University of Amsterdam

MLRC 2021 - Our Experience

Authors: Piyush Bagad, Danilo de Goede, Paul Hilders, Jesse Maas

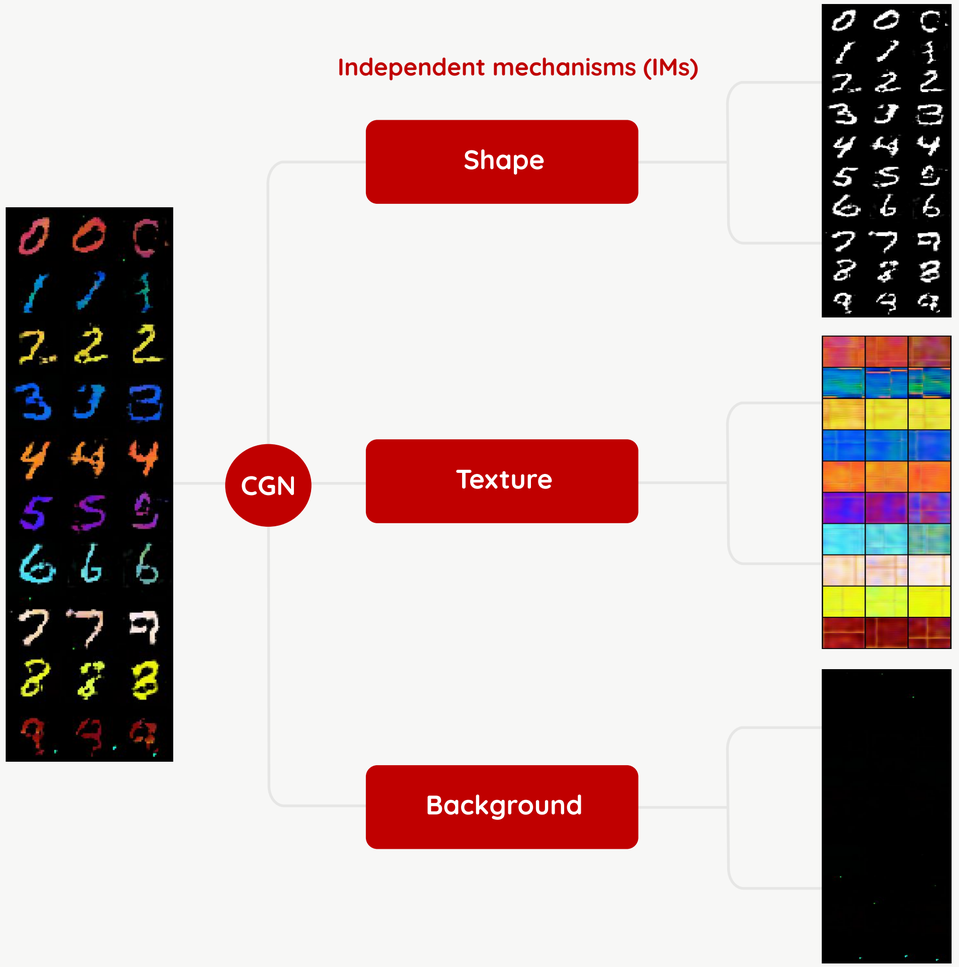
Supervisor: Christos Athanasiadis

Talk Overview

- Our work from FACT 2022
  - Context
  - The Counterfactual Generative Network (CGN)
  - Scope of Reproducibility
  - Our methodology and Results
- Lessons learnt, Tips and Suggestions

Context

- Deep Learning models tend to learn “shortcuts” that perform well on benchmarks.
- Shortcut learning causes models to be more sensitive to input perturbation and unseen input contexts.
- Sauer and Geiger (2021) propose an approach using a Counterfactual Generative Network.



Counterfactual Generative Network

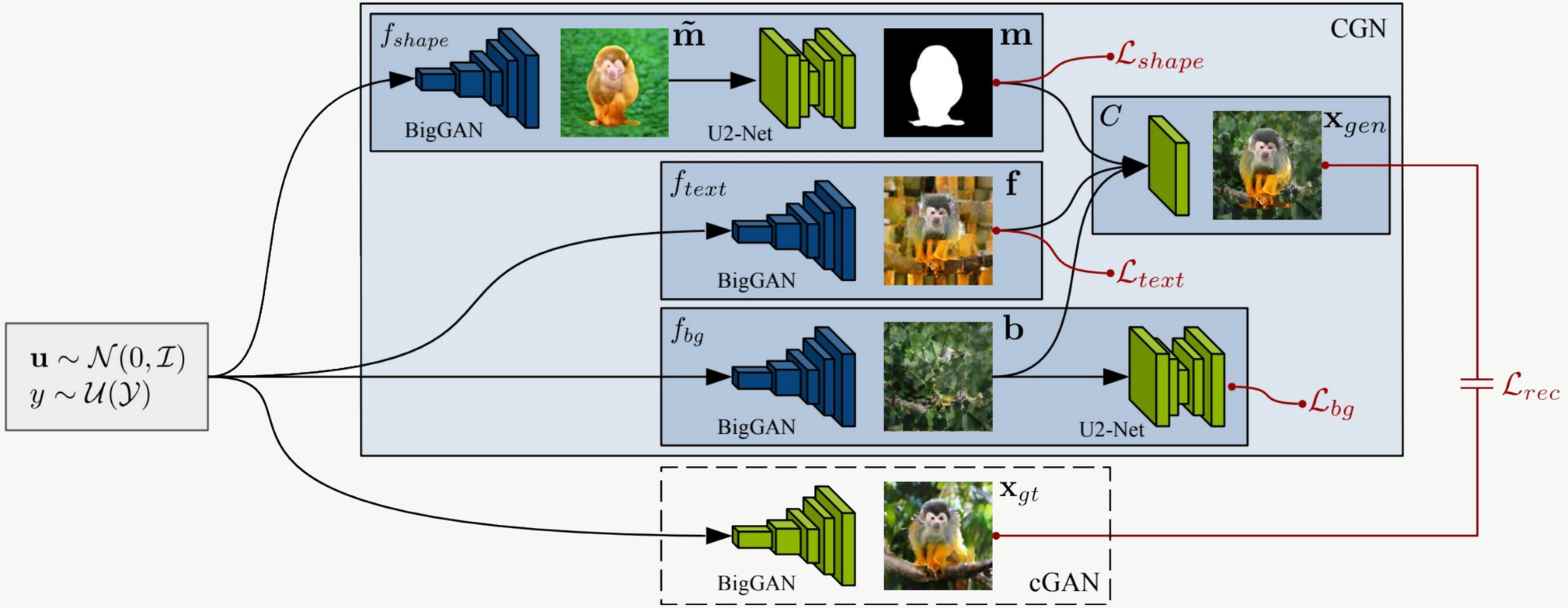


Figure 1. Architecture overview (ImageNet) of the Counterfactual Generative Network (Sauer and Geiger, 2021)

Counterfactual Generative Network

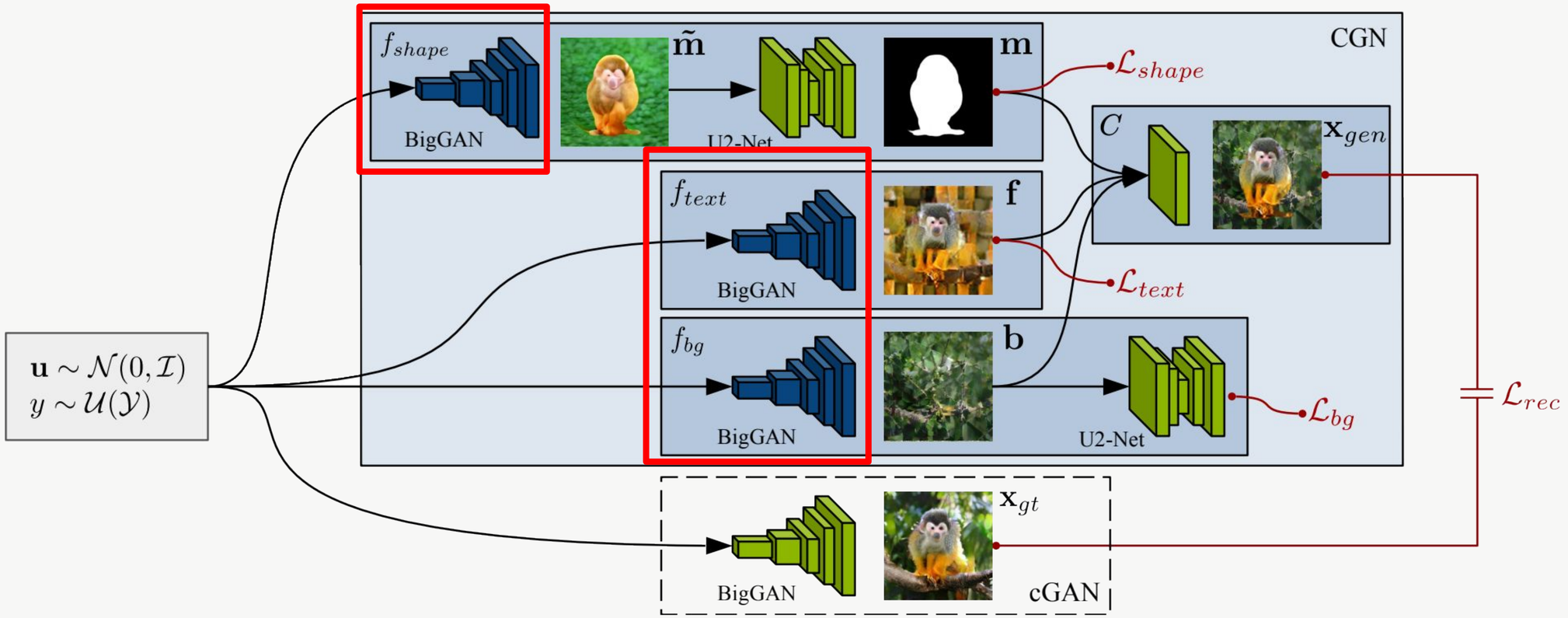


Figure 1. Architecture overview (ImageNet) of the Counterfactual Generative Network (Sauer and Geiger, 2021)

