

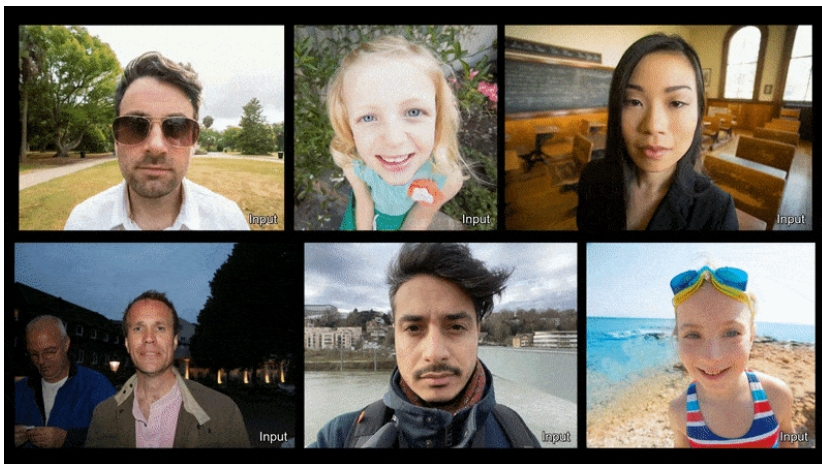
Correcting 'Selfie'-Based Facial Distortion, for Psychological and AI Development Purposes

Originally published at <https://blog.metaphysic.ai/correcting-selfie-based-facial-distortion-for-psychological-and-ai-development-purposes/>

February 28, 2023

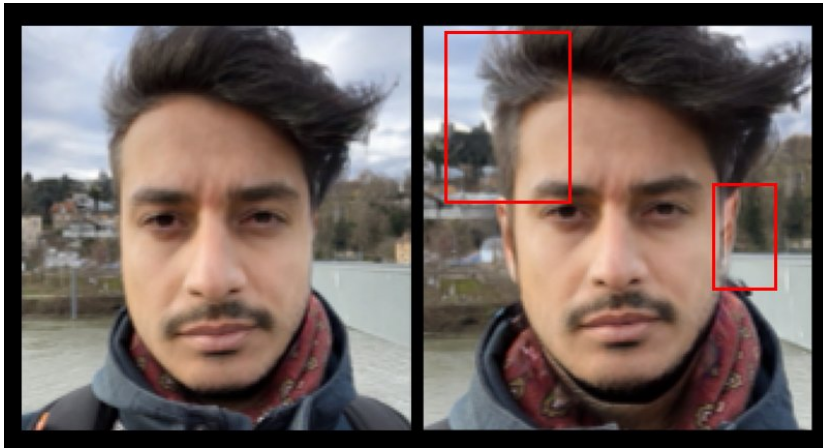


A new collaboration between Japan, Taiwan and the US offers a novel approach for 'flattening out' or foreshortening selfies that distort faces, because they are taken too near to the subject (i.e., at only arm's length):



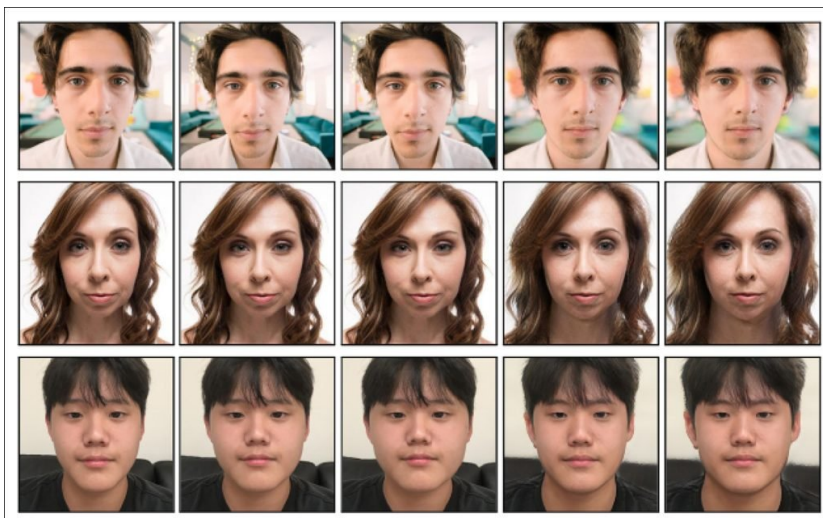
Examples of DISCO correcting distorted views of people. Source: <https://portrait-disco.github.io/>

Titled [DISCO](#), the new system is capable of actually restoring occluded parts of the face – features which have been hidden by the wide-angle nature of the lens in a typical portable device. The system even uses generative frameworks such as [Stable Diffusion](#) and DALL-E 2 to 'inpaint' newly-exposed background regions that may result from the synthesized alterations.



