MASTER’S DEGREE IN COMPUTER ENGINEERING

DESIGN AND IMPLEMENTATION OF A VISUAL ANOMALY DETECTION SYSTEM BASED ON ATTENTION

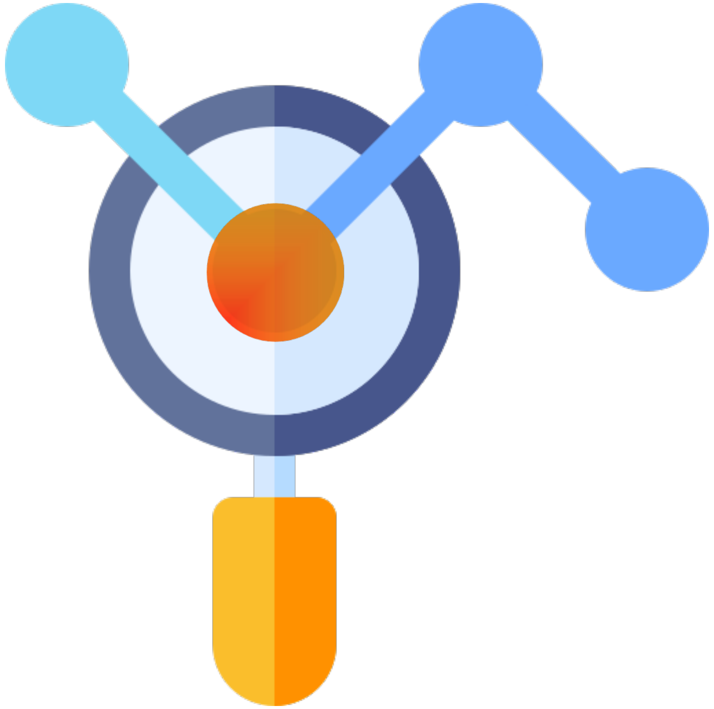
FILIPPO SCOTTO

Claudio GENNARO · Fabrizio FALCHI · Nicola MESSINA

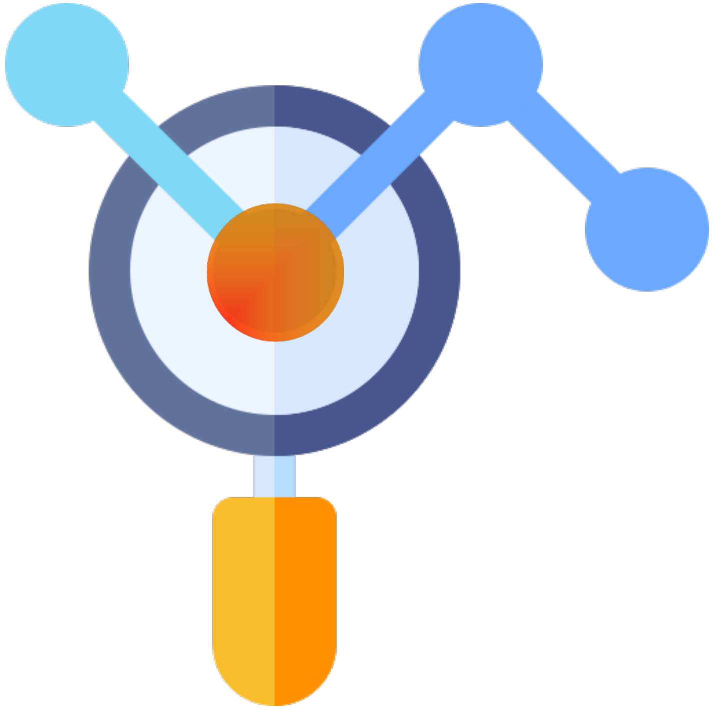
JULY, 2021

WHAT IS ANOMALY DETECTION?

Anomaly Detection is about finding abnormal events or patterns among normality (concerning a *specific context*).



WHAT IS ANOMALY DETECTION?

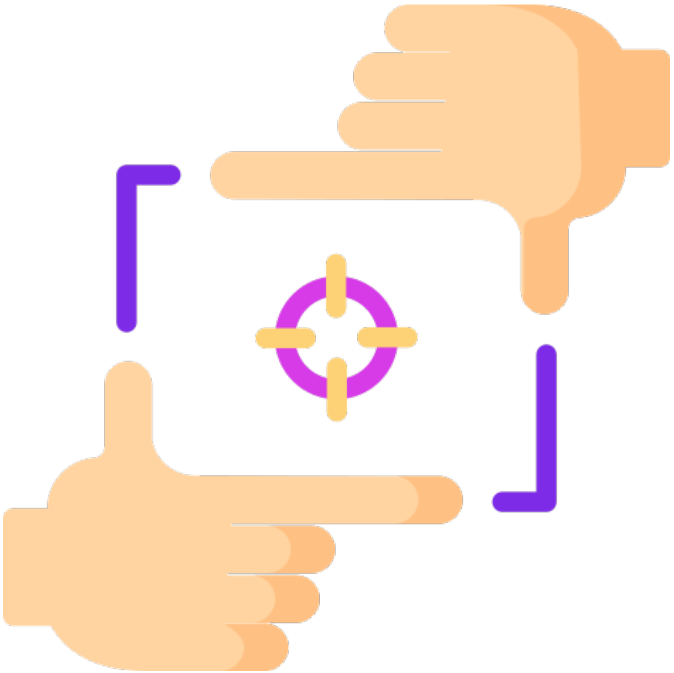


Anomaly Detection is about finding abnormal events or patterns among normality (concerning a *specific context*).

WHAT IS ATTENTION?

The **Attention Mechanism** aims to mimic the cognitive attention in order to enhance the *important parts* of an input and ignore the rest.

It gained popularity in **2017**, with the introduction of the **Transformer***, an architecture that relies solely on attention to solve *Natural Language Processing* (NLP) tasks.



* **Attention is All You Need** – Vaswani et al., *CoRR, abs/1706.03762, 2017*

THE PROPOSAL

Our proposal is to build an anomaly detection system using Attention Mechanisms to extract the features from the video frames.

ANOMALY DETECTION DATASET

ATTENTION-BASED NETWORK

ANOMALY DETECTION

Normal

Anomalous

AN IMAGE IS WORTH 16X16 WORDS [Dosovitskiy et al., 2020]

THE VISION TRANSFORMER (ViT)

The **Vision Transformer** is an attention-based architecture for *image classification*, which outperformed the state-of-the-art at the time of its publication.

AN IMAGE IS WORTH 16X16 WORDS [Dosovitskiy et al., 2020]

THE VISION TRANSFORMER (ViT)

The **Vision Transformer** is an attention-based architecture for *image classification*, which outperformed the state-of-the-art at the time of its publication.

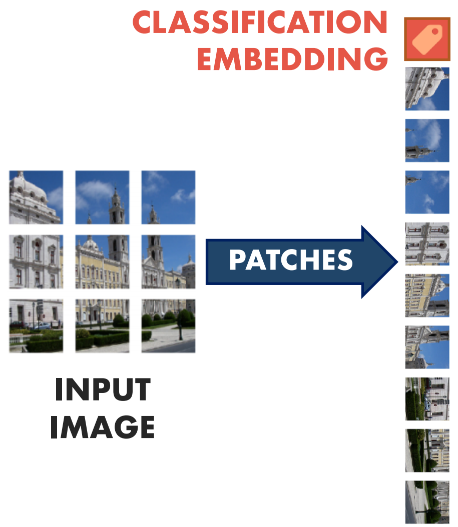
INPUT IMAGE IS SPLIT INTO PATCHES (each one is an embedding)

AN IMAGE IS WORTH 16X16 WORDS [Dosovitskiy et al., 2020]

THE VISION TRANSFORMER (ViT)

The **Vision Transformer** is an attention-based architecture for *image classification*, which outperformed the state-of-the-art at the time of its publication.

