Poster: Diffusion Based Conditional DeepFake Generation

Liyue Fan, Joseph Roberson
Department of Computer Science
UNC Charlotte
{liyue.fan, jrobe187}@charlotte.edu

***Abstract*—DeepFake technology has important implications on creative work and potential consequences in society. Research shows that the curation of DeepFake datasets can advance the development of DeepFake detection methods. Existing DeepFake datasets adopted emerging generative models, such as diffusion models. This study aims to understand "conditional generation", which produces new synthetic content based on input conditions. We propose to extend existing DeepFake datasets by incorporating recent results in diffusion based image editing and image morphing. Furthermore, we quantify the usefulness of generated datasets using state-of-the-art measures for generative quality and DeepFake detection. Our results show that samples generated by image editing and image morphing differ from existing face forgery datasets and provide interesting additions.**

***Index Terms*—DeepFake, Diffusion Models, Conditional Generation**

I. Introduction

Malicious applications or misuse of DeepFake techniques may lead to grave societal consequences, such as the spread of misinformation, and impersonation and defamation of public figures [1]. Significant efforts have been made by researchers to organize DeepFake detection challenges and to publicly release datasets, e.g., [2]–[5]. Despite those promising results, we identify a few limitations in diffusion based DeepFake datasets [5]. Firstly, conditional generation in [5] adopted standard approaches, without considering recent results in editing images and interpolating images in the latent space [6], [7]. Furthermore, usefulness evaluation of generated data is under-explored in [5] and prior research. In this study, we propose to extend existing diffusion based DeepFake datasets with new generation methods and to quantify the usefulness of generated data using state-of-the-art approaches.

Related Work. Curating DeepFake datasets requires significant efforts and can help advance the development and validation of DeepFake detection models. Several datasets have been publicly released for research purposes. To name a few, FaceForensics++ [2], Celeb-DF [3], and DFDC [4] provide real and manipulated video sequences obtained with a range of manipulation methods, such as FaceSwap, autoencoders, and StyleGAN. Recently, DiffusionFace dataset [5] was created and it contains synthetic images generated by 11 state-of-the-art diffusion based methods, including unconditional generation methods and conditional generation methods. Nevertheless, existing datasets do not take into consideration recent results in image editing [6] and image morphing [7], leaving out potential DeepFake generation methods.

This material is based upon work supported by the National Science Foundation under Grant Number 2027114. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

II. Conditional DeepFake Generation

Conditional image generation informs the data synthesis process with input conditions, producing contextually coherent fake content. This approach is essential for simulating realistic DeepFake scenarios where specific content or characteristics (such as identity and surroundings) are altered or preserved. Prior work [5] studied several conditional generation methods based on diffusion models, such as text-guided and image-guided image generation with Stable Diffusion, inpainting, and DiffSwap [8]. In this study, we propose to generate new DeepFake images with recent methods for image editing [6] and image morphing [7].

Generation Methods. We adopt the x-space guidance approach in [6] to edit images along the directions of local basis vectors. In addition, we adopt DiffMorpher [7] to create a smooth interpolation between two input images. Both methods leverage the latent space of diffusion models and generate semantically meaningful synthetic images based on real input. In this study, we edit an image along the top 2 directions of its local basis and generate 16 frames for morphing a pair of images (where the first and the last are input images).

Performance Metrics. We evaluate the generated datasets on the performance of DeepFake detection as well as quality metrics for generative models. For DeepFake detection, we employ a state-of-the-art method [9] that classifies real vs. fake images in the latent space of a CLIP:ViT model. To study the quality of generated data, we adopt the Fréchet Inception Distance (FID) [10] and the improved precision and recall measure (IPR) [11].

Preliminary Results. Table I and Table II report samples obtained from imaging editing and image morphing respectively. We observe that imaging editing along latent basis produces meaningful changes. Furthermore, image morphing in the latent space of stable diffusion produces smooth and natural transitions between two input images. Overall, these results validate those generation methods in producing plausible

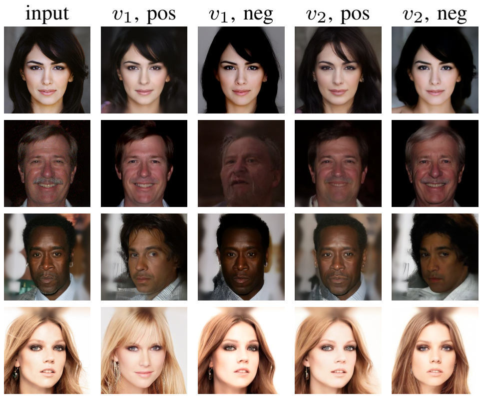


TABLE I

SAMPLES OF IMAGE EDITING (BEST VIEWED ZOOMED IN): WE ADOPTED TOP-2 LOCAL BASIS VECTORS AND EDITS WERE MADE IN BOTH POSITIVE AND NEGATIVE DIRECTIONS.

