Replication: Contrastive Learning and Data Augmentation in Traffic Classification Using a Flowpic Input Representation

A. Finamore, C. Wang, J. Krolikowski, J. M. Navarro, F. Cheng, D. Rossi
Huawei Technologies SASU, France

ACM Internet Measurement Conference (IMC)
Montreal, 24-26 Oct, 2023

HUAWEI

If I had three wishes for the genie

from my TMA22 keynote

1 More “code challenges”
They can be occasion to release data and put focus on specific problems

2 Create one permanent replicability track/workshop
- *Decouple study state-of-the-art from promoting new ideas*
- *Foster data/code sharing for the benefit of the community*

ACM IMC 2023 MONTRÉAL October 24 - 26

PAM 2023

PAM 2024

3 Federate universities/research centers for data access/sharing
Break the barrier of 1-to-1 cooperations

The data divide affects the whole measurements community
AI-driven measurement methods is just exacerbating it

Lot of literature for a REPLICABILITY study on TRAFFIC CLASSIFICATION to choose from

...But

Replicability Track:

IMC 2023 will trial a new Replicability Track for submissions that aim to reproduce or replicate results that have been previously published at IMC. Papers accepted to this track will be published in ACM SIGCOMM Computer Communication Review (CCR). Priority will be given to replicability studies, although reproducibility studies are also in scope. For the definitions, please see **ACM's site**. The authors of outstanding replicability papers may receive an invitation to present at the main conference. In that case, the paper would also be included in IMC's proceedings (rather than CCR).

Short paper @ IMC'22

Check for updates

A Few Shots Traffic Classification with mini-FlowPic Augmentations

Eyal Horowicz
eyalhorowicz@mail.tau.ac.il
Tel Aviv University
Israel

Tal Shapira*
talshapirala@gmail.com
Reichman University
Israel

Yuval Shavitt
shavitt@eng.tau.ac.il
Tel Aviv University
Israel

HUAWEI

Outline

1. Introduce the IMC22 paper and set our goals

2. Datasets and methodology

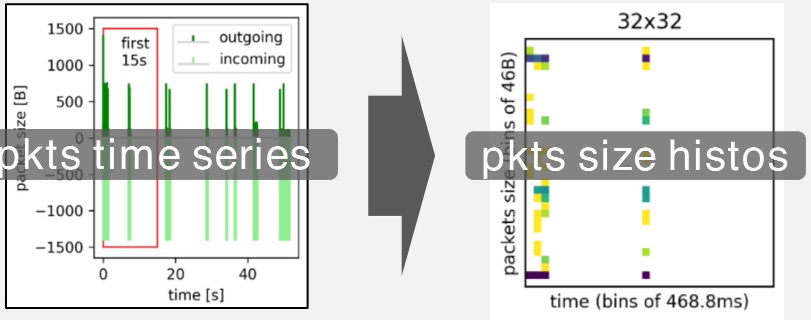
3. Results

4. Closing remarks

Outline

1. Introduce the IMC22 paper and set our goals
2. Datasets and methodology
3. Results
4. Closing remarks

IMC22 paper : TLDR (1/2)

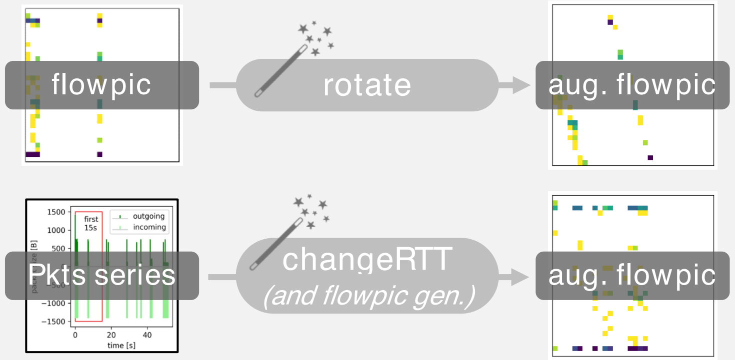


pkts time series → pkts size histos

1 Flowpic input representation

Image transformation: flowpic → rotate → aug. flowpic

Time series transformation: Pkts series → changeRTT *(and flowpic gen.)* → aug. flowpic



2 Data augmentation

3 Few-shot learning

Few samples (100 per class): Original trainset

transform

10x trainset size: Augmented trainset

→ Supervised training

4 Self-supervision

Original trainset → Contrastive learning training *(via augmentations)* → Supervised training

Few samples (10 per class)

Unsupervised Pretraining

Supervised Finetune

[1] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

IMC22 paper : TLDR (2/2)

Evaluation settings

- UCDAVIS-19 dataset [1]
  5 QUIC-based Google services
- Benchmarking flowpic computed from 15sec of traffic at different resolutions
  (32x32 → 1500x1500)
- 6 augmentations
  3 image-based, 3 time series-based
- 100 samples per class augmented 10 times
- Contrastive learning via SimCLR [2] and finetune with 10 labeled samples

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

IMC22 paper : TLDR (2/2)

Evaluation settings

- UCDAVIS-19 dataset [1]
  5 QUIC-based Google services
- Benchmarking flowpic computed from 15sec of traffic at different resolutions
  (32x32 → 1500x1500)
- 6 augmentations
  3 image-based, 3 time series-based
- 100 samples per class augmented 10 times
- Contrastive learning via SimCLR [2] and finetune with 10 labeled samples

Takeaways

- Time series transformations are superior wrt image transformations
- 100 labeled samples and a 32x32 flowpic are enough for good accuracy
- SimCLR performance almost on par with supervised training

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

Our goals

G0 **ML reference baseline** NEW
How complex is the problem? Do we really need DL?

G1 Reproduce IMC22 **augmentations benchmark** in supervised setting
+ statistical analysis to compare augmentations NEW

G2 Reproduce IMC22 **contrastive learning benchmark**
+ considering more scenarios NEW

G3 Replicate G1 with **3 alternative datasets** NEW

G4 Treat our paper a **“software deliverable”**
Contribute curated artifacts NEW

Outline

1. Introduce the IMC22 paper and set our goals

2. **Datasets and methodology**

3. Results

4. Closing remarks

HUAWEI

Datasets

Class imbalance

$$\rho = \frac{Max\,(flows\ per\ class)}{Min\,(flows\ per\ class)}$$

Name	Partition	Filter	Classes	Flows *All*	Flows *Min* (per class)	Flows *Max* (per class)	Flows *ρ* (class imbal.)	Pkts *Mean* (per flow)
UCDAVIS-19 [1]	Pretraining	*none*	5	6,439	592	1,915	3.2	6,653
	Human			83	15	20	1.3	7,666
	Script			150	30	30	1.0	7.131
MIRAGE-19 [2]	n.a.	*none*	20	122,007	1,986	11,737	5.9	23
		>10pkts		64,172	1,013	7,505	7.4	17
MIRAGE-22 [2]	n.a.	*none*	9	59,071	2,252	18,882	8.4	3,068
		>10pkts		26,773	970	4,437	4.6	6,598
		>1,000pkts		4,569	190	2,220	11.7	38,321
UTMOBILENET-21 [4]	4-into-1	*none*	17	34,378	159	5,591	35.2	664
		>10pkts	14	9,460	130	2,246	19.2	2,366

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] The MIRAGE project: https://traffic.comics.unina.it/mirage/
[3] UTMobileNetTraffic2021: A Labeled Public Network Traffic Dataset, IEEE Networking letters

Datasets

Class imbalance

$$\rho = \frac{Max\ (flows\ per\ class)}{Min\ (flows\ per\ class)}$$

Name	Partition	Filter	Classes	Flows All	Flows Min (per class)	Flows Max (per class)	Flows ρ (class imbal.)	Pkts Mean (per flow)
UCDAVIS-19 [1]	Pretraining			6,439	592	1,915	3.2	6,653
	Human	none	5	83	15			7,666
	Script			150	30	30	1.0	7.131
MIRAGE-19 [2]	n.a.	none	20	122,007	1,986	11,737	5.9	23
		>10pkts		64,172	1,013	7,505	7.4	17
MIRAGE-22 [2]	n.a.	none	9	59,071	2,252	18,882	8.4	3,068
		>10pkts		26,773	970	4,437	4.6	6,598
		>1,000pkts		4,569	190	2,220	11.7	38,321
UTMOBILENET-21 [4]	4-into-1	none	17	34,378	159	5,591	35.2	664
		>10pkts	14	9,460	130	2,246	19.2	2,366

Google Doc
Google Music
Google Drive
Google Search
YouTube

Very long flows

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] The MIRAGE project: https://traffic.comics.unina.it/mirage/
[3] UTMobileNetTraffic2021: A Labeled Public Network Traffic Dataset, IEEE Networking letters

Datasets

Class imbalance

$$\rho = \frac{Max\ (flows\ per\ class)}{Min\ (flows\ per\ class)}$$

Name	Partition	Filter	Classes	Flows All	Flows Min (per class)	Flows Max (per class)	Flows ρ (class imbal.)	Pkts Mean (per flow)
UCDAVIS-19 [1]	Pretraining	Large training set	5	6,439	592	1,915	3.2	6,653
	Human	Small testing sets		83	15	20	1.3	Light imbalance
	Script			150	30	30	1.0	7,131
MIRAGE-19 [2]	n.a.	*none*	20	122,007	1,986	11,737	5.9	23
		>10pkts		64,172	1,013	7,505	7.4	17
MIRAGE-22 [2]	n.a.	*none*	9	59,071	2,252	18,882	8.4	3,068
		>10pkts		26,773	970	4,437	4.6	6,598
		>1,000pkts		4,569	190	2,220	11.7	38,321
UTMOBILENET-21 [4]	4-into-1	*none*	17	34,378	159	5,591	35.2	664
		>10pkts	14	9,460	130	2,246	19.2	2,366

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] The MIRAGE project: https://traffic.comics.unina.it/mirage/
[3] UTMobileNetTraffic2021: A Labeled Public Network Traffic Dataset, IEEE Networking letters

HUAWEI

Datasets

Class imbalance

$$\rho = \frac{Max\ (flows\ per\ class)}{Min\ (flows\ per\ class)}$$

Name	Partition	Filter	Classes	Flows *All*	Flows *Min* (per class)	Flows *Max* (per class)	Flows *ρ* (class imbal.)	Pkts *Mean* (per flow)
UCDAVIS-19 [1]	Pretraining	*none*	5	6,439	592	1,915	3.2	6,653
	Human			83	15	20	1.3	7,666
	Script			150	30	30	1.0	7.131
MIRAGE-19 [2]	n.a.	*non*	Variety of Android apps 20	122,007	← Many flows [illegible]	[illegible]37	…but short [illegible]	→ 23
		>10p		64,172	1,013	7,505	7.4	17
MIRAGE-22 [2]	n.a.	*none*	Only video Meeting apps 9	59,071	2,252	18,882	8.4	3,068
		>10p		26,773	970	4,437	4.6	6,598
		>1,000pkts		4,569	190	2,220	11.7	38,321
UTMOBILENET-21 [4]	4-into-1	*non*	In between MIRAGE datasets 17	34,378	159	5,591	35.2	664
		>10p	14	9,460	130	2,246	19.2	2,366

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19

[2] The MIRAGE project: https://traffic.comics.unina.it/mirage/

[3] UTMobileNetTraffic2021: A Labeled Public Network Traffic Dataset, IEEE Networking letters

Datasets

Class imbalance

$$\rho = \frac{Max\,(flows\ per\ class)}{Min\,(flows\ per\ class)}$$

Name	**Partition**	**Filter**	**Classes**	**Flows** *All*	*Min* (per class)	*Max* (per class)	*ρ* (class imbal.)	**Pkts** *Mean* (per flow)
UCDAVIS-19 [1]	Pretraining			6,439	592	1,915	3.2	6,653
	Human	*none*	5	83	15	20	1.3	7,666
	Script			150	30	30	1.0	7.131
MIRAGE-19 [2]	n.a	*none*	20	122,007	1,986	11,737	5.9	23
		>10pkts		64,172	1,013	7,505	7.4	17
MIRAGE-22 [2]	n.a	*none*	9	59,071	2,252	18,882		3,068
		>10pkts		26,773	970	4,437	4.6	6,598
		>1,000pkts		4,569	190	2,220	11.7	38,321
UTMOBILENET-21 [4]	4-into-	*none*	17	34,378	159	5,591	35.2	664
		>10pkts	14	9,460	130	2,246	19.2	2,366

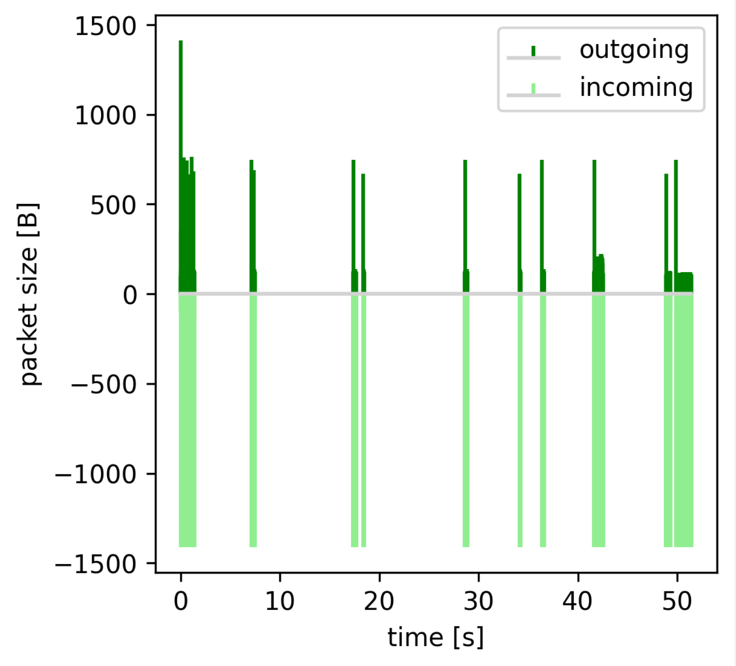
Data curation

Larger imbalance

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19
[2] The MIRAGE project: https://traffic.comics.unina.it/mirage/
[3] UTMobileNetTraffic2021: A Labeled Public Network Traffic Dataset, IEEE Networking letters

