IPTC-Amazon collaboration

Meeting

(June 15th, 2023)

POLITÉCNICA

iptc

www.iptc.upm.es

IPTC-Amazon: Agenda

Agenda:

- Projects:
  - Sign language motion generation from high level sign characteristics
  - Speaker diarization with multimodal inputs
  - Pose and spatial movement as input for dynamic content search & generation
  - Entangling AI-audio synthesis models and multimodal representations
  - Zero-shot sonorizing of video sequences
- Dissemination

Sign language motion generation and analysis

iptc

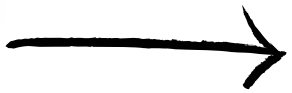
POLITÉCNICA

Index

- Our Objective
- What have we done?
- Previous Results & Errors
- The Solutions
- Evaluation Metrics
- Our Results

Our Objective

```
<hamnosys_manual>
        <hamfist/>
        <hambetween/>
        <hamfist/>
        <hamthumbacrossmod/>
        <hamextfingeru/>
        <hampalmd/>
        <hamshoulders/>
        <hamlrat/>
</hamnosys_manual>
```



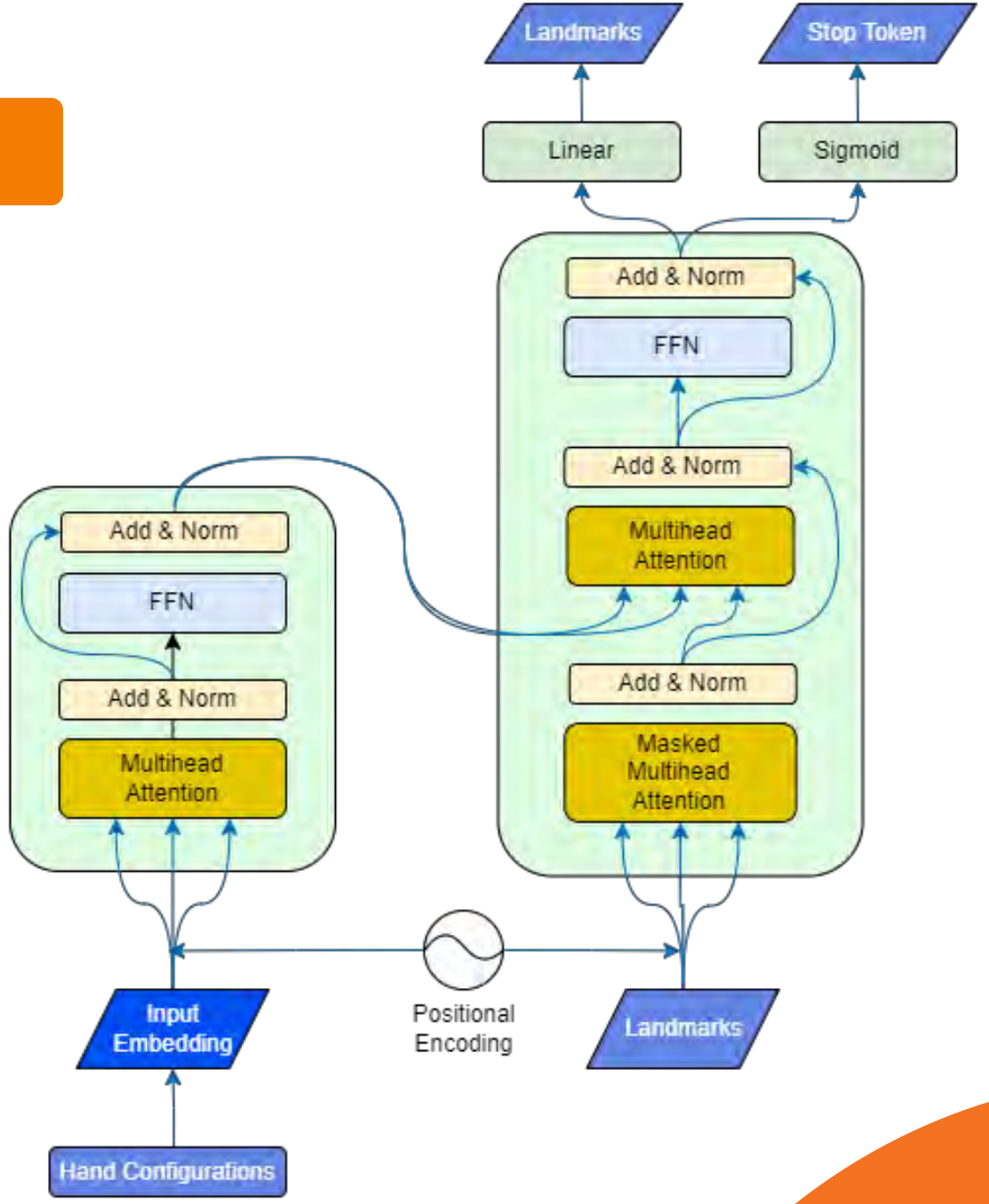
Hand configurations

Landmarks

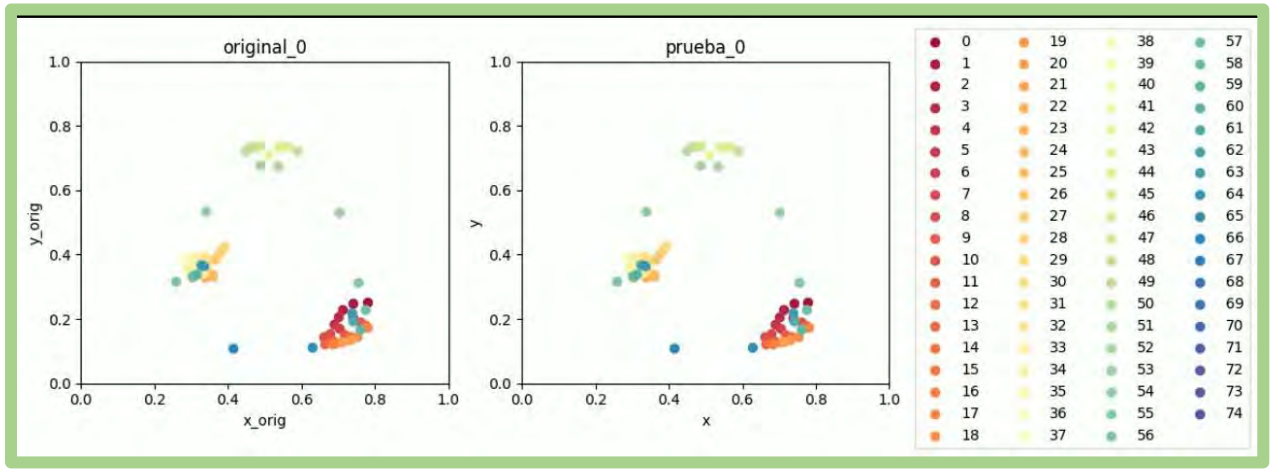
iptc

Transformer

- Based on a spanish to english translator transformer



Previous Results & Errors

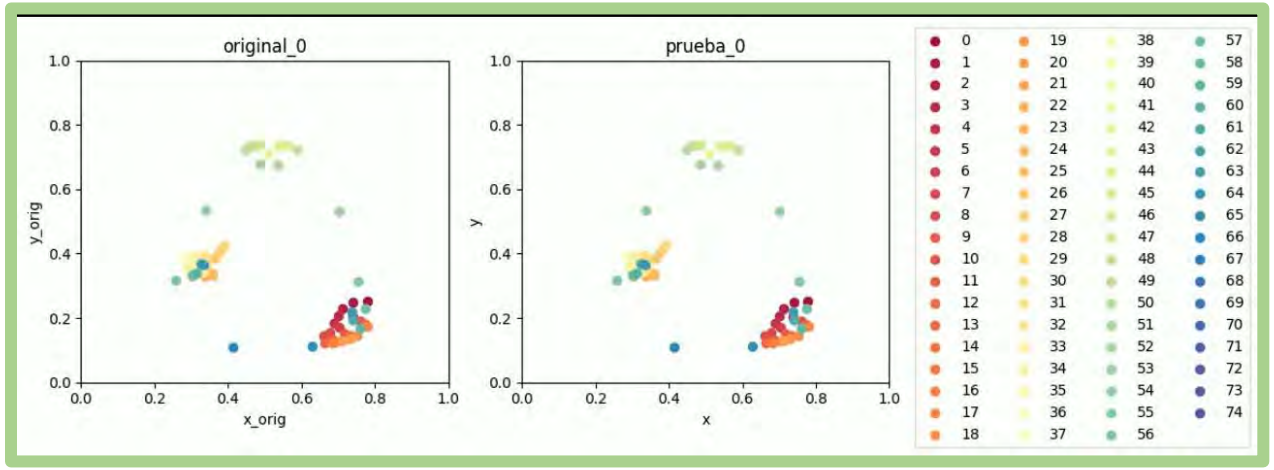


Database

- Filling with zeros
- Mediapipe and the mirror effect
- Interpolation

iptc

Previous Results & Errors

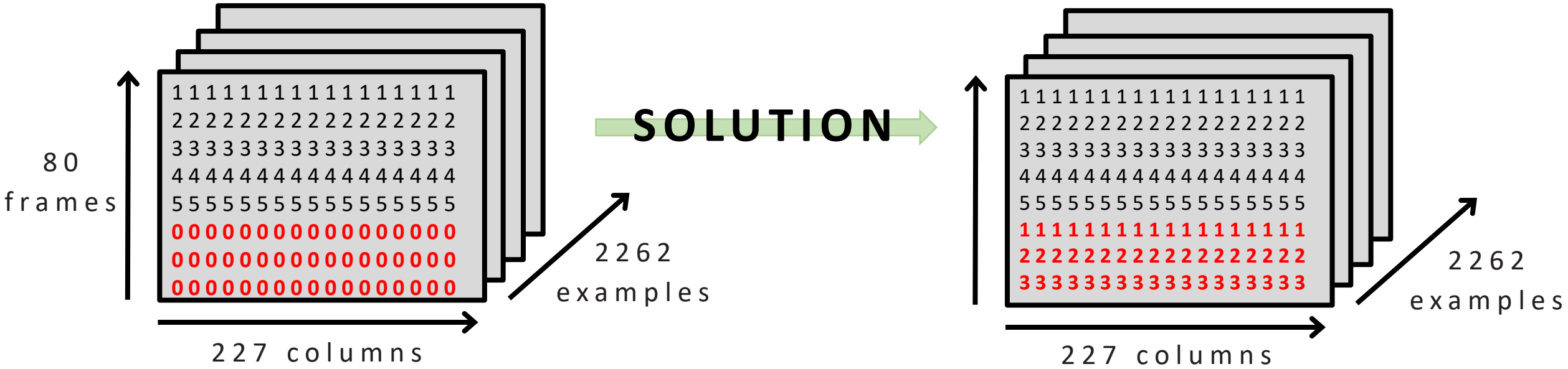


Transformer performance

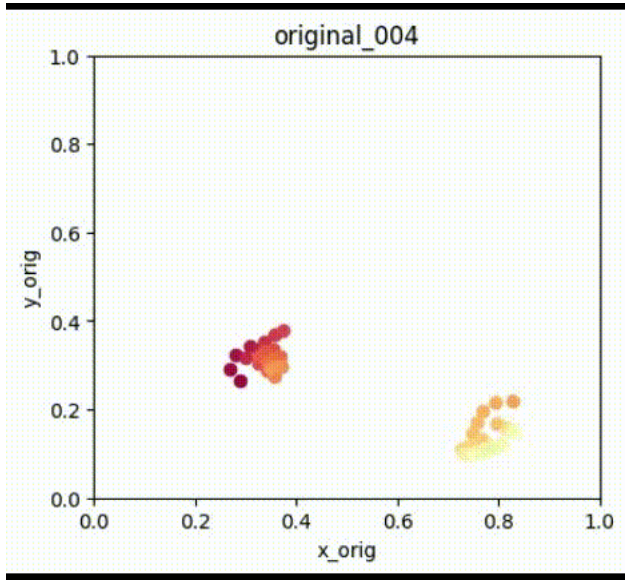
- Loss function
- The Optimizer
- The Dropout
- Batch size , number of epochs & number of heads
- Intermediate dimension
- The stop token

