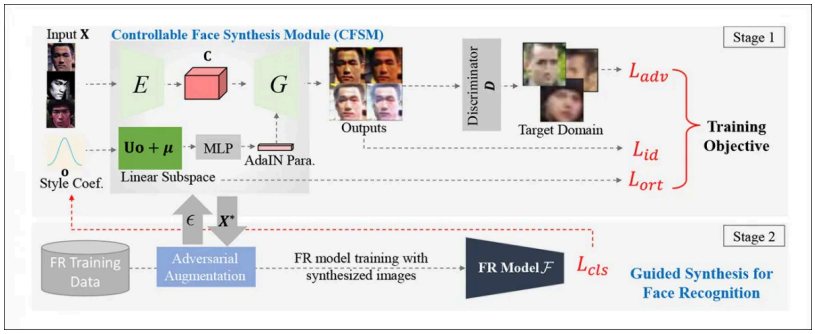
‘Degraded’ Synthetic Faces Could Help Improve Facial Image Recognition

Originally published at https://www.unite.ai/degraded-synthetic-faces-could-help-improve-facial-image-recognition/

Researchers from Michigan State University have devised a way for synthetic faces to take a break from the deepfakes scene and do some good in the world – by helping image recognition systems to become more accurate.

The new controllable face synthesis module (CFSM) they’ve devised is capable of regenerating faces in the style of real-world video surveillance footage, rather than relying on the uniformly higher-quality images used in popular open source datasets of celebrities, which do not reflect all the faults and shortcomings of genuine CCTV systems, such as facial blur, low resolution, and sensor noise – factors that can affect recognition accuracy.



Conceptual architecture for the Controllable Face Synthesis Module (CFSM). Source: http://cvlab.cse.msu.edu/pdfs/Liu_Kim_Jain_Liu_ECCV2022.pdf

CFSM is not intended specifically to authentically simulate head poses, expressions, or all the other usual traits that are the objective of deepfake systems, but rather to generate a range of alternative views in the style of the target recognition system, using style transfer.

The system is designed to mimic the style domain of the target system, and to adapt its output according to the resolution and range of ‘eccentricities’ therein. The use-case includes legacy systems that are not likely to be upgraded due to cost, but which can currently contribute little to the new generation of facial recognition technologies, due to poor quality of output that may once have been leading-edge.

Testing the system, the researchers found that it made notable gains on the state of the art in image recognition systems that have to deal with this kind of noisy and low-grade data.

Training the facial recognition models to adapt to the limitations of the target systems. Source: http://cvlab.cse.msu.edu/pdfs/Liu_Kim_Jain_Liu_ECCV2022_supp.pdf

They additionally found a useful by-product of the process – that the target datasets could now be characterized and compared to each other, making the comparing, benchmarking and generation of bespoke datasets for varying CCTV systems easier in the future.

Further, the method can be applied to existing datasets, performing *de facto* domain adaptation and making them more suitable for facial recognition systems.

The new paper is titled *Controllable and Guided Face Synthesis for Unconstrained Face Recognition*, is supported in part by the US Office of the Director of National Intelligence (ODNI, at IARPA), and comes from four researchers at the Computer Science & Engineering department at MSU.

Featured Content

Low-quality face recognition (LQFR) has become a notable area of study over the past few years. Because civic and municipal authorities built video surveillance systems to be resilient and long-lasting (not wanting to re-allocate resources to the problem periodically), many 'legacy' surveillance networks have become victims of technical debt, in terms of their adaptability as data sources for machine learning.

Varying levels of facial resolution across a range of historic and more recent video surveillance systems. Source: https://arxiv.org/pdf/1805.11519.pdf

Luckily, this is a task that diffusion models and other noise-based models are unusually well-adapted to solve. Many of the most popular and effective image synthesis systems of recent years perform upscaling of low-resolution images as part of their pipeline, while this is also absolutely essential to neural compression techniques (methods to save images and movies as neural data instead of bitmap data).

Part of the challenge of facial recognition is to obtain the maximum possible accuracy from the minimum number of *features* that can be extracted from the smallest and least promising low-resolution images. This constraint exists not only because it's useful to be able to identify (or create) a face at low resolution, but also because of technical limitations on the size of images that can pass through the emerging latent space of a model that's being trained in whatever VRAM is available on a local GPU.

In this sense, the term 'features' is confusing, since such features can also be obtained from a dataset of park benches. In the computer vision sector, 'features' refers to the distinguishing characteristics obtained from images – *any* images, whether it's the lineaments of a church, a mountain, or the disposition of *facial* features in a face dataset.

