. Section 4 presents the evaluation results, highlighting the improvements in both accuracy and model interpretability. Finally, Section 5 offers a discussion on the implications of our findings and potential avenues for future research.

2 RELATED WORKS

In the transformer architecture (Vaswani et al., 2017), attention mechanisms provides relative importance of different input features when transformers generate outputs, emphasizing the significance of attention scores in capturing contextual relationships within data. Attention allows transformer models to weigh the relevance of various input features dynamically, leading to improved performance in language, vision, and other domains. The effectiveness of attention scores is further enhanced by their potential ability to facilitate explainability, enabling researchers to understand which parts of the input data contribute most to the model's predictions (Bahdanau et al., 2014; Mullenbach et al., 2018; Serrano and Smith, 2019; Thorne et al., 2019). However, whether the attention mechanisms provides comprehensive explanations is debated (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Jain and Wallace, 2019).

In most of modern transformer models, a multiple set of attention mechanisms are used, such that multi-head attentions may learn different types of relationships between input representations. Different types of relationships may even represent semantic across a single hidden layer. Nonetheless, some studies have found that not all attention heads are necessary or important. Thus, there are methods to estimate importance of attention heads and prune the neural networks (Michel et al., 2019; Voita et al., 2019; Molchanov et al., 2019). While we have similar motivation, our approach is orthogonal and complementary to the neural network pruning. Specifically, they want to identify and potentially remove unimportant attention heads. Our methods regularize attention scores, at a more fine-grained level (i.e., pixel-level), that makes of each attention heads. Our motivation to regularize attention scores arises from a need to improve interpretability of transformers, which is broadly called explainable artificial intelligence (XAI).

For convolutional neural networks (CNNs), saliency maps[1] back-propagate gradients from the output score to the input features (Simonyan et al., 2013). Extensions of saliency maps – modified back-propagation or aggregation of multiple saliency maps – are developed due to the noisy and unreliable nature of vanilla saliency maps in certain conditions (Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2016). Gradient-based methods are also developed and incorporated to improve attention scores of transformers in language (Voita et al., 2019; Abnar and Zuidema, 2020) and computer vision (Chefer et al., 2021; Brocki and Chung, 2019; Brocki et al., 2024). The proliferation of saliency maps and related methods has then spawned a number of evaluation frameworks (Brocki and Chung, 2023b,a). In most of these studies, regularization is implicitly prized as it related to interpretability, feature selection, visual contrast, and human-centric evaluation.

In machine learning, regularization techniques are utilized to mitigate overfitting and improve explainability, that are especially well developed for weights of deep neural networks (DNNs). Classically, network pruning was used to improve generalization (LeCun et al., 1989; Thodberg, 1991). Popular regularization

[1] Also called importance, attribution, or relevance maps. When writing about general architectures and different methods, we call them an importance estimator.

methods, such as L_1 (Tibshirani, 1996), L_2 (Hoerl and Kennard, 1970), and elasticnet (Zou and Hastie, 2005) regularization add penalties to the loss function to constrain the model's complexity. L_2 penalties can help regularize model weights towards 0 (Krogh and Hertz, 1991; Ash and Adams, 2020) or initialization (Kumar et al., 2023). Such modified loss functions has been also used for feature selection (Rahangdale and Raut, 2019; Lemhadri et al., 2021). Dropout randomly deactivates neurons during training to promote robustness and reduce co-adaptation among units (Srivastava et al., 2014). MixUp trains a DNN by using similar samples as data augmentation (Zhang et al., 2018). Total variation regularization and related methods have been proposed (Bredies et al., 2010; Ren et al., 2013; Kobler et al., 2020; Zhang and Guo, 2023). Peer-regularized networks (PeerNet) and related methods leverage dependent samples to increase the model performance and mitigate adversarial attacks (Svoboda et al., 2019; Sun et al., 2019). The theoretical foundations of regularization in deep learning have also been an area of active research (Achille and Soatto, 2018; Taheri et al., 2020). Regularization methods not only enhance the stability of the training process but also improve the model's ability to generalize (Hernández-García and König, 2018; Zhang et al., 2020).

By resampling with replacement, the bootstrap mimics the data generation process which help estimate the sampling distribution of a statistic without relying on traditional assumptions (Efron, 1979; Efron and Tibshirani, 1994). The computational advancements in recent years have made it feasible to apply the bootstrap methods to large datasets and to increase the downstream model performance (Kleiner et al., 2014). Bootstrapping has been shown to be effective in scenarios where labeled data is scarce (Chai and Jin, 2024). The bootstrapping has been applied to neural processes (NP) to improved the classification and regression models to be more robust and generalizable. Particularly, bootstrapping neural processes (BNP) estimate the residuals and obtaining the paired residual-bootstrapped samples, followed by bagging. Instead of sampling the residual of the data, NP with stochastic attention replaces deterministic attention modules with Bayesian Attention Modules (BAM) to sample attention weights from the Weibull distribution. By modeling the uncertainty directly, they show that the stochastic attention improves NP compared to the deterministic baselines.

In contrast, we aim to regularize attention scores by directly generating or resampling the input features. To the best of our knowledge, we are introducing for the first time the use of bootstrapping to regularize attention scores.

3 METHODS AND MATERIALS

3.1 Proposed Methods

We propose *Attention Regularization* by the Bootstrap (ARB) to better estimate attention scores and enhance the explainability of transformers. In brief, we bootstrap an input sample $\mathbf{X}$ to obtain bootstrapped samples $\mathbf{X}^*$. With a pre-trained transformer model f, we compute attention scores that would arise by chance $\mathbf{A}^*$. Attention scores derived from a bootstrap sample, which is constructed to not contain systematic structure, do not exhibit meaningful patterns. By estimating a null distribution of such unimportant attention scores through repeated bootstrapping, we can evaluate the observed attention scores. This comparison allows us to determine whether the observed attention scores significantly deviate from null attention scores that would be expected by random chance, thus identifying truly relevant features.

The Bootstrap

Our bootstrapping approach offers flexibility in its implementation, allowing for both (a) nonparametric and (b) parametric bootstrap sampling techniques (Step 2 in Algorithm 1). In the non-parametric bootstrap, input features are randomly resampled with replacement (Efron, 1979; Efron and Tibshirani, 1994). This relies on the robustness provided by resampling the actual dataset, thereby preserving first-order characteristics of observed samples. In contrast, the parametric bootstrap constructs input features from a distribution, effectively simulating feature sets under a specific probabilistic model assumption (Davison and Hinkley, 1997). Specifically, we demonstrate constructing bootstrap samples by drawing values from a normal distribution with mean and variance corresponding to that of each RGB channel of an image (Fig. 1).

Uncertainty Measures

Attention scores computed from the observed sample $\mathbf{A}$ and the bootstrap samples $\mathbf{A}^*$ are converted to z-statistics $\mathbf{Z}$ and $\mathbf{Z}^*$, respectively. The mean and standard deviation of the bootstrap sample $\mathbf{A}^*$ are estimated by

$$\hat{\mu} = \frac{1}{mB} \sum_{i=1}^{m} \sum_{j=1}^{B} a_{ij}^*, \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{mB} \sum_{i=1}^{m} \sum_{j=1}^{B} (a_{ij}^* - \hat{\mu})^2. \tag{2}$$

where index i goes over all the attention scores in the image sample of size m and index j goes over all boot-

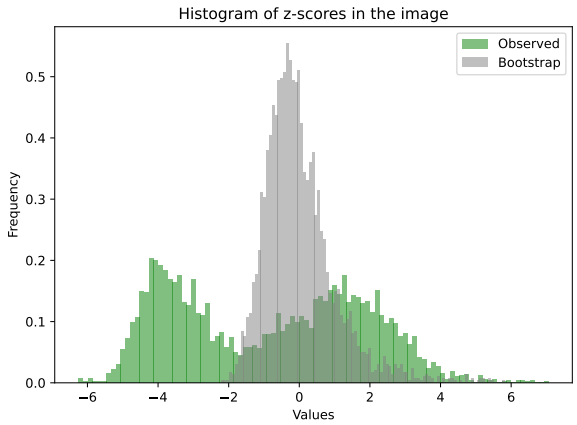


Figure 1: Histogram of z-statistics from the observed attention score sample corresponding to one of the images we use in our analysis and from the derived bootstrap sample.

strap samples with the total number B. Then, attention scores are converted to z-statistics as follows:

$$z_i = \frac{a_i - \hat{\mu}}{\hat{\sigma}}, \quad z^*_{ib} = \frac{a^*_{ib} - \hat{\mu}}{\hat{\sigma}}. \tag{3}$$

We measure the uncertainty of observed attention scores by comparing the distributions $\mathbf{Z}$ and $\mathbf{Z}^*$ (Fig. 1). First, note that the z-statistics are standardized by attention scores from the bootstrap samples (mean 0 and standard deviation 1), such that their magnitudes and signs are meaningful. For example, negative z-statistics indicate attention scores smaller than the mean that would arise by chance.

Second, given an attention score, the statistical significance (p-value) is estimated as the probability that z_i is lower than a value in the bootstrap distribution established by $\mathbf{Z}^*$:

$$p_i = \frac{\#\{(z_j + z^*_{jb}) > z_i;\ j = 1, \ldots, m,\ b = 1, \ldots, B\}}{mB}. \tag{4}$$

Third, we compute the local false discovery rate (LFDR), defined as the probability that a given attention score is attributable to the bootstrap distribution (Efron et al., 2001; Storey, 2001). LFDR can be computed directly from the observed z-statistics as

$$l_i = \Pr(z_i = 0 \mid \mathbf{Z}^*), \tag{5}$$

which assumes the most conservative hyper-parameter $\pi_0 = 1$. More accurate estimates of the probability of null scores $\pi_0 \in [0, 1]$ can be obtained following the approach of Storey and Tibshirani (2003).

Algorithm 1 Estimating Uncertainty of Attention Scores

Require: Input sample $\mathbf{X}$ and pretrained model f

1: Compute the observed attention scores $\mathbf{A} = (a_1, \ldots, a_m)^T$ using model f
2: Create a bootstrap sample $\mathbf{X}^*_b$ by
   1. a) resampling $\mathbf{X}$ with replacement, or
   2. b) sampling from a parametric distribution
3: For $\mathbf{X}^*_b$, compute the null attention scores $\mathbf{A}^*_b = (a^{*1}_b, \ldots, a^{*m}_b)^T$
4: Repeat steps 2–3 for $b = 1, \ldots, B$ to obtain $B \times m$ null attention scores $\mathbf{A}^* = (a^*_1, \ldots, a^*_B)$
5: Estimate mean μ and standard deviation σ^2 of $\mathbf{A}^*$
6: Compute standardized z-statistics for $\mathbf{A}$ and $\mathbf{A}^*$ according to Eq. (3)
7: Compute p-values for the observed attention scores from Eq. (4)
8: Compute LFDR for the observed attention scores from Eq. (5)

Regularization Techniques

Based on uncertainty statistics, we introduce several shrinkage approaches. As an initial step, we apply the simple z-statistics by setting to zero any attention scores with $z \leq 0$ and then apply one of the methods described below[2].

While we present those methods for clarity, it is possible to develop related shrinkage methods.

p-thresholding: We threshold attention scores based on p-values using a threshold p_{th}

$$\tilde{a}^p_i = \begin{cases} 0, & \text{if } p_i > p_{\text{th}} \\ a_i, & \text{otherwise} \end{cases} \tag{6}$$

l-thresholding: Similarly, l_i is used to regularize attention scores.

$$\tilde{a}^l_i = \begin{cases} 0, & \text{if } l_i > l_{\text{th}} \\ a_i, & \text{otherwise} \end{cases} \tag{7}$$

In the aforementioned p- and l-thresholding methods, it becomes apparent that choosing p_{th} and l_{th} controls regularization. To automate threshold selections, we can estimate the proportion of null scores $\pi_0 \in [0, 1]$

[2]This z-statistics procedure can be regarded as a regularization technique on its own with a threshold z_{th}. While we use $z_{\text{th}} = 0$ here, identical behaviors can be achieved by considering z-statistics in combination of p-values and LFDRs.

using the method of Storey and Tibshirani (2003) and zero out attention scores below the corresponding percentile. We call it the π_0 **thresholding**.

3.2 Quantitative Evaluation

Simulation Studies

We conduct simulation studies to create random i.i.d. pixels and evaluate the corresponding attention scores. By knowing which pixels are purely noise, we can evaluate whether regularization helps to obtain attenuated attention scores. After applying the proposed regularization methods, we quantify shrinkage, sparsity, and suppression factor. Furthermore, sensitivity and specificity are measured while varying the regularization thresholds.

The images are perturbed by an injection of a 100×100 size square consisting of generated noise pixels into a random location in the image such that the square fits in the image completely. Alternatively, we have also conducted simulation study with a different noise pattern which is randomly scattered around the image (diffuse) (Appendix H). Thus, we can easily frame the noisy part of the image, which we call the region of interest (ROI). Pixels are sampled independently for each channel from a normal distribution whose mean (μ) and standard deviation (σ) match those of the corresponding image pixel values. The DINO ViT backbone (Caron et al., 2021) with the patch size of 8×8 pixels and the backbone fine-tuned on the ImageNet subsets as described in Brocki et al. (2024). In Appendix I we show how the proposed regularization method works with a different type of ViT (DINOv2 with the patch size of 14×14).

The attention map is constructed from the attention weights for the CLS token extracted from the last transformer encoder layer of the ViT and averaged over all attention heads. The attention map consisting of patches is further refined to match the image size using the nearest-neighbor interpolation scheme and rescaled using min-max normalization. We construct our bootstrap sample from a normal distribution with mean and standard deviation corresponding to those of distribution of pixels in each RGB channel. We consider the effects of using non-parametric bootstrap method in Appendix G: both parametric and non-parametric bootstrap methods result in highly comparable operating characteristics. Generally we use one bootstrap sample ($B = 1$) and study the impact of other values of B hyperparameter (as well as the width of the distribution) in Appendix F.

The patches may have an internal *structure* that are quite different from the structure of the rest of the image in some cases, while the null attention scores are computed for a homogeneous bootstrap image. Hence in general the distribution of attention scores in ROI deviate from the null distribution. Since our goal is to simulate the noise in the images, we filter out those perturbed images that has large mean z scores in ROI and concentrate on the region $|z| \leq 1$. In Appendix C we show the distribution of mean z scores for all the perturbed images in our study, as well as the distribution of selected images in the z range of interest.

Analysis and Evaluation Metrics

To analyze the regularization effectiveness in our simulation quantitatively we rely on the percentiles q of attention scores in ROI w.r.t. the scores in the rest of the image. The *mean percentile* $\langle q \rangle$ of scores in ROI before and after regularization ($\langle \tilde{q} \rangle$) is a point estimate that reflects the strength of the regularization method for different noise levels.

To evaluate the power of different shrinkage methods cumulatively over sets of images we introduce the *average suppression factor* D, which is computed as a sum of mean percentiles in ROI after regularization divided by the sum of mean percentiles before regularization

$$D = \frac{\sum_k \langle \tilde{q} \rangle_k}{\sum_k \langle q \rangle_k}, \tag{8}$$

where index k denotes a single image and the sum goes over all the images in a set. The closer D is to 1 the worse is the efficiency of the shrinkage method on average for a given set, while $D = 0$ is the perfect case in terms of noise reduction.