iptc

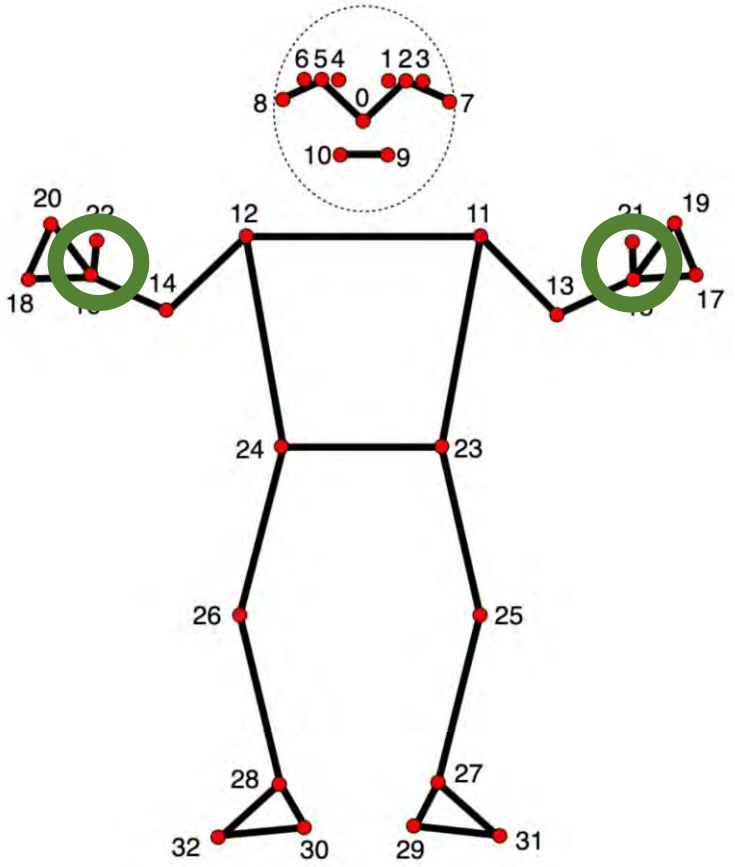
DATABASE
Filling with zeros



DATABASE
Mirror Effect: Mediapipe

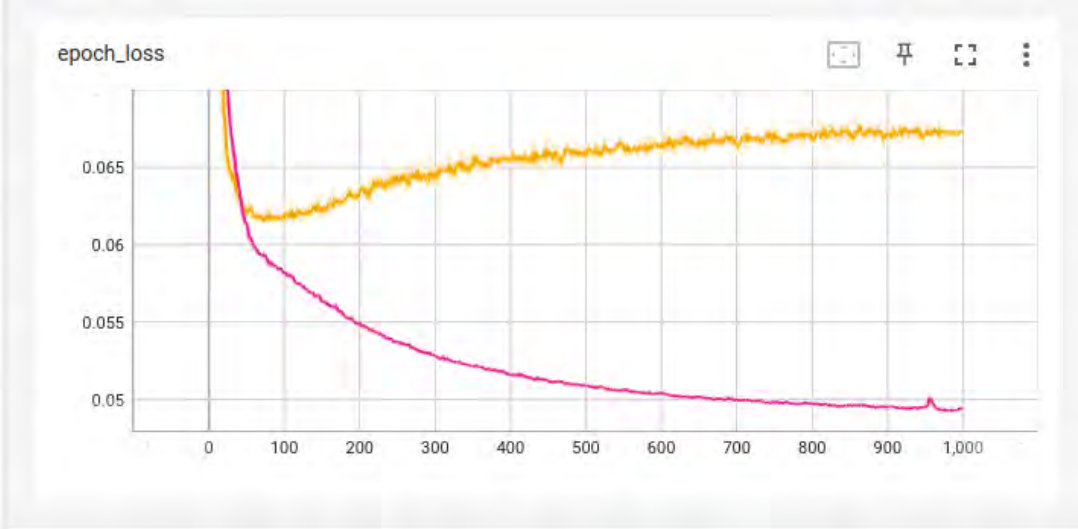


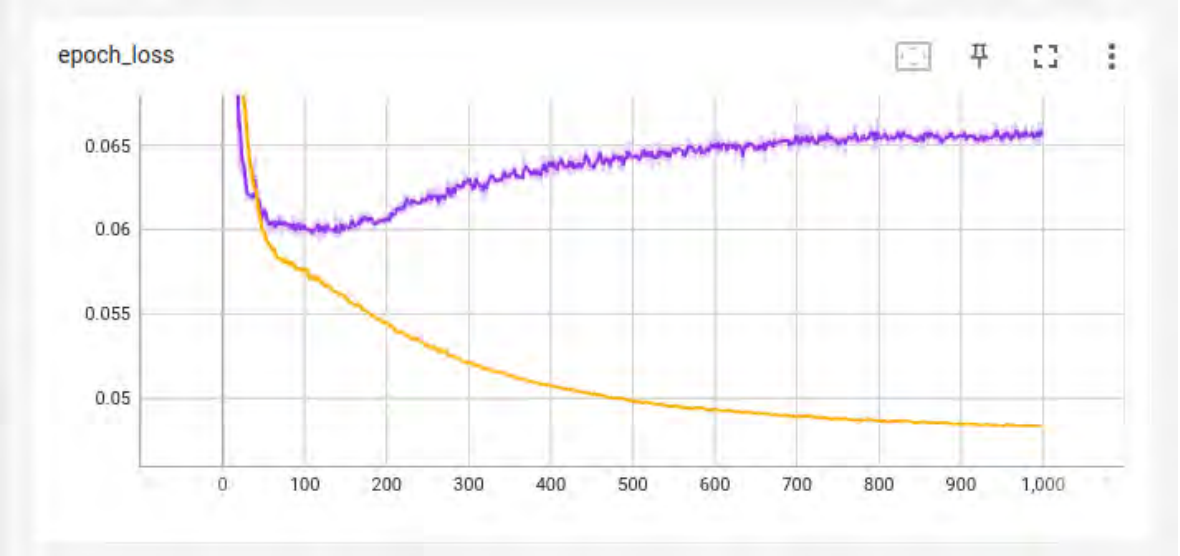
SOLUTION



DATABASE
Interpolation

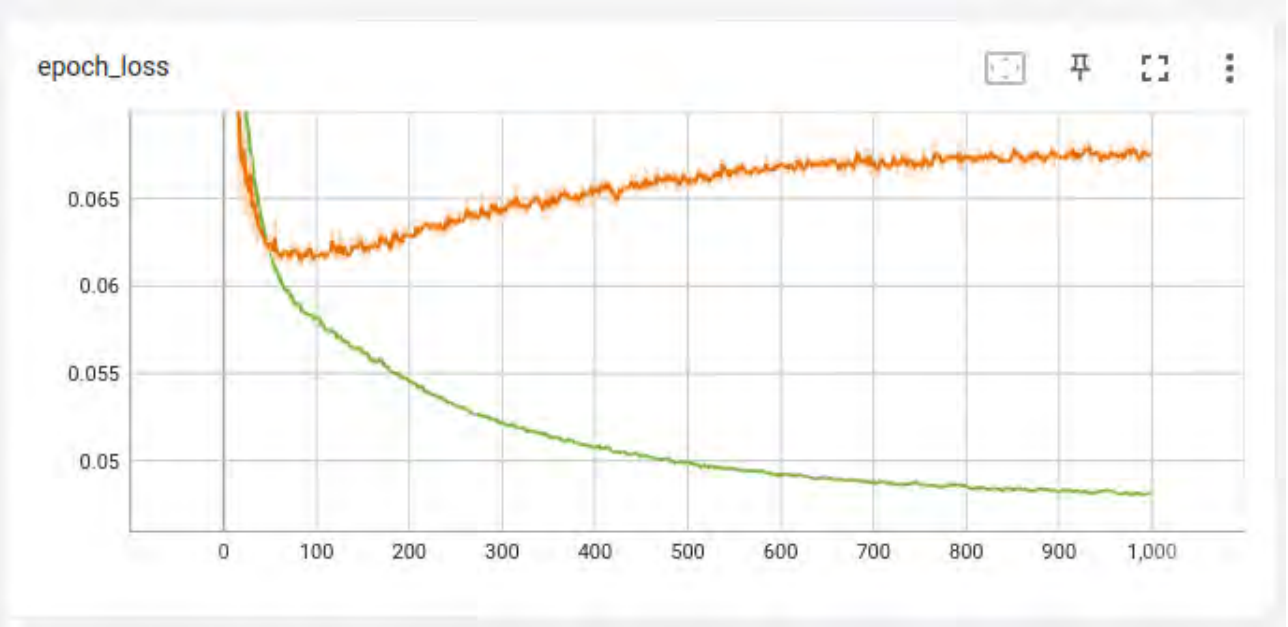
NO INTERPOLATION VS ONE ROW OF INTERPOLATION





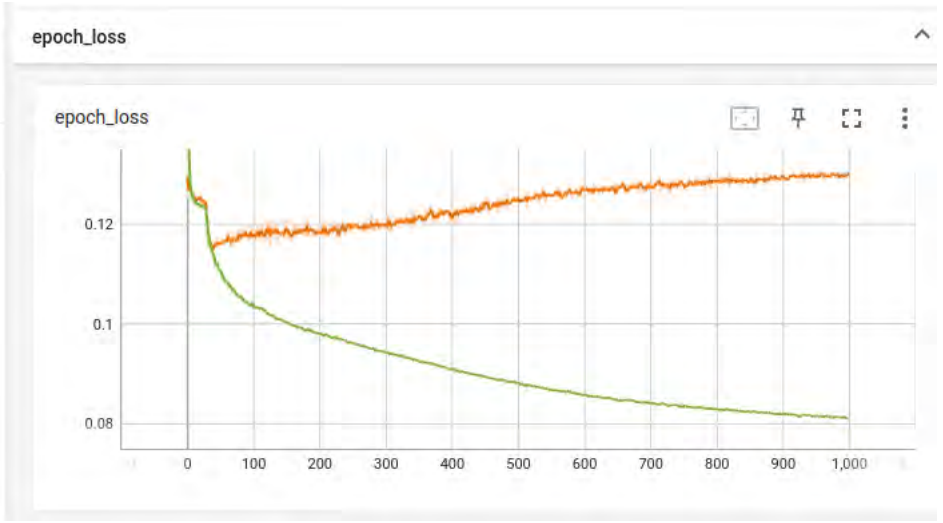
TRANSFORMER PERFORMANCE
Changing the parameters

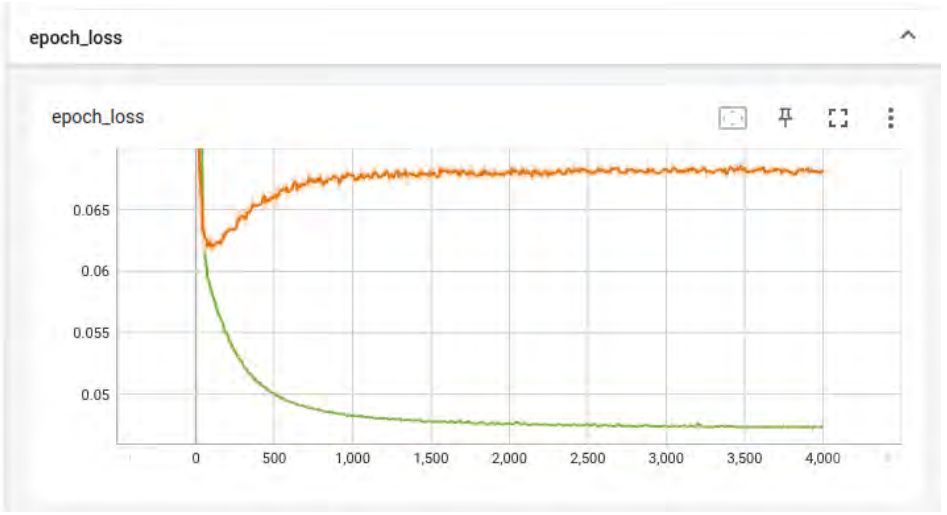
- Batch Size → 32
- Number of Epochs → 1000
- Number of Heads → 6
- Intermediate Dimension → 4096
- DropOut → 0.3
- Learning Rate → 0.001



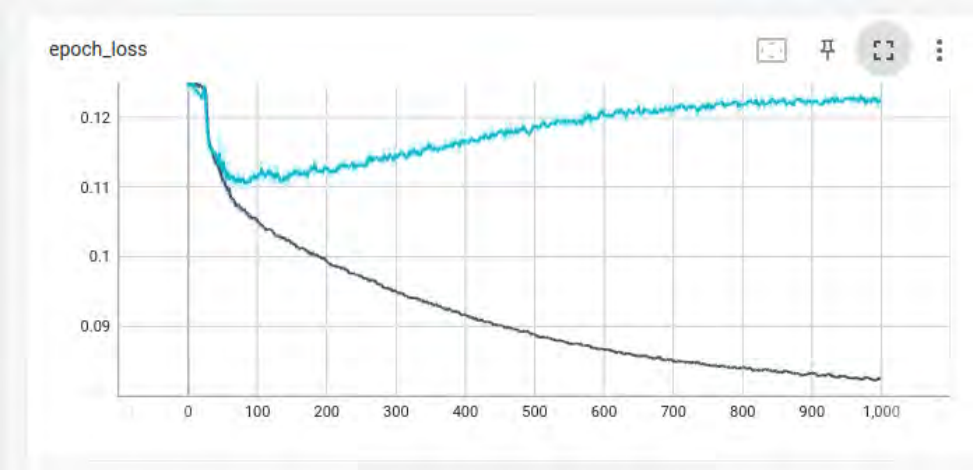
TRANSFORMER PERFORMANCE
Changing the parameters

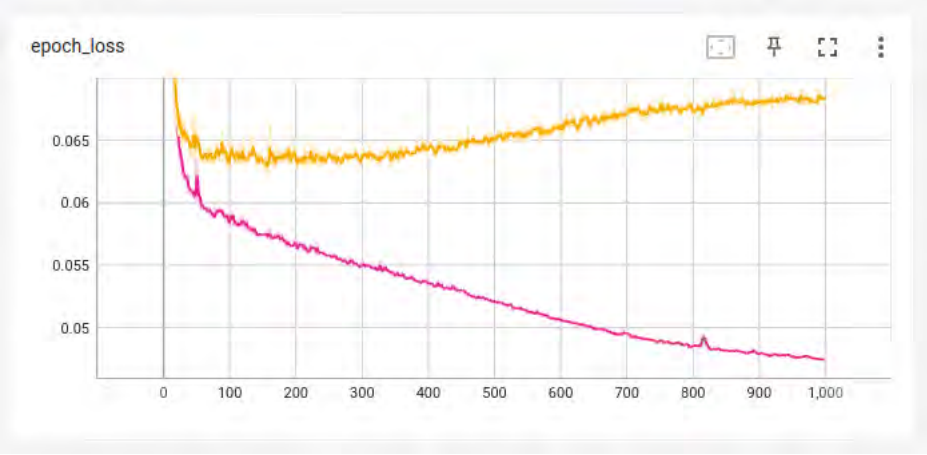
DROPOUT 0.0 VS DROPOUT 0.3



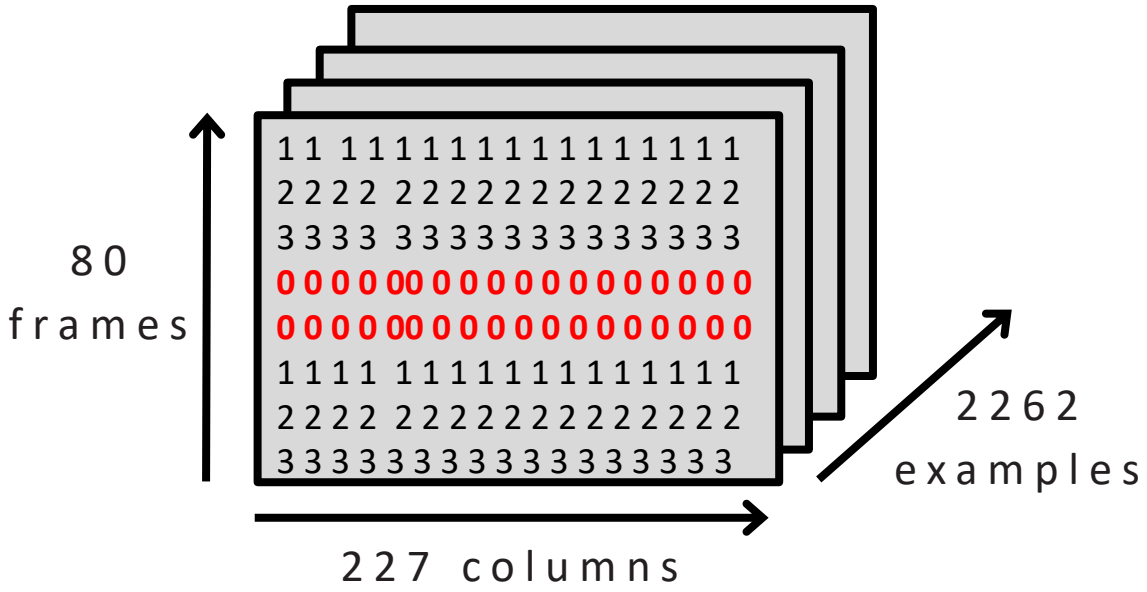


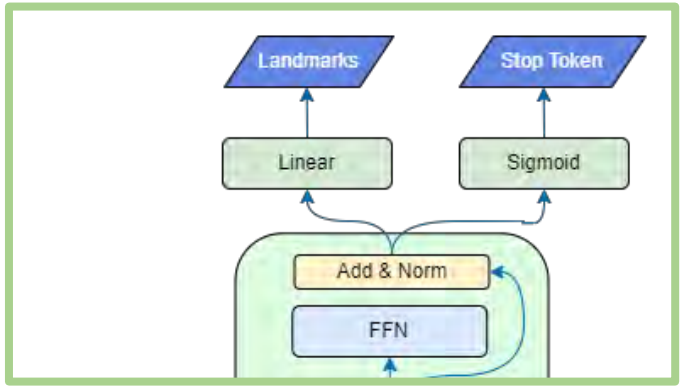
INTERMEDIATE DIM 2048 VS INTERMEDIATE DIM 4096





