https://doi.org/10.2352/EI.2025.37.10.IPAS-237

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*

Abstract

Assessing distances between images and image datasets is a fundamental task in vision-based research. It is a challenging open problem in the literature and despite the criticism it receives, the most ubiquitous method remains the Fréchet Inception Distance. The Inception network is trained on a specific labeled dataset, ImageNet, which has caused the core of its criticism in the most recent research. Improvements were shown by moving to self-supervision learning over ImageNet, leaving the training data domain as an open question. We make that last leap and provide the first analysis on domain-specific feature training and its effects on feature distance, on the widely-researched facial image domain. We provide our findings and insights on this domain specialization for Fréchet distance and image neighborhoods, supported by extensive experiments and in-depth user studies.

Introduction and Related Work

Measuring distances between datasets is a valuable yet challenging task, in particular for complex signals such as images. It is crucial for understanding data distributions and domain gaps for transfer learning and generalization. It is also important for developing generative networks that recently gained in popularity [1] and that are prone to hallucination [2].

The most ubiquitous approach to measuring dataset distance is the widely used Fréchet inception distance (FID) [3]. It computes the Fréchet statistical distance [4] between the datasets' image features, extracted by an ImageNet-trained [5] Inception network [6]. A plethora of similar solutions emerged in the literature, notably extensions to conditional inputs [7] and adversarial robustness [8]. Binkowski *et al.* [9] propose Kernel Inception Distance (KID), a modified distance on the same feature space. It has certain theoretical advantages but in practice correlates closely with FID. Centered Kernel Alignment (CKA) [10] is another FID alternative, but both the paper's results and user survey show it performs in a similar way as FID. sFID [11] simply mimics FID on intermediate feature maps to improve spatial information that is more fine-grained. StyleGAN-XL [12] even computes rFID (random-FID) on the features of a randomly initialized network as an additional metric, with the idea originating from Naeem *et al.* [13], where the objective is to be more general by being task-agnostic. rFID results, however, tend to have an erratic behavior in practice with extremely large values. Zhang *et al.* [14] have shown that training is crucial as random networks achieve significantly worse performance, and that random networks focus more on low-level information [15], further supporting the use of trained features for Fréchet distance.

More specialized methods have been proposed that are different from FID. Kynkäänniemi *et al.* [16] investigated precision and recall between feature spaces as complementary metrics to FID. The pair can illustrate certain trade-offs but do not give a single score that can be used as an optimization target. We note that the experimental evaluation on faces only uses the standard ImageNet FID [16]. Precision and recall definitions are refined to density and coverage in [13], to better adapt to image manifolds that need to be estimated through only a limited number of sample points. Another approach to visualize feature space drift is based on SVCCA [17], however, it cannot readily scale to large datasets. Lastly and most similar to FID, Ramtoula *et al.* [18] create a histogram per neuron that contains the activation values of that neuron across network layers. The histogram of an image can be compared to the average histogram of a dataset to obtain a similarity metric. While this approach has the advantage of applicability to a single image, high-level information in cross-neuron dependencies as well as location information are lost.

FID thus remains the most practical and ubiquitous metric in recent literature [19, 15], despite its numerous shortcomings. The most simple to resolve are that it can be affected by resizing when anti-aliasing is omitted [19] and that its estimator has statistical bias that is model dependent [20]. FID relies on an ImageNet-trained Inception network that can be more sensitive to texture than to shape [21]. This bias is due to aggressive random cropping in data augmentation and can be reduced by using more natural augmentations like image distortions [22]. However, the underlying ImageNet training causes inherent limitations [23], such as a bias towards only the most salient object in a multi-object image [24], because ImageNet is meant for single object learning. Most recently, [25] criticizes the strong relation between Inception features and ImageNet classes, particularly as ImageNet does *not* contain human or human face classes while FID is most commonly used in studying generative models for face synthesis.

The goal of Morozov *et al.* [26] is to explore replacing supervised ImageNet feature extractors with self-supervised ones. The results show certain improvements in FID. The investigation supports the use of self-supervision and concludes with the open question of using self-supervised features that are *domain specific*, which was left to future research [26].

Our goal is to analyze how specializing the feature space impacts feature distances. We collect a novel facial dataset for our self-supervised learning to guarantee the independence from public datasets, and high image quality. We conduct extensive experiments and three user studies with 342 responses from 26 participants.

