YouTubePD: A Multimodal Benchmark for Parkinson's Disease Analysis Supplementary Material

In this appendix, we include (1) discussion regarding our YouTubePD benchmark detail, release, and licensing, (2) additional analysis of the dataset, (3) additional illustration and details about our methods, (4) additional experimental results and analysis, (5) more discussion on limitations and future work, and (6) potential negative societal impact.

A1 Benchmark Detail, Release, and Licensing

We provide our code and benchmark at `https://uiuc-yuxiong-lab.github.io/YouTubePD`. We release our benchmark under the CC0 license. Here, we describe how we publicly release our benchmark:

1. We include all the code used in the process of converting publicly available YouTube videos into our benchmark.
2. We include all our annotations and extracted landmarks. Note that offensive content is not included in our dataset, as all the sources are publicly available interviews of public figures speaking openly about their experiences with Parkinson's Disease.

Despite that all the videos used in our benchmark are publicly available YouTube videos, we are also actively taking steps to approach the public figures involved to respect their autonomy and privacy. This ensures that we uphold the highest standards of ethical data usage. We strive to balance open access and reproducibility with respect for privacy, all while providing a resource that could significantly advance the analysis and understanding of Parkinson's Disease (PD).

A2 Dataset: Additional Analysis

A2.1 Dataset Statistics

In Table A1, we summarize the severity label distribution in YouTubePD. This includes severity labels for the overall subject in each video and severity labels for each of the 14 important facial regions informative for PD analysis. The number of annotations varies depending on severity levels and regions.

We also summarize the demographic distribution in YouTubePD, split between PD-positive and healthy control (HC), or PD-negative, subjects. Table A2 provides the country, race, and gender statistics of PD-positive subjects in YouTubePD. Similarly, Table A3 provides the country, race, and gender statistics of HC subjects in YouTubePD.

We would like to provide additional details in our annotation process, particularly regarding how we denote the severity of PD. Our annotation strategy utilizes a detailed scale, ranging from 0 to 5, where 0 signifies a healthy individual, and 5 corresponds to severe PD. We do not apply the Unified Parkinson's Disease Rating Scale (UPDRS) [4] for facial expression. This decision is based on the clinician's suggestion, since an accurate UPDRS facial expression rating would require more information (e.g., observing the subject's facial expression pattern at rest or when not talking) than

Region/Severity	0	1	2	3	4	5
Overall Video	187	15	8	10	17	7
Forehead	187	12	3	4	16	18
Left nasolabial fold	187	9	6	10	20	10
Right nasolabial fold	187	11	7	10	21	8
Right lip crease	187	1	0	3	7	3
Left lip crease	187	1	0	2	10	3
Left outer eye	187	4	2	1	6	17
Right outer eye	187	2	2	2	5	16
Between eyebrows	187	5	1	3	5	3
Right above between eyebrows	187	3	3	4	12	5
Right eye	187	15	4	3	8	6
Left eye	187	16	4	3	8	6
Mouth	187	1	0	1	5	0
Right eyebrow	187	6	5	8	11	12
Left eyebrow	187	6	4	7	10	12
Total Annotations	2808	107	49	71	161	126

Table A1: Distribution of severity labels in YouTubePD for the overall video-level analysis and for all 14 facial regions, with 0 denoting the absence of PD and 5 indicating a severe form of PD.

Country	Count	Race	Count	Gender	Count
United States	12	White/Caucasian	13	Male	15
United Kingdom	3	Black/African descent	3	Female	1
Canada	1				

Table A2: Country/race/gender statistics of PD-positive public figures in YouTubePD.

Country	Count	Race	Count	Gender	Count
South Africa	1	South African+Swiss-German	1	Male	68
United States	52	Black/African Descent+Filipino	1	Female	23
United Kingdom	12	Black/African Descent	11		
Israel	3	White/Caucasian	63		
Russia/Canada	1	Indian	1		
Sweden	2	Latinx	4		
Kenya/Mexico	2	Asian	9		
Brazil	2	Black/African Descent+Samoan	1		
Serbia	1				
South Korea	2				
Puerto Rico	1				
Mexico	1				
Japan	1				
Canada	4				
Denmark	1				
Russia	2				
Germany	1				
France	2				

Table A3: Country/race/gender statistics of healthy control or PD-negative public figures in YouTubePD.

facial expression videos contain. This strategy also allows for a finer classification. In addition, we do not apply UPDRS because facial expression and audio have distinct UPDRS standards. We instead use the holistic severity and confidence annotation based on the video. Doing so ensures the label consistency between the audio data and other modalities, thereby facilitating multimodal PD classification.

