# Digital Signal and Image Processing

- **Audio classification**
- **Image classification**
- **Image retrieval**

Alberto Usai, 886731
Geralda Nushi, 873269
Tommaso Redaelli, 830442

*Corso di Digital Signal and Image Preprocessing*
*Università Milano Bicocca - A.A. 2022/20223*

# Audio classification

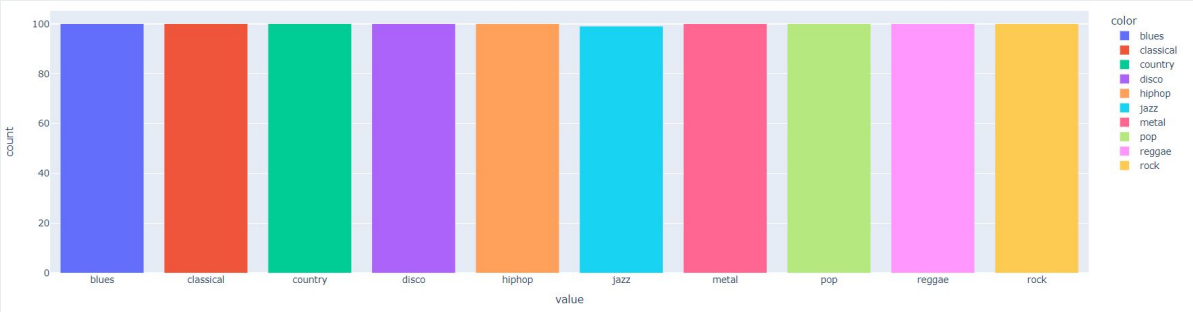
1. Dataset
2. Classic machine learning
3. Deep learning methods
4. Evaluation and conclusion

# Dataset

**GTZAN Dataset** – Music Genre Classification

**Number of classes:** 10 *{ “rock”, “jazz”, “blues”, “disco” …. }*
**Audio structure:** length: *30 sec* | rate: *22050*
**Provided features:** *{ “chroma”, …. }*

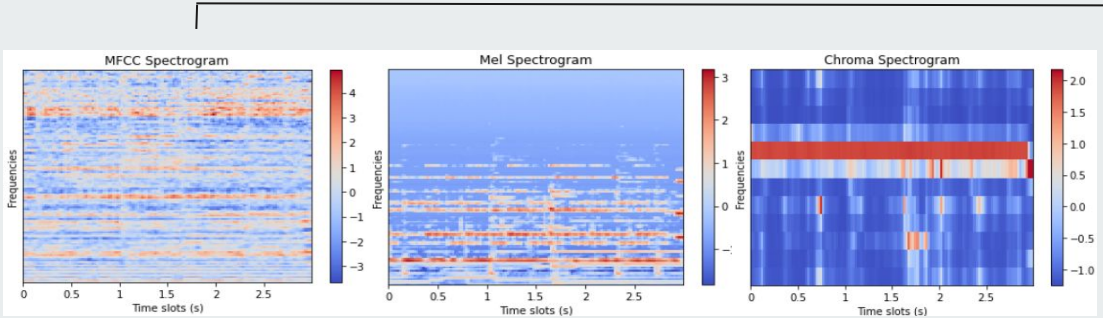
**Labels distr.**

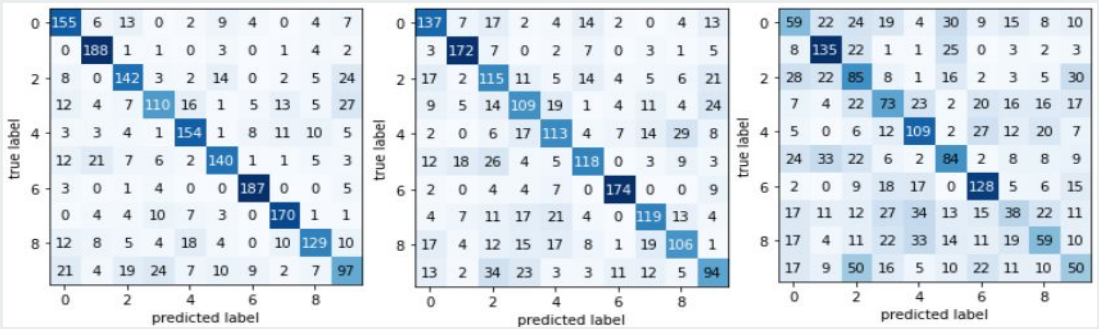


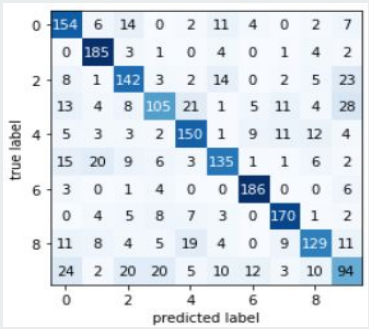
# Classic ML methods

SVM on:

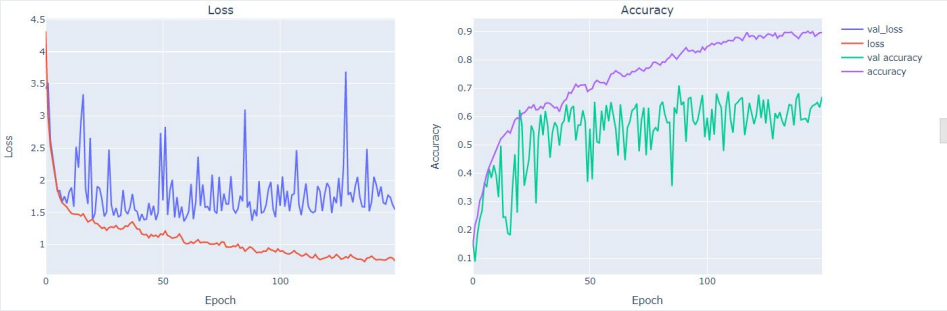
- MFCC
- MEL
- CHROMA
- MEL + CHROMA

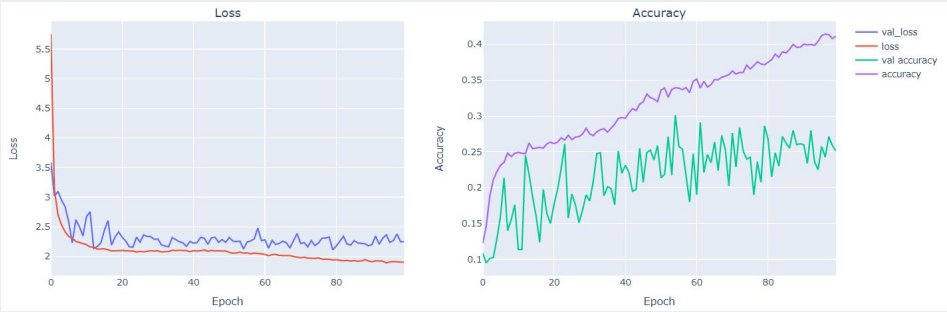


# Deep learning methods 1/3



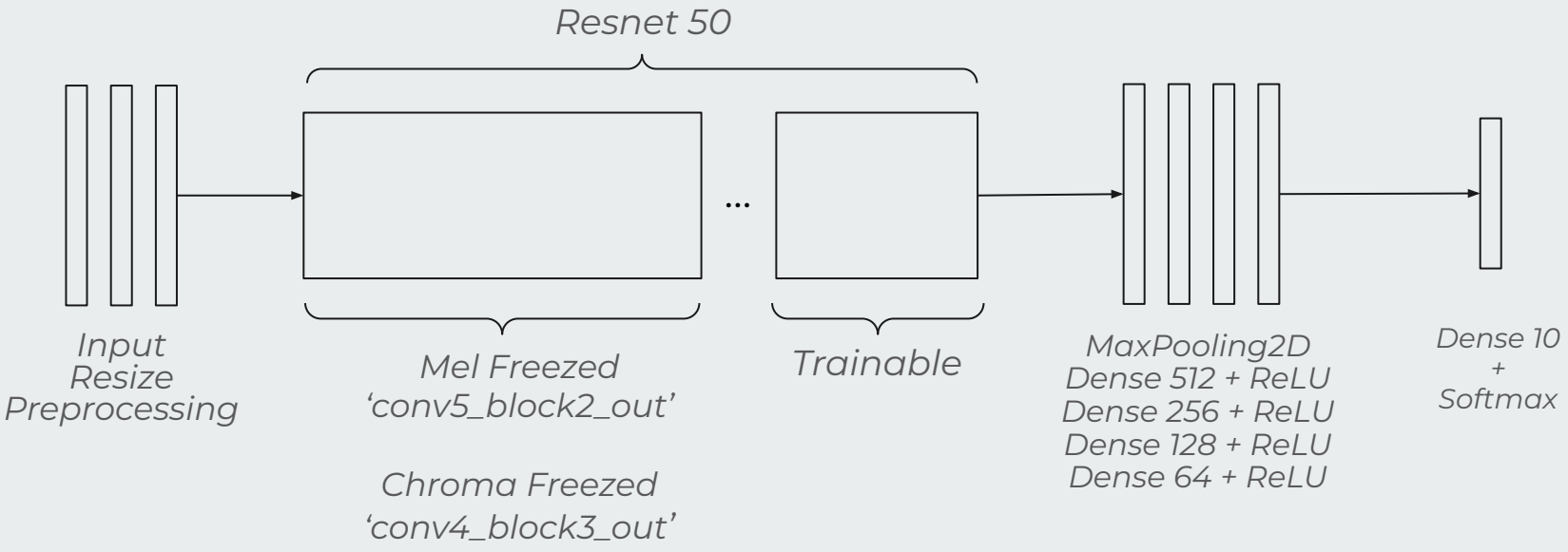

***CNN ON CHROMA SPECTROGRAM***

| | Train | Val | Test |
|---|---|---|---|
| Acc. | 0.7501 | 0.6513 | 0.6612 |
| Loss | 1.0320 | 1.3605 | 1.5576 |

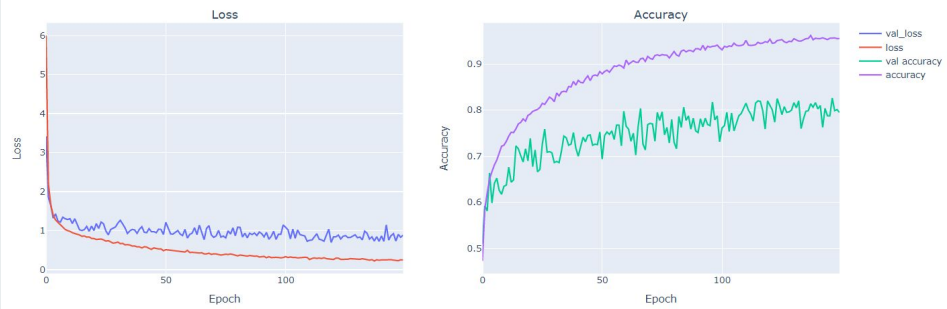
***CNN ON CHROMA SPECTROGRAM***

| | Train | Val | Test |
|---|---|---|---|
| Acc. | 0.3714 | 0.2862 | 0.2863 |
| Loss | 1.9355 | 2.1075 | 2.2351 |

# Resnet 50

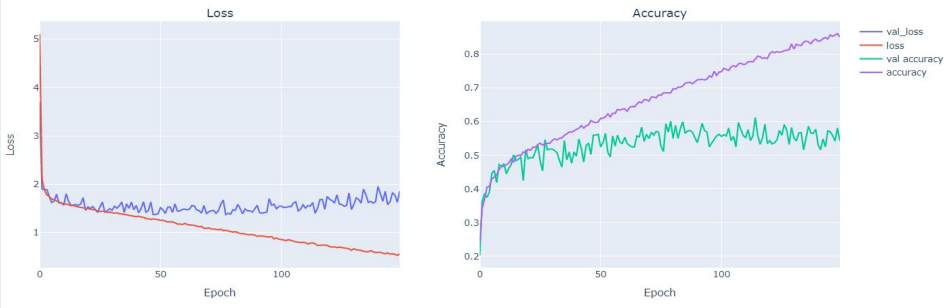

Resnet 50
...
Input
Resize
Preprocessing
Mel Freezed
'conv5_block2_out'
Chroma Freezed
'conv4_block3_out'
Trainable
MaxPooling2D
Dense 512 + ReLU
Dense 256 + ReLU
Dense 128 + ReLU
Dense 64 + ReLU
Dense 10
+
Softmax

