

# Malerhände

This is about the differentiation of illustrator in the Wenceslas bible.

## Features

### Face detection

- deepface
  - dlib
  - mtcnn
  - mediapipe
  - opencv
  - retinaface
  - ssd
  - yolov8
  - yunet
- **insightface** (also reimplemented by deepface retinaface, but that is worse)

Only insightface does properly report the face landmarks. It's also the best performing so we use that.

This topic is described in [paper\\_humanities\\_wenzelfacedetection](#).

### Face comparison features

- VGG-Face
- Facenet
- Facenet512
- OpenFace
- DeepID
- ArcFace
- Dlib
- SFace
- GhostFaceNet

# Other comparison features

- ccv
- lbp
- lpips
- hu moments

## Data

Face are detected with insightface (Tiles with 3000 tilesize, scale factor 1, and rotation factor  $0^\circ$  and  $\pm 45^\circ$ ) and face landmarks (for alignment are stored). Tags come from the groundtruth if it can be matched with detected faces.

## Transform

Derived from that we have two basic extractions, normal (no tag) and transformed (**t**).

normal	transformed
	

Transforms are certainly helpful for face comparison but should also be helpful to non-face features which are not rotation invariant. The drawback is that the face detection algorithm should provide face landmarks with which to align the faces. Not all of them provide face landmarks so this prevents the use of some face detectors, although the current best from the prior paper, *insightface*, does provide face landmarks.

## Context

This is the additional image information around the face which is included in the extracted image, either with 50 pixel context on each side from the source image (**c50**) or no additional context (**nc**), see above for **nc** examples.

c50	c50t
-----	------

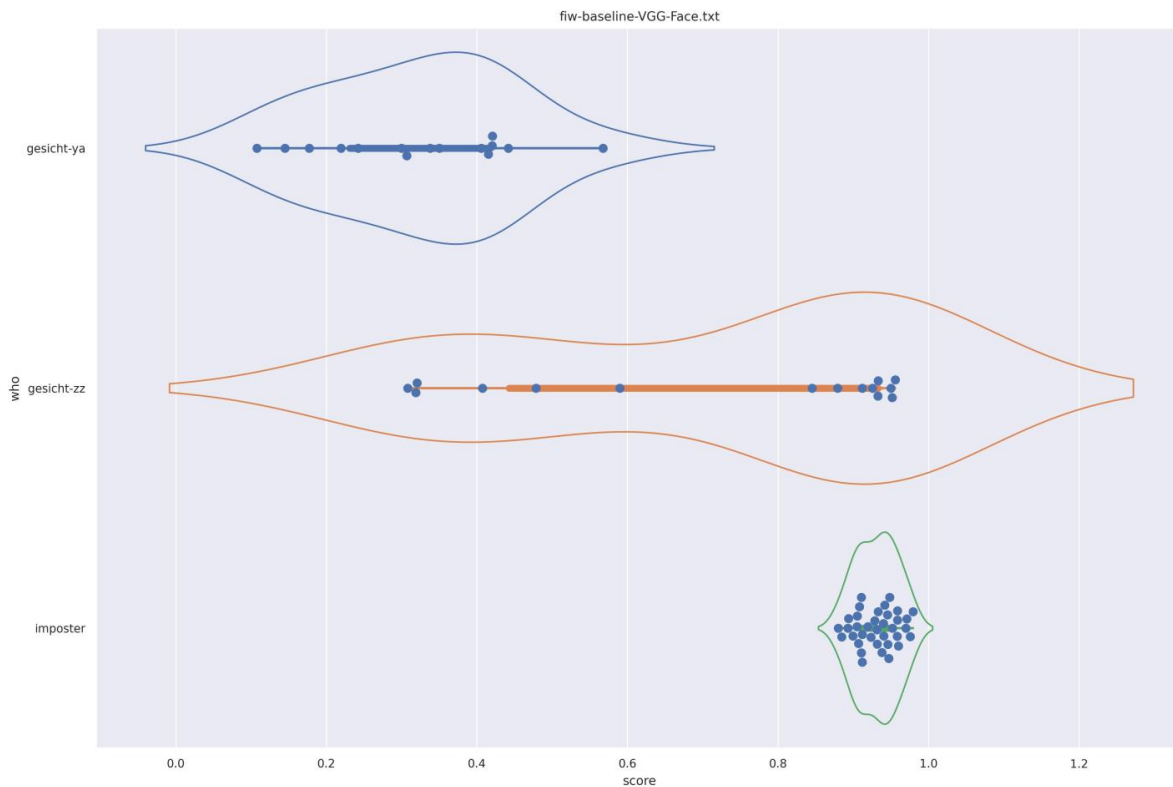


This is used for face comparison, as face algorithm often have an assumption about face to image ratios which the images without context might exceed. For texture features the extra context is likely detrimental as it will change even if the person does not.

## First Evaluation

The basic idea is this, if a method can not differentiate between two figures (Wenzel vs. Bademagd) then it is useless in trying to differentiate between potential painters of a single figure (wenzel).

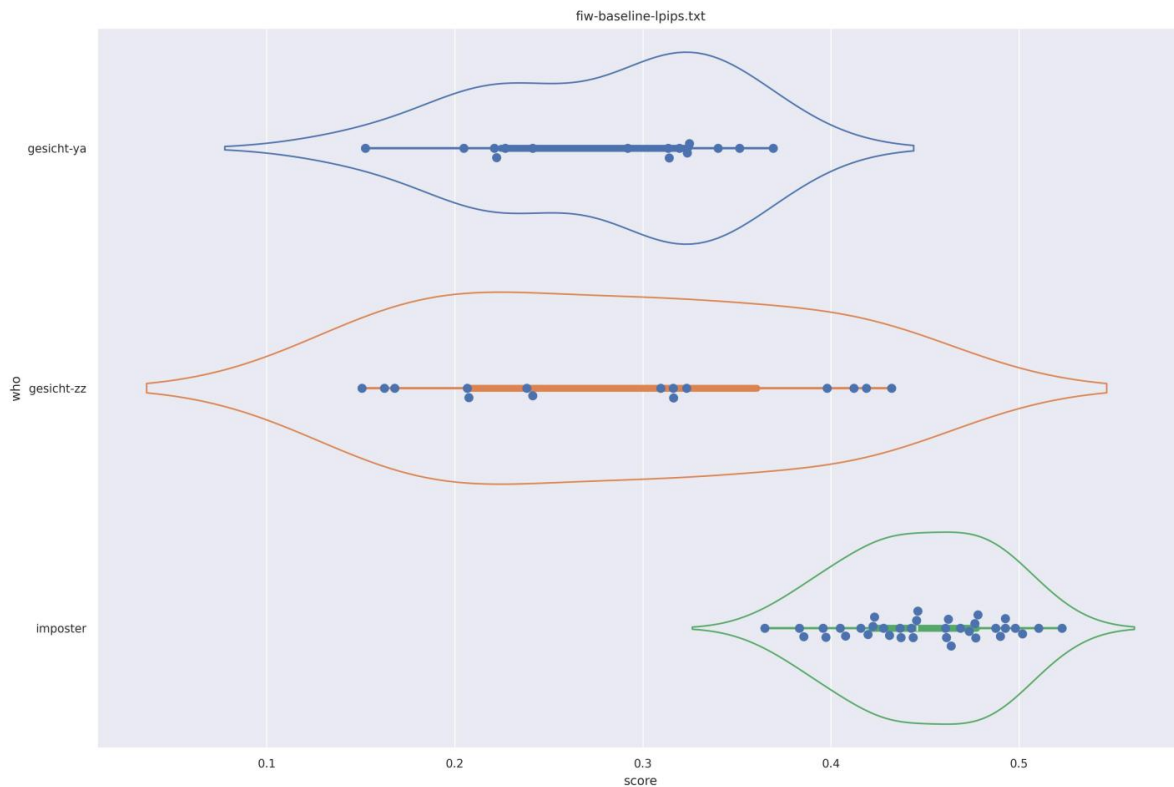
As an attempt for clarify what is expected here is a quick experiment using real faces (as we know face recognition works on them). There are two classes, one is a composed of 6 faces from a single person (Yassar Arafat) and one is composed of two persons with 3 images each (Zico and Zoran Djindjic). Between the two clases will be a real imposter comparison. The single person class (ya) has true genuine comparisons and the two person class (zz) has genuine as well as imposter comparisons.