In addition, we provide (i) illustrations of our facial landmarks and regions in detail in Figures A1, A2, and A3; (ii) the longitudinal statistics of our PD videos showing their distribution in time in Figure A4.

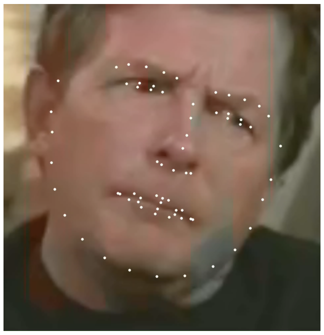


Figure A1: Original landmark extraction.

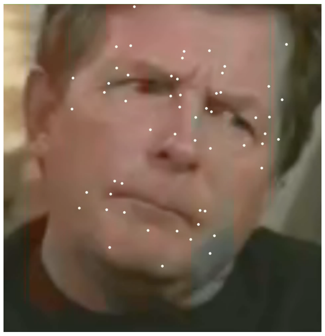


Figure A2: Interpolated landmarks.

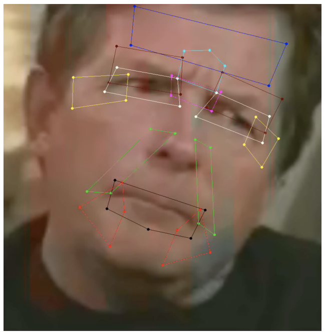


Figure A3: Visualized regions from interpolated landmarks.

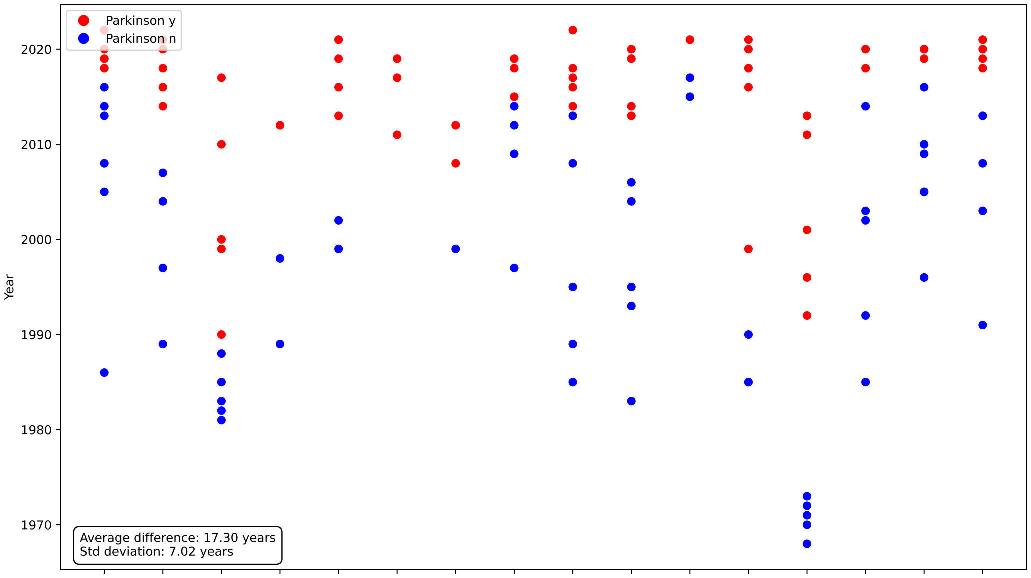


Figure A4: Longitudinal data for the time gap between PD-negative and PD-positive videos. The x-axis represents public figures, while the y-axis represents the time in years. Red dots denote videos with PD-positive labels, and blue dots denote videos with PD-negative labels.

A2.2 Are Annotated Regions Correlated to PD Severity?

To understand if the annotated regions are informative for PD severity, we investigate the correlation between the 14 annotated facial regions and a patient's annotated PD severity level. We accomplish this by training a linear classifier that takes as input the annotated informative region index and predicts the severity level. More specifically, for each video, the input is a binary vector which maps the clinician-annotated facial region indices, while the output is matched to the annotated severity level of the video. As a control experiment, we instead train on random region-level annotations as input. We find that the linear model trained on the actual region annotation achieves 70% test accuracy, while the model trained on random annotations achieves 15% test accuracy. This validates that the severity level is predictable from the annotated informative regions, but not from random region annotations. Therefore, *the annotated regions and PD severity are correlated.* This is also consistent with our approach that leverages region-level information for improved PD classification performance.