TRANSFORMER PERFORMANCE
The stop token





- Intercalate ceros between frames
- Separate training for stop token and generation of landmarks

EVALUATION METRICS

DTW

DTW → Test Dataset

DTW → Validation Dataset

DTW → Mean Value

VIDEOS

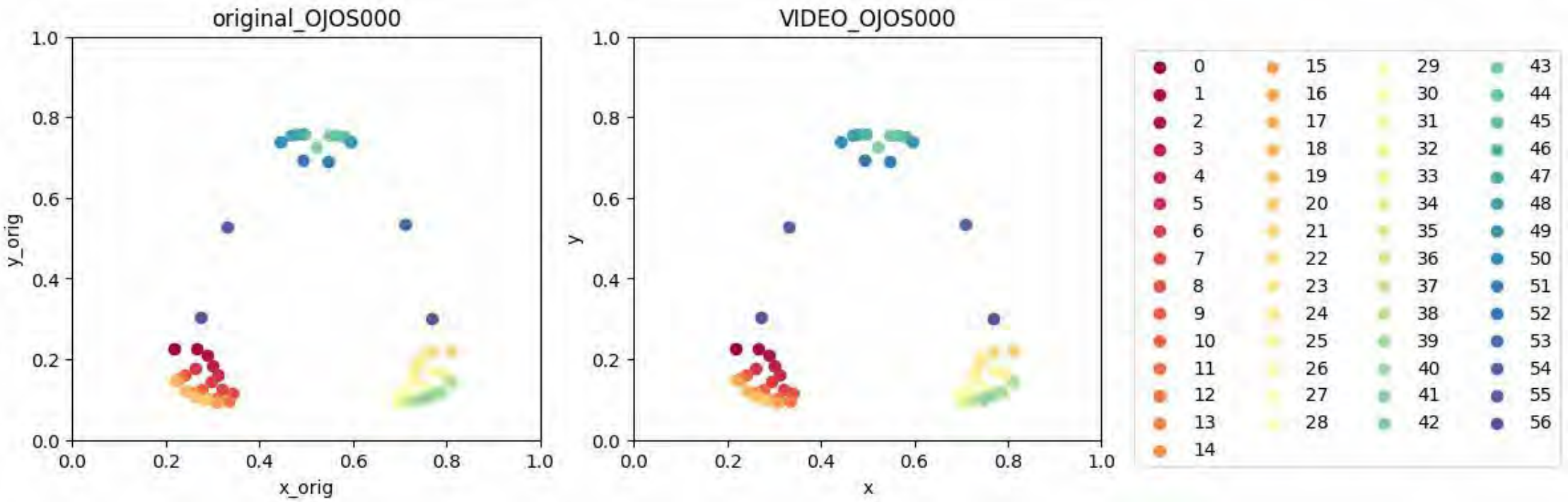
Generation of landmarks

Generation of videos

Comparing original vs predicted

iptc

RESULTS



MEAN DTW → 7,112

iptc

Future Lines

1. Transformer
   - Data Augmentation
   - The Stop Token
   - Evaluation Metrics
   - Analysis of the parameters
2. Publish the results
   - Creation of the database
   - Transformer and its performance

Speaker diarization with multimodal inputs

iptc

POLITÉCNICA

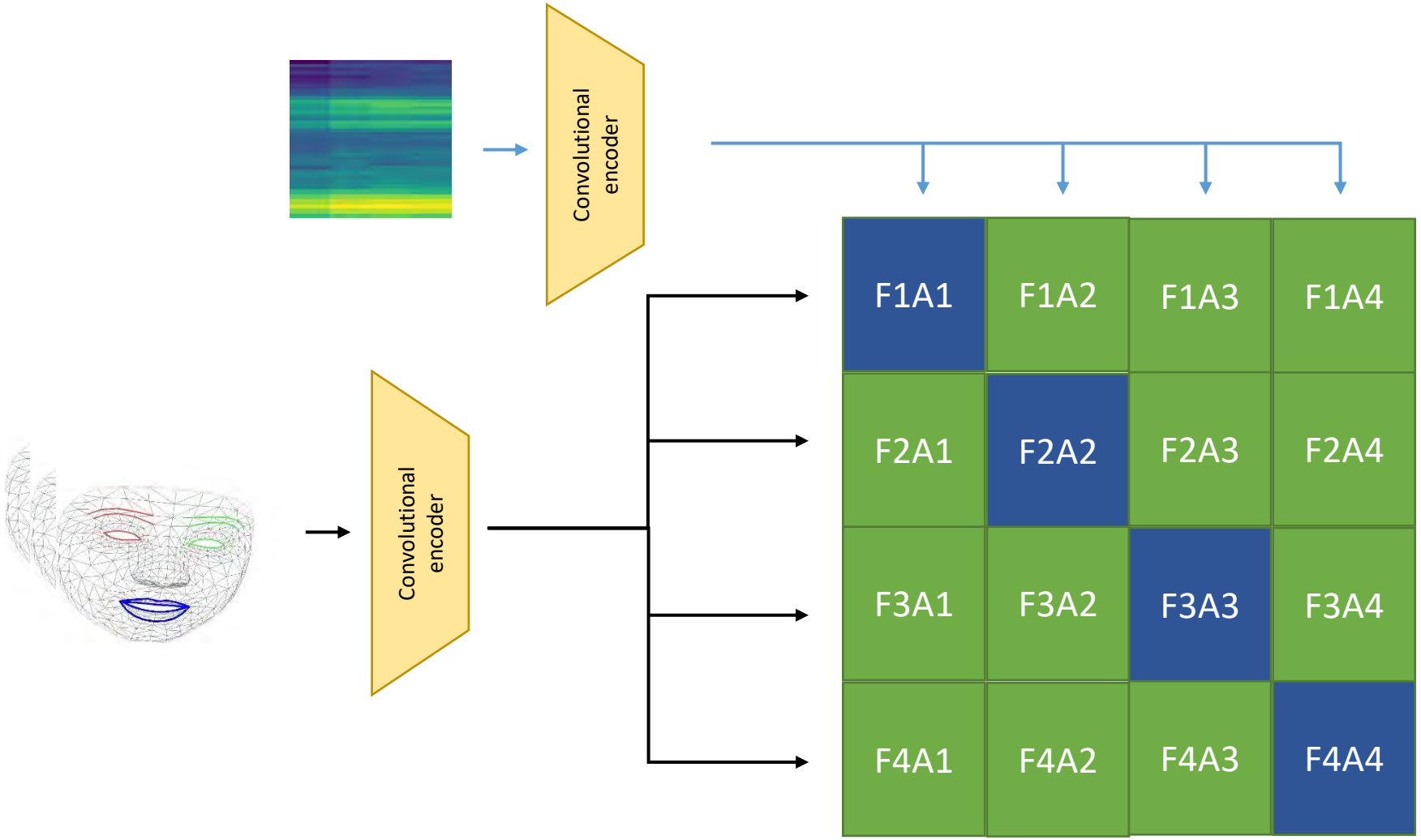
IPTC-Amazon: Index

Index:

- Previous approach to the problem
- Updated approach
- Results
- Improvements and future lines
- Conclussions

Approach to the problem

Initial approach



Approach to the problem

Initial approach

Problems found:

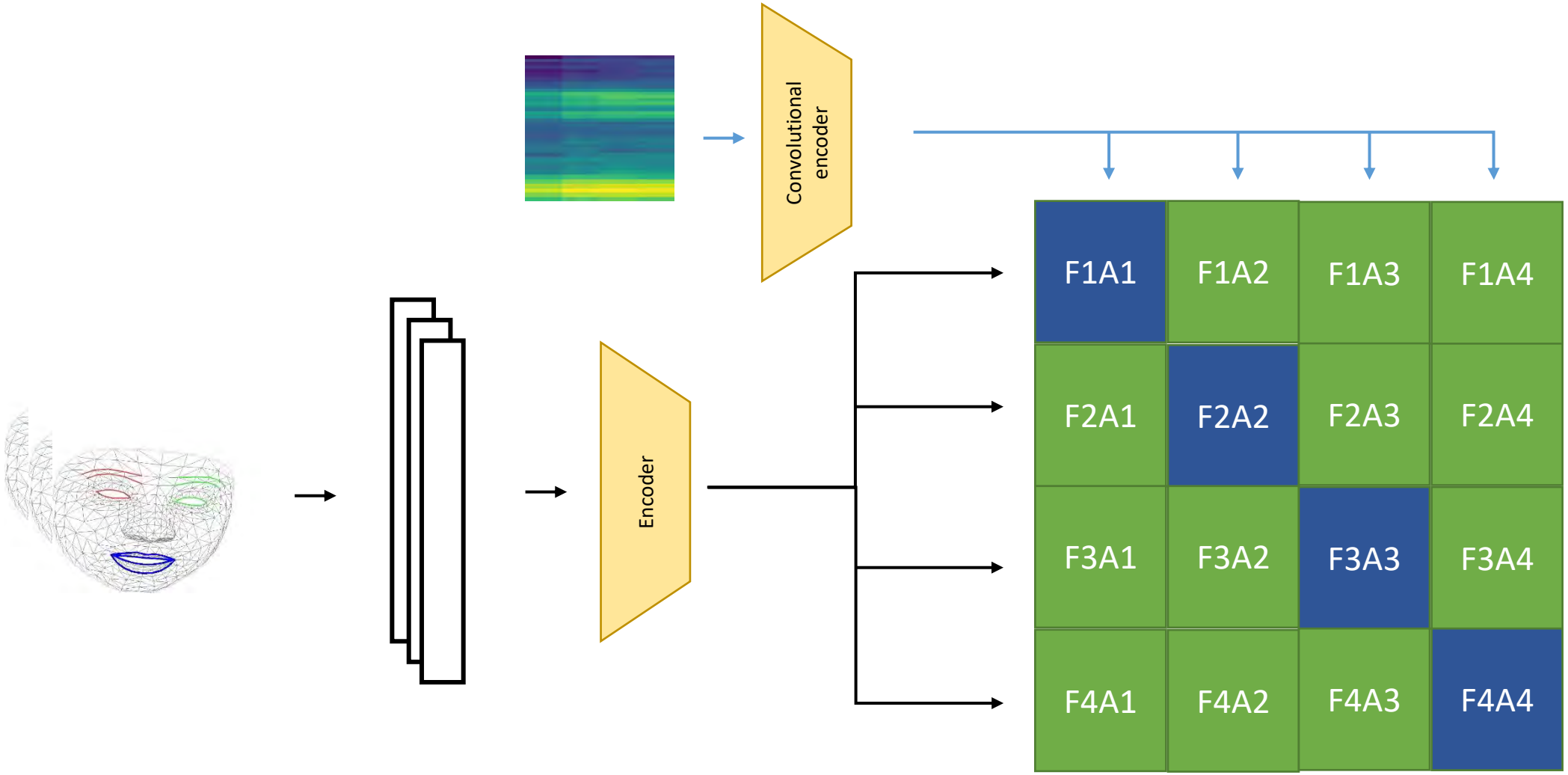
- Using **convolutional networks** for the mediapipe-generated images was a bad idea due to the high similarity between all of them.

