Data Analytics:
Gen-AI

Project Assignment

Domenico Ciuonzo, Ten. Track Assistant Prof.

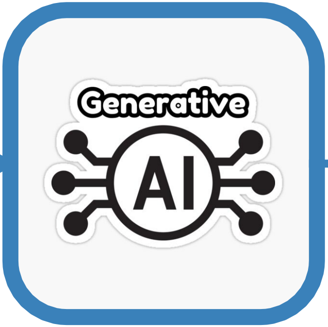
University of Napoli Federico II, Italy

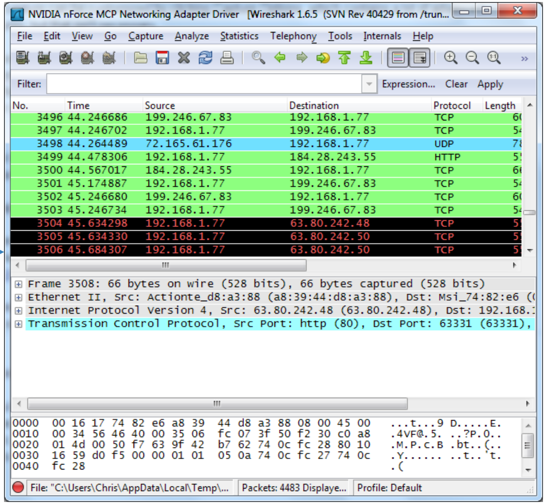
domenico.ciuonzo@unina.it

GenAI: Context and Motivation in Networking

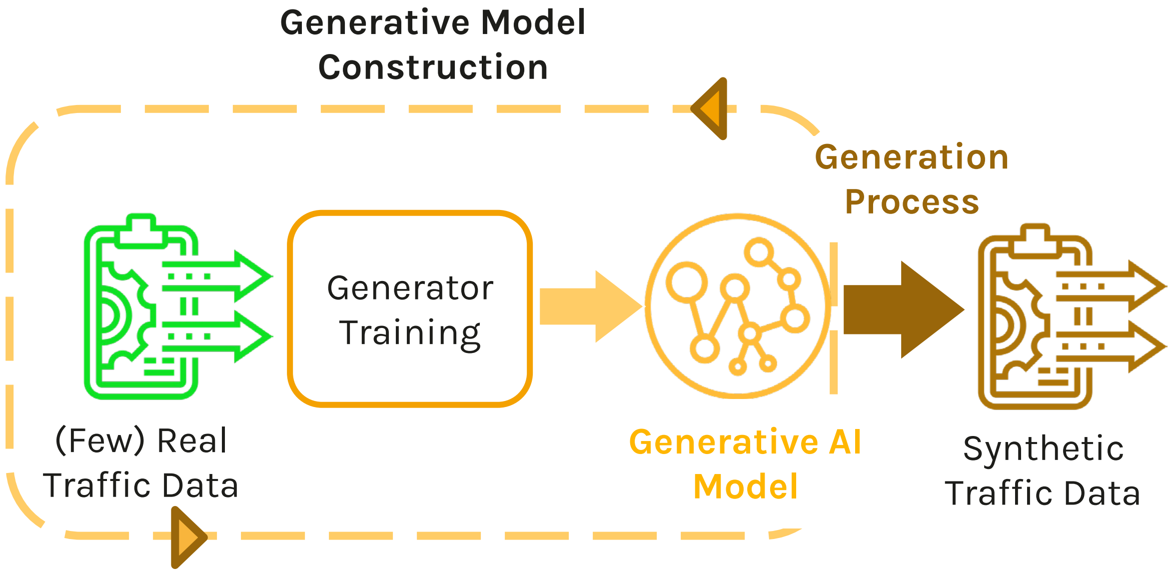
How to bring network traffic generation to the computer vision level (e.g., Dall·E)?

"Generate Netflix traffic data"