Flowpics: *computation*

1 Get pkts time series

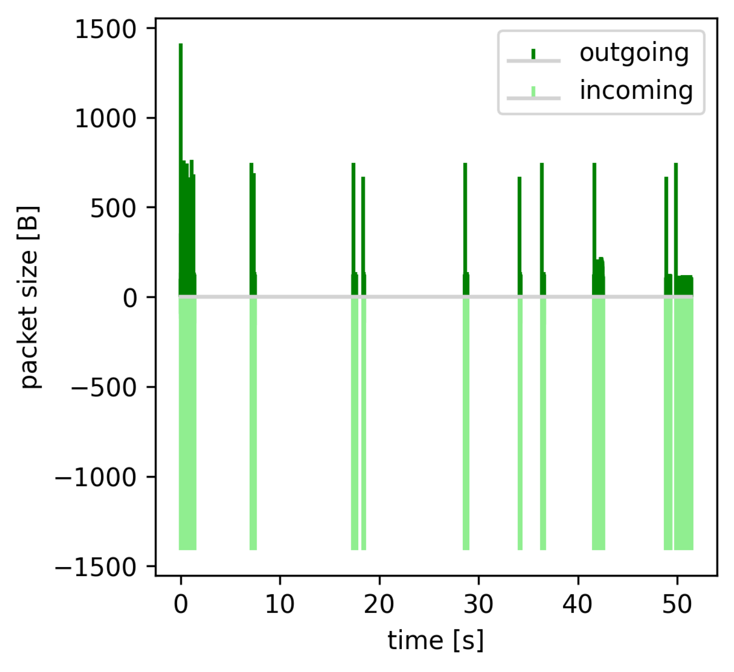


Example of a YouTube flow

HUAWEI

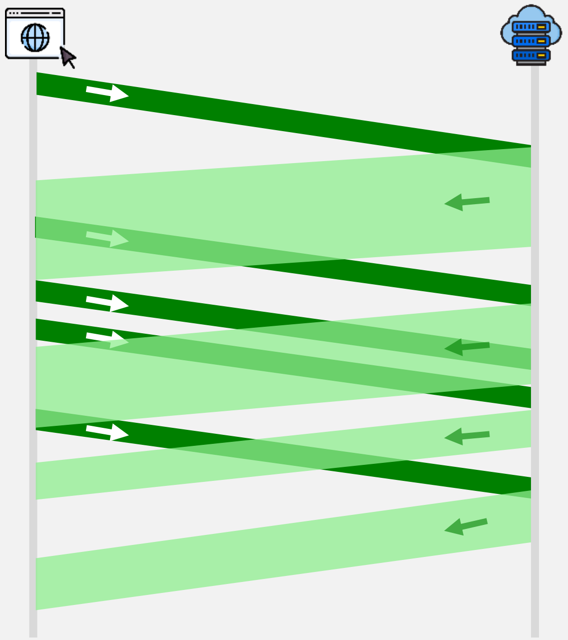
Flowpics: *computation*

1 Get pkts time series



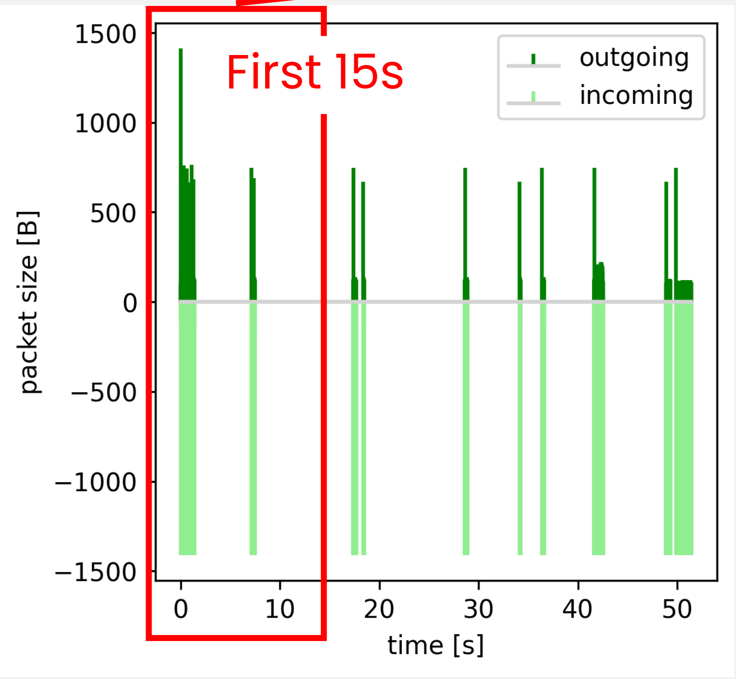
Example of a YouTube flow

2 Pkts size histograms



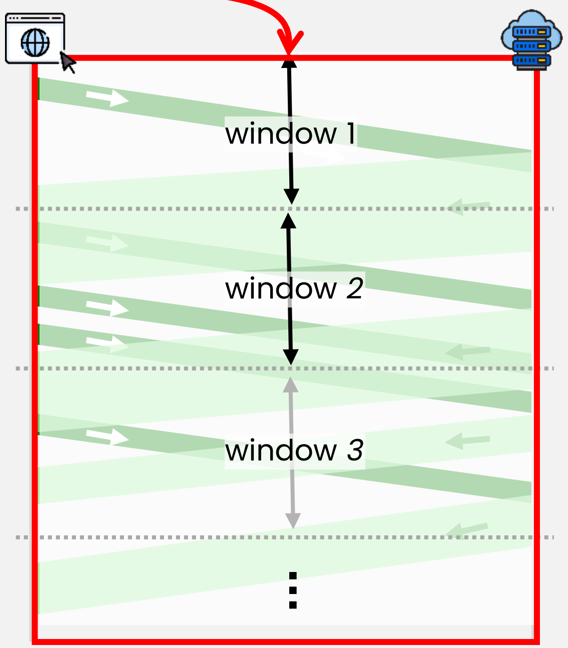
Flowpics: *computation*

1 Get pkts time series



Example of a YouTube flow

2 Pkts size histograms



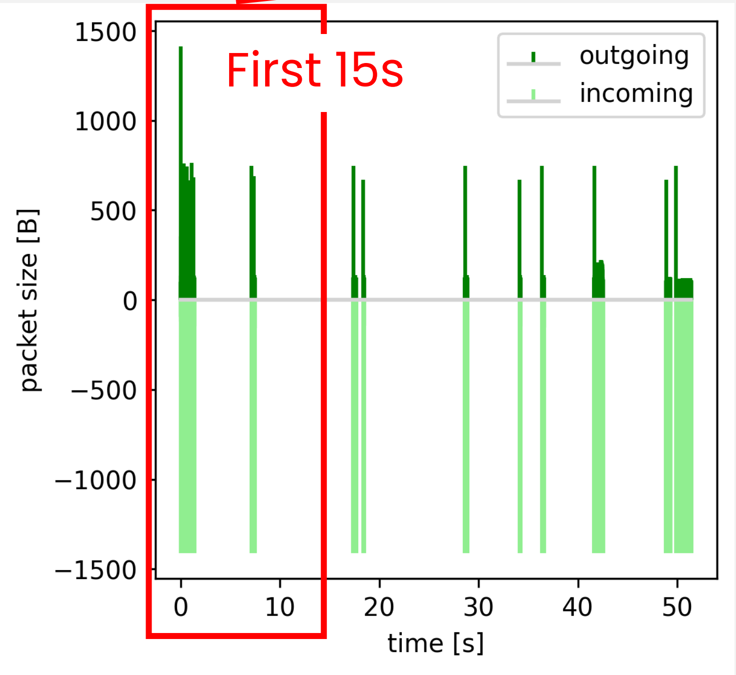
For 32x32 resolution

Window size of 15s/32 = *468ms*

HUAWEI

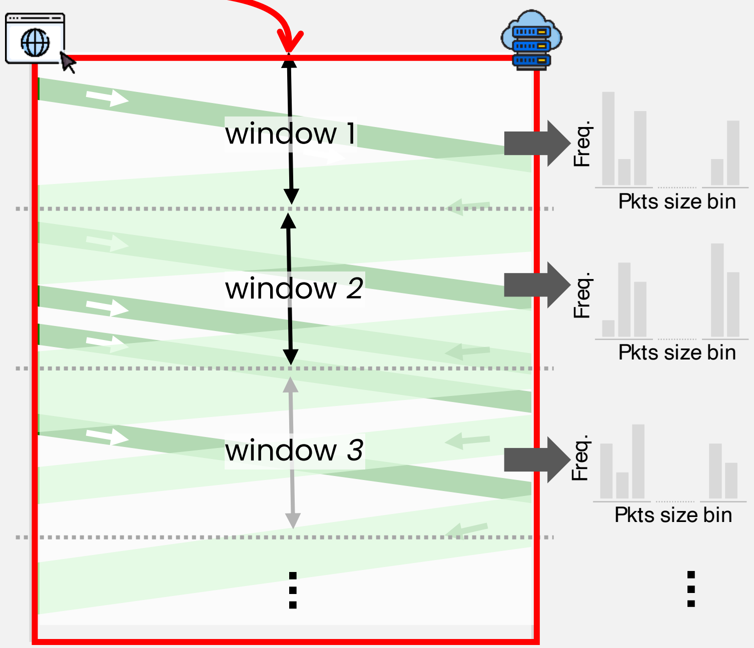
Flowpics: *computation*

1 Get pkts time series



Example of a YouTube flow

2 Pkts size histograms



For 32x32 resolution

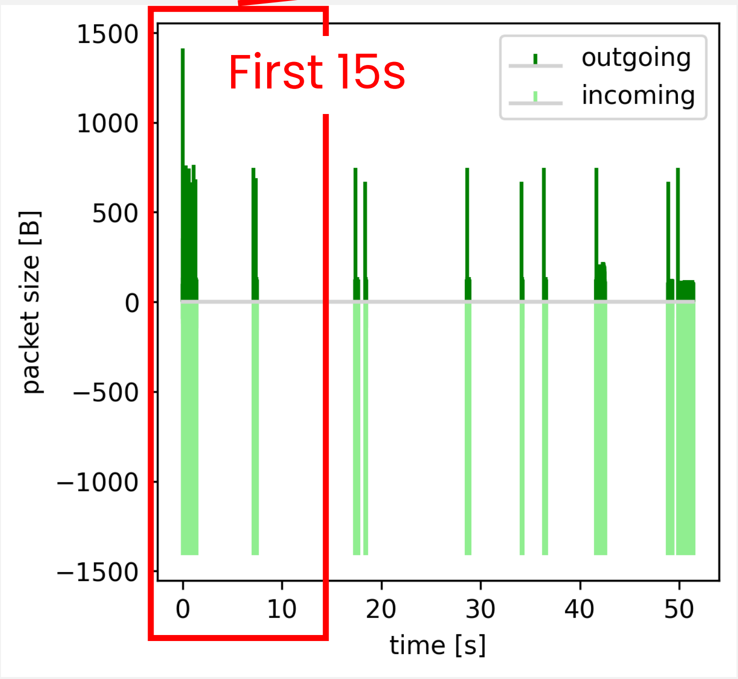
Window size of 15s/32 = *468ms*

Packets bin or ceil(1500/32) = *47B*

HUAWEI

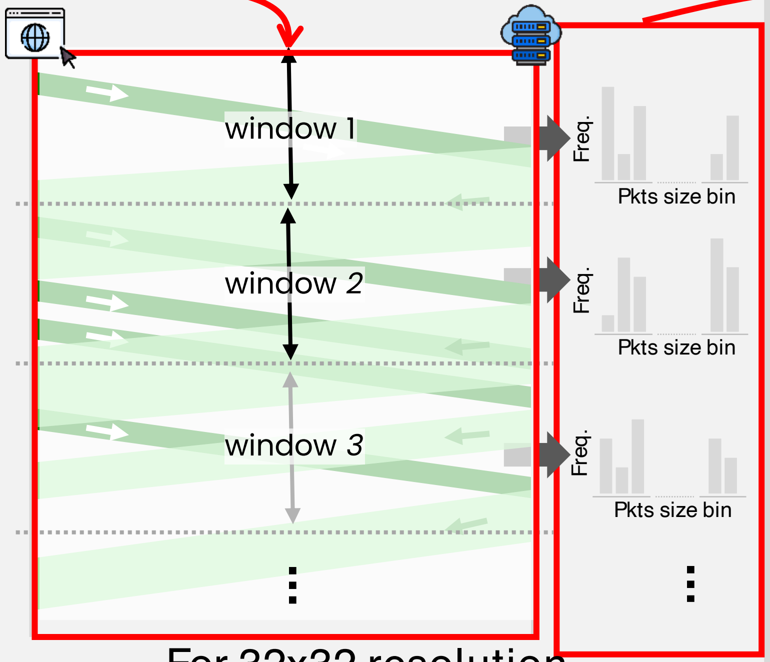
Flowpics: *computation*

1 Get pkts time series



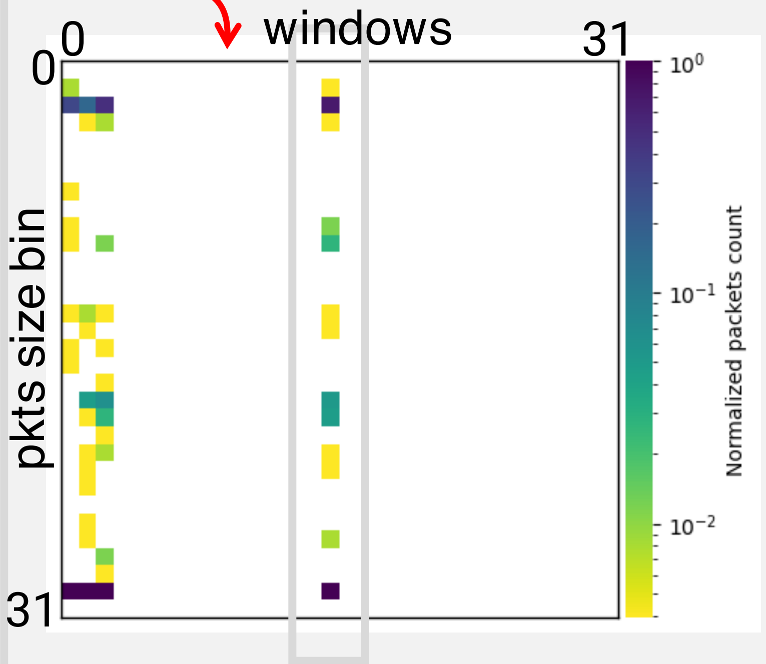
Example of a YouTube flow

2 Pkts size histograms



For 32x32 resolution
Window size of 15s/32 = *468ms*
Packets bin or ceil(1500/32) = *47B*