Counterfactual Generative Network

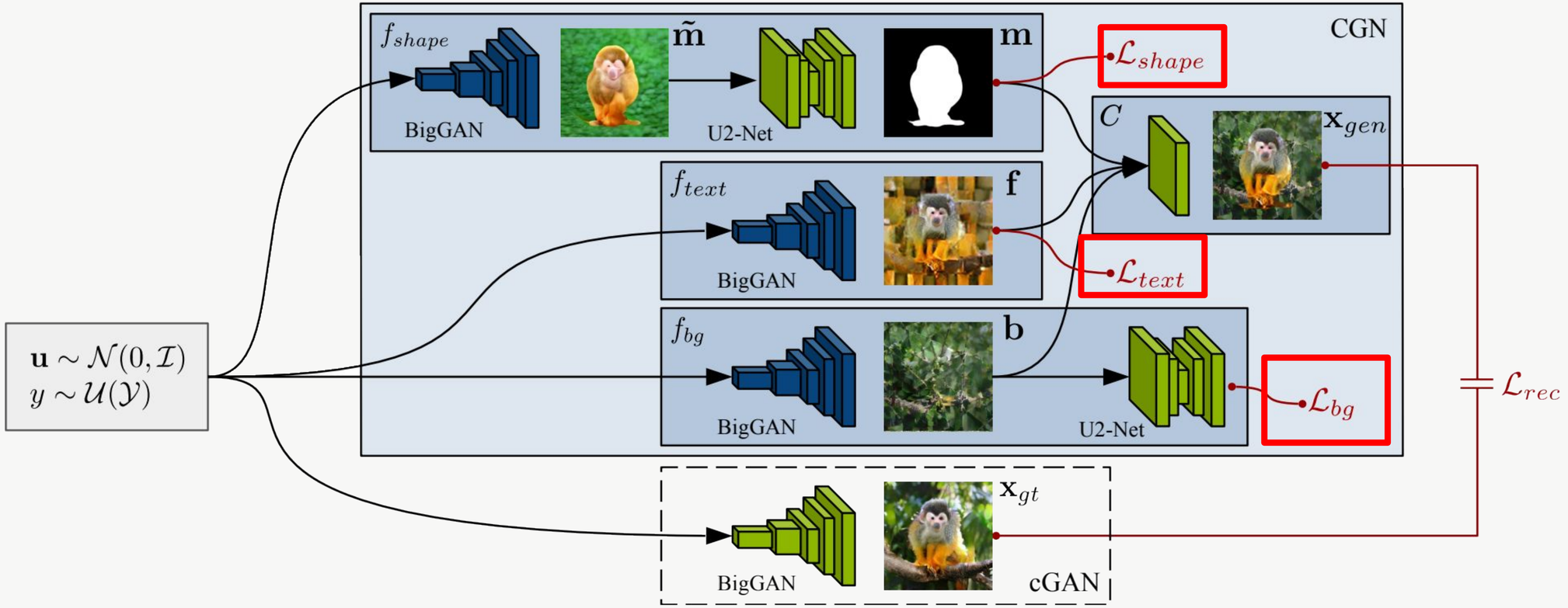


Figure 1. Architecture overview (ImageNet) of the Counterfactual Generative Network (Sauer and Geiger, 2021)

Counterfactual Generative Network

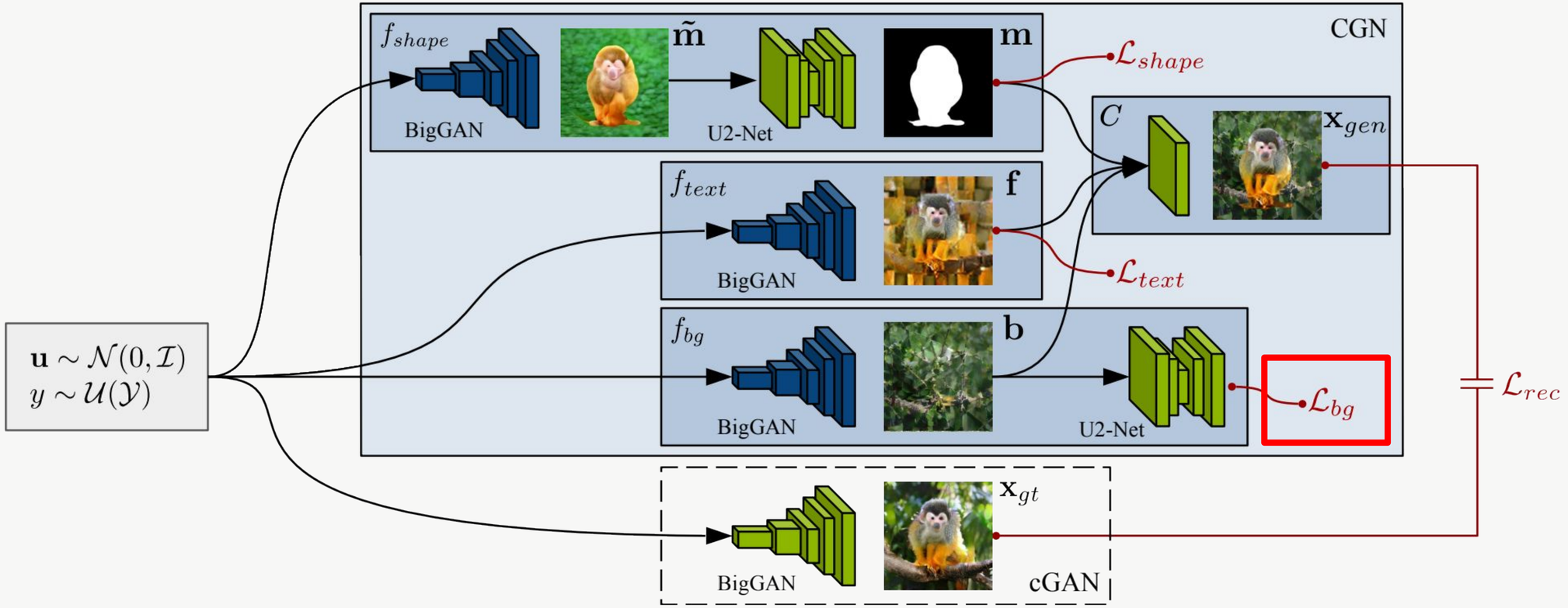


Figure 1. Architecture overview (ImageNet) of the Counterfactual Generative Network (Sauer and Geiger, 2021)

Counterfactual Generative Network

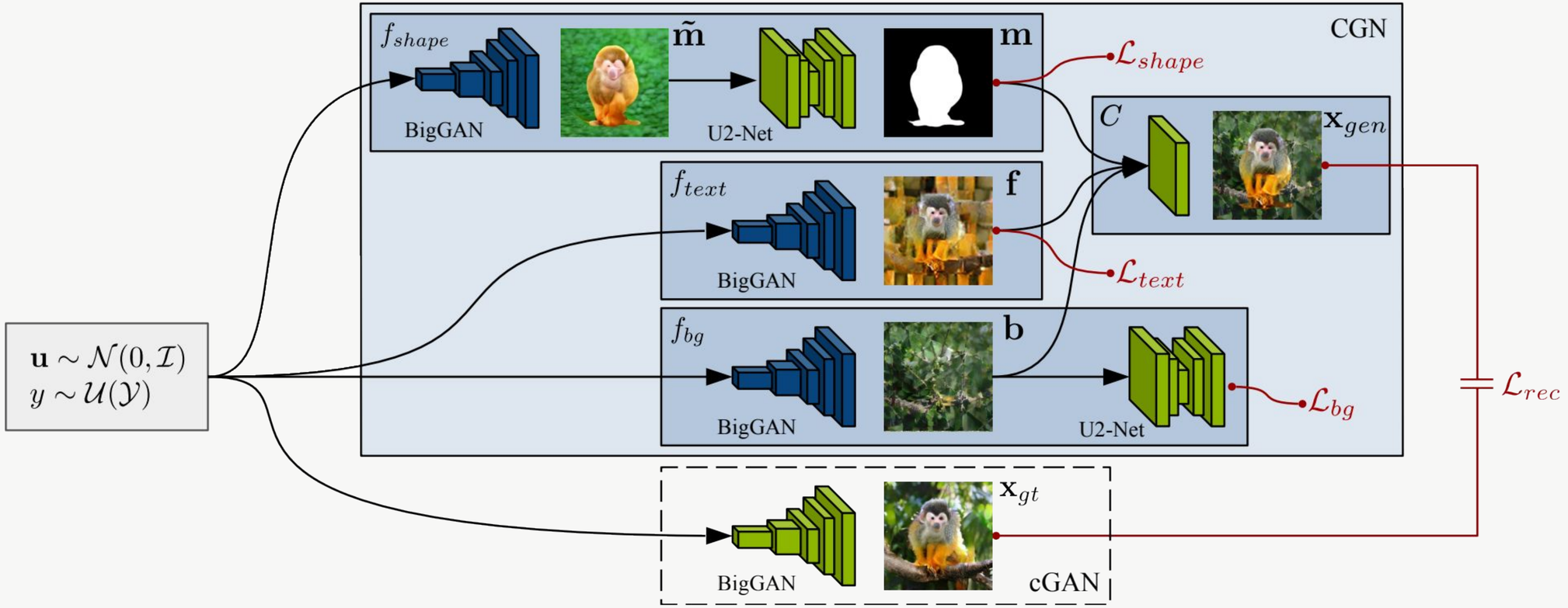
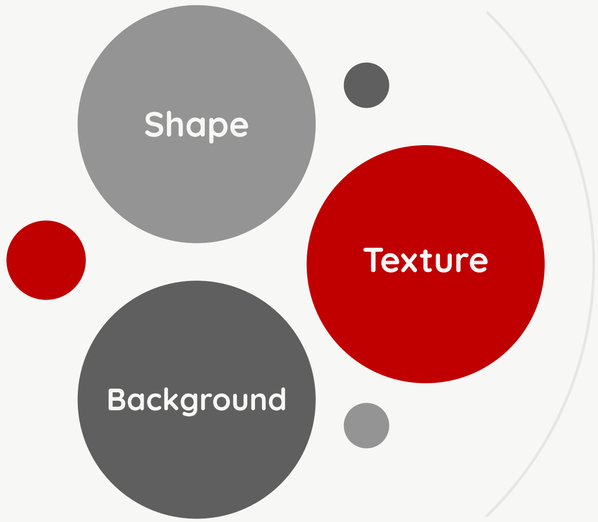