To measure how regularization affects the rest of the image, which is treated as a positive signal, we employ the concepts of *sensitivity* and *specificity*. We define sensitivity as the measure of how well a shrinkage method removes the noise (ROI) for a given image and calculate it according to the following formula

$$\mathrm{Se} = 1 - \left. \frac{\sum_i \tilde{a}_i}{\sum_i a_i} \right|_{\mathrm{ROI}}, \tag{9}$$

where a denotes the attention scores before regularization and $\tilde{a}$ the attention scores after regularization. Thus, the sensitivity is equal to 1 if the noise is completely removed and is equal to 0 if it remains exactly the same. In its turn the specificity is defined as the measure of how intact is the rest of the image after regularization and is calculated as follows

$$\mathrm{Sp} = \left. \frac{\sum_i \tilde{a}_i}{\sum_i a_i} \right|_{\mathrm{Rest}}. \tag{10}$$

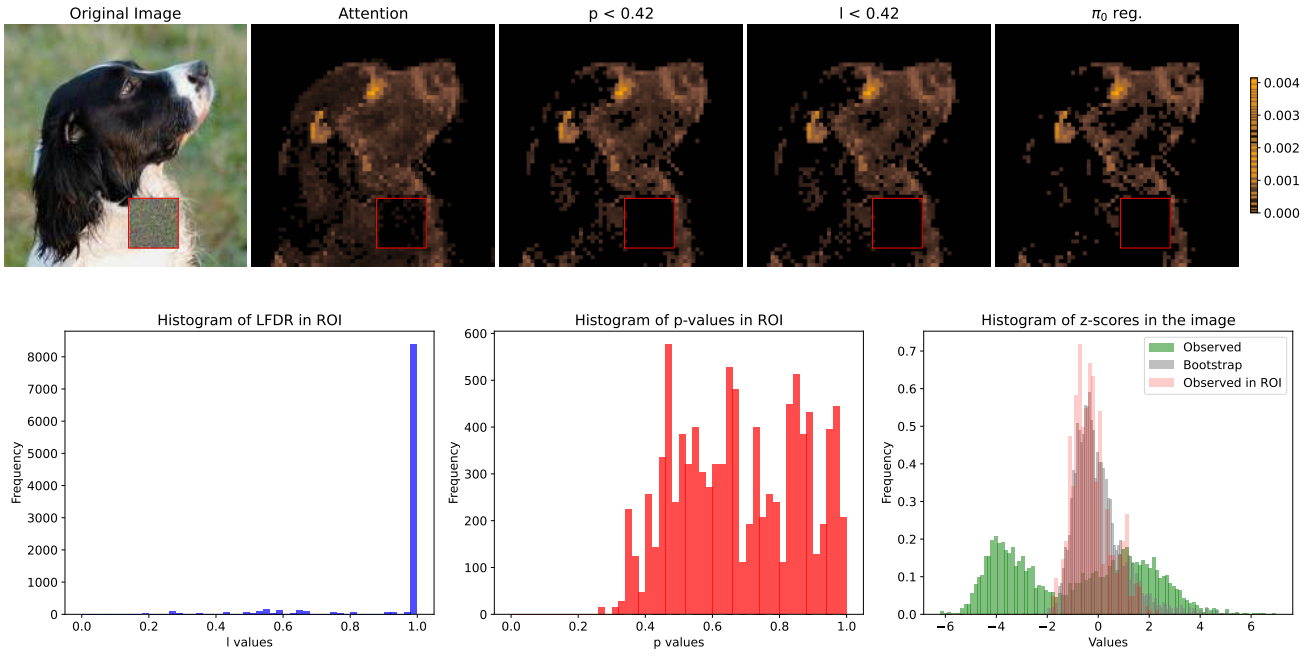


Figure 2: Example of a perturbed image (`n02102040_821.JPEG`) with the attention map before and after regularization via different shrinkage methods: p-thresholding and l-thresholding with thresholds set at the 10-th percentile for p-values and LFDR respectively.

We use these metrics to illustrate the balance between noise reduction and loss of important features controlled by the regularization parameter p_{th} or l_{th}.

3.3 Application to Medical Images

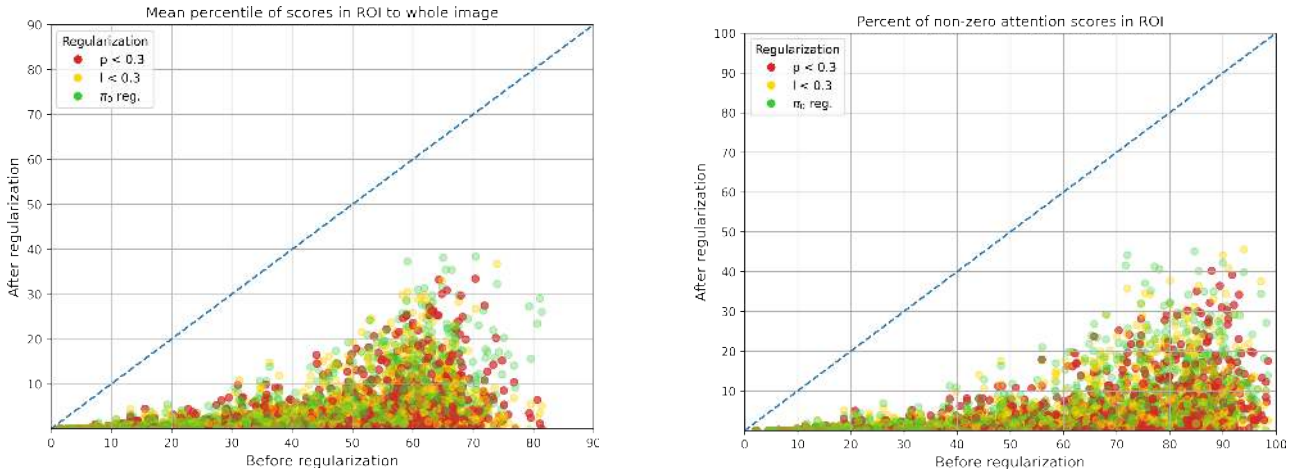
The proposed *Attention Regularization* can be used to denoise attention maps for various real-world applications. Besides the natural images in the ImageNet, we apply these regularization techniques to the images from lung cancer screening dataset IQ-OTH/NCCD (Al-Yasriy et al., 2020) containing 1097 images of normal, malignant and benign cases (a resolution of 512×512). In Appendix B we show several examples of the attention maps obtained with different shrinkage methods for different types of images and also briefly address the efficiency of regularization in Fig. S6.

4 RESULTS

In Fig. 2 we demonstrate a simulation study example from the Imagenette validation dataset and visualize the corresponding attention maps before and after regularization. We further show the distributions of the quantities that are used for regularization, namely LFDR, p-values and z-scores. The top row of the figure shows from left to right the original image with a noise patch at $(350, 250)$, the attention map extracted for that image followed by the attention map after various regularization methods: p-thresholding, l-thresholding and π_0 regularization. We use $p_{\text{th}} = 0.42$ and $l_{\text{th}} = 0.42$ that correspond to the 10th-percentile of p-values and LFDR in ROI.

The mean percentile of attention scores in ROI w.r.t. the whole attention map before regularization is 16.24, while for the displayed shrinkage methods these values are $[1.90, 1.90, 0.54]$ in the corresponding order. The bottom row shows from left to right the histograms of LFDR values in ROI, the histogram for p-values in ROI and the superimposed histograms of z-scores from the attention scores (green), from the bootstrap distribution (gray) and from the attention scores in ROI (red), which correspond to noise. Several other examples of simulation and the corresponding attention maps (including the ones that show the results of regularization for different threshold values) can be found in Appendix A.

In Fig. 3a we show the mean percentile of attention scores in ROI with respect to the attention scores in the whole image before and after different regularization methods. Fig. 3b shows the percentage of non-zero attention scores in ROI before and after these regularization methods. Each dot denotes the results of using one method for one perturbed image. Red dots show the results of p-thresholding shrinkage method for $p_{\text{th}} = 0.3$, yellow dots – l-thresholding with $l_{\text{th}} = 0.3$ and green dots – π_0-thresholding for p-values. The threshold values (except for the π_0 threshold) are



(a) (b)

Figure 3: Regularization efficiency in ROI for various images from all the categories of the Imagenette validation subset expressed in terms of: a) mean percentile of scores in ROI w.r.t. the whole image; b) percentage of non-zero attention scores in ROI. Each dot denotes the results of using a method for a perturbed image. Thresholding values $p_{\text{th}} = 0.3$ and $l_{\text{th}} = 0.3$. The blue dashed line indicates no regularization.

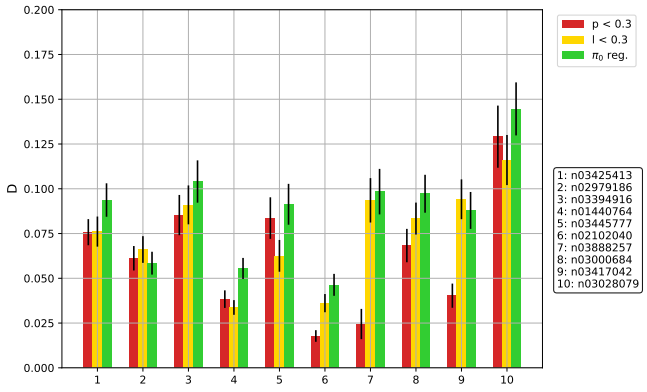


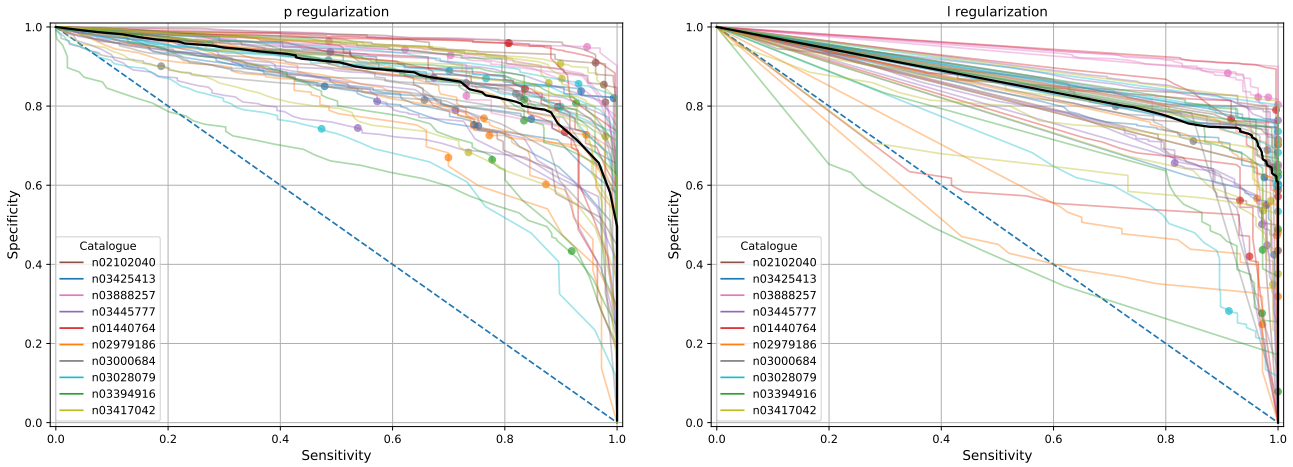
Figure 4: The suppression factor D for different categories of the Imagenette validation subset and different shrinkage methods. Lower D corresponds to better regularization.

taken as rather arbitrary non-extreme values and do not have any specific meaning. The blue dashed line corresponds to the same values before and after regularization, thus any successful regularization method is expected to produce points below that line.

Overall, our evaluation suggests that our bootstrap regularization works well for different image categories and that the results of different regularization methods correlate with each other. This can be also seen in Fig. 4, in which we show the average suppression factor D for different categories and methods. While some categories of images are regularized a bit more efficiently than the others, overall we achieve $\sim 10\%$ noise level w.r.t. the case before regularization. One can notice that different regularization methods give quite similar results as expected since the p-values and LFDR are monotonically increasing. It is worth noting that quantitatively similar results to the ones presented in Figs. 3 and 4 are achieved even if we include noise simulation cases with mean $z > 1$, as demonstrated in Appendix E.

Figure 5 depicts the specificity-sensitivity curves for different images that undergo p-regularization (left subplot) and l-regularization (right subplot) with different threshold values. For this study we extracted 5 random images from each category of Imagenette validation subset (denoted with different colors) which have the mean z-score in ROI between -1 and 1. The curves connect the results for 50 threshold values that span between 0 and 1.

The dots correspond to the sensitivity-specificity values for each presented image that are regularized with the π_0 threshold estimated for p-values and l-values accordingly. The black solid curve shows the median of all the displayed curves. The blue dashed line corresponds to the regularization process that shrinks all the attention scores in the image equally, hence any successful regularization curve should fall above that line. We see that the p-thresholding regularization is a bit more conservative than the l-thresholding and the regularization efficiency changes less steeply with the change of the threshold. This pattern follows from the fact that p-values are distributed more uniformly than the LFDR values, as can be seen in the example of Fig. 2 and similar figures in Appendix A (we also



(a) (b)

Figure 5: Sensitivity vs. specificity curves for 5 random images from each of the categories of the Imagenette validation subset (denoted with different colors) regularized via p-thresholding (*left*) and l-thresholding (*right*) for 50 values of the corresponding threshold spanning logarithmically from 0 to 1. The dots correspond to the π_0-threshold value for each curve. The black solid line shows the median of all the curves.

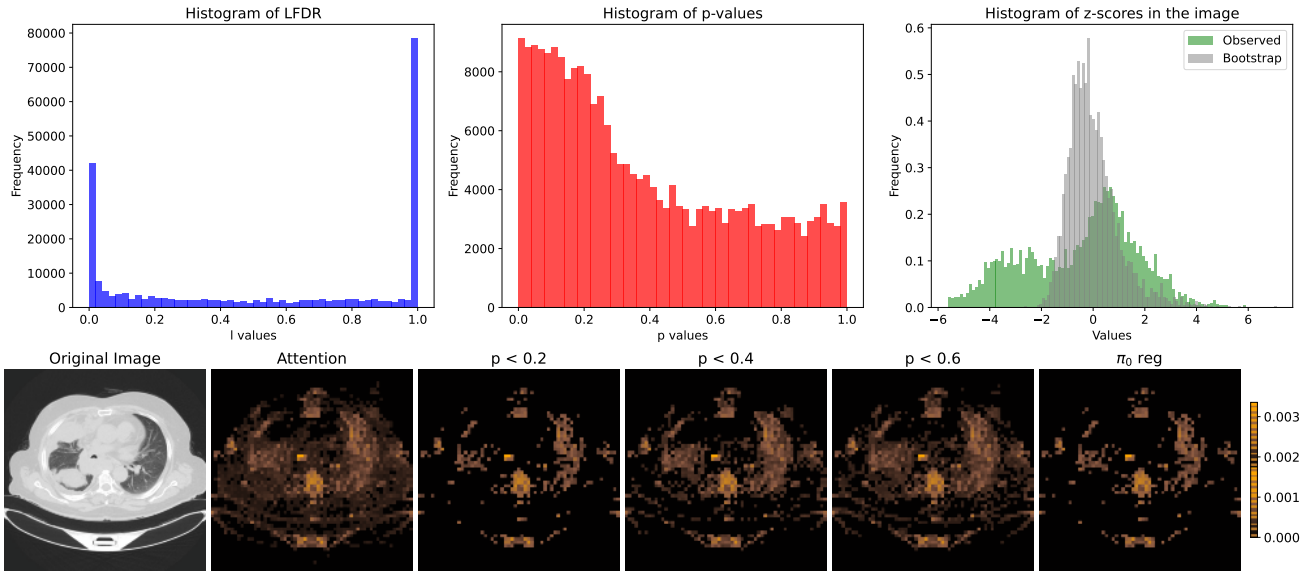


Figure 6: Results of the proposed attention regularization on the lung cancer medical images (malignant case 21). The top row shows the histograms of LFDRs, p-values, and z-scores. In the bottom row, regularized attention maps with various p-thresholds are shown alongside the CT scan and the original attention map. In the rightmost, a π_0 threshold is automatically estimated.

demonstrate the uniformity of p-values and LFDR in Fig. S8 in Appendix D).