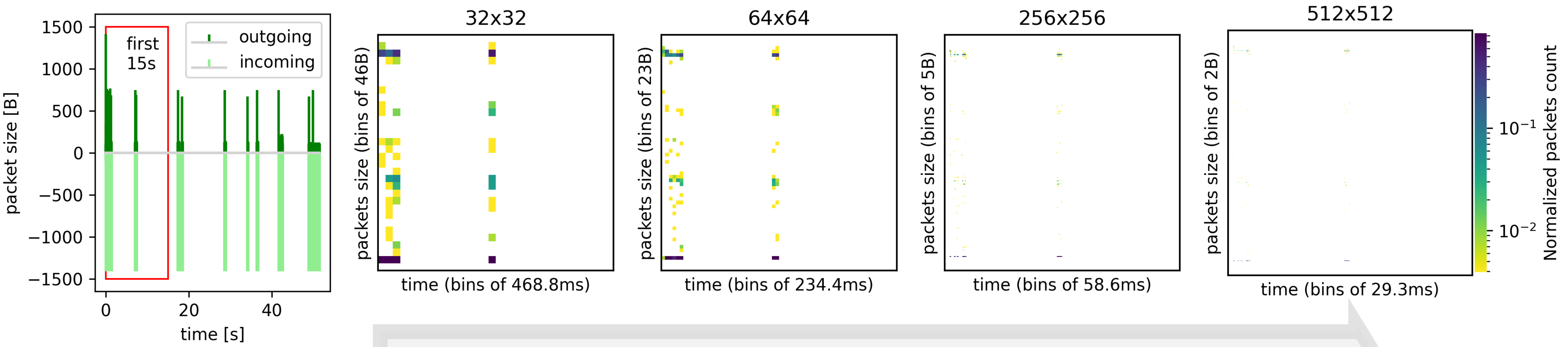
3 Stack histograms



Each column is a frequency histogram of a different window

HUAWEI

Flowpics: *resolution*



Sparsity proportional to image resolution

mini-flowpic

IMC22 paper contrasts ~~32x32~~ against 1500x1500

Experimental settings
Augmentations

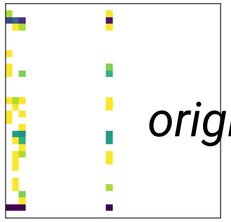
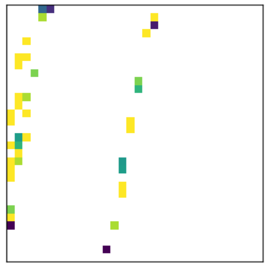
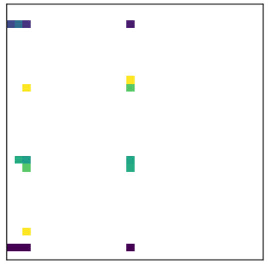
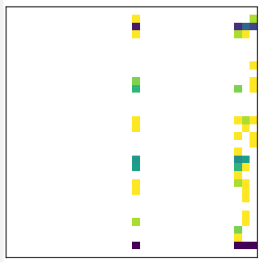

original

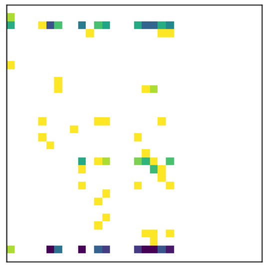
Image-based

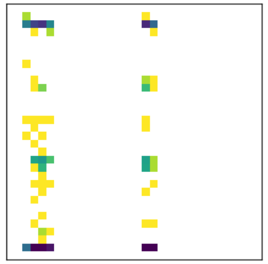
Rotate


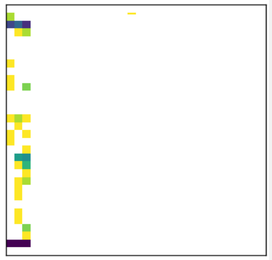
Color jitter


Horizontal flip


Time series-based

Change RTT

Multiply packets timestamp by a random factor

Time shift

Add a random factor to Packets timestamp

Packet loss

Remove a window of packets

Experimental settings (1/3)

Dataset folds

UCDAVIS-19	pretraining	Script	Human

Experimental settings (1/3)

Dataset folds

UCDAVIS-19

pretraining

Script

Human

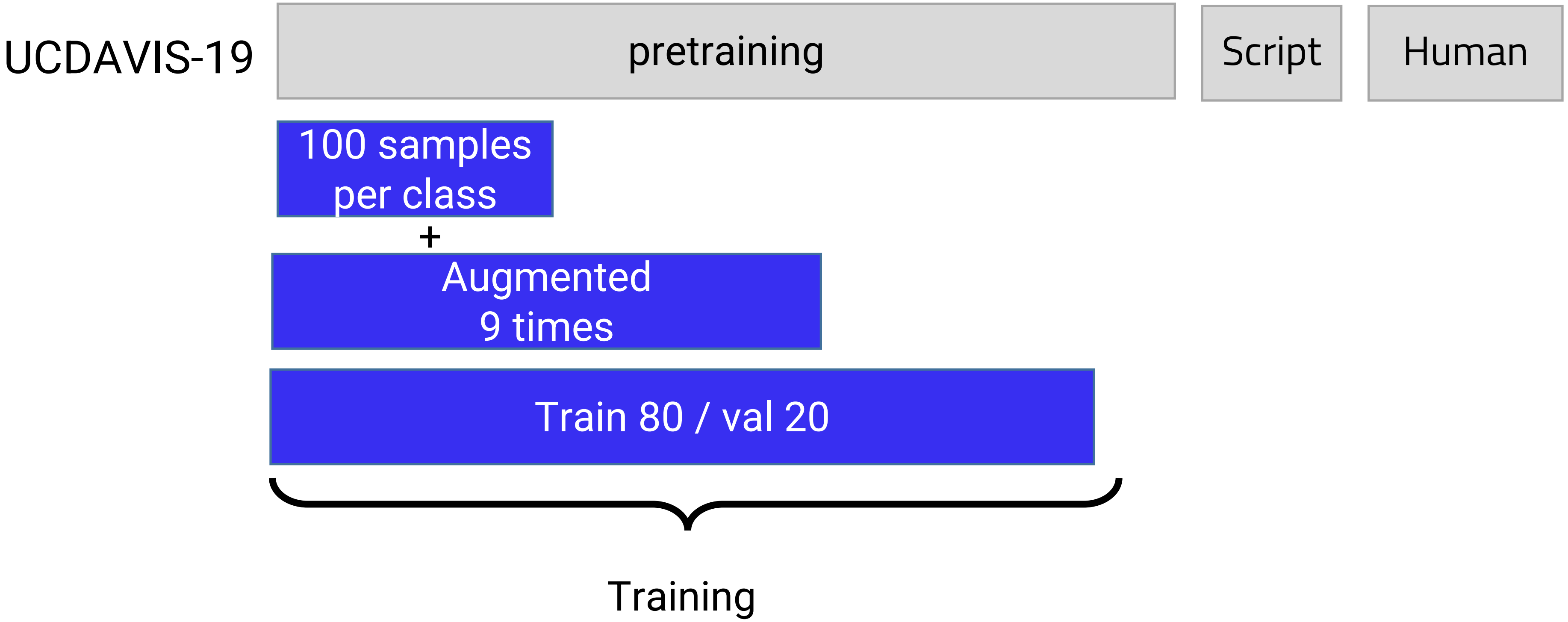
100 samples per class

+

Augmented 9 times

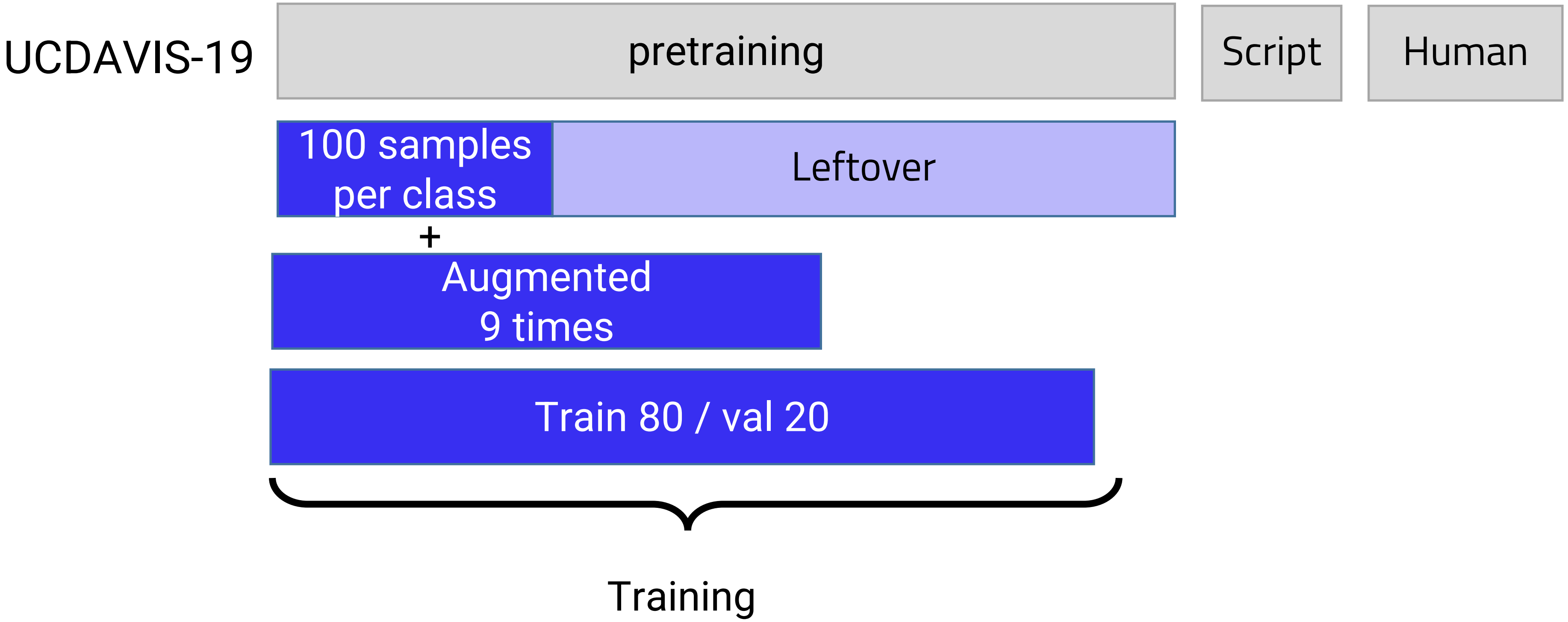
Experimental settings (1/3)

Dataset folds



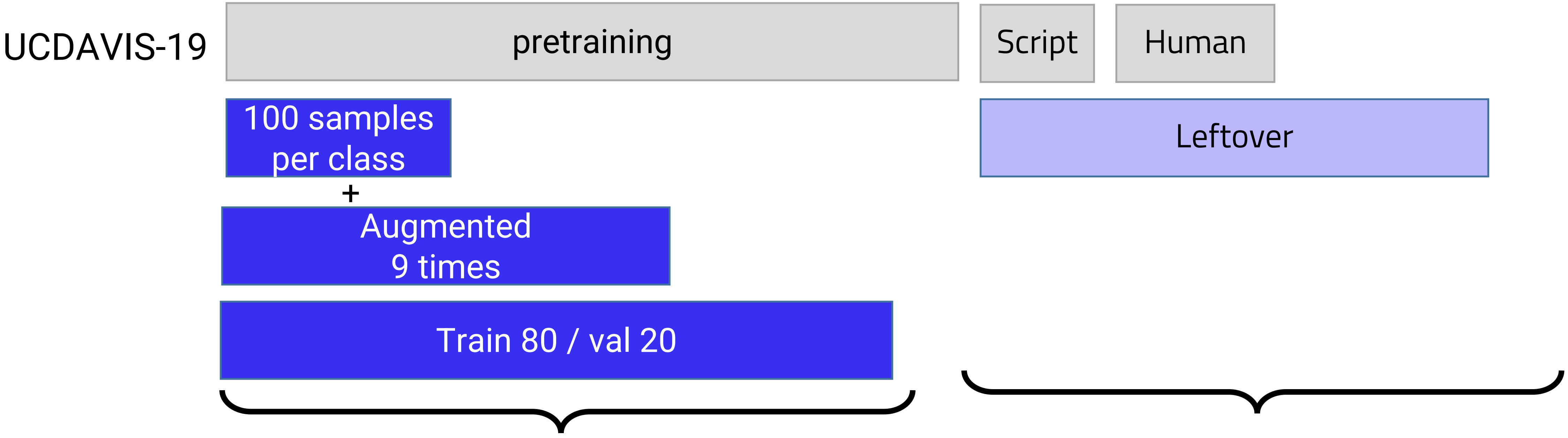
Experimental settings (1/3)

Dataset folds



Experimental settings (1/3)

Dataset folds



Training

Testing

Performance metric: accuracy

HUAWEI

Experimental settings (1/3)

Dataset folds

UCDAVIS-19

pretraining | Script | Human

100 samples per class
+
Augmented 9 times

Leftover

Train 80 / val 20

Training

Testing

Performance metric: accuracy

Other datasets

Train (80)	Val (10)	Test (10)

Performance metric: F1 score

Experimental settings (2/3)

Modeling framework and Artifacts

Created a framework to

- Trigger multiple modeling campaigns
- Fine-grained tracking of model training/inference performance
- Collect model artifacts
- Bind modeling to dataset splits

Experimental settings (2/3)

Modeling framework and Artifacts

Created a framework to

- Trigger multiple modeling campaigns
- Fine-grained tracking of model training/inference performance
- Collect model artifacts
- Bind modeling to dataset splits

Created 13 campaigns for a total of 2,760 experiments

Code artifacts

https://github.com/tcbenchstack/tcbench

Data artifacts

https://doi.org/10.6084/m9.figshare.c.6849252.v3

Documentation

https://tcbenchstack.github.io/tcbench/papers/imc23/

Experimental settings (3/3)
More from IMC22 *paper's authors*

The IMC22 paper has a github repo *https://github.com/eyalho/mini-flowpic-traffic-classification*

...but available code is not usable

- *Code only for* SimCLR *pretraining*
- *Network architectures and training are not the same as in the paper*
- *As is, the code is mixing training includes also testing samples*

We contacted IMC22 paper's authors mostly during camera ready

...but we received only short and delayed answers

Outline

1. Introduce the IMC22 paper and set our goals
2. Datasets and methodology
3. Results
    1. ML baseline
    2. Supervision
    3. Contrastive learning
4. Closing remarks

ML Baseline

G0 ML baseline

	Input (size)	Model	Paper	Accuracy 95th CI	
				Script	Human
	Flowpic (32x32)	CNN LeNet5	IMC22	98.67 *n.a.*	92.40 *n.a.*
(a)	Flowpic (32x32)	XGBoost	Ours	96.80±0.37	73.65±2.14
(b)	Time series (3x10)	XGBoost	Ours	94.53±0.56	66.91±1.40

