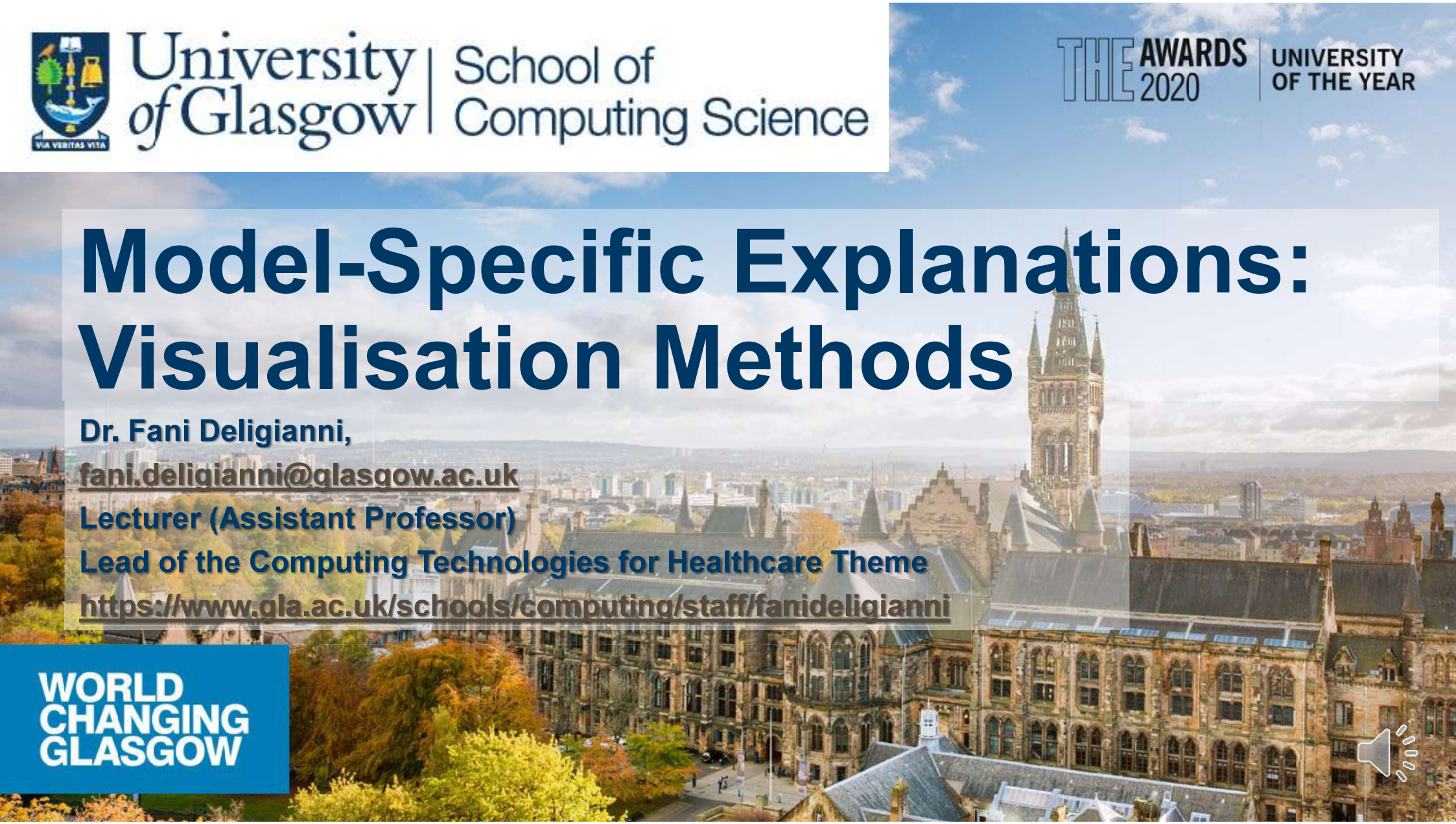


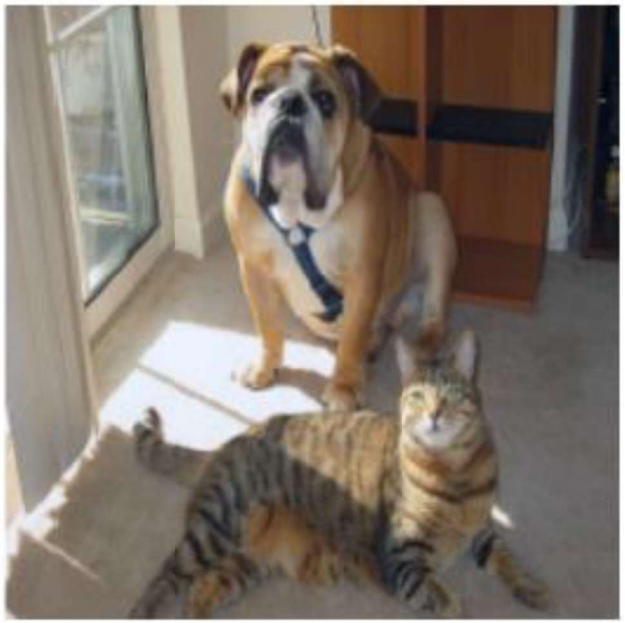
Model-Specific Explanations: Visualisation Methods

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Model-Specific Approaches

- Guided Backpropagation
- Class Activations Maps