Solutions proposed:

- Use **graph convolutional networks**
- Use directly the **landmarks** provided by mediapipe

Approach to the problem

Actual approach

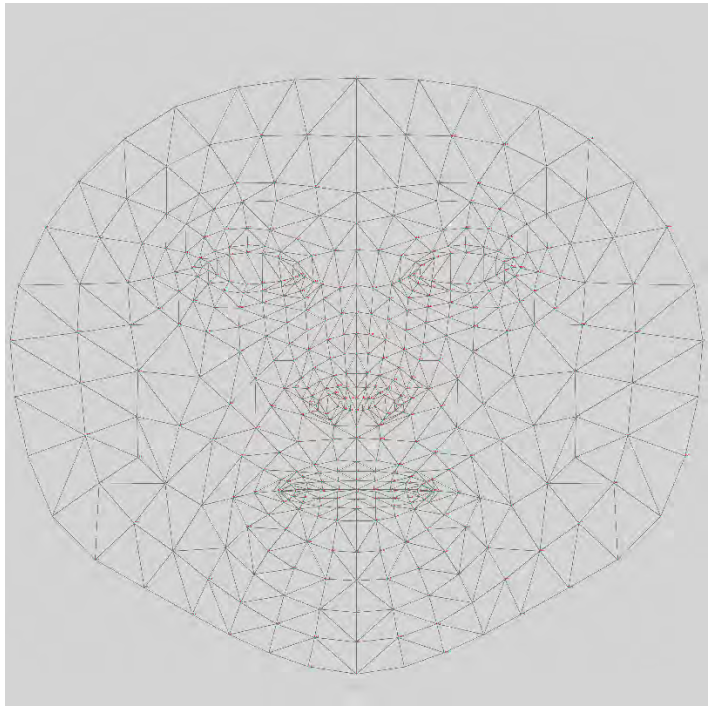


Approach to the problem

Actual approach: experiments

1) <u>Raw landmarks:</u>

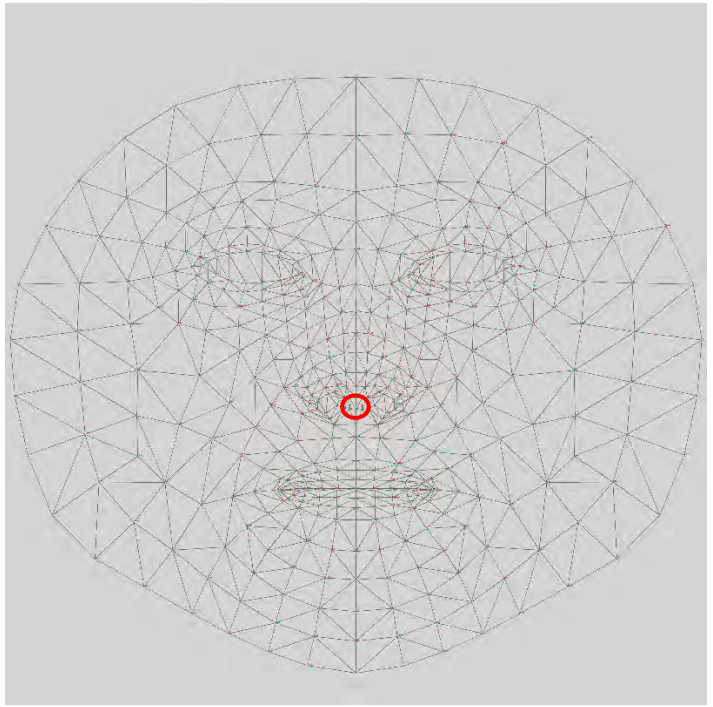
Using the 478 x,y and z landmarks concatenated into a big [1, 1434] tensor



<u>2) Distance from a point:</u>

Using the 478 landmarks distance from a fixed point of the face concatenated.

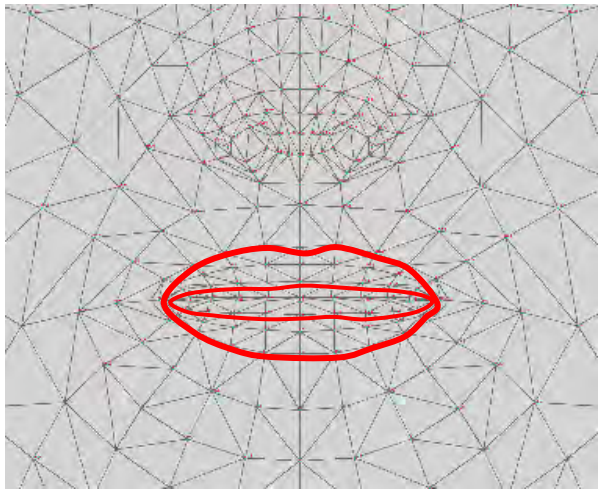
The point chosen was the 82, corresponding to the nose



<u>3) Distance from a point to the lips:</u>

Using the 21 landmarks (lips) distance from a fixed point of the face concatenated.

The point chosen was again the 82, corresponding to the nose

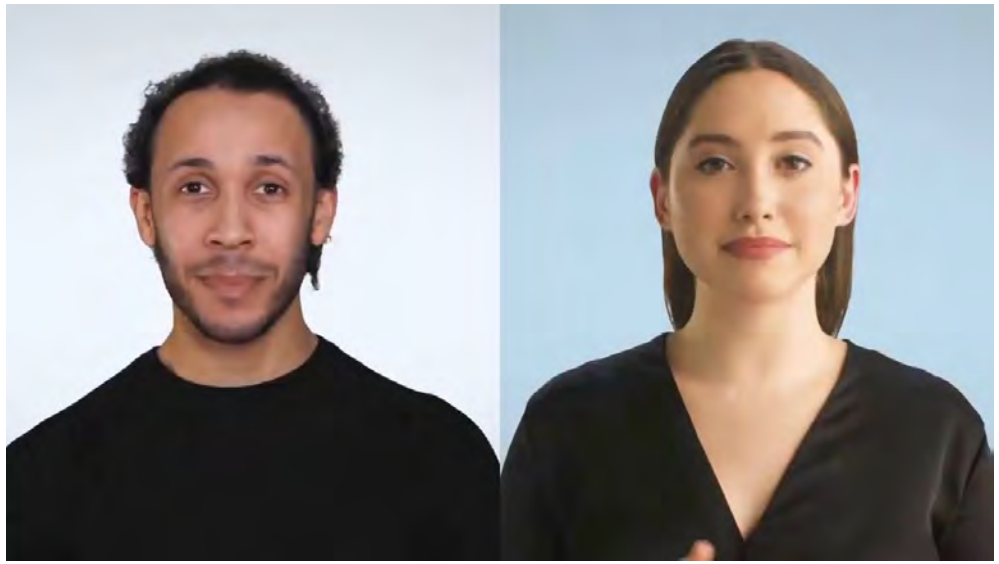


iptc

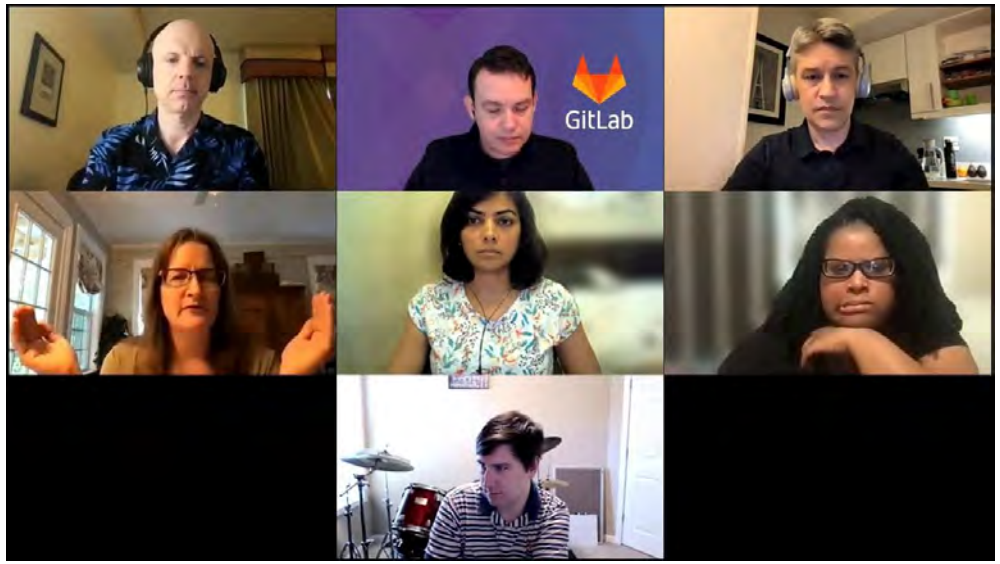
Results

Testing

After training the network and obtaining low losses, we move on to testing part. To do so we will use 2 scenarios: 2-face experiment and 7-face experiment.