This DISCO conversion, enlarged from source materials in the project's video resources, essentially adds geometry that was not present in the picture, such as a higher volume of hair content, greater detail around the scarf/neck area, and ears that were entirely missing in the original photo.

DISCO uses a 3D-aware version of a [Generative Adversarial Network](#) (GAN) to obtain a more complete relationship than prior approaches between a face that is trained into a framework and the coverage (with associated distortion, the more coverage there is) provided by the lens on the capture device.



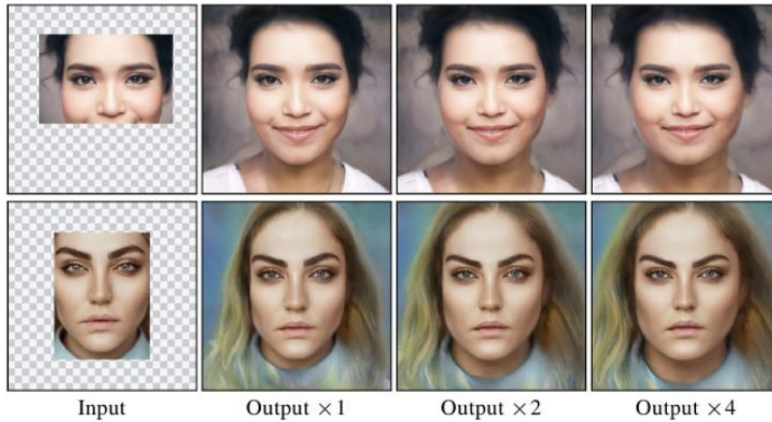
From the new paper, examples of the perspective foreshortening achieved by DISCO. Source: <https://arxiv.org/pdf/2302.12253.pdf>

In addition to facilitating an artificial [foreshortening](#) of perspective, DISCO's deeper understanding of this relationship also allows for more accurate editing and facial completion in GAN-based architectures – both hot pursuits in image synthesis.

Editing



Face completion



Editing and face completion are improved by DISCO's superior mapping of the relationship between the camera's field-of-view and the resulting face data, the researchers of the new paper claim.

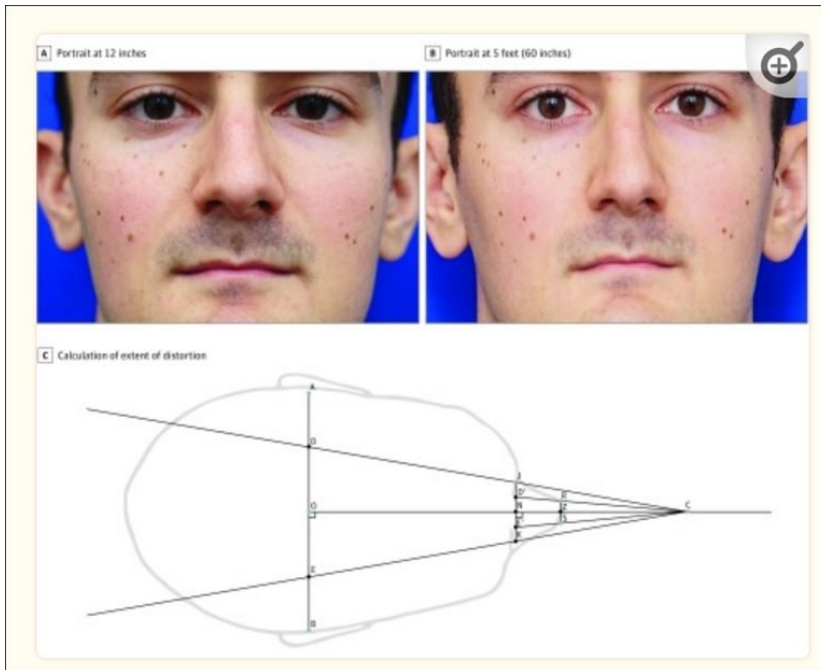
In theory, a system such as this could be used to 'equalize' or normalize all photos in a training dataset, so that – despite the wide variety of sources from which web-scraped data is obtained for hyperscale training – both the source faces and their ultimate application would be consistent among themselves.