- Gradient Weighted Class Activation Maps (GRAD-CAM)
- Guided Grad-CAM

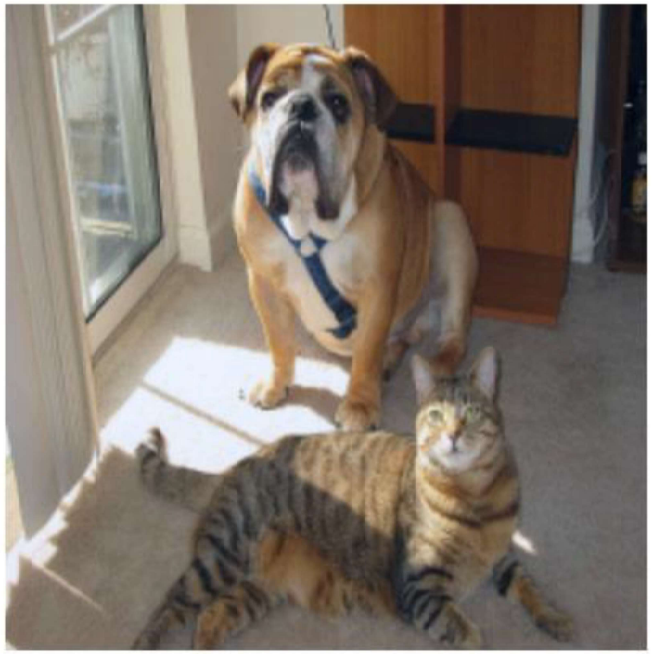
Good Visual Explanation

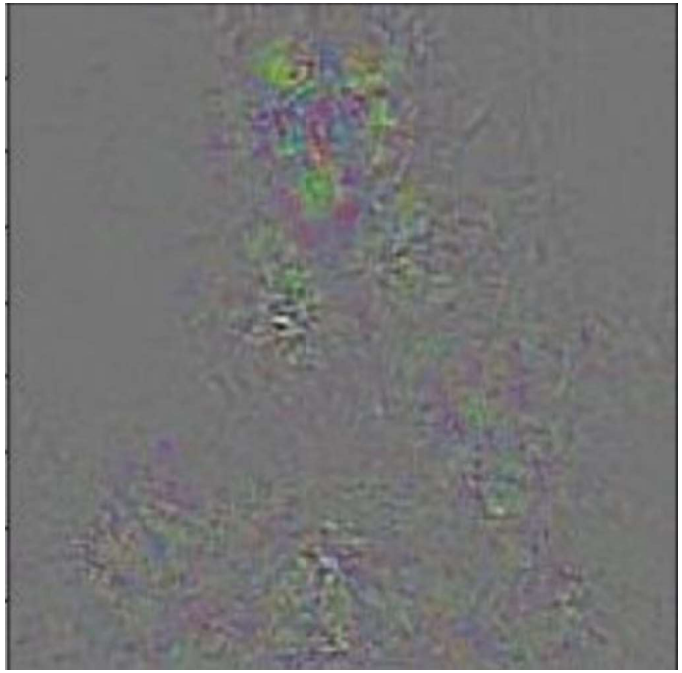


- Class-discriminative
- High-resolution

Selvaraju et al. ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, IJCV, 2021