These are the *expected* outcomes for these classes. The single person class should be well separated from the imposter distribution and the mixes class should span the range. This is a bit of an 'optimal' scenario, but what should be visible is a distribution outside the imposter distribution.

The bademagd vs. wenzel is certainly a proper imposter distribution. The bademagd class might well be, as it is not intended to be the likeness of a single person. The Wenzel class should be either genuine fi the by the same artist, as the artist endeavours to depict a single (idealized) person. In the case of mulitple painters the version of Wenceslas bewtween painters might well differ and we would get a mixed class, which should non-the-less show a distribution spanning outside the imposter range.

When it comes to the use of non-face features we can have a quick look at lpips (the best texture feature from this test) on the real images test.



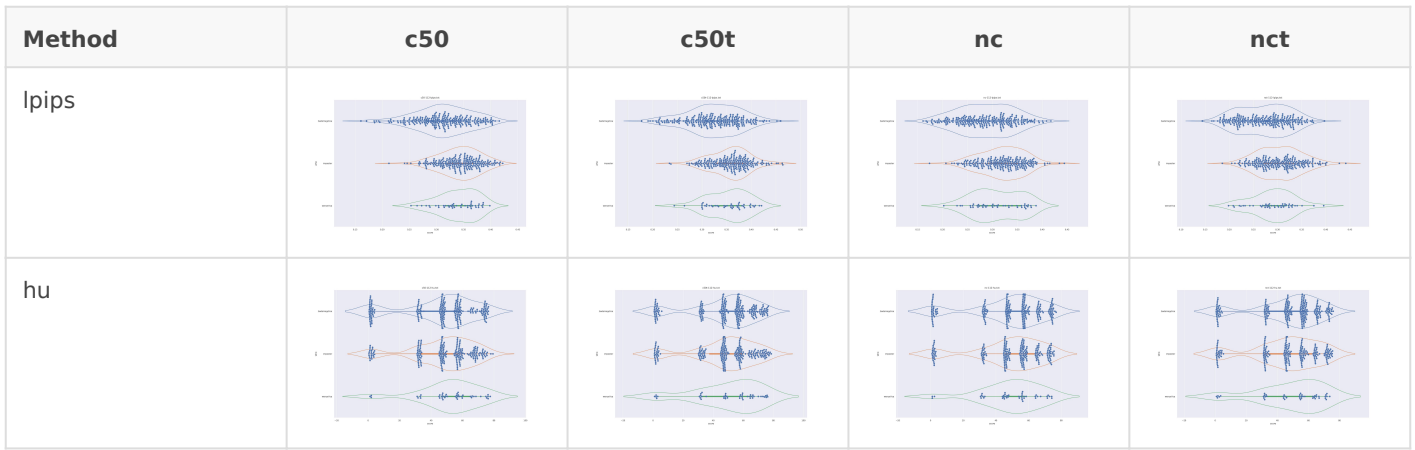
This distribution is a lot less nice but in a way reflects reality better, as the imposter and genuine distributions overlap. The basic distributions explained above however are still very much there. This also showcases that these texture features *can* differentiate faces.

## Non-face features

First let's look at the non-face recognition feature comparison. The expectation is that the transform may help, and that context does not help.

The display is bademagd on top (blue), then imposter comparisons (bademagd vs. wenzel in orange) and then wenzel at the bottom (green).