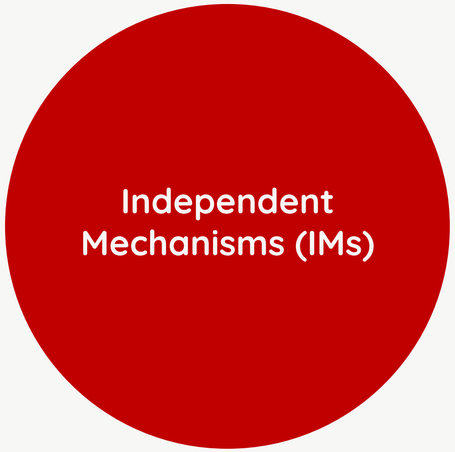
Figure 1. Architecture overview (ImageNet) of the Counterfactual Generative Network (Sauer and Geiger, 2021)

Scope of Reproducibility

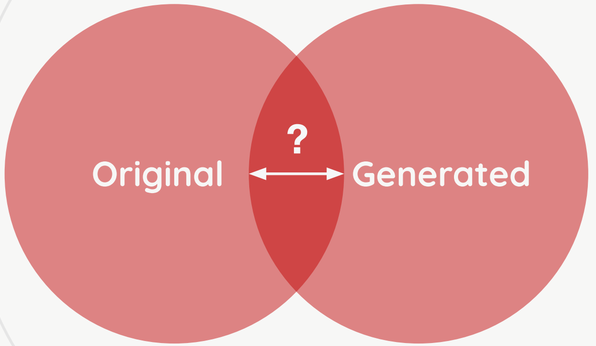
Main claims of the original paper



High-Quality Counterfactuals (HQC)



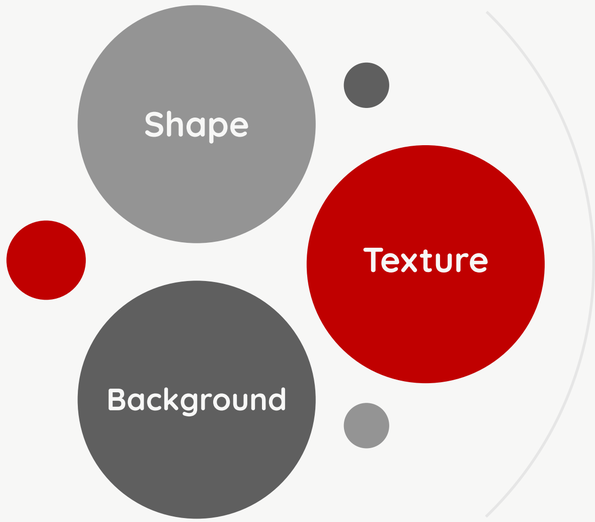
Inductive Bias Requirements (IBR)



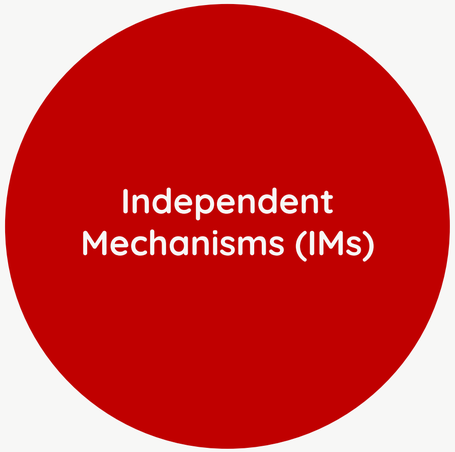
Out-of-Distribution Robustness (ODR)

Scope of Reproducibility

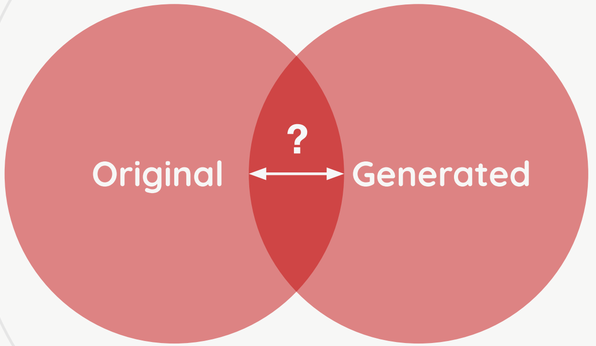
Main claims of the original paper



High-Quality Counterfactuals (HQC)



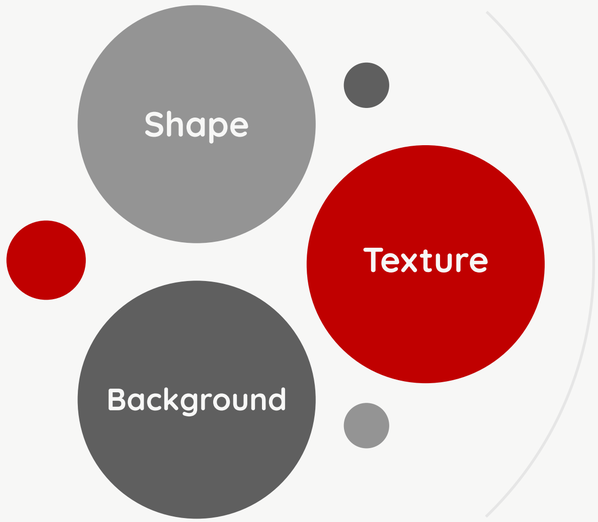
Inductive Bias Requirements (IBR)



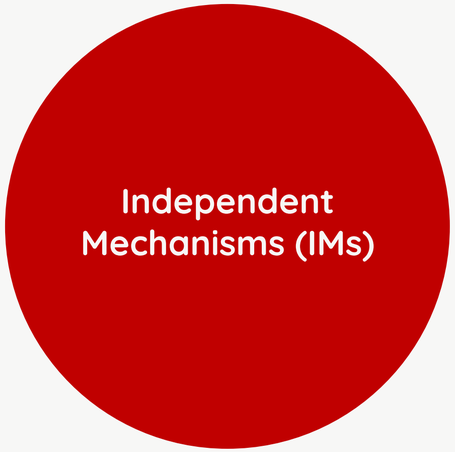
Out-of-Distribution Robustness (ODR)

Scope of Reproducibility

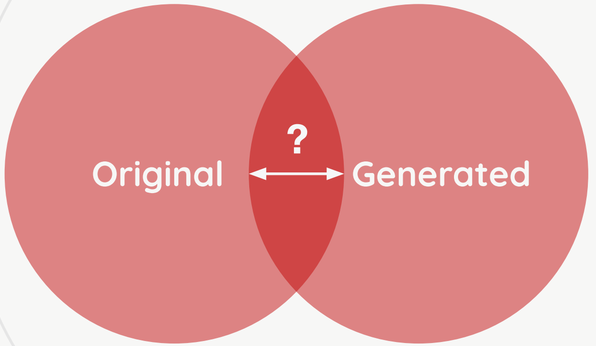
Main claims of the original paper



High-Quality Counterfactuals (HQC)



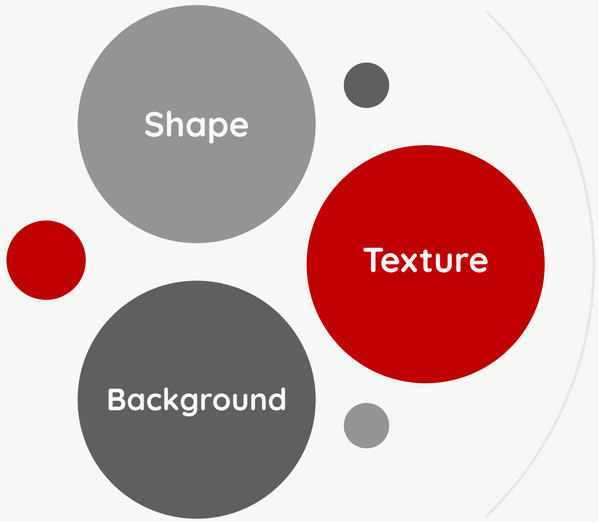
Inductive Bias Requirements (IBR)



