Towards Visual Explanations of Movie Scenario Segmentation

Xu Linning

The Chinese University of Hong Kong

1155128851@link.cuhk.edu.hk

Class Activation Maps (CAM)

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba.
Learning Deep Features for Discriminative Localization. CVPR'16

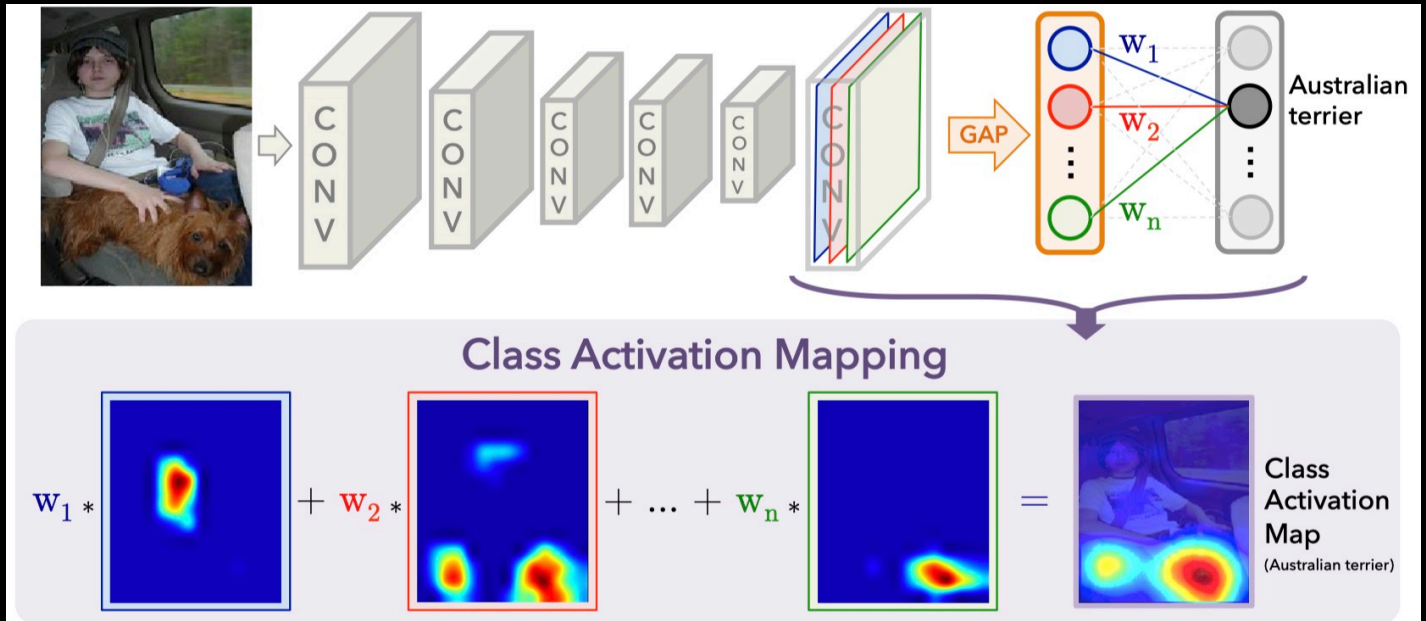


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

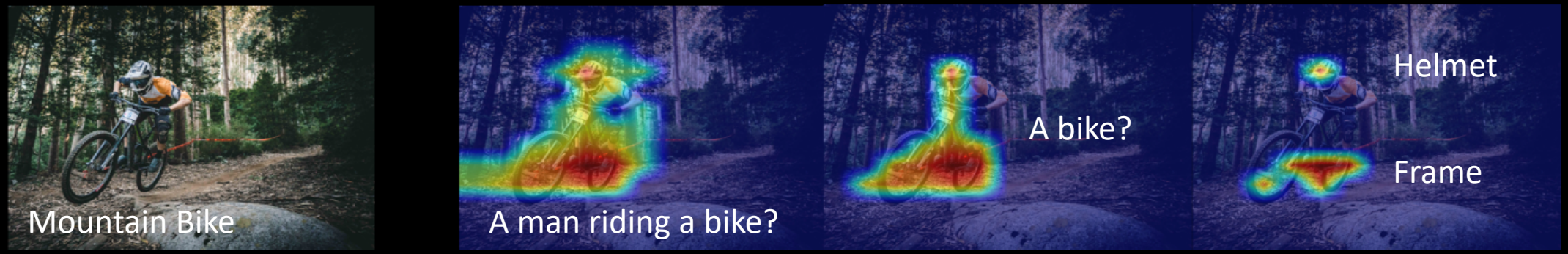
Class Activation Maps (CAM)

