

# 1 ROBUSTNESS TESTING FOR IDENTIFIED CONCEPTS

## 1.1 TARGETED ATTACK

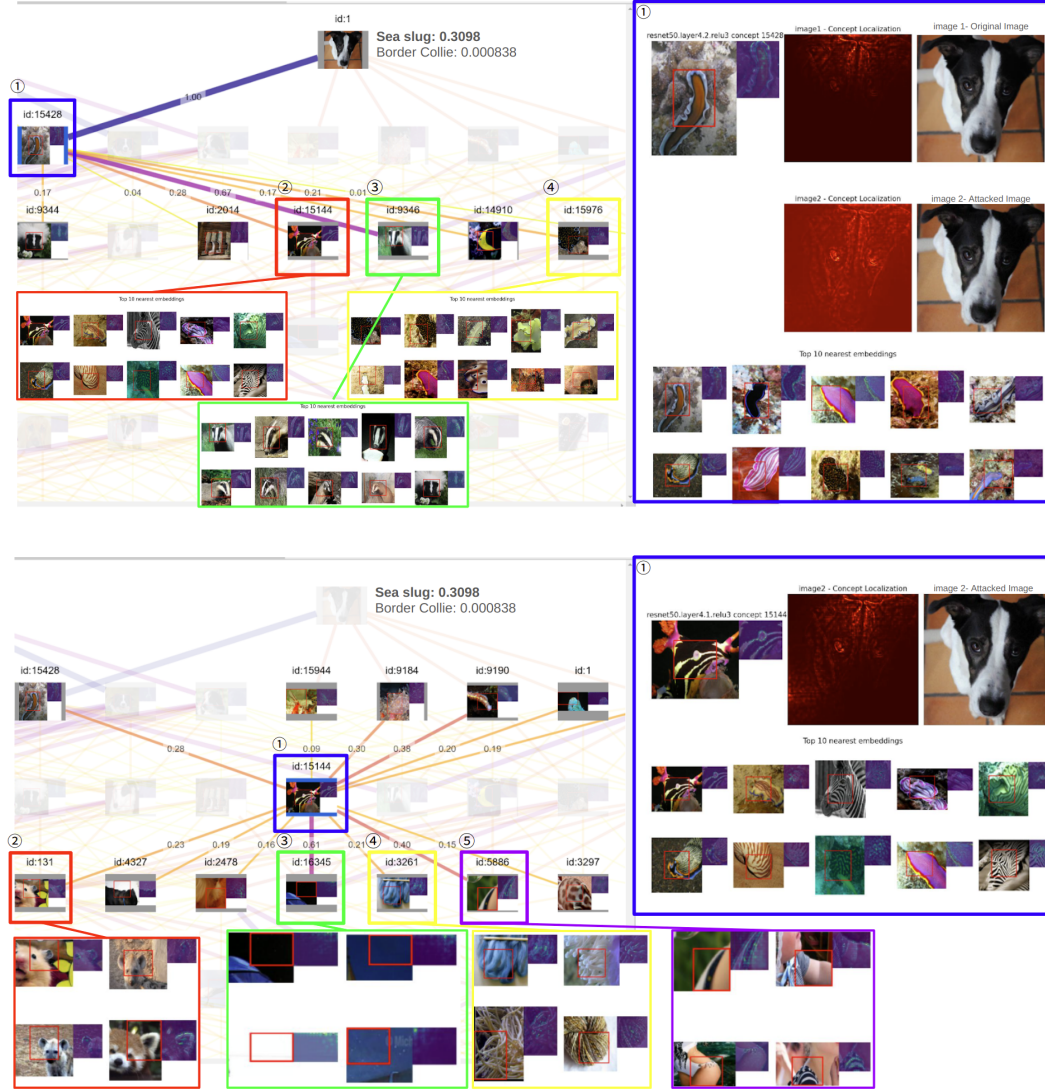


Figure 1: Causal explanation graph for Targeted Adversarial Attack. We attacked the border collie image to be identified as a sea slug. **Top:** The sea slug-related concept (①) was dominant at the last layer. At Layer 4.1, the sea slug concept (①) was most influenced by three key concepts: the stripe concept (②), black-white furry head concept (③), and the slurp body concept (④). Interestingly, the black-white furry head concept, which closely resembles a critical concept used in the correct classification (cosine similarity > 0.9), was also dominant in forming the corrupted higher-level concept of the sea slug. This suggests that the targeted adversarial attack might build corrupted higher-level concepts by combining non-corrupted features with corrupted features. **Bottom:** A similar pattern is evident with the stripe concept (①) at Layer 4.0, which was influenced by the black-white round ear concept (②), the monochrome background corner concept (③), the wrinkle concept (④), and the black-white stripe concept (⑤).