# Deep learning methods 2/3




***ResNet50 ON CHROMA SPECTROGRAM***

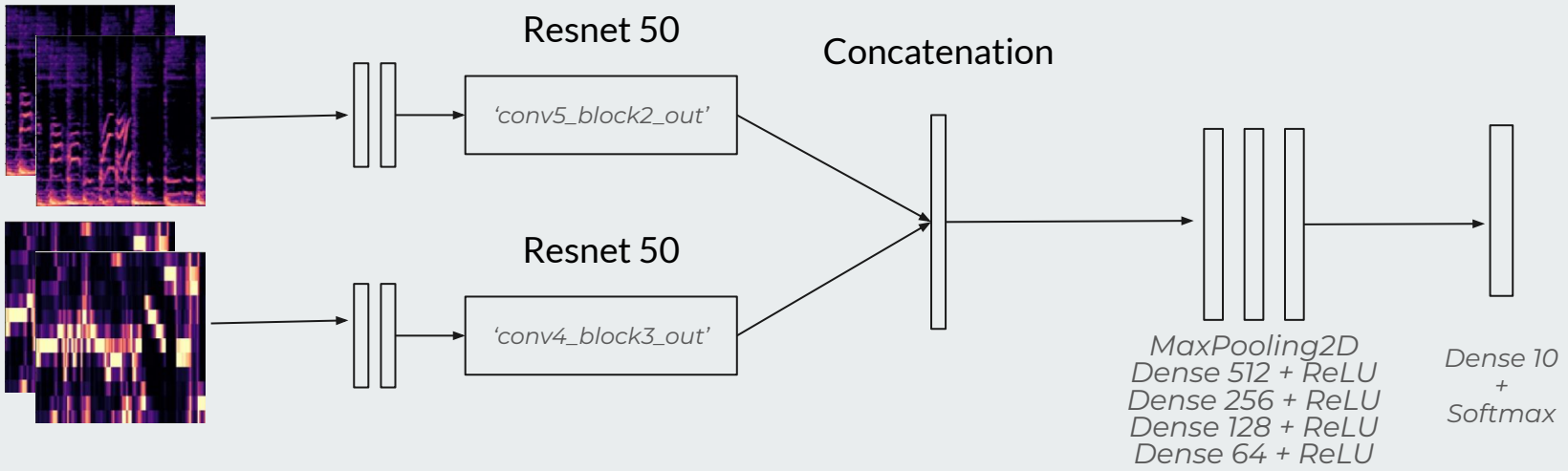
| | Train | Val | Test |
|---|---|---|---|
| Acc. | 0.9469 | 0.8200 | 0.9527 |
| Loss | 0.2778 | 0.6968 | 0.7838 |




***ResNet50 ON CHROMA SPECTROGRAM***

| | Train | Val | Test |
|---|---|---|---|
| Acc. | 0.6853 | 0.5938 | 0.5505 |
| Loss | 1.0467 | 1.3730 | 1.7829 |

# Deep learning methods 3/3



# Possible improvements

| | |
|---|---|
| Features → | - Different parameters (hop_length, n_mel, n_chroma, n_mfcc)<br>- Audio augmentation techniques<br>- Harmonic and Percussive spectrogram |
| CNN Networks → | - Use of MFCC spectrograms<br>- Different freezed layers in resnet50 |
| Double Input Network → | - Optimization of Data Generator<br>- Use of Mel + MFCC spectrograms |

# Image classification

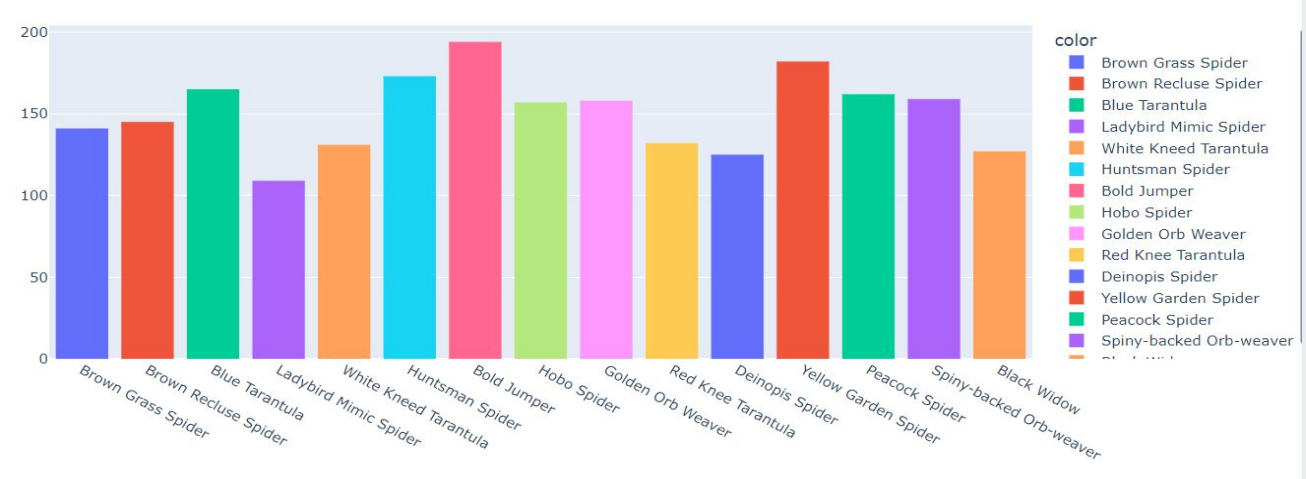
1. Dataset
2. CNN transfer-learning and fine-tuning
3. Ensemble models
4. Evaluation and conclusion