At last, we demonstrate the proposed regularization methods on the CT scans of lungs from IQ-OTH/NCCD (Al-Yasriy et al., 2020). Fig. 6 shows the distribution of attention scores and their corresponding p-values and LFDRs from one malignant case of lung cancer. More examples and analyses are provided in Appendix B.

5 CONCLUSION

We consider the attention score in transformer architectures as a noisy realization. Given all input features are assigned non-zero attention scores, we have developed a regularization method to suppresses noise while preserving salient features. We detailed how the boot-

strap can be used in several different approaches that can be tailored to achieve the desired level of noise suppression. Due to the non-linearity and complexity of attention scores, the bootstrap is suitable to estimate the null distribution which can not be obtained otherwise.

We demonstrate the efficiency of these methods by simulating images. Our results highlight that both p-thresholding and l-thresholding regularization methods can achieve strong noise suppression across a wide range of thresholding hyper-parameters with p-regularization being slightly more conservative in terms of preserving the image features. The π_0-threshold estimation provides a remarkable balance between sensitivity and specificity without manual tuning.

This study opens several directions for future work. Although we focused on particular realizations of parametric and non-parametric bootstrapping, further research is needed to investigate other forms of bootstrap sampling. Given a long history of the bootstrap (Efron and Tibshirani, 1994), we expect to see even more accurate and efficient implementation. An interesting avenue for future studies would be to explore this regularization framework for different types of images, as different applications may require calibration to be useful in real world.

By grounding attention refinement in statistical principles, this work advances interpretability and robustness in Transformer models and opens the door to more reliable deployment of attention-driven architectures across diverse domains.

Acknowledgements

This work was funded by the SONATA BIS [2023/50/E/ST6/00694] from Narodowe Centrum Nauki, with computational resources of Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw [GDM-3540].

References

Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. ACL.

Achille, A. and Soatto, S. (2018). Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897–2905.

Al-Yasriy, H. F., Al-Husieny, M. S., Mohsen, F. Y., Khalil, E. A., and Hassan, Z. S. (2020). Diagnosis of lung cancer based on ct scans using cnn. In *IOP Conference Series: Materials Science and Engineering*, volume 928. IOP Publishing.

Ash, J. and Adams, R. P. (2020). On warm-starting neural network training. In *Advances in Neural Information Processing Systems*, volume 33, pages 3884–3894.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bredies, K., Kunisch, K., and Pock, T. (2010). Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526.

Brocki, L., Binda, J., and Chung, N. C. (2024). Class-discriminative attention maps for vision transformers. *Transactions on Machine Learning Research*.

Brocki, L. and Chung, N. C. (2019). Concept saliency maps to visualize relevant features in deep generative models. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 1771–1778. IEEE.

Brocki, L. and Chung, N. C. (2023a). Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. *Pattern Recognition Letters*, 176:131–139.

Brocki, L. and Chung, N. C. (2023b). Fidelity of interpretability methods and perturbation artifacts in neural networks. In Maughan, K., Liu, R., and Burns, T. F., editors, *Tiny Papers at International Conference on Learning Representations 2023*. OpenReview.net.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*. Pretrained models and code available at `https://github.com/facebookresearch/dino`. Accessed: 2025-09-13.

Chai, Y. and Jin, L. (2024). Deep learning in data science: Theoretical foundations, practical applications, and comparative analysis. *Applied and Computational Engineering*, 69(1):1–6.

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional

transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Hernández-García, A. and König, P. (2018). Further advantages of data augmentation on convolutional neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 95–103. Springer International Publishing.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):795–816.

Kobler, E., Effland, A., Kunisch, K., and Pock, T. (2020). Total deep variation: A stable regularizer for inverse problems.

Krogh, A. and Hertz, J. (1991). A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

Kumar, S., Marklund, H., and Van Roy, B. (2023). Maintaining plasticity in continual learning via regenerative regularization.

LeCun, Y., Denker, J., and Solla, S. (1989). Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. (2021). Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29.

Michel, P., Levy, O., and Neubig, G. (2019). Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32.

Molchanov, P., Mallya, A., Tyree, S., Frosio, I., and Kautz, J. (2019). Importance estimation for neural network pruning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2023). Dinov2: Learning robust visual features without supervision.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rahangdale, A. and Raut, S. (2019). Deep neural network regularization for feature selection in learning-to-rank. *IEEE Access*, 7:53988–54006.

Ren, Z., He, C., and Zhang, Q. (2013). Fractional order total variation regularization for image super-resolution. *Signal Processing*, 93(9):2408–2421.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A

simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Storey, J. D. (2001). *The false discovery rate: A Bayesian interpretation and the q-value*. Department of Statistics, Stanford University.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Sun, K., Yu, B., Lin, Z., and Zhu, Z. (2019). Patch-level neighborhood interpolation: A general and effective graph-based regularization strategy.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Svoboda, J., Masci, J., Monti, F., Bronstein, M., and Guibas, L. (2019). Peernets: Exploiting peer wisdom against adversarial attacks. In *International Conference on Learning Representations*.

Taheri, M., Xie, F., and Lederer, J. (2020). Statistical guarantees for regularized neural networks.

Thodberg, H. H. (1991). Improving generalization of neural networks through pruning. *International Journal of Neural Systems*, 01(04):317–326.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2019). Generating token-level explanations for natural language inference. *arXiv preprint arXiv:1904.10717*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. *arXiv e-prints*.

Yao, Y. and Ochoa, A. (2023). Limitations of principal components in quantitative genetic association models for human studies. *Elife*, 12:e79238.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhang, J. and Guo, W. (2023). A new regularization for deep learning-based segmentation of images with fine structures and low contrast. *Sensors*, 23(4):1887.

Zhang, Y., Qu, H., Metaxas, D., and Chen, C. (2020). Local regularizer improves generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6861–6868.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Examples of noise simulation and regularization

In Fig. S1 we show the examples of regularization in our noise simulation study for different regularization methods and images from different categories. The figure layout and styling are exactly the same as in Fig. 2. For p- and l-thresholds we again use the 10th quantile of the respective statistic.

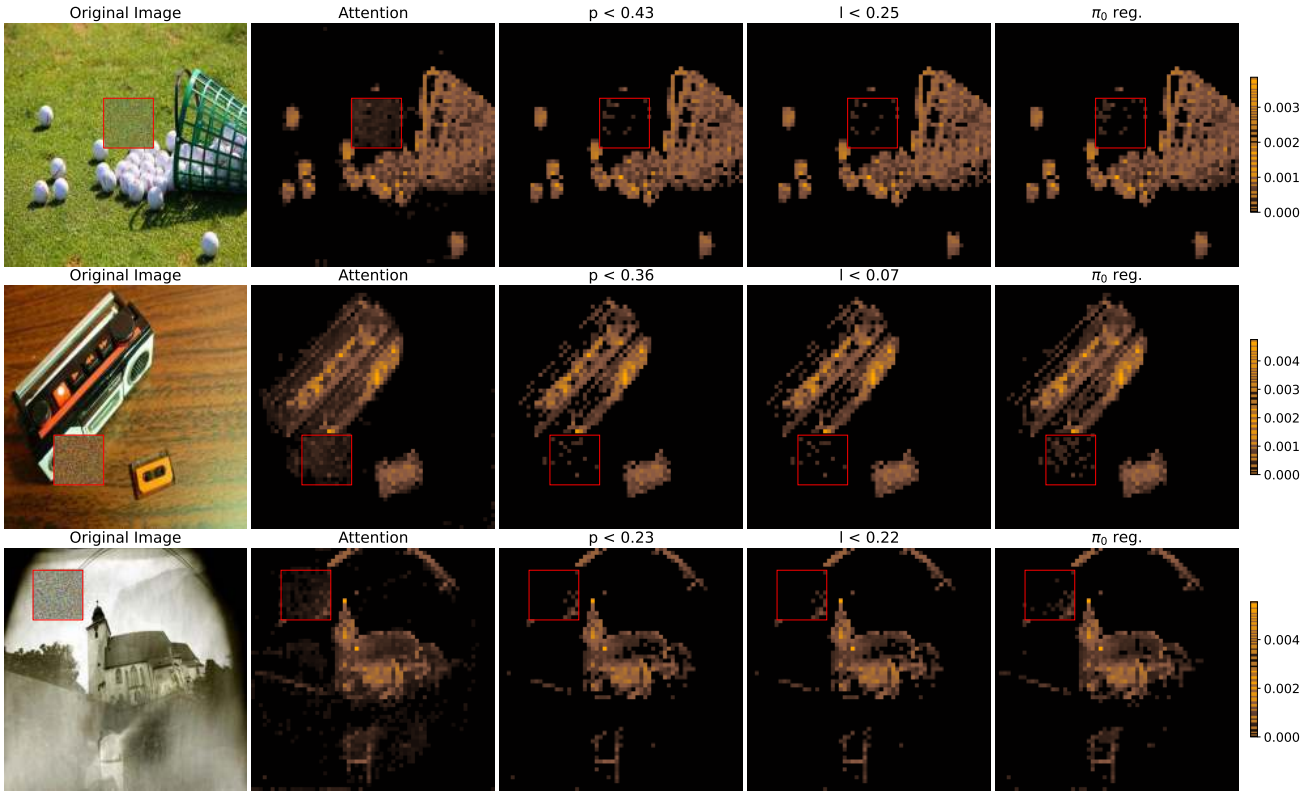


Figure S1: Examples of perturbed images with the attention maps before and after regularization via different shrinkage methods: p-thresholding and l-thresholding with thresholds set at the 10-th percentile for p-values and LFDR respectively, and π_0-thresholding with p-values. The names of the images, the coordinates of the noise patch (bottom left corner of the patch) and the mean z in ROI are as follows (from top to bottom): `n03445777_3192.JPEG`, $(200, 150)$, $\langle z \rangle = 0.98$; `n02979186_9450.JPEG`, $(100, 300)$, $\langle z \rangle = 1.03$; `n03028079_8361.JPEG`, $(58, 44)$, $\langle z \rangle = 0.68$. The noise patch in the image and attention maps is highlighted with a red frame.

In Figs. S2–S5, we show the results of regularization for various images and for different threshold values for p- and l-regularization (middle row and bottom row respectively), as well as the corresponding histograms of p-values, LFDR and z-scores (top row), similarly to Fig. 2 (bottom). The last attention map on the right corresponds to π_0-thresholding.

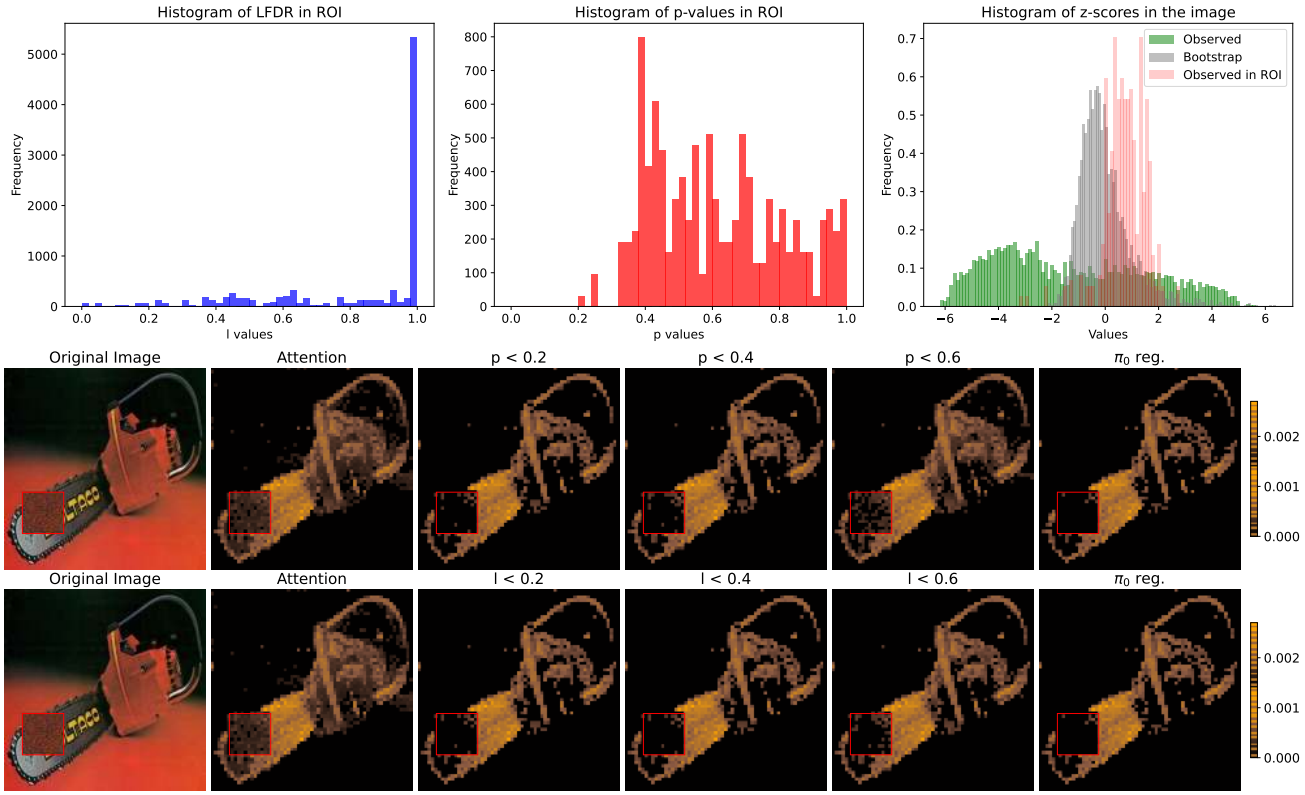


Figure S2: Example of a perturbed image (`n03000684_5970.JPEG`) with the attention map before and after regularization via p-thresholding (middle row) and l-thresholding (bottom row) with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image. The noise patch in the image and attention maps is highlighted with a red frame $(44, 300)$. The mean z-score in ROI is $\langle z \rangle = 0.54$.

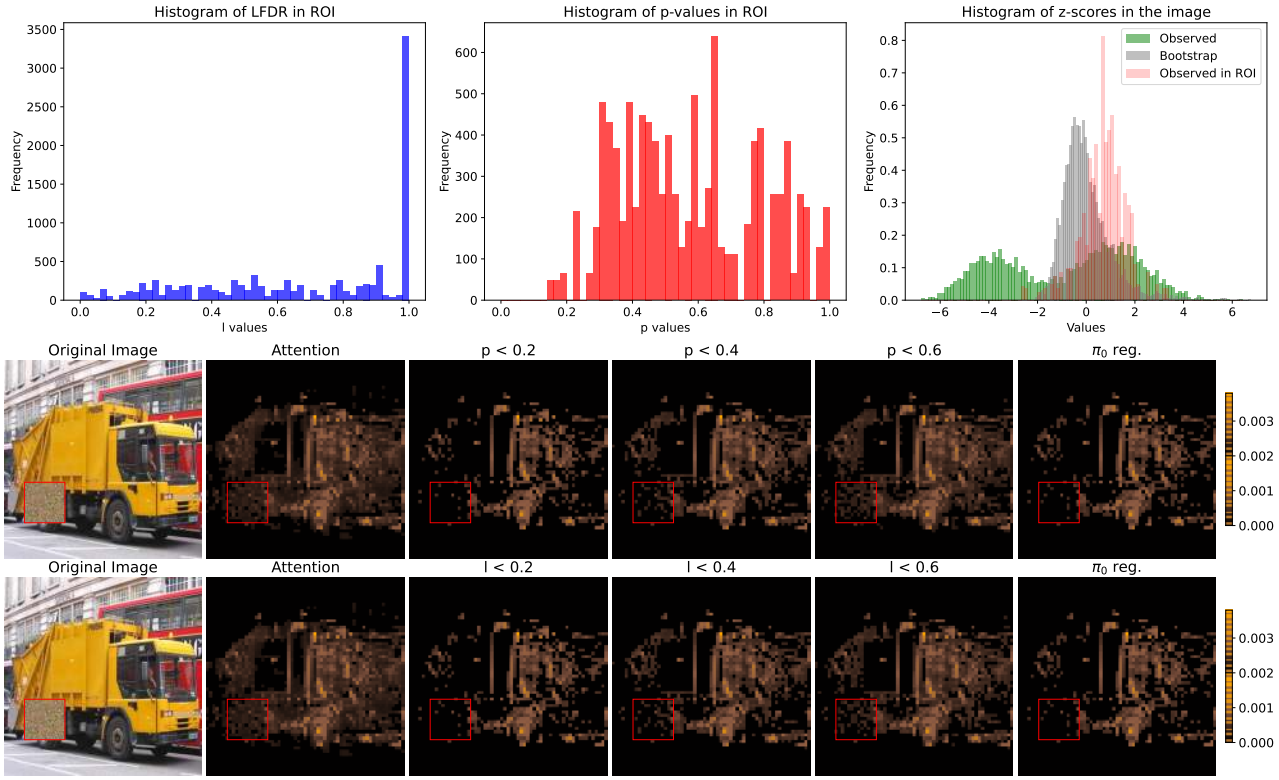


Figure S3: Example of a perturbed image (`n03417042_3300.JPEG`) with the attention map before and after regularization via p-thresholding (middle row) and l-thresholding (bottom row) with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image. The noise patch in the image and attention maps is highlighted with a red frame $(50, 300)$. The mean z-score in ROI is $\langle z \rangle = 0.88$.