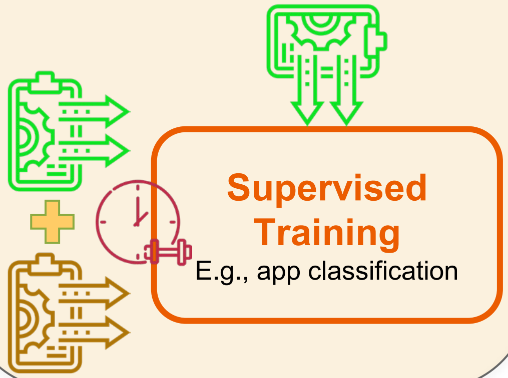


GenAI: Application

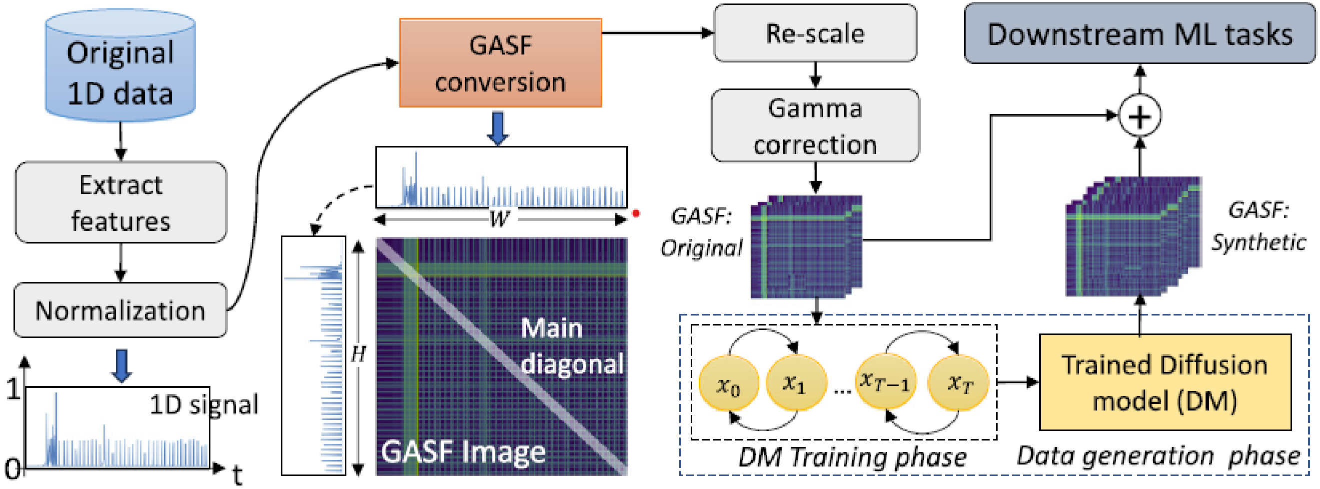


[1] Assessing synthetic traffic data validity/high-fidelity

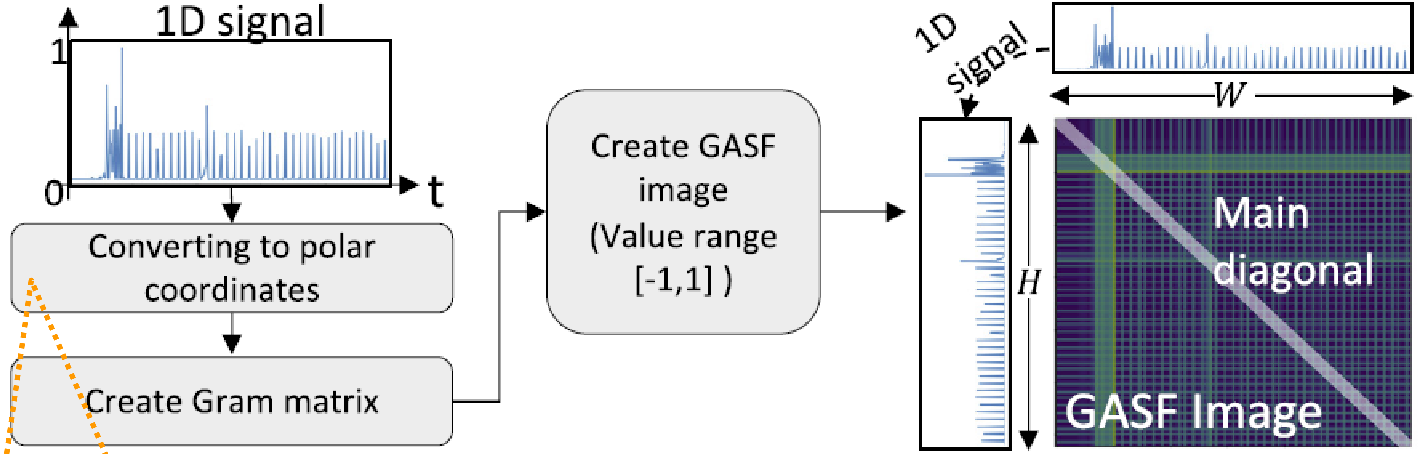
[2] Augmenting the training set of a downstream task



NetDiffus: Network traffic generation by diffusion models through time-series imaging