# Dataset

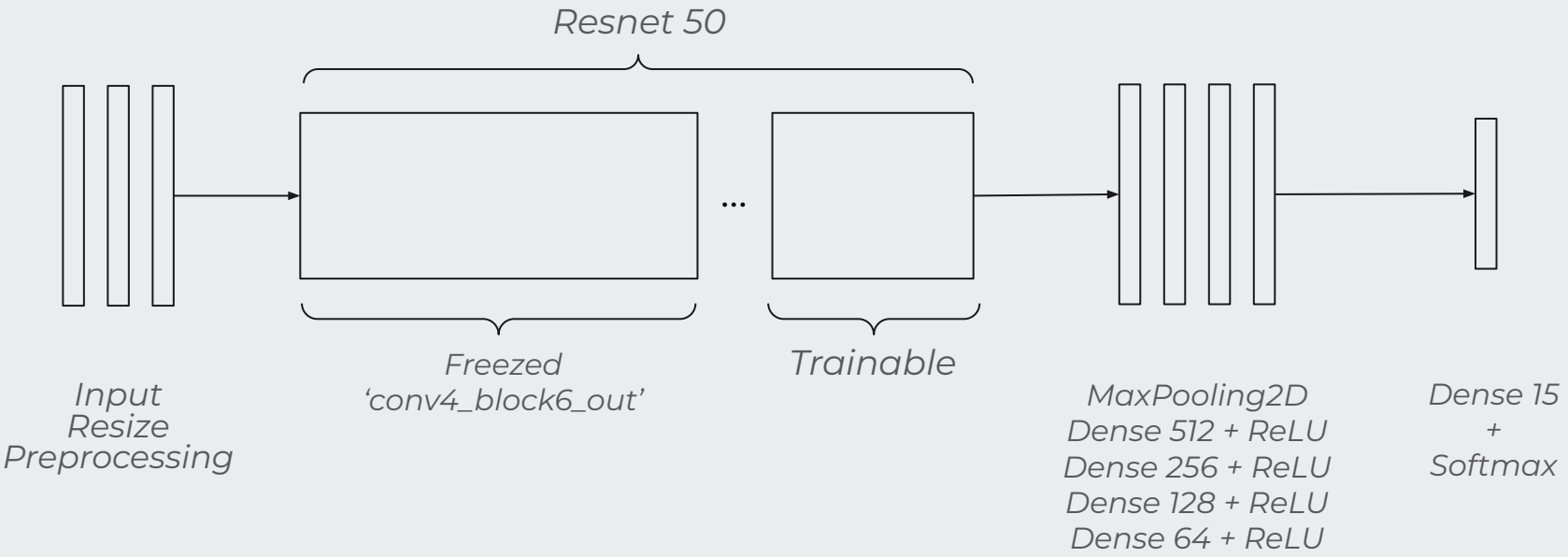
**YIKES! Spiders** – 15 Species Classification Dataset

**Numero di classi:** 15 *{ "Hobo Spider", "Black Widow", .... }*
**Struttura immagini:** 224 x 224 x 3

**Labels distr.**



# Resnet 50


Resnet 50
...
Input
Resize
Preprocessing
Freezed
'conv4_block6_out'
Trainable
MaxPooling2D
Dense 512 + ReLU
Dense 256 + ReLU
Dense 128 + ReLU
Dense 64 + ReLU
Dense 15
+
Softmax

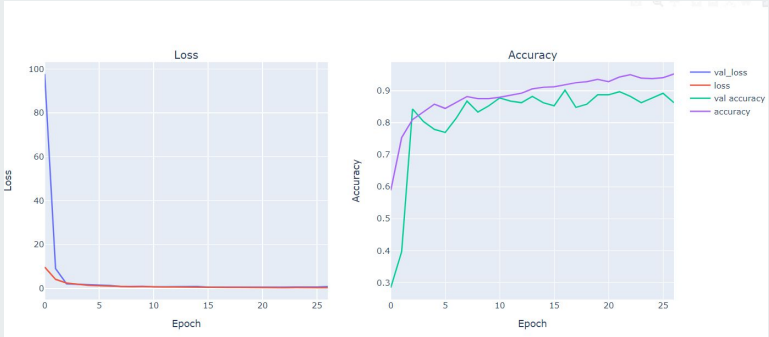
# Results

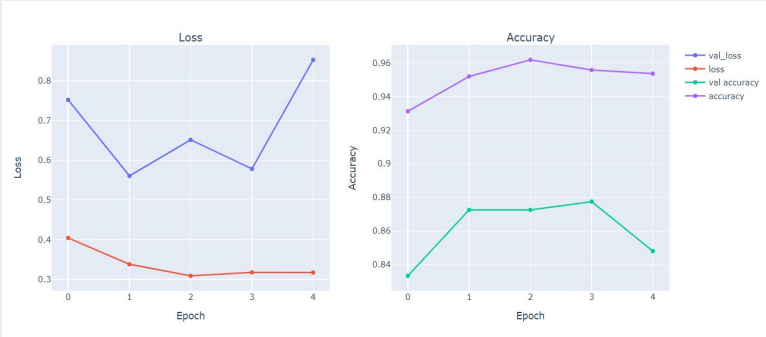
*Resnet50*

| | Train | Val |
|---|---|---|
| Acc. | 0.9432 | 0.8971 |
| Loss | 0.3969 | 0.5613 |