2-face experiment



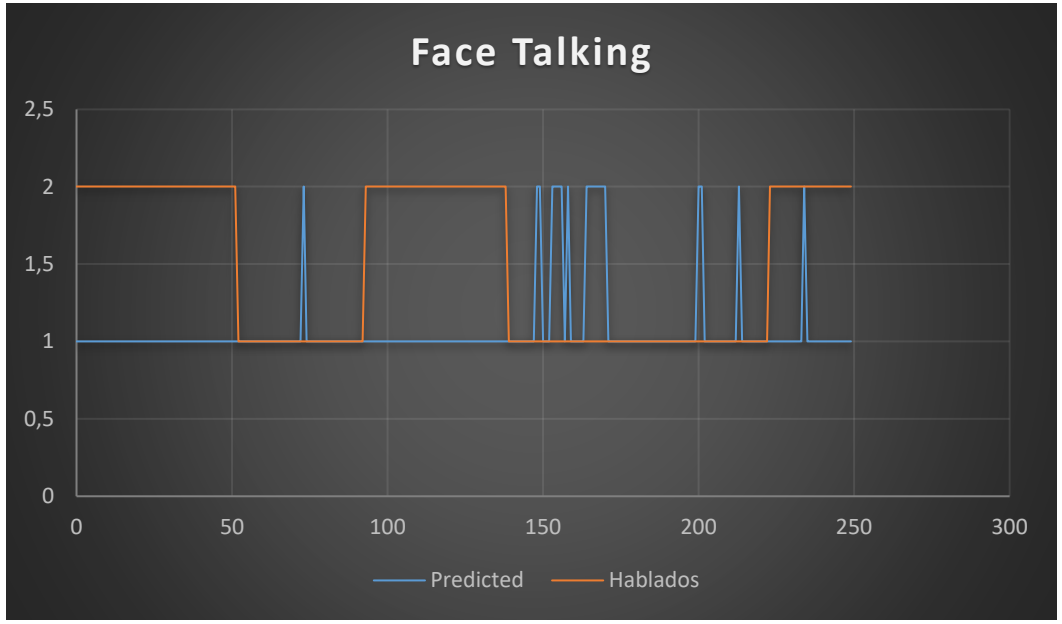
7-face experiment

iptc

Results

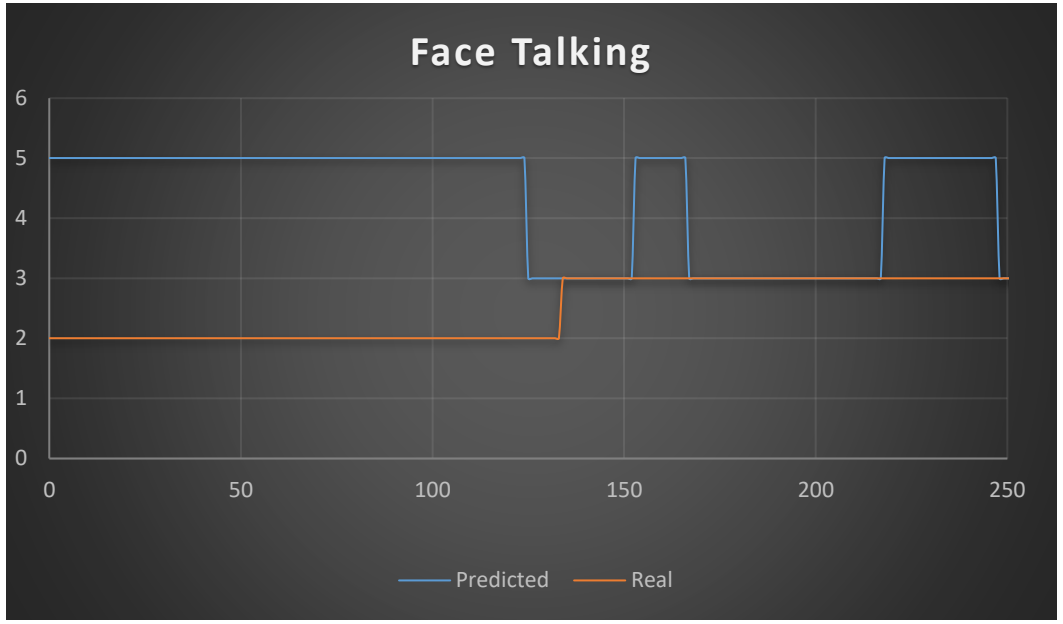
Raw Landmarks

Test with 2 faces



Accuracy
0.432

Test with 7 faces

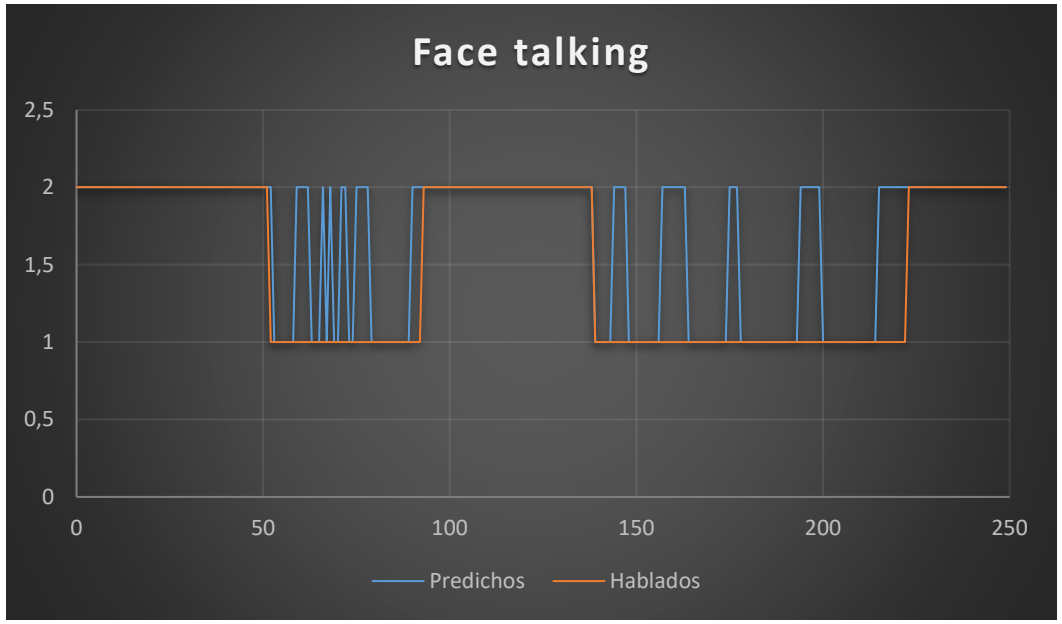


Accuracy
0.356

Results

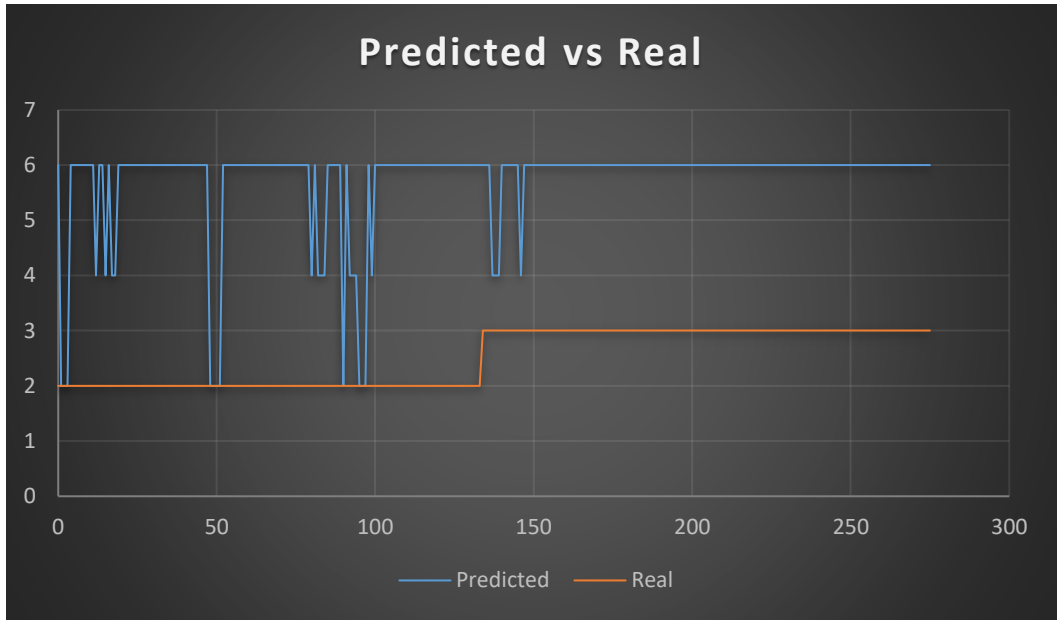
Distance

Test with 2 faces



Accuracy
0.824

Test with 7 faces



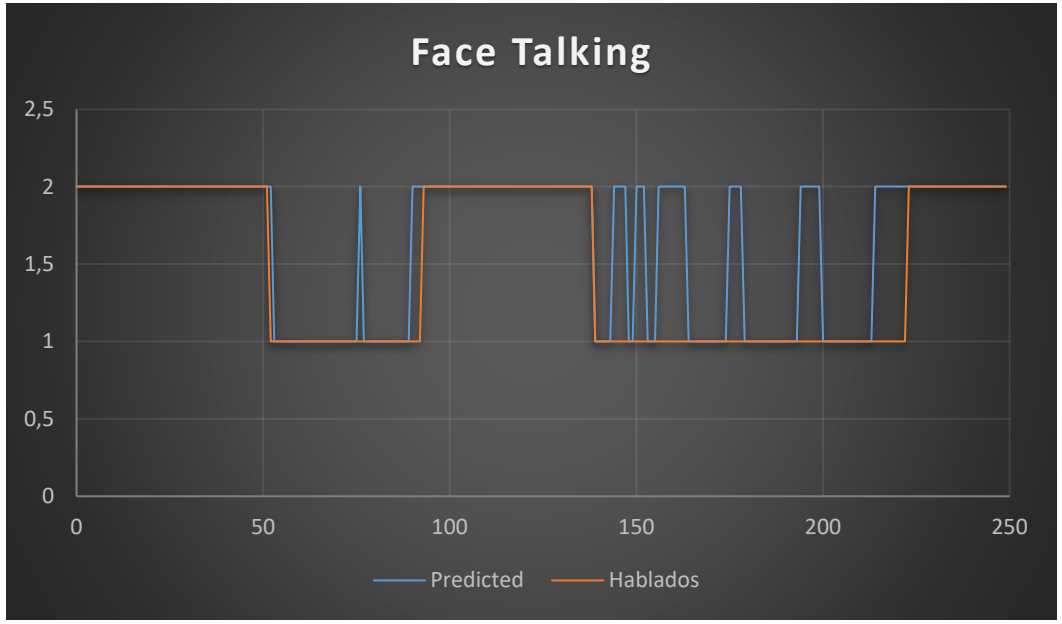
Accuracy
0.04

iptc

Results

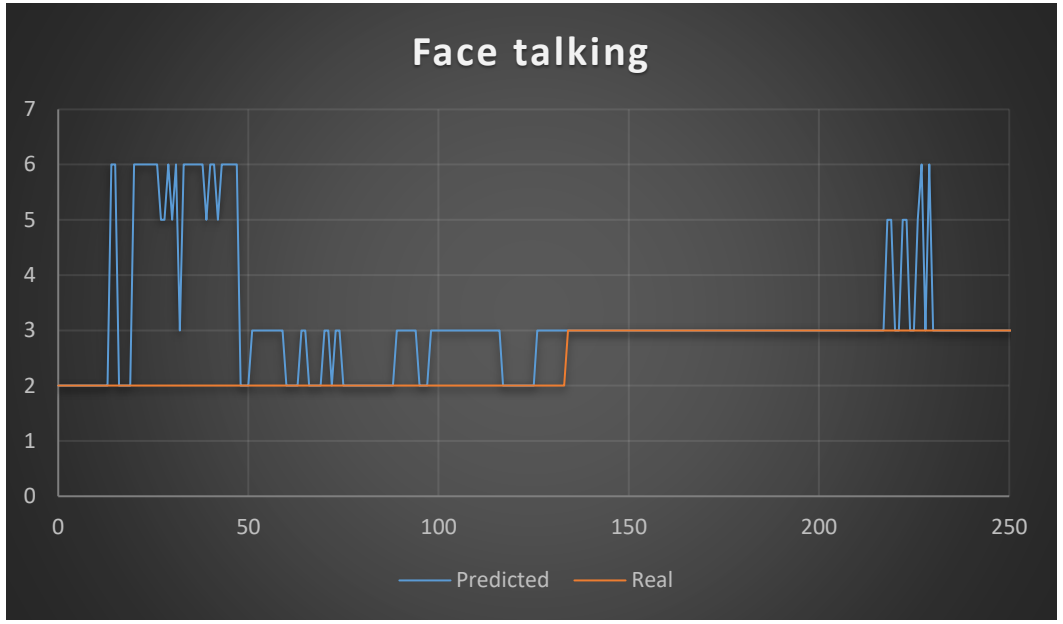
Distances lips

Test with 2 faces



Accuracy
0.844

Test with 7 faces



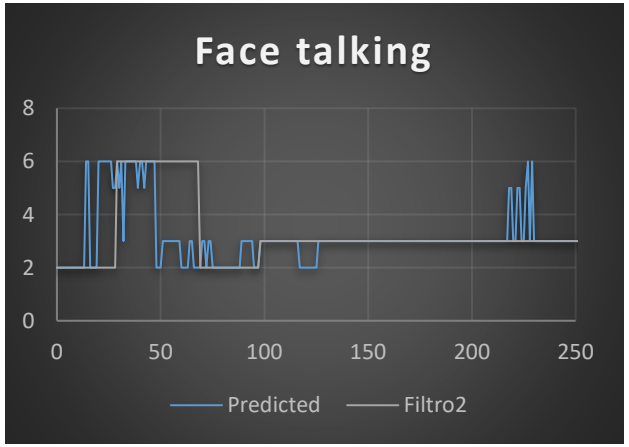
Accuracy
0.692

iptc

Improvements and future

Filtering the results

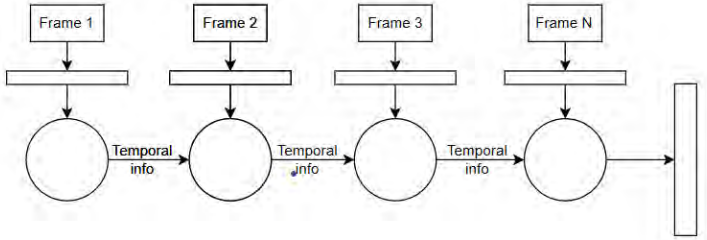
To filter spurious miss predictions

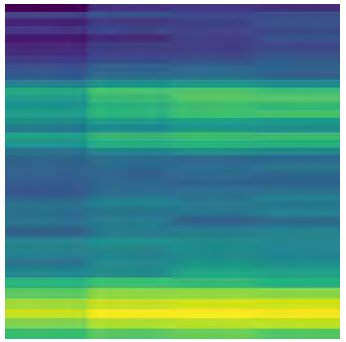


Accuracy
0.725

Improving the resolution of the images

Including temporal component in images





Temporal information

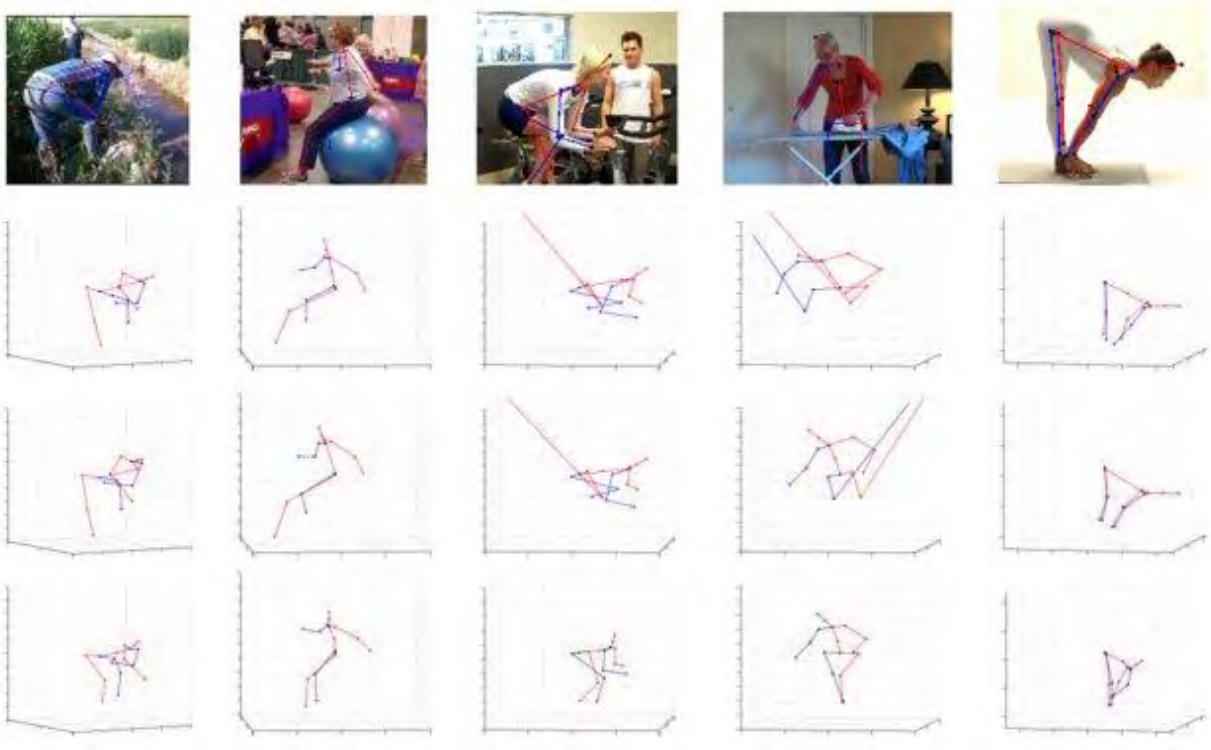
Pose and spatial movement as input for dynamic content search & generation

iptc

POLITÉCNICA

Pose and spatial movement as input for dynamic content search & generation

Main purpose? To explore the potential of posture correctness analysis and multimodal feedback delivery for different applications (ergonomics, yoga, others).



Tasks

1) Task scope definition ✓
2) Dataset search and evaluation. ✓
3) State of the art on building postural models and postural analysis. ✓
4) Setting up an environment for posture classification from images. ✓
5) Model concept proposal for posture analysis, based on angles. Built from reference datasets and literature. Limited scope.
6) Multimodal feedback by using virtual assets (e.g. avatar)
7) Incremental prototype set up.

iptc

Pose and spatial movement as input for dynamic content search & generation

CLIP as classifier

DEEPER ANALYSIS OF CLIP AS CLASSIFIER:

- **Low zero-shot performance**
- **Significant improvement after fine tuning**
  - Metrics on **82 classes**:

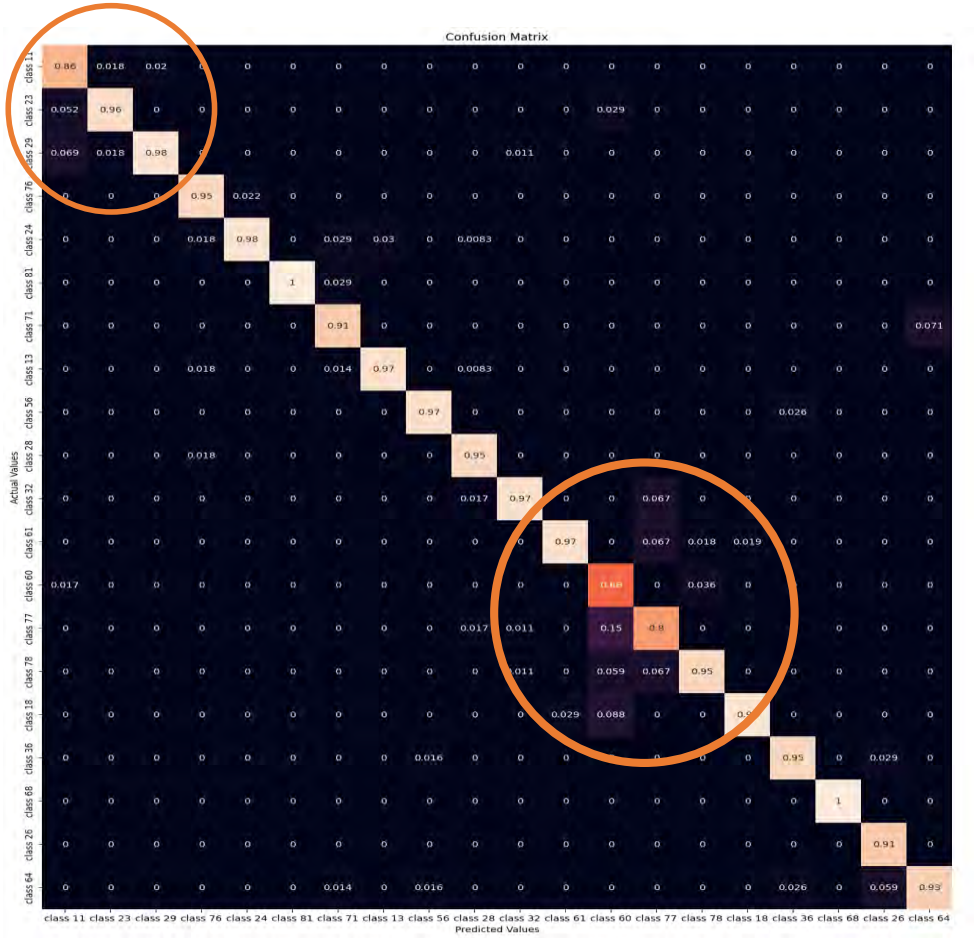
	Precision	Recall	F1-score	MCC	Support
Weighted avg	0,861	0,857	0,856	0.855	3826