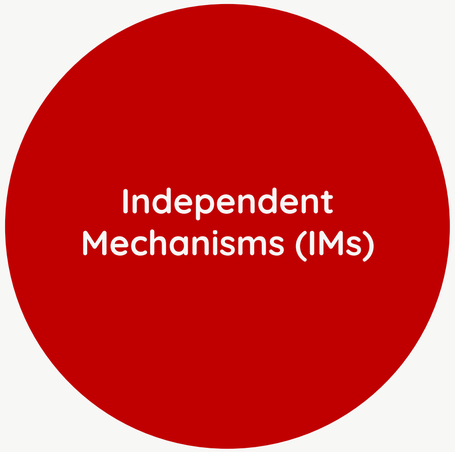
Out-of-Distribution Robustness (ODR)

Scope of Reproducibility

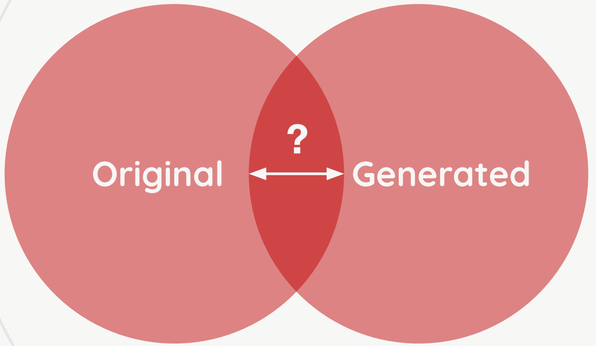
Main claims of the original paper



High-Quality Counterfactuals (HQC)



Inductive Bias Requirements (IBR)

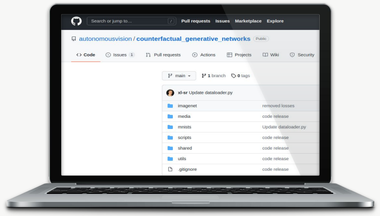


Out-of-Distribution Robustness (ODR)

Methodology

Models

The CGN model is publicly available on GitHub



Datasets

Various variants of MNIST and ImageNet

Experimental Setup + Metrics

Re-implement based on description of paper

Computational Requirements

112 + 48 GPU hours on a 1080Ti node (Lisa)

SURF SARA

Experimental results of reproducibility study

Claim 1: High-Quality Counterfactuals (HQC)

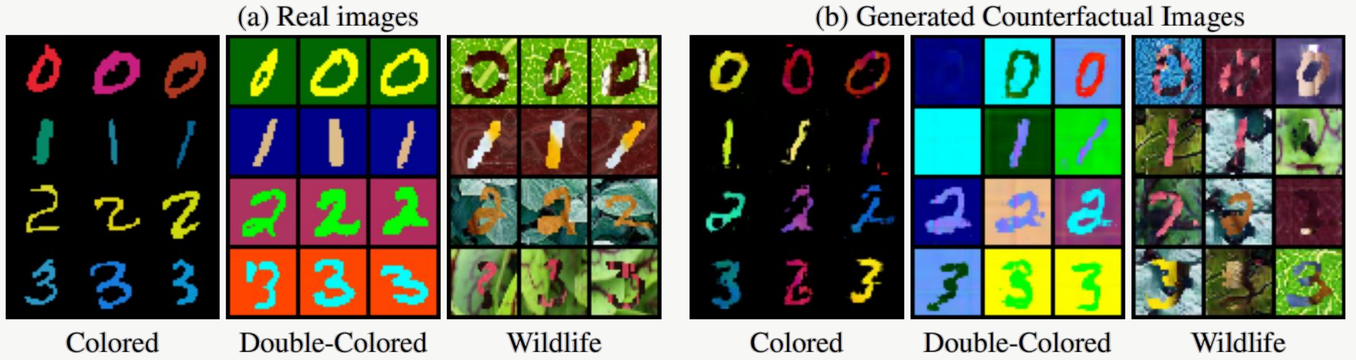


***Figure 2**. Reproduced qualitative results on MNIST variants*

Shape	Racer	Trench coat	Turtle	Vase	Malinois	Barrel
Texture	Clock	Cab	Cauliflower	Elephant	Viper	Piggy bank
Background	Toucan	Coral reef	Mushroom	Alp	Spider	Ibex

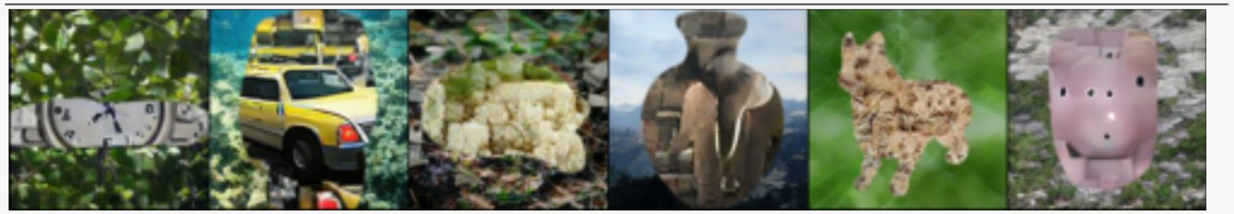


***Figure 3**. Reproduced qualitative results on ImageNet*

Claim 1: High-Quality Counterfactuals (HQC)

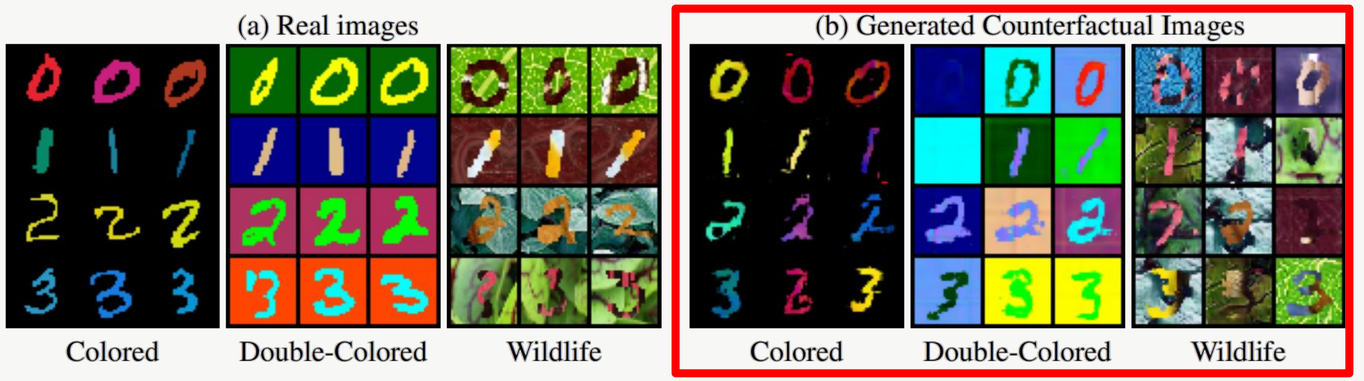


***Figure 2**. Reproduced qualitative results on MNIST variants*

Shape	Racer	Trench coat	Turtle	Vase	Malinois	Barrel
Texture	Clock	Cab	Cauliflower	Elephant	Viper	Piggy bank
Background	Toucan	Coral reef	Mushroom	Alp	Spider	Ibex

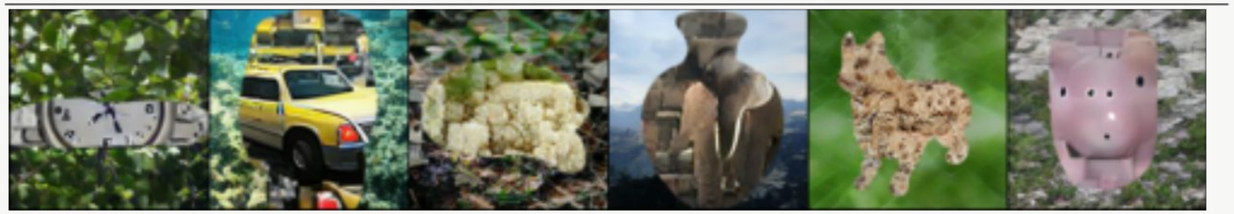


***Figure 3**. Reproduced qualitative results on ImageNet*

Claim 1: High-Quality Counterfactuals (HQC)