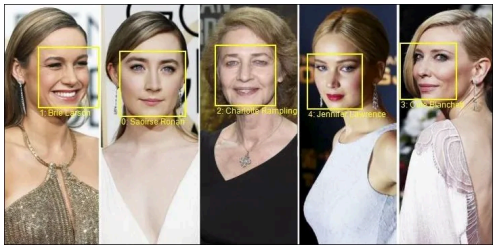
Since computer vision algorithms are now adept at upscaling images and video footage, various methods have been proposed to 'enhance' low-resolution or otherwise degraded legacy surveillance material, to the point that it might be possible to use such augmentations for legal purposes, such as placing a particular person at a scene, in relation to a crime investigation.

Besides the possibility of misidentification, which has occasionally gathered headlines, in theory it should not be necessary to hyper-resolve or otherwise transform low-resolution footage in order to make a positive identification of an individual, since a facial recognition system keying in on low-level features should not need that level of resolution and clarity. Further, such transformations are expensive in practice, and raise additional, recurrent questions around their potential validity and legality.

The Need for More 'Down-At-Heel' Celebrities

It would be more useful if a facial recognition system could derive features (i.e. machine learning features of *human* features) from the output of legacy systems as they stand, by understanding better the relationship between 'high resolution' identity and the degraded images that are available in implacable (and often irreplaceable) existing video surveillance frameworks.

The problem here is one of standards: common web-gathered datasets such as MS-Celeb-1M and WebFace260M (among several others), have been latched onto by the research community because they provide consistent benchmarks against which researchers can measure their incremental or major progress against the current state of the art.



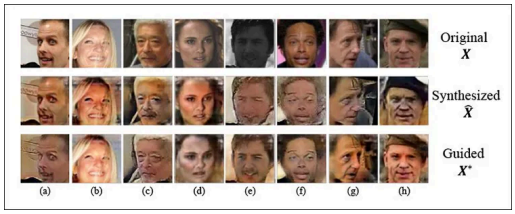
Examples from Microsoft's popular MS-Celeb1m dataset. Source: https://www.microsoft.com/en-us/research/project/ms-celeb-1m-challenge-recognizing-one-million-celebrities-real-world/

However, the authors argue that facial recognition (FR) algorithms trained on these datasets are unsuitable material for the visual 'domains' of the output from many older surveillance systems.

The paper states*:

> *'[State-of-the-art] (SoTA) FR models do not work well on real-world surveillance imagery (unconstrained) due to the domain shift issue, that is, the large-scale training datasets (semi-constrained) obtained via web-crawled celebrity faces lack in-the-wild variations, such as inherent sensor noise, low resolution, motion blur, turbulence effect, etc.*
>
> *'For instance, 1:1 verification accuracy reported by one of the SoTA models on unconstrained IJB-S dataset is about 30% lower than on semi-constrained LFW.*
>
> *'A potential remedy to such a performance gap is to assemble a large-scale unconstrained face dataset. However, constructing such a training dataset with tens of thousands of subjects is prohibitively difficult with high manual labeling cost.'*

The paper recounts various prior methods that have attempted to 'match' the variegated types of outputs from historical or low-cost surveillance systems, but note that these have dealt with 'blind' augmentations. By contrast, CFSM receives direct feedback from the real-world output of the target system during training, and adapts itself via style transfer to mimic that domain.



Actress Natalie Portman, no stranger to the handful of datasets that dominate the computer vision community, features among the identities in this example of CFSM performing style-matched domain adaptation based on feedback from the domain of the actual target model.

The architecture designed by the authors utilizes Fast Gradient Sign Method (FGSM) to individuate and 'import' the obtained styles and characteristics from true output of the target system. The part of the pipeline devoted to image generation will subsequently improve and become more faithful to the target system with training. This feedback from the low dimensional style space of the target system is low-level in nature, and corresponds to the broadest derived visual descriptors.

The authors comment:

> *'With the feedback from the FR model, the synthesized images are more beneficial to the FR performance, leading to significantly improved generalization capabilities of the FR models trained with them.'*

Tests

The researchers used MSU's own prior work as a template for testing their system. Based on the same experimental protocols, they used MS-Celeb-1m, which consists exclusively of web-trawled celebrity photographs, as the labeled training dataset. For fairness, they also included MS1M-V2, which contains 3.9 million images featuring 85,700 classes.

The target data was the WiderFace dataset, from the Chinese University of Hong Kong. This is a particularly diverse set of images designed for face detection tasks in challenging situations. 70,000 images from this set were used.

For evaluation, the system was tested against four face recognition benchmarks:: IJB-B, IJB-C, IJB-S, and TinyFace.

CFSM was trained with ~10% of training data from MS-Celeb-1m, around 0.4 million images, for 125,000 iterations at 32 batch size under the Adam optimizer at a (very low) learning rate of 1e-4.