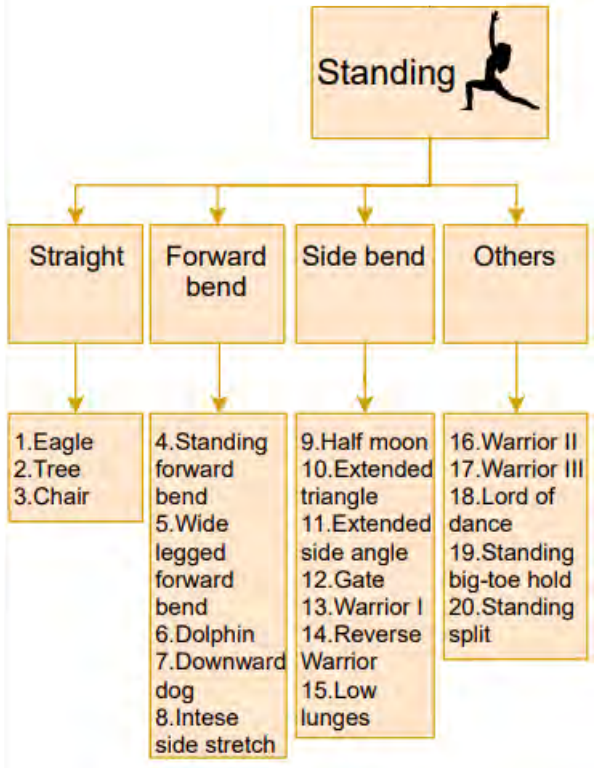
  - Confusion matrix:
    - Problem: Too big, 82 classes.
    - Solution: **Representation based on hierarchical order groups** to find common problems classifying.

Pose and spatial movement as input for dynamic content search & generation

CLIP as classifier

- **Representation based on hierarchical order groups**
  - Standing group example:

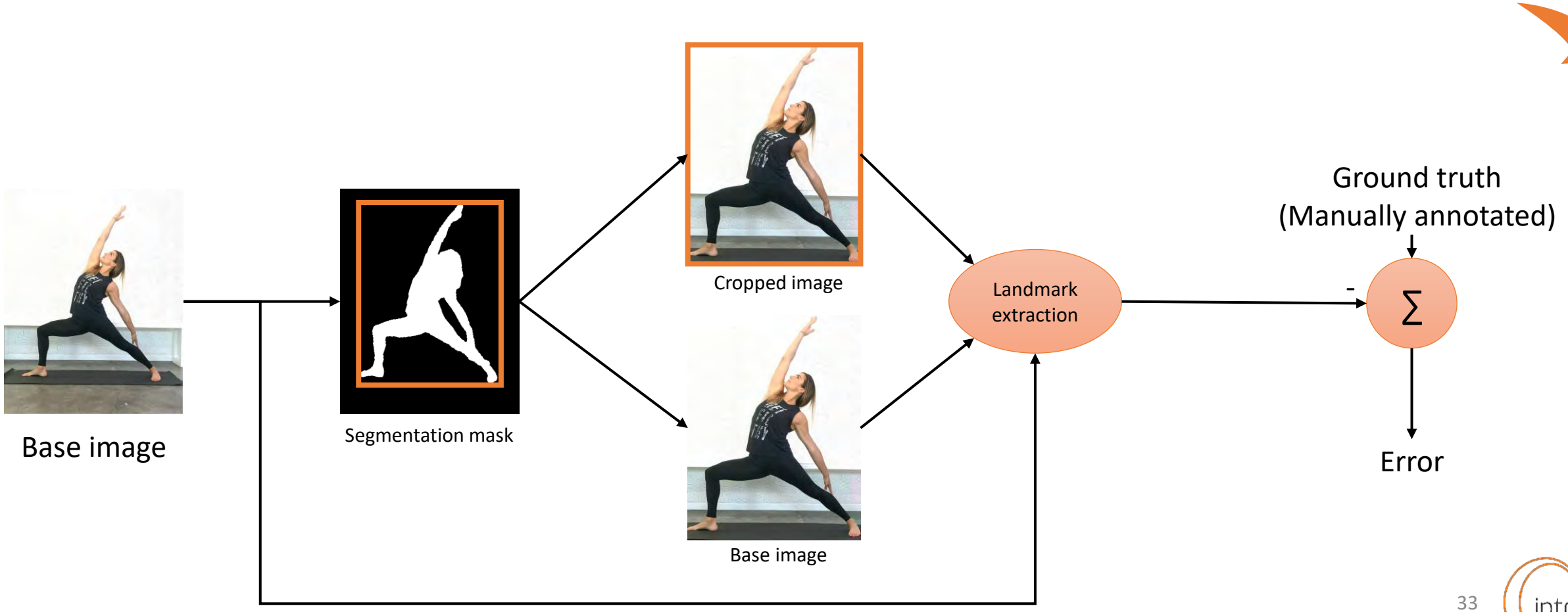




Pose and spatial movement as input for dynamic content search & generation

Mediapipe for pose evaluation

- **Mediapipe for landmarks extraction and angle computing.**
- **From previous approximation -> No significant error difference between cropping & cropping + reshaping**



Pose and spatial movement as input for dynamic content search & generation

Mediapipe for pose evaluation

- **Pipeline applied on 66 images of each class -> Analysis based on 5412 images**

Analysis based on 66 images by posture

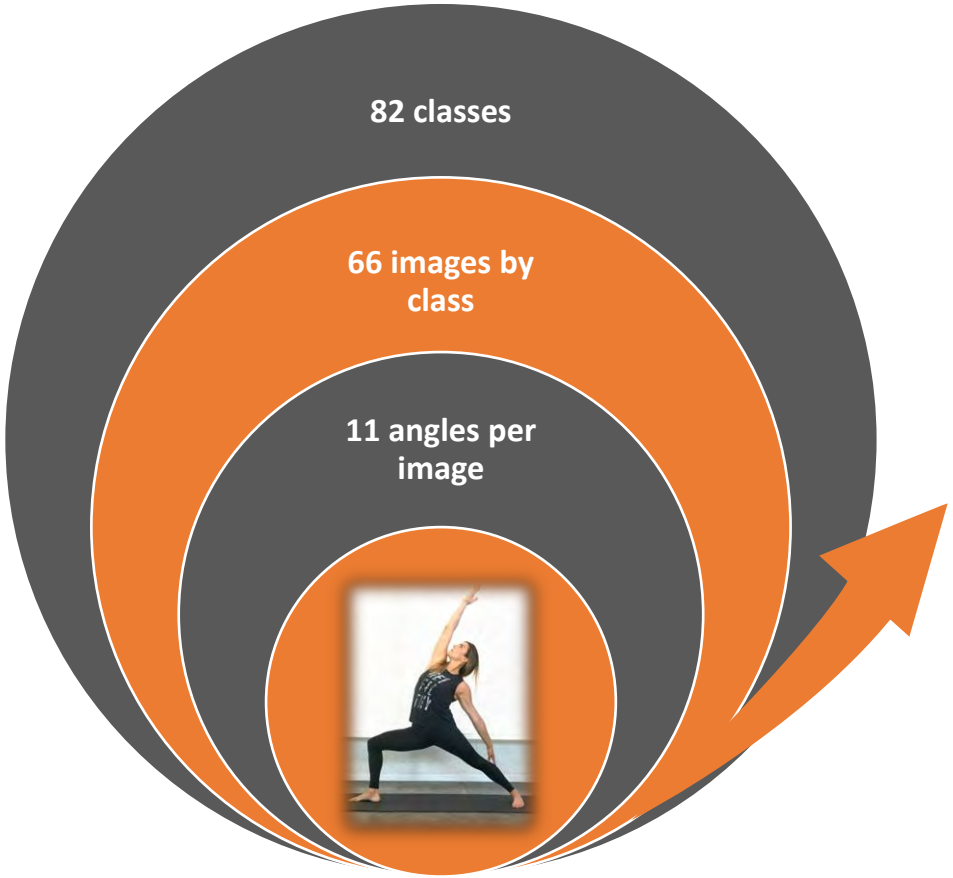
iptc

Pose and spatial movement as input for dynamic content search & generation

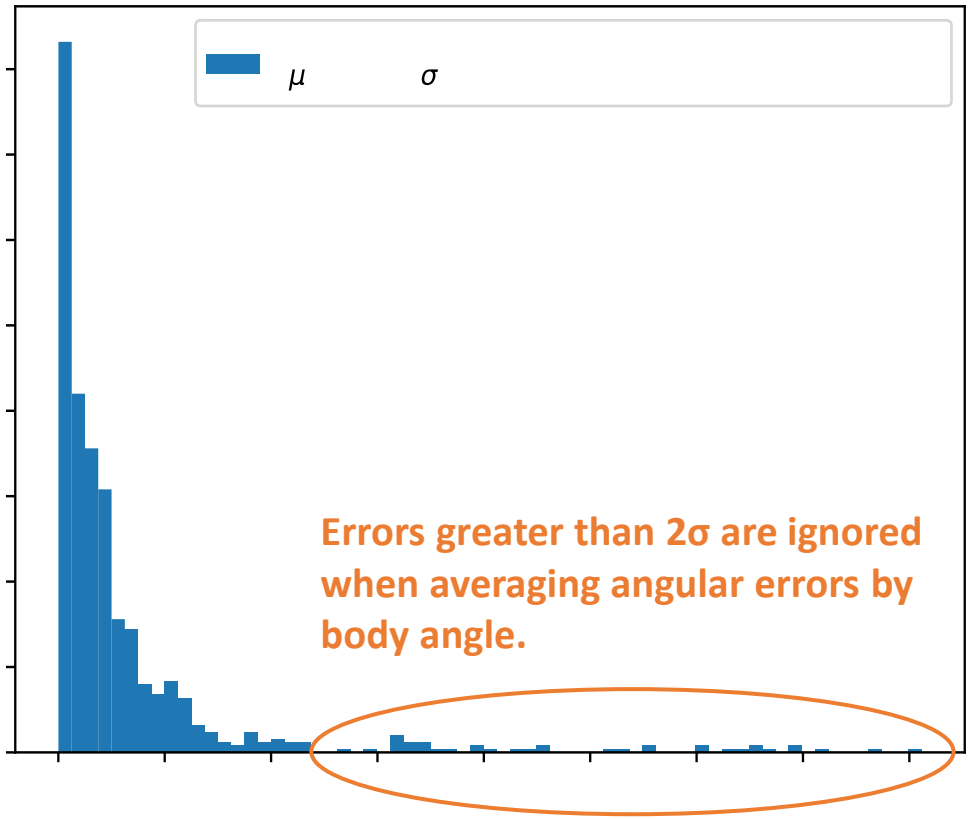
Mediapipe for pose evaluation

ANOVA & t-test analysis in progress

- **Pipeline applied to 66 images of each class:**
  - 5412 images
  - 11 body angles per image



Reverse warrior pose distribution of all angle errors

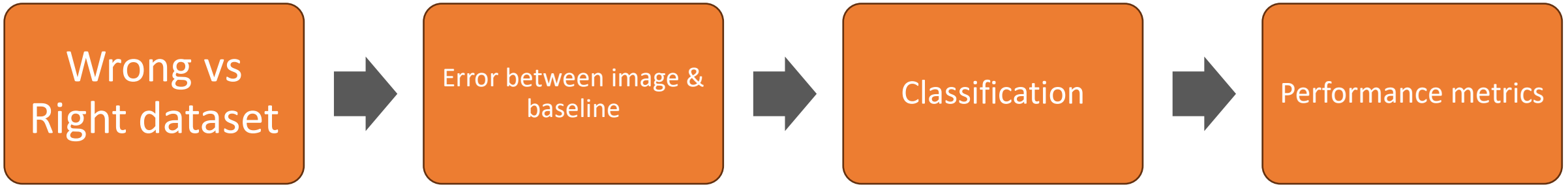


iptc

Pose and spatial movement as input for dynamic content search & generation

Binary classifier

- **Objective: Binary classifier for good/bad posture differentiation**
  - Creation of a data set of 26 classes of postures, each of which is subdivided with examples of correct and incorrect execution.



  - Problems finding the correct classification criteria.
    - Counting criteria: Nº of errors per joint > mean error per joint & Nº of errors per joint > mean standard deviation per posture.