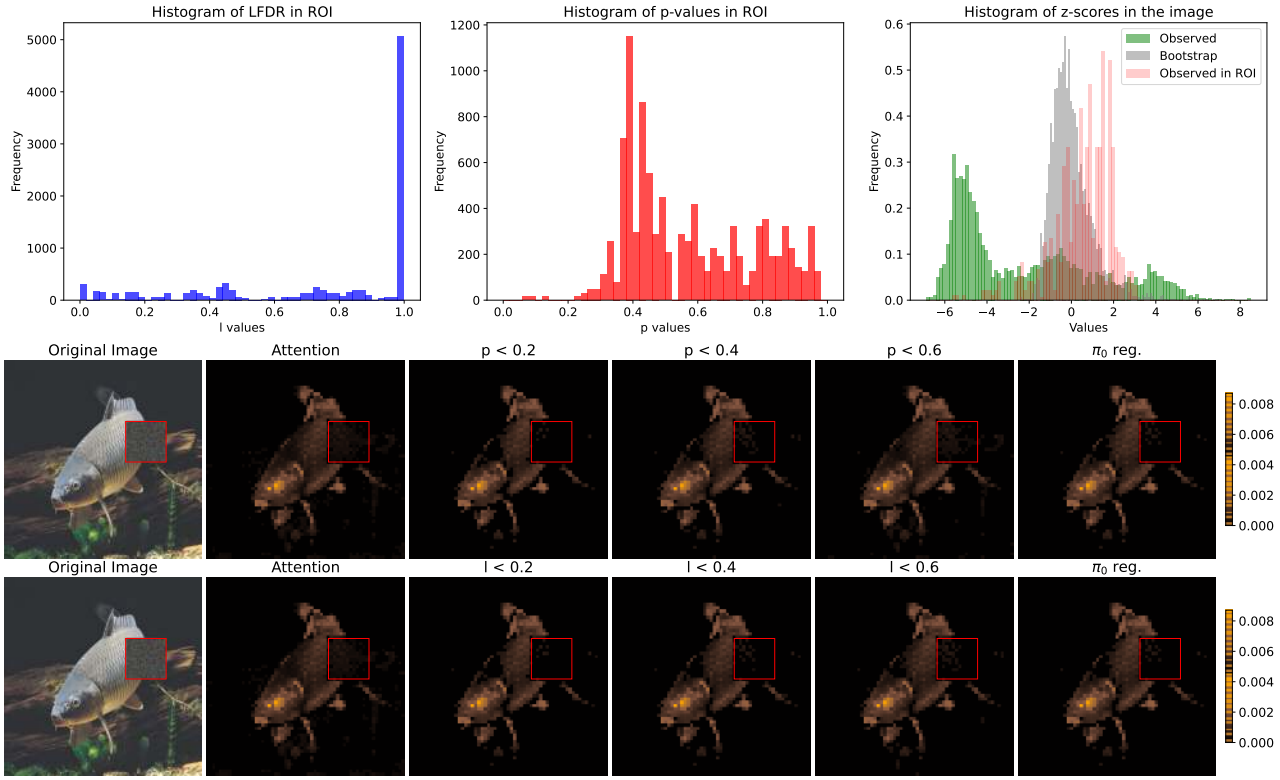


Figure S4: Example of a perturbed image (`n01440764_1310`) with the attention map before and after regularization via p-thresholding (middle row) and l-thresholding (bottom row) with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image. The noise patch in the image and attention maps is highlighted with a red frame $(300, 150)$. The mean z-score in ROI is $\langle z \rangle = 0.8$.

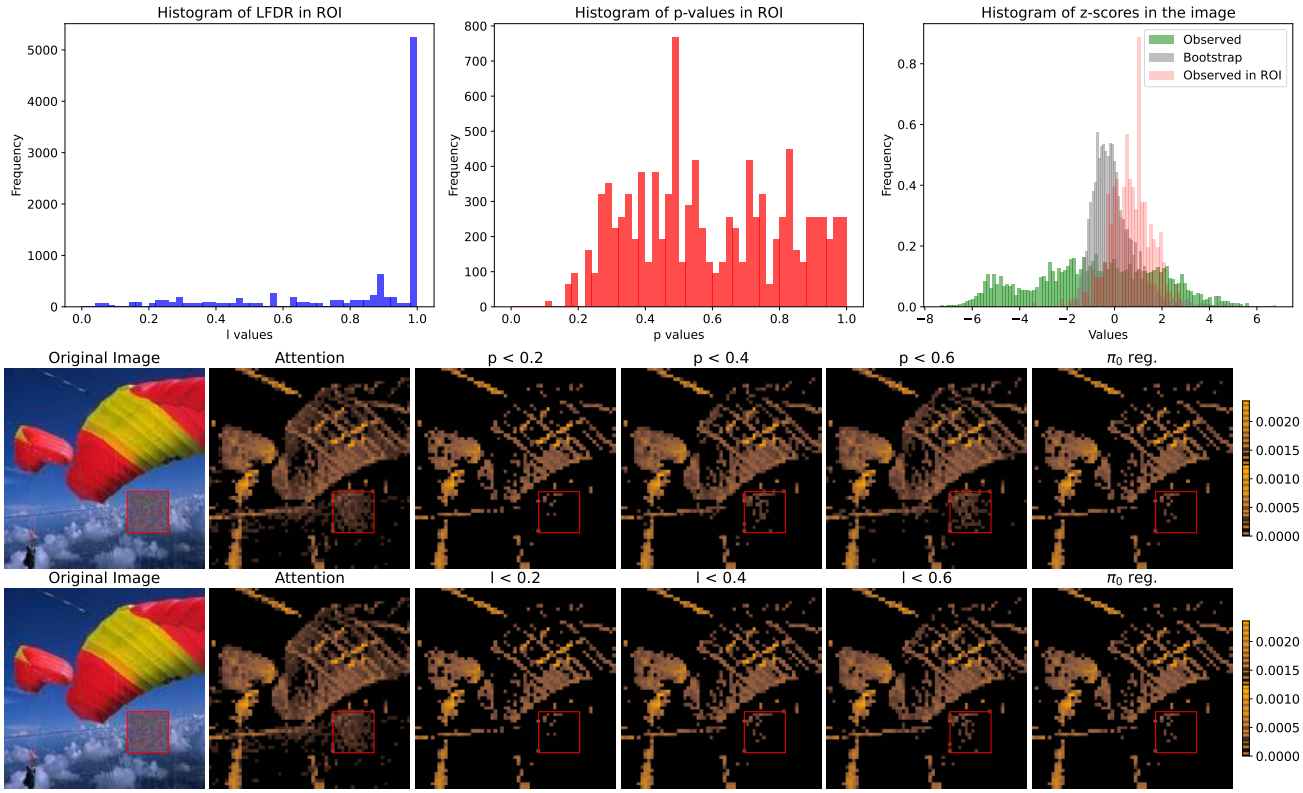


Figure S5: Example of a perturbed image (`n03888257_6331.JPEG`) with the attention map before and after regularization via p-thresholding (middle row) and l-thresholding (bottom row) with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image. The noise patch in the image and attention maps is highlighted with a red frame $(300, 300)$. The mean z-score in ROI is $\langle z \rangle = 0.71$.

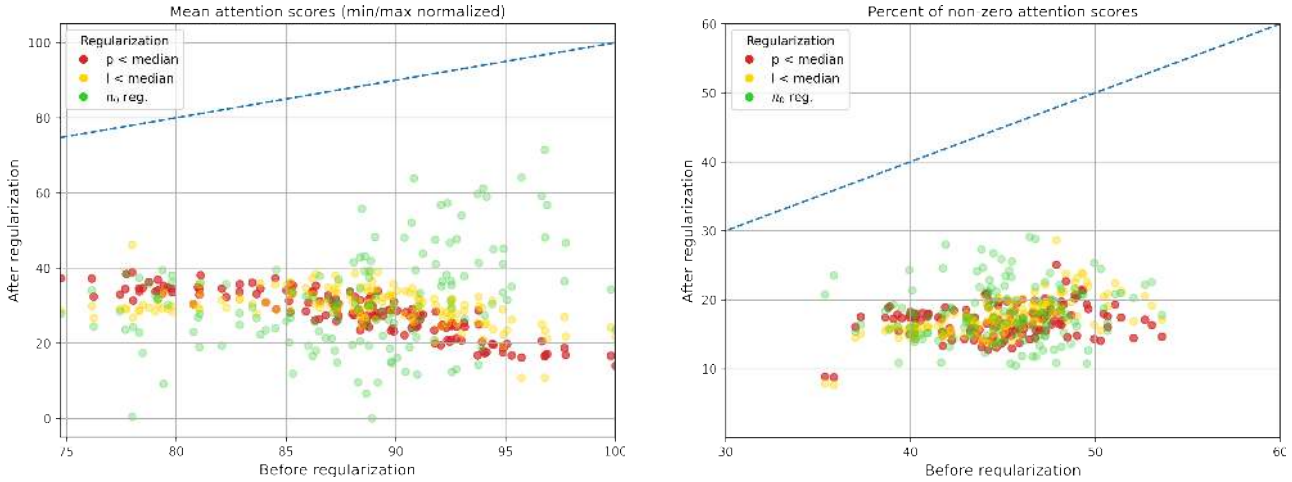
B Evaluations on the medical image dataset

We apply our regularization method to the images from the IQ-OTH/NCCD dataset as described in Section 3.3. We use the same ViT model that we use for noise simulation study.

In Fig. S6 we show the results of the regularization that we perform on 50 random images from each category (benign, malignant and normal): left plot shows the mean attention scores before and after regularization rescaled via min-max normalization for three different methods, right plot shows the percentage of non-zero attention scores before and after regularization. The blue dashed line corresponds to the same values before and after regularization. For p- and l-thresholds we use the median values of the respective statistic distributions.

In Fig. S12 we demonstrate several examples of attention maps for different images from the considered dataset before and after regularization via different methods. Each pair of images is taken from each subset, namely malignant, benign and normal in the presented order from top to bottom. For p- and l-thresholds we again use the median values of the respective statistic distributions.

In Figs. S13–S16 we present the examples of regularization using different p- and l-thresholds for medical images similar to what we display in Figs. S2–S5. In addition for each example we also show the distribution of attention scores before and after regularization for different p- and l-thresholds.

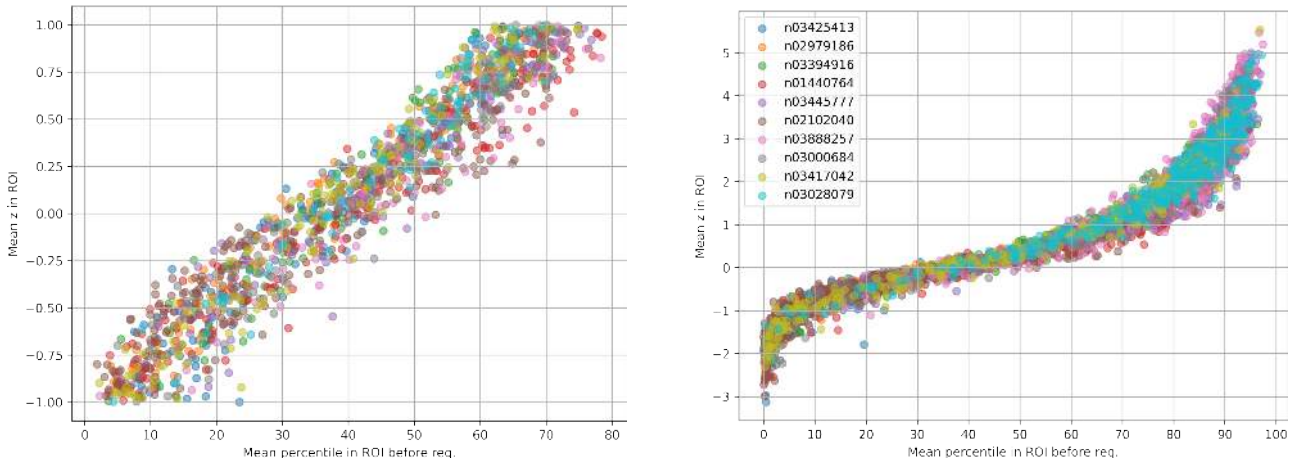


(a) (b)

Figure S6: Regularization efficiency in for various images from the IQ-OTH/NCCD dataset expressed in terms of: a) mean attention scores before and after regularization scaled via min-max normalization b) percentage of non-zero attention scores before and after regularization. Each dot denotes the results of using one method for one perturbed image: p-thresholding (red), l-thresholding (yellow), π_0-thresholding (green). For each point we use the corresponding median of the p-value and LFDR distributions as the respective regularization threshold. The blue dashed line corresponds to the same metric values before and after regularization.

C Mean z-scores in ROI

In Fig. S7 we show the mean z-scores in ROI vs. mean percentile of attention scores in ROI w.r.t. the rest of the image before regularization for our noise simulation study. The left plot shows the distribution of z-scores for the cases that we actually include in our study described in Sec. 3.2, where we select the perturbed images with $|z| \leq 1$, while the right plot shows the z-score distribution for all the perturbed images that we obtained. The results of regularization for all the images are presented in Appendix E.



(a) (b)

Figure S7: Mean z-scores in ROI vs. mean percentile of attention scores in ROI w.r.t. the rest of the image before regularization for mean $|z| \leq 1$ (left) and for all of our noise simulation cases (right). Each point corresponds to one perturbed image with color denoting the category of the image from the Imagenette validation dataset.