(a) Flattened flowpic; (b) concat first 10 values of packet size, direction and inter arrival time
Our results are aggregation of 15 experiments (5 splits x 3 seeds)

G0

ML baseline

	Input (size)	Model	Paper	Accuracy 95th CI	
				Script	Human
	Flowpic (32x32)	CNN LeNet5	IMC22	98.67 *n.a.*	92.40 *n.a.*
(a)	Flowpic (32x32)	XGBoost	Ours	-1.87 96.80±0.37	73.65±2.14
(b)	Time series (3x10)	XGBoost	Ours	-4.14 94.53±0.56	66.91±1.40

(a) Flattened flowpic; (b) concat first 10 values of packet size, direction and inter arrival time
Our results are aggregation of 15 experiments (5 splits x 3 seeds)

- On **Script**, results on par with flowpic but lower performance with time series *(10 pkts -vs- 15s of traffic)*

HUAWEI

G0 ML baseline

Input (size)	Model	Paper	Accuracy 95th CI	
			Script	Human
Flowpic (32x32)	CNN LeNet5	IMC22	98.67 *n.a.*	92.40 *n.a.*
(a) Flowpic (32x32)	XGBoost	Ours	-1.87 96.80±0.37	-18.75 73.65±2.14
(b) Time series (3x10)	XGBoost	Ours	-4.14 94.53±0.56	-25.49 66.91±1.40

(a) Flattened flowpic; (b) concat first 10 values of packet size, direction and inter arrival time
Our results are aggregation of 15 experiments (5 splits x 3 seeds)

- On **Script**, results on par with flowpic but lower performance with time series *(10 pkts -vs- 15s of traffic)*
- On **Human**, unexpectedly large differences

G0 ML baseline

Input (size)	Model	Paper	Accuracy 95th CI	
			Script	Human
Flowpic (32x32)	CNN LeNet5	IMC22	98.67 *n.a.*	92.40 *n.a.*
(a) Flowpic (32x32)	XGBoost	Ours	96.80±0.37 (-1.87)	73.65±2.14 (-18.75)
(b) Time series (3x10)	XGBoost	Ours	94.53±0.56 (-4.14)	66.91±1.40 (-25.49)

(a) Flattened flowpic; (b) concat first 10 values of packet size, direction and inter arrival time
Our results are aggregation of 15 experiments (5 splits x 3 seeds)

- On **Script**, results on par with flowpic but lower performance with time series *(10 pkts -vs- 15s of traffic)*
- On **Human**, unexpectedly large differences

Be very cautious to understand the cause of the performance discrepancy

HUAWEI

Supervised settings

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87.07	77.30	71.65±1.98	71.08±1.51	68.19±0.97	96.76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.56±0.55	97.16±0.62	94.93±0.68	82.93	74.93	68.00	68.43±2.82	70.20±1.99	69.08±1.72	96.93±0.56	96.46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	71.89±1.59	71.08±1.33	97.02±0.50	97.51±0.46	97.67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean difi				-2.05	-2.26	-0.63				-21.96	-13.27	-9.13			

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87.07	77.30	71.65±1.98	71.08±1.51	68.19±0.97	96.76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.56±0.55	97.16±0.62	94.93±0.68	82.93	74.93	68.00	68.43±2.82	70.20±1.99	69.08±1.72	96.93±0.56	96.46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	71.89±1.59	71.08±1.33	97.02±0.50	97.51±0.46	97.67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean diff	2.21			-2.05	-2.26	-0.63	12.19			-21.96	-13.27	-9.13			

From IMC22 evaluation

- 32x32 is superior to higher resolutions

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87.07	77.30	71.65±1.98	71.08±1.51	68.19±0.97	96.76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.5[illegible]	97.16±0.62	94.93±0.68	82.93	74.93	68.00	68.43±2.82	70.20±1.99	69.08±1.72	96.93±0.56	96.46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	71.89±1.59	71.08±1.33	97.02±0.50	97.51±0.46	97.67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean diff				-2.05	-2.26	-0.63				-21.96	-13.27	-9.13			

6.61

From IMC22 evaluation

- 32x32 is superior to higher resolutions
- Contained difference between Script and Human partitions

HUAWEI

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87.07	77.30	71.65±1.98	71.08±1.51	68.19±0.97	96.76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.56±0.55	[illegible]	94.93±0.68	82.93	74.93	68.00	68.43±2.82	[illegible]	69.08±1.72	96.93±0.56	96.46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	71.89±1.59	71.08±1.33	97.02±0.50	97.51±0.46	97.67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean dif				-2.05	-2.26	-0.63				-21.96	-13.27	-9.13			

0.81

-0.64

From IMC22 evaluation

- 32x32 is superior to higher resolutions
- Contained difference between Script and Human partitions

From Our evaluation

- Small differences between resolutions
  but 1 model @1500x1500 takes ~20min vs <1min @32x32

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87.07	77.30	71.65±1.98	71.08±1.51	68.19±0.97	96.76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.56±0.55	97.16±0.62	94.93±0.68	82.93	74.93	68.00	68.43±2.82	70.20±1.99	69.08±1.72	96.93±0.56	96.46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	71.89±1.59	71.08±1.33	97.02±0.50	97.51±0.46	97.67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean dif				-2.05	-2.26	-0.63				-21.96	-13.27	-9.13			

From IMC22 evaluation

- 32x32 is superior to higher resolutions
- Contained difference between Script and Human partitions

From Our evaluation

- Small differences between resolutions
  but 1 model @1500x1500 takes ~20min vs <1min @32x32
- Confirmed discrepancy observed via XGBoost

G1 Benchmark augmentations in supervised setting

Each ours value is an aggregation of 15 experiments (5 splits x 3 seeds)

	Test on **Script**						Test on **Human**						Test on **Leftover**		
	IMC22			Ours			IMC22			Ours			Ours		
flowpic res.	32	64	1500	32	64	1500	32	64	1500	32	64	1500	32	64	1500
No augment.	98.67	99.10	96.22	95.64±0.37	95.87±0.29	94.93±0.72	92.40	85.60	73.30	68.84±1.45	69.08±1.35	69.32±1.63	95.78±0.29	96.09±0.38	95.79±0.51
Rotate	98.60	98.87	94.89	96.31±0.44	96.93±0.46	95.69±0.39	93.73	87	[illegible]	71.65±1.98	71.08±1.51	68.19±0.97	[illegible].76±0.35	97.00±0.38	95.79±0.31
Horizontal flip	98.93	99.27	97.33	95.47±0.45	96.00±0.59	94.86±0.79	94.67	79.33	87.90	69.40±1.63	70.52±2.03	73.90±1.06	95.68±0.40	96.32±0.59	95.97±0.80
Color jitter	96.73	96.40	94.00	97.56±0.55	97.16±0.62	94.93±0.68	82.93	74.93	68.0	[illegible]	70.20±1.99	69.08±1.72	96.93±0.56	[illegible].46±0.46	95.47±0.49
Packet loss	98.73	99.60	96.22	96.89±0.52	96.84±0.63	95.96±0.51	90.93	85.60	84.00	70.68±1.35	71.33±1.45	71.08±1.13	96.99±0.39	97.25±0.39	96.84±0.49
Time shift	99.13	99.53	97.56	96.71±0.60	97.16±0.49	96.89±0.27	92.80	87.30	77.30	70.36±1.63	[illegible]	71.08±1.33	97.02±0.50	97.51±0.46	[illegible].67±0.29
Change RTT	99.40	100.00	98.44	97.29±0.35	97.02±0.46	96.93±0.31	96.40	88.60	90.70	70.76±1.99	71.49±1.59	71.97±1.08	98.38±0.18	97.97±0.39	98.19±0.22
Mean diff				-2.05	-2.26	-0.63				-21.96	-13.27	-9.13			

-0.23

-0.23

-0.79

From IMC22 evaluation

- 32x32 is superior to higher resolutions
- Contained difference between Script and Human partitions

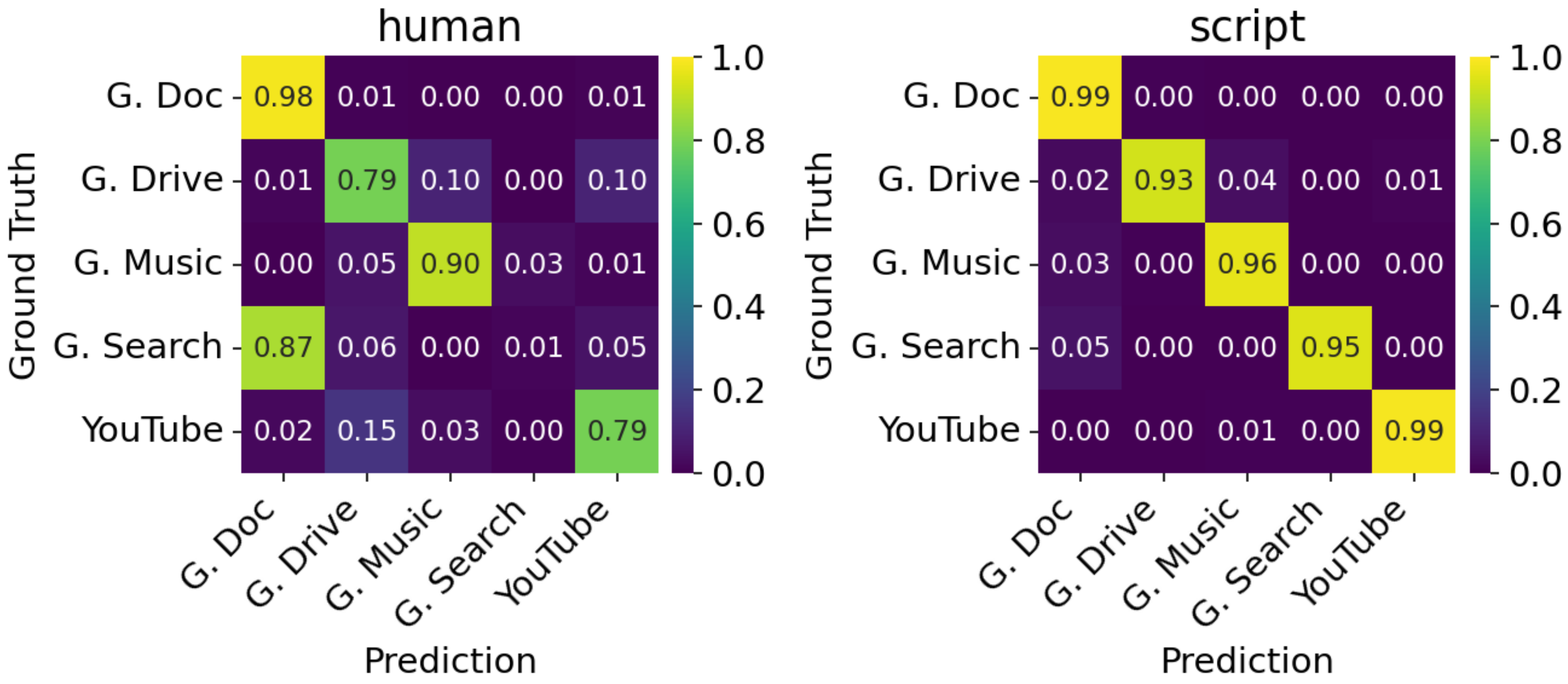
From Our evaluation

- Small differences between resolutions
  but 1 model @1500x1500 takes ~20min vs <1min @32x32
- Confirmed discrepancy observed via XGBoost
- Leftover is consistent with Script

…so, what’s the problem with Human?

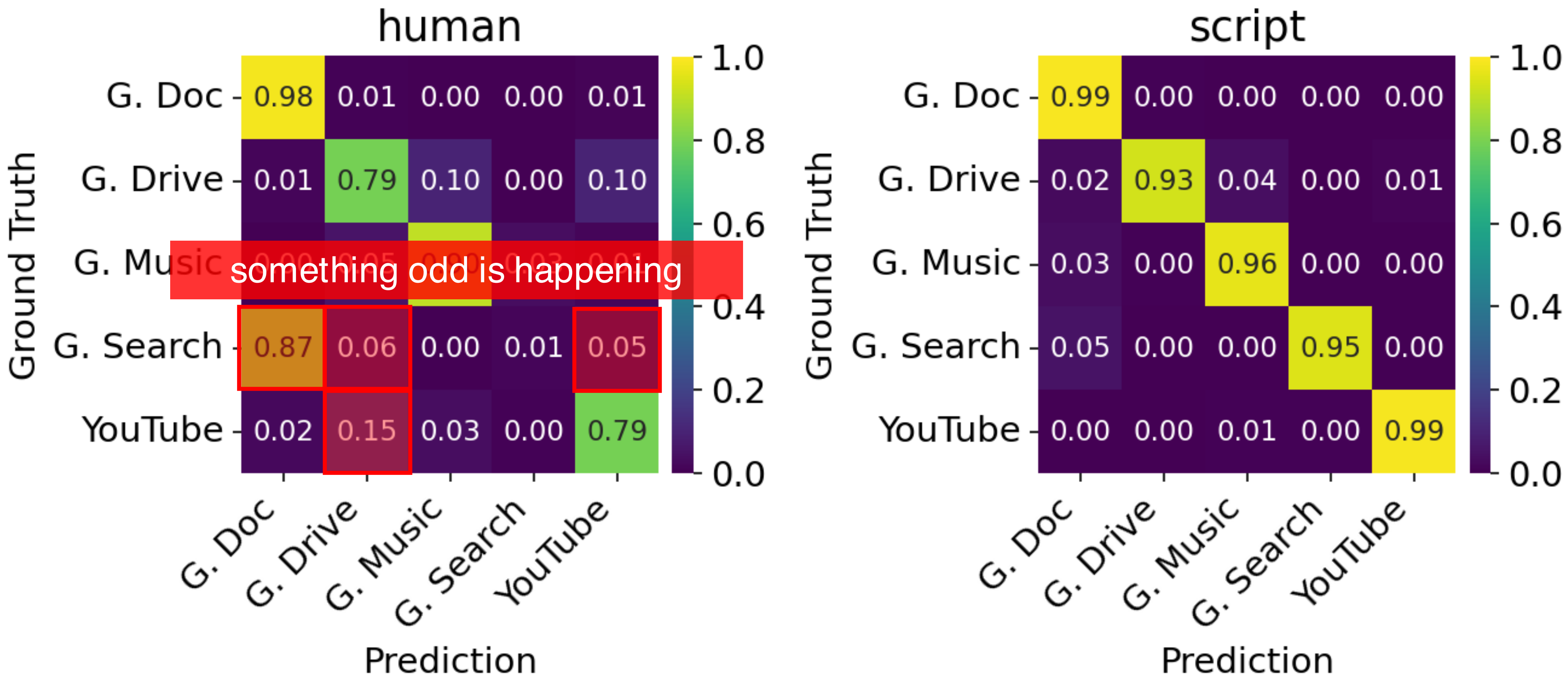
Investigating human-vs-script performance gap