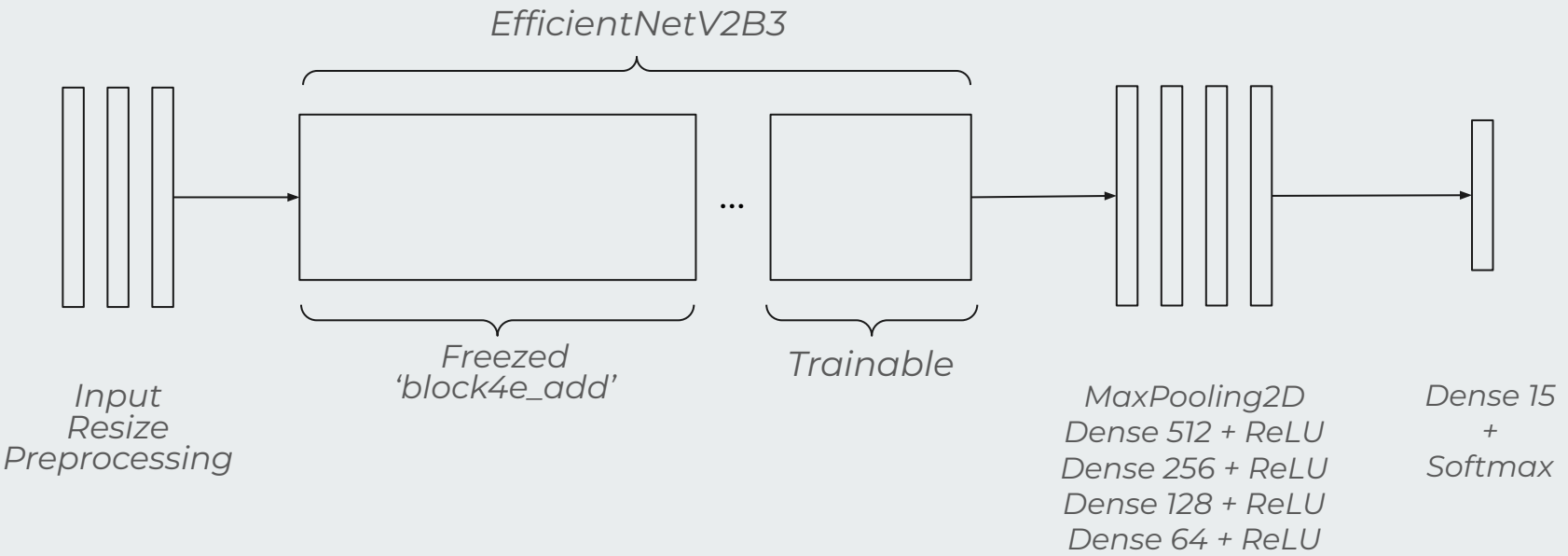
*Resnet50 with fine-tune*

| | Train | Val |
|---|---|---|
| Acc. | 0.9519 | 0.8725 |
| Loss | 0.3380 | 0.5605 |



# EfficientNetV2B3



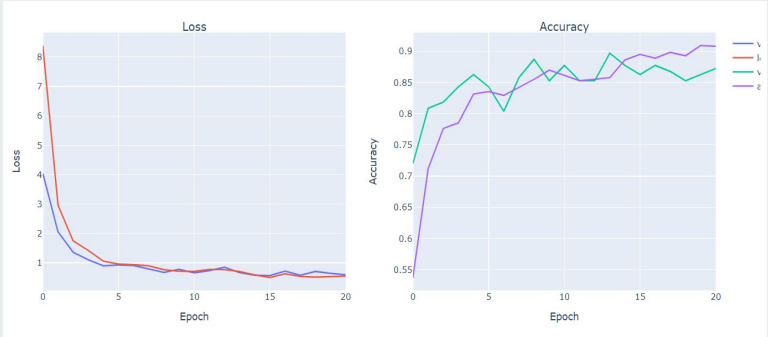
# Results

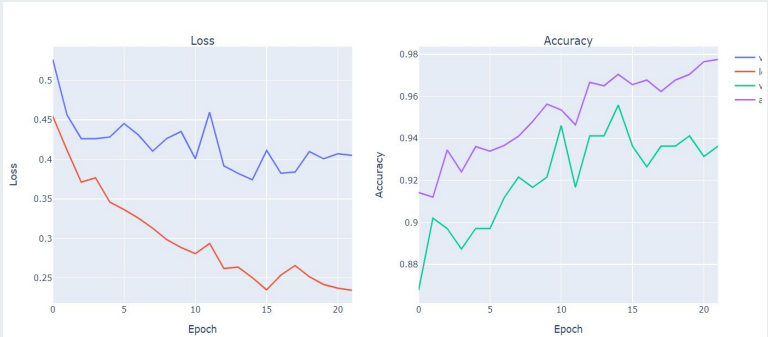
*EfficientNetV2B3*

| | Train | Val |
|---|---|---|
| Acc. | 0.8951 | 0.8627 |
| Loss | 0.5005 | 0.5661 |

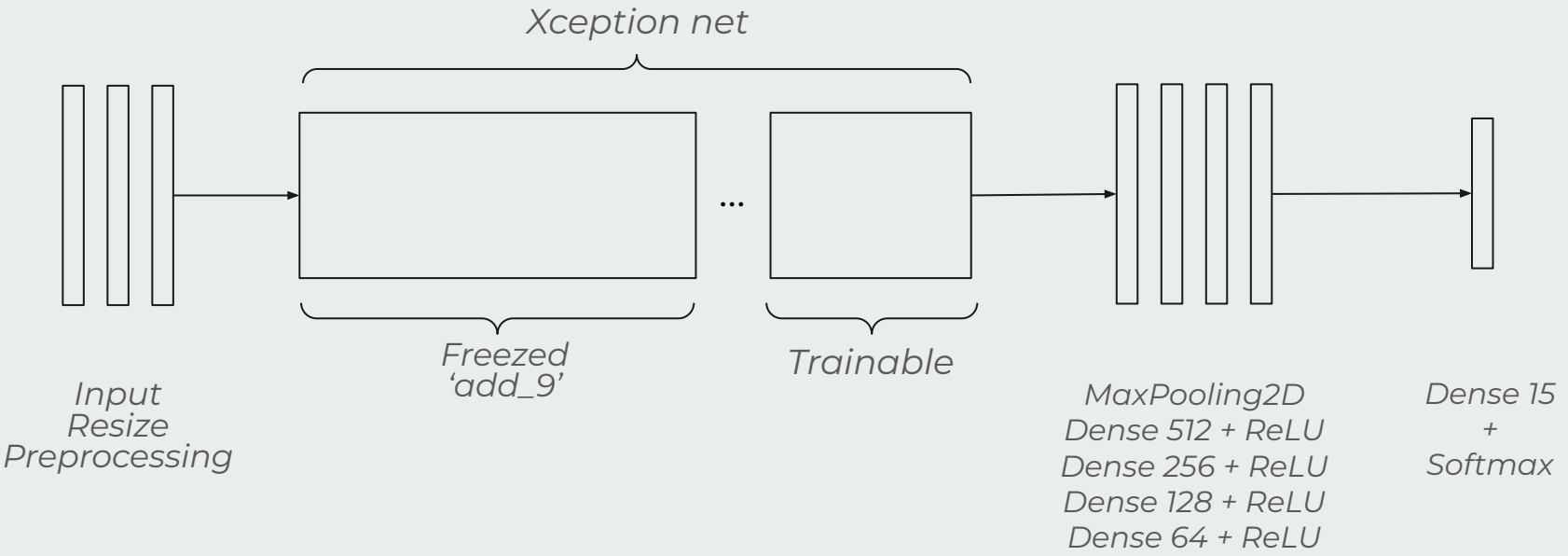
*EfficientNetV2B3 with fine-tune*

| | Train | Val |
|---|---|---|
| Acc. | 0.9705 | 0.9559 |
| Loss | 0.2503 | 0.3742 |