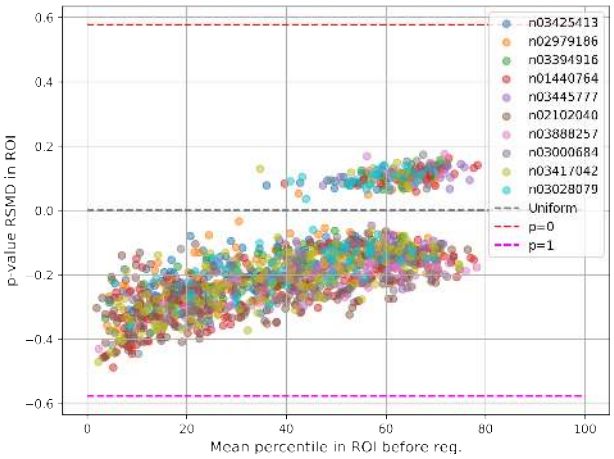
D Uniformity of p-values and LFDR

We use signed root mean square deviation (sRMSD) (Yao and Ochoa, 2023) to measure the uniformity of p-value distributions defined as follows:

$$\mathrm{sRMSD}(P) = \mathrm{Sign}\left[\tfrac{1}{2} - \mathrm{med}(P)\right] \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_{(i)} - \tfrac{i-0.5}{n}\right)^2}, \tag{11}$$

where P is the set of p-values, $\mathrm{med}(P)$ denotes the median of the set and the sum goes over all the values in the set. The sign shows whether the distribution is tilted towards 0 (positive) or 1 (negative) w.r.t. the perfectly uniform case in which sRMSD is zero. The absolute upper limit for sRMSD is $\sqrt{1/3}$.

In Fig. S8 we show the distributions of sRSMD values for p-values in ROI in our noise simulation study. Each dot corresponds to one perturbed image and the color denotes the category. The distribution of sRSMD values is plotted against the mean percentile in ROI before regularization to indicate how the shape of the distribution of these statistics changes with the level of noise. The distributions tend to be more skewed towards 0 as the mean percentile in ROI grows, which implies that the attention scores are further away from the bootstrap distribution.

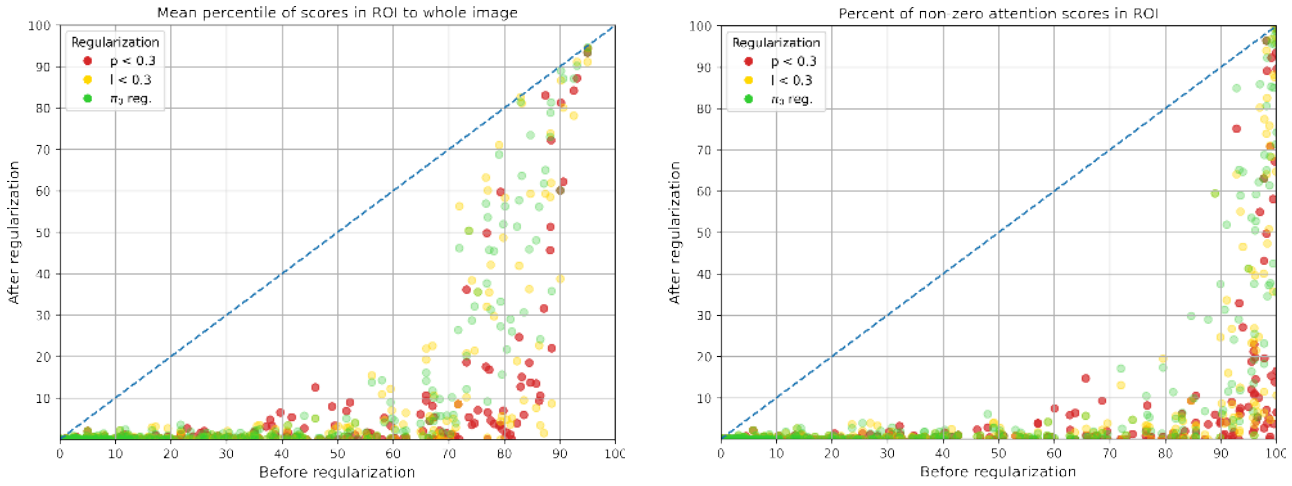


(a)

Figure S8: sRSMDs in ROI for p-values vs. the mean percentile in ROI. Each point corresponds to one perturbed image and the color denotes the category of this image. Gray dashed line corresponds to a perfect uniformity of scores, while the red and magenta dashed lines correspond to the values being concentrated around 0 and 1 respectively. See text for more details.

E Regularization efficiency for images not filtered by $|z| \leq 1$

In this appendix we present the results similar to the ones described in Section 4 (Figs. 3 and 4), but for all the perturbed images without z-filtering (Fig. S7, right). Without z-filtering several images contain noise samples that are interpreted as positive signal by ViT and receive high attention scores (which result in large mean z values in ROI). Here we present the mean percentile of scores in ROI and the percentage of non-zero attention scores in ROI before and after regularization separately for different categories: n02102040 (dogs) in Fig. S9 and n03028079 (churches) in Fig. S10. All the figure layout and styling as well as the details of the study including the values of hyperparameters are the same as for results presented in Fig. 3. We show these two particular categories of images as they manifest different distributions of average attention scores in ROI before regularization. The results for the suppression factor D for different categories, similar to Fig. 4, are shown in Fig. S11. Even though we do not separate the images with high mean z-scores in ROI, the regularization efficiency on average is still quite good.



(a) (b)

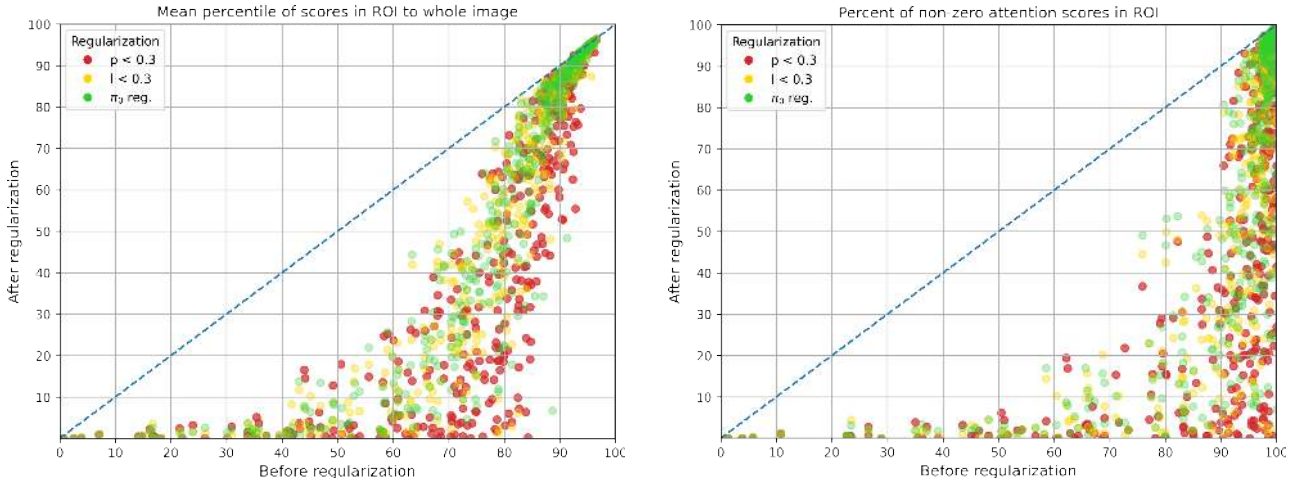
Figure S9: Regularization efficiency in ROI for various images from a particular category of the Imagenette validation subset (n02102040) without z-filtering expressed in terms of: a) mean percentile of scores in ROI w.r.t. the whole image before and after regularization; b) percentage of non-zero attention scores in ROI before and after regularization. Each dot denotes the results of using one method for one perturbed image: p-thresholding (red), l-thresholding (yellow), π_0-thresholding (green). The threshold values for this example are $p_{\text{th}} = 0.3$ and $l_{\text{th}} = 0.3$. The blue dashed line corresponds to the same metric values before and after regularization.

F The impact of bootstrap hyperparameters

We examine how the choice of the bootstrap hyperparameters affects the regularization. We perform the same study as in Sec. 3.2 with the same parametric bootstrap method (normal distribution) and vary independently the number of bootstrap samples B and the width of the bootstrap distribution, which we parameterize as $\sigma_B^j = \omega \times \sigma^j$, where the index j denotes RGB channels, σ^j is the standard deviation of the image pixels in each channel and ω is the width parameter that we actually alter. We use the average suppression factor D from Eq. (8) as the main metric to evaluate the impact of different hyperparameter values on regularization.

In Fig. S17 we present the results of this study for different values of B (left plot) and ω (right plot). We show the D factors calculated on the whole Imagenette validation dataset (1286 images after filtering for mean $|z| \leq 1$ in ROI, see Sec. 3.2) with 3 different regularization methods (similarly to Fig. 4) and for the same threshold values as in our main study. Recall that the lower is the average D factor the better is the noise reduction in our experiments. The left plot in Fig. S17 illustrates that the increase of B leads to a weaker noise reduction for all the methods, which is especially pronounced for the methods based on p-values (p-thresholding and π_0-thresholding), as the increase of the number of bootstrap z-scores below a certain attention z-score shifts its p-value closer to 0 and enlarge the population of p-values below the threshold. The LFDR values are rather dependent on the shape of the null distribution, hence the bootstrap sample enlargement leaves l-values more intact. In Fig. S18 we show the histograms of l-values, p-values (in ROI) and z scores, as well as the attention maps before and after regularization, for the same perturbed image and threshold values as in Fig. 2, but for $B = 10$. Note that sensitivity is traded off against specificity with the increase of B.

The right plot in Fig. S17 shows that the wider is the bootstrap distribution the more attention scores in ROI are attributed to noise, hence it leads to the decrease of the average D factor. In Fig. S19 we show the attention maps and statistics histogram as in Fig. S18, but for $B = 1$ and $\omega = 4$. Comparing the attention maps between the two cases (as well as the case in Fig. 2) we can again notice the trade off between sensitivity and specificity. These considerations can be used to adjust the hyperparameters of the bootstrap distribution to reach a desired level of regularization.



(a) (b)

Figure S10: Regularization efficiency in ROI for various images from a particular category of the Imagenette validation subset (n03028079) without z-filtering expressed in terms of: a) mean percentile of scores in ROI w.r.t. the whole image before and after regularization; b) percentage of non-zero attention scores in ROI before and after regularization. Each dot denotes the results of using one method for one perturbed image: p-thresholding (red), l-thresholding (yellow), π_0-thresholding (green). The threshold values for this example are $p_{\text{th}} = 0.3$ and $l_{\text{th}} = 0.3$. The blue dashed line corresponds to the same metric values before and after regularization.

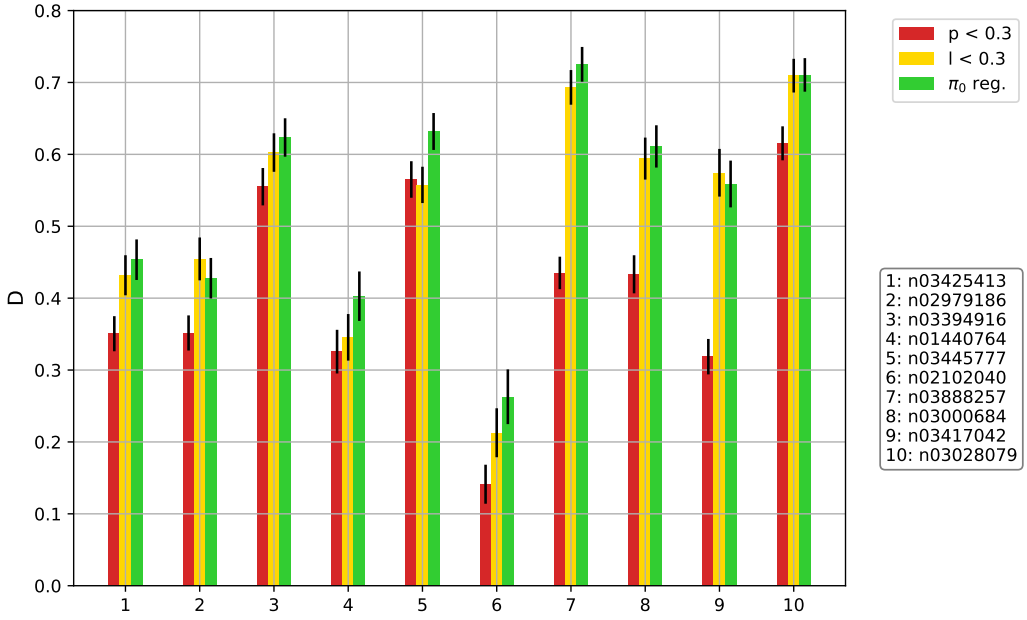


Figure S11: The average suppression factor D for different categories of the Imagenette validation subset without z-filtering and different shrinkage methods. The lower is D the better average regularization is achieved with $D = 1$ being the case when all the points in Fig. 3a located on the dashed blue curve. The relative uncertainty of the results are shown with the black lines.

G Evaluation with non-parametric bootstrap

We conduct a study similar to the one described in Sec. 3.2, but with a non-parametric bootstrapping. We prepare the null image by sampling with repetitions from the set of pixels of the original image. In Fig. S20 we show the examples of the bootstrap images sampled from the original image (left plot) via normal bootstrapping that we use in all our studies (center plot) and via non-parametric bootstrapping (right plot).

We calculate the average D factor for 1286 images from the Imagenette validation dataset (with mean $|z| \leq 1$ in ROI) for parametric and non-parametric bootstrapping. For both bootstrap methods we obtain the same values of $D = [0.057; 0.069; 0.081] \pm 0.003$ for three shrinkage methods (p-thresholding, l-thresholding and π_0-thresholding) within the margin of error.

H Diffuse noise simulation

In Sec. 3.2 we describe the noise simulation procedure, in which we inject a square of the gaussian noise into a random location in an image. Here we perform a different type of noise simulation, in which the noise patches are spread over the image in a more random and diffuse manner.

For that we generate a $\mathcal{N}(0, 1)$ field of the size of the image, smooth it using an FFT-based gaussian filter $S \propto \exp\left(-(x_f + y_f)^2 * \lambda^2\right)$, where $||x_f + y_f||$ is the distance from the center in the frequency space and λ is the clustering parameter. We further apply min-max normalization, select the top N_p pixels and create an ROI mask out of them. For the sake of comparison with our main noise simulation method we choose $N_p = 100 \times 100$. We take $\lambda = 20$, which produces multiple small, but visually noticeable clusters of noise. These clusters are filled with the same gaussian noise generated with the mean and standard deviation of the distribution of pixels in each RGB channel as we used in all of our experiments.

In Fig. S21 we show an example of an image (`n02102040 821.JPEG`, which we commonly adopt as a benchmark image in our studies) perturbed with the injection of such diffuse noise, as well as the corresponding attention maps before and after regularization and the histograms of l-,p- and z-values (similarly to Fig. 2).

In Fig. S22 we show the average suppression factors D calculated for each category of the Imagenette validation dataset (similarly to Fig. 4). We injected the diffuse noise into each image and selected those with the mean $|z| \leq 1$ in ROI (resulting in the total of 528 images). Comparing this plot with Fig. 4 one can notice that the average regularization efficiency for all methods in the case of diffuse noise is a few times worse. The reason is the presence of multiple small clusters within the objects of attention in the images - these clusters would typically contain high attention scores and assigned low p- and l-values. This is well illustrated in Fig. S21. The efficiency of regularization should depend on the size of the noise clusters, however the analysis of this dependence is out of scope of our paper.

I Evaluation with DINOv2

We conduct a study of how the proposed regularization method works with a different type of ViT. For that we adopt a DINOv2 ViT with a different patch size of 14×14 pixels (Oquab et al., 2023).

In Fig. S23 we show the attention maps before and after regularization for the same benchmark image and the same threshold values as in Fig. 2 (`n02102040_821.JPEG`), as well as the corresponding histograms of l-, p- and z-values. Compared to our base ViT model the z-scores in ROI are less sparsely distributed because of a larger patch size. In fact, DINOv2 ViT displays a strong tendency to disregard the injected square noise regions, hence the attention scores in ROI are mostly low even prior to regularization (as can be seen e.g. in Fig. S24, to be compared with Fig. S2). Nevertheless, bootstrapping regularization is efficient in removing the noisy attention scores further, as demonstrated in Fig. S25, where we show the average D factors for each category of Imagenette validation dataset. The perturbed images that we included in this study were not filtered by the mean $|z|$ in ROI.