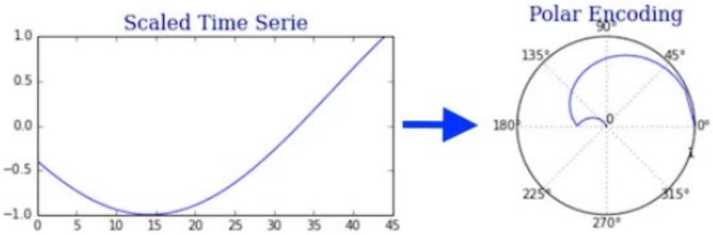


NetDiffus: GASF conversion

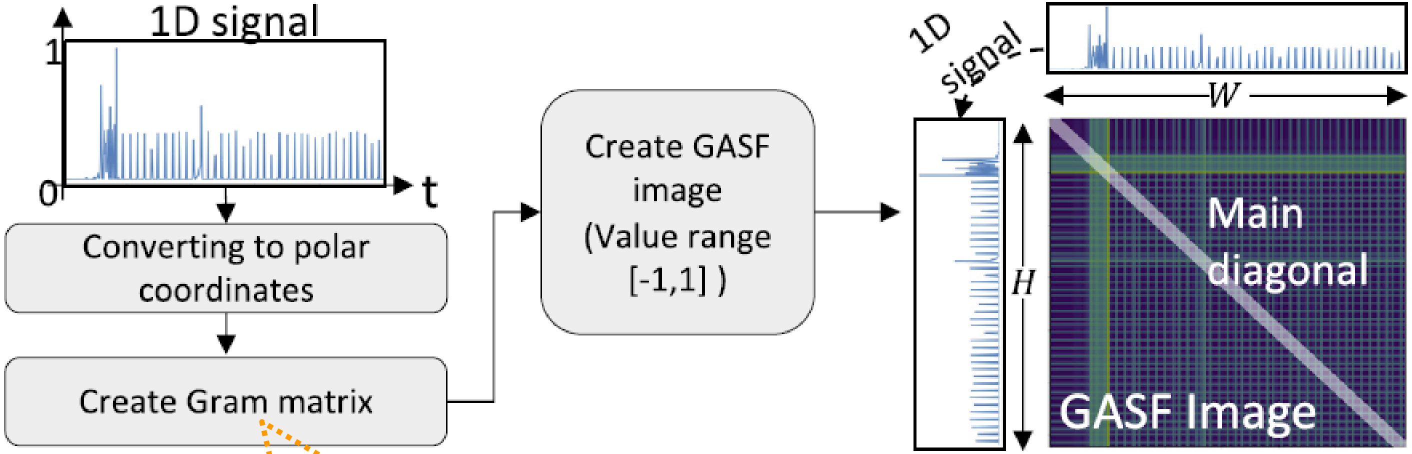


$$\theta_i = \arccos(x_i), \; x_i \in X$$

$$rad_i = \frac{t_i}{C}$$



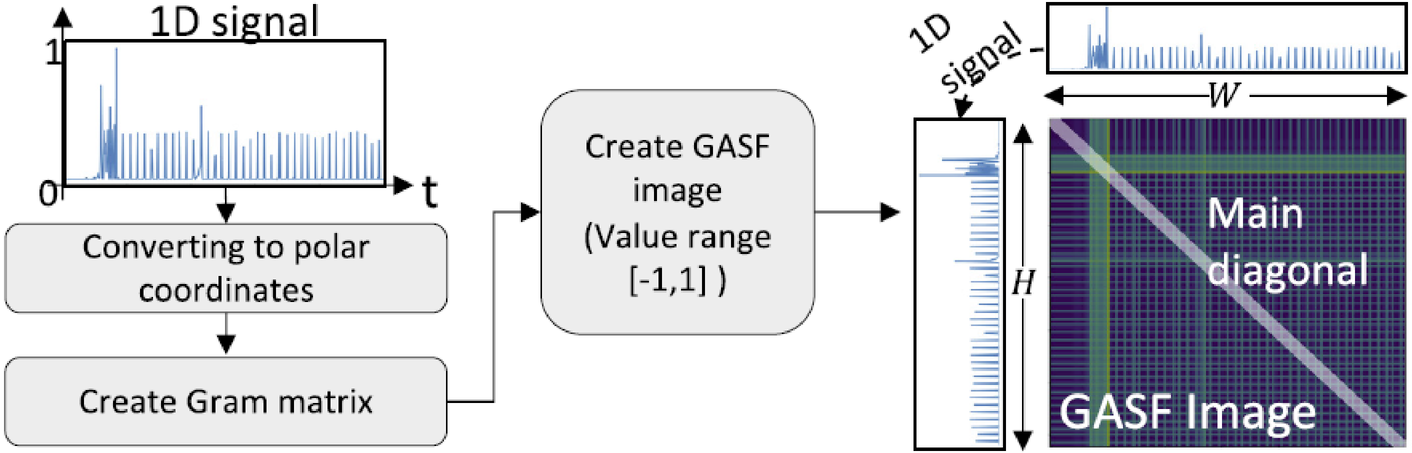
NetDiffus: GASF conversion



$$GASF = [\cos(\theta_i + \theta_j)]$$
$$= X'.X - \sqrt{I - X^2}'.\sqrt{I - X^2}$$

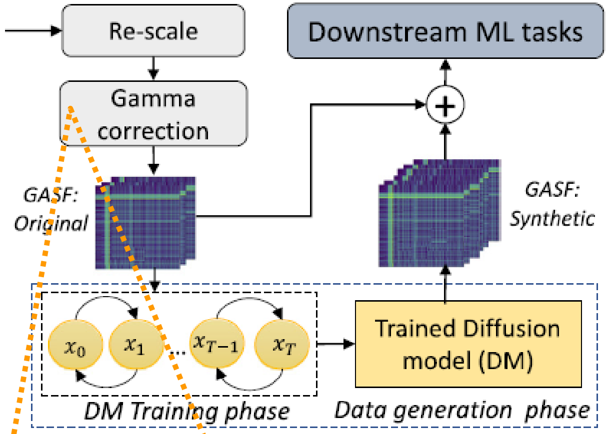
$$G = \begin{pmatrix} \cos(\phi_1 + \phi_1) & cos(\phi_1 + \phi_2) & \dots & cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & cos(\phi_2 + \phi_2) & \dots & cos(\phi_2 + \phi_n) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\phi_n + \phi_1) & cos(\phi_n + \phi_2) & \dots & cos(\phi_n + \phi_n) \end{pmatrix}$$