# Xception net



# Results

*Xception*

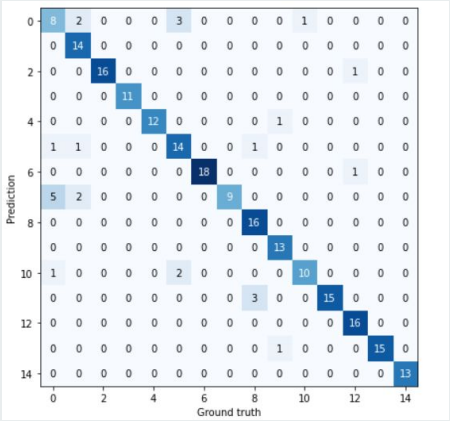
| | Train | Val |
|---|---|---|
| Acc. | 0.9656 | 0.9216 |
| Loss | 0.3819 | 0.4922 |

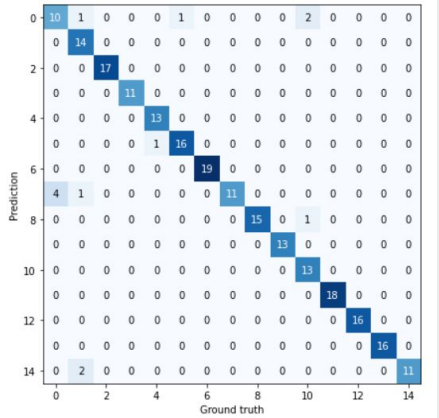
*Xception with fine-tune*

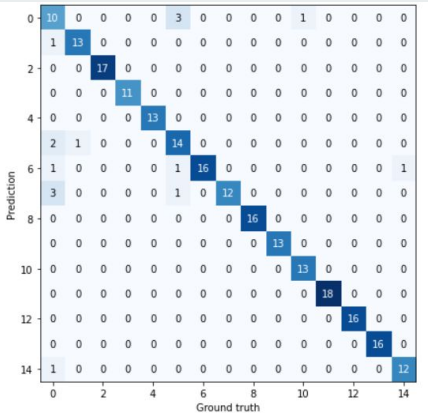
| | Train | Val |
|---|---|---|
| Acc. | 0.9814 | 0.9369 |
| Loss | 0.1968 | 0.3869 |


# Evaluation

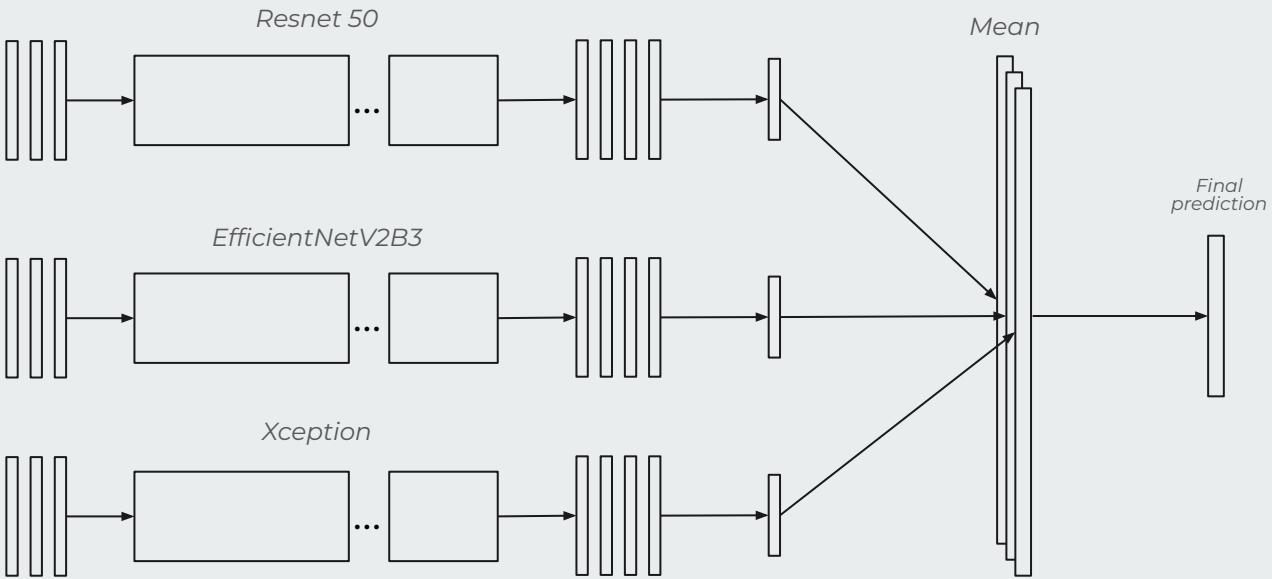
| | *Resnet 50 (*)* | *EfficientNetV2B3* | *Xception net* |
|---|---|---|---|
| Accuracy | 0.9071 | 0.9425 | 0.9292 |
| Loss | 0.4670 | 0.3415 | 0.3227 |