- **INPUT IMAGE IS SPLIT INTO PATCHES** (each one is an embedding)
- **ADDITIONAL EMBEDDING FOR CLASSIFICATION** (from NLP)

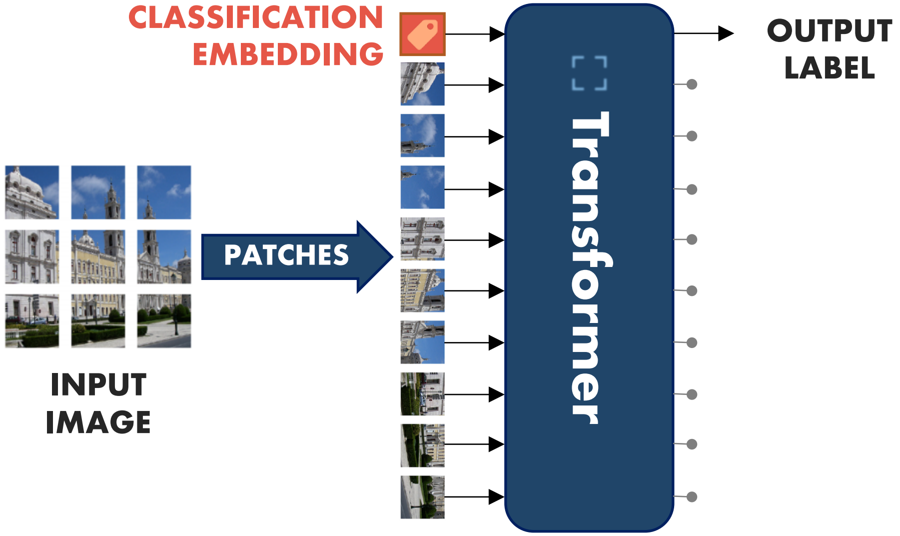


AN IMAGE IS WORTH 16X16 WORDS [Dosovitskiy et al., 2020]

THE VISION TRANSFORMER (ViT)

The **Vision Transformer** is an attention-based architecture for *image classification*, which outperformed the state-of-the-art at the time of its publication.

- **INPUT IMAGE IS SPLIT INTO PATCHES** (each one is an embedding)
- **ADDITIONAL EMBEDDING FOR CLASSIFICATION** (from NLP)
- **THE RESULTING VECTORS ARE FED INTO A TRANSFORMER**



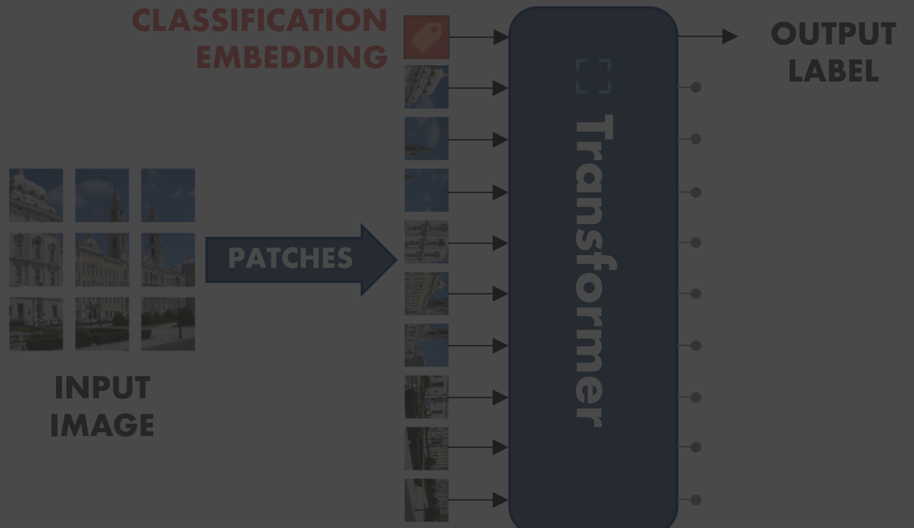
AN IMAGE IS WORTH 16X16 WORDS [Dosovitskiy et al., 2020]

THE VISION TRANSFORMER (ViT)

88.55% ACCURACY ON IMAGENET!

The **Vision Transformer** is an attention-based architecture for *image classification*, which outperformed the state-of-the-art at the time of its publication.

- **INPUT IMAGE IS SPLIT INTO PATCHES** (each one is an embedding)
- **ADDITIONAL EMBEDDING FOR CLASSIFICATION** (from NLP)
- **THE RESULTING VECTORS ARE FED INTO A TRANSFORMER**



ELEPHANT IN THE ROOM

THE PROBLEM WITH VISION TRANSFORMERS

The Vision Transformer looks very promising, but its peak performance comes at a **cost**:

ELEPHANT IN THE ROOM

THE PROBLEM WITH VISION TRANSFORMERS

The Vision Transformer looks very promising, but its peak performance comes at a **cost**:

HUGE AMOUNT OF DATA

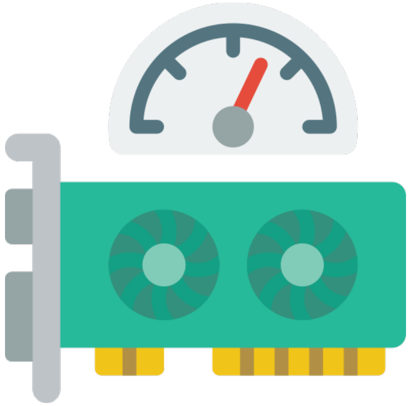
300+ million images

ELEPHANT IN THE ROOM

THE PROBLEM WITH VISION TRANSFORMERS

The Vision Transformer looks very promising, but its peak performance comes at a **cost**:

HUGE AMOUNT OF DATA
300+ million images



A LOT OF COMPUTING POWER

DINO TRAINING METHOD

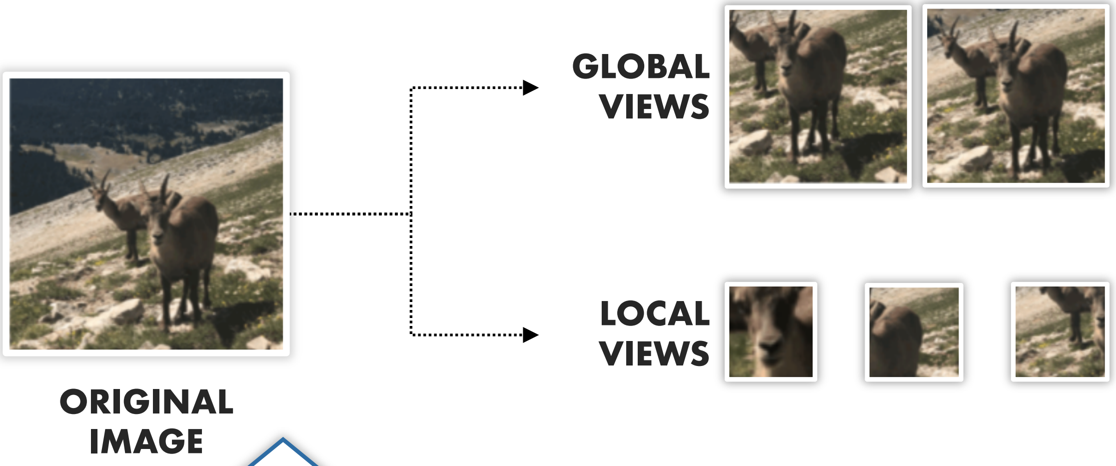
Researchers at Facebook AI, tried to solve the problem by introducing a new training method called «*Self-DIstillation with NO labels*» (DINO).

- **AD HOC DATA AUGMENTATION**
- **KNOWLEDGE DISTILLATION**
- **SELF-SUPERVISED LEARNING** (no labels)

EMERGING PROPERTIES IN SELF-SUPERVISED ViTs [Caron et al., 2021]

DINO TRAINING METHOD

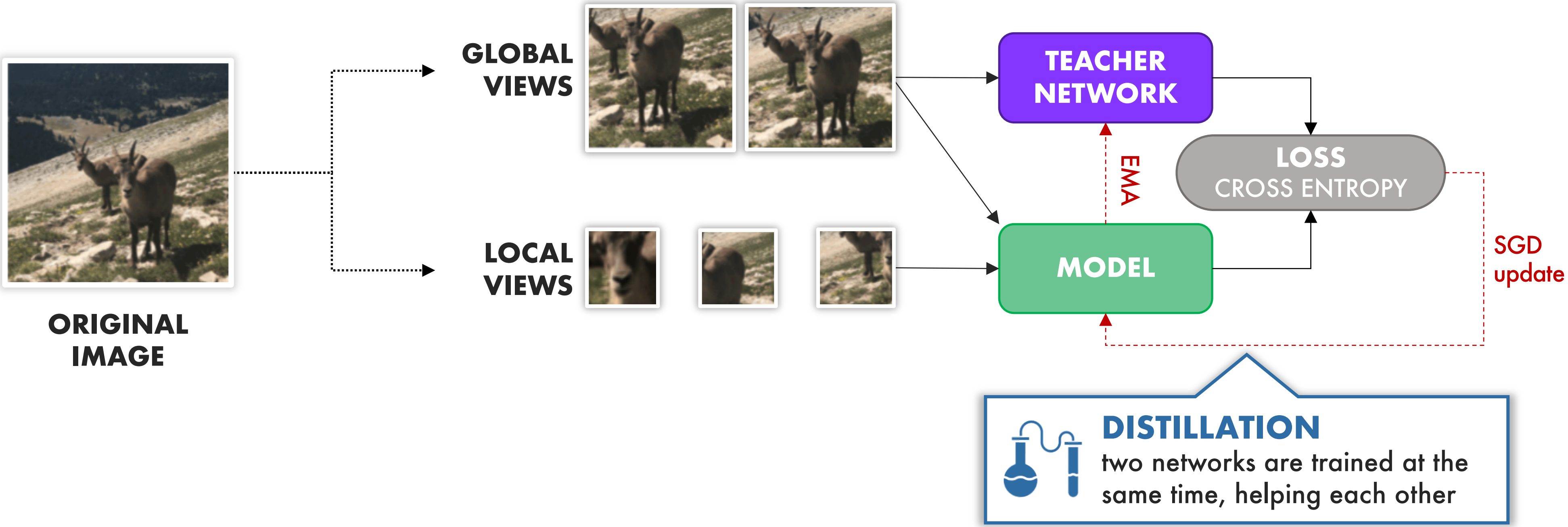
Researchers at Facebook AI, tried to solve the problem by introducing a new training method called «*Self-DIstillation with NO labels*» (DINO).