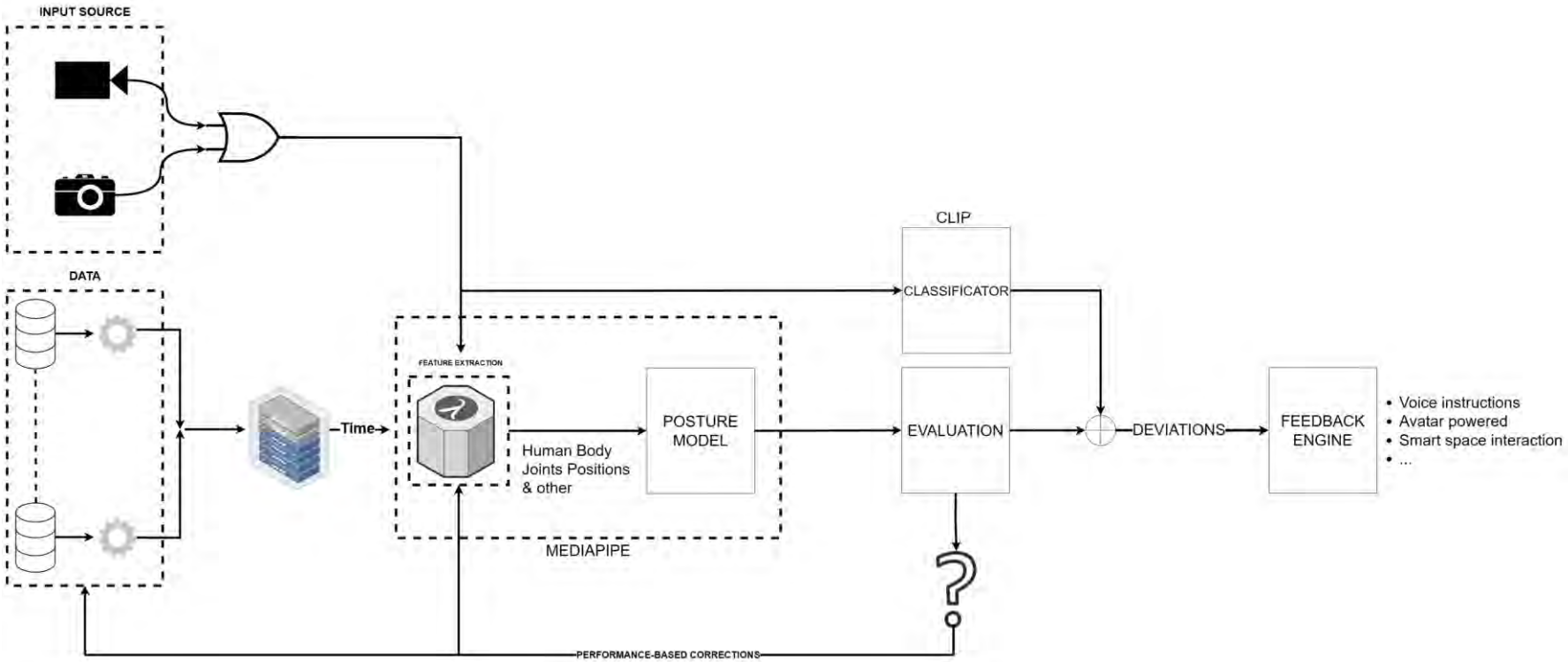
Image predictions	Wrong image predictions	Precision
199	89	0.553

Good posture classification

Pose and spatial movement as input for dynamic content search & generation

Next Steps

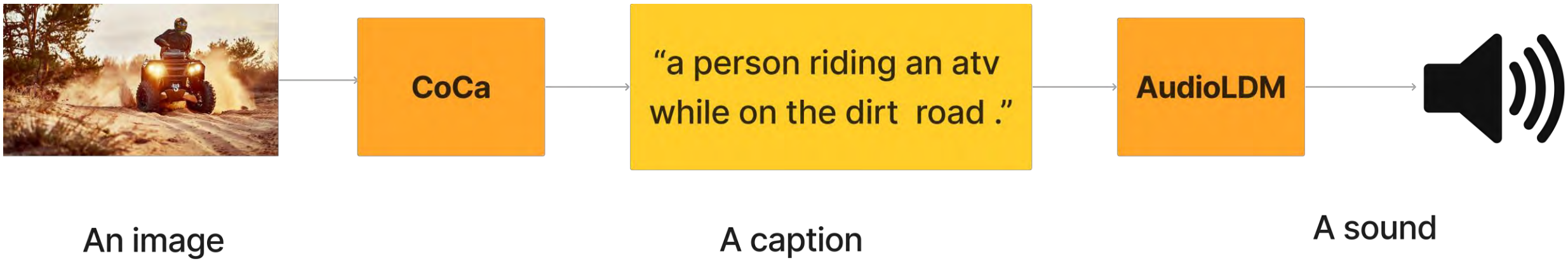
- Improving performance differentiating between "correct" and "wrong" postures.
  - Gathering a specific dataset to train a classifier?
  - Ma
- Setting-up a proof-of-concept system prototype with (some) feedback on posture evaluation.



Audio Generation using Deep Learning

Sound Generation from Images

Basic Pipeline

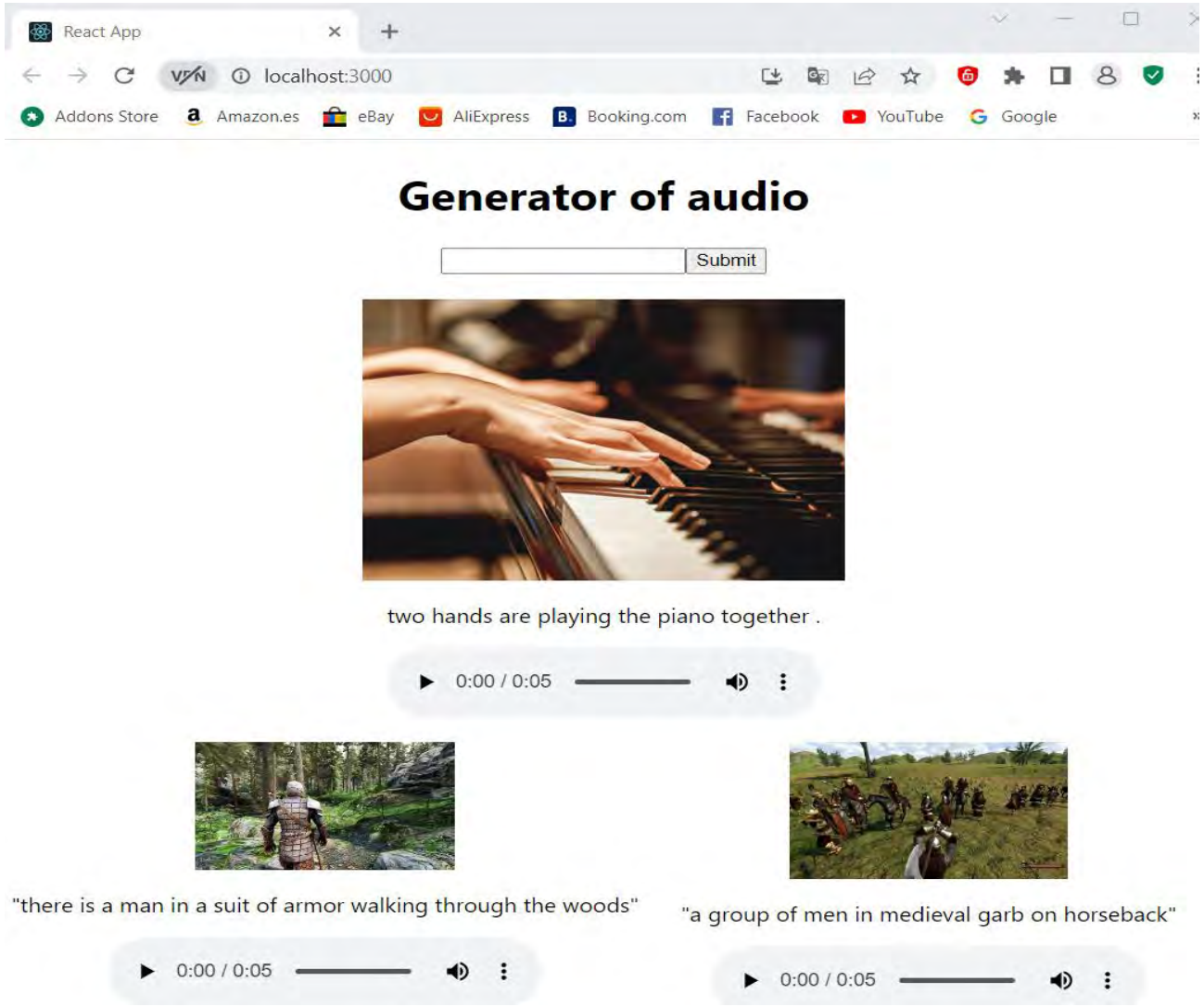


CoCa: https://arxiv.org/abs/2205.01917

AudioLDM: https://arxiv.org/abs/2301.12503

iptc

DEMO



https://github.com/Laurafdez/Generator-of-audio-from-images

Two research lines

RESEARCH LINE 1: WITHOUT CAPTIONS

CoCa

~~"a person riding an atv while on the dirt road ."~~

AudioLDM

An image

A caption

A sound

RESEARCH LINE 2: USING CAPTIONS

CoCa

"a person riding an atv while on the dirt road ."

AudioLDM

An image

A caption

A sound

VS

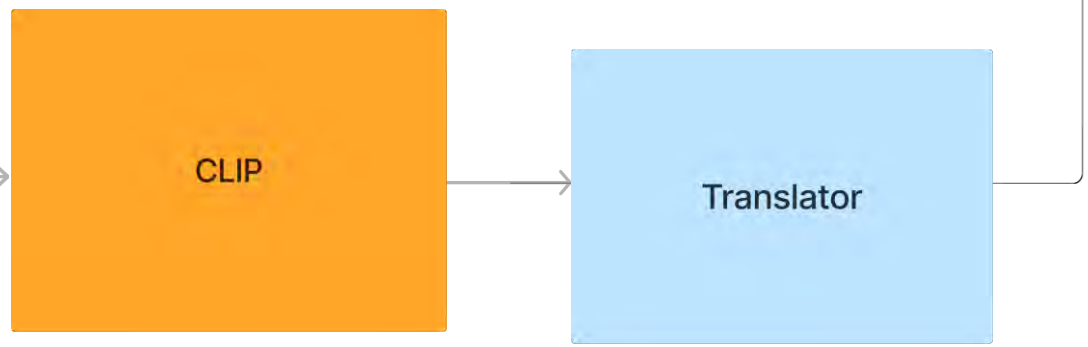
CoCa: https://arxiv.org/abs/2205.01917

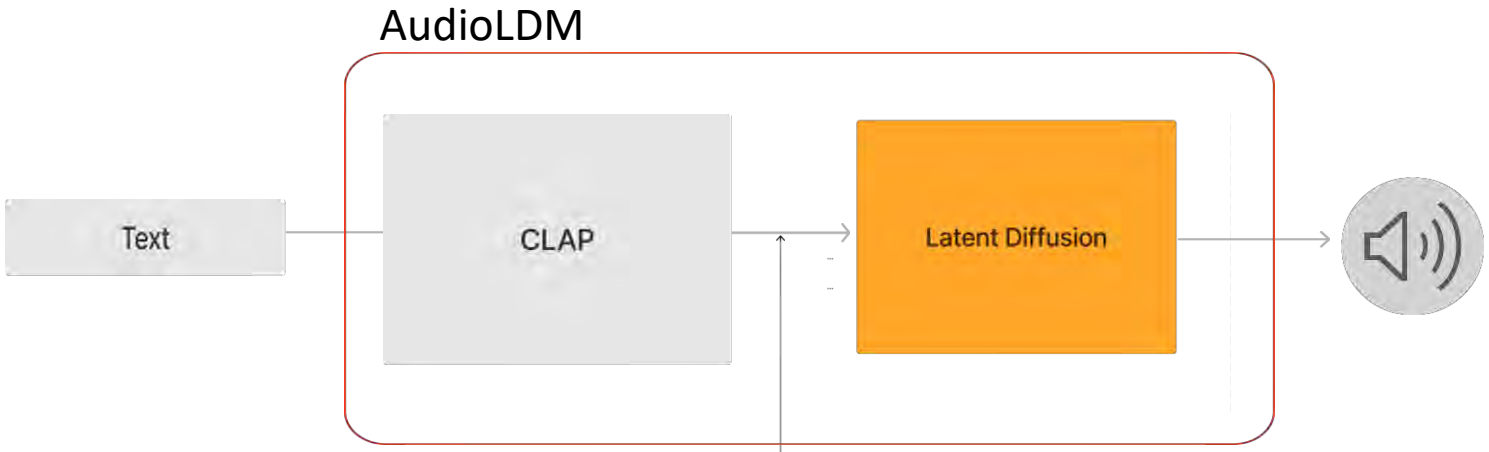
AudioLDM: https://arxiv.org/abs/2301.12503

Research Line 1: Pipeline without using captions

Audio Generation from Images

- Eliminate the need for CoCa to describe images.
- AudioLDM utilizes CLAP to generate audio representations.
- Translate CLIP embeddings into CLAP embeddings.

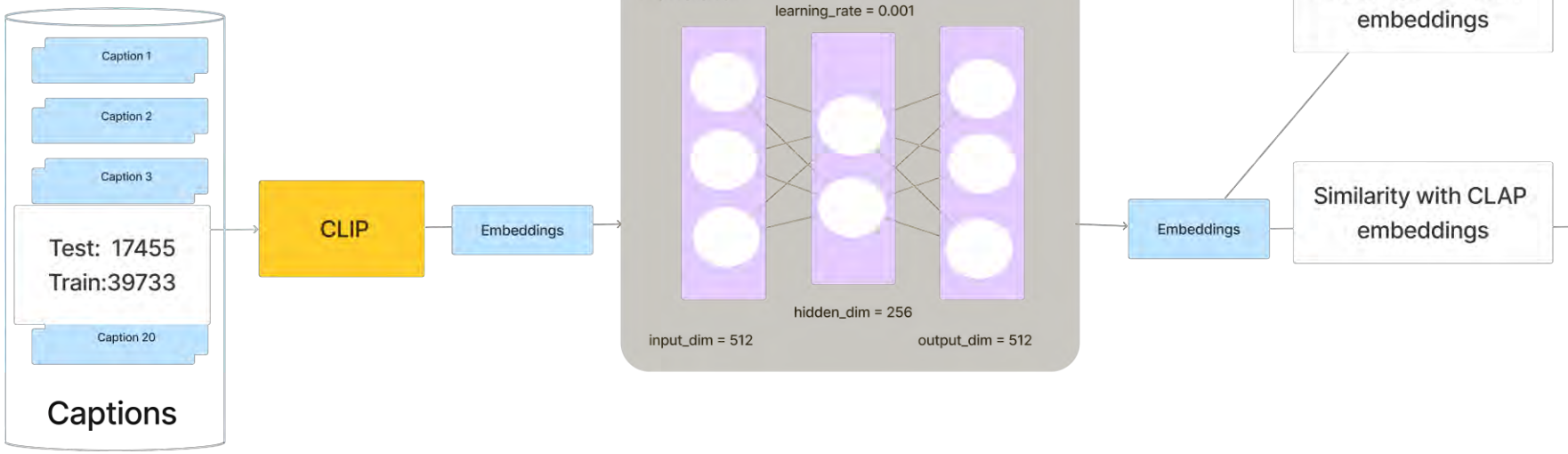


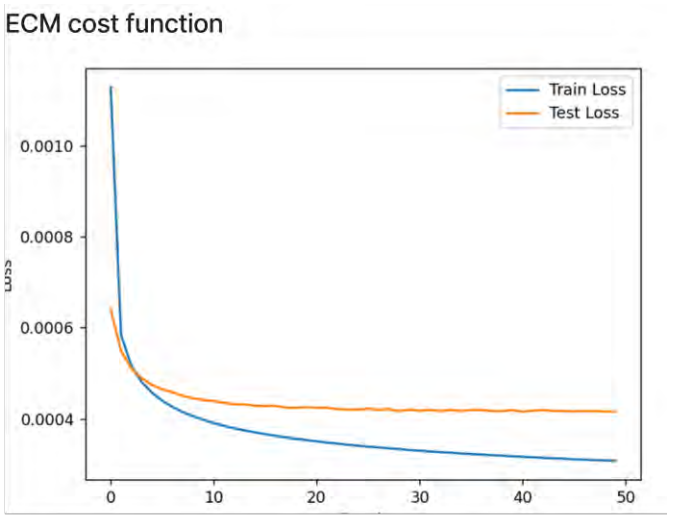


Research Line 1: Translator

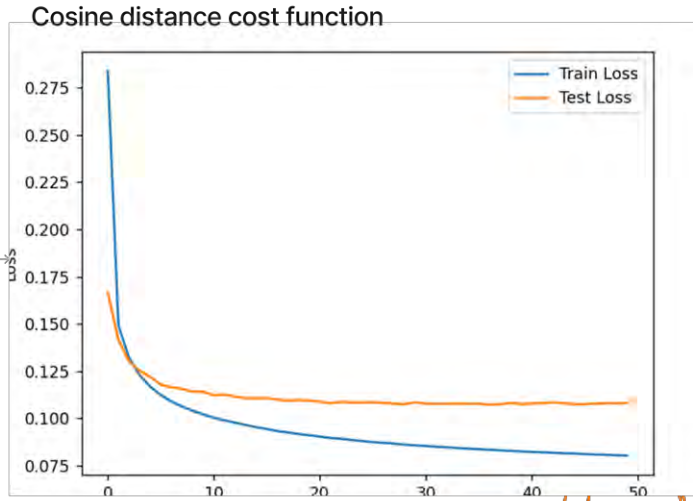
- **Audiocaps dataset texts:** Text data from the Audiocaps dataset used for training AudioLDM.