# Ensemble model

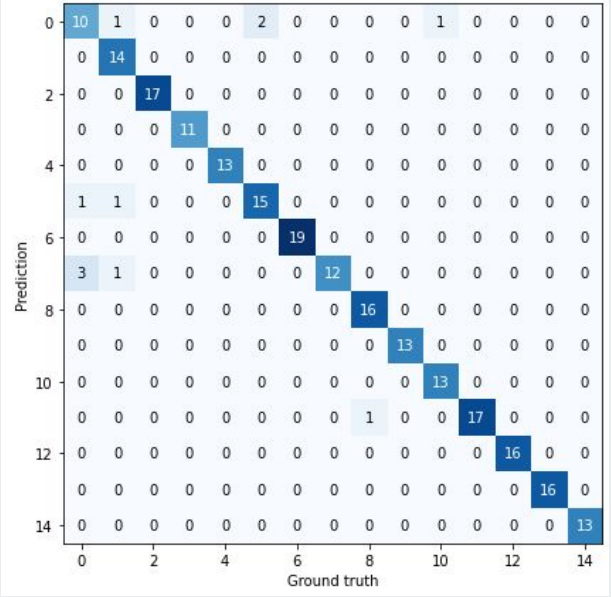


# Evaluation

*Classification report*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.71 | 0.71 | 0.71 | 14 |
| 1 | 0.82 | 1.00 | 0.90 | 14 |
| 2 | 1.00 | 1.00 | 1.00 | 17 |
| 3 | 1.00 | 1.00 | 1.00 | 11 |
| 4 | 1.00 | 1.00 | 1.00 | 13 |
| 5 | 0.88 | 0.88 | 0.88 | 17 |
| 6 | 1.00 | 1.00 | 1.00 | 19 |
| 7 | 1.00 | 0.75 | 0.86 | 16 |
| 8 | 0.94 | 1.00 | 0.97 | 16 |
| 9 | 1.00 | 1.00 | 1.00 | 13 |
| 10 | 0.93 | 1.00 | 0.96 | 13 |
| 11 | 1.00 | 0.94 | 0.97 | 18 |
| 12 | 1.00 | 1.00 | 1.00 | 16 |
| 13 | 1.00 | 1.00 | 1.00 | 16 |
| 14 | 1.00 | 1.00 | 1.00 | 13 |
| accuracy | | | 0.95 | 226 |
| macro avg | 0.95 | 0.95 | 0.95 | 226 |
| weighted avg | 0.95 | 0.95 | 0.95 | 226 |

*Confusion matrix*



# Possible improvements

*Resnet 50* →
- Random search
- Different freezed layers

*EfficientNetV2B3* →
- Random search
- Different freezed layers

*Xception net* →
- Random search
- Different freezed layers

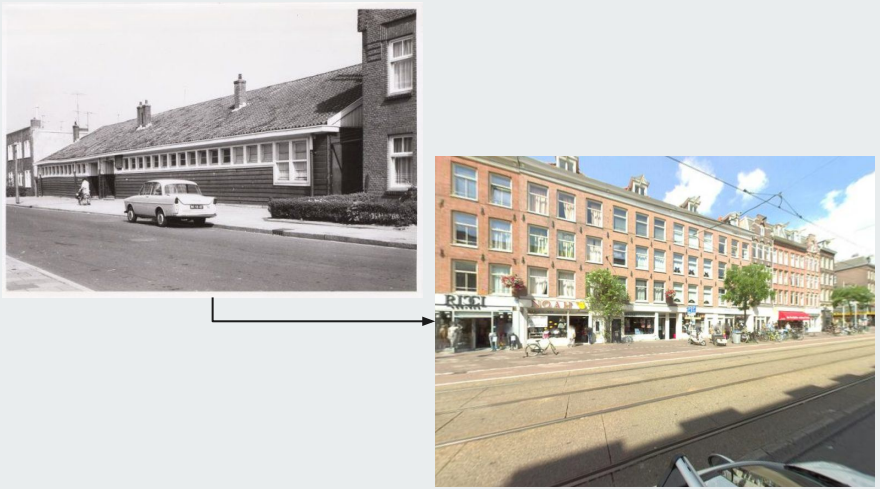
# Image retrieval

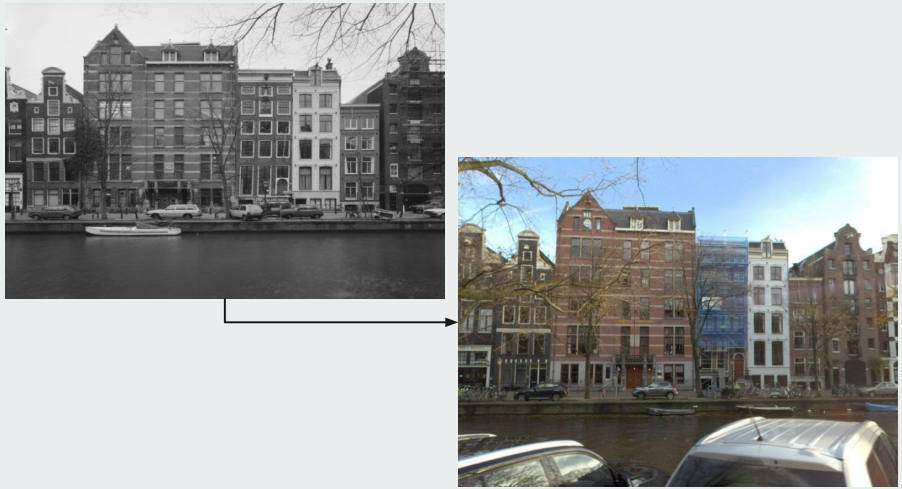
1. Dataset
2. SIFT matches
3. CNN features
4. Siamese Net
5. VLAD descriptors
6. Evaluation and conclusion

# Dataset

**AmsterTime** – A Visual Place Recognition Benchmark Dataset for Severe Domain Shift

**Struttura dataset:** 1239 Pairs of old and new image for the same place
**Tipo immagini:** Various, both RGB and grayscale