The target facial recognition model used a modification of ResNet-50 for the backbone, with ArcFace loss function enabled during training. Additionally, a model was trained with CFSM as an ablation and comparative exercise (noted as 'ArcFace' in the results table below).

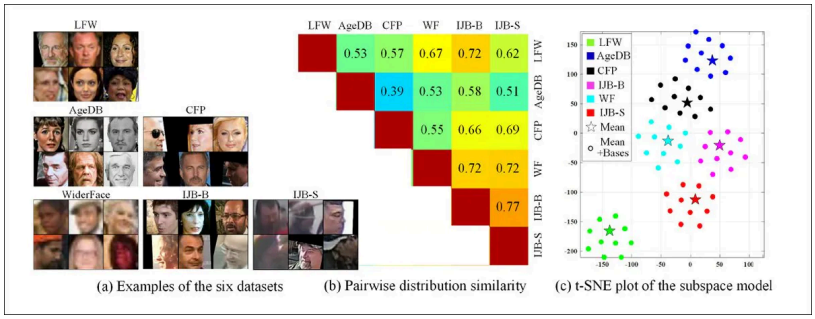
Method	Labeled Train Data	Backbone	IJB-S V2S				IJB-S V2B				IJB-S V2V				TinyFace	
			Rank1	Rank5	1%	10%	Rank1	Rank5	1%	10%	Rank1	Rank5	1%	10%	Rank1	Rank5
C-FAN [20]	MS1M-*	ResNet-64	50.82	61.16	16.44	24.19	53.04	62.67	27.40	29.70	10.05	17.55	0.11	0.68	–	–
MAIN [21]	MS1M-*	ResNet-64	58.14	64.11	21.47	–	59.26	65.93	32.07	–	22.25	34.16	0.19	–	–	–
PFE [68]	MS1M-*	ResNet-64	50.16	58.33	31.88	35.33	53.60	61.75	35.99	39.82	9.20	20.82	0.84	2.83	–	–
ArcFace [12]	MS1MV2	ResNet-50	50.39	60.42	32.39	42.99	52.25	61.19	34.87	43.50	–	–	–	–	–	–
Shi *et al.* [60]	MS1MV2-*	ResNet-50	59.29	66.91	39.92	50.49	60.58	67.70	32.39	44.32	17.35	28.34	1.16	5.37	–	–
ArcFace [12]	MS1MV2-*	ResNet-50	58.78	66.40	40.99	50.45	60.66	67.43	43.12	51.38	14.81	26.72	2.51	5.72	62.21	66.85
ArcFace+Ours*	MS1MV2-*	ResNet-50	61.69	68.33	43.99	53.34	62.20	69.50	44.38	53.49	18.14	31.34	2.09	4.51	62.39	67.36
ArcFace+Ours	MS1MV2-*	ResNet-50	**63.86**	**69.95**	**47.86**	**56.44**	**65.95**	**71.16**	**47.28**	**57.24**	**21.38**	**35.11**	**2.96**	**7.41**	**63.01**	**68.21**
AdaFace [43]	WebFace12M	IResNet-100	71.35	76.24	59.40	**66.34**	71.93	76.56	59.37	**66.68**	36.71	50.03	4.62	11.84	72.29	74.97
AdaFace+Ours	WebFace12M	IResNet-100	**72.54**	**77.59**	**60.94**	66.02	**72.65**	**78.18**	**60.26**	65.88	**39.14**	**50.91**	**5.05**	**13.17**	**73.87**	**76.77**

Results from the primary tests for CFSM. Higher numbers are better.

The authors comment on the primary results:

> *'ArcFace model outperforms all the baselines in both face identification and verification tasks, and achieves a new SoTA performance.'*

The ability to extract domains from the various characteristics of legacy or under-specced surveillance systems also enables the authors to compare and evaluate the distribution similarity among these frameworks, and to present each system in terms of a visual style that could be leveraged in subsequent work.



(a) Examples of the six datasets (b) Pairwise distribution similarity (c) t-SNE plot of the subspace model

Examples from various datasets exhibit clear differences in style.

The authors note additionally that their system could make worthwhile use of some technologies that have, to date, been viewed solely as problems to be resolved by the research and vision community:

> *'[CFSM] shows that adversarial manipulation could go beyond being an attacker, and serve to increase recognition accuracies in vision tasks. Meanwhile, we define a dataset similarity metric based on the learned style bases, which capture the style differences in a label or predictor agnostic way.*
>
> *'We believe that our research has presented the power of a controllable and guided face synthesis model for unconstrained FR and provides an understanding of dataset differences.'*

** My conversion of the authors' inline citations to hyperlinks.*

First published 1st August 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More