LEGAN: Disentangled Manipulation of Directional Lighting and Facial Expressions whilst Leveraging Human Perceptual Judgements

Paper # 21

Sandipan Banerjee, Ajjen Joshi, Prashant Mahajan*, Sneha Bhattacharya*, Survi Kyal, Taniya Mishra*

:) Affectiva

Affectiva Inc., Boston, USA

*Work done while at Affectiva

OVERVIEW

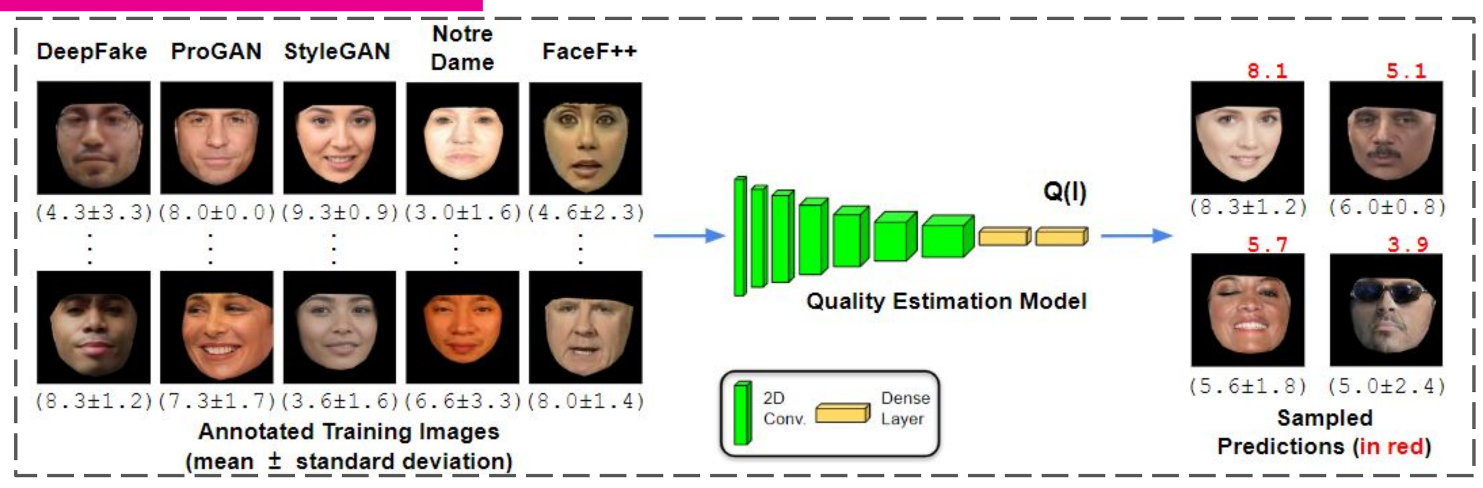
- Existing naturalness metrics either generate a single score for the **whole dataset** (FID [1]) or compute dissimilarity among **image pairs** (LPIPS [2]).
- Our quality metric rates the **naturalness of individual synthetic face images** in vacuum, serving as an auto substitute for human judgement.
- We directly plug this metric into **LEGAN**, our framework for disentangled lighting and expression manipulation, as an **auxiliary discriminator**.
- Using a set of hourglass nets, **LEGAN separates the attribute sub-spaces** & performs the desired translation while preserving identity.

CONTRIBUTIONS

- We build a **quality estimation model (Q)** to directly evaluate the perceived quality of GAN-generated images, and **release the dataset** of synthetic images along with their crowd-sourced quality annotations.
- When used in training, **Q improves the perceptual quality** of images synthesized by not only **LEGAN** but other **off-the-shelf GANs** as well.
- **Q** can also be used to **filter face images** synthesized by different models.
- **LEGAN** can be utilized as **data augmenter** to improve model performance on downstream tasks like **face verification** and **expression recognition**.