Confusion matrixes



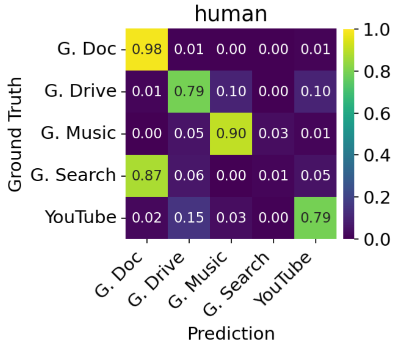
Investigating human-vs-script performance gap

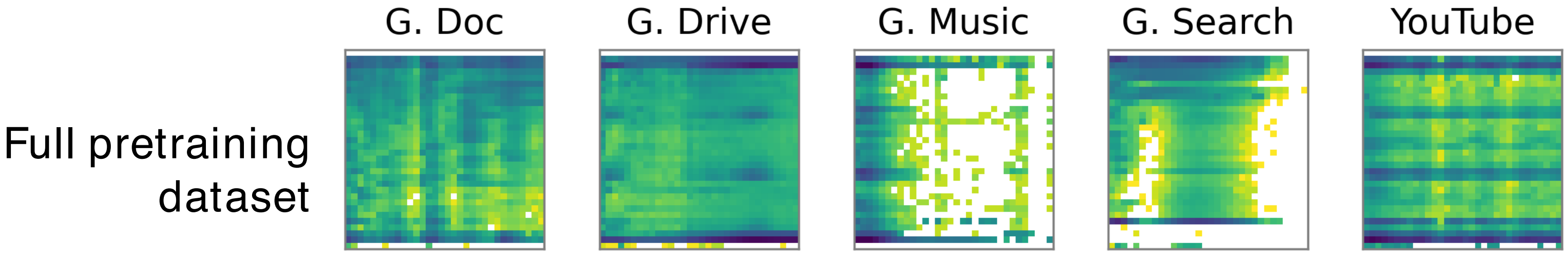
Confusion matrixes

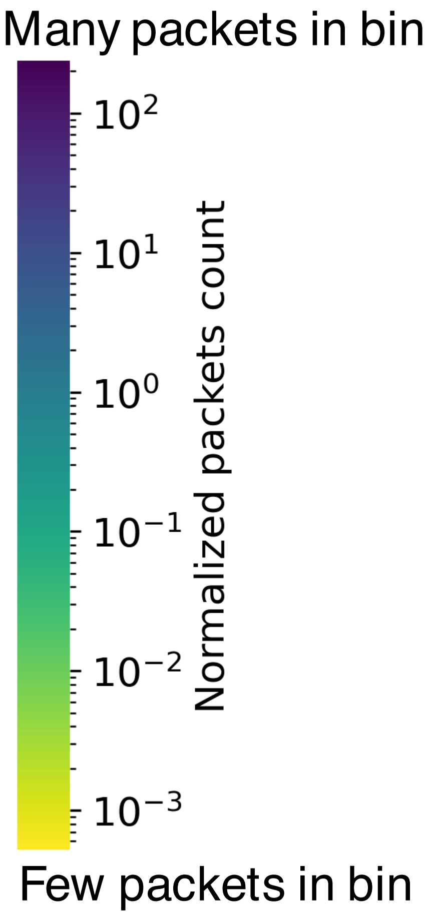


Investigating human-vs-script performance gap

Average flowpics

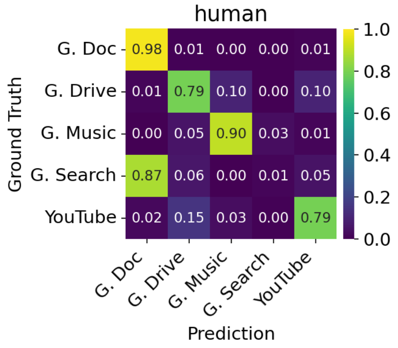


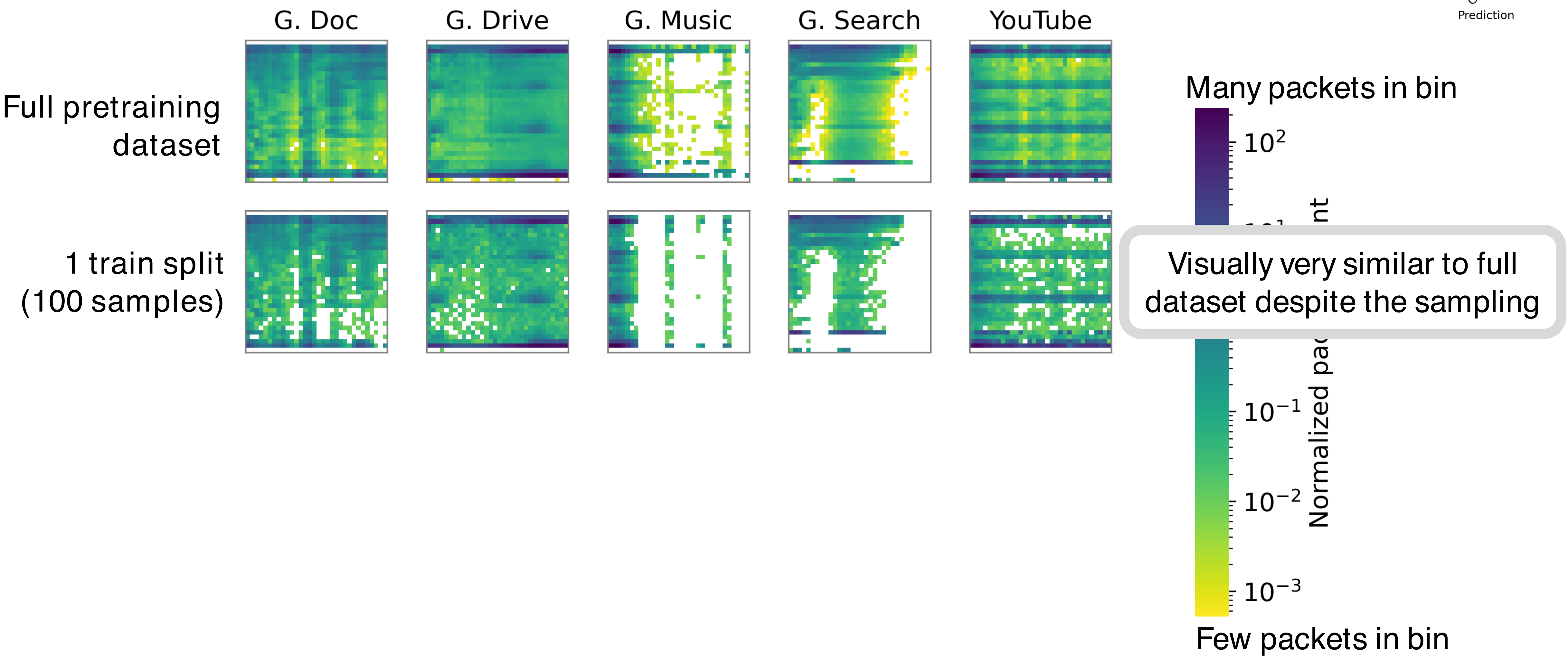




Investigating human-vs-script performance gap

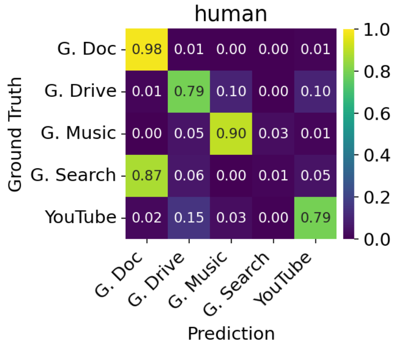
Average flowpics





Investigating human-vs-script performance gap

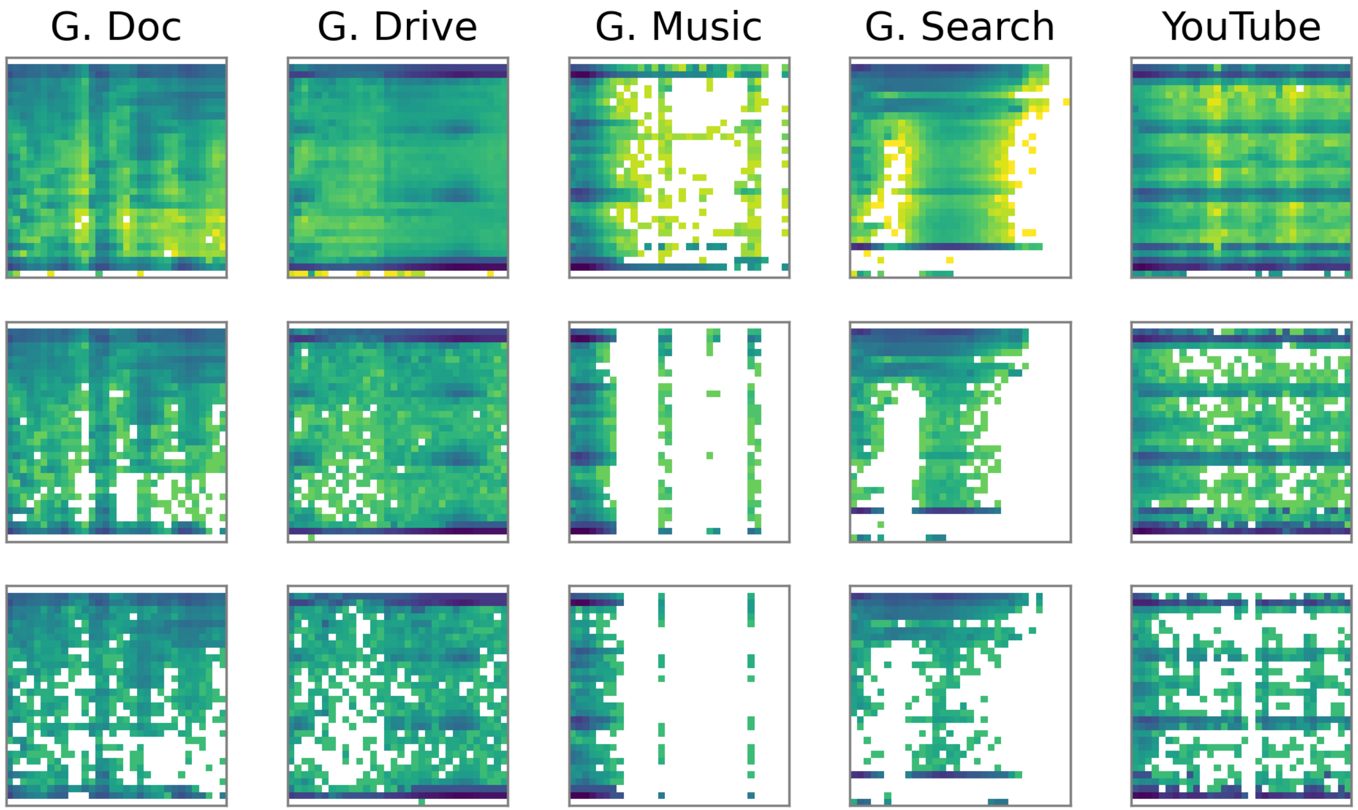
Average flowpics



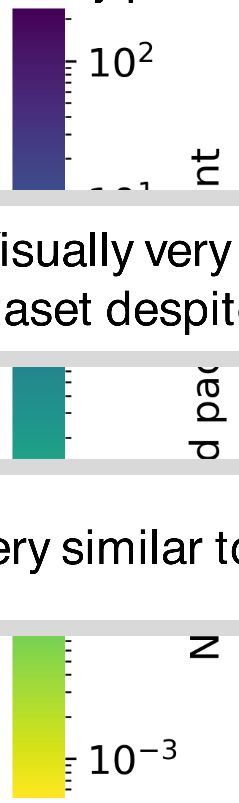
Full pretraining dataset

1 train split (100 samples)

Full **Script** partition



Many packets in bin



Few packets in bin

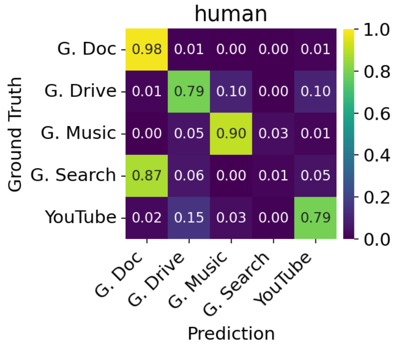
Visually very similar to full dataset despite the sampling