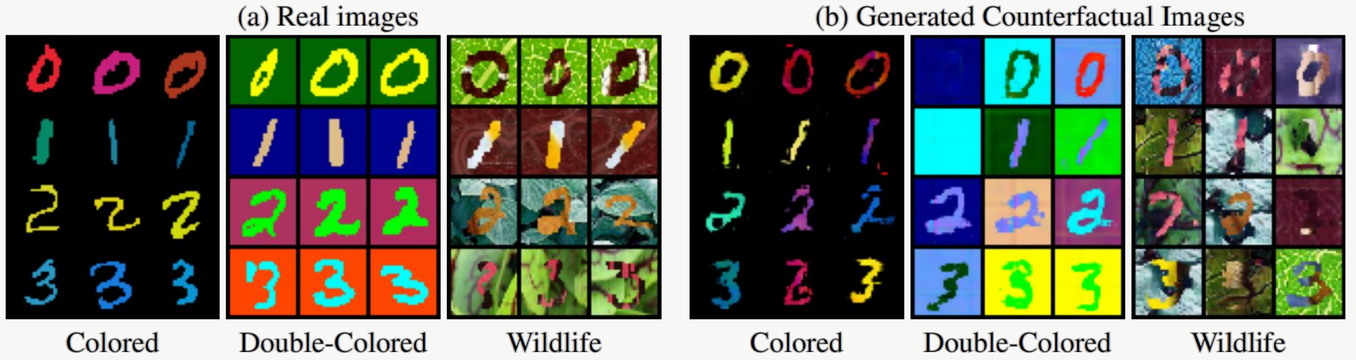


***Figure 2**. Reproduced qualitative results on MNIST variants*

Shape	Racer	Trench coat	Turtle	Vase	Malinois	Barrel
Texture	Clock	Cab	Cauliflower	Elephant	Viper	Piggy bank
Background	Toucan	Coral reef	Mushroom	Alp	Spider	Ibex

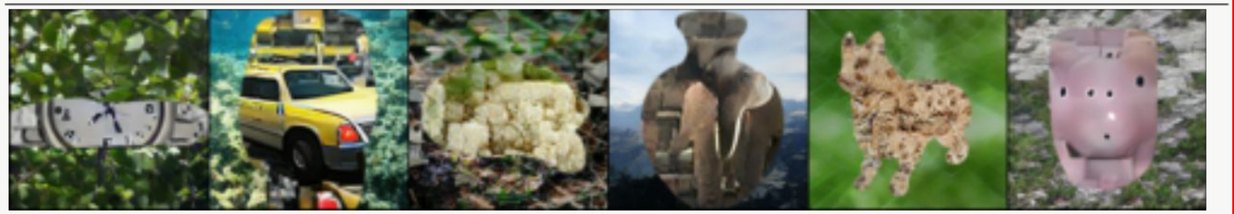
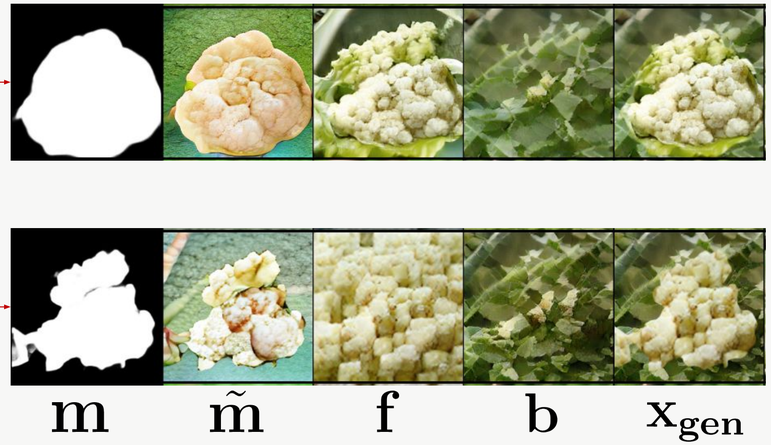


***Figure 3**. Reproduced qualitative results on ImageNet*

Claim 2: Inductive Bias Requirements (IBR)

$\mathcal{L}_{shape}$	$\mathcal{L}_{text}$	$\mathcal{L}_{bg}$	$\mathcal{L}_{rec}$	IS ⇑	μ_{mask}
✗	✓	✓	✓	100.8 | 85.9	0.3 | 0.2
✓	✗	✓	✓	186.5 | 198.4	0.4 | 0.9
✓	✓	✗	✓	200.9 | 195.6	0.1 | 0.1
✓	✓	✓	✗	19.3 | 38.4	0.4 | 0.3
✓	✓	✓	✓	156.1 | 130.2	0.3 | 0.3
BigGAN (Upper Bound)				202.9	-

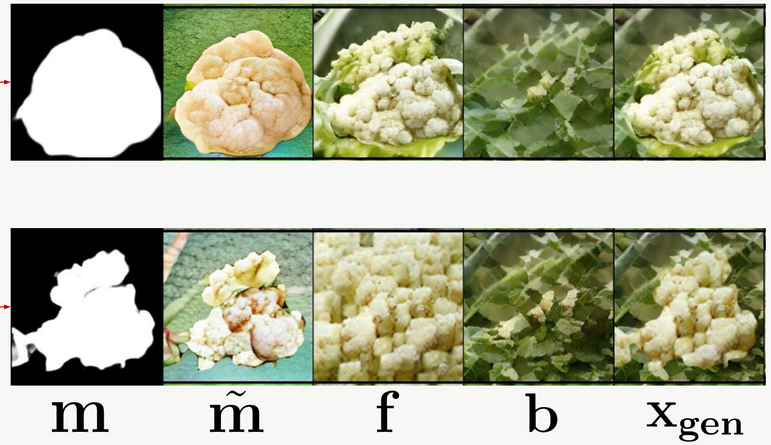
***Table 1**. Reproduced loss ablation study.*



Claim 2: Inductive Bias Requirements (IBR)

$\mathcal{L}_{shape}$	$\mathcal{L}_{text}$	$\mathcal{L}_{bg}$	$\mathcal{L}_{rec}$	IS ⇑	μ_{mask}
✗	✓	✓	✓	100.8 | 85.9	0.3 | 0.2
✓	✗	✓	✓	186.5 | 198.4	0.4 | 0.9
✓	✓	✗	✓	200.9 | 195.6	0.1 | 0.1
✓	✓	✓	✗	19.3 | 38.4	0.4 | 0.3
✓	✓	✓	✓	156.1 | 130.2	0.3 | 0.3
BigGAN (Upper Bound)				202.9	-

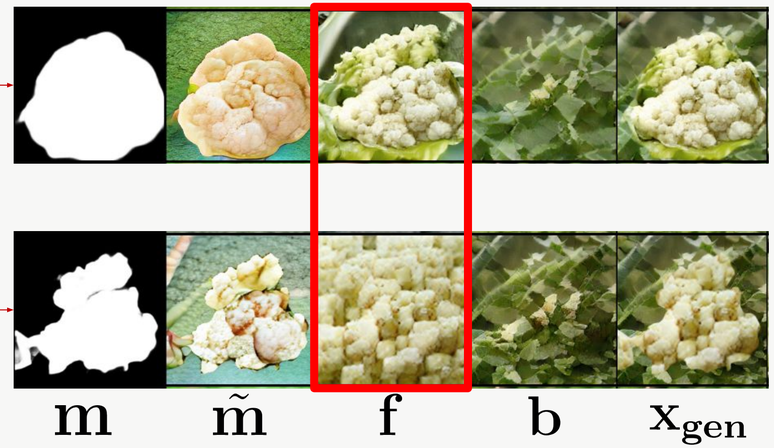
***Table 1**. Reproduced loss ablation study.*



Claim 2: Inductive Bias Requirements (IBR)

$\mathcal{L}_{shape}$	$\mathcal{L}_{text}$	$\mathcal{L}_{bg}$	$\mathcal{L}_{rec}$	IS ⇑	μ_{mask}
✗	✓	✓	✓	100.8 | 85.9	0.3 | 0.2
✓	✗	✓	✓	186.5 | 198.4	0.4 | 0.9
✓	✓	✗	✓	200.9 | 195.6	0.1 | 0.1
✓	✓	✓	✗	19.3 | 38.4	0.4 | 0.3
✓	✓	✓	✓	156.1 | 130.2	0.3 | 0.3
BigGAN (Upper Bound)				202.9	-

***Table 1**. Reproduced loss ablation study.*



Claim 3: Out-of-Distribution Robustness (ODR)

***Table 2**. Reproduced qualitative results on MNIST variants.*

Setting	C-MNIST		DC-MNIST		W-MNIST	
	Train ⇑	Test ⇑	Train ⇑	Test ⇑	Train ⇑	Test ⇑
Original	99.7 | 99.5	37.6 | 35.9	100 | 100	10.5 | 10.3	100 | 100	10.8 | 10.1
GAN	99.6 | 99.8	32.5 | 40.7	100 | 100	10.6 | 10.8	99.9 | 100	11.2 | 10.4
CGN	99.4 | 99.7	92.3 | 95.1	94.8 | 97.4	86.5 | 89.0	95.5 | 99.2	81.4 | 85.7
O + GAN	99.6 | 99.8	41.5 | 40.7	100 | 100	10.0 | 10.8	100 | 100	11.1 | 10.4
O + CGN	99.2 | 99.7	95.9 | 95.1	96.9 | 97.4	85.5 | 89.0	96.8 | 99.2	62.8 | 85.7

***Table 3**.Shape biases of independent classifiers*

Trained on	Shape Bias	top-1 ⇑	top-5 ⇑
IN + GCN/Shape	54.8	74.0	91.7
IN + GCN/Text	16.7		
IN + GCN/Bg	22.9		
IN-mini + GCN/Shape	58.8	56.5	79.3
IN-mini + GCN/Text	22.6		
IN-mini + GCN/Bg	24.7		

***Table 4**. Evaluation of robustness against adversarially chosen backgrounds*

Trained on	IN-9 ⇑	Mixed-Same ⇑	Mixed-Rand ⇑	BG-Gap ⇓
IN	95.6	86.2	78.9	7.3
SIN	89.2	73.1	63.7	9.4
IN + SIN	94.7	85.9	78.5	7.4
Mixed-Rand	73.3	71.5	71.3	0.2
IN + CGN	94.2	83.4	80.1	3.3
IN-mini + CGN	89.4	75.4	66.7	5.0