PERCEPTUAL QUALITY ESTIMATION

- **Dataset (URL)**: we collected face images generated using five different GAN & 3D model based synthesis approaches. After pre-processing, we ended up with **37K synth. images**.
- **Perceptual annotation**: each image was scored for **naturalness** by 3 **human raters**. We used the **mean** (m) & **standard deviation** (std) from these ratings as the perceptual label.

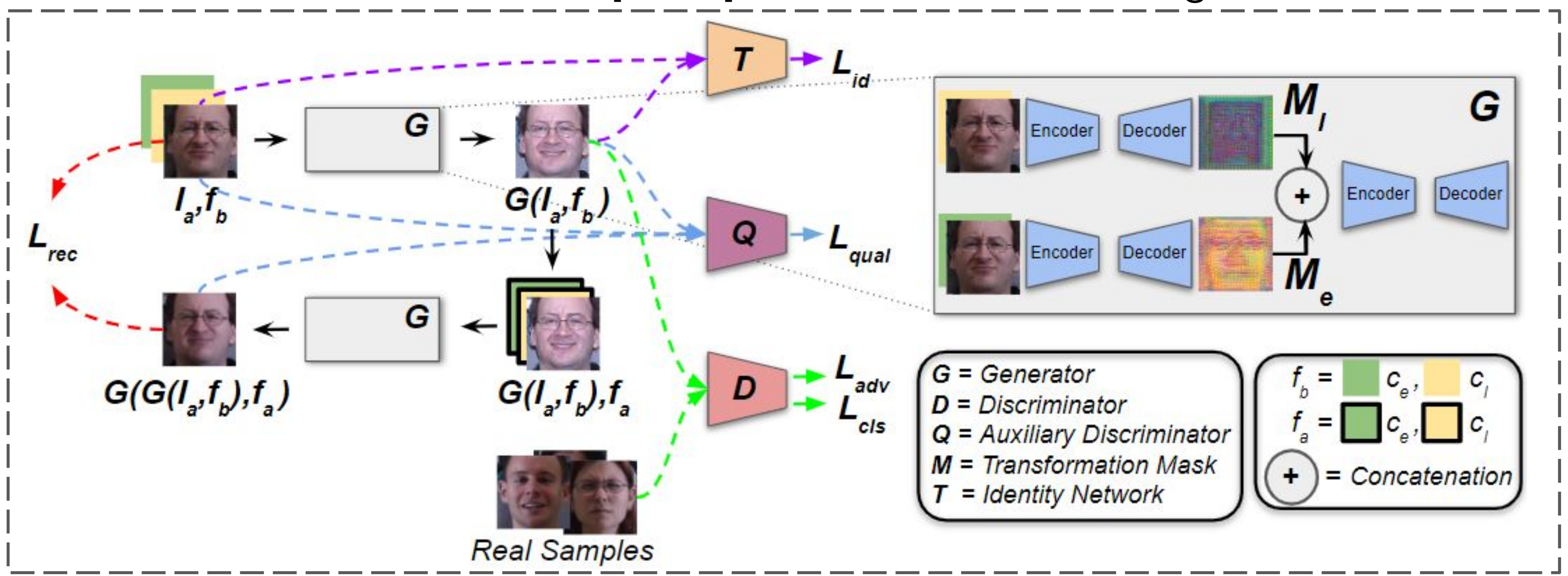


- **Quality estimation model (Q)**: as a **cheap proxy for human annotation**, we train a CNN with the images & their (m, std) labels. To capture the subjectiveness in visual perception, we formulated a **margin based loss function** for training.

LEGAN: UTILIZING Q FOR LIGHTING & EXPRESSION MANIPULATION

- **Architecture**: LEGAN is composed of generator (G) and discriminator (D) networks, while Q serves as an auxiliary module for estimating quality of the synthesized images during training. Similar to other image-to-image translation models, LEGAN **does not require paired data** for training.