NetDiffus: GASF conversion

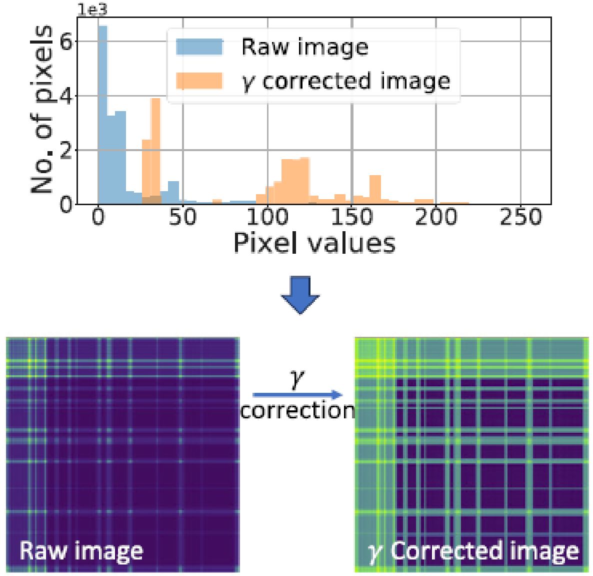


- constant values in a certain segment of the 1D signal are represented as constant pixel value patches on the GASF image
- the main diagonal directly corresponds to the initial 1D trace
- symmetric around the main diagonal

NetDiffus: Gamma correction

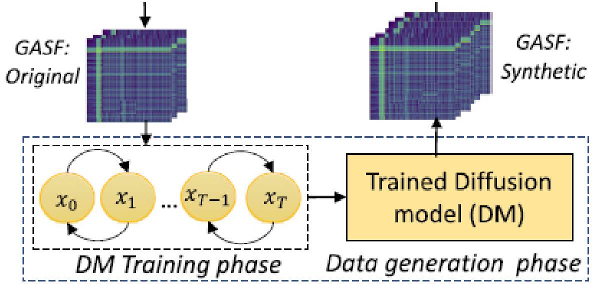


As we operate in 2D domain, enhancing the contrasts of GASF images can further highlight subtle feature variations that can be effectively learnt by DMs and improve the fidelity. We leverage standard gamma correction on raw GASF images according to the equation, $I_c = A * I_r^{\gamma}$, where I_r, I_c, A and γ are gamma corrected image, raw image, a constant and gamma variable respectively. We empirically set $\gamma = 0.25$ and $A = 1$. Fig. 3(a) shows the histogram distribution of sample raw and gamma-corrected images. We notice that this process separates the pixel values into distinct ranges increasing the image contrasts and emphasizing the feature variations.

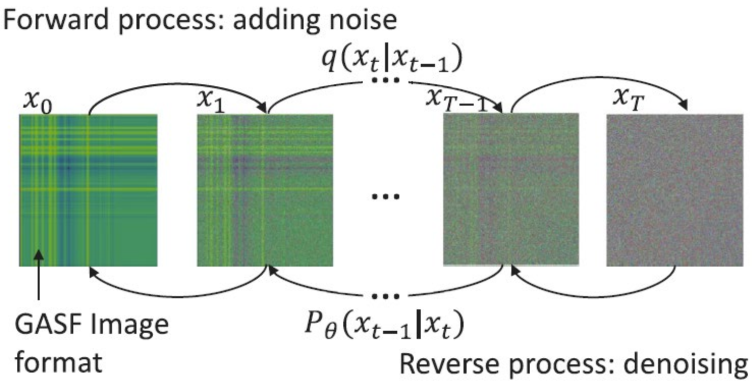


(a) Histogram distribution between raw and γ corrected image

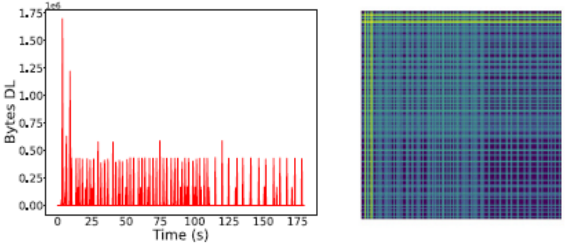
NetDiffus: Diffusion models



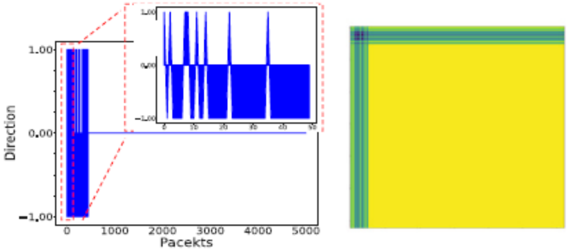
Overview of Diffusion processes:



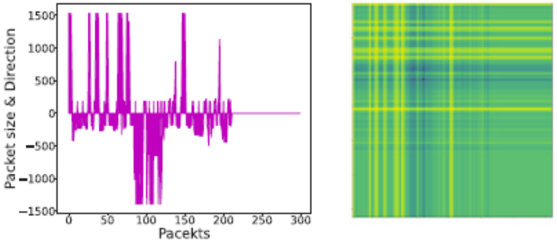
NetDiffus: Evaluation (GASF images)