A series of quantitative and visual comparisons on the new system prove, the authors of the new work assert, that DISCO represents improved performance over existing methods, and therefore it would currently seem to be the state-of-the-art in 'selfie fixing'.

The [new paper](#) is titled *Portrait Distortion Correction with Perspective-Aware 3D GANs*, and comes from seven researchers across various institutions, including the University of Tokyo, Japan's National Institute of Informatics (NII), Taiwan's National Yang Ming Chiao Tung University, the University of Maryland, and the image-focused US technology company Snap Inc..

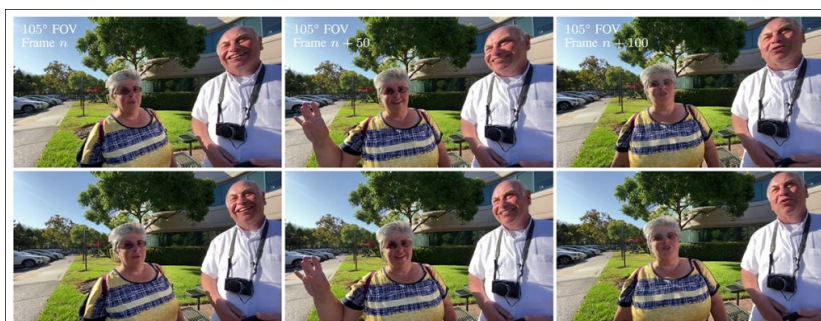
The Need for a Clearer View

Perhaps surprisingly, this is a fertile and well-followed investigative trend in computer vision. For one reason, the 'selfie effect' has been [shown](#) in recent times to be a psychologically destabilizing influence on some people, who tend to view the inevitably warped stance of self-taken (i.e., hand-held) self-portraits as a new universal standard in appearance – despite the obvious ways that this set-up does not represent how the person is typically seen by other people (unless the other person is seeing you from an extraordinarily close angle).



A 2018 paper from the US emphasized the extent to which individuals may become dysmorphic due to being 'misrepresented' by the typically wide-angle lenses and close-quarters distortion of a selfie. Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876805/>

Indeed, as the paper notes, there is an [active level of invention](#) in the academic and industrial research community that's aimed at remediating the problem, such as one pending patent for software that can change the field-of-view (FOV) effects [based on the camera disposition and orientation](#). For the purposes of computer vision, distorted selfies are a mixed blessing. On the one hand, their ease-of-use and wide accessibility mean that the amount of face-based material available for AI training has massively increased since the smartphone era began. On the other hand, the fact that so many facial images are taken in this way is effectively setting a new standard of facial representation which is specific to 'selfie-culture', rather than generally useful for a wider application in computer vision technologies such as generative image synthesis and facial identification systems that will work on smartphones (where the faces are very close) and on general, more 'distant' placements (where the capture equipment is further away from the subject, and their face will be more evenly represented).



A 2021 paper from Google offered a method of correcting the distorted faces that result when people dwell near the periphery of a wide-FOV shot, where distortion is maximized. Source: <https://arxiv.org/pdf/2111.09950.pdf>

For this reason, there are [currently calls from researchers](#) for increased use of metadata in machine learning training sets that feature faces, so that the amount of distortion that can be expected in a face image will be a more governable and rational factor for training new models. In professional photography, this problem is easy to avoid, since subjects will be placed centrally, and, in cases where 'median' reproduction of faces is desired, a 50mm lens (or equivalent) will be used, since this is the closest objective focal length to the way that the human eye operates, and represents a 'default' or base human view. Conversely, a 50mm equivalent lens on a typical smartphone would capture only a portion of the face.