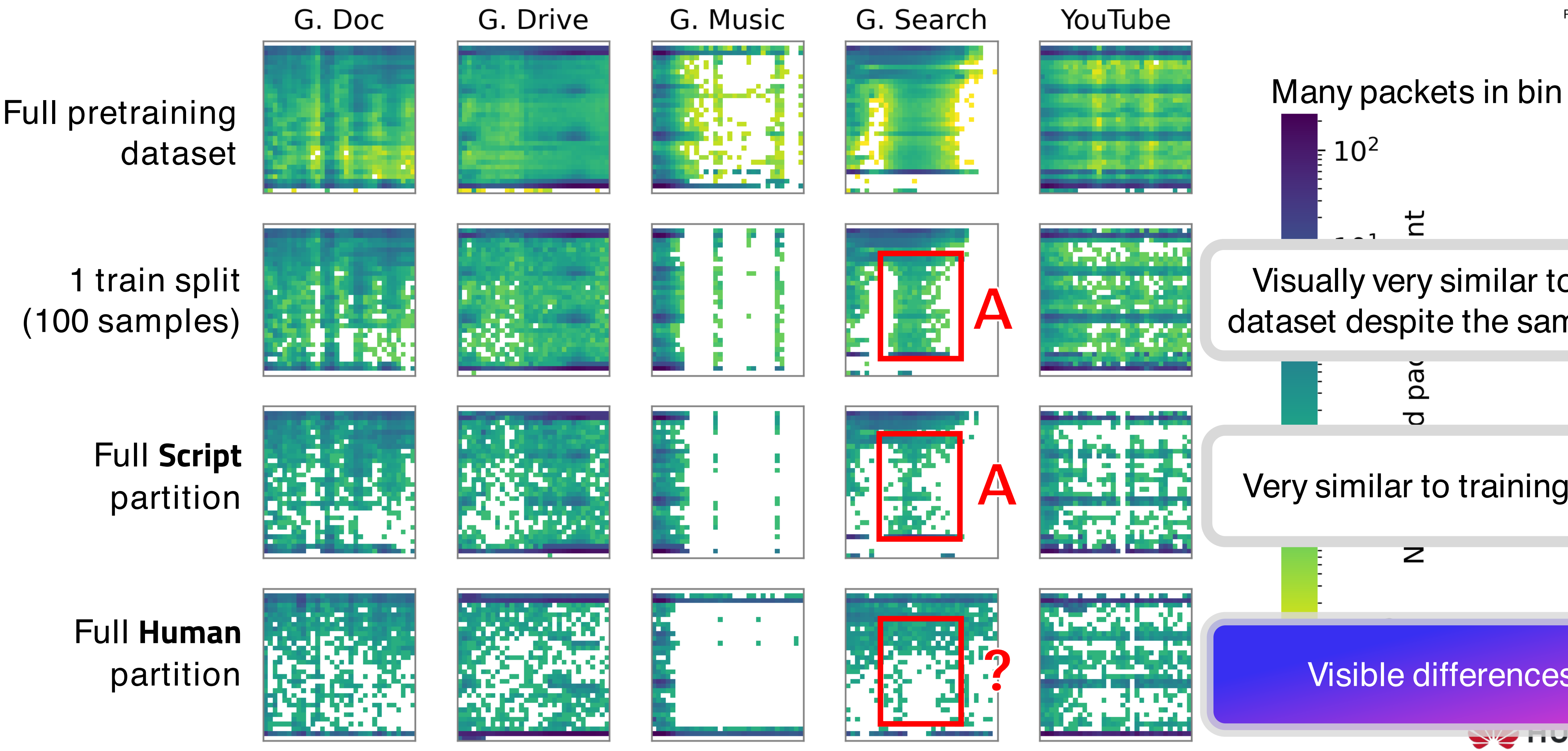
Very similar to training split

HUAWEI

Investigating human-vs-script performance gap

Average flowpics





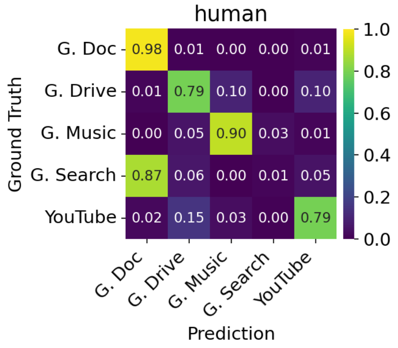
Visually very similar to full dataset despite the sampling

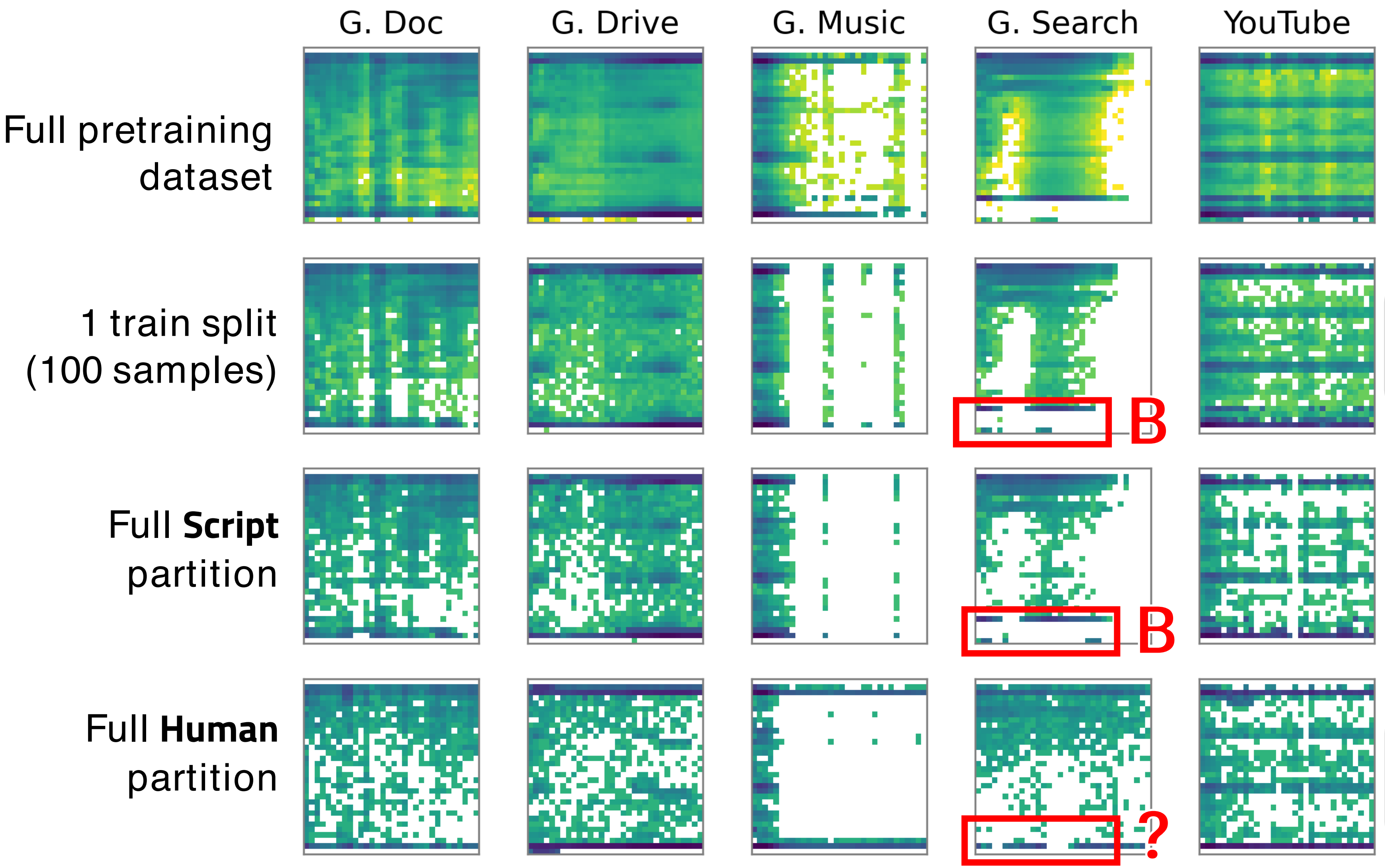
Very similar to training split

Visible differences

Investigating human-vs-script performance gap

Average flowpics





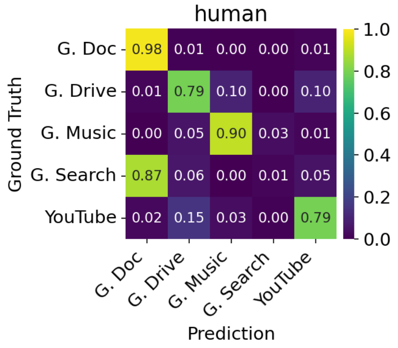
Visually very similar to full dataset despite the sampling

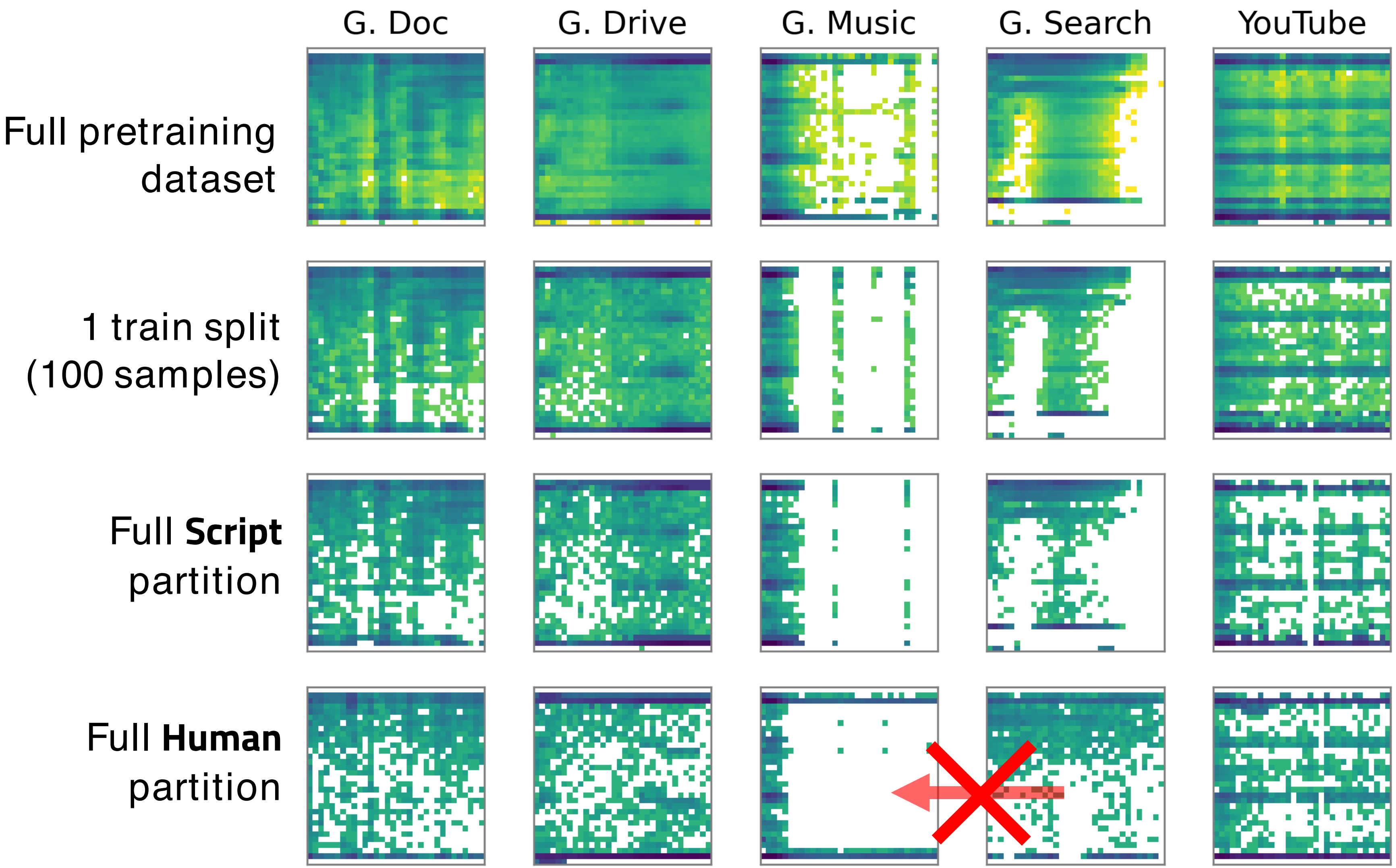
Very similar to training split

Visible differences

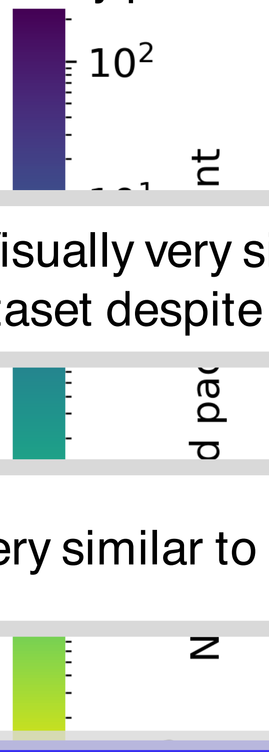
Investigating human-vs-script performance gap

Average flowpics





Many packets in bin



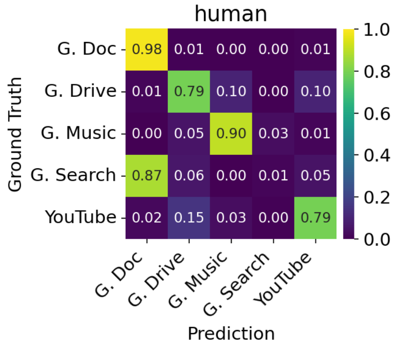
Visually very similar to full dataset despite the sampling

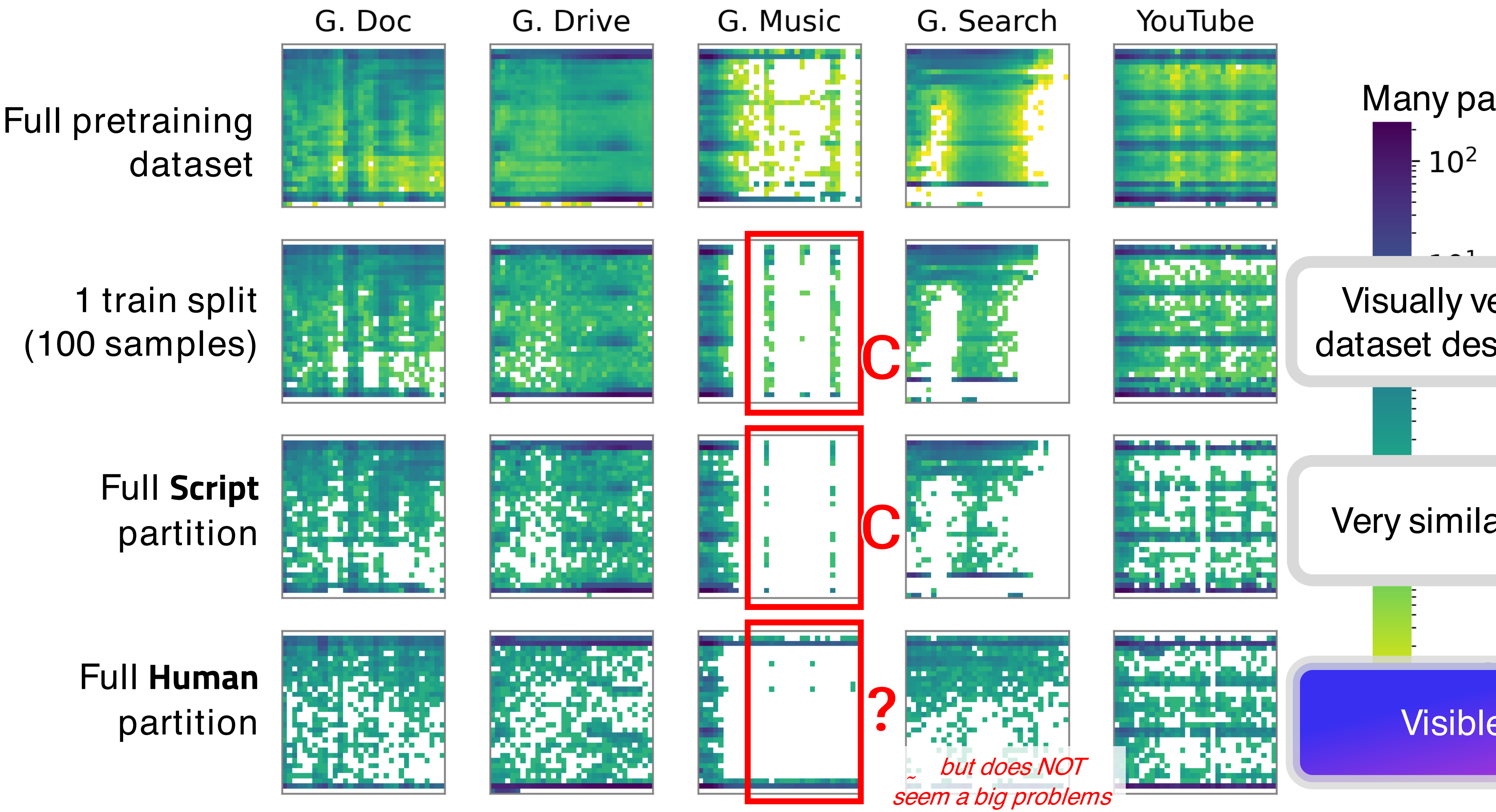
Very similar to training split

Visible differences

Investigating human-vs-script performance gap

Average flowpics





Visually very similar to full dataset despite the sampling

Very similar to training split

Visible differences

but does NOT seem a big problems

UCDAVIS-19 human partition suffers from a data shift

confirmed by

1. More analysis of the dataset
2. Replication of results of [1]

check our paper appendix 😉

Unclear why this did not affect IMC22 *paper results*

[1] How to Achieve High Classification Accuracy with Just a Few Labels: A Semi-supervised Approach Using Sampled Packets, ICDM19