(a) **D1-YT**: Class 1

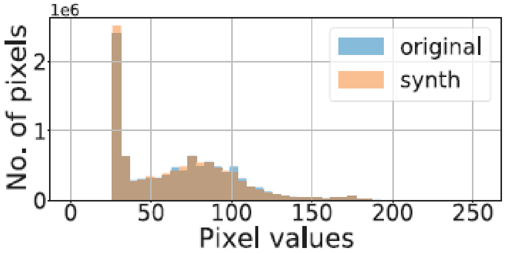


(b) **D2-DF**: Class 1

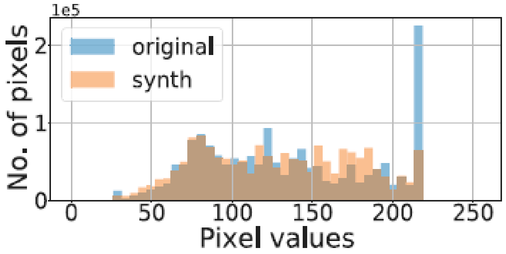


(c) **D3-Google**: Class 1

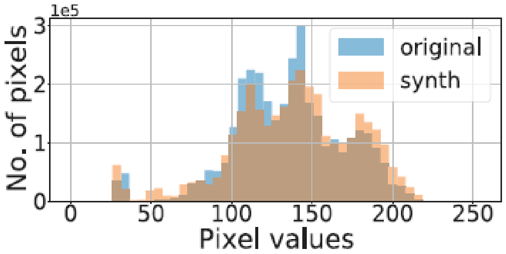
NetDiffus: Evaluation (Fidelity)



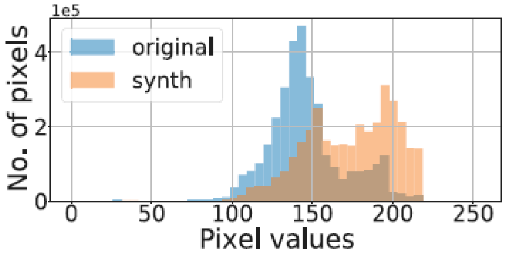
(a) D1-(YT, Stan, Netflix)



(b) D2-DF



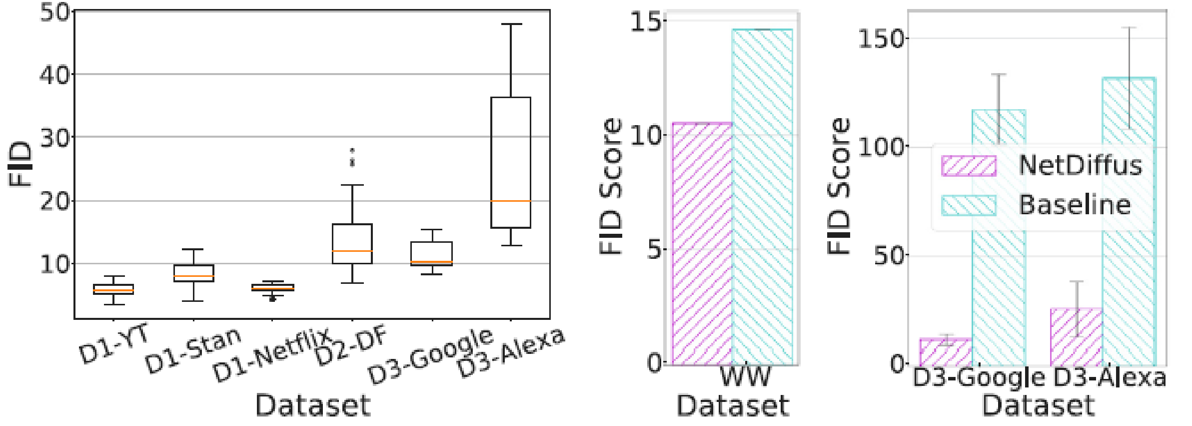
(c) D3-Google



(d) D3-Alexa

How accurate are the generated synthetic GASF images?

NetDiffus: Evaluation (Fidelity)

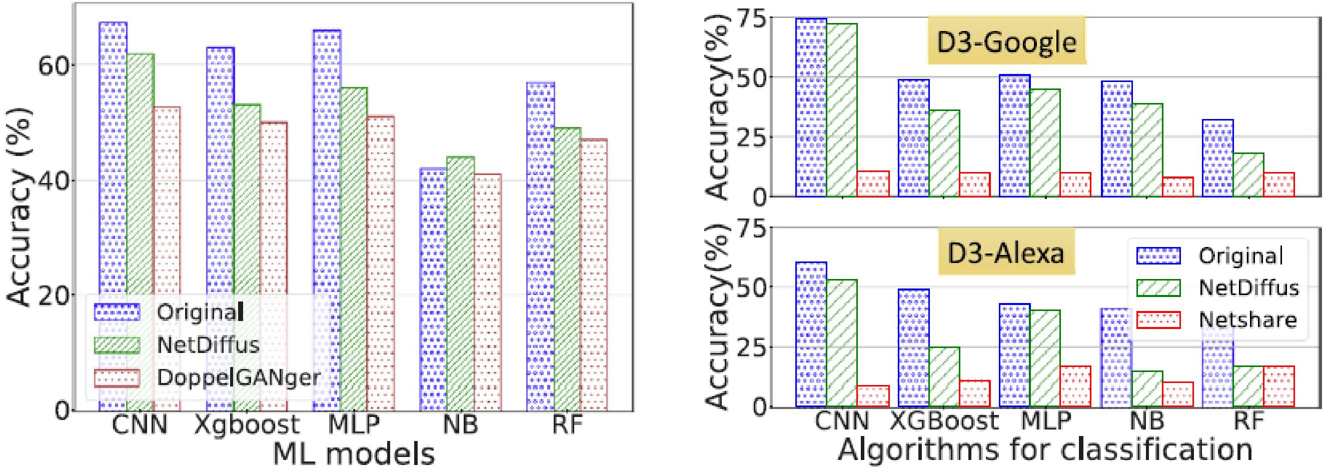


(a) FID score by datasets (b) DG (c) Netshare

How accurate are the generated synthetic GASF images?