AUGMENTATION
images are cropped using two data augmentation strategies

DINO TRAINING METHOD

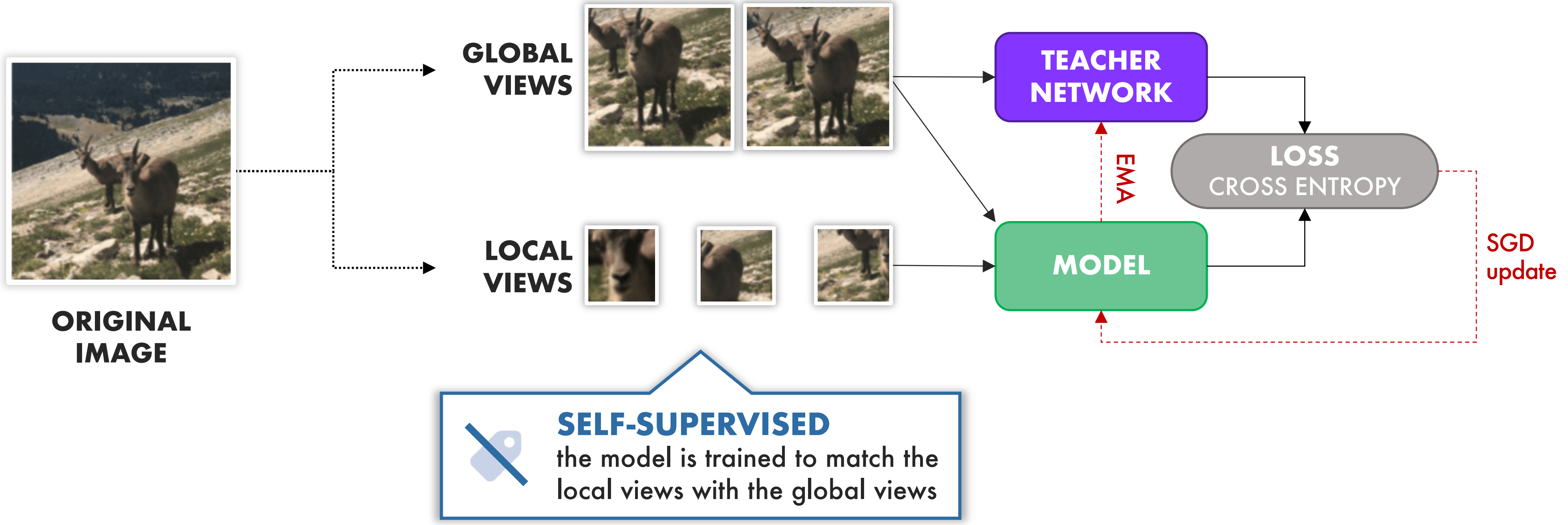
Researchers at Facebook AI, tried to solve the problem by introducing a new training method called «*Self-DIstillation with NO labels*» (DINO).



EMERGING PROPERTIES IN SELF-SUPERVISED ViTs [Caron et al., 2021]

DINO TRAINING METHOD

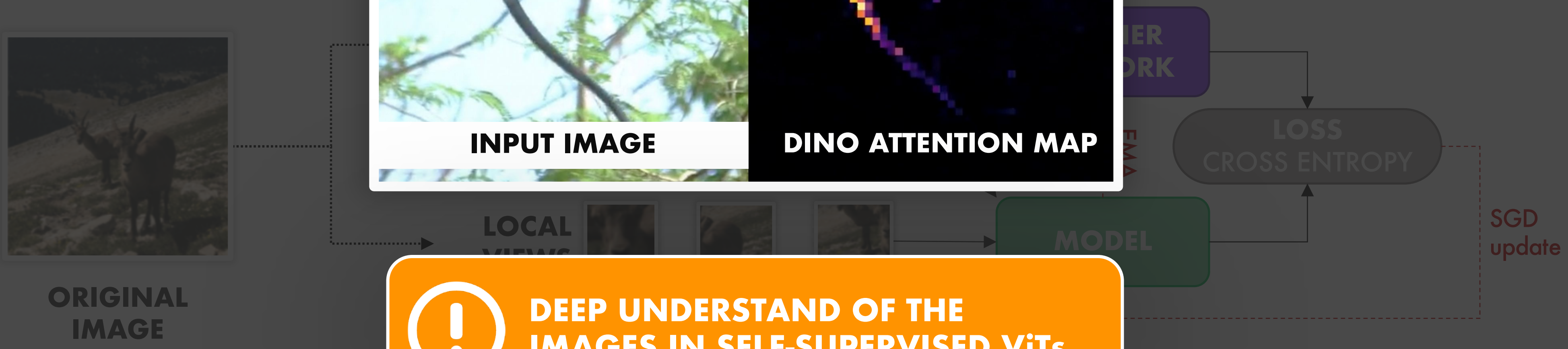
Researchers at Facebook AI, tried to solve the problem by introducing a new training method called «*Self-DIstillation with NO labels*» (DINO).



EMERGING PROPERTIES IN SELF-SUPERVISED ViTs [Caron et al., 2021]

DINO TRAINING METHOD

Researchers at Facebook AI, … w training method called «*Self-DIstillation with NO labels*» (…



DEEP UNDERSTAND OF THE IMAGES IN SELF-SUPERVISED ViTs.

WHAT ABOUT THE DATA?

The authors of DINO relied on a moderately big data set to achieve such results.

1.2 MILLION IMAGES

better than 300 Million, but still too much

AN ANOMALY DETECTION SYSTEM BASED ON THE DINO FEATURES

ANOMALY DETECTION WITH DINO

The idea is to build an anomaly detection system using a DINO pre-trained ViT **Features Extractor** (we called those features *DINO Features*).

ANOMALY DETECTION DATASET

DINO PRE-TRAINED VIT

AN ANOMALY DETECTION SYSTEM BASED ON THE DINO FEATURES

ANOMALY DETECTION WITH DINO

The idea is to build an anomaly detection system using a DINO pre-trained ViT **Features Extractor** (we called those features *DINO Features*).

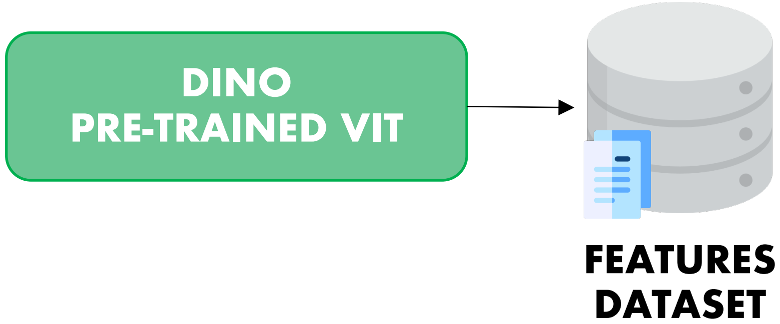
AN ANOMALY DETECTION SYSTEM BASED ON THE DINO FEATURES

ANOMALY DETECTION WITH DINO

The idea is to build an anomaly detection system using a DINO pre-trained ViT **Features Extractor** (we called those features *DINO Features*).

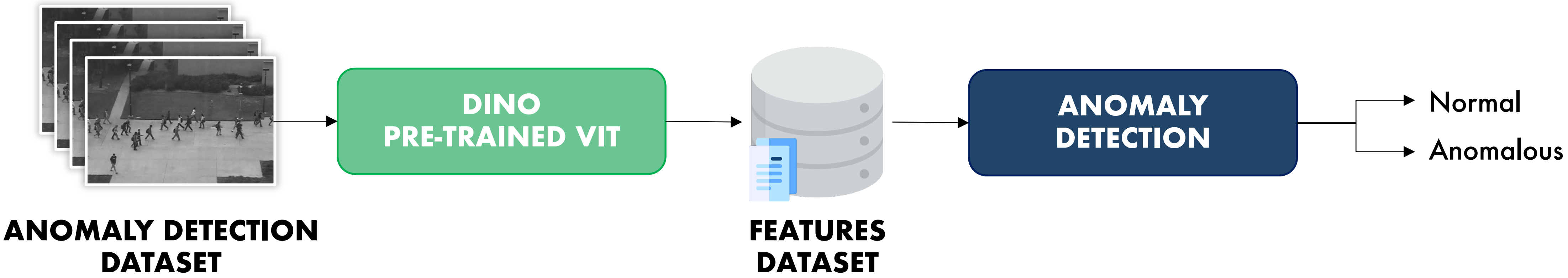
ANOMALY DETECTION WITH DINO

The idea is to build an anomaly detection system using a DINO pre-trained ViT **Features Extractor** (we called those features *DINO Features*).