Backpropagation

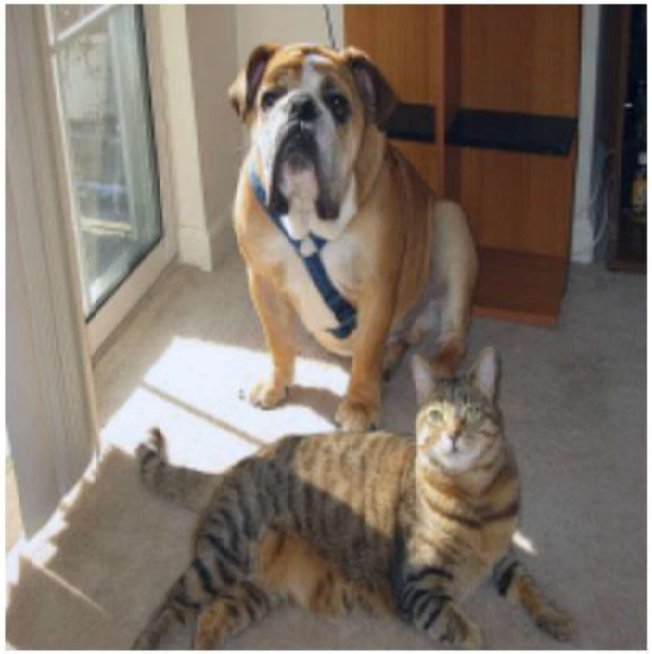




Class-discriminative ✗

High-resolution ✗

Guided Backpropagation



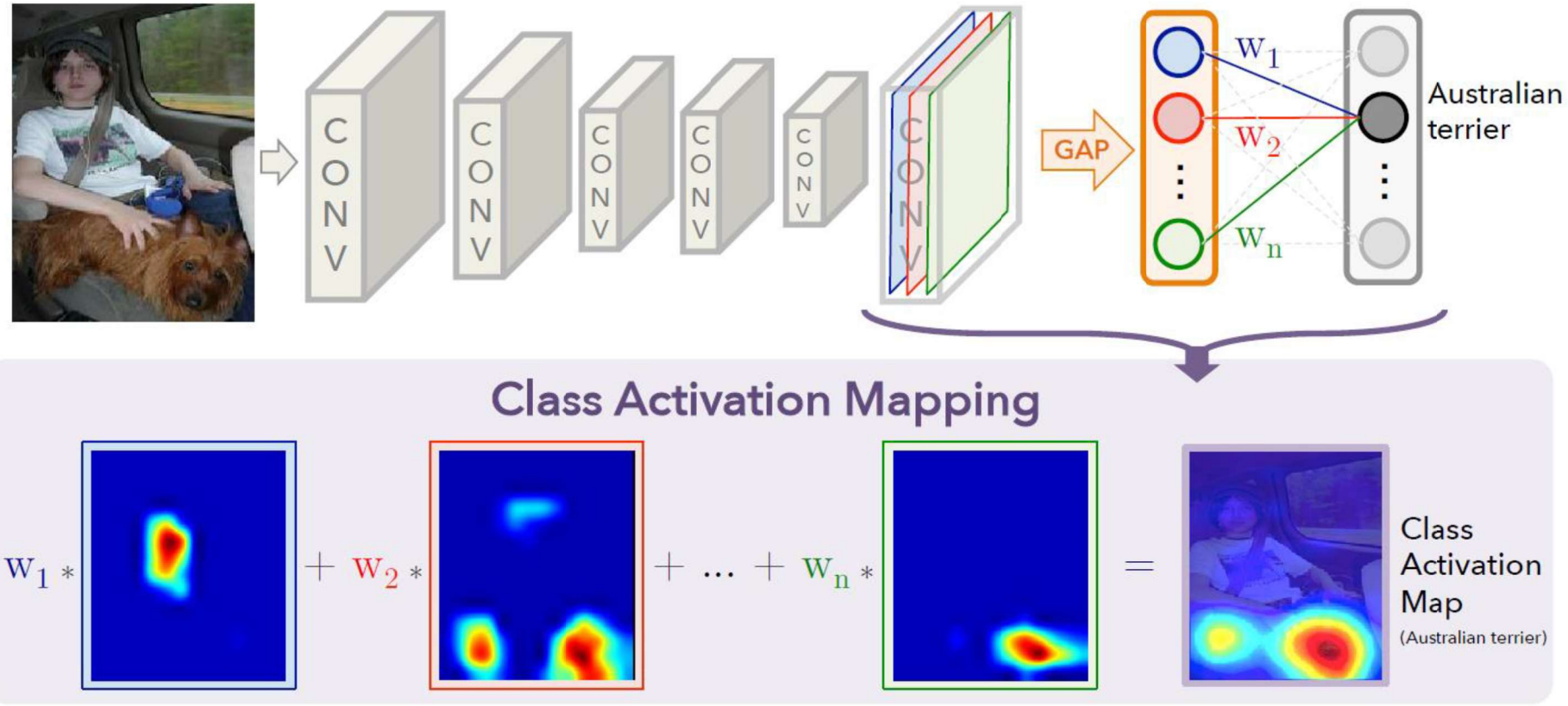
Guided Backpropagation ‘Cat’