HUAWEI

G1

Benchmark augmentations in supervised setting

Ranking augmentations

In the **IMC22** paper states that

- Change RTT is the best performing augmentations
- Time series augmentations are better than image transformations

...but no confidence reported

G1

Benchmark augmentations in supervised setting

Ranking augmentations

In the **IMC22** paper states that

- Change RTT is the best performing augmentations
- Time series augmentations are better than image transformations

...but no confidence reported

We study **augmentations performance** via **critical distance [1]**

- For the same input configuration, rank augmentations from best (1) to worse (7)
- Compute average rank for each augmentation
- Use a pair-wise post-hoc Nemenyi test based and CD to assess statistical similarity

Critical Distance (CD) = $q_\alpha \sqrt{\frac{k(k+1)}{6N}}$

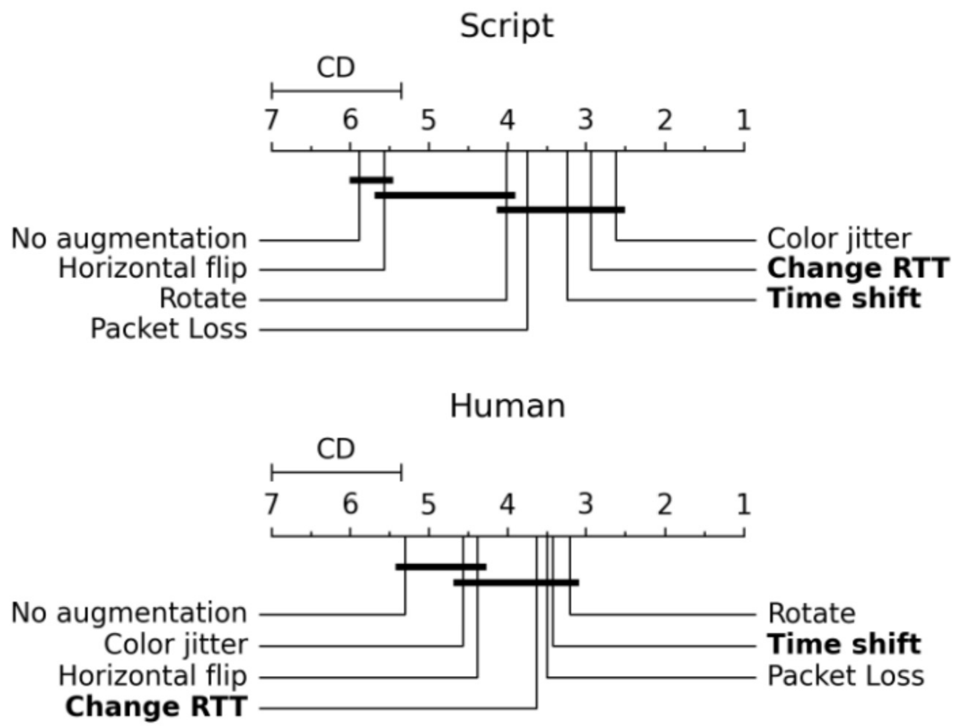
k : number of augmentations
N : number of experiments
q_α : studentized range statistic

[1] Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research

G1

Benchmark augmentations in supervised setting

Ranking augmentations

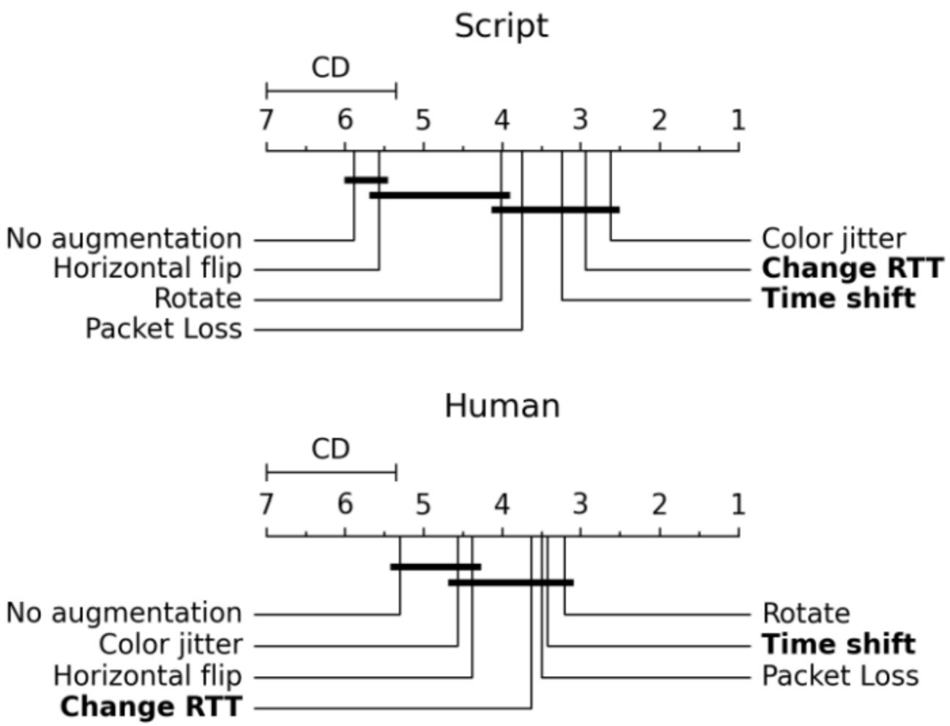


Augmentations connected by horizontal lines are NOT statistically different

[1] Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research

G1 Benchmark augmentations in supervised setting

Ranking augmentations



Augmentations connected by horizontal lines are NOT statistically different

Takeaway

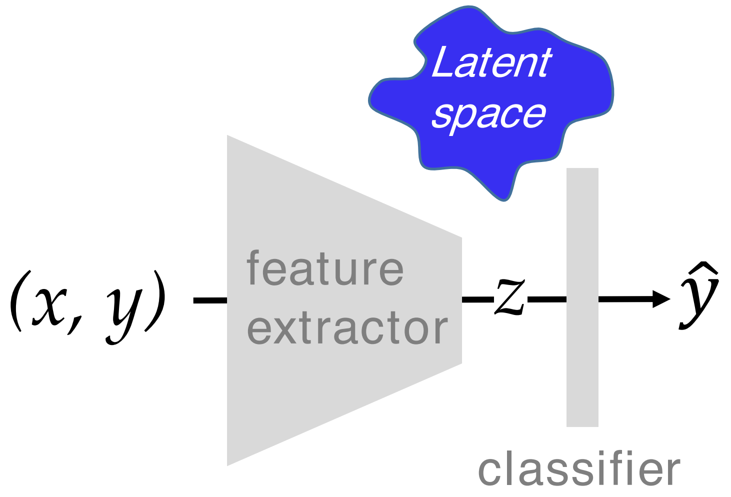
- Augmentations improve performance
- Time series augmentations are not statistically different from image augmentations

[1] Statistical comparisons of classifiers over multiple data sets. The Journal of Machine learning research

Contrastive learning settings

HUAWEI

Supervised -vs- Contrastive learning



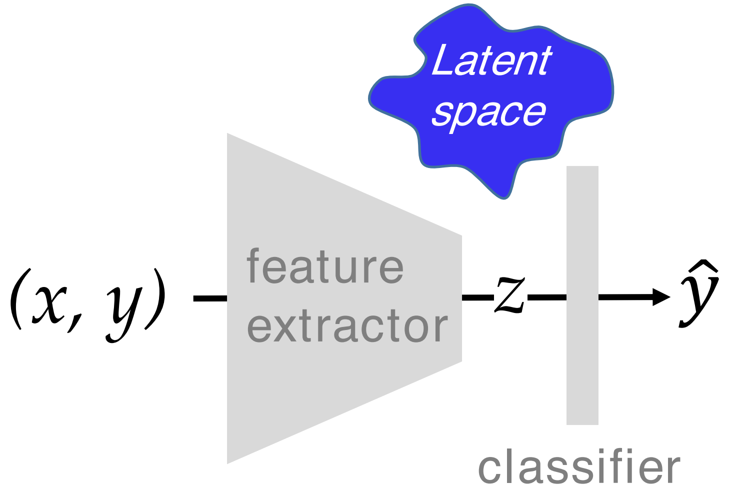
In **supervised** training

Good separation in the latent space leads to good performance

...but

- The (cross entropy) loss is computed after the classifier
- The latent space geometry is *indirectly* controlled

Supervised -vs- Contrastive learning

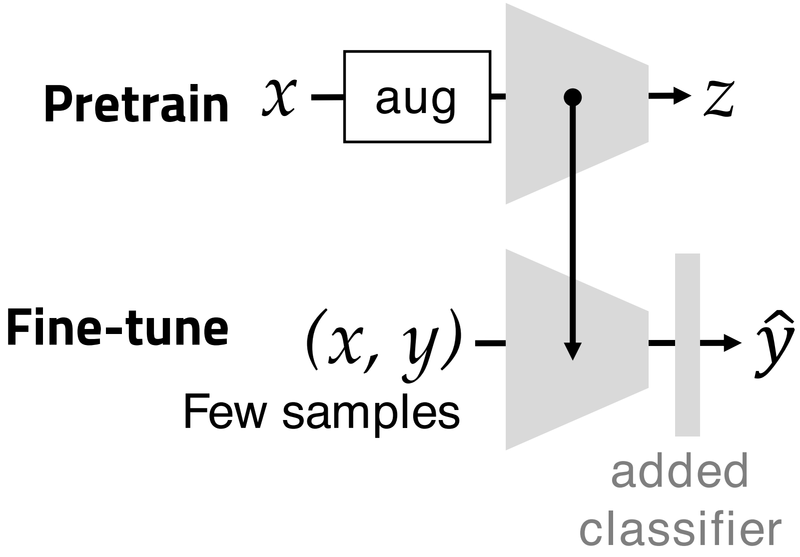


In **supervised** training

Good separation in the latent space leads to good performance

…but

- The (cross entropy) loss is computed after the classifier
- The latent space geometry is *indirectly* controlled

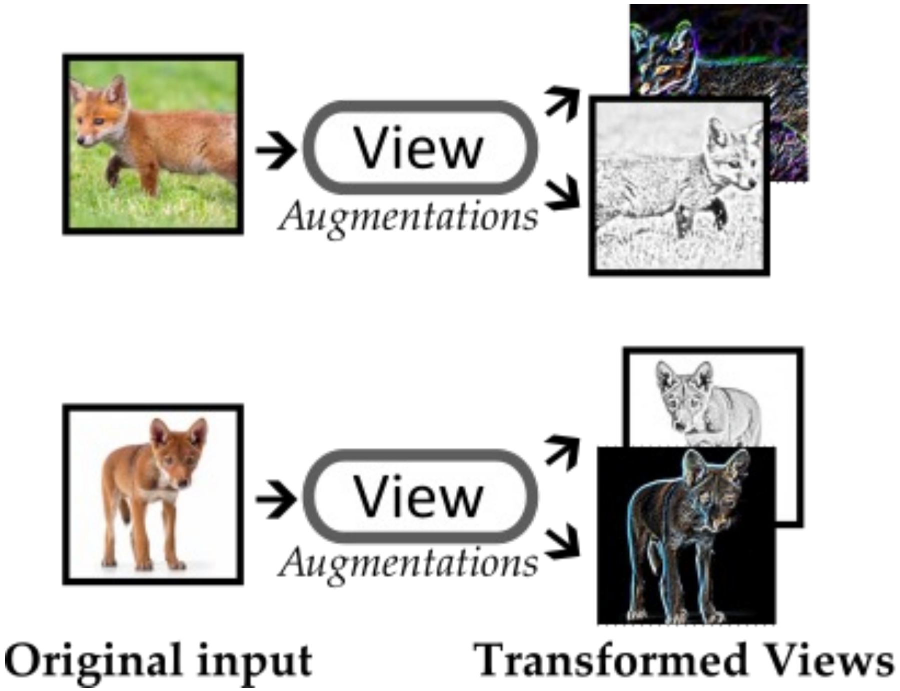


In **contrastive learning** training

- **First** a model is trained in an **unsupervised** manner controlling the latent space geometry
- **Then** the learned representation is finetuned with a **few labeled samples** for the specific classification task