ANOMALY DETECTION WITH DINO

The idea is to build an anomaly detection system using a DINO pre-trained ViT **Features Extractor** (we called those features *DINO Features*).



UCSD PEDESTRIAN DATASET

In this thesis, we considered the UCSD Pedestrian Dataset **PED2**.

- Stationary videos of pedestrian walkways;
- Anomalies are due to the circulation of *non-pedestrian entities*;
- Split into a *training set* (only normal videos) and a *testing set* (contains anomalies).

ANOMALY DETECTION IN CROWDED SCENES [Mahadevan et al., 2010]

UCSD PEDESTRIAN DATASET

In this thesis, … Datase… nd PED2.

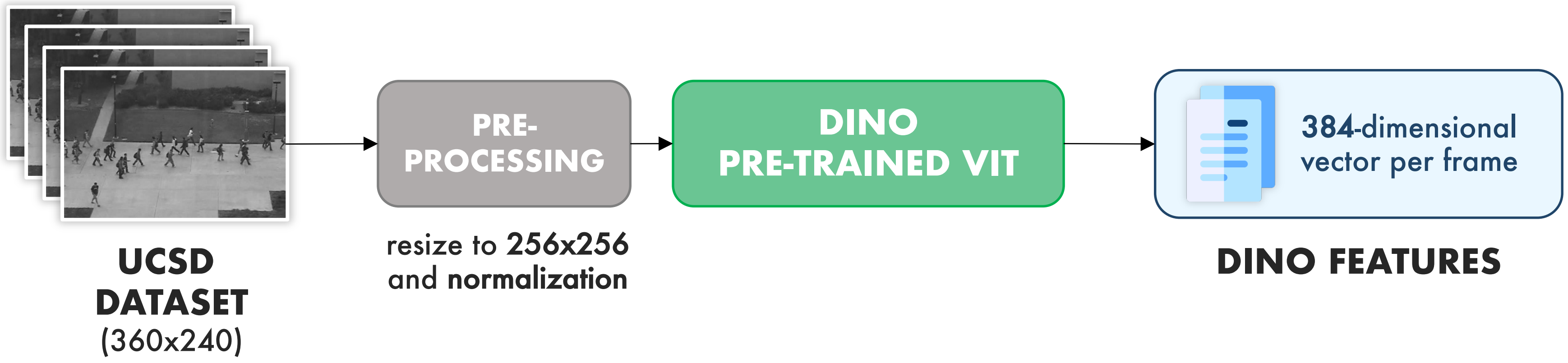
- Stationar…
- Anomalie… *estrian*…
- Both split…) and a…

NORMAL

ANOMALOUS

FROM UCSD PEDESTRIAN TO FEATURES SPACE

THE DINO FEATURES



FROM UCSD PEDESTRIAN TO FEATURES SPACE

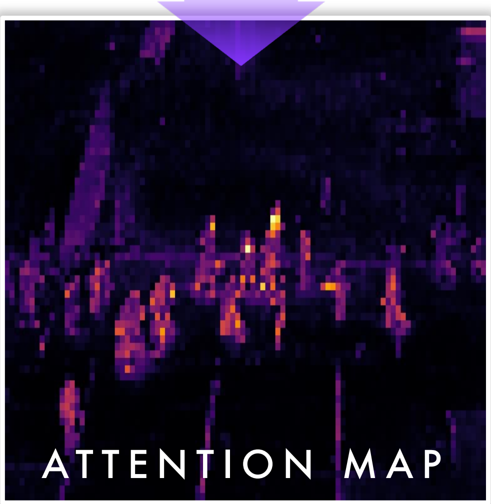
THE DINO FEATURES

UCSD DATASET
(360x240)

PRE-PROCESSING

resize to **256x256** and **normalization**

DINO PRE-TRAINED VIT



384-dimensional vector per frame

DINO FEATURES

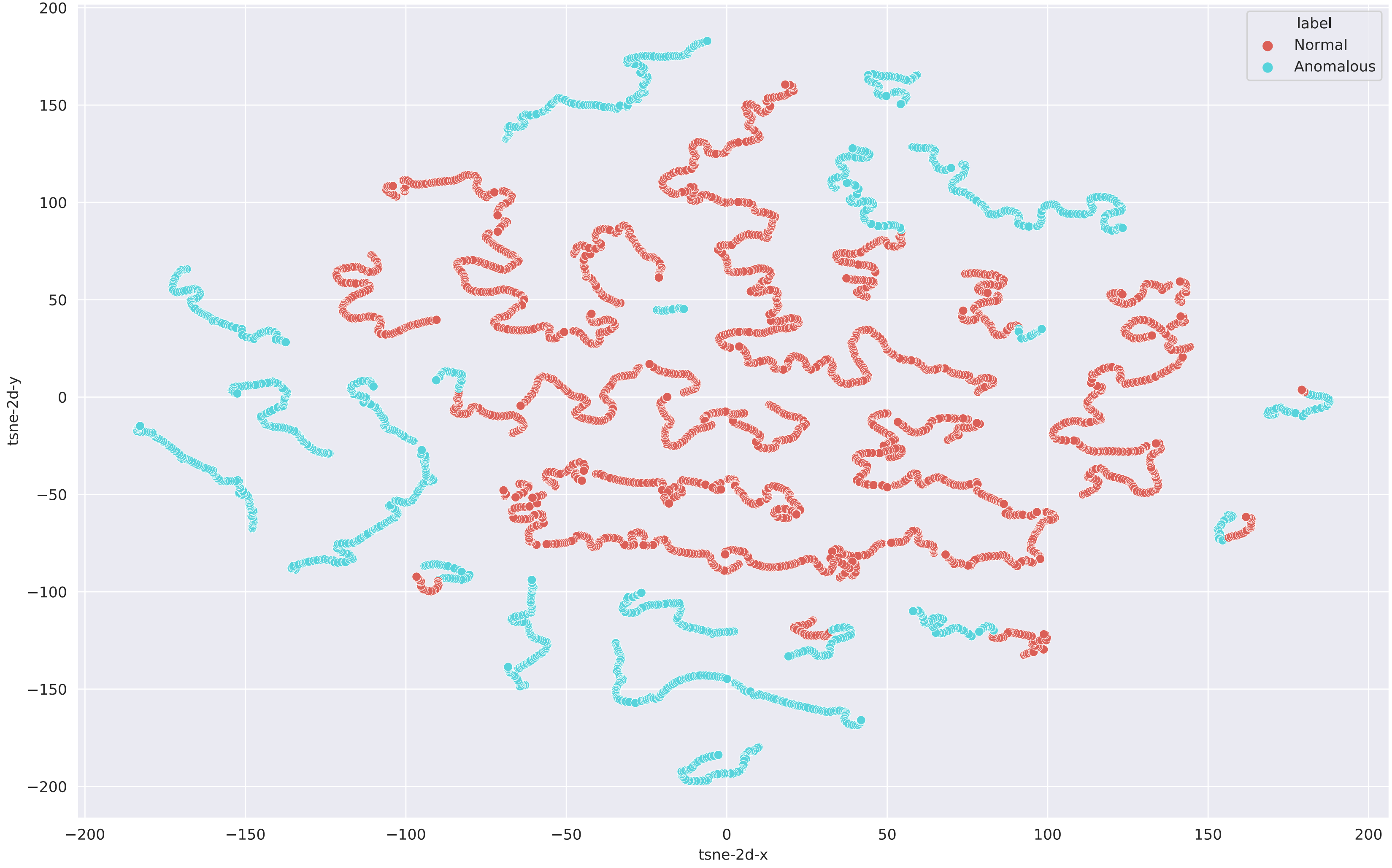
ARE THEY ANY GOOD?

ANALYSIS OF THE DINO FEATURES

Several strategies were tried to test the quality of the DINO Features. Here are reported the results from the *t-SNE* analysis perfomed on every single features vectors (one features vector per frame) and on their *aggregation* (one features vector per video).