– Adaptive Threshold Strategy

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \quad \Rightarrow \quad M_c^*(x,y) = \begin{cases} M_c(x,y) & \text{if } M_c(x,y) > \theta \\ 0 & \text{otherwise} \end{cases}$$

$$O_c(x,y) = \alpha M_c^*(x,y) + \beta R(x,y)$$

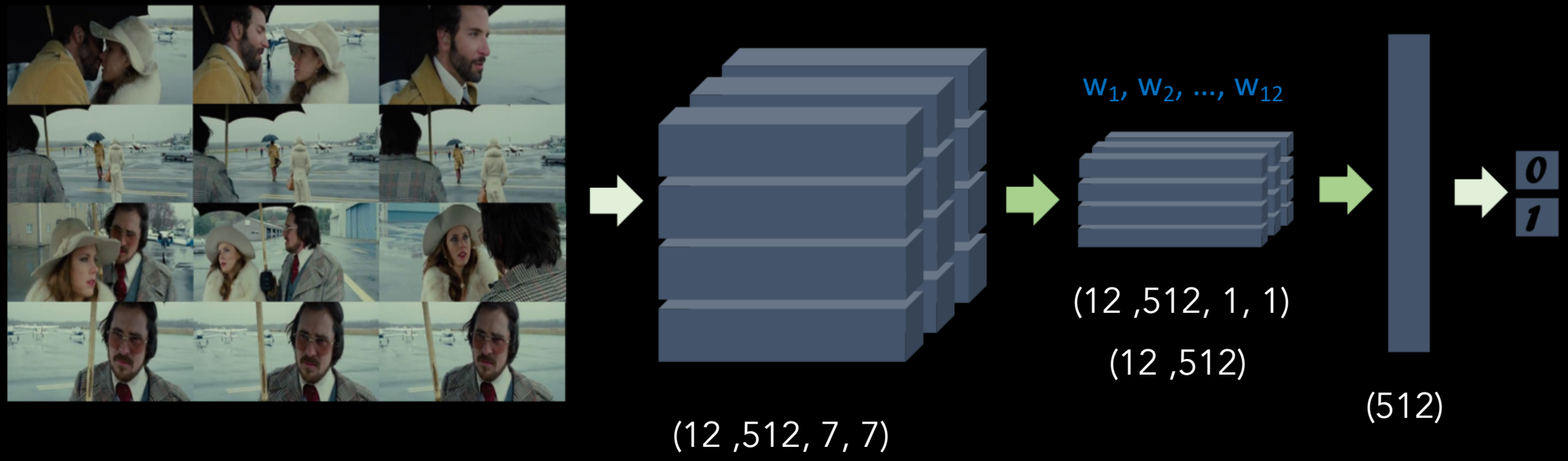
Original | θ = 0.2 | θ = 0.5 | θ = 0.8



CAM for Movie Setting

– Modified CAM & Network Structure

$$M_c(x,y) = \sum_k w_k^c f_k(x,y) \quad \Rightarrow \quad M_c(x,y,z) = \sum_z w_z^c \sum_k w_k^c f_k(x,y,z)$$



Ground Truth = 0

Paranorman (2012)

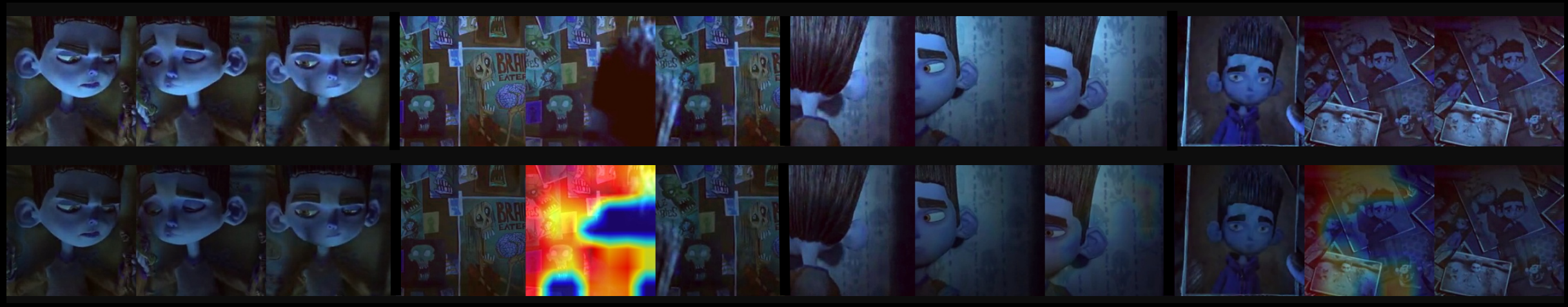
Prediction = 0

Prediction = 1

Ground Truth = 0

Paranorman (2012)

