paragraph

Quantitative Layer-wise Analysis. Based on mean activation differences ($\Delta\mu_n$) and normalized z-scores (z_n), the aggregated behavior of neurons in each network layer has been analyzed. These metrics make it possible to identify units with differential responses between correctly classified (**OK**) images and misclassified as *castle* (**KO**) images.

The results show that:

- In the initial layers (`ReLU` and `layer1`), the distributions of $\Delta\mu_n$ are centered around zero and show limited variation. Z-score values are generally low, although a few neurons exhibit slightly higher deviations.
- From `layer2` onwards, the dispersion of $\Delta\mu_n$ increases, and z-scores begin to highlight neurons with more pronounced behaviors. This could indicate the emergence of internal patterns differentiating OK and KO.
- The deeper layers (`layer3` and especially `layer4`) present a much wider distribution, with long tails and extreme values both in activation and inhibition. The z-score heatmaps show a higher density of neurons with elevated absolute values ($|z_n| > 2$), suggesting that high-level semantic encoding could play an important role in predictive confusion.

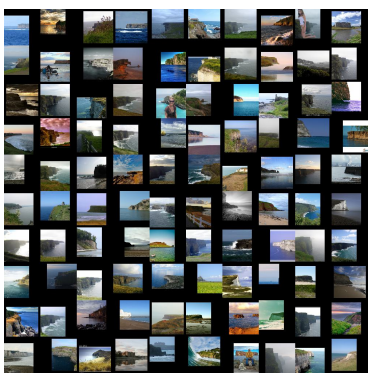
This statistical pattern provides an objective criterion to identify neurons with anomalous behaviors, i.e., with marked activation differences between correct and incorrect classifications. To qualitatively interpret these anomalies, a ranking has been generated of the neurons with the highest z-score values in the `layer4` layer, along with their semantic descriptions obtained from the visual stimuli that most strongly activate them.

Table **??** shows that many of these neurons are associated with landscape-related elements such as water, birds, or natural scenes. This reinforces the hypothesis presented in Figure **??**, where several *palace* images misclassified as *castle* featured open or natural environments.

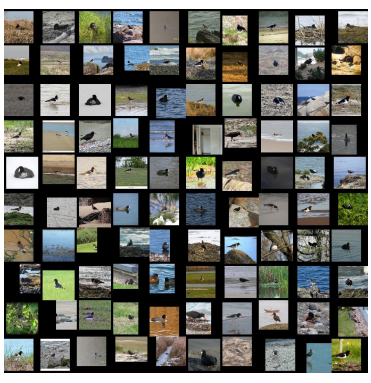
Table 1: Conflicting neurons with landscape-related semantics in the `layer4` layer.

#	Neuron	$\Delta\mu$	z	Description
1	1858	+2.209	+2.05	*"water and landscapes"*
2	1096	+0.936	+2.10	*"birds, water, and sand."*
3	744	+1.176	+2.30	*"a variety of animals, frogs and alligators, in their natural habitats."*

To complement the numerical analysis above, the following visual mosaics correspond to the neurons with the highest positive z-score deviations in the `layer4` layer, all of them related to landscapes, aquatic scenes, or natural environments. These images represent the visual stimuli that most strongly activate each of these neurons.



(a) Neuron 1858



(b) Neuron 1096