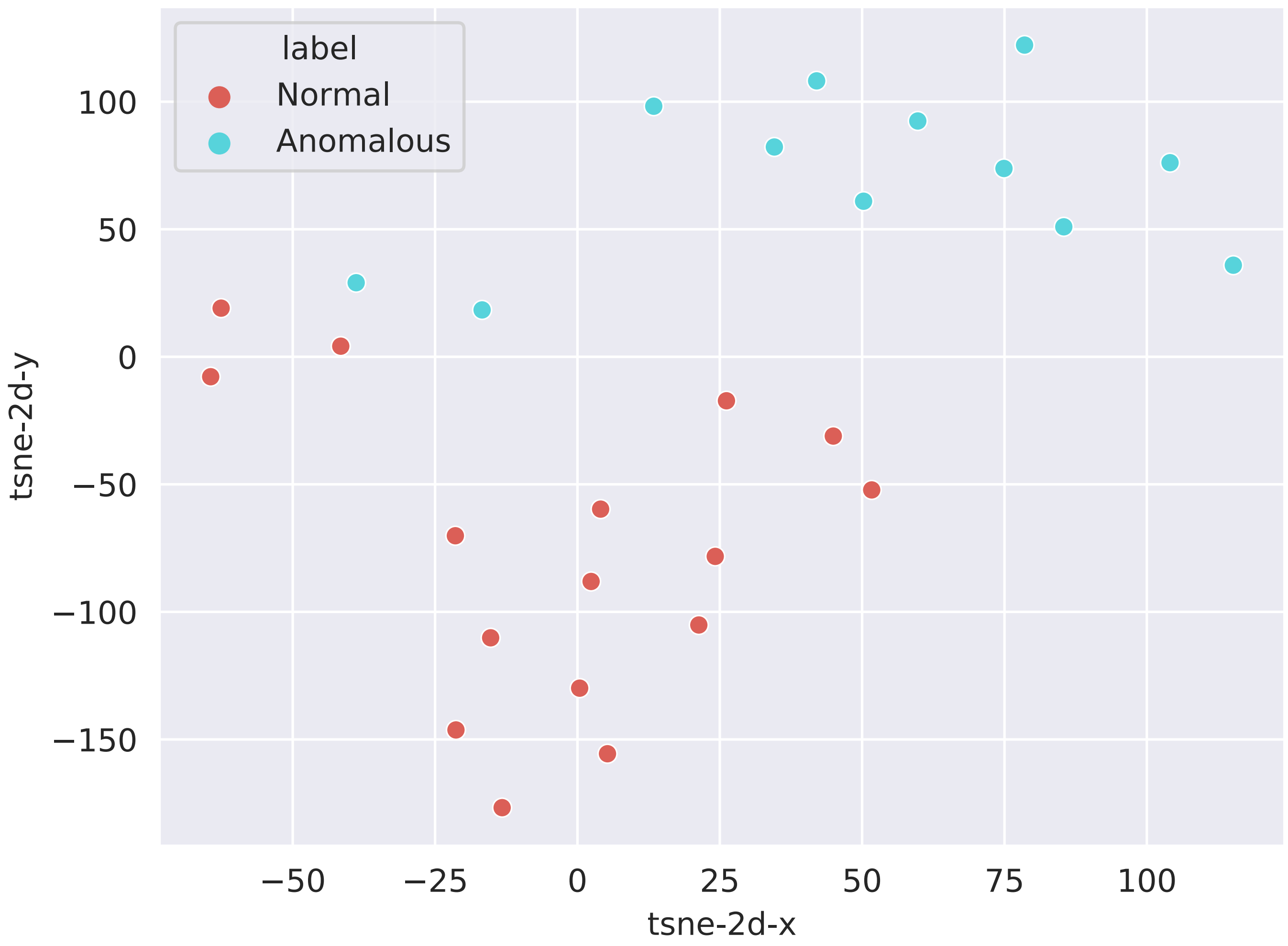
SIDE NOTE: the t-distributed Stochastic Neighbor Embedding (**t-SNE**) is a statistical visual tool for visualizing high-dimensional data by giving to each point a location in a two-dimensional map.

T-SNE ON ALL THE FEATURES VECTORS FROM PED2



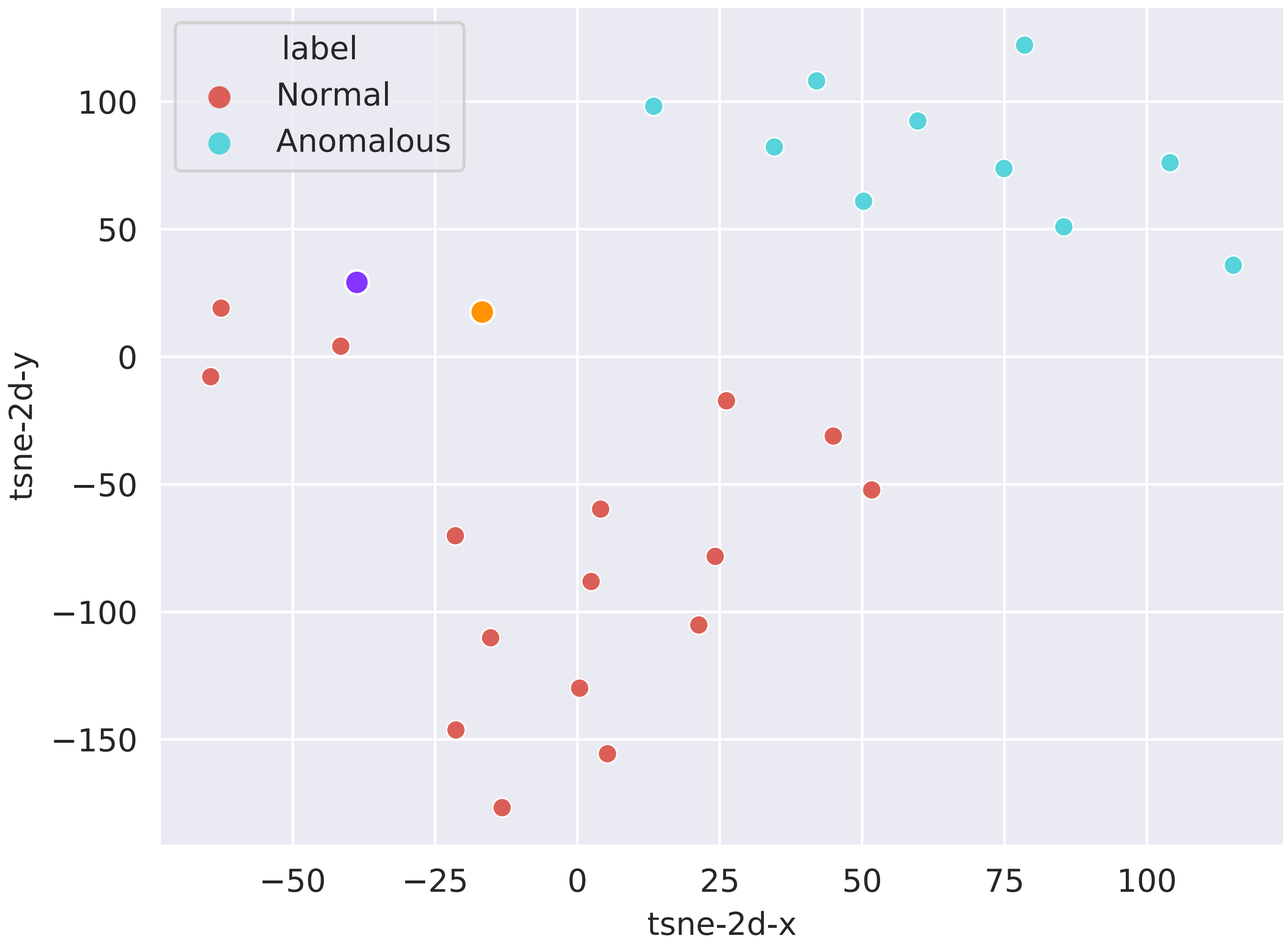
The t-SNE was performed by setting `numiter=5000` and `perplexity=10`.

T-SNE ON AGGREGATED (MAX) FEATURES VECTORS FROM PED2



The t-SNE was performed by setting `numiter=5000` and `perplexity=10` the features vectors were aggregated using the `max`.