Guided Backpropagation ‘Dog’

Class-discriminative ✗

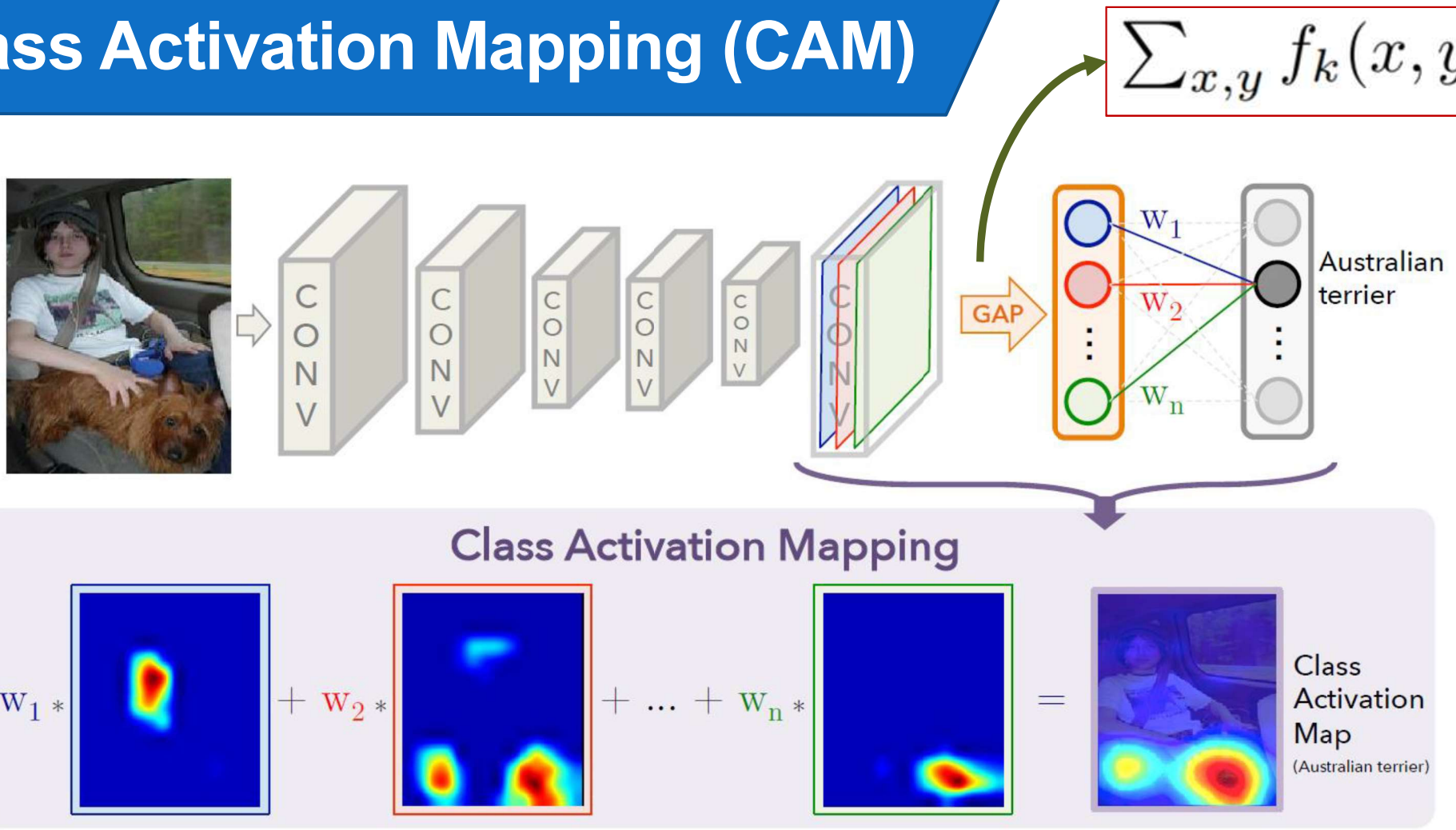
High-resolution ✓

Class Activation Mapping (CAM)