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

Table 1. Classification accuracy (%) of Head networks [27] and MLPs trained on CelebA-HQ features and the corresponding classes. CelebA-HQ features are extracted by: Inception (trained on ImageNet), DINO trained on ImageNet (I) and on Faces (F). We test on 3 CelebA-HQ classes and a Faces class (gender) that we collected to have a fully separate test set.

Methodology

Feature-learning independent dataset

To train our feature extractor, it is important to rely on a completely external dataset. The reason is that common facial image datasets are often used in training image generators, and any distance metric should be disentangled from them. We thus collect an in-house facial image dataset to train our feature extractor through self-supervision. We create a 30,000 image training set, in accordance with the size of CelebA-HQ [28], which we call Faces. The images are all center-cropped, with no occlusions, and manually curated to ensure quality. By training a feature extractor on our held-out dataset, we lay the basis for an independent metric built over those features. This enables benchmarking on the commonly used public datasets, and on public image generators trained on them. To promote better fairness, our dataset is balanced across six ethnicities (latino hispanic, asian, middle eastern, black, indian, and white) [29]. We further leave out an additional 21,000 images that we label for gender and use as a test set in a separate experiment to evaluate the extracted features.

Self-supervised feature learning

Self-supervised learning can improve feature extraction performance [26]. Another advantage is to reduce biases and errors coming from the choice and assignment of labels for supervised learning [25]. We exploit the simple yet effective state-of-the-art DINO [27] method for self-supervised learning on our dataset. DINO builds on knowledge distillation between teacher and student networks, and fundamental self-supervised learning data augmentation strategies, notably extending on SwAV [30]. For all of our experiments, we configure the feature embedding to have 2048 dimensions, aligning with the Inception [6] architecture for direct comparisons. We train for 100 epochs on one 24GB NVIDIA RTX 3090 GPU with a batch size of 16, and all other settings follow DINO's approach.

Fréchet distance over feature spaces

The Fréchet [4] distance F between two Gaussian distributions $\mathscr{N}(\mu_1,\Sigma_1)$ and $\mathscr{N}(\mu_2,\Sigma_2)$ is given by

$$F(\mu_1,\Sigma_1,\mu_2,\Sigma_2) = ||\mu_1-\mu_2||_2^2 + Tr(\Sigma_1+\Sigma_2-2(\Sigma_1\Sigma_2)^{\frac{1}{2}}),$$

where $Tr(\cdot)$ is the matrix trace. This formulation is then adapted to measure the distance between two datasets $\mathscr{D}_1$ and $\mathscr{D}_2$. This is achieved by exploiting the Inception [6] network's feature extractor trained on ImageNet in a supervised manner, and called FID [3]. The feature extractor takes an image as input and generates its embedding in a feature space. Generally, for any feature extractor $f(\cdot)$ we can define the Fréchet distance between datasets as $F(\mu^f_{\mathscr{D}_1},\Sigma^f_{\mathscr{D}_1},\mu^f_{\mathscr{D}_2},\Sigma^f_{\mathscr{D}_2})$, where $\mu^f_{\mathscr{D}_i}$ and $\Sigma^f_{\mathscr{D}_i}$ are the mean and covariance of the best-fit Gaussian over the feature distribution of dataset i, obtained by the feature extractor $f(\cdot)$. In our experiments, we study the effects of $f(\cdot)$ on the distance metric, with a focus on domain-specific specialized features. We denote the Fréchet distance computed over our DINO Faces feature space by FDD.

Experimental Evaluation

Datasets

In addition to CelebAHQ and FFHQ (sizes 30,000 and 70,000, respectively), we curate a separate face dataset of 30,000 samples, as explained earlier in the paper. We generate 10,000 samples using a PGGAN trained on CelebAHQ, and 10,000 samples using a StyleGAN2 trained on FFHQ two times: without truncation and with a truncation value of 0.7. For StyleGAN2 we fix the random seed, thus samples in both datasets match each other in terms of the attributes of synthetic people. Lastly, we utilize two datasets that do not contain images of human faces. We use all 5'558 cat images from AFHQv2-Cats, which has a preprocessing similar to the face datasets. We also use the 8'042 images from the test set of Stanford Cars. It is the only dataset where the images do not have equal width and height of 1024 pixels, so we resize all images to squares as preprocessing.

Are our self-learned features sufficient?