Claim 3: Out-of-Distribution Robustness (ODR)

***Table 2**. Reproduced qualitative results on MNIST variants.*

Setting	C-MNIST		DC-MNIST		W-MNIST	
	Train ⇑	Test ⇑	Train ⇑	Test ⇑	Train ⇑	Test ⇑
Original	99.7 | 99.5	37.6 | 35.9	100 | 100	10.5 | 10.3	100 | 100	10.8 | 10.1
GAN	99.6 | [illegible]	32.5 | 40.7	100 | 100	10.6 | 10.8	[illegible] | 100	11.2 | 10.4
CGN	99.4 | 99.7	[illegible] | 95.1	94.8 | 97.4	86.5 | [illegible]	95.5 | 99.2	81.4 | 85.7
O + GAN	99.6 | 99.8	41.5 | 40.7	100 | 100	10.0 | 10.8	100 | 100	11.1 | 10.4
O + CGN	99.2 | 99.7	95.9 | 95.1	96.[illegible] | [illegible]7.4	85.5 | 89.0	96.8 | 99.2	62.8 | 85.7

***Table 3**.Shape biases of independent classifiers*

Trained on	Shape Bias	top-1 ⇑	top-5 ⇑
IN + GCN/Shape	[illegible]		
IN + GCN/Text	16.7	74.0	91.7
IN + GCN/Bg	22.9		
IN-mini + GCN/Shape	58.8		
IN-mini + GCN/Text	22.6	56.5	79.3
IN-mini + GCN/Bg	24.7		

***Table 4**. Evaluation of robustness against adversarially chosen backgrounds*

Trained on	IN-9 ⇑	Mixed-Same ⇑	Mixed-Rand ⇑	BG-Gap ⇓
IN	95.6	86.2	78.9	7.3
SIN	89.2	73.1	63.7	9.4
IN + SIN	94.7	85.9	78.5	7.4
Mixed-Rand	73.3	71.5	71.3	0.2
IN + CGN	94.2	83.4	80.1	3.3
IN-mini + CGN	89.4	75.4	66.7	5.0

Results beyond original paper

Explainability analysis: Visualizing features

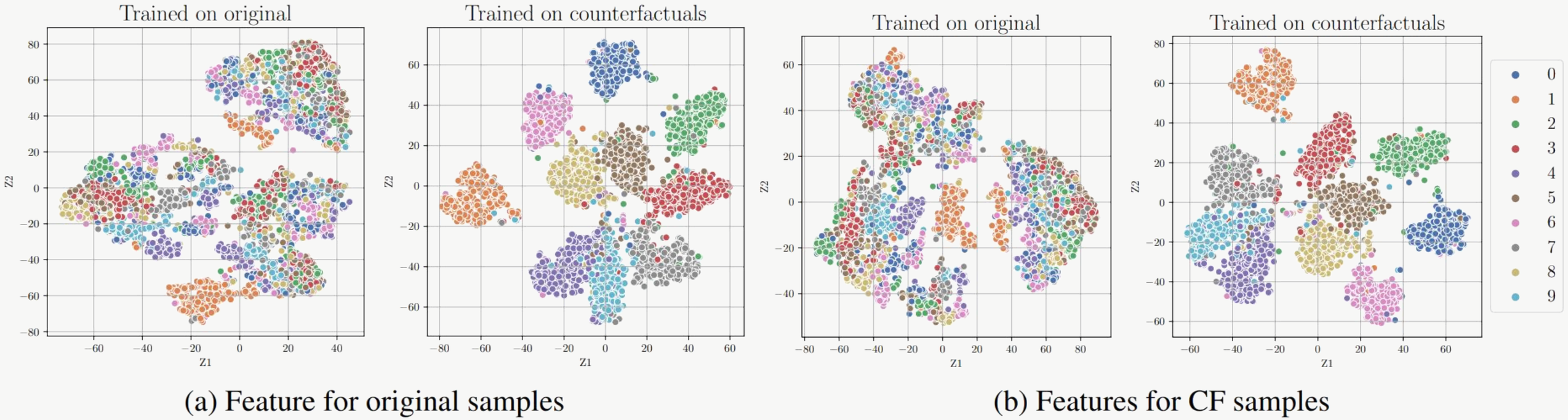


Figure 4. *Feature space visualization of a CNN classifier trained on on colored MNIST variants*

Explainability analysis: Visualizing features

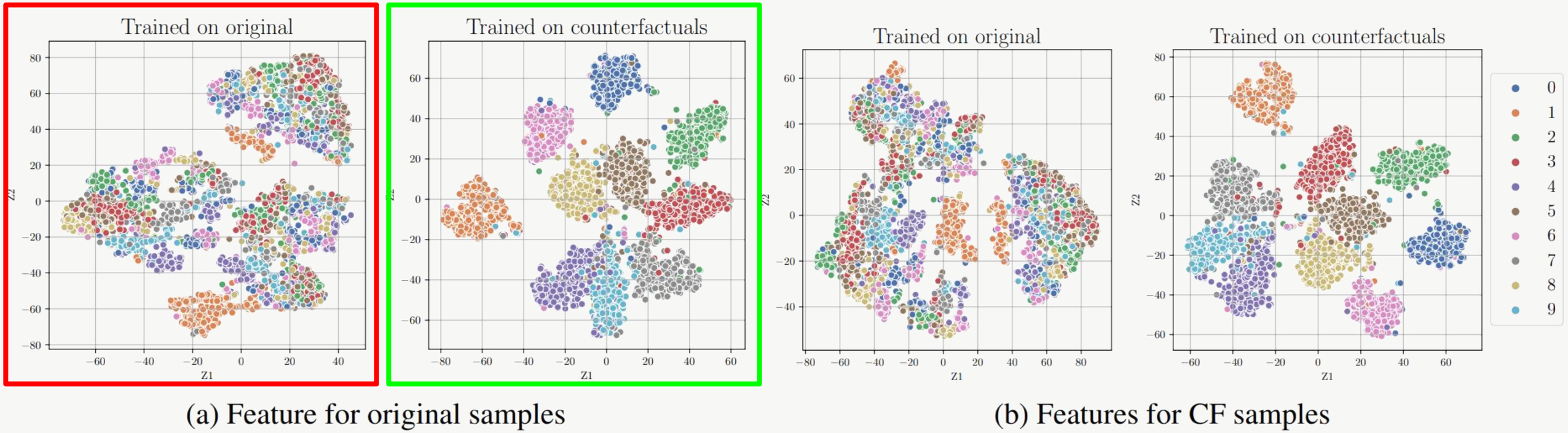


Figure 4. *Feature space visualization of a CNN classifier trained on on colored MNIST variants*

Explainability analysis: Visualizing features

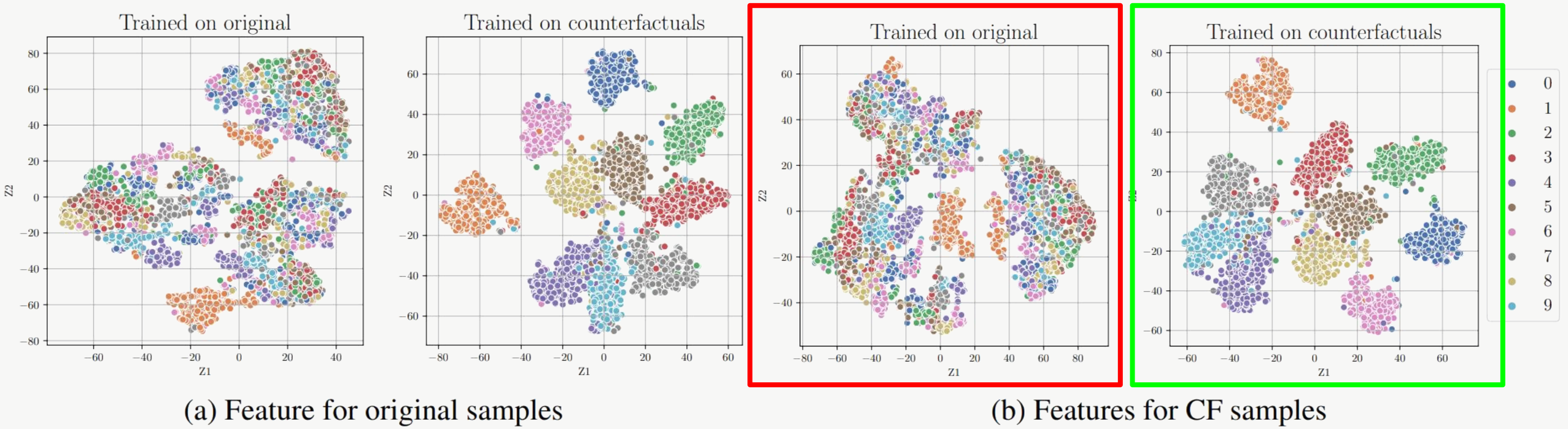


Figure 4. *Feature space visualization of a CNN classifier trained on on colored MNIST variants*