Zhou et al. 'Learning Deep Features for Discriminative Localization', CVPR, 2016

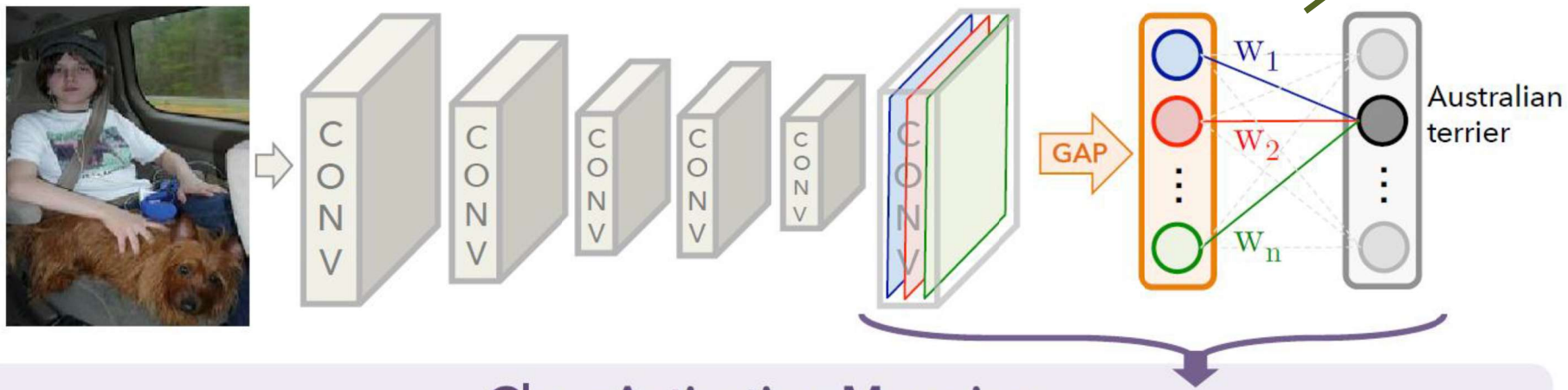
Class Activation Mapping (CAM)