Prediction = 0



Prediction = 1

Ground Truth = 1

Shame (2011)

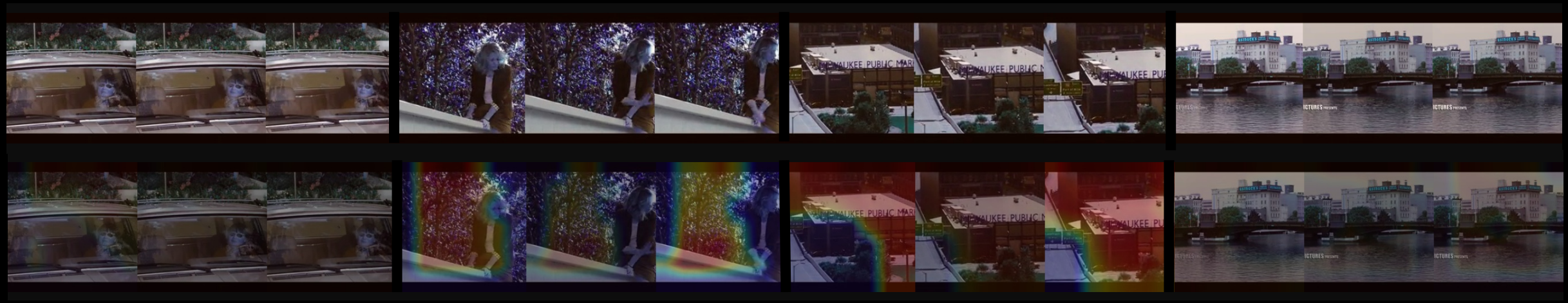
Prediction = 1

Prediction = 0

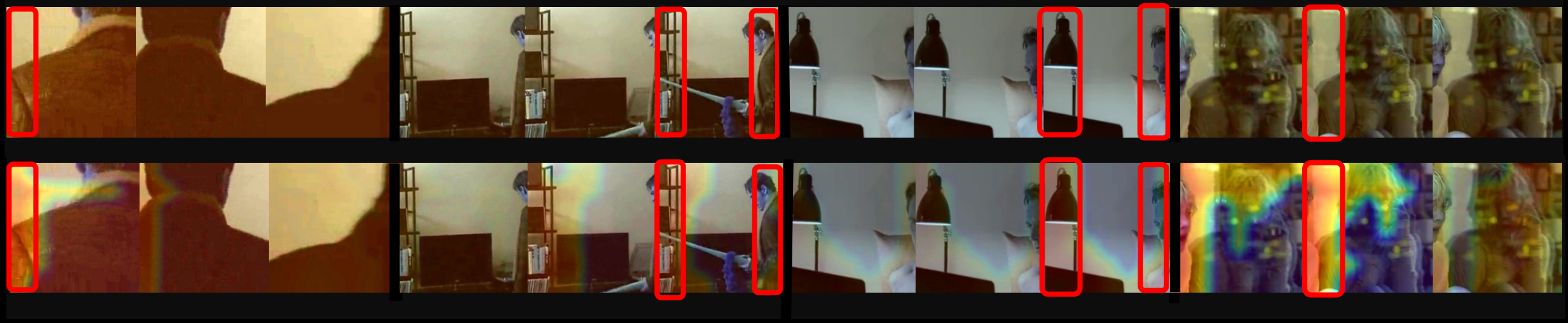
Ground Truth = 1

Shame (2011)

Prediction = 1



Prediction = 0



Summary

"Spot the Difference" Game

- *Class-discriminative?*
- *Object-oriented?*
- *NO !*

- *Looks at the Space Layout*
- *Sensitive to minor change (hue / lightness)*
- *Ignoring minor content changes*

→ *Condition on Human Interactions, Types of different filmmaking techniques (e.g., long shots/close-up), etc. More informative annotations are desired*
→ *Incorporation the temporal information in sequence*
→ *Proper Image Transformation*