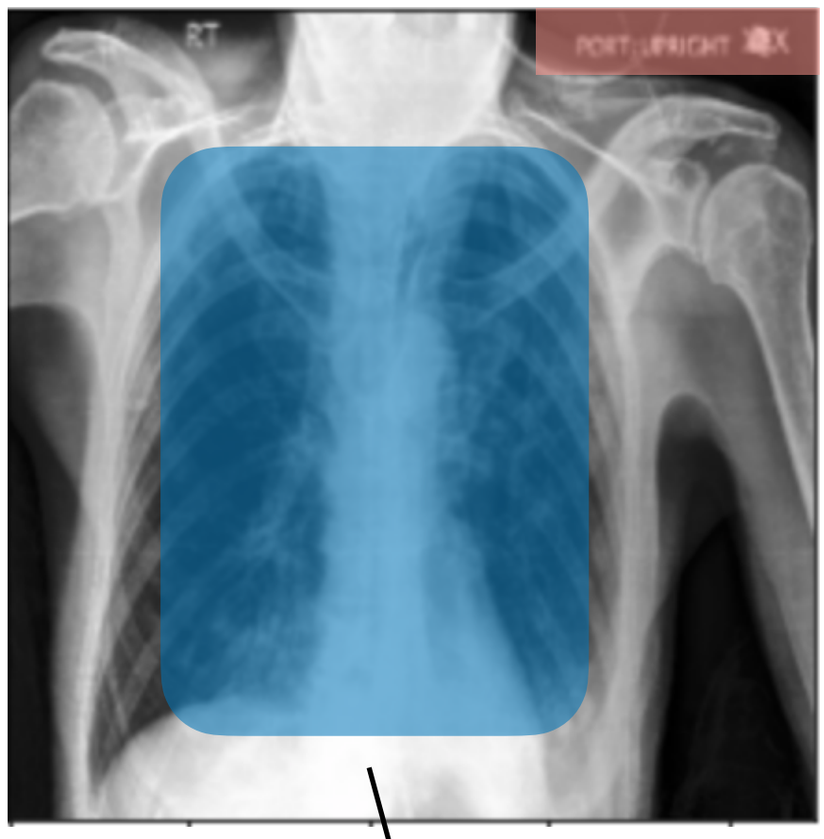
Chroma-VAE: Mitigating Shortcut Learning with Generative Classifiers

Wanqian Yang • Polina Kirichenko • Micah Goldblum • Andrew Gordon Wilson
New York University

Lazy models take shortcuts!

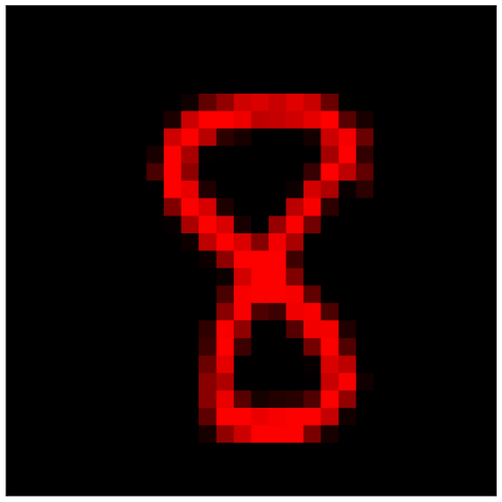


→ Shortcut: **Background**

Intended: **Pneumonia Signal**

Zech et al. (2018)

Lazy models take shortcuts!



Shortcut: **Color**
Intended: **Digit**

Shortcut: **Background**
Intended: **Pneumonia Signal**

Highly correlated feature at train time, but fails on distribution shifts

Can we learn a shortcut-invariant representation?