NetDiffus: Evaluation (Downstream task #1)

D3: Traffic generated by IoT smart-home devices (Google or Alexa)



(a) DG: WW dataset

(b) Netshare: D3-G & D3-A

Fig. 7. Comparison with baselines G:Google, A:Alexa.

Train on real data -> test on real data (original)

Train on synthetic data -> test on real data (Netdiffus, DoppelGANger/Netshare)

NetDiffus: Evaluation (Downstream task #2)

Early-classification of traffic

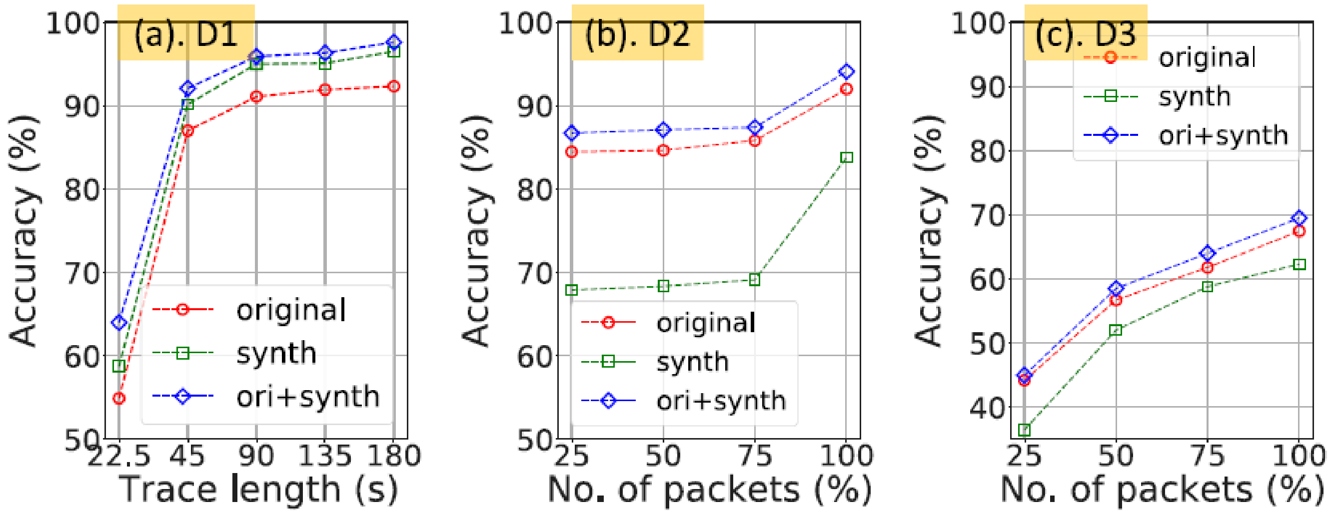
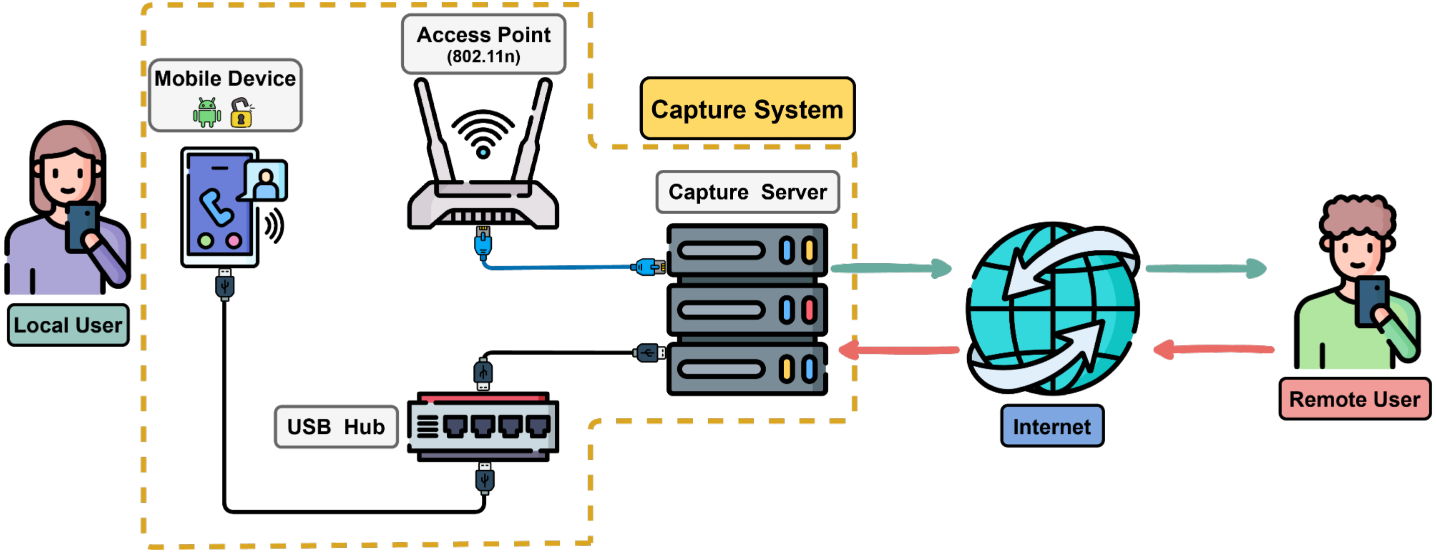


Fig. 12. Performance of L3 classification for different trace lengths/No. of packets.