Explainability analysis: Visualizing features

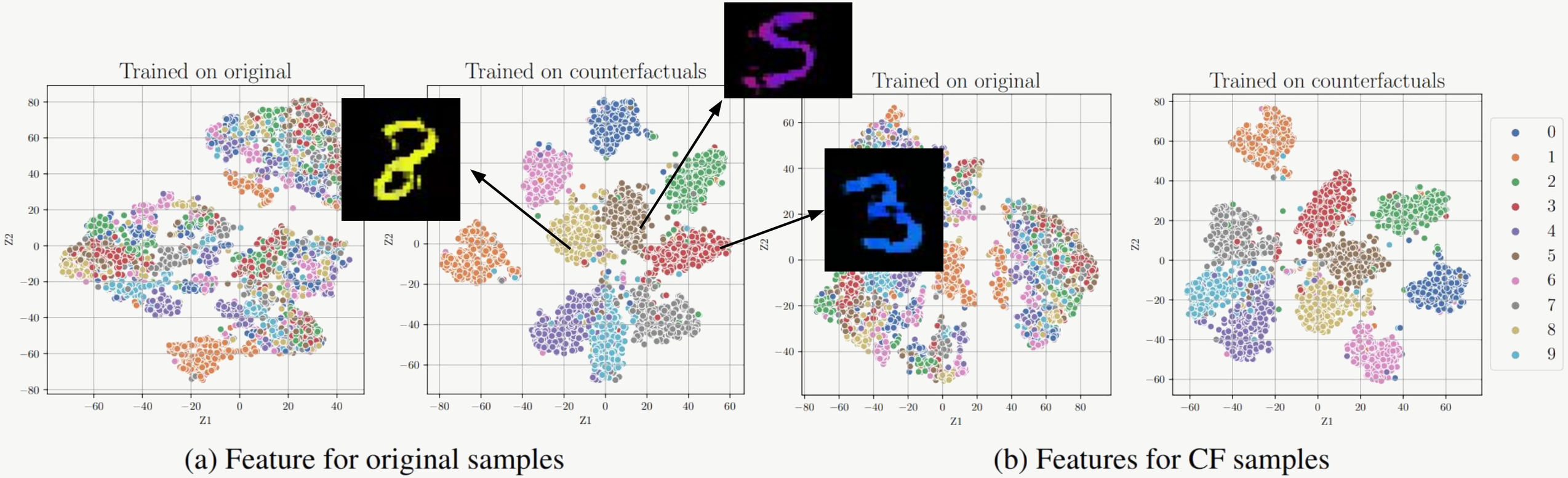


Figure 4. *Feature space visualization of a CNN classifier trained on on colored MNIST variants*

Explainability analysis: What does the model focus on?

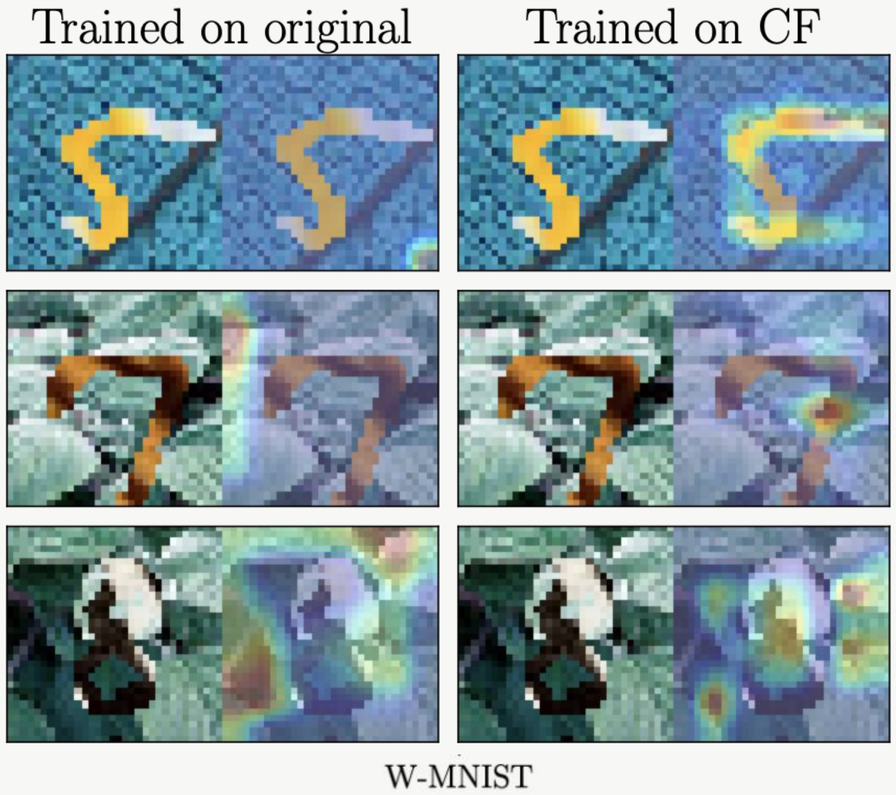


***Figure 5**. GradCAM heatmap visualized on W-MNIST samples*

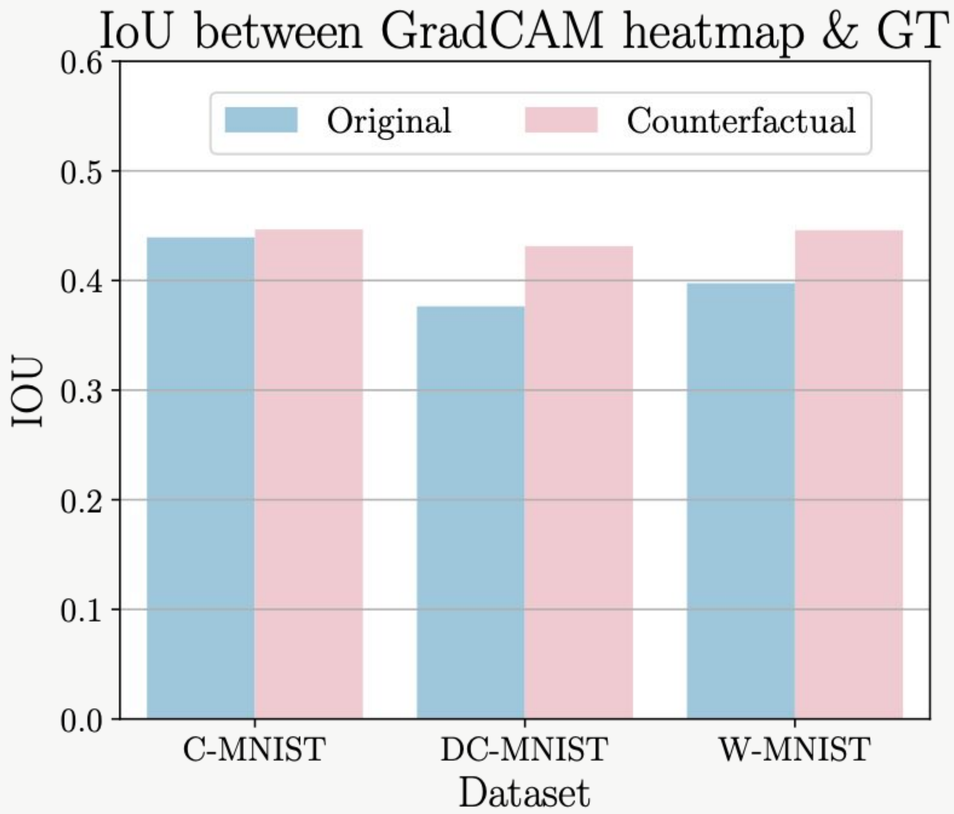


***Figure 6**. Metric to quantify areas where the model focuses on*

[Demo for ImageNet]

Explainability analysis: What does the model focus on?