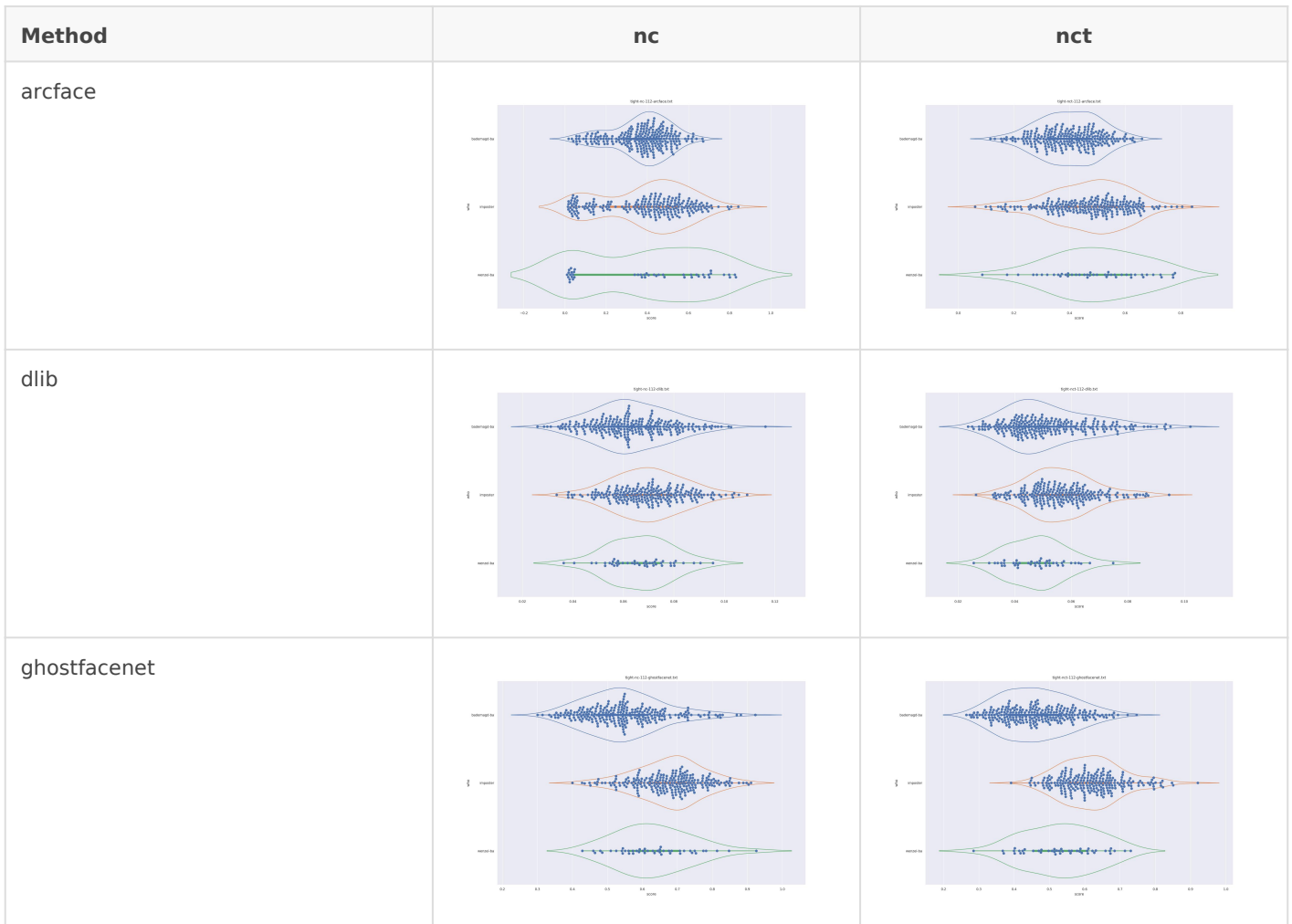
Method	c50	c50t	nc	nct
CCV				
lbp				

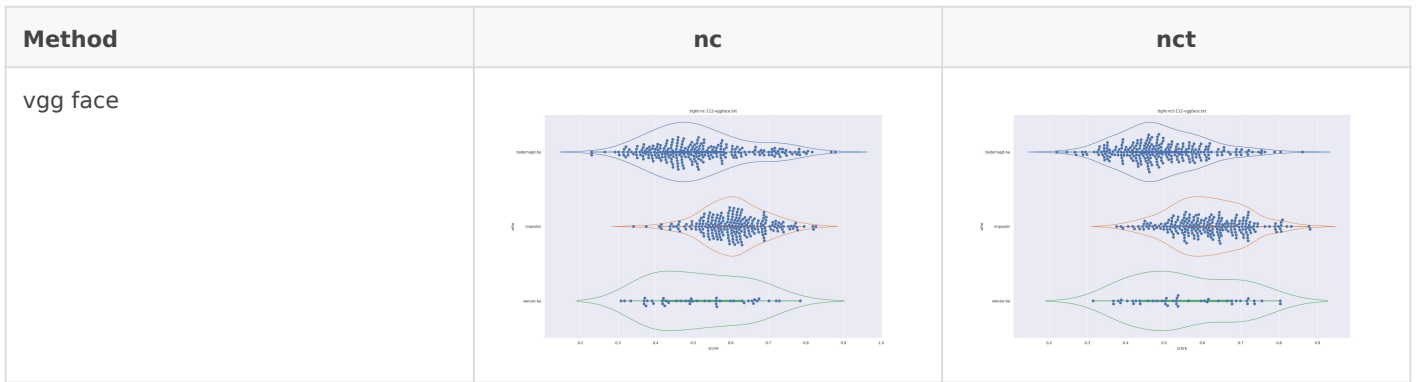


Overall all comparisons are very similar in that the genuine and imposter distributions are basically the same. In other words these methods are poor at differentiating the figures in the Wenceslas bibel. The usage of transforms or context slightly changes the distributions but the aforementioned statement still holds true.

# Face features

**This is from older data which is basically nc and nct.** *The new stuff runs very slowly, no idea why*





Overall the performance is rather bad. Transform seems to show minor improvements: pulling the wenzel distribution a bit away from the imposter distribution for ghostfacenet and resolving the clumps in the wenzel distribution for arcface (presumably those are groupings of similarly angles faces).

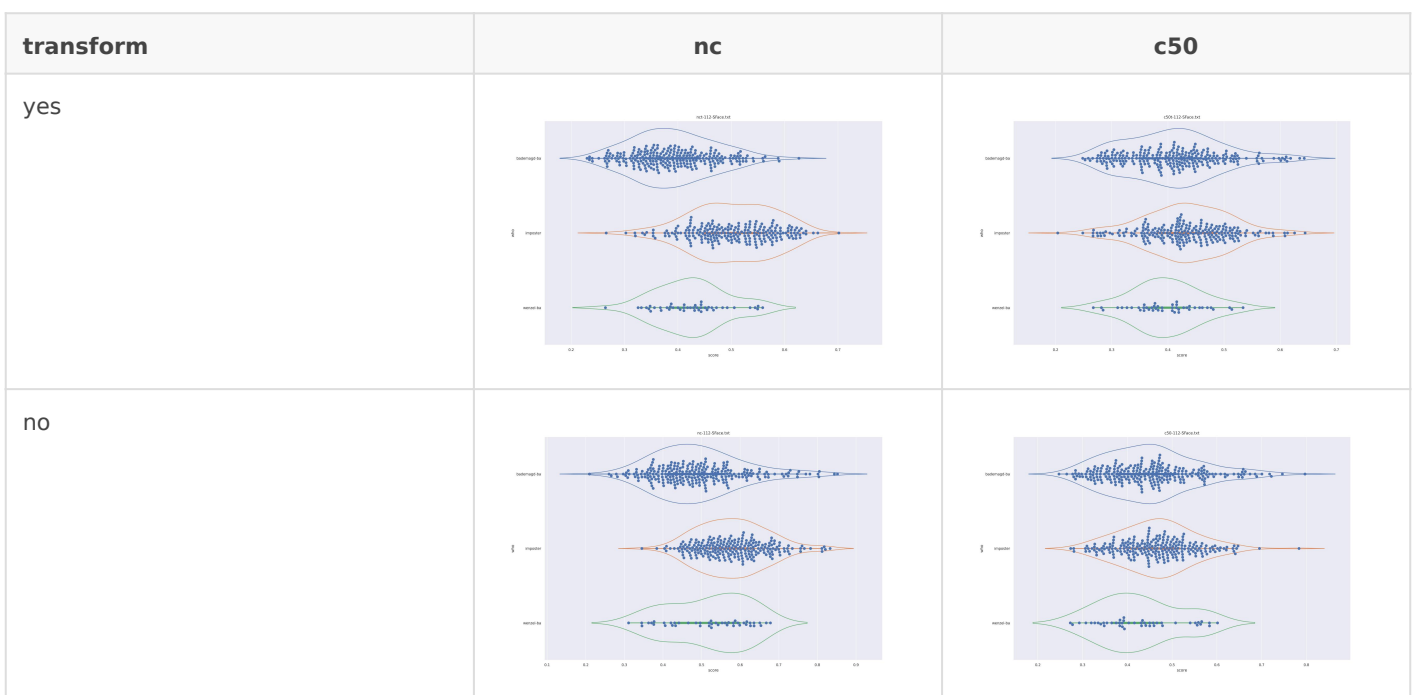
In any case the separation between imposter and genuine or mixed genuine seems to be rather bad.