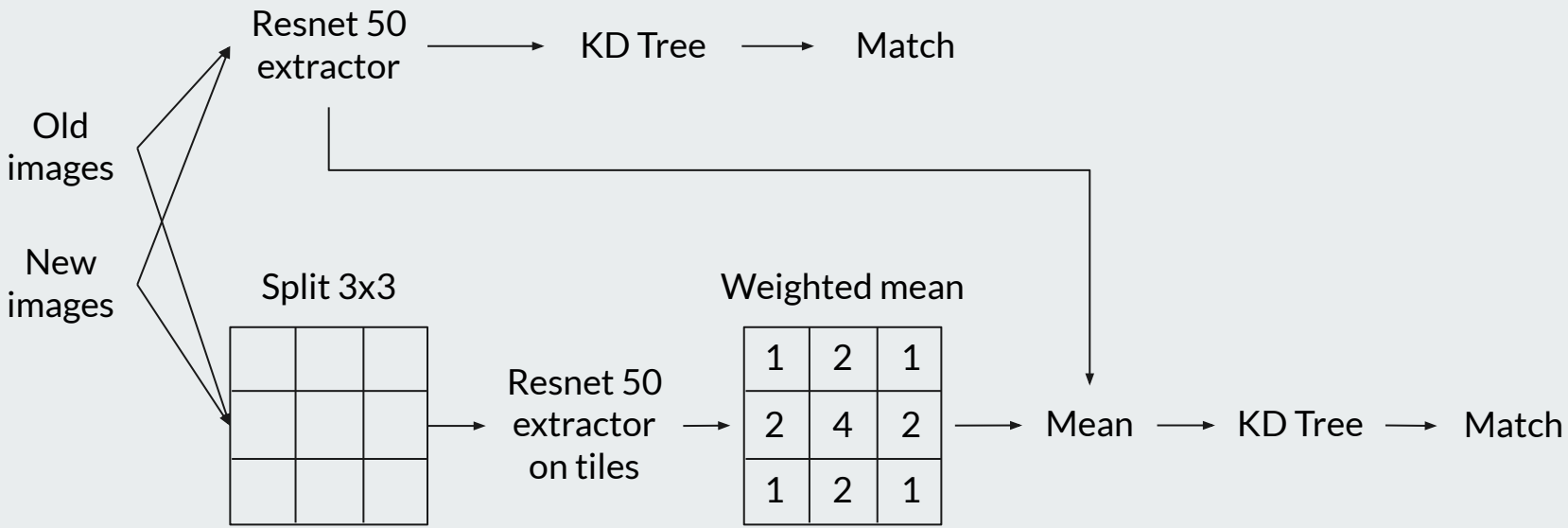


# SIFT descriptors

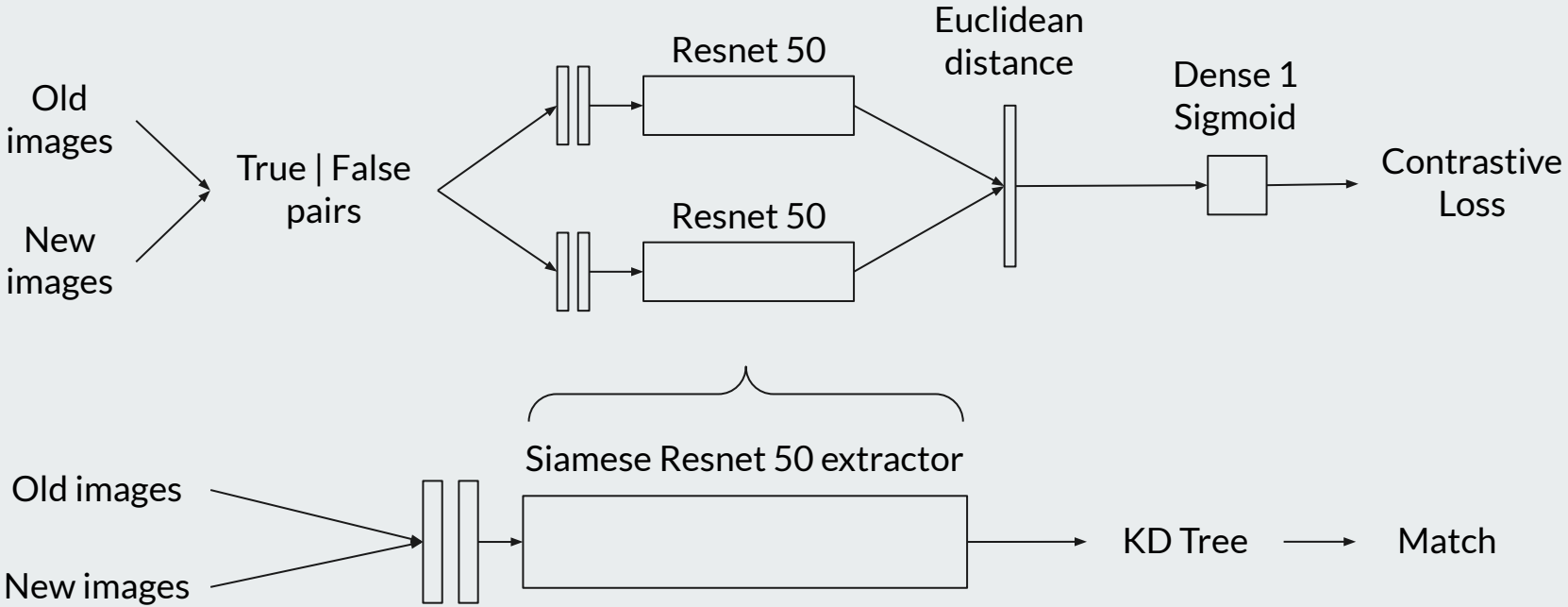

**Sorting matches metrics:**

1. *By number of matches*
2. *By sum of matches' similarity*
3. *By avg of matches' similarity*

# CNN features



# Siamese Net


Old images
New images
True | False pairs
Resnet 50
Resnet 50
Euclidean distance
Dense 1 Sigmoid
Contrastive Loss
Siamese Resnet 50 extractor
Old images
New images
KD Tree
Match

# VLAD descriptors [ https://arxiv.org/pdf/1808.05022.pdf ]

Old images  New images

SIFT extraction

KMeans
k = 128 clusters

*Sum of residuals*

$$\tilde{v}^i = \sum_{\forall x \in X : q(x) = k_i} \frac{x - \mu_i}{\|x - \mu_i\|}$$

*Normalization*

$$\hat{v}^i = \frac{\tilde{v}^i - m}{\sigma}$$

Concatenation
[ n × k × m ]

KDTree

Match

# Evaluation

| ***Method*** | ***Top-1 acc.*** | ***Top-3 acc.*** | ***Top-5 acc.*** | ***Top-10 acc.*** | ***Top-30 acc.*** | ***Top-50 acc.*** |
|---|---|---|---|---|---|---|
| SIFT [ *Metric 1* ] | 2.5 % | 5.25 % | 6.5 % | 11.0 % | 23.5 % | 29.25 % |
| CNN | 1.25 % | 2.75 % | 3.25 % | 6.5 % | 15.5 % | 20.25 % |
| Loc CNN | 1.75 % | 4.25 % | 5.25 % | 8.5 % | 19.5 % | 27.5 % |
| Siamese Net | 0.25 % | 1.0 % | 1.5 % | 2.5 % | 8.0 % | 13.75 % |
| **VLAD** | **4.5 %** | **10.25 %** | **13.0 %** | **20.0 %** | **34.5 %** | **42.75 %** |

# Possible improvements

| | |
|---|---|
| SIFT matches → | - Different KP matcher<br>- Different image comparison techniques<br>- Different distance sorting metrics |
| CNN features → | - Different networks<br>- Different freezed layers |
| Siamese Net → | - Different features distance metrics |
| VLAD descriptors → | - Different K + cluster evaluation<br>- Different cluster methods<br>- Use localized-features representation |