Furthermore, we can examine the learned weights and biases of the linear model to understand how the model has learned to classify. We find that in general, higher severity patients have positive annotations on a larger number of informative regions (more symptoms), and vice versa. The model also leverages different facial regions to determine severity at different levels; for example, very severe cases could be easily distinguished via eyes and lips feature. Figure 3 in the main paper

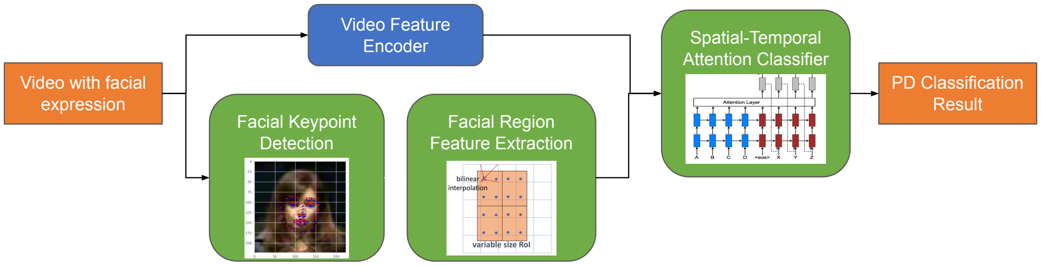

Video with facial expression
Video Feature Encoder
Facial Keypoint Detection
Facial Region Feature Extraction
bilinear interpolation
variable size RoI
Spatial-Temporal Attention Classifier
Attention Layer
PD Classification Result

Figure A5: Illustration of our baseline method for the facial expression-based PD classification. An input video is processed through two branches, one for video features and the other for region features. These features are then aggregated with a spatial-temporal attention classifier to obtain the PD classification result.

visualizes the most informative regions at each severity level. Corresponding regions are indicated by highlighted facial creases in the figure and bounding boxes in the input video frames.

A3 Methods: Additional Illustration and Details

A3.1 Model Architecture Illustration

We provide an additional illustration of our baseline method for the first task in Figure A5. An input video is processed through two branches, one for video features and the other for region features. These features are then aggregated with a spatial-temporal attention classifier to obtain the PD classification result.

A3.2 Implementation Details

For the facial-expression-based classification method, we use $\lambda = 0.9$, emphasizing the binary portion of the loss. We use all 14 regions, i.e., $|R| = 14$. We use 8 frames from each video and a batch size of 32. We train our models with the Adam [8] optimizer with a learning rate of 0.001. More details are provided in the code. All experiments corresponding to this task are conducted on a single 16GB NVIDIA V100 GPU.

For the audio baseline, we utilize the entire audio and average the representations to obtain a single 768-dimensional audio-level feature vector. Afterwards, we use a batch size of 64 and feed the input through an MLP with one hidden layer of size 1,024. We train the model using Adam with a learning rate of 0.0003. Similarly, for the landmark baseline, we use 8 frames to maintain consistency with the other modalities, and use the same hyperparameters for Adam. For both modalities, we use $\lambda = 0.5$. All experiments are conducted on a single NVIDIA 3060 GPU.

For the multimodal fusion baseline, we use 8 frames from each video. The batch size is set as 16 and trained for 50 epochs. We train the model using Adam with a learning rate of 0.02. All experiments are conducted on a single NVIDIA 4090 GPU.

For PD progression synthesis, we follow the settings used in the official implementations [3, 7, 10, 11, 12, 14, 15]. In addition, we make modifications to two methods to align with our task setting. Specifically, we replace the age classifier in HRFAE [14] with our trained PD binary classification model. As for JoJoGAN [3], we iteratively sample images during training to utilize all available images. All experiments are conducted on a single NVIDIA 4090 GPU.