Interestingly for ghostfacenet the 'bademagd' distribution seems to be better have a wider range than wenzel. In most cases Wenzel seems to span roughly the same range as bademagd which it should not, even if a couple of painters painted faces.

It could be argued here that Wenzel has not been drawn by a single painter (given expected results above). However, since we don't know if the methods used can properly separate imposter from genuine it is equally likely that the methods are simply a bad fit for the task.

The following are with the new data

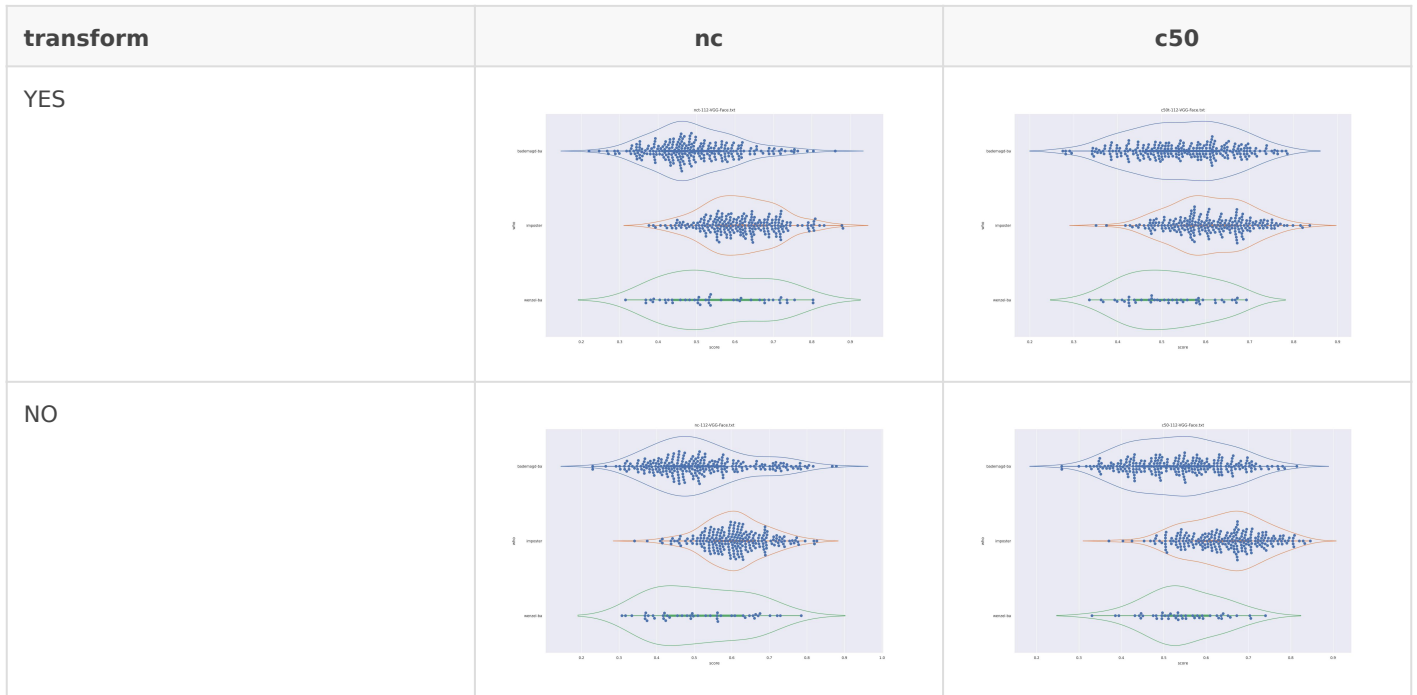
## SFace (nct)



It's quite clear that context is detrimental to the performance here. The use of transformation seems less useful as barely a difference is in overall distribution is shown. There are minor changes but overall the behaviour does not seem to change.

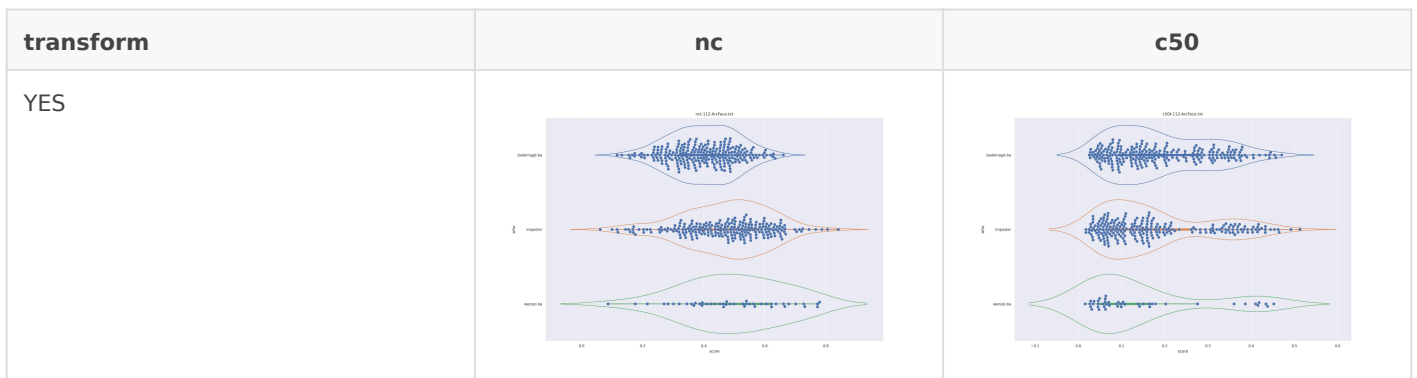
The version without transforms does show a bit of a split between values (the hump in the kde estimate of the distribution). It is minor and only appears in the non transformed version. As such it might well be the difference in angle rather than a difference in features. For this reason the use of transformations to align faces seems to be best.