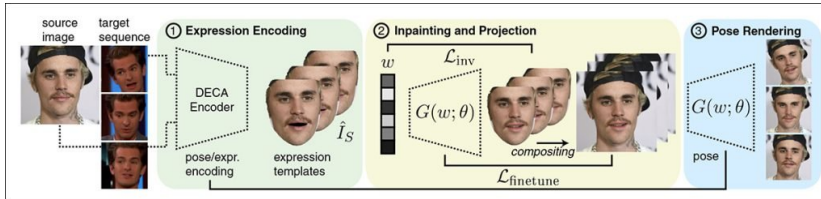
At the highest FOV, or 'focal length' of a camera lens (right, in image above), the face is flattened into near-isometric low-relief. As the FOV widens, distortion increases. The appearance of the individual is severely altered at the widest FOV. In the second part of the captions, we can see that the camera is much further away at the start, and much nearer by the time the image is severely distorted. For a famous example of this 'push/pull' syndrome, see the 'Jaws' image below. Source: <https://archive.is/uU8jl> (title of sub-Reddit in archive link contains NSFW word)

Parsing a large number of such 'distorted' faces through training for generative systems such as Stable Diffusion will result in systems that are most familiar with these skewed views, and most disposed to reproduce them, if such views occupy a majority of the data, and cannot be 'balanced' by parallel data that has more neutral perspective.

Similarly, facial ID systems are frequently inflexible, most especially when they have been specifically designed for smartphone ranges which are likely to have a limited and wide series of FOV coverage over the lifetime of the product. The data obtained from such systems is only likely to be transferable to other similar systems, because faces change so radically in appearance as FOV changes.

Tackling 'Selfie-Warp'

DISCO uses 3D [GAN inversion](#) to correct portrait distortion. GAN inversion in general is the process of 'projecting' novel data into a trained generative network so that it can benefit from the network's acquired knowledge about the facial domain, and have transformations performed on it.

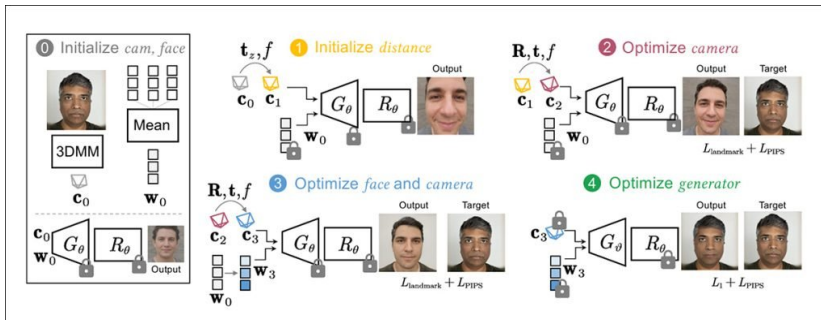


From a 2022 Stanford paper, the process of 3D GAN inversion allows genuine spatial modeling of face-based entities trained into the network. Source: <https://arxiv.org/pdf/2203.13441.pdf>

A 3D-aware GAN takes account of additional factors besides the 2D representation of trained images, so that the system has some conception of volume. This facilitates changes in the 3D X/Y/Z coordinate space, and allows for deeper transformations than just style transfer or minor modifications within a pixel-based [latent representation](#) of a face.

DISCO improves upon prior methods, such as *Pivotal Tuning for Latent-based Editing of Real Images* ([PIT](#)), by separating the optimization of face and camera parameter information during training. Optimization in this sense is equivalent to 'fitting' – the process of conforming related but very different data so that the final latent codes have high instrumentality, and factors such as field-of-view become [disentangled](#) from the face data that they affect.

This separation is initially achieved by mapping the real face data to a virtual, old-school parametric CGI model, called a [3D Morphable Model](#) (3DMM). 3DMMs are commonly used as a relatively quick and cheap method of mapping flat pixel images into 3D space.