A4 Experiments: Additional Results and Analysis

A4.1 Additional Ablations

Alternative audio representations. We additionally explore alternative feature representations for audios beyond masked auto-encoders (MAE) [9] presented in the main paper – namely, wave2vec (W2V) 2.0 [1], a deep feature extractor for audios that also uses self-supervised learning, and MEL-frequency cepstral coefficients (MFCC) [2], which have demonstrated reasonable success in general audio processing tasks. We find empirically that MAE features work the best with regards to metrics and consistency, as shown in Table A4. In general, wav2vec features remain competitive with the MAE features though trailing slightly, while MFCC performs considerably worse likely due to its inability to express complex features necessary for the task, given the simplicity of the classification model. Note that the hand-crafted MFCC achieves a relatively high performance on AUROC, since it effectively makes random guess among all classes. The other *learned* features are more biased by the data imbalance between 0 and all other classes. This is due to the fact that AUROC is computed as an unweighted average of one-vs-all calculations, leading to MFCC to incorrectly appear competitive on that metric.

Audio Feature	Top-1 Acc ↑	F1 ↑	AUROC ↑	MSE ↓
MAE [9]	47.79(±1.66)	0.16(±0.01)	0.50(±0.02)	6.26 (±0.39)
W2V [1]	57.90(±7.46)	0.14(±0.02)	0.43(±0.04)	4.50(±0.93)
MFCC [2]	39.42(±9.66)	0.11(±0.03)	0.49(±0.06)	6.48(±1.80)

Table A4: Ablation study on audio representations for multiclass classification on YouTubePD. 'MAE' denotes masked auto-encoders presented in the main paper; 'W2V' denotes wave2vec 2.0; 'MFCC' denotes MEL-frequency cepstral coefficients. We find that MAE provides the most stable and consistent results – we prioritize F1 and AUROC, as the other metrics are influenced by the data imbalance.

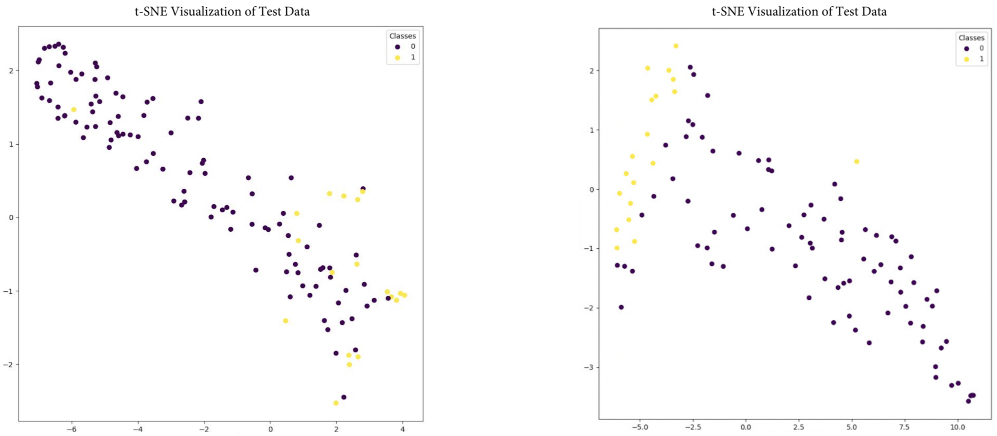
Multimodal fusion strategies. We conduct an ablation study on the multimodal fusion strategy. Specifically, we explore two different strategies to produce the logits for the facial expression video and facial landmark modalities. One way is frame concatenation (FC) where we concatenate the frame features, and the other is frame voting (FV) where we perform voting to aggregate the result of each frame. Note that FV is used for results in Table 3. For FC, we concatenate frames' features to one vector representing the whole video (either image features from ResNet or landmark coordinates). Then, we train a video-level classifier to obtain the video logits. For FV, we train a frame-level classifier for each frame and average the predictions as video-level logits. For both strategies, the video-level logits from different modalities are averaged to get the final prediction. As shown in Table A5, with FC, the multimodal performance is even lower than the single facial expression modality on F1 and AUROC metrics. By contrast, the FV strategy helps to improve performance.

Audio Feature	Top-1 Acc ↑	F1 ↑	AUROC ↑	MSE ↓
VGGFace [5]	78.20(±3.13)	0.23(±0.02)	0.68(±0.01)	2.29(±0.77)
Multimodal (FC)	79.23(±1.94)	0.21(±0.02)	0.69(±0.01)	1.76(±0.18)
Multimodal (FV)	82.75(±2.85)	0.28(±0.02)	0.80(±0.03)	1.40(±0.25)

Table A5: Ablation study on multimodal fusion strategies for multiclass classification on YouTubePD. 'FC' denotes frame concatenation, and 'FV' denotes frame voting. We find that the frame voting strategy improves the fusion performance, while the frame concatenation strategy even leads to a decrease in performance on F1 and AUROC metrics, compared with the single facial expression modality.