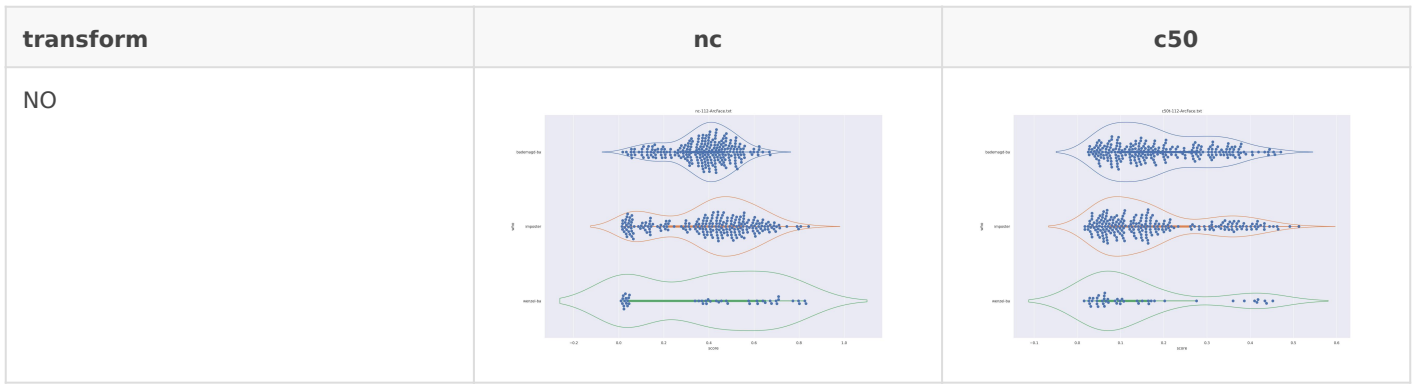
## VGG Face (*any*)



Not much different from SFace, they look all rather similar and no clear separation of impostor and wenzel distributions. Interestingly, no-context rather has some bademagd faces move away from the impostor distribution.

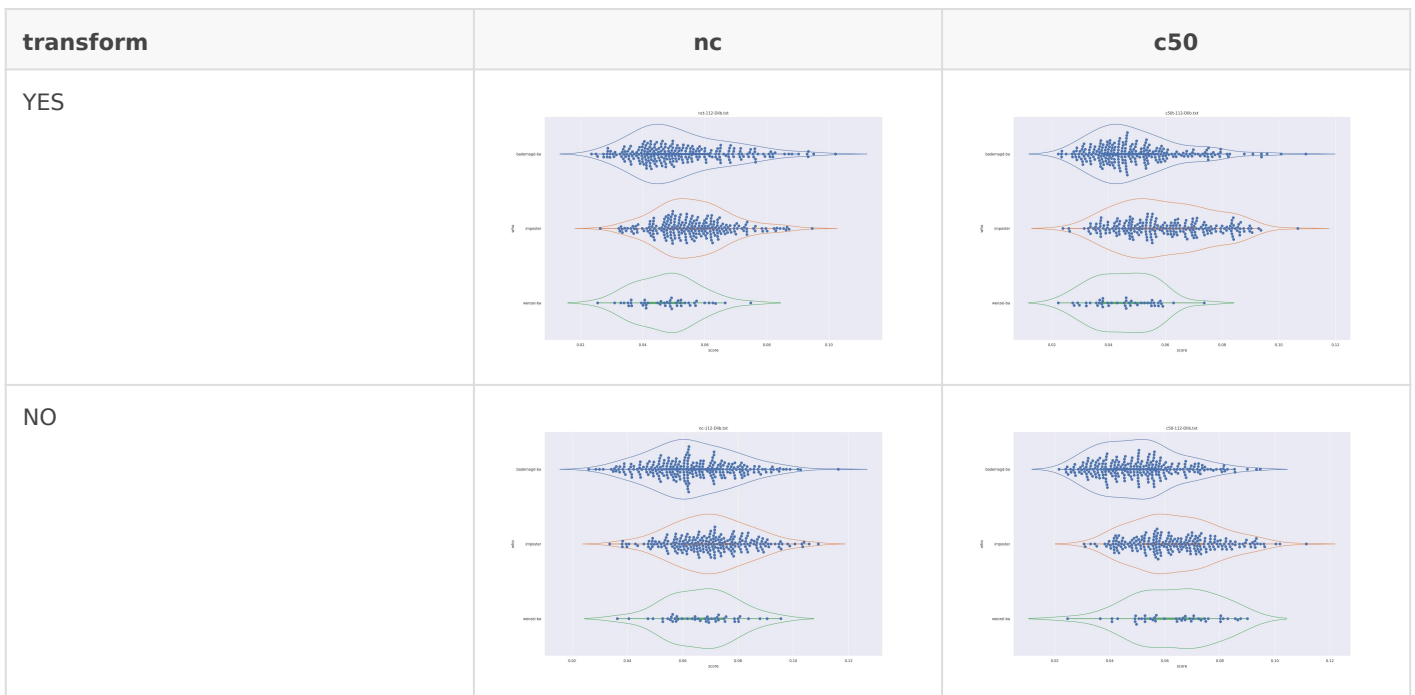
## ArcFace (*nct*)





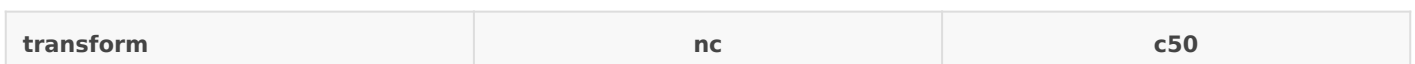
So far interesting as we see the two humps which would indicate a double distribution (genuine and imposter). But confusingly this is present in imposter also so clearly must have a different origin. The no context with transform seems to be the most 'normal' distribution so the nct version should be preferred.

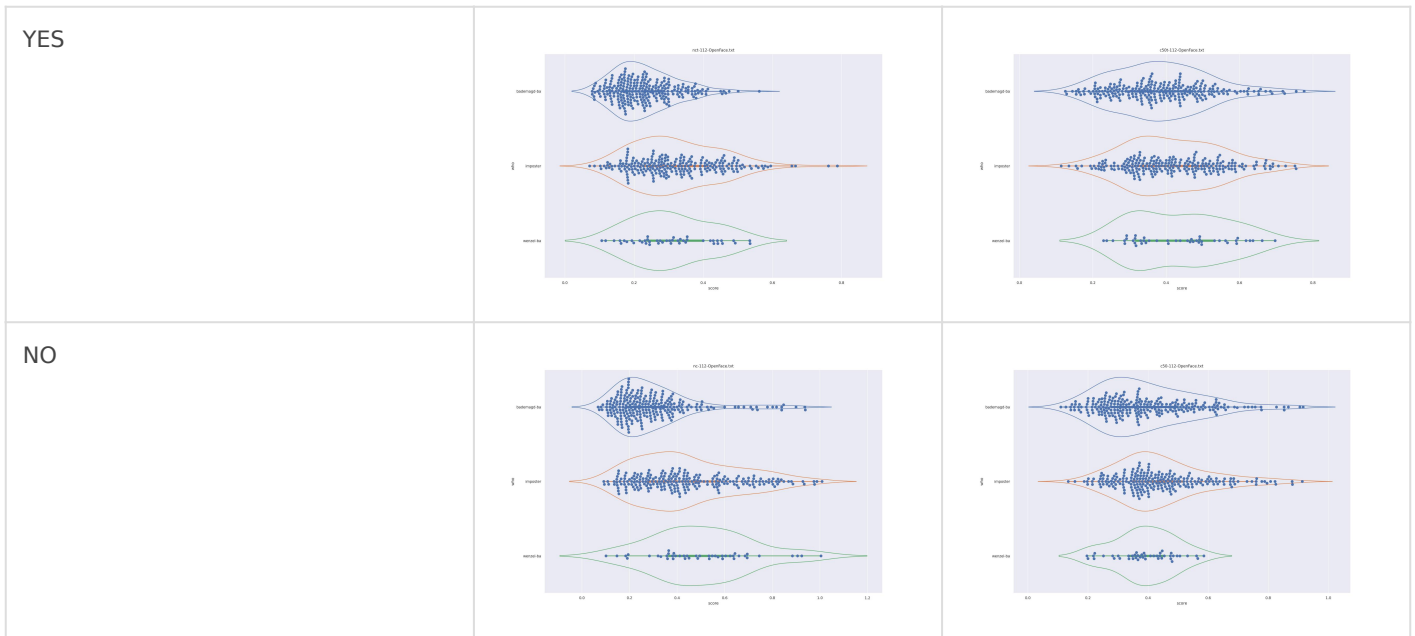
## Dlib (*any transformed*)