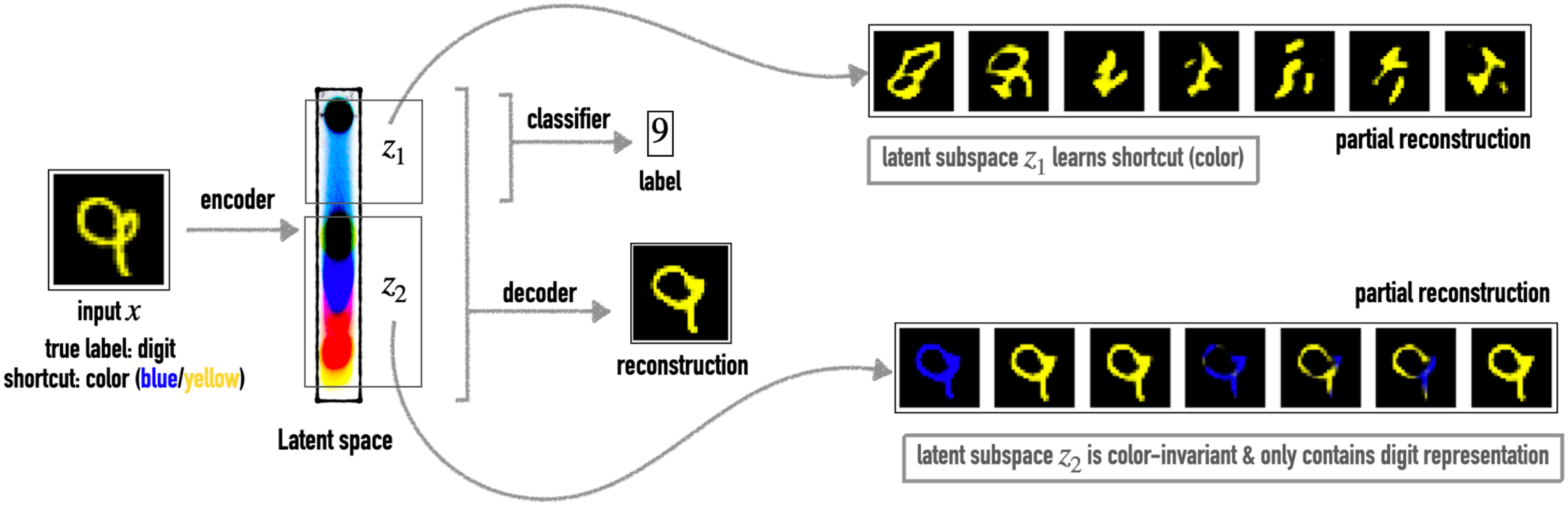
Can we do so without relying on additional data?

#1: Shortcuts are efficiently compressed

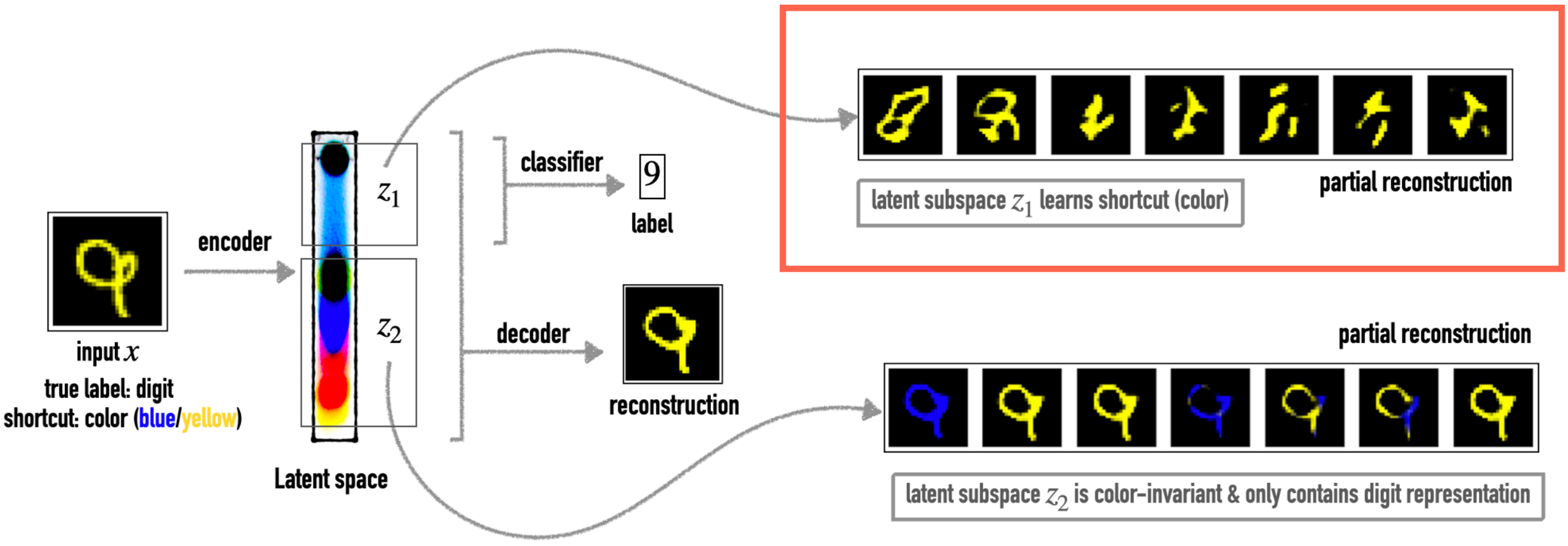
Original image
with shortcut
(blue patch)