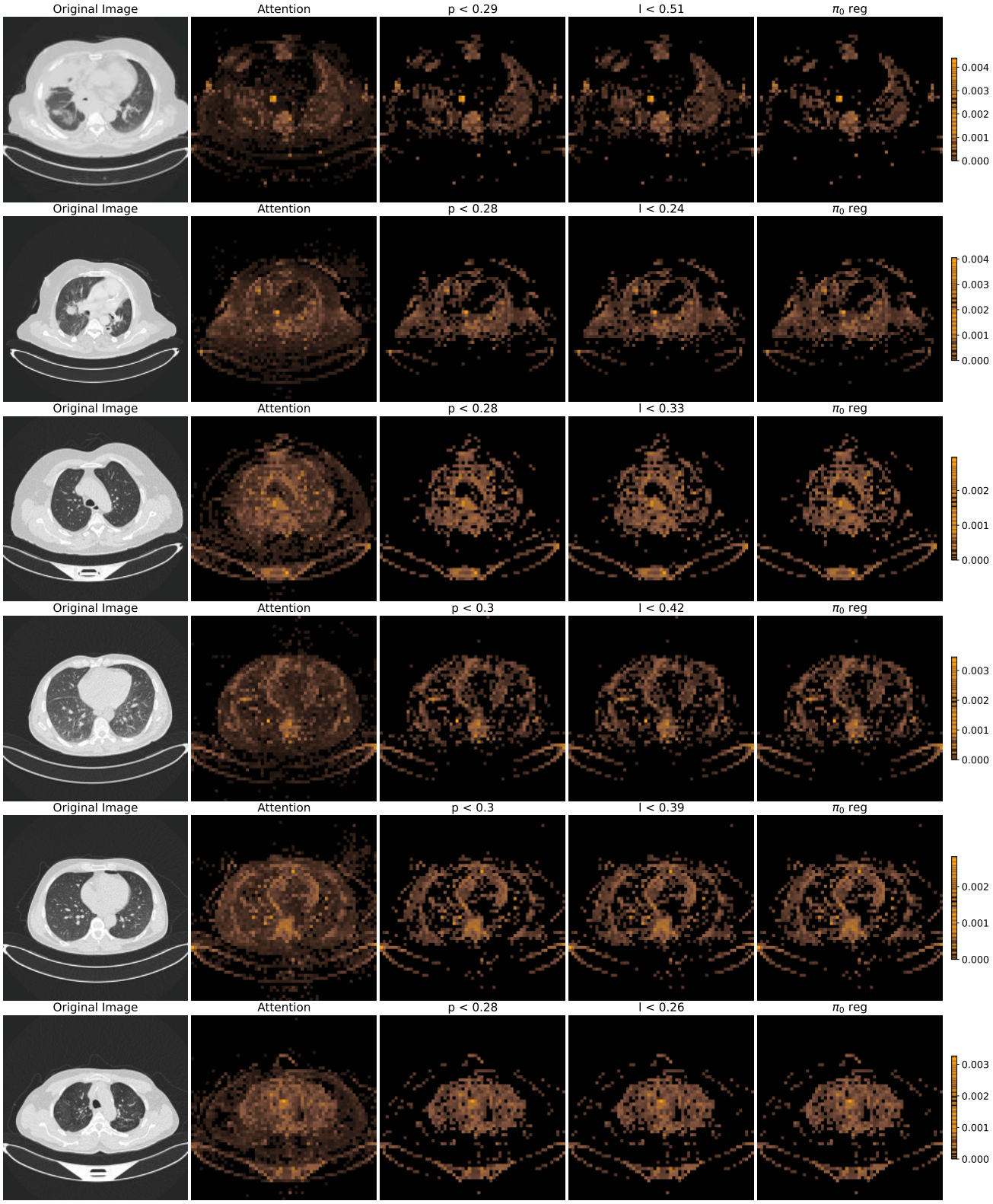


Figure S12: The examples of applying bootstrap regularization to different images from the IQ-OTH/NCCD dataset from top to bottom: malignant case (23 and 90), benign case (28 and 44) and normal case (9 and 18). The original image is displayed on the left of each row and is succeeded by the attention map before regularization and the attentions maps after: p-thresholding, l-thresholding and π_0-thresholding regularizations. For p- and l-thresholds we use the median values of the respective statistic distributions.

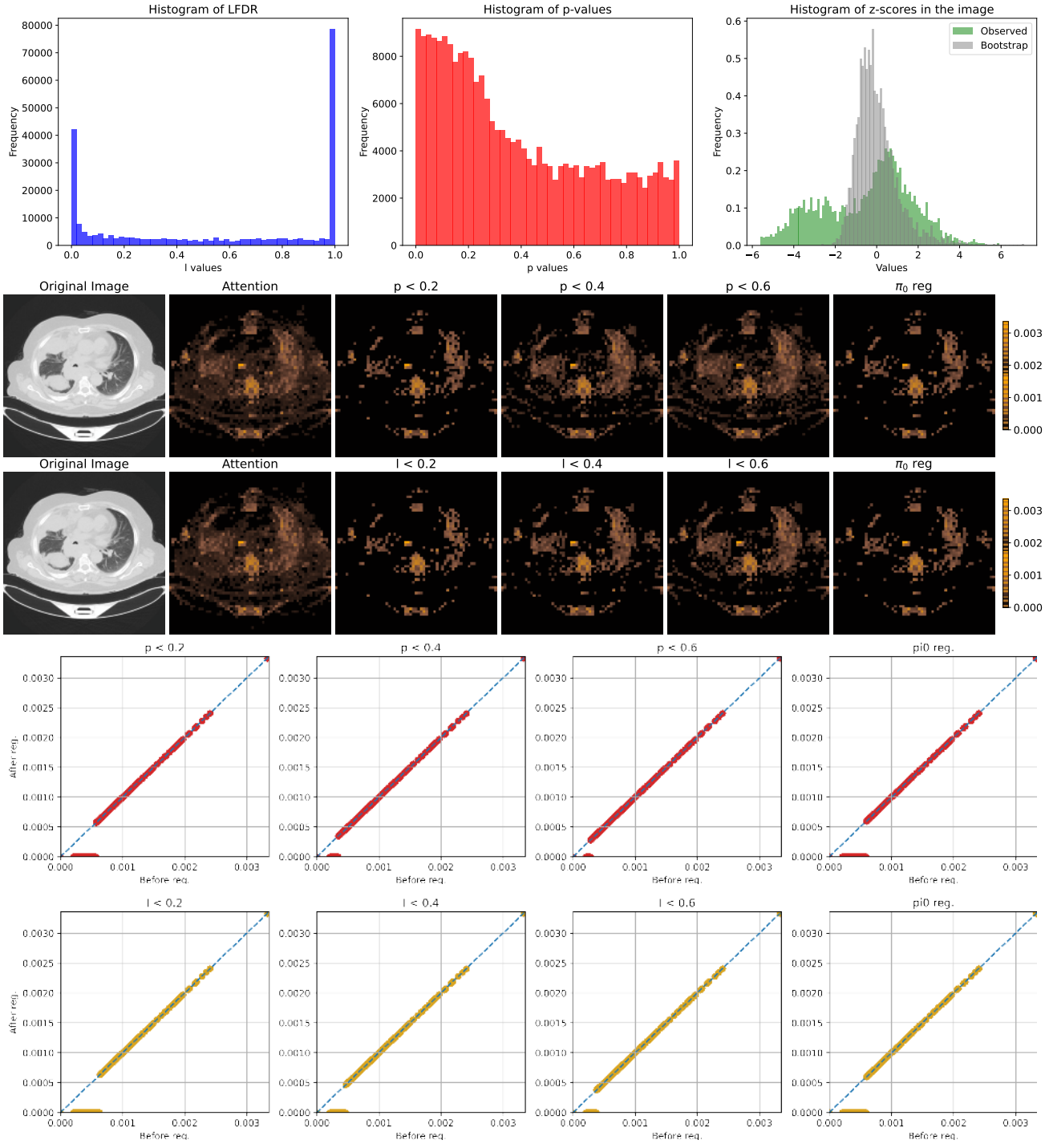


Figure S13: Results of applying p-thresholding (second row) and l-thresholding (third row) regularizations to the malignant case 21 image with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values, LFDR values and z-scores of the observed attention scores (green) and bootstrap attention scores (gray). The two bottom rows display the attention scores after regularization vs. before regularization for the regularized attention maps considered above, for p-thresholding (red) and l-thresholding (gold). The blue dashed line corresponds to the same attention score values before and after regularization.

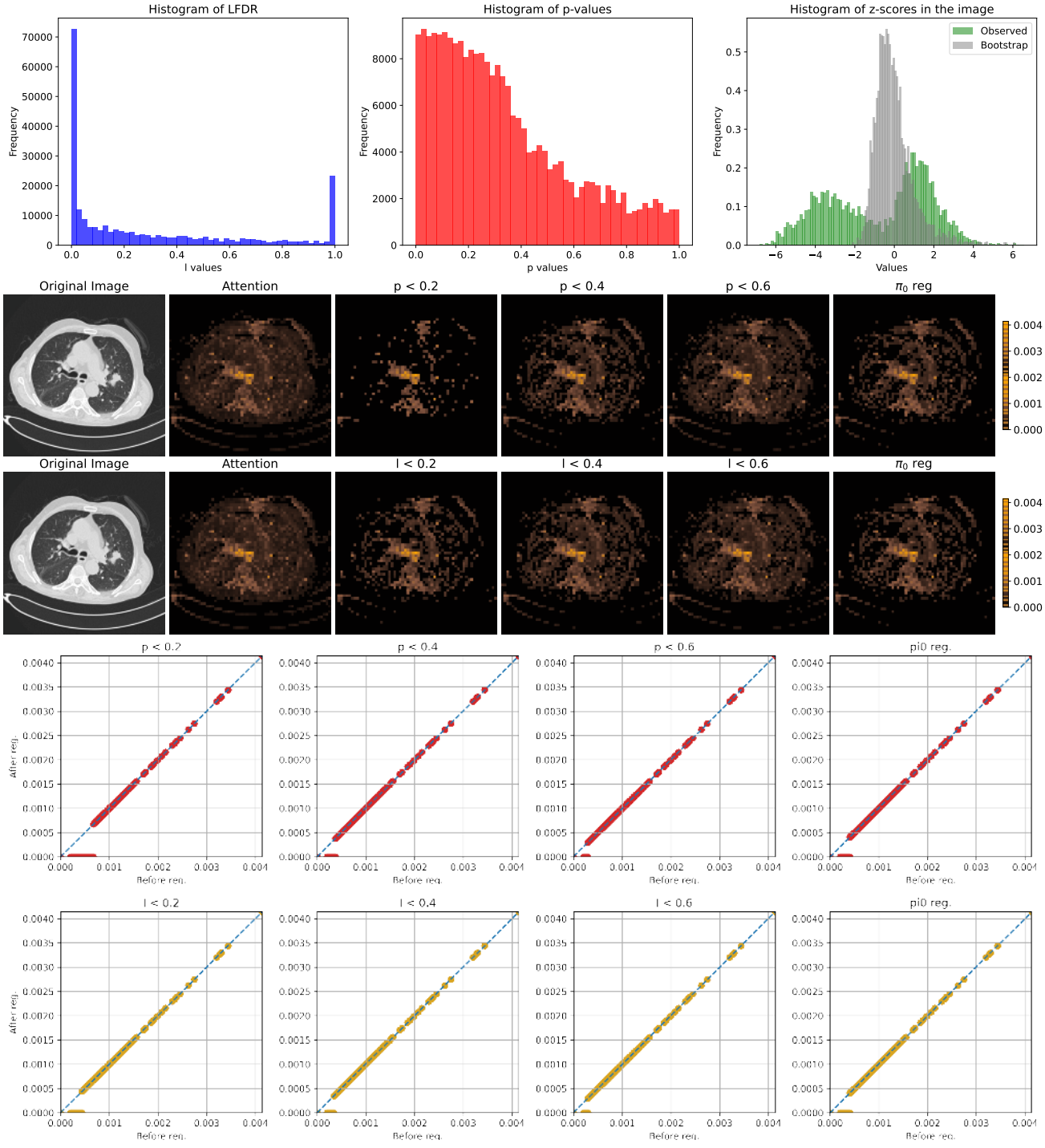


Figure S14: Results of applying p-thresholding (second row) and l-thresholding (third row) regularizations to the malignant case 73 image with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values, LFDR values and z-scores of the observed attention scores (green) and bootstrap attention scores (gray). The two bottom rows display the attention scores after regularization vs. before regularization for the regularized attention maps considered above, for p-thresholding (red) and l-thresholding (gold). The blue dashed line corresponds to the same attention score values before and after regularization.

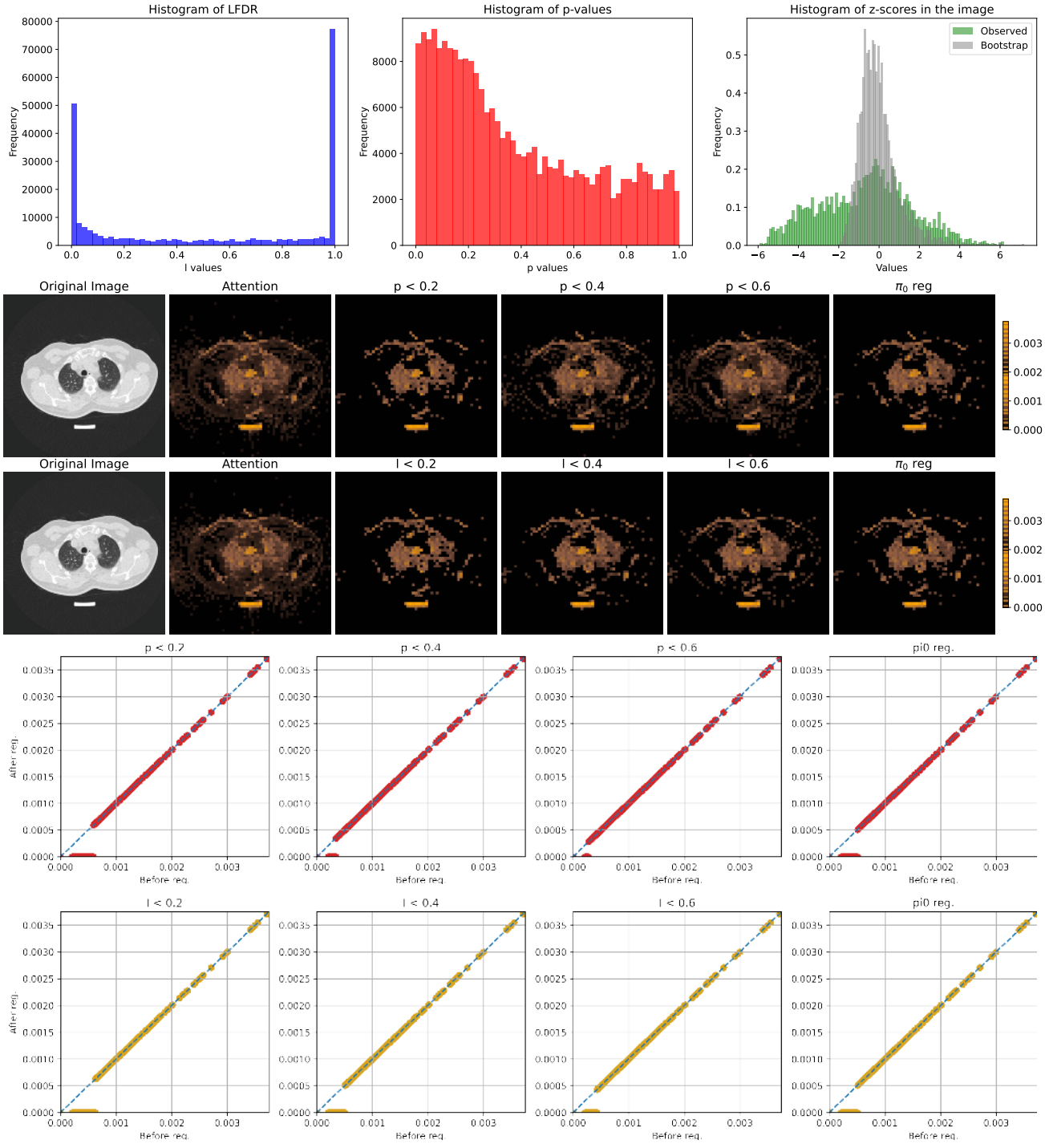


Figure S15: Results of applying p-thresholding (second row) and l-thresholding (third row) regularizations to the benign case 5 image with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values, LFDR values and z-scores of the observed attention scores (green) and bootstrap attention scores (gray). The two bottom rows display the attention scores after regularization vs. before regularization for the regularized attention maps considered above, for p-thresholding (red) and l-thresholding (gold). The blue dashed line corresponds to the same attention score values before and after regularization.