Again not much difference between options. If anything the wenzel distribution seems to be more narrow when using transforms (which makes sense as the transformation aligns the faces) so should be preferred.

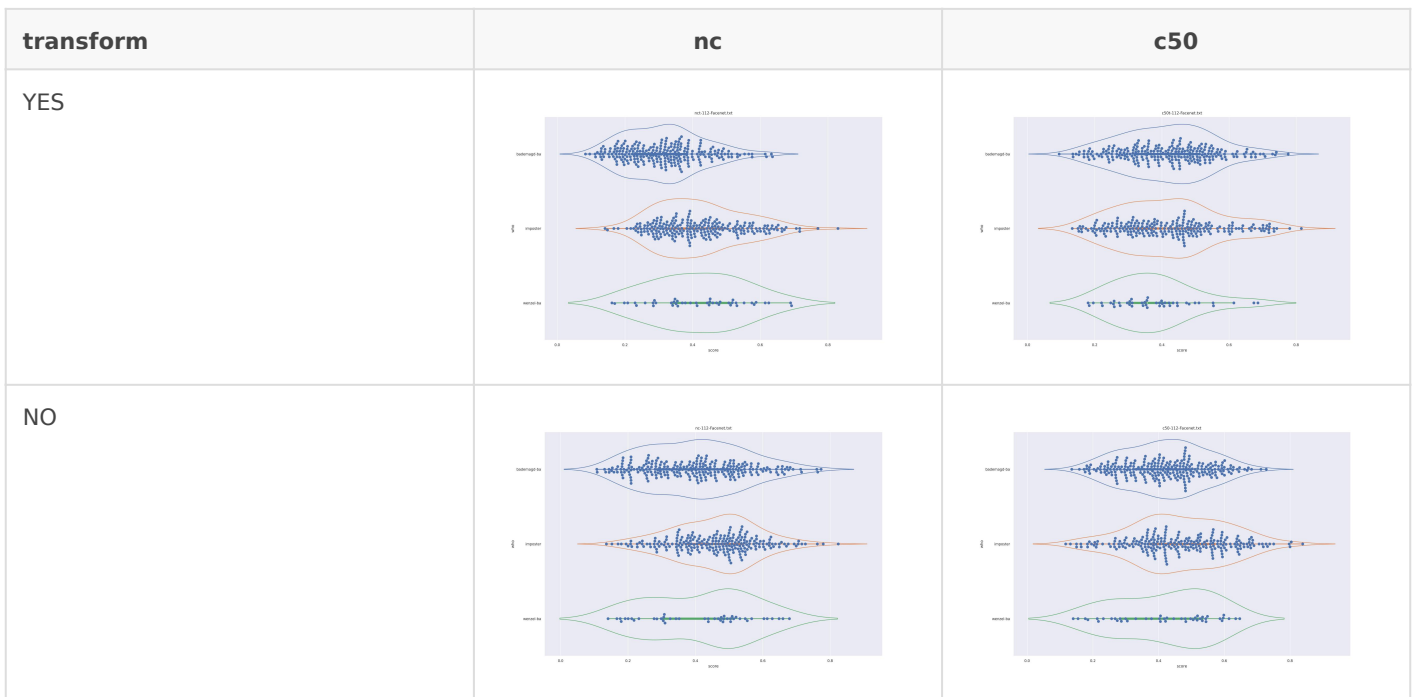
## OpenFace (*any*)





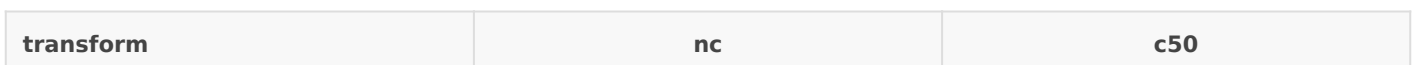
Overall not much of a difference.

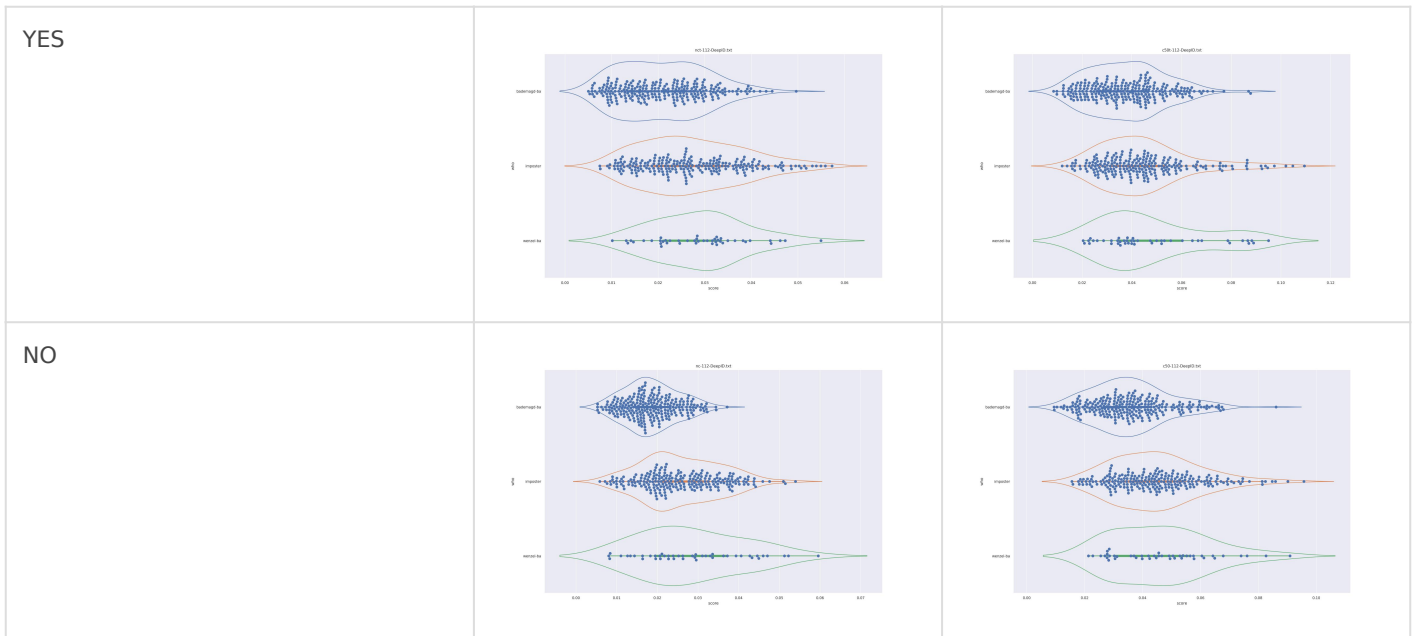
## FaceNet (*any transformed*)



The double humps (presumably due to face rotation) are removed from transformed faces so this is preferable. As for nc or c50, not clear distinction is visible.