A4.2 t-SNE Visualizations

We provide qualitative results for the performance of our facial-expression-based classification model. We use t-SNE [13] on both YouTubePD and the private clinical dataset [6], shown in Figure A6. Our approach is able to clearly separate the PD-positive and PD-negative classes on both distributions.



(a) t-SNE visualization for YouTubePD.

(b) t-SNE visualization for the facial activation set with clinical patients.

Figure A6: t-SNE visualizations of our learned facial expression representation for healthy (class 0) and PD-positive (class 1) subjects on YouTubePD and the clinical dataset [6]. Intriguingly, we observe a more pronounced separation between the PD-positive and PD-negative classes on clinical data, demonstrating the generalizability of our learned representation from the in-the-wild YouTube videos.

A5 More Discussion on Limitations and Future Work

Limitations. In the main paper, we have briefly discussed the limitations of our work. Here, we provide a more in-depth discussion. First, as our dataset is composed of publicly available YouTube videos of public figures, the subjects and video samples in the dataset may not adequately represent the wide range of individuals affected by PD. The videos primarily capture interview scenarios, which may not effectively showcase the indicative symptoms of subjects, compared with the motor tasks and instructions typically used in medical studies. Furthermore, there is a demographic bias in the dataset subjects, as they are all public figures (predominantly male) with very few available details about their treatment course and disease progression. Meanwhile, we are not aware of the treatment or treatment response experienced by these individuals.

Second, it is necessary to conduct further investigation and analysis of the performance and deployment of models developed using our benchmark in real-world clinical scenarios. In the main paper, we have demonstrated that our method developed on the benchmark exhibits promising results on a clinical dataset. More comprehensive evaluation on additional clinical datasets would validate the broad generalizability of our benchmark and associated models to practical medical applications.

Finally, the size of the dataset is relatively small compared with typical computer vision datasets, due to the inherent challenges involved in collecting PD data. The small dataset size increases the difficulty of developing and training larger models from scratch, often necessitating some form of finetuning to achieve reasonable performance.

Future work. These limitations open up a wide scope of future advancements and progress in this field. As mentioned previously, while our benchmark represents a strong first step, further comprehensive datasets and benchmarks are necessary to thoroughly evaluate the performance and generalizability of methodologies prior to their deployment. Moreover, our findings highlight the potential of developing multimodal frameworks that leverage various examination modalities and track complementary symptoms, such as facial expression, speech, posture, and gait for PD classification. Although PD classification has been the primary focus (as in our first two proposed tasks), we note that there are interesting unexplored directions in this realm, particularly in generative tasks like progression synthesis (as in our third proposed task), which can serve as effective augmentation and learning techniques.

A6 Ethics Discussion

A6.1 Personally Identifiable Information and Informed Consent

YouTubePD may include personally identifiable information (PII) or sensitive personally identifiable information. The data we collect from YouTube include facial expressions, PD identity, and audios. However, we would like to highlight that the public figures chose to make their struggles with PD public and discussed their disease and diagnosis in front of cameras. By willingly revealing their identifiable faces and voices, the public figures do not intend to keep their PD information fully private. We believe that the concern regarding a breach of privacy is not a newly raised issue specific to our work, as the possibility of any misuse of these videos already exists.

The central question we posed to ourselves was whether sharing these videos with our research community, along with annotations of facial expressions, would amplify the risk of misuse. We are of the opinion that this action does not escalate the aforementioned risk. To further address this matter, we took the initiative to contact the public figures involved and requested permission. This step was taken proactively, particularly in the event that the public figures had regrets about their previous decision to go public and now wished to make a different choice.

Regarding PII and sensitive PII, we are fully aware of the sensitive nature of the data we are working with. In order to safeguard individuals' privacy, we have sought both guidance from the Institutional Review Board (IRB) Office and legal guidance from the Legal Department at University of Illinois Urbana-Champaign to ensure compliance with regulations. Furthermore, in line with ethical norms, we have made efforts to obtain explicit consent from each public figure featured in the videos. The consent form clearly includes our intention in using these data and how these data are expected to be used. We remove videos of public figures who wish not to be part of our dataset. In addition, we acknowledge the concern about the potential for individuals and their families to feel uncomfortable with the label of "illness." While we respect this sensitivity, we emphasize that our intention is to contribute to a better understanding of PD, its impact on facial expressions, facial landmarks, and audios, and the potential for technological advancements. We approach this research with the utmost respect for the individuals involved and strive to contribute positively to the discourse around the disease.