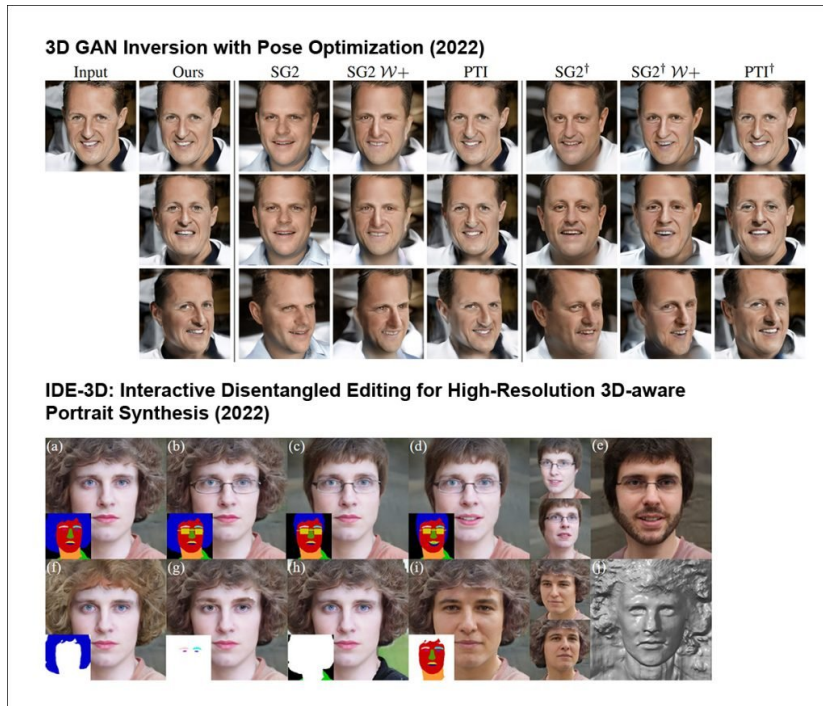
After this, as we can see in the image above, the distance between the camera lens and the image is parametrized, based on known data (though this can also be based on metadata in images which may contain such information, which can be made into an explicit training stream during pre-processing of the data).

Then, as we can see in the middle lower section of the image above, the aforementioned parallel optimization occurs, before similar processes are applied also to the virtual camera and the generator module that will finally output the altered images.

Approach

Though a number of prior approaches have used camera parameters to control apparent FOV in rendered images, these have side-stepped some of the more complex problems involved, by avoiding face images that have the perspective problems which come with typical selfie set-ups (bigger noses, disappearing ears, distorted general appearance).

Effectively, such systems have defaulted back to the friendlier 50mm lens standard, which is a correct photographic ideal, but not how people are actually taking pictures of themselves these days.

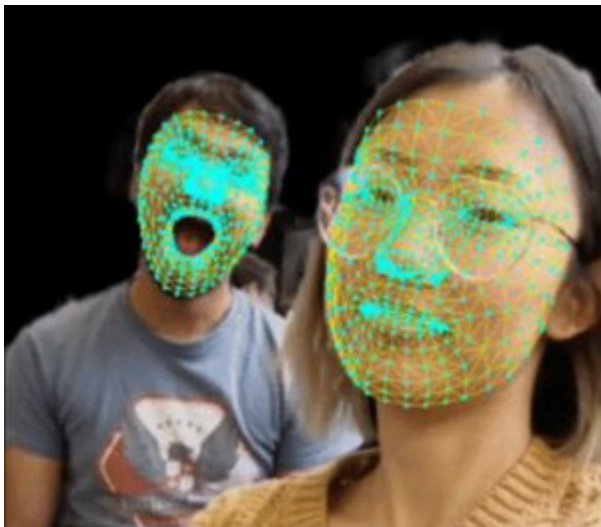


Prior methods weed out 'difficult' selfie-style, high-FOV images before training. Sources: <https://arxiv.org/pdf/2210.07301.pdf> and <https://arxiv.org/pdf/2205.15517.pdf>

DISCO addresses these previously-avoided challenges in three ways: by parametrizing the focal length (instead of excluding 'challenging' data from being trained); through *optimization scheduling*, which accounts for the shortfall in progress between the rapid development of the face image's latent code and the slower optimization of the camera lens parameters; and *landmark regularization*.

The latter is perhaps the most radical innovation: by default, a GAN uses a photometric [loss function](#) which is unaware of the problems of lens distortion, and is simply expecting a 'default' image, and which lets the image itself set the focal standard.

Therefore the researchers used Google Research's [MediaPipe framework](#) to calculate dense facial landmarks for the input faces.