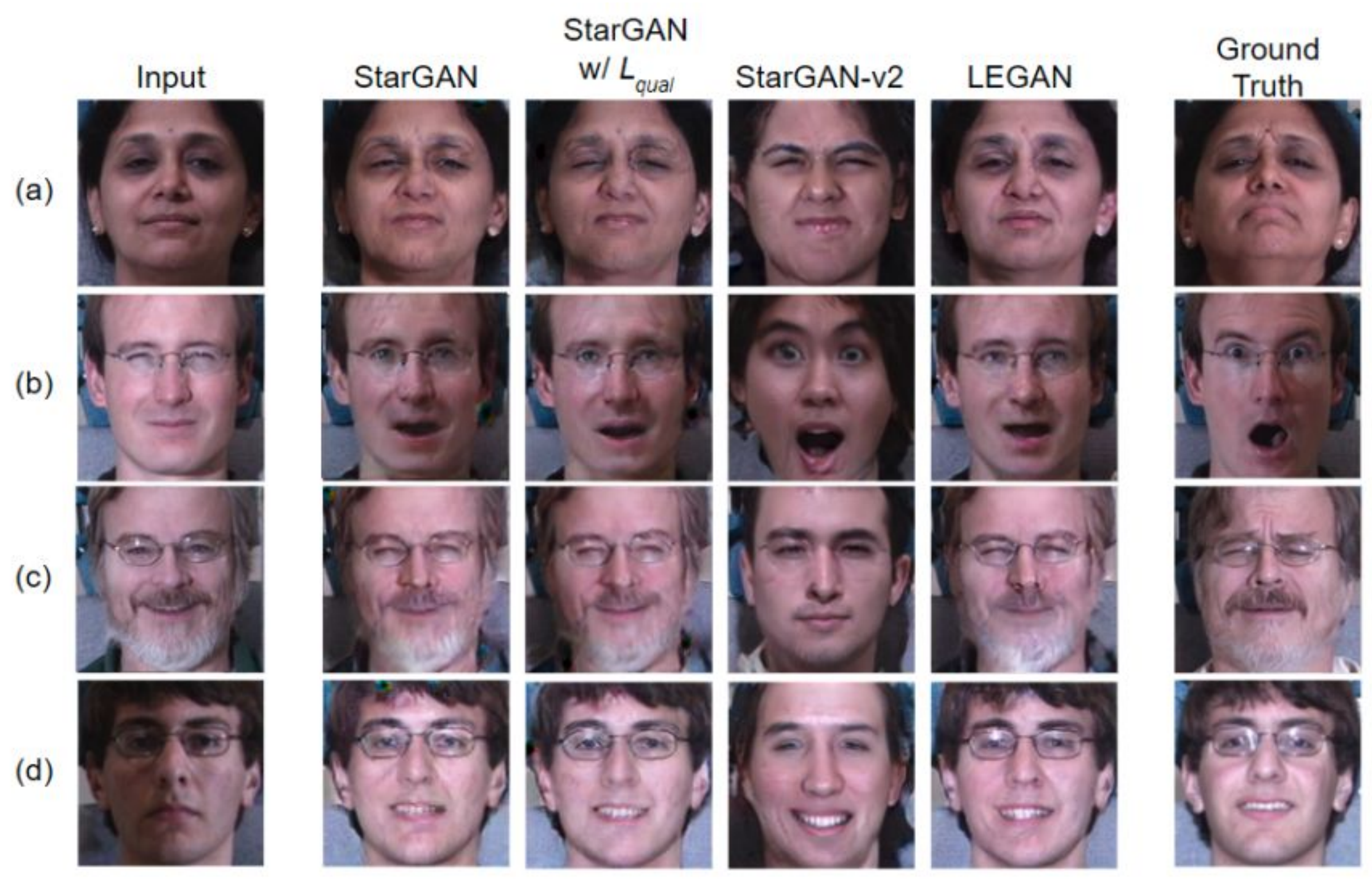
(1) **G**: takes an input image with target attributes and generates **disentangled transformation masks** in lighting and expression sub-spaces using a pair of hourglass nets. A third hourglass generates the final output from these masks.
(2) **D**: takes the synthetic sample and generates predictions based on its **realness** and **target attribute(s) association** (unpaired data formulation).
(3) **Q**: is pre-trained on perceptual data. Kept **frozen** during LEGAN training.