We evaluate whether our self-supervised features extract sufficient information relevant to faces. We train MLPs and Head networks on top of ImageNet-trained *Inception* features (used by FID), *DINO (I)* features from the ImageNet-trained DINO, and *DINO (F)* features from the Faces-trained DINO. The MLPs and Heads are trained on the CelebA-HQ training set annotations to predict different binary classes (Blond, Young, Gender) based on input features. We show the results in Table 1 on the CelebA-HQ test set and on an additional test set for gender from a fully independent source (see discussion above on datasets). We note that accuracies are significantly high, indicating that the networks extract sufficient features to enable classification. The results with the self-supervised DINO are on-par with Inception results, even surpassing them when testing on the independent curated test set, rather than the test set of CelebA-HQ. We emphasize, however, that the results highlight that the features are *sufficient* but not that they are *necessary*, in other words, some could nonetheless be irrelevant.

Benchmarking Fréchet distance results

We run benchmarking experiments for Fréchet distance computed over features from Inception (trained on ImageNet with supervision), SwAV (trained on ImageNet with self supervision), and DINO (trained on Faces with self supervision) networks. The distances are computed for 19 image sets (Fig. 1) with respect to CelebA-HQ images, for 5k samples.

With DINO specialized to the facial domain with standard-

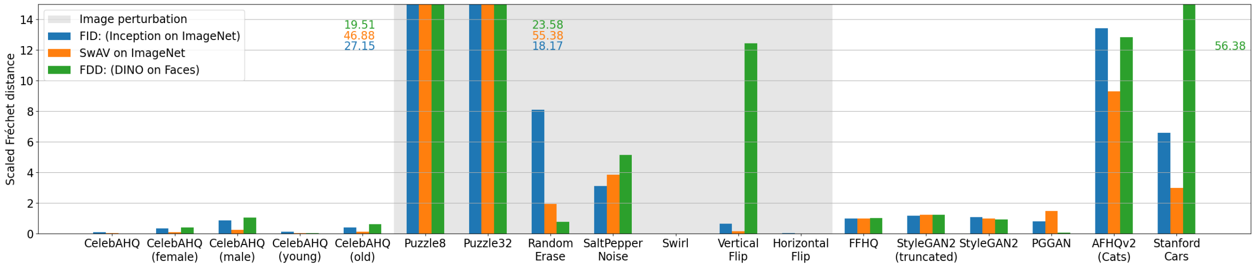


Figure 1. *Rescaled Fréchet distances computed on Inception features (trained on ImageNet), SwAV features (trained on ImageNet), and DINO features (trained on Faces). Each distance is between the x-axis sets and CelebA-HQ data (5k samples). For better readability, we rescale all values with a ratio fixed per method and determined on an independent dataset.*

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

Table 2. *User study rating how well images from different distributions correspond to random CelebA-HQ sets (1-5 score), the corresponding rescaled Fréchet distances (FID, FDD), and Pearson and Spearman correlation.*

ized faces, the distance is large when images are flipped vertically, while Inception and SwAV distances remain surprisingly small (smaller than the distance relative to FFHQ [31], which also contains faces). For random erasing of small patches, the distance is the smallest for DINO, which can extract high-level facial features rather than only fine-granularity generalized ones, due to its specialization to faces. Meanwhile, Inception distance caused by random erasing is even larger than the Inception distance between Cars and CelebA-HQ. Lastly, we note the large distance for DINO on car images, which are completely out of domain. This is not the case with cats, where facial features remain correlated to human facial features and are aligned in the same way in preprocessing. For the remaining setups, the distances obtained by the different approaches remain, on average, closely tied.

We also observe similar trends in Fig. 2 when tracking the training of the SemanticStyleGAN [32] with Fréchet distance on Inception and DINO. We only note that FDD is larger and drops faster than FID, as it is more sensitive to the low-quality faces initially synthesized. We conduct a user study (with 10 images per class) to obtain ratings on how well an image corresponds to the CelebA-HQ distribution represented by randomly sampled sets of 9 images at a time (Table 2). While the variation in scores over classes such as male and female is interesting, we mostly note that FID and FDD (relative to CelebA-HQ) strongly correlated with the participants' answers (lower distance correlates with higher correspondence).