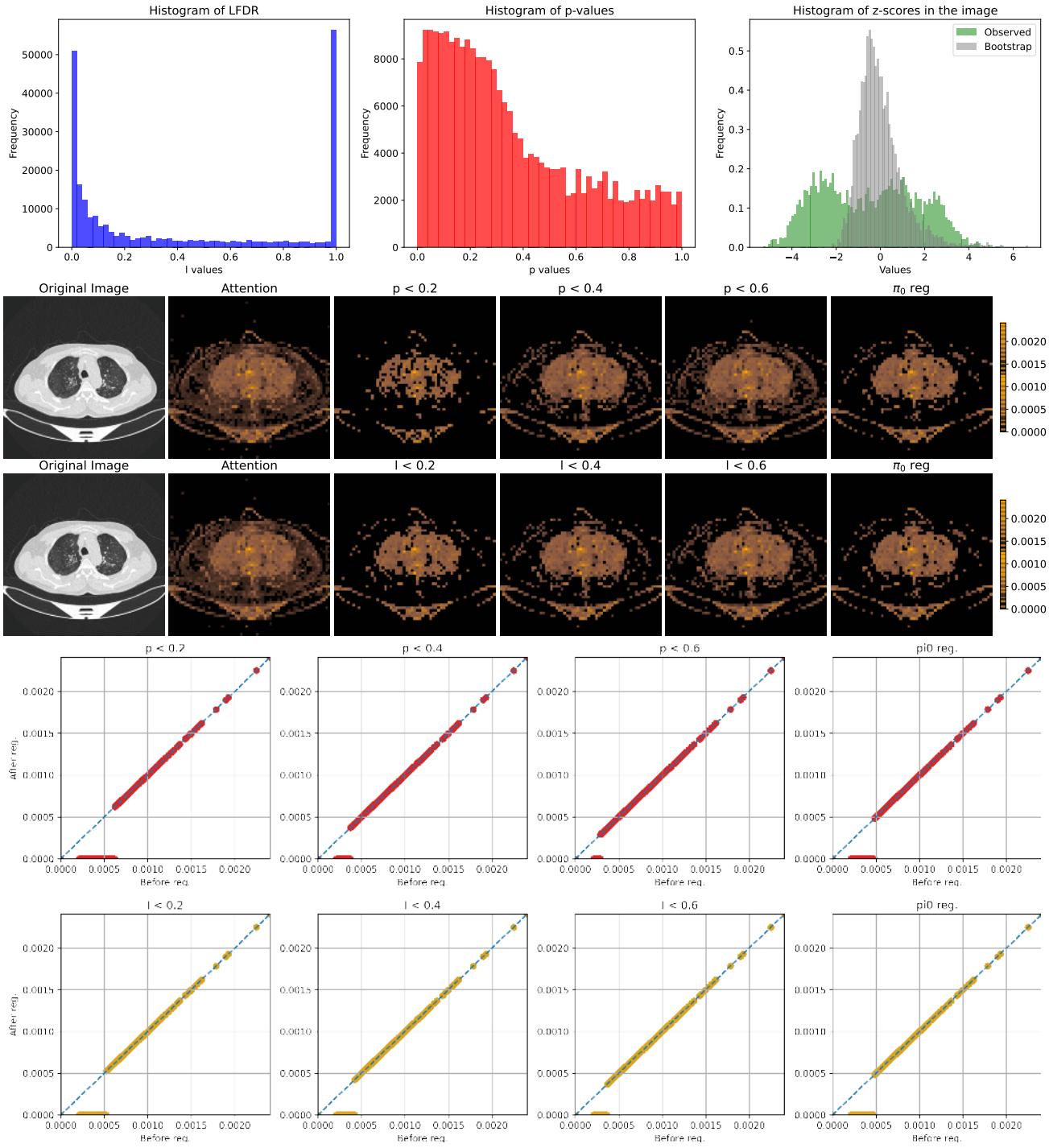
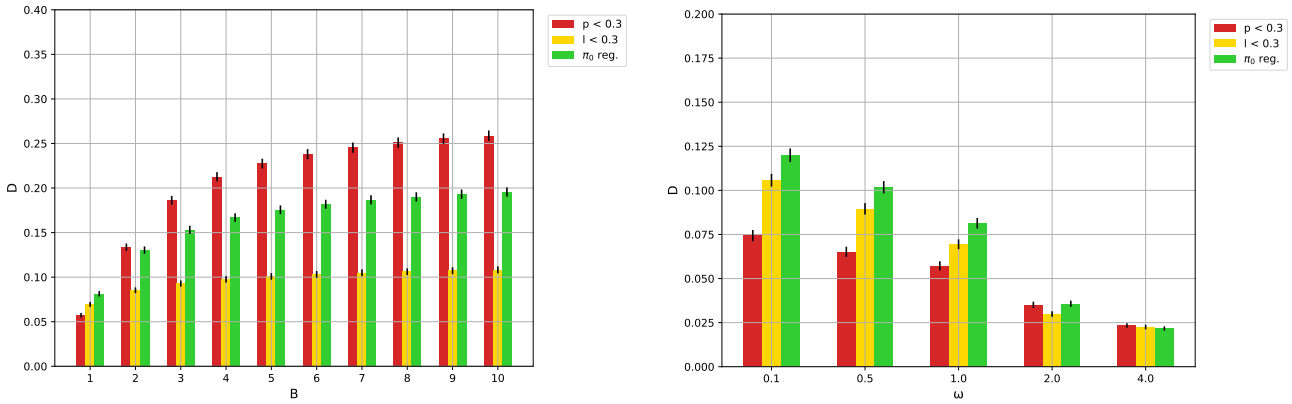


Figure S16: Results of applying p-thresholding (second row) and l-thresholding (third row) regularizations to the normal case 17 image with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values, LFDR values and z-scores of the observed attention scores (green) and bootstrap attention scores (gray). The two bottom rows display the attention scores after regularization vs. before regularization for the regularized attention maps considered above, for p-thresholding (red) and l-thresholding (gold). The blue dashed line corresponds to the same attention score values before and after regularization.



(a) (b)

Figure S17: The average suppression factor D for different values of the B hyperparameter (*left*) and for different values of the ω hyperparameter (*right*) for all the images from the Imagenette validation subset (with the mean $|z|$ in ROI ≤ 1) and different shrinkage methods. The relative uncertainty of the results are shown with the black lines. See text for more details.

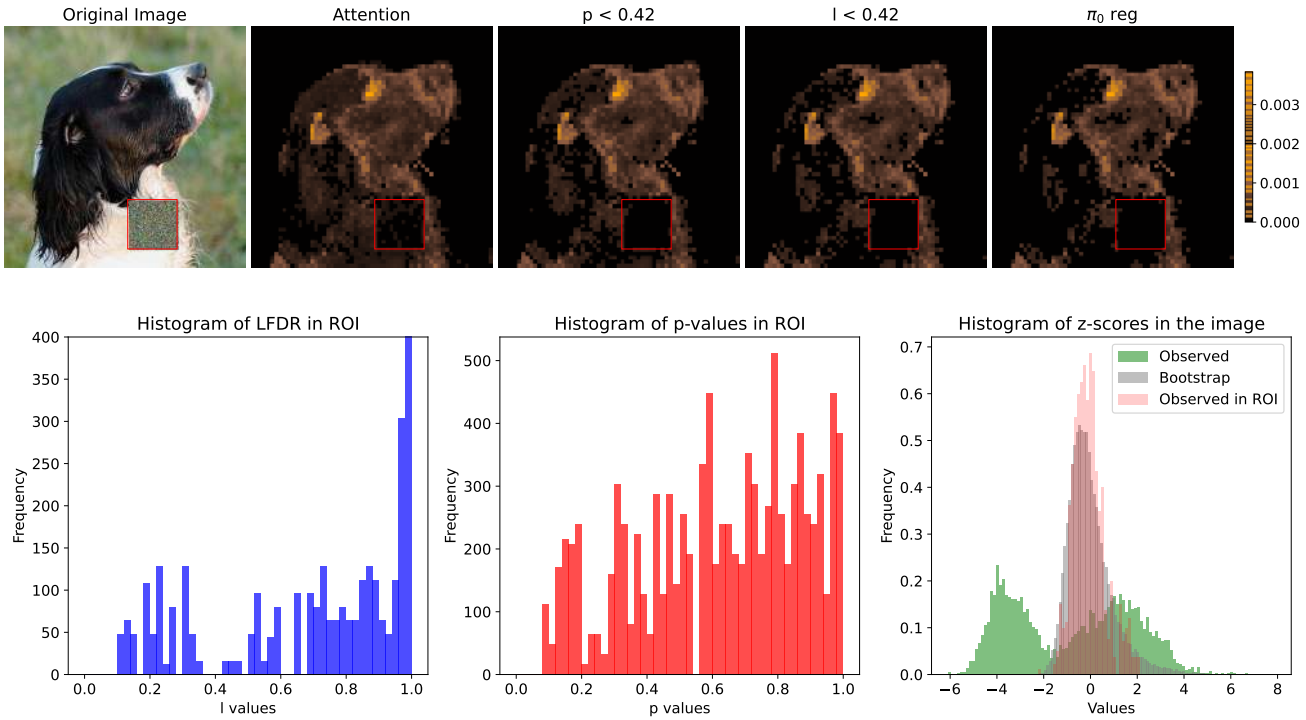


Figure S18: Attention map before and after regularization via different shrinkage methods (p-thresholding, l-thresholding and π_0-thresholding) for the same image as in Fig. 2 (`n02102040_821.JPEG`), but for $B = 10$ (top row); Histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image (bottom row). The noise patch in the image and attention maps is highlighted with a red frame.

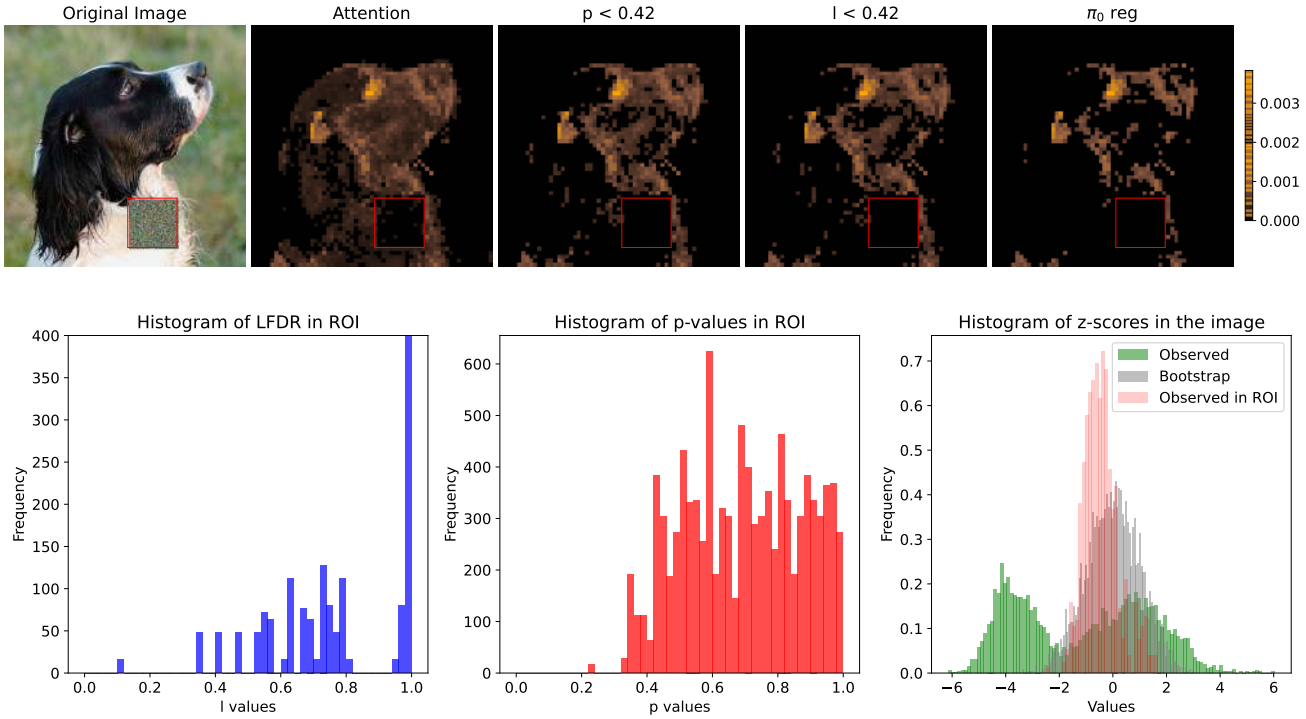
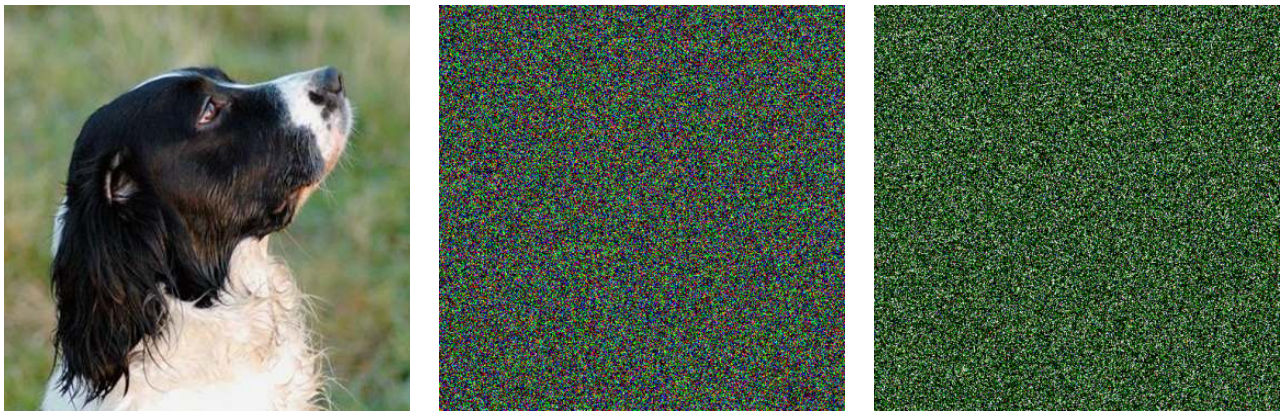


Figure S19: Attention map before and after regularization via different shrinkage methods (p-thresholding, l-thresholding and π_0-thresholding) for the same image as in Fig. 2 (`n02102040_821.JPEG`), but for std factor = 4 (top row); Histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image (bottom row). The noise patch in the image and attention maps is highlighted with a red frame.