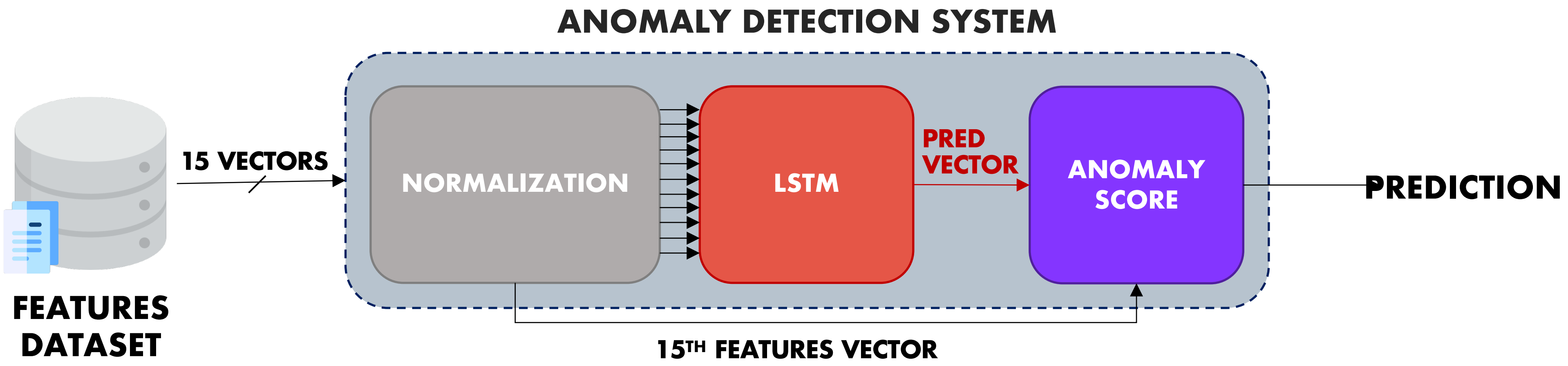
T-SNE ON AGGREGATED (MAX) FEATURES VECTORS FROM PED2



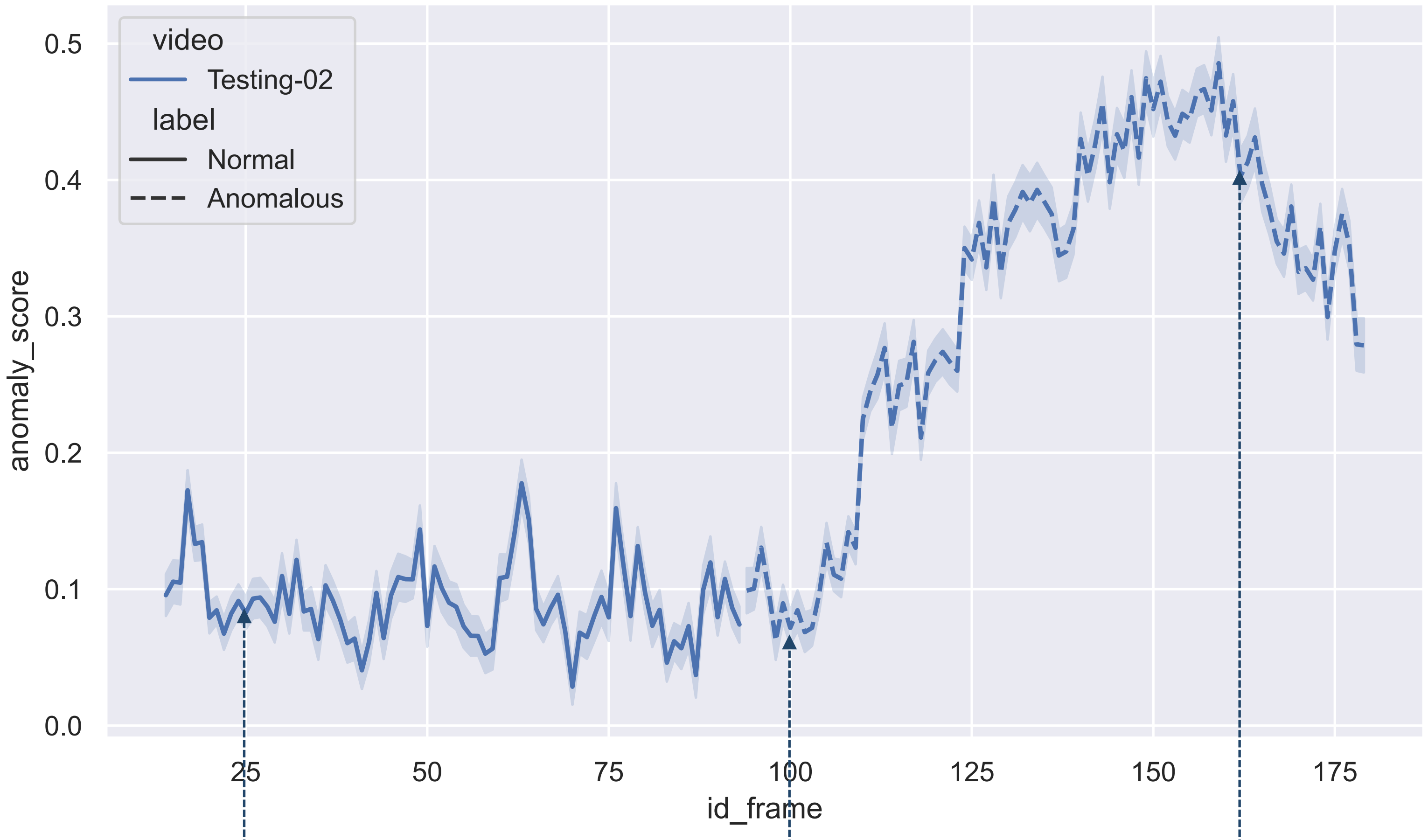
The t-SNE was performed by setting `numiter=5000` and `perplexity=10` the features vectors were aggregated using the `max`.

DETECTING ANOMALIES BY PREDICTING THE DINO FEATURES

THE ANOMALY DETECTION SYSTEM

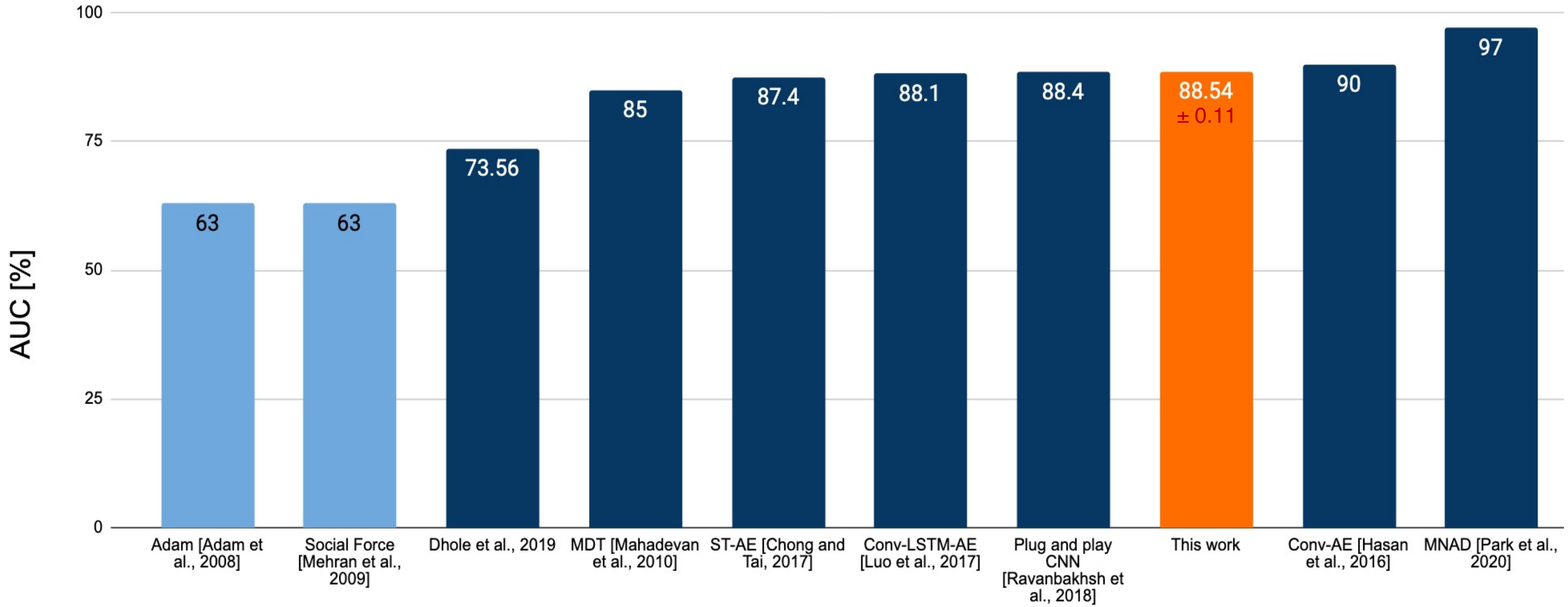


AN EXAMPLE OF PREDICTION FROM PED2



The **anomaly score** was obtained by normalizing the distance between the predicted features vector from the *LSTM* and the actual features vector. The error is expressed using the *standard deviation* over 60 repetitions.

COMPARISON WITH LITERATURE, AUC SCORE ON PED2



DON'T USE CONVOLUTIONAL NEURAL NETWORKS

THIS WORK WITH 95% CI [60 REPETITIONS]

SOURCE: A SURVEY OF SINGLE-SCENE VIDEO ANOMALY DETECTION [Ramachandra, 2020]

CONCLUSIONS

Convolutional Neural Networks (CNNs) are still the best option for visual anomaly detection, however, it is important to remind that Vision Transformers are new technology (**2020**), whereas CNNs have been widely adopted since the early 2010s.

The biggest limitations today is the necessity for **high computational power** and **tons of data examples**.