Train on real data -> test on real data (original)
Train on synthetic data -> test on real data (synth)
Train on synthetic + real data -> test on real data (ori+synth)

Data collection: Mirage Architecture



https://traffic.comics.unina.it/mirage/

Data collection: Mirage Architecture

We release the dataset in **JSON** format, consisting in a JSON file for each capture that provides **3 types of information for each biflow**.

Biflow

Per-packet Data

- Timestamp
- Src/Dst port
- **Packet Direction**
- L3/L4 Header Size
- **L3/L4 Payload Size**
- **Inter-arrival Time**
- TCP Flags
- TCP Window Size
- L4 Payload Bytes (raw)

Flow Metadata

- **Android package name**
- **Activity**
- Android pkg category
- Android pkg vers. code/name
- Device's MAC address
- Labeling criterion
- BF/UF/DF Number of packets
- BF/UF/DF Total B in L3/L4 packets/payload
- BF/UF/DF Duration (s)
- UF/DF MSS/WS

Flow Features

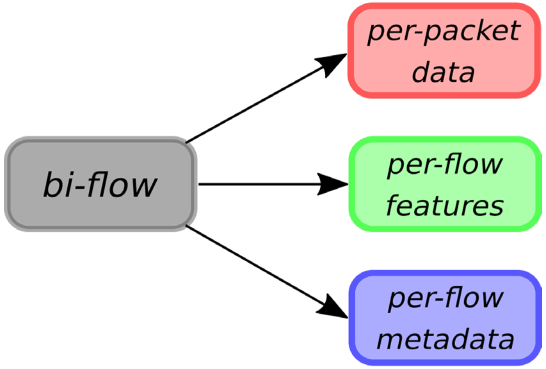
17 Statistical Features (e.g., mean, min, max) on

- BF/UF/DF L3 Packet Length
- BF/UF/DF Inter-arrival Time

Dataset #1: Mirage-2019

A **public human-generated** dataset for mobile traffic analysis

- **40** Android apps
- **16** different categories
- **No less than 2500** bi-flows for each app
- Each bi-flow is labeled with the **Android package-name** of generating app

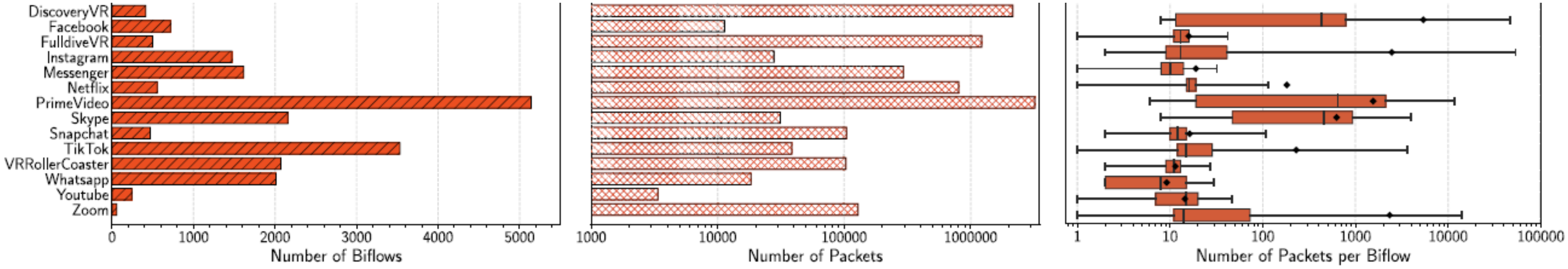


MIRAGE-2019 is released in **JSON** format with information at **different granularities**

Dataset #2: Mirage-Video

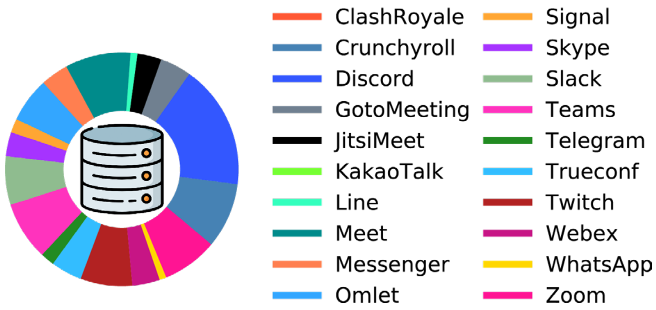
- **Crowdsourcing project** → more than 240 participants (students and researchers)
- **Long Time coverage** → Apr. 2021 - Dec. 2023
- **Real** and **human-generated** traffic
- **4 Android devices** running Android 10 (e.g., Xiaomi Mi 10 Lite)
- **20** popular **Android apps**
- **5 User Activities**: Audiocall, Chat, Online Gaming, Video Streaming, Videocall
- **Reliable app labeling** via netstat