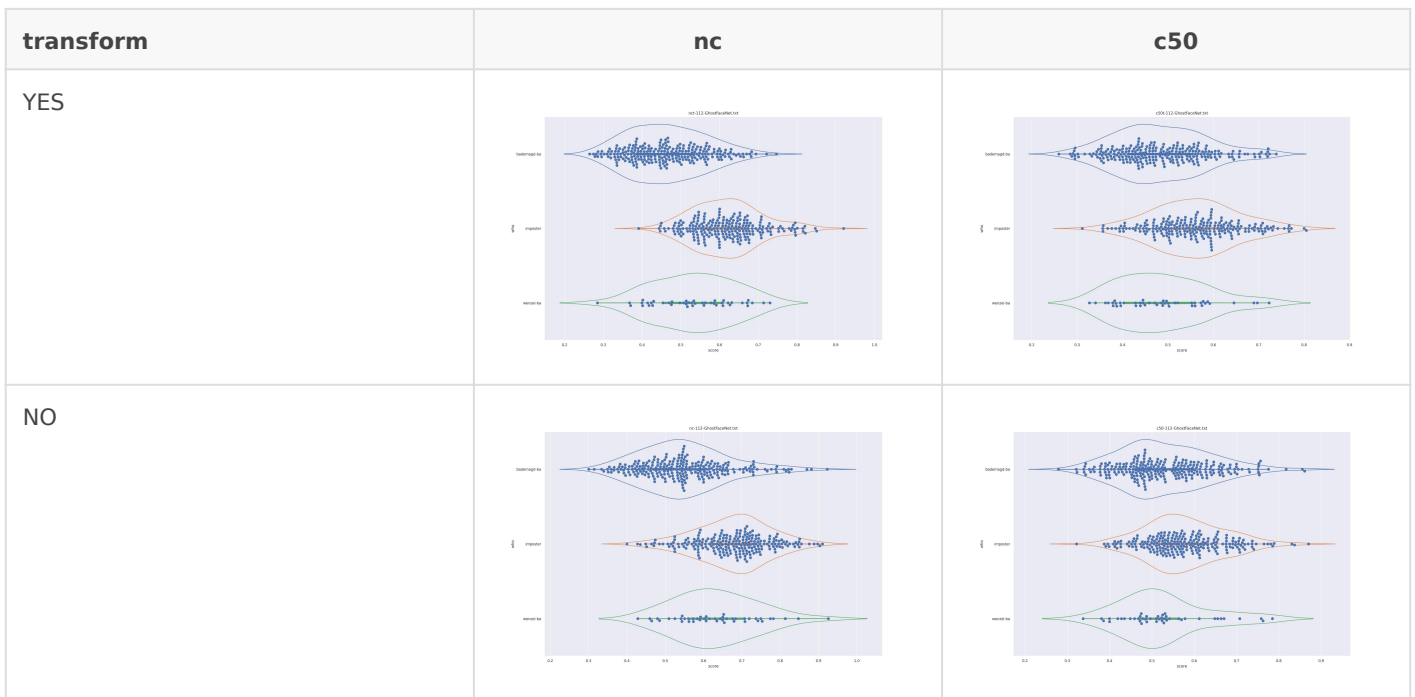
## DeepID (*any*)





No real differences I would say.

## GhostFaceNet (*nct*)



The transformed modes seem to slightly better separate the imposter and wenzel distribution, the nc does so even more.

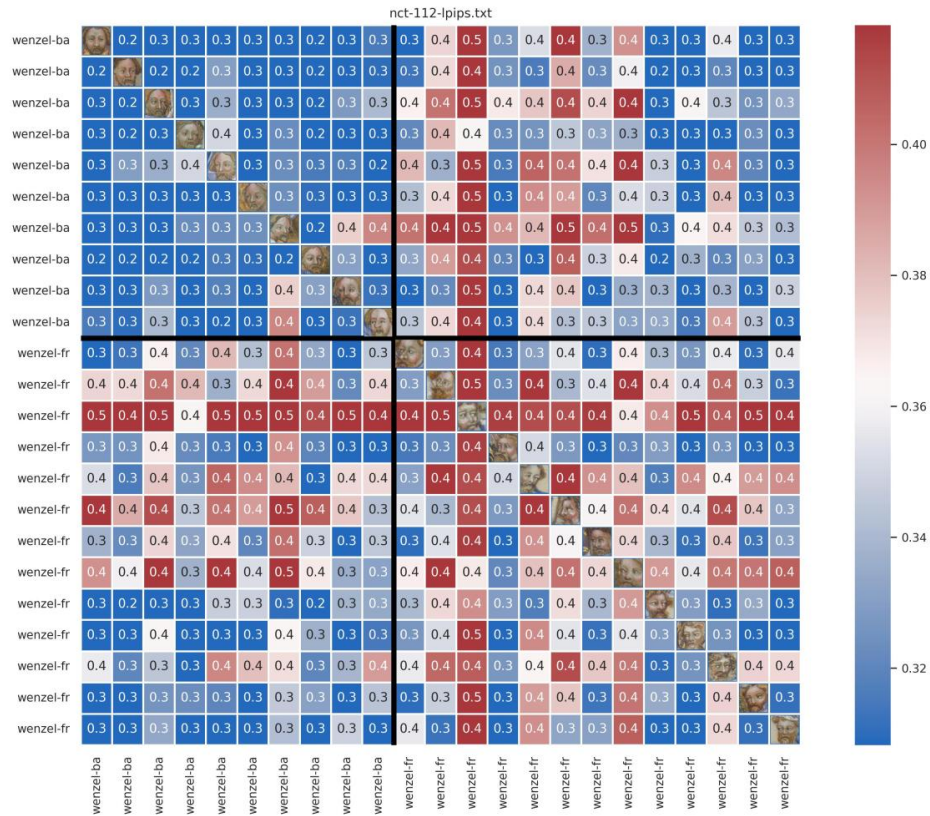
## Conclusion so far

The consensus seems to be that *nct* is the best option for all.