Investigating photorealism correlation

We expand with an analysis of the connection between FID/FDD relative to CelebA-HQ and the photorealism of image distributions. We conduct a second user study to obtain opinion scores on photorealism based on 10 images per category. The results of FID and FDD (Table 3) are closely related on the different sets. For FFHQ and truncated StyleGAN2 [1] images, the distances match well with opinion scores, however, they diverge for untruncated StyleGAN2 and PGGAN [33], indicating that participant opinions are strongly affected by visual artifacts while the distance metrics focus more on content distributions. This is even more observable with PGGAN FDD; as PGGAN is trained on CelebA-HQ, its synthetic-image distribution matches better with it and leads to a low FDD despite lower visual quality. This further supports the claim that the specialized FDD focuses on high-level abstract information.

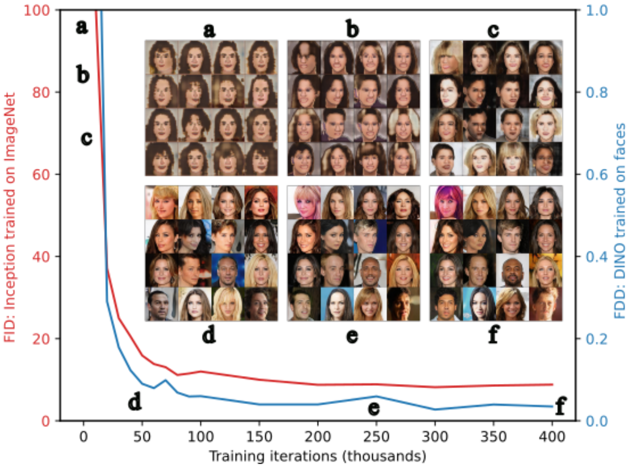
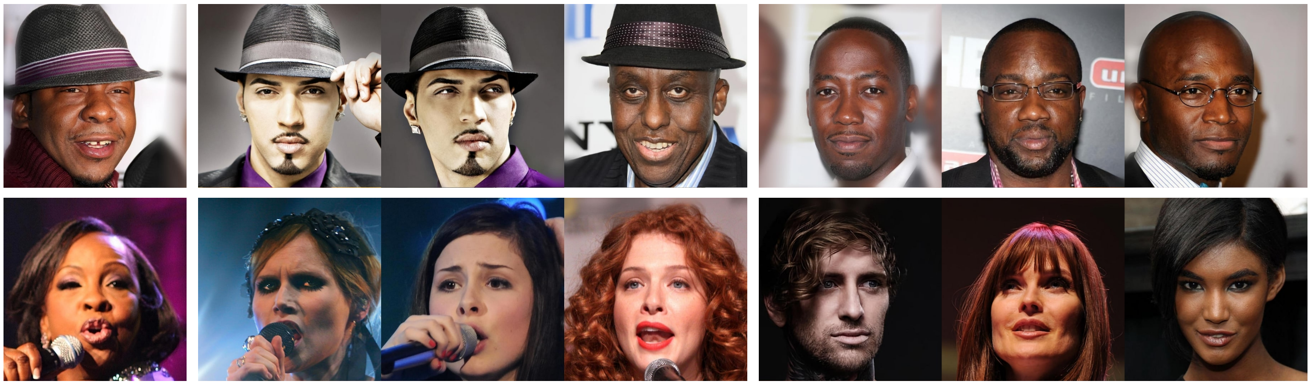


Figure 2. *Fréchet distances computed on ImageNet-trained Inception features and on Faces-trained DINO features, between synthetically generated images and CelebA-HQ images.*

Deeper dive into feature space neighborhoods

Finally, we narrow down to an image-level analysis of the feature spaces. We exploit local neighborhoods to analyze the feature-space landscapes. We select reference images and find their respective nearest neighbors in each of the two spaces. Sample results are shown in Fig. 3. We conduct a third user study where participants are asked to select which feature space induces neighbor images that are more similar to the reference (Table 4). The Inception space is by a large margin better for Stanford Cars images, as expected. For cats or random CelebA-HQ images, the



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

Figure 3. *Samples from our user study on feature space neighborhoods. For each reference image, we show its nearest neighbors in Inception and DINO feature spaces. Inception is biased towards objects (hat and microphone), while DINO can be perturbed by occluding objects (bottom). Therefore, Inception neighbors are not similar to the person, but simply wear similar hats.*