A6.2 Negative Societal Impact

While our work provides promising advancements in AI-assisted analysis and severity evaluation of PD, we recognize that it may also present potential negative societal impacts that deserve careful consideration.

The first concern pertains to privacy. The videos we use for our work are publicly available, featuring public figures who openly discuss their experience with PD. However, widespread use of similar technology could raise issues of privacy, as individuals may not wish to have their health condition detected or revealed, even inadvertently, through casual video or audio footage. As healthcare professionals and researchers, it is critical to respect patient privacy and consent in all facets of care and study. Second, there is a risk of misuse or over-reliance on our technology. While the goal of our work is to aid the detection of PD, it should not replace the professional diagnosis of healthcare providers. Misinterpretation or misuse of this technology may lead to false positives or negatives, causing unnecessary distress or false reassurance. Lastly, issues of inequity may also arise. Access to advanced diagnostic tools such as the one we propose may be limited, due to geographic location, financial constraints, or digital literacy. As such, this technology could inadvertently widen the healthcare disparity between different socioeconomic groups.

In light of these potential societal impacts, it is essential that proper protocols and measures are put in place to guide the ethical use of such technologies. This includes clear communication about the tool's intended use, rigorous validation processes, and ongoing dialogue about equitable access to and use of these technological advances.

A6.3 Mitigating Bias and Negative Societal Impacts

Some ethical risks exist in the originally publicly available YouTube videos. We are aware that we cannot mitigate those risks to zero. There will be a rest risk. Again, we emphasize that our intention is to enhance the better understanding of PD, its effects on facial expressions, facial landmarks, and audios, as well as explore the potential for technological advancements in this field. We aim to ensure the highest possible standards of ethical conduct in downstream research.

A6.4 Responsibility of AI-Assisted Systems

As mentioned in Section 7 of the main paper, our benchmark aims to inspire a PD early screening tool based on modern machine learning and computer vision techniques. This tool would assist primary clinicians in identifying individuals who may be displaying early physical signs indicative of an evolving Parkinson's syndrome. These individuals can then undergo further neurological evaluation as clinically indicated. This proactive approach will expedite diagnosis and treatment, potentially leading to improved outcomes. On the other hand, while we acknowledge the potential of facial videos and audios in aiding PD detection, we do not advocate for clinicians to rely solely on these modalities for diagnosis. Instead, if positive detection results emerge from facial videos and audios, we recommend that patients seek medical attention at an earlier stage and obtain a more comprehensive diagnosis using additional assessments, such as Dopamine Transporter Scan (DAT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET).

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Conference on Neural Information Processing (NeurIPS)*, 2020. 5

[2] Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch, and Sara Sandabad. Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. In *International Conference on Electrical and Information Technologies (ICEIT)*, 2015. 5

[3] Min Jin Chong and David Forsyth. JoJoGan: One shot face stylization. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[4] Christopher G Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T Stebbins, Matthew B Stern, Barbara C Tilley, Richard Dodel, Robert Holloway Bruno Dubois, Joseph Jankovic, Jaime Kulisevsky, Anthony E Lang, Andrew Lees, Sue Leurgans, Peter A LeWitt, David Nyenhuis, C Warren Olanow, Olivier Rascol, Anette Schrag, Jeanne A Teresi, Jacobus J Van Hilten, and Nancy LaPelle. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale MDS-UPDRS: Scale presentation and clinimetric testing results. In *Movement Disorders*, 2008. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[6] Trung-Hieu Hoang, Mona Zehni, Huaijin Xu, George Heintz, Christopher Zallek, and Minh N. Do. Towards a comprehensive solution for a vision-based digitized neurological examination. *IEEE Journal of Biomedical and Health Informatics*, 2022. 5, 6

[7] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Conference on Neural Information Processing (NeurIPS)*, 2020. 4

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[9] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. In *HEAR: Holistic Evaluation of Audio Representations (Neural Information Processing Systems 2021 Competition)*, Proceedings of Machine Learning Research, 2022. 5

[10] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[11] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 2020. 4

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. In *Journal of Machine Learning Research*, 2008. 5

[14] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. In *International Conference on Pattern Recognition (ICPR)*, 2021. 4

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 4