Grad-CAM activations as bottleneck layer increases

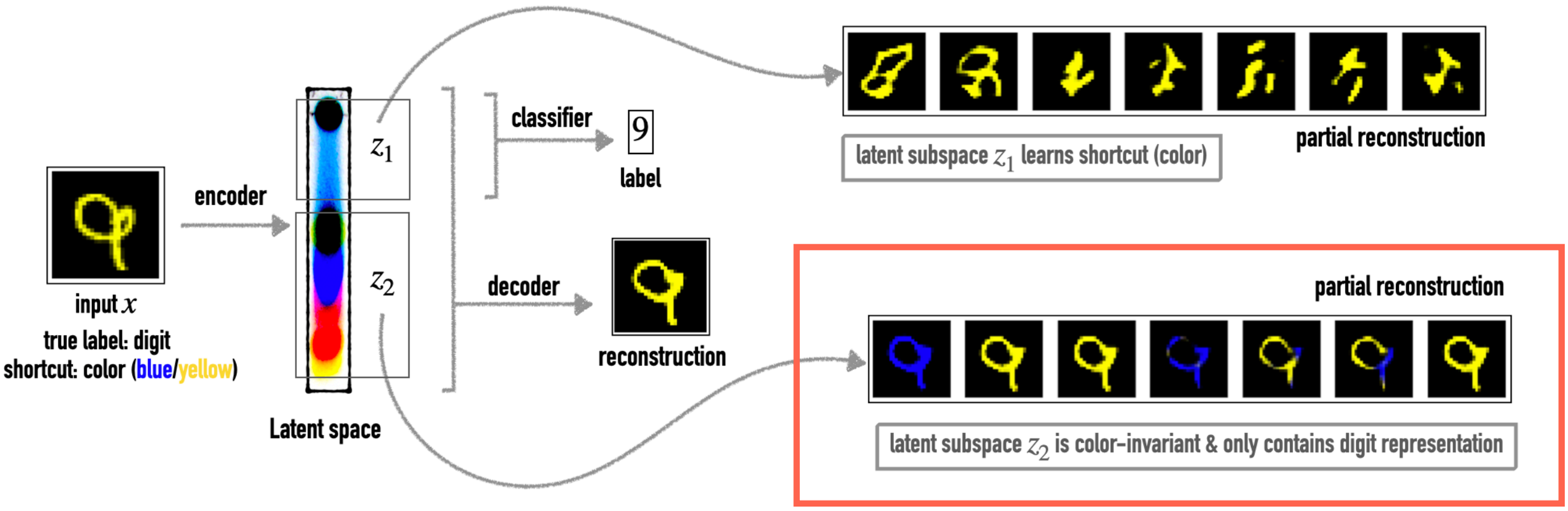
#2: Exploiting this using Chroma-VAE



#2: Exploiting this using Chroma-VAE



#2: Exploiting this using Chroma-VAE



Come chat with us!

Paper link somewhere down below (we assume)