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

Table 3. *User study rating the photorealism of images from various sources (1-5 score), and the corresponding rescaled Fréchet distances relative to CelebA-HQ. *PGGAN is trained on CelebA-HQ, while the other models are trained on FFHQ.*

	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

Table 4. *User study selecting which of Inception or DINO nearest neighbors are most similar to the reference. Numbers reported are mean vote percentages. The first 4 rows are based on image similarity, "P." refers to person similarity.*

Inception space is more in accordance with human perception. However, when asking which neighbor set contains *people* who are more similar to the reference *person*, the DINO space correlates closer to the user choices for images with accessories. We note, however, that these aggregated results hide part of the analysis. Depicted in Fig. 3, we observe that, indeed, Inception is excessively biased towards focusing on objects rather than faces. But for DINO, the lack of such bias did not guarantee the desired face similarity results, as seen in the bottom row. The specialization of DINO features on faces makes them untrained for other objects, which risk becoming similar to an adversarial attack.

Conclusion and Key Take-aways

We analyze an open question on feature-space distance, particularly, the effects on Fréchet distance, and neighborhoods, of specializing the feature extractor to the facial domain. Our experiments and user studies support the following findings.

(1) Specialists become better at abstraction. Our experiments highlight that our specialized feature extractor can learn abstract concepts pertaining to faces. The generalist focuses more on fine-granularity features that can be exploited across tasks, making it more sensitive to spatially localized loss of information and less sensitive to global changes like an upside-down face, as shown in Fig. 1.

(2) Feature distance does not equate to photorealism. Fréchet distance measures statistics over dataset image features. This is affected both by photorealism and degradations, but also by the general content distribution across the dataset. When computing the Fréchet distance *relative* to a base dataset, it is important to use a high-quality one and to ensure that the base dataset contains a fair representation of desirable content. We emphasize that the vanilla distance is a holistic *image*-based distance rather than a *face* or identity distance.

(3) Noticing can be easier than not noticing. While we can train specialists for features relevant to a specialized domain, this does not guarantee their ability to dismiss all irrelevant information. Facing novel content in their input can act as adversarial attacks perturbing the specialized network (Fig. 3).

(4) The risk of smaller specialized datasets. Modern networks are large and this can lead to rich representations emerging even in randomly initialized ones. As the lottery ticket hypothesis [36] hints, multiple paths lead to the final representation, enough for coincidental features to appear. Training improves this representation making it more practical. Particularly, training over a massive dataset such as ImageNet constrains the behavior of the feature extractor across its many paths. This advantage can be lost when training a large specialist network on smaller domain-specific datasets, possibly leading to the weakness described in (3).

Our findings fill a gap in the literature, highlighting the trade-offs between general and specialized feature extractors. One avenue for future research is hybrid training, preserving well constrained extractors with low-granularity features and robustness, while learning abstract specialized features.

References

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *Computer Vision and Pattern Recognition*, 2020.

[2] M. El Helou and S. Süsstrunk, “BIGPrior: Towards decoupling learned prior hallucination and data fidelity in image restoration,” *IEEE Transactions on Image Processing*, 2022.

[3] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” in *Neural Information Processing Systems*, 2017.

[4] M. Fréchet, “Sur la distance de deux lois de probabilité,” in *Annales de l'ISUP*, vol. 6, 1957, pp. 183–198.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition*, 2009.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception architecture for computer vision,” in *Computer Vision and Pattern Recognition*, 2016.

[7] M. Soloveitchik, T. Diskin, E. Morin, and A. Wiesel, “Conditional Fréchet Inception distance,” *arXiv preprint arXiv:2103.11521*, 2021.

[8] M. Alfarra, J. Pérez, A. Frühstück, P. Torr, P. Wonka, and B. Ghanem, “On the robustness of quality measures for GANs,” in *European Conference on Computer Vision*, 2022.

[9] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *International Conference on Learning Representations*, 2018.

[10] M. Yang, C. Yang, Y. Zhang, Q. Bai, Y. Shen, and B. Dai, “Revisiting the evaluation of image synthesis with GANs,” *arXiv preprint arXiv:2304.01999*, 2023.

[11] C. Nash, J. Menick, S. Dieleman, and P. Battaglia, “Generating images with sparse representations,” in *International Conference on Machine Learning*, 2021.