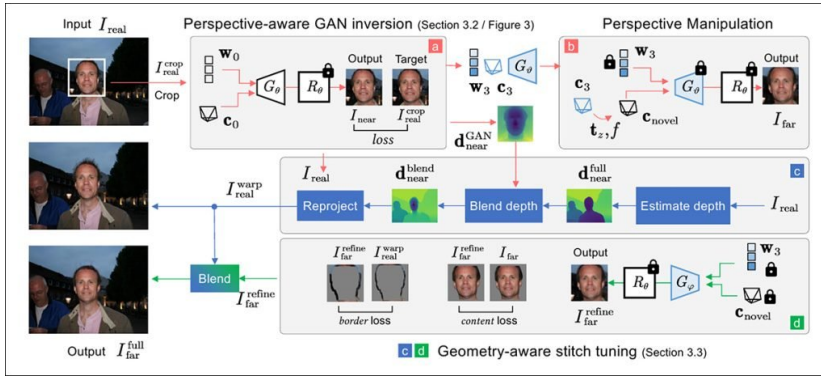


Dense facial landmarks captured by Google Research's MediaPipe project, used for DISCO. Source: <https://arxiv.org/pdf/1906.08172.pdf>

The way that the landmarks change relate to the focal length. For instance, in a very wide-angle picture of a person, the subject's eyes may appear notably larger, providing one possible 'anchor definition' for that focal length.

During inversion (the point at which a novel image is projected into the trained system, so that it can be manipulated), the uncertainty-based landmark loss built into the [LPIPS](#) loss metric is used for optimization. During [fine-tuning](#) of the generator, LPIPS is also used, together with [L1 loss](#).

Since 3D GANs only take cropped faces as input, the researchers had to devise a method to 're-stitch' altered images back into a more complete image. The algorithm therefore aligns and blends [MiDaS](#)-calculated depth for the face with the estimated depth for the 'fuller' image.



Workflow for the final composition in DISCO.

The composited image is then re-projected to the same camera parameters as the 3D GAN itself, and the generator module is fine-tuned to modify and re-blend all the related borders. The end result is a ‘virtual’ image apparently captured from a greater distance.

This, finally, provides a focal length mapping that permits the user to ‘scrub’ through diverse FOVs – an old-school optical versatility made famous by Steven Spielberg in one particularly effective ‘push/pull’ shot from *Jaws* (1975).



The famous ‘push/pull’ shot from *Jaws* (1975), in which a zoom lens is ‘racked’ from its lowest to its widest focal distance while the camera itself is physically dollied up towards the actor, keeping the head the same size while changing its apparent perspective. Source: <https://www.youtube.com/watch?v=GqymBzfuftc>

Training and Tests

The camera parameters for DISCO are estimated using a 2019 China/Microsoft [collaboration](#); a 2020 US/China/Facebook [project](#) was used to accomplish the necessary inpainting (also used as a competitor – see below); and in the case of damaged backgrounds, Stable Diffusion or DALL-E 2 are used to inpaint these.

For a testing round, the researchers used the [EG3D](#) dataset trained on [FFHQ](#). DISCO was tested using the [Caltech Multi-Distance Portraits](#) (CMDP, ‘28’ in results table below) Dataset; the [USC perspective portrait](#) database (‘94’ in results included below); and a collection of ‘in-the-wild’ images compiled by the researchers themselves.

The system was pitted against the implementation of the Caltech and USC collections, both [2D warping](#)-based methods.

Since the methodologies differ, and official implementations were not available across the board, the authors concocted ‘equivalent’ standards, and additionally tested against the inpainting method used in DISCO, an unnamed process featured in the paper *3D Photography using Context-aware Layered Depth Inpainting*, from Virginia Tech, National Tsing Hua University, and Facebook (indicated as ‘68’).

Tests were on the CMDP dataset, and here DISCO ‘performs well’ in landmark metrics (according to the authors), and is comparable to the Caltech approach. Besides LPIPS, quantitative metrics used are Peak signal-to-noise ratio ([PSNR](#)) and Structural Similarity Index ([SSIM](#)); the ‘LMK-E’ metric is unreferenced and unexplained in the paper.