- **Loss function**: The full loss is a weighted sum of following:

(1) $\mathbf{L_{adv}}$: **D**'s weights are leveraged to tune **G**'s hallucinations to match distribution of real data and **produce realistic samples** as training progresses.
(2) $\mathbf{L_{cls}}$: ensures the **target class association** of a synthetic vector is preserved in the attribute space, using cross entropy over **D**'s softmax prediction.
(3) $\mathbf{L_{rec}}$: maintains **structural integrity** by cyclically reconstructing the input image from the translated output, comparing the two in pixel space.
(4) $\mathbf{L_{id}}$: preserves **subject identity** by minimizing the distance between representations of the input & output images in the **LightCNN-29** [3] feature space.
(5) $\mathbf{L_{qual}}$: optimizes the perceptual quality of the translated output in the forward phase while preserving the same for the reconstructed input in the cyclic phase using Q's prediction.

EXPERIMENTAL RESULTS



Models	FID [45] ↓	LPIPS [103] ↓	SSIM [94] ↑	Match Score [44, 22] ↑	Quality Score ↑	Human Preference ↑
StarGAN [25]	38.745	0.126	0.559	0.635	5.200	22.3%
StarGAN w/ L_{qual}	34.045	0.123	0.567	0.647	5.391	34.7%
StarGAN-v2 [26]	54.842	0.212	0.415	0.202	5.172	3.75%
LEGAN	**29.964**	**0.120**	**0.649**	**0.649**	**5.853**	**39.3%**
Real Images	12.931	-	-	0.739	5.921	-

- **Training data**: We use frontal face images from the MultiPIE dataset.
- **Improving perceptual quality**: Adding Q to the training framework improves visual quality and removes blob-like artifacts [4] from synthesized images (StarGAN: d).
- **Improving off-the-shelf StarGAN**: When added to the training framework of StarGAN [5], Q improves its performance on almost all metrics (compare rows 1 & 2).
- **Correlation with existing metrics & human judgement**: As can be seen in columns (2, 3, 6) & (6, 7) in the table above, our quality metric is well correlated with FID and LPIPS, and naturalness ratings provided by human annotators.

- **Improving face verification**: When training data (CASIA-WebFace) is augmented with LEGAN's synthetic images, model performance improves on IJB-B and LFW datasets.

Training Data	Real Images [100] (# Identities)	Synthetic Images (# Identities)	IJB-B [96] Performance	LFW [46] Performance
Original	439,999 (10,575)	0	0.954 ± 0.002	0.966 ± 0.002
Augmented	439,999 (10,575)	439,999 (10,575)	**0.967 ± 0.001**	**0.972 ± 0.001**

- **Improving emotion recognition**: adding synth. images with targeted emotions alleviates class imbalance also improves model performance on the AffectNet dataset.

Training Data	Real Images [66]	Synthetic Images	'Neutral'	'Happy'	'Surprise'	'Disgust'
Original	204,325	0	0.851 ± 0.005	0.955 ± 0.001	0.873 ± 0.004	0.887 ± 0.005
Augmented	204,325	279,324	**0.868 ± 0.005**	**0.956 ± 0.001**	**0.890 ± 0.003**	**0.897 ± 0.001**

[1] M. Heusel, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium", in NeurIPS, 2017.
[2] R. Zhang, et al. "The unreasonable effectiveness of deep features as a perceptual metric", in CVPR, 2018.
[3] X. Wu, et al. "A light cnn for deep face representation with noisy labels", in IEEE Trans. on Information Forensics and Security (TIFS), 2018.
[4] T. Karras, et al. "Analyzing and Improving the Image Quality of StyleGAN", in CVPR, 2020.
[5] Y. Choi, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation", in CVPR, 2018.