[12] A. Sauer, K. Schwarz, and A. Geiger, “StyleGAN-XL: Scaling StyleGAN to large diverse datasets,” in *SIGGRAPH*, 2022.

[13] M. Naeem, S. Oh, Y. Uh, Y. Choi, and J. Yoo, “Reliable fidelity and diversity metrics for generative models,” in *International Conference on Machine Learning*, 2020.

[14] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Computer Vision and Pattern Recognition*, 2018.

[15] J. Lee, J.-H. Kim, and J.-S. Lee, “Demystifying randomly initialized networks for evaluating generative models,” in *AAAI Conference on Artificial Intelligence*, 2023.

[16] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved precision and recall metric for assessing generative models,” in *Neural Information Processing Systems*, 2019.

[17] M. El Helou, F. Dümbgen, and S. Süsstrunk, “AL2: Progressive activation loss for learning general representations in classification neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[18] B. Ramtoula, M. Gadd, P. Newman, and D. De Martini, “Visual DNA: Representing and comparing images using distributions of neuron activations,” in *Computer Vision and Pattern Recognition*, 2023.

[19] G. Parmar, R. Zhang, and J.-Y. Zhu, “On aliased resizing and surprising subtleties in GAN evaluation,” in *Computer Vision and Pattern Recognition*, 2022.

[20] M. J. Chong and D. Forsyth, “Effectively unbiased FID and Inception score and where to find them,” in *Computer Vision and Pattern Recognition*, 2020.

[21] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, and W. Brendel, “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” in *International Conference on Learning Representations*, 2019.

[22] K. Hermann, T. Chen, and S. Kornblith, “The origins and prevalence of texture bias in convolutional neural networks,” in *Neural Information Processing Systems*, 2020.

[23] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, “A study on the evaluation of generative models,” *arXiv preprint arXiv:2206.10935*, 2022.

[24] S. Steenkiste, K. Kurach, J. Schmidhuber, and S. Gelly, “Investigating object compositionality in generative adversarial networks,” in *Neural Networks*, 2020.

[25] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, “The role of ImageNet classes in Fréchet Inception distance,” in *International Conference on Learning Representations*, 2023.

[26] S. Morozov, A. Voynov, and A. Babenko, “On self-supervised image representations for GAN evaluation,” in *International Conference on Learning Representations*, 2021.

[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *IEEE International Conference on Computer Vision*, 2021.

[28] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “MaskGAN: Towards diverse and interactive facial image manipulation,” in *Computer Vision and Pattern Recognition*, 2020.

[29] K. Karkkainen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” in *Winter Conference on Applications of Computer Vision*, 2021.

[30] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” in *Neural Information Processing Systems*, 2020.

[31] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Computer Vision and Pattern Recognition*, 2019.

[32] Y. Shi, X. Yang, Y. Wan, and X. Shen, “SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing,” in *Computer Vision and Pattern Recognition*, 2022.

[33] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in *International Conference on Learning Representations*, 2018.

[34] Y. Choi, Y. Uh, J. Yoo, and J. W. Ha, “StarGAN: Diverse image synthesis for multiple domains,” in *Computer Vision and Pattern Recognition*, 2020.

[35] J. Krause, J. Deng, M. Stark, and L. Fei-Fei, “Collecting a large-scale dataset of fine-grained cars,” in *Computer Vision and Pattern Recognition Workshops*, 2013.

[36] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” in *International Conference on Learning Representations*, 2019.

Author Biography

Doruk Cetin is a software research engineer at Align Technology. Previously, he received his BS in computer engineering from the Middle East Technical University (2018) and his MS in computer science from

ETH Zurich (2021).

Benedikt Schesch is pursuing his MS in computer science at ETH Zurich, with a primary research focus on machine learning. Previously, he received his BS in computer science from ETH Zurich (2021). His work on this paper was done during an internship in ETH Zurich's Media Technology Center.

Petar Stamenkovic is a research engineer at ETH Zurich's Media Technology Center. Previously, he received his BS in electrical engineering from the University of Belgrade (2018) and his MS in electrical engineering and information technology from ETH Zurich (2021).

Majed El Helou is an established researcher at ETH Zurich's Media Technology Center, working on and supervising industry-based machine learning and computer vision projects. His main focus is on controllable and interpretable generative AI and visual media representation/understanding. Previously, he was a PhD/postdoc at EPFL's Image and Visual Representation Lab IVRL.