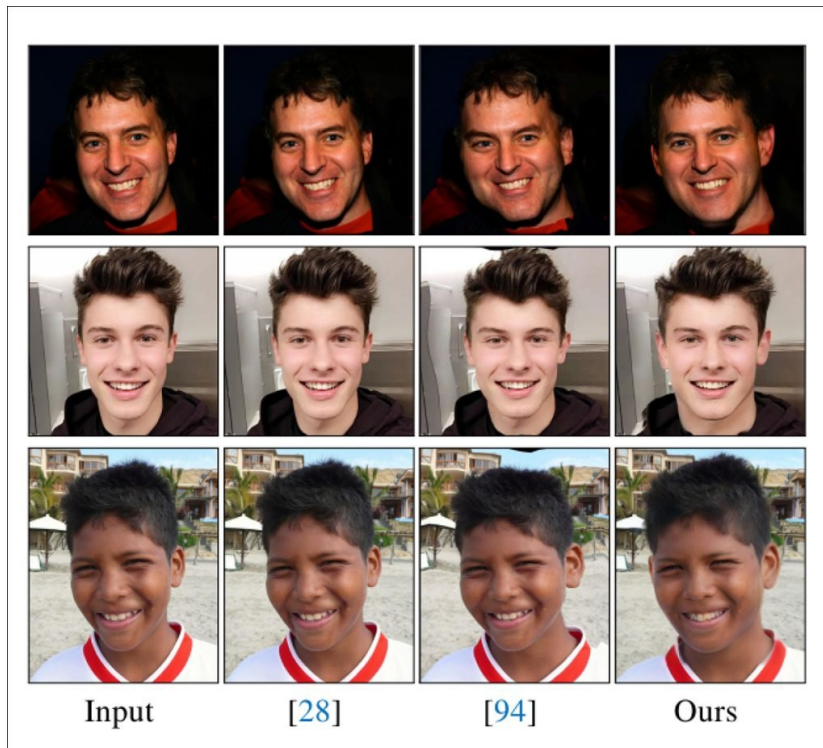
Method	LMK-E↓	PSNR↑	SSIM↑	LPIPS↓
Input	0.23	12.49	0.653	0.265
[28] ¹	0.18	14.81	0.695	0.201
[28] ²	0.16	13.30	0.684	0.225
Ours	0.14	14.26	0.694	0.209

Results from the quantitative round.

These results are from evaluations on 43 faces projected at various focal lengths, from 60cm to 480cm.

(It should be noted here that in the ordinary course of events, the researchers behind truly innovative systems tend to go to extraordinary lengths to provide like-for-like equivalency in testing rounds. Very often – as in this case – the lack of publicly available data and code from prior systems renders such tests exercises in ‘submission completism’ rather than constituting a valid and reproducible evaluation method. This seems to be the case for DISCO, where a particularly tortuous set of workarounds was needed in order to provide any former frameworks to test against. We can perhaps consider more the innovations present in the work rather than the quantitative results.)

Finally, for a round of qualitative evaluation, the authors present some direct comparisons:



Results from the qualitative testing round.

Of these, they state:

‘Note that with the help of the 3D GAN, our method can generate occluded parts in the original input images, such as ears. We further demonstrate this advantage and show the perspective distortion correction results at different [distances].

‘These visual results show that 3D GAN inversion is an effective way of portrait perspective correction compared to the flow-based warping methods.’

Conclusion

Understanding the extent to which faces are distorted by wide-angle lenses is a valuable pursuit, both in the development of more flexible facial ID systems, and in the evolution of generative systems, which currently have a more ‘generic’ or imaginative conception of the physics behind these distortions (usually based on labels, and/or on comparison with thousands or millions of other face images present in hyperscale datasets such as LAION).

Being able to quantify the extent to which a face is ‘under pressure’ from extreme FOVs could enable the rational development of flexible and accurate generative systems trained on much lower volumes of data, and which could provide the end-user with genuine instrumentality over FOV, much as a photographer can pick and choose lenses to suit their subject and objectives.

In practice, systems such as DISCO tend to obtain requisite funding through more immediately enticing capitalization prospects, such as ‘selfie correction’ apps and filters that can operate on edge devices (i.e., smart phones), and provide the user with a dumbed-down way of altering their own images.

However, the effort needed to arrive at such functionality may, as a collateral benefit, be immensely useful in the deeper stratas of the human image synthesis research sector.