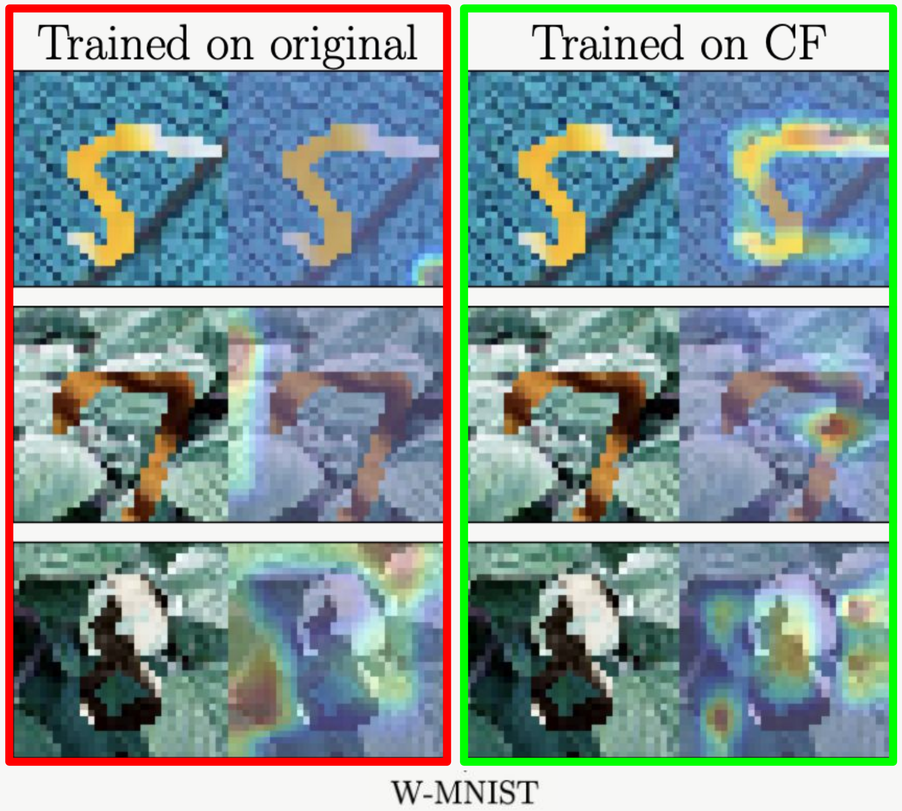


***Figure 5**. GradCAM heatmap visualized on W-MNIST samples*

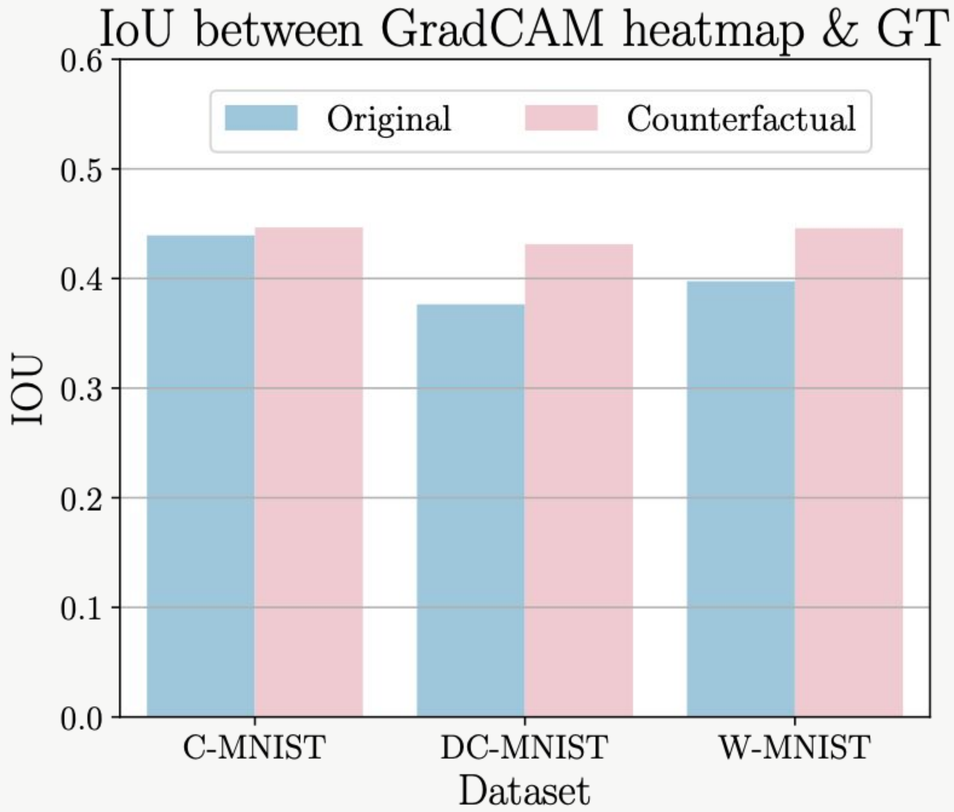


***Figure 6**. Metric to quantify areas where the model focuses on*

[Demo for ImageNet]

Explainability analysis: What does the model focus on?

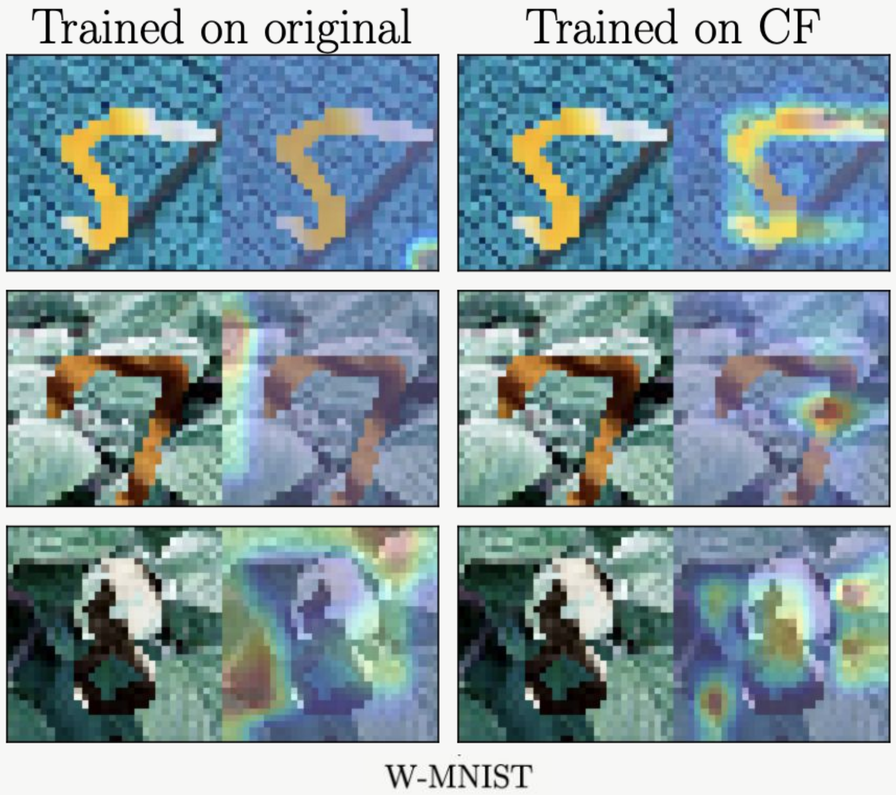


***Figure 5**. GradCAM heatmap visualized on W-MNIST samples*

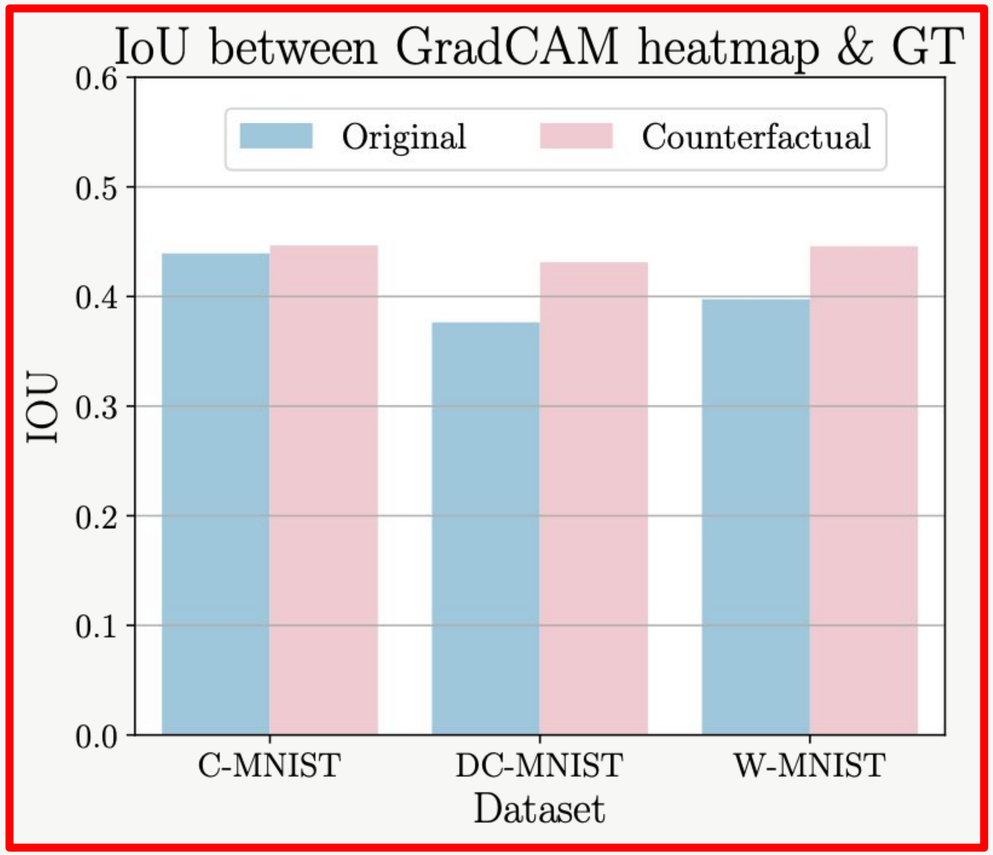


***Figure 6**. Metric to quantify areas where the model focuses on*

[Demo for ImageNet]

Our experience & Lessons Learned

Fairness, Research & Management

Great overview of topics such as fairness, accountability and AI ethics in general!

Great (first) research experience unlike other course assignments!

Managing time/workload under a deadline

Collaborating in a group over a research project

Tips & Suggestions: Research

Several aspects to research (specific to reproducibility)

- Read and understand thoroughly
- Identify key contributions of the paper
- Identify key experiments that support these
- Identify drawbacks and possible extensions
- Coding
- Experimenting
- Writing and presenting
- Submitting
- …

Ask the right questions!

Qualitative analysis generally helps to get a nice intuition beyond numbers!

Look at reproducibility papers of previous years for inspiration and structure!

Do not fixate on reproducing the exact numbers. Looks for matching trends!

Tips & Suggestions: Management

It helps to appoint a lead for each broad vertical – but the lead should not do everything!

Communication is the key!

- Setup a chat for real-time comms (Discord/WhatsApp)
- It helps to meet regularly (daily - albeit for 15 mins)

Start writing early and not just a day before the deadline! We started in week 1.

Learn from each other!

Q&A

Authors: Piyush Bagad, Danilo de Goede, Paul Hilders, Jesse Maas

Supervisor: Christos Athanasiadis