Facial Image Feature Analysis and its Specialization for Fréchet Distance and Neighborhoods

Doruk Cetin[1], Benedikt Schesch[2], Petar Stamenkovic[2], Majed El Helou[2]
[1]Align Technology Zürich, Switzerland, [2]Media Technology Center, ETH Zürich, Switzerland

Motivation

- Measuring distances between datasets is a valuable yet challenging task
- FID remains the most practical and ubiquitous metric, despite its numerous shortcomings
- Kynkäänniemi et al. criticize the strong relation between Inception features and ImageNet classes
- Morozov et al. explore replacing supervised ImageNet feature extractors with self-supervised ones
- **We make the last leap:** first analysis on domain-specific feature training and its effects on feature distance – on the widely-researched facial image domain

Methodology

Feature-learning independent dataset

- 30k samples, same size as CelebA-HQ
- No occlusions, manually curated
- Balanced across six ethnicities (latino hispanic, asian, black, middle eastern, indian, white)

Self-supervised feature learning

- DINO for self-supervised learning, state-of-the-art vision transformer model
- Feature embedding of 2048 dimensions, same size as Inception architecture

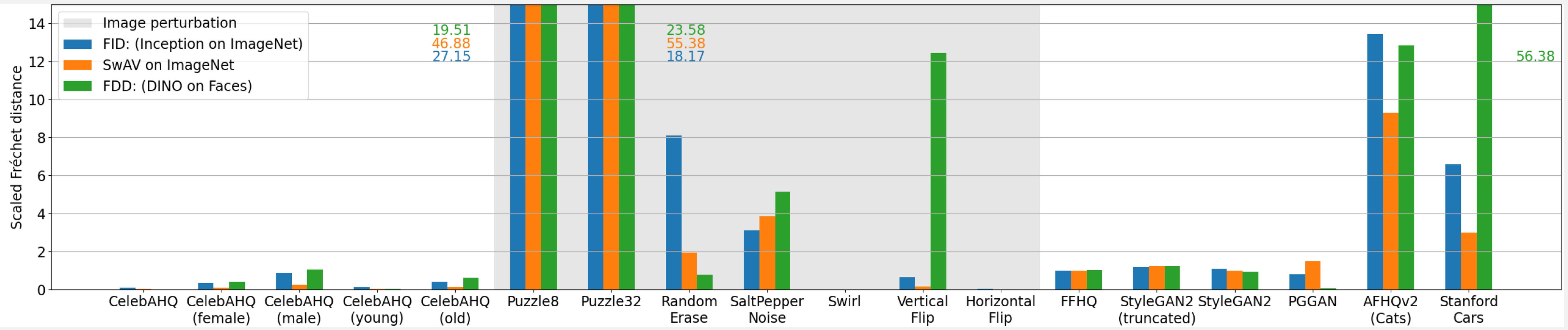
Experiments

Tab. Classification accuracies for binary CelebA-HQ annotations

Method /vs./ Test	Blond	Young	Gender	Gender
Inception + Head	93.54	**85.58**	**96.44**	84.92
Inception + MLP	92.83	83.90	96.25	84.22
DINO (I) + Head	90.63	83.08	94.33	**86.40**
DINO (I) + MLP	91.37	83.25	94.96	85.71
DINO (F) + Head	**93.85**	82.54	92.56	85.86
DINO (F) + MLP	93.92	83.06	93.02	86.00

Results with self-supervised DINO are on-par with Inception: our self-learned features are sufficient

Fig. Rescaled Fréchet distances on Inception (trained on ImageNet), SwAV (trained on ImageNet), and DINO features (trained on Faces)



Distance on DINO features …

- is large when images are flipped vertically → more sensitivity to global changes
- is smallest for random erasing of small patches → specialized to faces, high-level features
- grow larger moving from faces to cats to cars → more sensitivity to out-of-domain data
- is similar to other approaches for remaining setups, on average

Tab. User study results on distribution matching (1-5 score)

Image source distribution	μ	σ	FID	FDD
CelebA-HQ (class: male)	2.00	1.09	0.87	1.06
CelebA-HQ (class: female)	2.52	1.15	0.34	0.40
CelebA-HQ (class: young)	2.43	1.20	0.12	0.06
CelebA-HQ (class: old)	2.28	1.16	0.43	0.63
StyleGAN2 (untruncated)	1.92	1.00	1.09	0.94
StyleGAN2 (0.7 truncated)	2.16	1.10	1.20	1.23
r-correlation to survey μ	1.00	-	-0.83	-0.79
ρ-correlation to survey μ	1.00	-	-0.77	-0.71

FID and FDD both strongly correlated with the participants' answers

Tab. User study results on photorealism (1-5 score)

Image source distribution	μ	σ	FID	FDD
FFHQ dataset samples	4.12	1.10	0.99	1.02
StyleGAN2 (0.7 truncated)	4.03	1.13	1.20	1.23
StyleGAN2 (untruncated)	3.19	1.44	1.09	0.94
PGGAN* dataset samples	1.93	1.11	0.83	0.09

Distances highly diverge on PGGAN
→ participant opinions are strongly affected by visual artifacts, while distance metrics focus on content distributions

Fig. Samples from our user study on feature space neighborhoods

(d) Reference (e) Nearest neighbors in the Inception space (FID) (f) Nearest neighbors in the DINO space (FDD)

Tab. User study results on similarity (% avg. votes)

	Subset	Inception	DINO	σ
Image Sim	CelebA-HQ (accessories)	59	41	20
	CelebA-HQ (random)	72	28	14
	AFHQv2-Cats [34]	69	31	29
	Stanford Cars [35]	92	8	4
P.	CelebA-HQ (accessories)	42	58	24

- Inception is excessively biased towards focusing on objects rather than faces
- Lack of such bias for DINO did not guarantee the desired face similarity

Conclusion

1. **Specialists become better at abstraction.** Generalists focus more on fine-granularity features.
2. **Feature distance does not equate to photorealism.** Quality and distribution of the base dataset matters.
3. **Noticing can be easier than not noticing.** Novel content in input can act as adversarial attacks.
4. **The risk of smaller specialized datasets.** Multiple paths lead to the final representation and training over a large dataset constrains the behavior of the feature extractor across its many paths.

arXiv

References

- Heusel et al. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium" NIPS 2017
- Kynkäänniemi et al. "The Role of ImageNet Classes in Fréchet Inception Distance" ICLR 2023
- Morozov et al. "On self-supervised image representations for GAN evaluation" ICLR 2021
- Caron et al. "Emerging properties in self-supervised vision transformers" ICCV 2021
- Frankle et al. "The lottery ticket hypothesis: Finding sparse, trainable neural networks" ICLR 2019
- Szegedy et al. "Rethinking the Inception architecture for computer vision" CVPR 2016
- Lee et al. "MaskGAN: Towards diverse and interactive facial image manipulation" CVPR 2020
- Karras et al. "Analyzing and improving the image quality of StyleGAN" CVPR 2020
- Karras et al. "Progressive growing of GANs for improved quality, stability, and variation" ICLR 2018