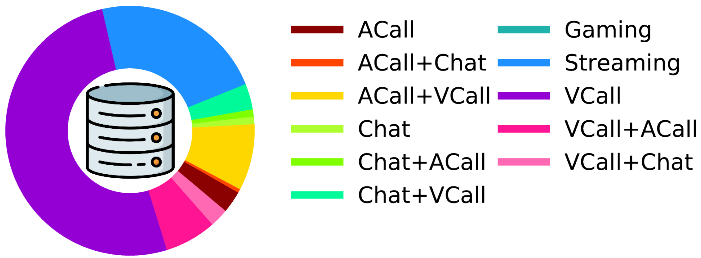
14 Video Android Apps
4 Video Categories
(Cloud VR, Short Video, Video Chat, Video on Demand)



(a) Biflows. (b) Packets. (c) Distribution of biflow lengths.

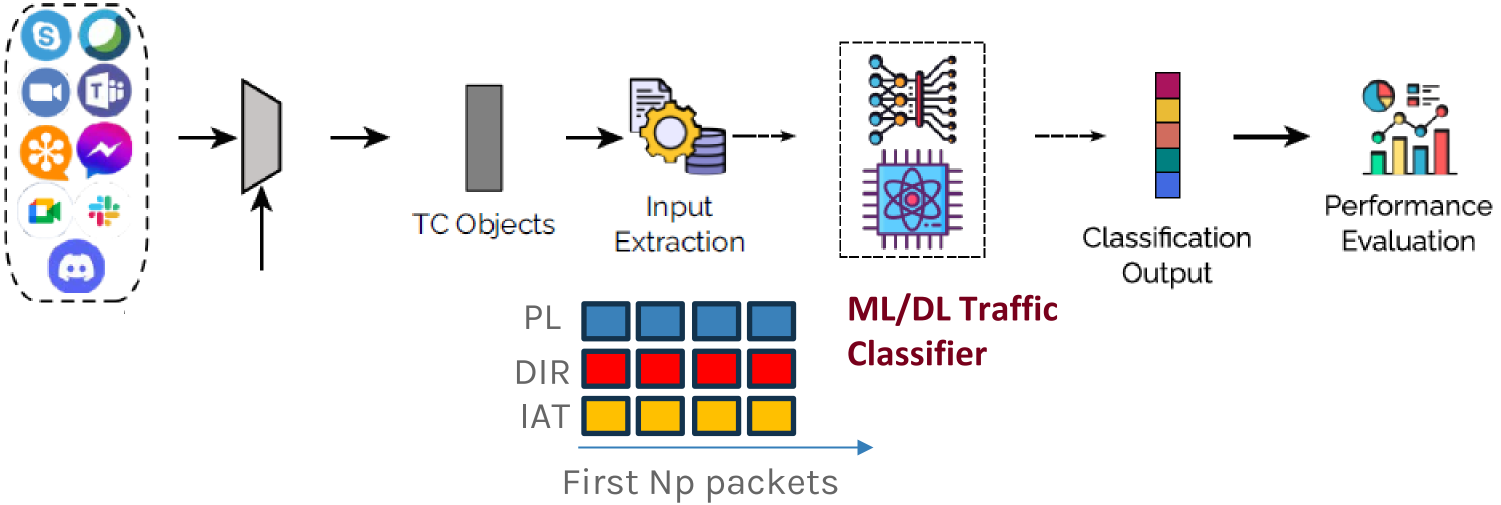
Dataset #3: Mirage-AppxAct





- **Crowdsourcing project** → more than 240 participants (students and researchers)
- **Long Time coverage** → Apr. 2021 - Dec. 2023
- **Real** and **human-generated** traffic
- **4 Android devices** running Android 10 (e.g., Xiaomi Mi 10 Lite)
- **20** popular **Android apps**
- **5 User Activities**: Audiocall, Chat, Online Gaming, Video Streaming, Videocall
- **Reliable app labeling** via netstat

Datasets #1-#3: Commonly-used Format for TC



Riferimenti

- https://github.com/Nirhoshan/NetDiffus
- https://medium.com/analytics-vidhya/encoding-time-series-as-images-b043becbdbf3
- https://traffic.comics.unina.it/mirage/

- Sivaroopan, N., Bandara, D., Madarasingha, C., Jourjon, G., Jayasumana, A. P., & Thilakarathna, K. (2024). Netdiffus: Network traffic generation by diffusion models through time-series imaging. Computer Networks, 251, 110616.

- Aceto, G., Ciuonzo, D., Montieri, A., Persico, V., & Pescapé, A. (2019, October). MIRAGE: Mobile-app traffic capture and ground-truth creation. In 2019 4th International conference on computing, communications and security (ICCCS) (pp. 1-8). IEEE. **[MIRAGE-2019]**

- Montieri, A., Bovenzi, G., Aceto, G., Ciuonzo, D., Persico, V., & Pescapè, A. (2021). Packet-level prediction of mobile-app traffic using multitask deep learning. Computer Networks, 200, 108529. [**MIRAGE-VIDEO]**

- I. Guarino, D. Ciuonzo, A. Montieri, A. Pescapè, MIRAGE-APP× ACT-2024: a Novel Dataset for Mobile App and Activity Traffic Analysis, 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2024) . [**MIRAGE-APPxACT]**