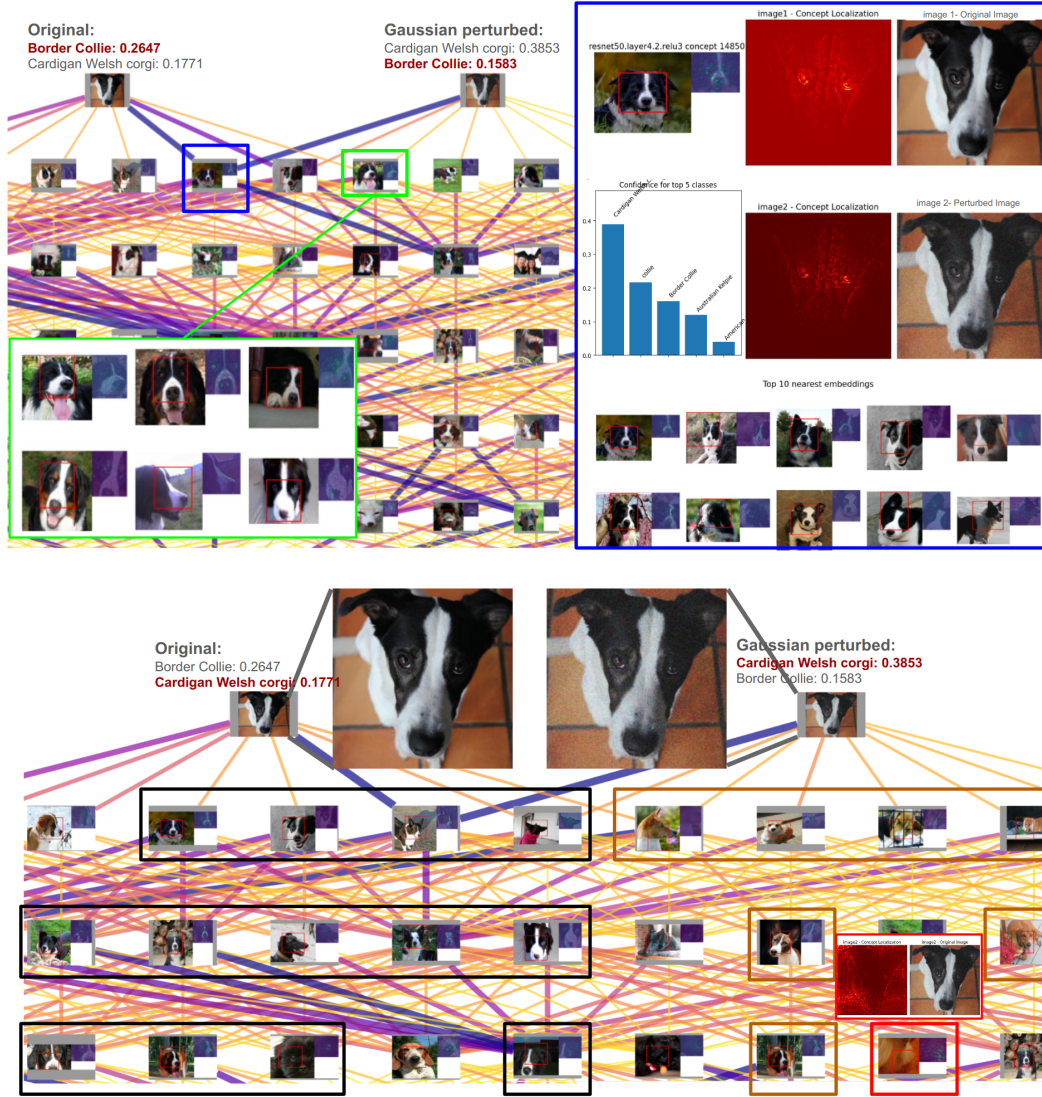


Figure 2: The causal explanation graphs for Gaussian Noise Attack. The model misclassified the border collie image as a Cardigan Welsh corgi due to the heavy Gaussian noise ( $\sigma = 70$ , considering that RGB values are integers ranging from 0 to 255). **Top:** Causal explanation graph for class ‘Border Collie’. Despite the misclassification of the perturbed image, the key concepts essential for correctly identifying the image as a border collie remained intact. This demonstrates the robustness of our method in preserving the underlying causal structure, even under significant noise perturbation. **Bottom:** Causal explanation graph for class ‘Cardigan Welsh corgi’. When the image was perturbed with Gaussian noise, brown corgi-related concepts appeared throughout the layers. The graphs revealed that the brown background color played a role in forming the corrupted “brown dog” concept, contributing to the misclassification.