The proposed architecture could be **extended** in the future, *e.g. by replacing the LSTM with a Transformer that works with aggregation of features vector to predict the next vector*.

THANK YOU,

QUESTIONS?

APPENDIX

BIBLIOGRAPHY

1. **Attention is All You Need** – *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. CoRR, abs/1706.03762, 2017*
2. **An image is worth 16x16 words** – *Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. CoRR, abs/2010.11929, 2020*
3. **Emerging properties in self-supervised vision transformers** – *Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, 2021*
4. **Anomaly detection in crowded scenes** – *Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010*

BIBLIOGRAPHY

5. **Learning Memory-guided Normality for Anomaly Detection** – *Hyunjong Park, Jongyoun Noh, Bumsub Ham. DBLP:journals/corr/abs-2003-13228, 2020*
6. **A Survey of Single-Scene Video Anomaly Detection** – *Bharathkumar Ramachandra, Michael J. Jones, Ranga Raju Vatsavai. DBLP:journals/corr/abs-2004-05993, 2020*

ADDITIONAL SLIDES

WHAT IS COMPUTER VISION?

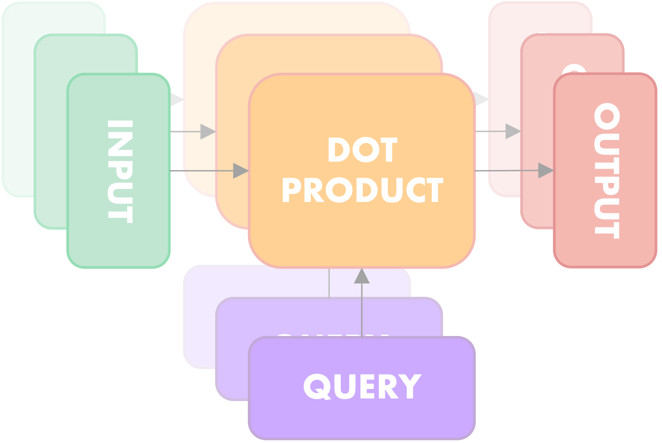
Introduced in the 1960s, **Computer Vision** is an interdisciplinary scientific field that aims to understand and reproduce tasks own of the *human visual system*.

- **OBJECT DETECTION**
- **IMAGE CLASSIFICATION**
- **IMAGE CAPTIONING**

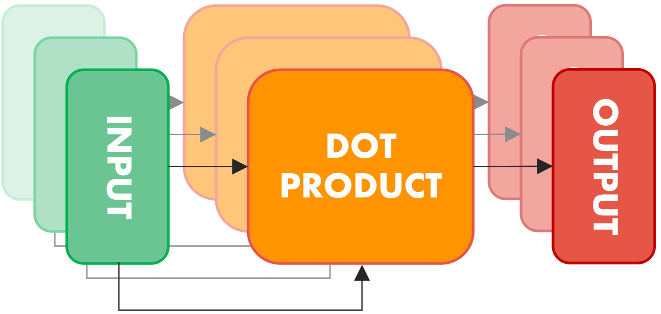
ATTENTION IN PRACTICE

… BUT WHAT ACTUALLY IS ATTENTION?

Mathematically speaking, the Attention Mechanism is just a series of *dot products* of the input sequence.



BASIC ATTENTION
w.r.t. a given query



SELF-ATTENTION
w.r.t. itself

ATTENTION IS ALL YOU NEED [Vaswani et al., 2017]

THE TRANSFORMER (1/2)

The **Transformer** is a neural network architecture that relies solely on *self-attention* to perform *Natural Language Processing (NLP)* tasks. In particular, it was proposed for language-translation.

TRADITIONAL APPROACHES (e.g. Recurrent Neural Networks)

«The quick brown fox jumps over the lazy dog»

Importance comes from closeness

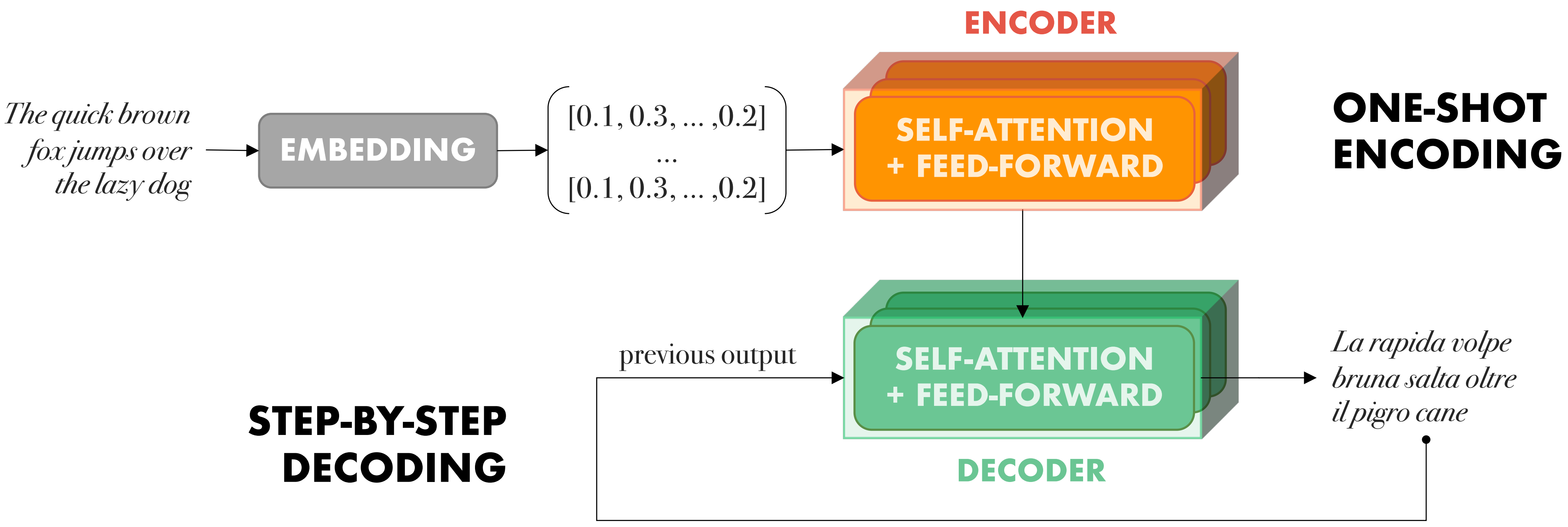
SELF-ATTENTION

«The quick brown fox jumps over the lazy dog»

Importance w.r.t. whole sentence

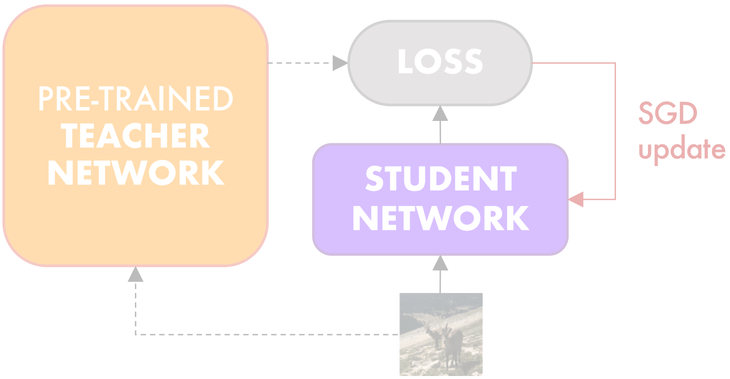
ATTENTION IS ALL YOU NEED [Vaswani et al., 2017]

THE TRANSFORMER (2/2)

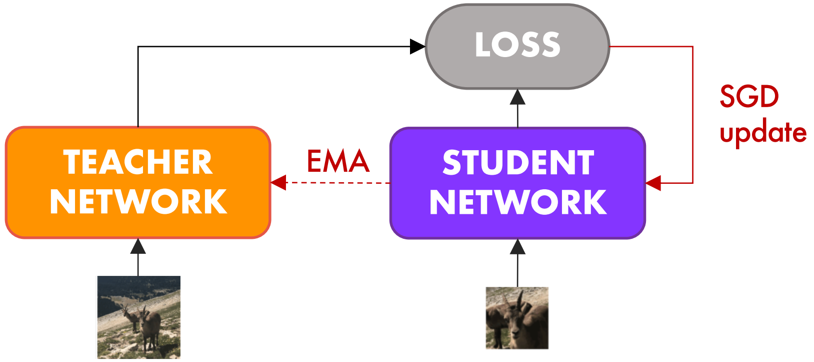


KNOWLEDGE DISTILLATION IN DINO

In general *distillation* is used to transfer the knowledge from a *teacher network* to a *student network*. In DINO both the network share the **same architecture** and they are **trained at the same time**.



TRADITIONAL



DINO APPROACH
codistillation

WHAT ABOUT THE DATA?