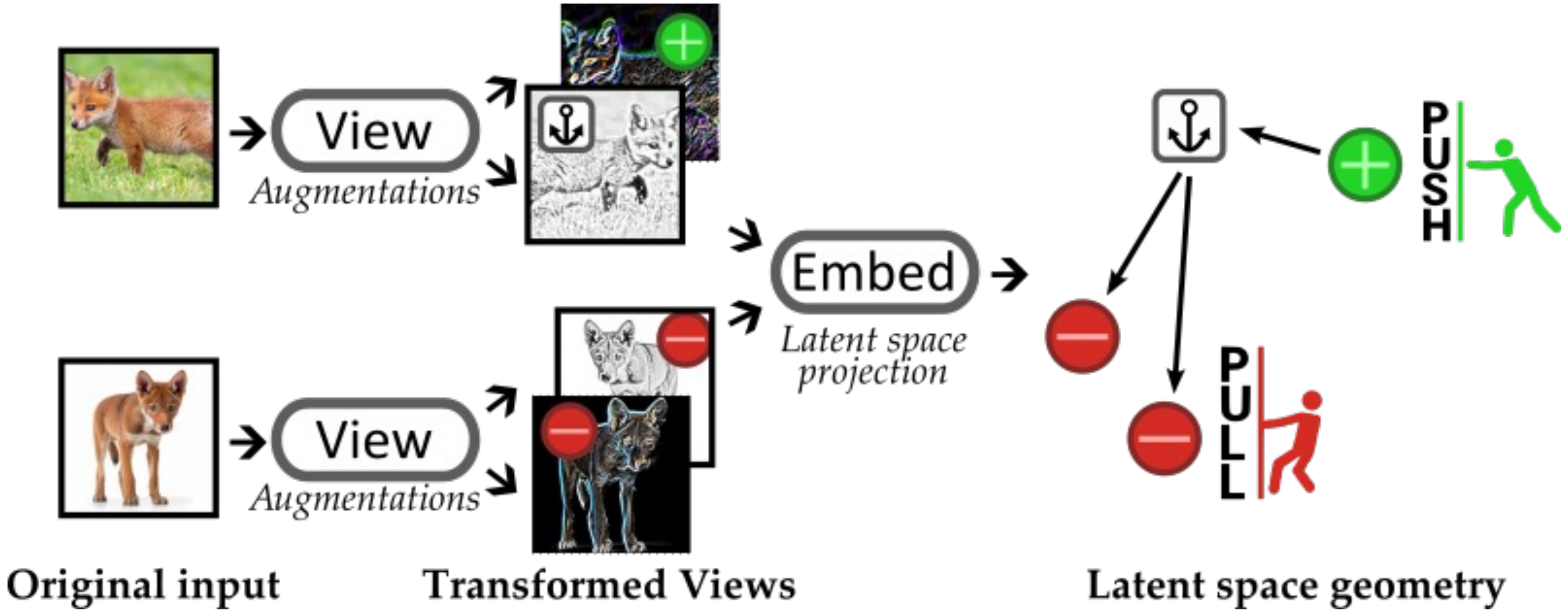
Self-supervision in contrastive learning

Base principle: In the absence of a label, a sample can only be similar to itself



Self-supervision in contrastive learning

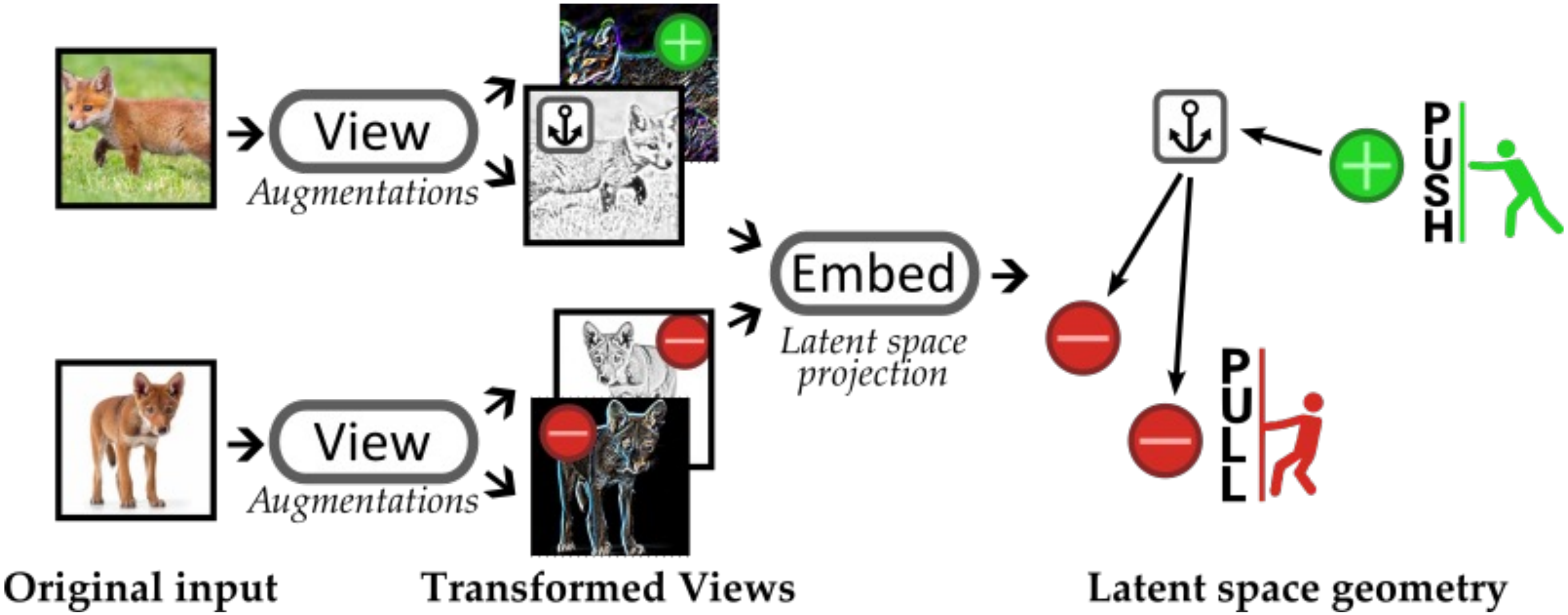
Base principle: In the absence of a label, a sample can only be similar to itself



HUAWEI

Self-supervision in contrastive learning

Base principle: In the absence of a label, a sample can only be similar to itself



- Positive and anchor form *their own class* → harder problem than supervision
- The better the representation, the smaller the trainset to finetune a classifier

G2 Contrastive learning + finetuning (1/2)

Small pretraining

- *Which algorithm?* SimCLR [1]
- *Which augmentations?* TimeShift and ChangeRTT
- *Which dataset size?* 100 samples for pretrain, 10 for finetune

[1] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

G2 Contrastive learning + finetuning (1/2)

Small pretraining

- *Which algorithm?* SimCLR [1]
- *Which augmentations?* TimeShift and ChangeRTT
- *Which dataset size?* 100 samples for pretrain, 10 for finetune

1st augment.	**IMC22**	Change RTT	Packet loss		Change rtt		Color jitter
2nd augment.		Time shift	Color jitter	Rotate	Color jitter	Rotate	Rotate
Test on **Script**	94.5	92.18±0.31	90.17±0.41	91.94±0.30	91.72±0.36	**92.38±0.32**	91.79±0.34
Test on **Human**	~80.0	74.69±1.13	73.67±1.24	71.22±1.20	**75.56±1.23**	74.33±1.26	71.64±1.23

[1] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

G2 Contrastive learning + finetuning (1/2)

Small pretraining

- *Which algorithm?* SimCLR [1]
- *Which augmentations?* TimeShift and ChangeRTT
- *Which dataset size?* 100 samples for pretrain, 10 for finetune

1st augment.	IMC22	Change RTT	Packet loss		Change rtt		Color jitter
2nd augment.		Time shift	Color jitter	Rotate	Color jitter	Rotate	Rotate
Test on **Script**	94.5	92.18±0.31	90.17±0.41	91.94±0.30	91.72±0.36	**92.38**±0.32	91.79±0.34
Test on **Human**	~80.0	74.69±1.13	73.67±1.24	71.22±1.20	**75.56**±1.23	74.33±1.26	71.64±1.23

Takeaways

- On **Script**, performance are comparable to IMC22
- On **Human**, still evident performance gap
- Any transformation pair is qualitative equivalent

[1] A Simple Framework for Contrastive Learning of Visual Representations, ICML20

G2

Contrastive learning + finetuning (2/2)

Large pretraining

Lifting the constraint of 100 samples per class → 80/20 train/val split on the whole pretraining

- Script improves in supervised setting
- Human improves in contrastive learning setting

		Script	Human
Supervised	No augmentation	98.37±0.19	72.95±0.96
	Rotate	98.47±0.25	73.73±1.09
	Horizontal flip	98.20±0.15	74.58±1.16
	Color jitter	98.63±0.21	72.47±1.02
	Packet loss	98.63±0.19	73.43±1.25
	Time shift	98.60±0.22	73.25±1.17
	Change rtt	98.33±0.16	72.47±1.04
SimCLR + fine-tuning		93.90±0.74	80.45±2.37

G2 Contrastive learning + finetuning (2/2)

Large pretraining

Lifting the constraint of 100 samples per class → 80/20 train/val split on the whole pretraining

- Script improves in supervised setting
- Human improves in contrastive learning setting

		Script	Human
Supervised	No augmentation	98.37±0.19	72.95±0.96
	Rotate	98.47±0.25	73.73±1.09
	Horizontal flip	98.20±0.15	74.58±1.16
	Color jitter	98.63±0.21	72.47±1.02
	Packet loss	98.63±0.19	73.43±1.25
	Time shift	98.60±0.22	73.25±1.17
	Change rtt	98.33±0.16	72.47±1.04
SimCLR + fine-tuning		93.90±0.74	80.45±2.37

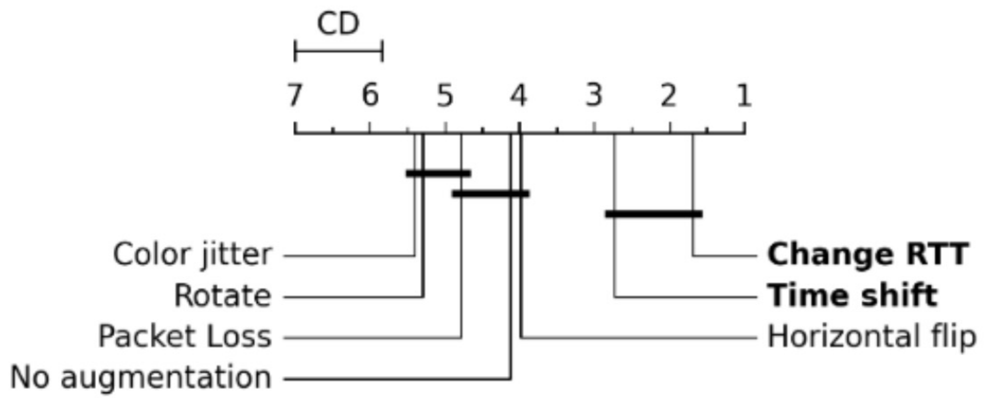
Takeaways

- Augmentations are not the final replacement for real samples
- Contrastive learning can help to reduce data shift (?)

Other datasets

G3 Benchmarking augmentations on other datasets

Augmentations	MIRAGE-22 (>10pkts)	MIRAGE-22 (>1000pkts)	UTMOBILENET-21 (>10pkts)	MIRAGE-19 (>10pkts)
No augmentation	90.97±1.15	83.35±3.13	79.82±1.53	69.91±1.57
Rotate	88.25±1.20	**87.32**±2.24	79.45±1.28	60.35±1.17
Horizontal flip	91.90±0.84	83.82±2.26	80.03±1.33	69.78±1.28
Color jitter	89.77±1.16	81.40±3.62	78.68±2.14	67.00±1.11
Packet loss	92.34±1.10	87.19±2.52	72.07±1.73	67.55±1.46
Time shift	**92.80**±1.21	86.73±3.88	**81.91**±2.21	**70.33**±1.26
Change RTT	**93.75**±0.83	**91.48**±2.12	**81.32**±1.54	**74.28**±1.22



Takeaways

Change RTT and Time Shift are better than other augmentations

HUAWEI

Want more?

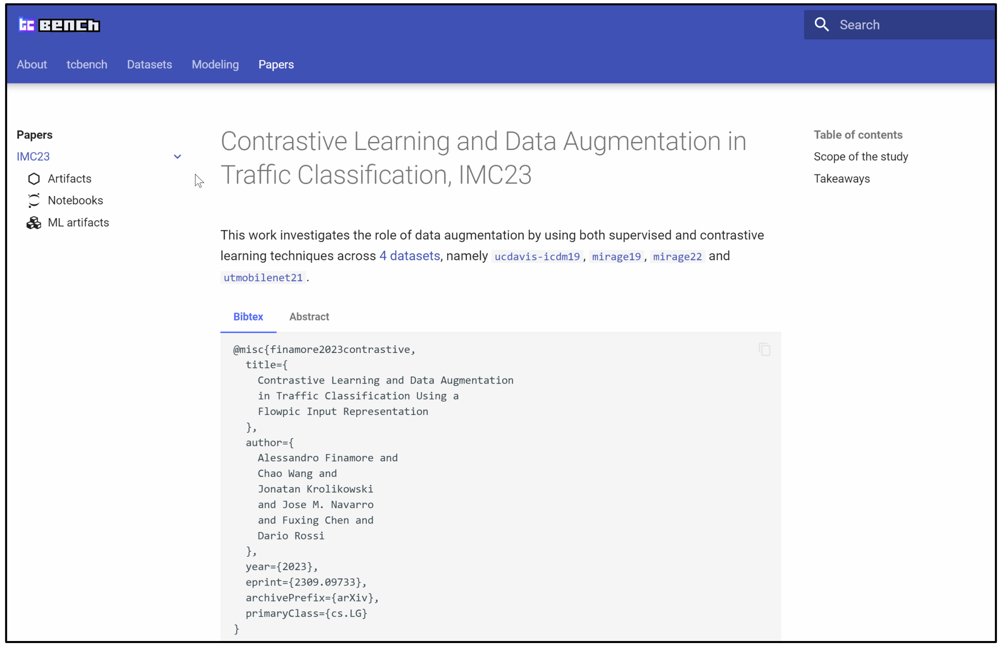
- Analysis of dropout
- Analysis of SimCLR projection layers
- …and other details

…in the paper

Outline

1. Introduce the IMC22 paper and set our goals
2. Datasets and methodology
3. Results
4. Closing remarks

Closing remarks



Replication is incredibly hard
…but worth if **geared toward community contributions**

Qualitatively our results are aligned with the IMC22 paper but the UCDAVIS-19 **data shift has an impact**

There is **space for more research** in the areas touched by our paper (check our paper for inspiration 😉)

Thank you

alessandro.finamore@huawei.com
mail@afinamore.io

https://afinamore.io
https://prc-ai4net.github.io/

Code artifacts

https://github.com/tcbenchstack/tcbench

Data artifacts

https://doi.org/10.6084/m9.figshare.c.6849252.v3

Documentation

https://tcbenchstack.github.io/tcbench/papers/imc23/