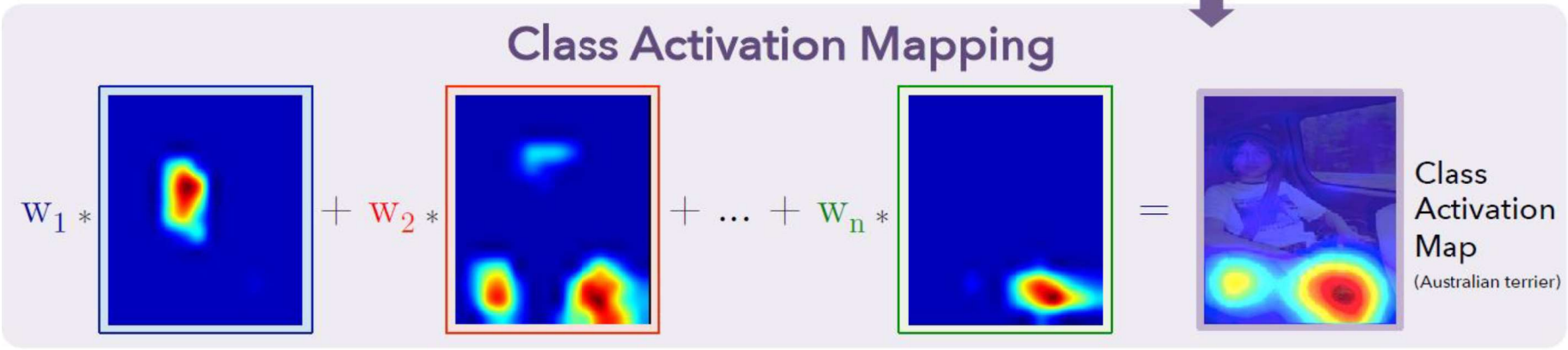


Zhou et al. 'Learning Deep Features for Discriminative Localization', CVPR, 2016

Class Activation Mapping (CAM)

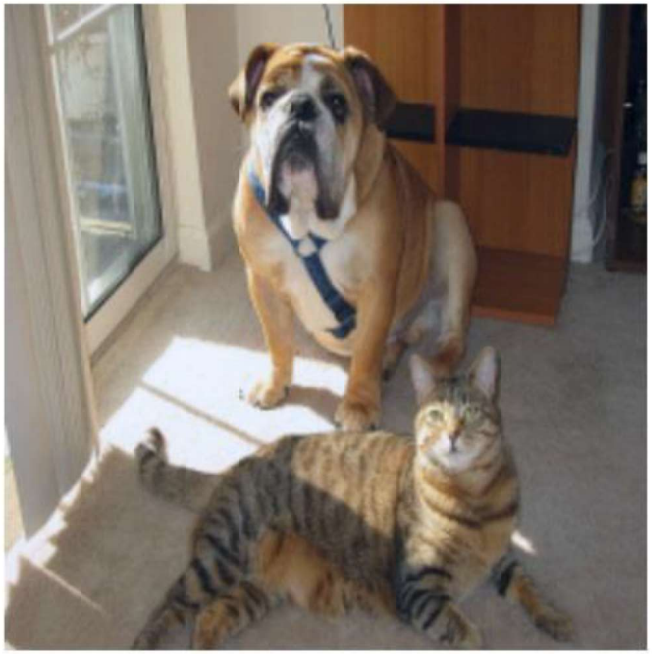
$$\sum_{x,y} \sum_{k} w_k^c f_k(x,y)$$



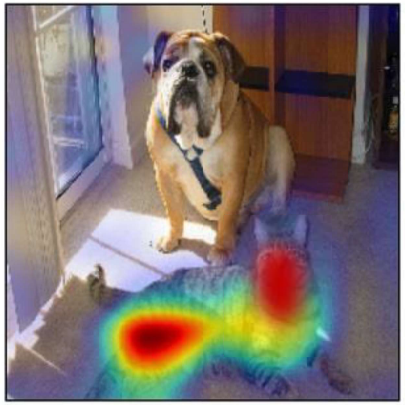


Zhou et al. 'Learning Deep Features for Discriminative Localization', CVPR, 2016

CAM



Guided Backpropagation ‘Cat’



CAM ‘Cat’

Guided Backpropagation ‘Dog’

CAM ‘Dog’

Class-discriminative ✓

High-resolution ✗

Requires GAP ✗

CAM Characteristics

- Class discriminative local explainability approach
- Weekly-Supervised CNN architectures
- Post-hoc Explanations
- Architecture needs to change to involve a convolutional layer followed by a Global Average Pooling Layer

Summary

- Backprobagation techniques can produce high resolution representations.
- They have been criticized that they are not decision specific
- Class activation maps produced class discriminative explanations
- CAM requires a specific architecture that involves a Global average pooling layer after the last convolutional layer, followed by a dense layer
- CAM trades off performance in order to provide local explanations

References

- Zhou et al. ‘Learning Deep Features for Discriminative Localization’, CVPR, 2016.
- Selvaraju et al. ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization’, International Journal of Computer Vision, 2019.