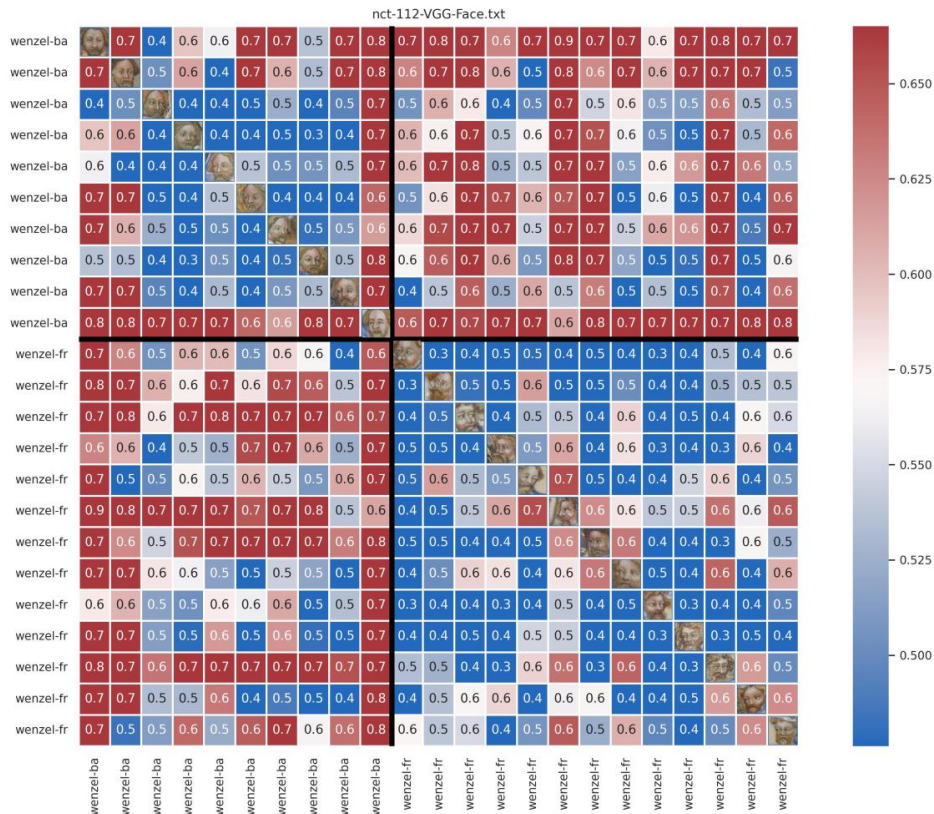
# A different view --- Cherry Picking

I randomly took -ba above and while looking ad comparison of wenzel faces between painters saw that the intra distribution depends strongly on painter combined with feature:

## FR vs. BA using Iperms and VGG-Face



These are the lperms results. A very cohesive Balaam, less so for Frana. Interestingly the Frana vs. Balaam look very similar to the intra-Fr comparisons.



Same experiment with VGG-Face. Pretty much the reverse of the IPIPs results. Internally cohesive Frana with less cohesive Balaam. Although it has to be said that intra Balaam looks more cohesive than inter comparisons.

These two examples show that we can be prone to cherry picking here. Since we have no real objective way to specify which method is superior (we don't have a ground truth on which to compare) we have to be careful not to cherry pick the method based on the desired outcome. A possible way forward would then be to not use a single one but use all (or at least many) of them in an ensemble and vote on it. The uniformity of the vote can also be used as a sort of 'certainty' score for the decision.

## Using EER

Well assuming Wenzel is genuine and bademagd is not wenzel we can calculate the EER to find which feature is best. Fr is included as we can be relatively certain that the frana pages are actually from the frana workshop.

FILE	BA	FR
c50-112-ArcFace.txt	42.19%	<b>18.82%</b>
c50-112-ccv.txt	43.75%	46.15%
c50-112-DeepID.txt	52.97%	52.81%

FILE	BA	FR
c50-112-Dlib.txt	54.38%	36.30%
c50-112-Facenet.txt	46.67%	28.17%
c50-112-GhostFaceNet.txt	33.33%	<b>24.56%</b>
c50-112-hu.txt	50.44%	50.11%
c50-112-lbp.txt	42.05%	42.38%
c50-112-lpips.txt	50.33%	40.42%
c50-112-OpenFace.txt	44.44%	47.61%
c50-112-SFace.txt	37.78%	30.77%
c50-112-VGG-Face.txt	28.73%	<b>20.58%</b>

FILE	BA	FR
c50t-112-ArcFace.txt	39.27%	38.13%
c50t-112-ccv.txt	37.43%	47.67%
c50t-112-DeepID.txt	47.19%	47.53%
c50t-112-Dlib.txt	33.63%	47.08%
c50t-112-Facenet.txt	44.44%	29.49%
c50t-112-GhostFaceNet.txt	33.33%	37.79%
c50t-112-hu.txt	49.55%	50.17%
c50t-112-lbp.txt	45.09%	48.21%
c50t-112-lpips.txt	49.02%	42.86%
c50t-112-OpenFace.txt	53.33%	48.72%
c50t-112-SFace.txt	37.14%	48.65%
c50t-112-VGG-Face.txt	33.33%	33.03%

FILE	BA	FR
nc-112-ArcFace.txt	48.26%	<b>21.88%</b>
nc-112-ccv.txt	51.11%	38.01%
nc-112-DeepID.txt	53.97%	61.95%
nc-112-Dlib.txt	46.01%	34.86%
nc-112-Facenet.txt	48.89%	27.86%
nc-112-GhostFaceNet.txt	37.83%	25.64%
nc-112-hu.txt	50.11%	48.97%

FILE	BA	FR
nc-112-lbp.txt	45.82%	48.95%
nc-112-lpips.txt	42.22%	46.15%
nc-112-OpenFace.txt	57.78%	39.63%
nc-112-SFace.txt	42.22%	28.40%
nc-112-VGG-Face.txt	32.73%	<b>16.05%</b>

FILE	BA	FR
nct-112-ArcFace.txt	50.00%	33.27%
nct-112-ccv.txt	46.67%	43.59%
nct-112-DeepID.txt	57.06%	45.39%
nct-112-Dlib.txt	36.83%	47.21%
nct-112-Facenet.txt	52.73%	39.23%
nct-112-GhostFaceNet.txt	35.56%	35.90%
nct-112-hu.txt	49.88%	49.25%
nct-112-lbp.txt	48.89%	50.53%
nct-112-lpips.txt	46.67%	47.34%
nct-112-OpenFace.txt	50.55%	47.83%
nct-112-SFace.txt	26.91%	43.59%
nct-112-VGG-Face.txt	40.00%	28.60%

Frana produces better results here, interestingly with different connotations than what I would have read from the graphs above (which were BA). Here it seems that transformations are detrimental and that context is better for some.

Overall FR seems to have a much better EER, some are almost decent.

---

Revision #42

Created 5 March 2025 14:10:01 by hh

Updated 22 May 2025 09:32:18 by hh