- **Translator generates embeddings:** The Translator model generates embeddings for comparison with CLAP embeddings.





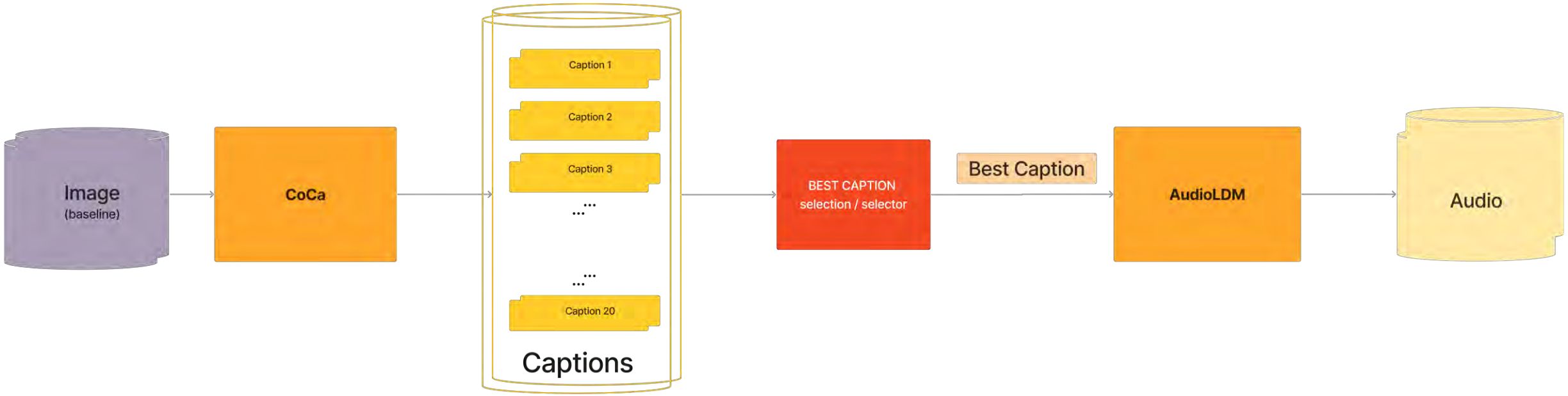
Training error average: 0.0961 Test error average: 0.1122



Training error average: 0.095 Test error average: 0.100

Research Line 2: Pipeline using captions

Audio Generation from Images



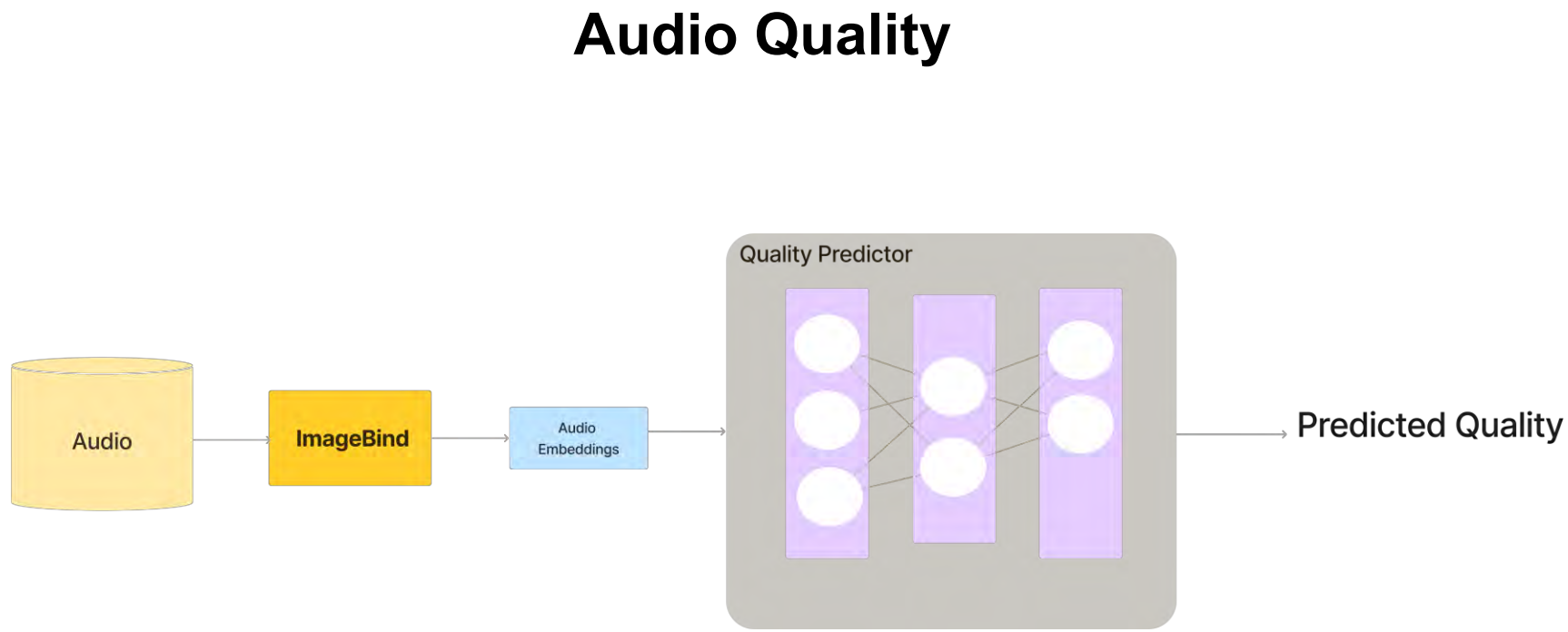
CoCa: https://arxiv.org/abs/2205.01917

AudioLDM: https://arxiv.org/abs/2301.12503

iptc

Objective Metrics

Through embeddings

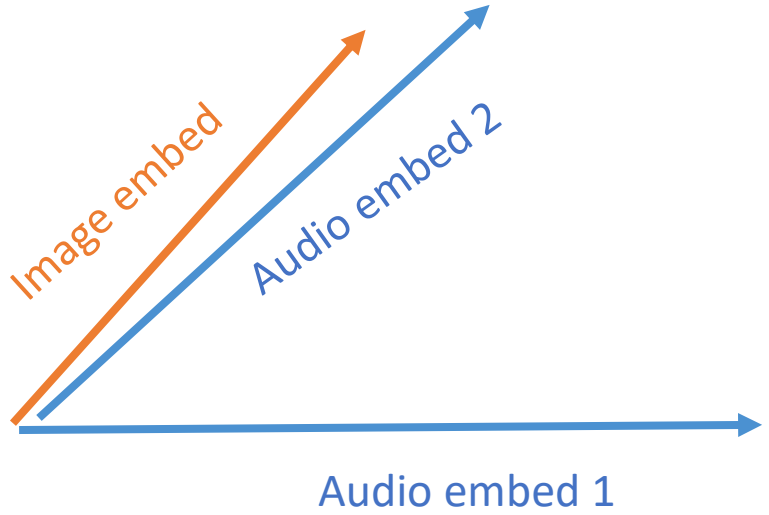


ImageBind: https://arxiv.org/abs/2305.05665

Objective Metrics

Through embeddings

Image-Audio Consistency



ImageBind: https://arxiv.org/abs/2305.05665

Subjective Metrics

Working on…

- Audio Quality
- Audio-Image Coherence
- Caption Reflects Visual Content
- Caption Reflects Acoustic Content

High, Medium, or Low

How does it sound to you?

“a person walking away from the water in gaming .”

iptc

Future work

Ongoing activities

- **Develop quality metrics for both Audio and Image-Audio Coherence to be used in Research Lines 1 & 2**
- **Complete and Evaluate Research Line 1: Image-to-audio without captions**
- **Evaluate different strategies for caption selection in Research Line 2.**

Technical Report

Audio Generation from Images

iptc

Dissemination: Publications

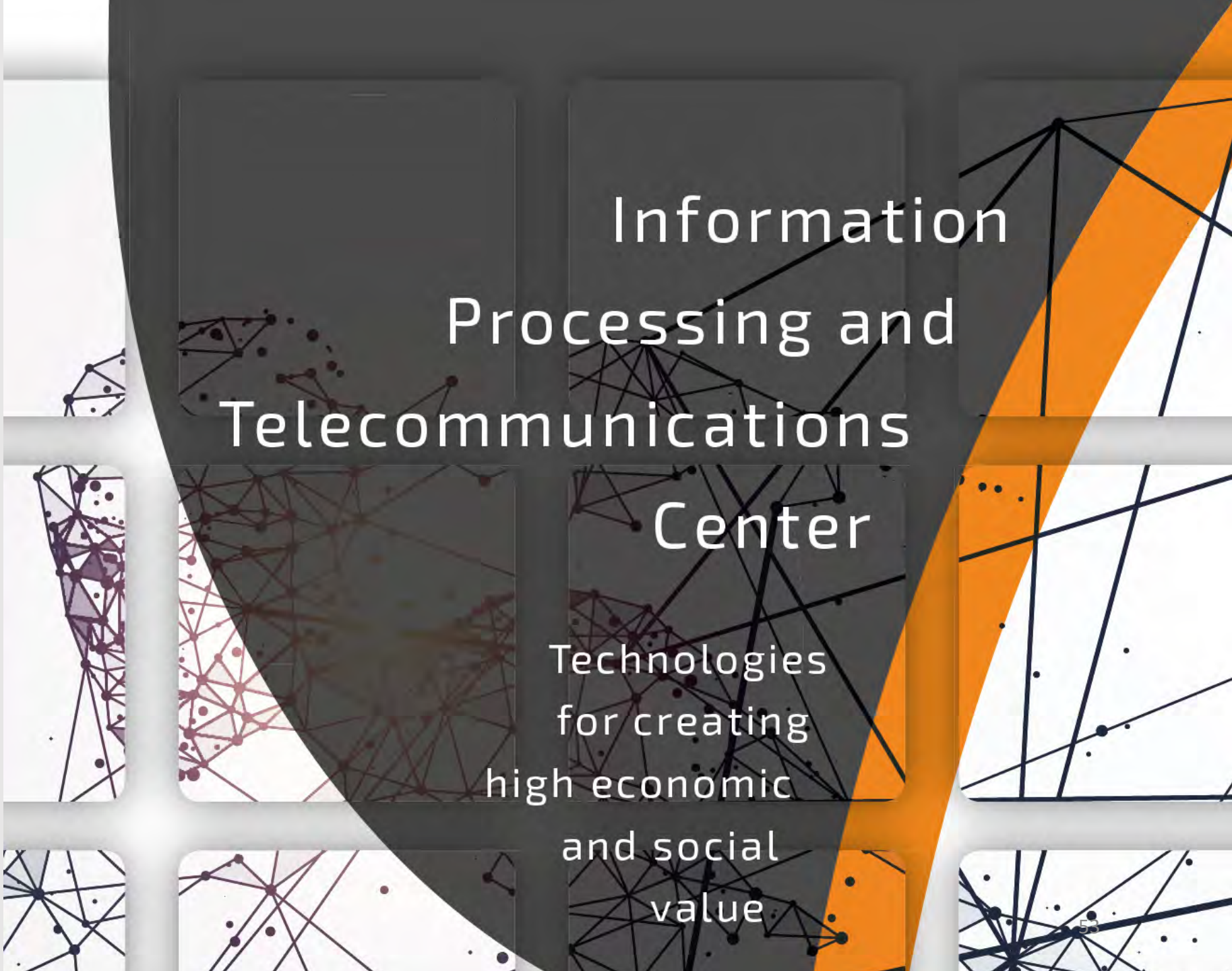
- **Acknowledgement:**
  - "XXX's scholarship has been supported by Amazon through the IPTC-Amazon collaboration initiative“
  - Promote joint authorship (IPTC-Amazon)
- **Possible targets (Open Access):**
  - Applied Sciences. (IF: 2.838, ISSN: 2076-3417) — Special Issue "Audio, Speech and Language Processing“
  - Biosensors (ISSN: 2079-6374, IF: 5.972).
  - Online conferences:
    - ECSA-10 (Electronic Conference on Sensors and Applications. 2023)
    - …

General meeting in September/October

- **Dissemination: open event**
- **Possibility to organize an onsite (+ online) event:**
  - Detailed presentation of all projects.
  - General presentation of the main results.
  - Presentation of Amazon at IPTC-UPM.

Great Summer and Vacations!!!

Many thanks



POLITÉCNICA

iptc

www.iptc.upm.es