(a) (b) (c)

Figure S20: Examples of bootstrap images generated from the original image (`n02102040_821.JPEG`, subfigure a) via parametric (gaussian) bootstrap (b) and non-parametric bootstrap (c).

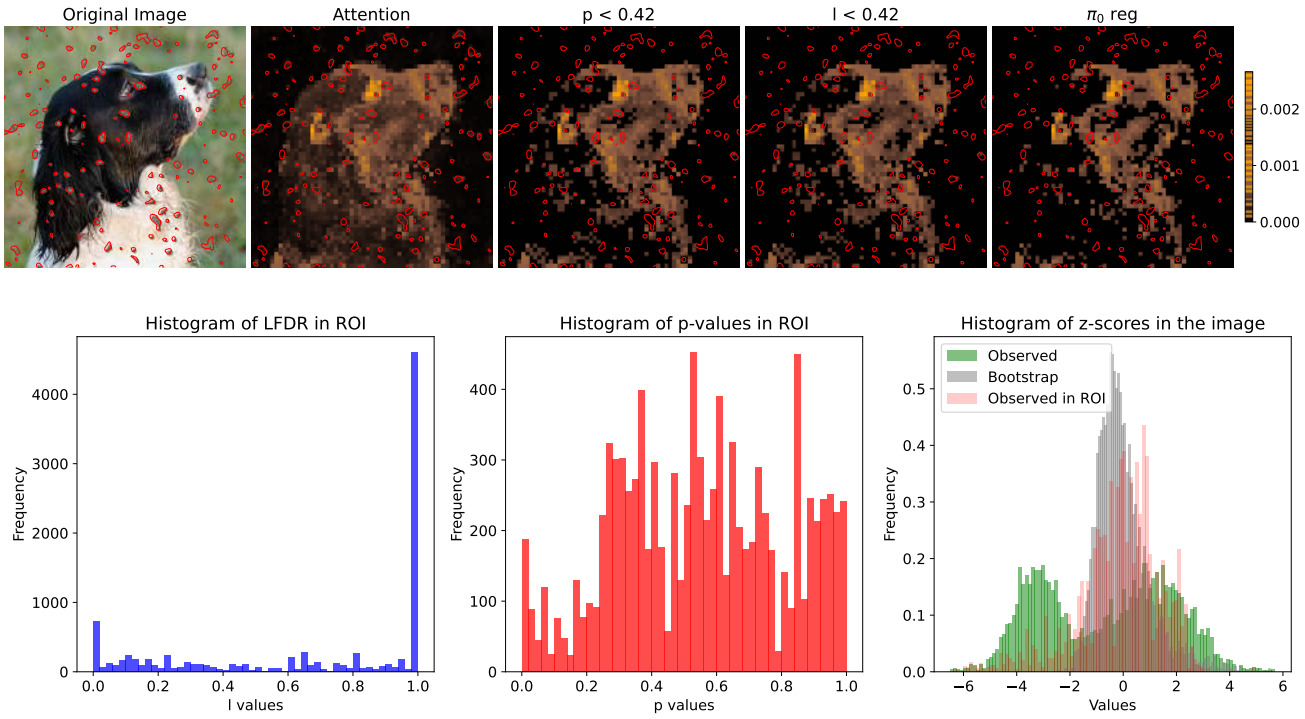


Figure S21: Attention map before and after regularization via different shrinkage methods (p-thresholding, l-thresholding and π_0-thresholding) for an image (`n02102040_821.JPEG`) perturbed with diffuse gaussian noise, (top row); Histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image (bottom row). The noise patches in the image and attention maps are highlighted with a red frame. See text for more details.

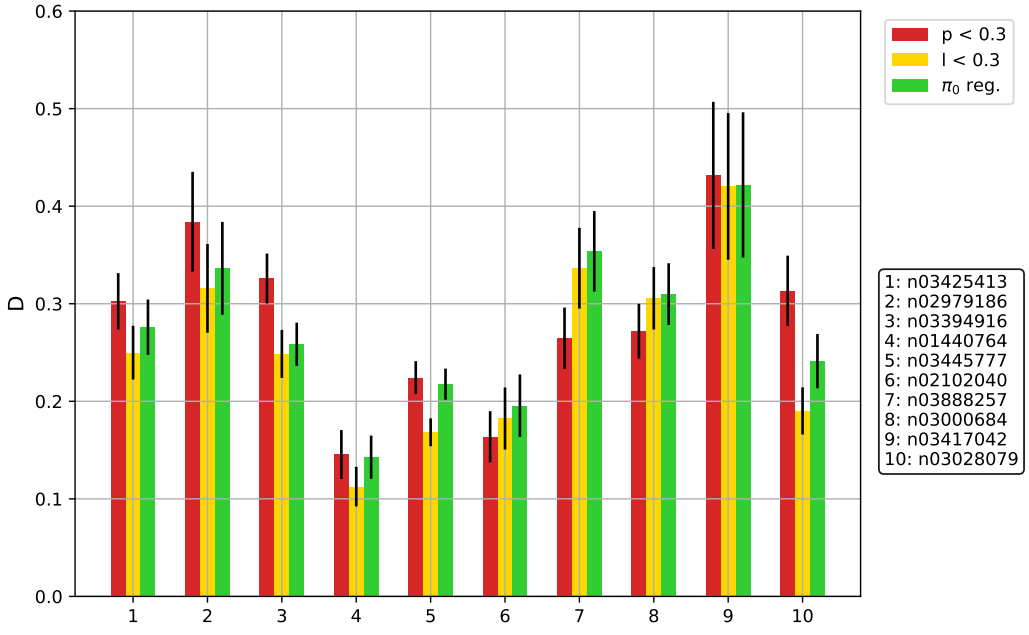


Figure S22: The average suppression factor D for different categories of images from the Imagenette validation subset perturbed with diffuse gaussian noise (with the mean $|z|$ in ROI ≤ 1) and different shrinkage methods. The relative uncertainty of the results are shown with the black lines. See text for more details.

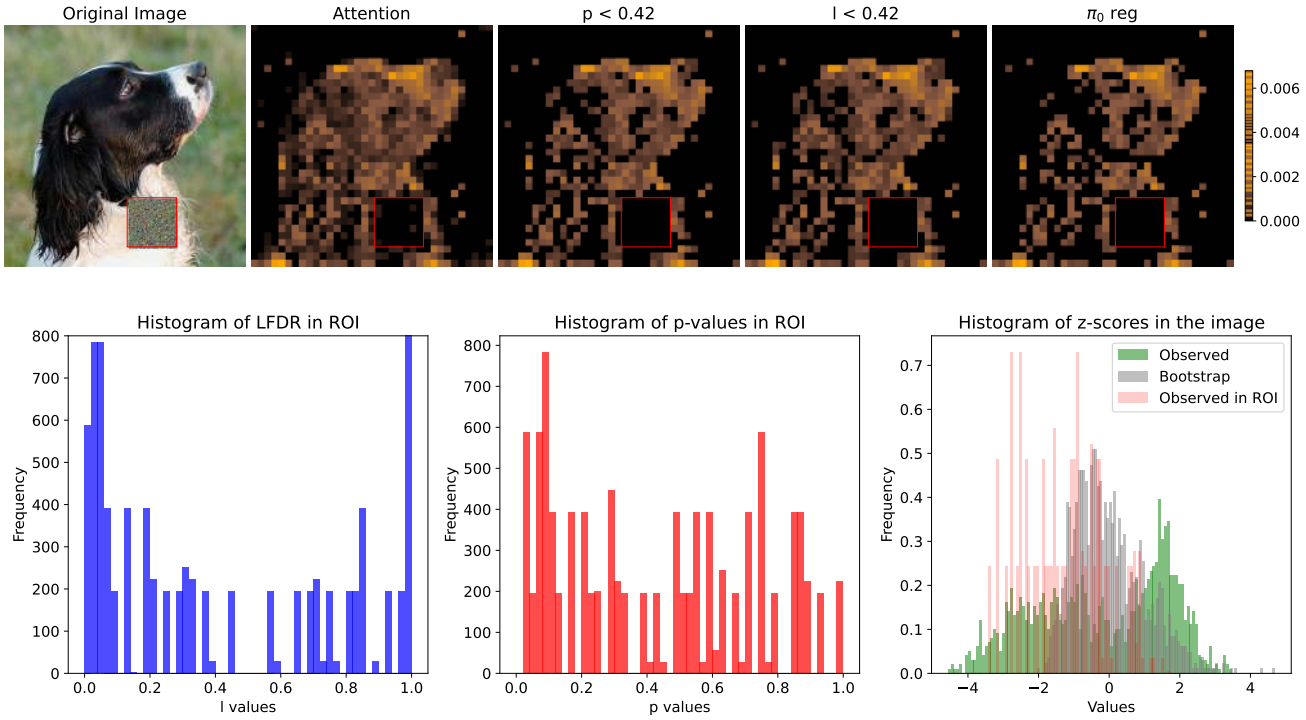


Figure S23: Attention map before and after regularization via different shrinkage methods (p-thresholding, l-thresholding and π_0-thresholding) for the same image as in Fig. 2 (`n02102040_821.JPEG`), but processed with a DINOv2 ViT with patch size 14×14 (top row); Histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image (bottom row). The noise patch in the image and attention maps is highlighted with a red frame.

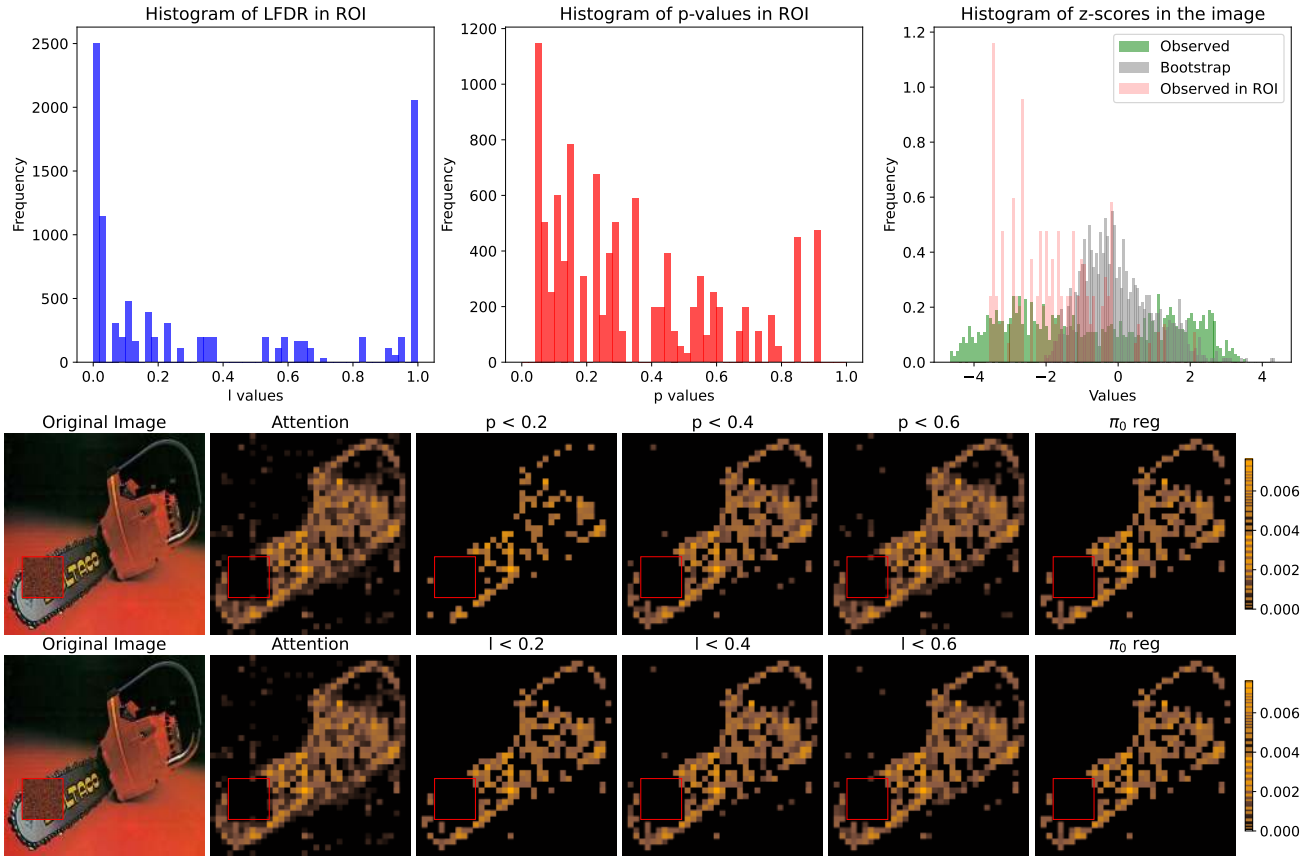


Figure S24: Example of a perturbed image (`n03000684_5970.JPEG`) with the attention map before and after regularization with the DINO v2 ViT (patch size 14×14) via p-thresholding (middle row) and l-thresholding (bottom row) with the respective thresholds varying from 0.2 to 0.6. The last attention map in each of the rows to the right corresponds to π_0-thresholding. The top row shows the histograms of p-values in ROI, LFDR values in ROI and z-scores of the observed attention scores (green), bootstrap attention scores (gray) and the attention scores observed in ROI (red) corresponding to the image. The noise patch in the image and attention maps is highlighted with a red frame $(44, 300)$.

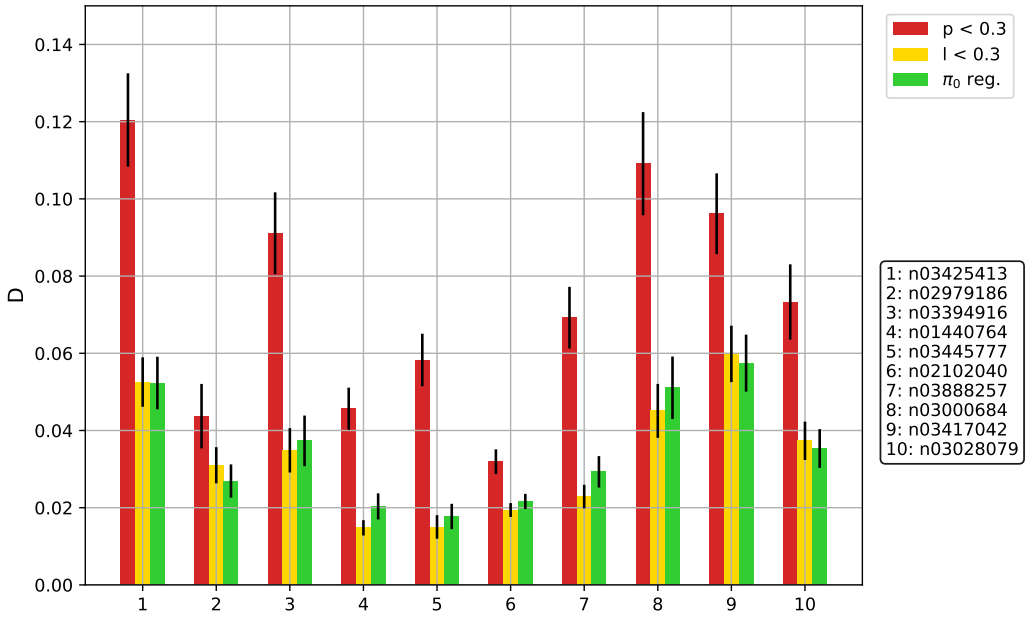


Figure S25: The average suppression factor D for different categories of images from the Imagenette validation subset processed with the DINO-v2 ViT model with patch size 14 and for different shrinkage methods (without z-filtering). The relative uncertainty of the results are shown with the black lines.