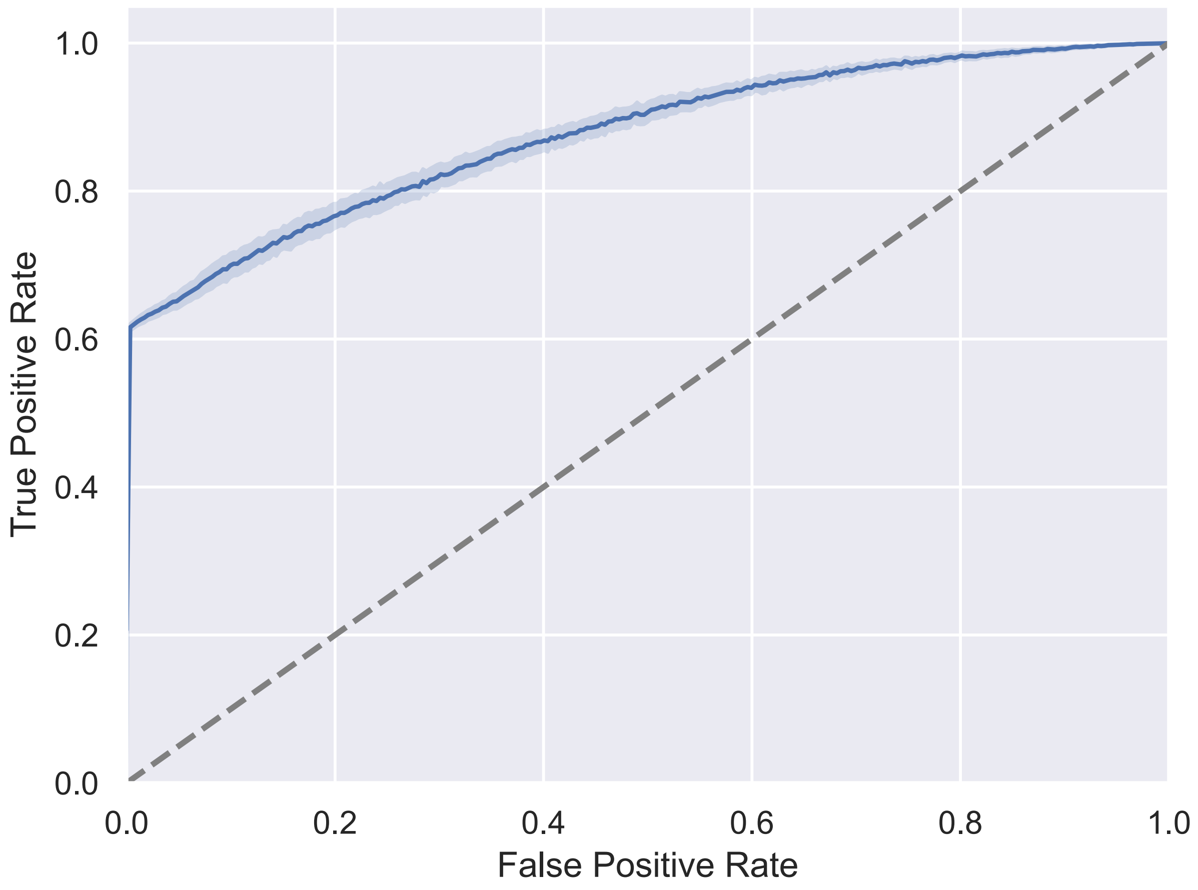
Let's find out how much data the authors of DINO required to achieve such results.

1.2 MILLION IMAGES

k-NN on **IMAGENET** with
80.1% ACCURACY

DETECTING ANOMALIES ON PED2 BY PREDICTING THE DINO FEATURES

ROC CURVE



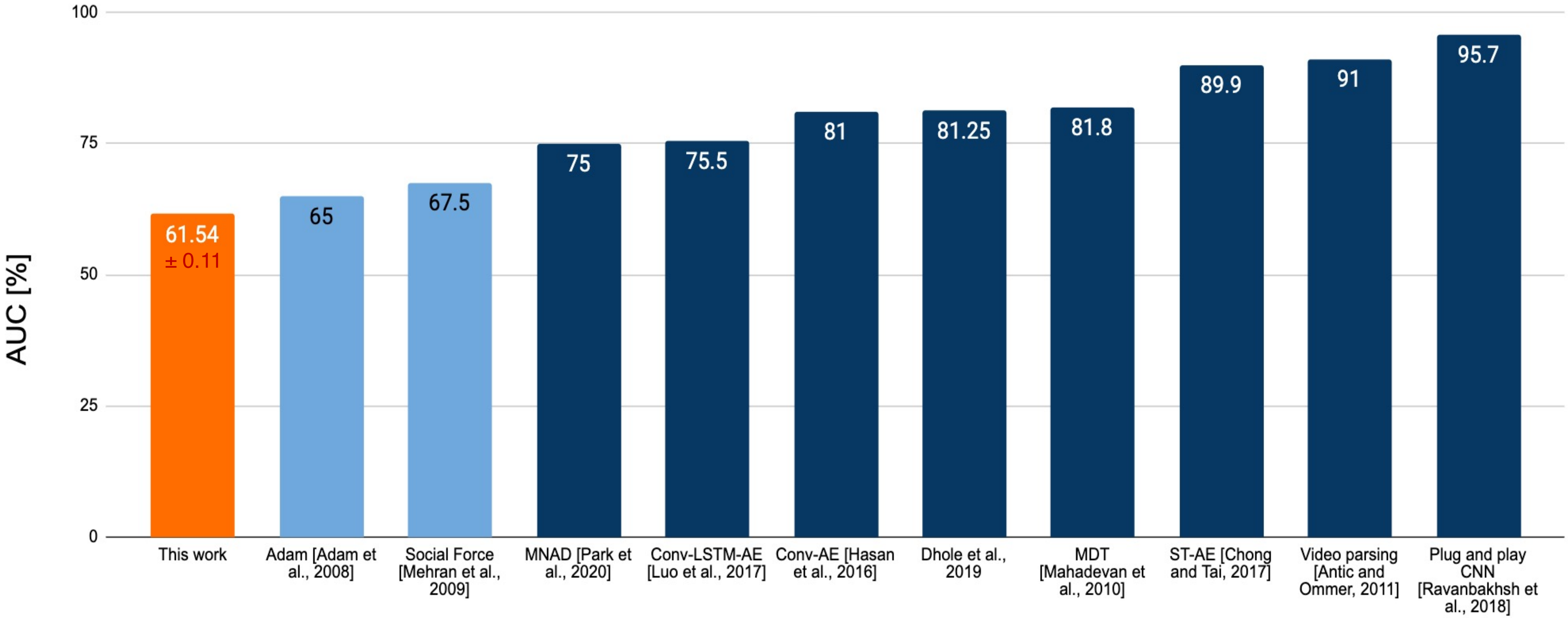
The average ROC curve was obtained by considering the mean curve from the set of ROCs obtained over 60 repetitions. The error is expressed using the *standard deviation*.

A SURVEY OF SINGLE-SCENE VIDEO ANOMALY DETECTION [Ramachandra, 2020]

LITERATURE COMPARISON

METHOD	CONVOLUTIONAL	PED1, AUC (%)	PED2, AUC (%)
Adam [Adam et al., **2008**]	✗	65.00	63.00
Social Force [Mehran et al., **2009**]	✗	67.50	63.00
MDT [Mahadevan et al., **2010**]	✗	81.80	85.00
Video parsing [Antic and Ommer, **2011**]	✗	91.00	92.00
Conv-AE [Hasan et al., **2016**]	✓	81.00	90.00
CAE(FR) [Sabokroua et al., **2017**]	✓	–	81.04
Nazare et al., **2018**	✓	64.06	88.93
Future frame prediciton [Liu et al., **2018**]	✓	83.10	95.40
Plug and play CNN [Ravanbakhsh et al., **2018**]	✓	95.70	88.40
Siamese distance learning [Ramachandra et al., **2020**]	✓	86.00	94.00
MNAD [Park et al., **2020**]	✓	–	**97.00**

COMPARISON WITH LITERATURE, AUC SCORE ON PED1



DON’T USE CONVOLUTIONAL NEURAL NETWORKS

THIS WORK WITH 95% CI [60 REPETITIONS]

SOURCE: «A SURVEY OF SINGLE-SCENE VIDEO ANOMALY DETECTION» [Ramachandra, 2020]

DETECTING ANOMALIES BY PREDICTING THE DINO FEATURES

COMPARISON WITH A RESNET-50

Our proposal failed to achieve the state-of-the-art performance in terms of AUC score. However, when compared with an analogous approach that relies on **ResNet-50** (*convolutional neural network*) to extract the features, the difference is quite impressive.

Dataset	DINO Features AUC [%]	ResNet-50 Features AUC [%]
PED2	**88.54** ± 0.22	70.98 ± 0.26

MEAN AUC WITH 95% CI (over 60 repetitions)