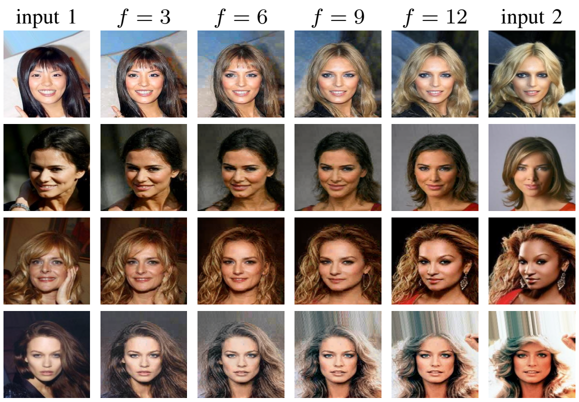


TABLE II

SAMPLES OF IMAGE MORPHING (BEST VIEWED ZOOMED IN): EACH ROW SHOWS INTERMEDIATE INTERPOLATIONS BETWEEN TWO INPUT IMAGES.

synthetic images. At the conference, we will present evaluation results on DeepFake detection and quality metrics on the newly generated datasets.

REFERENCES

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[2] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[4] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[5] Z. Chen, K. Sun, Z. Zhou, X. Lin, X. Sun, L. Cao, and R. Ji, "Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis," *arXiv preprint arXiv:2403.18471*, 2024.

[6] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 129–24 142, 2023.

[7] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan, "Diffmorpher: Unleashing the capability of diffusion models for image morphing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7912–7921.

[8] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577.

[9] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[10] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[11] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

Diffusion Based Conditional DeepFake Generation

Liyue Fan PhD, Joseph Roberson
Liyue.fan@charlotte.edu, jrob187@charlotte.edu
University of North Carolina at Charlotte

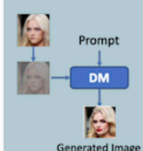
UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Introduction

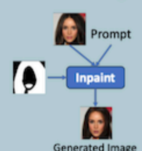
- Malicious applications of DeepFake techniques may lead to grave consequences
- Significant efforts have been made to organize DeepFake detection challenges and to curate datasets [1-4]
- **Conditional generation** informs the data synthesis process with input conditions
- Recent datasets [4] include four types of generation method:

Text-guided

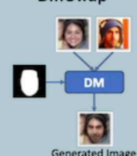
Image-guided



Inpainting



DiffSwap



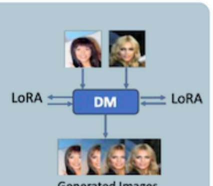
Motivation of this study

1. Recent results in latent space editing and interpolation [5,6] may produce new DeepFake datasets
2. Evaluation on DeepFake detection and generative quality may quantify the usefulness of DeepFake datasets

Generation Method

Image Morphing

- [6] generates smooth transition between two input images
- LoRA parameters and latent noises are interpolated



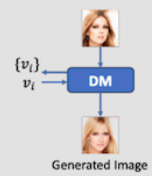
Generation Method (cont.)

Image Editing

- Recent work [5] proposes x-space guidance:

$$\tilde{x}_t = x_t + \gamma[\epsilon_\theta(x_t + v) - \epsilon_\theta(x_t)]$$

- Images can be edited in the latent space along any local basis vectors, e.g., v



Results

Qualitative Evaluation

Image Editing samples:

Image Morphing samples:

Quality metrics of generated datasets

- FID: distributional similarity between real and fake images
- Improved Precision and Recall (IPR): trade-off between sample quality and manifold coverage

Dataset	FID ⇓	Precision ⇑	Recall ⇑
IM $s = 3$	18.569	0.732	0.381
IM $s = 6$	41.100	0.637	0.180
IM $s = 9$	41.184	0.626	0.183
IM $s = 12$	18.860	0.739	0.373
IE v_1,pos	35.145	0.774	0.418
IE v_1,neg	28.253	0.75	0.417
IE v_2,pos	30.865	0.762	0.420
IE v_2,neg	29.337	0.774	0.435

⇑: higher is better; ⇓: lower is better

Results (cont.)

DeepFake Detection on generated datasets

- Classifies real and fake images on CLIP-ViT embeddings [7]
- Model is trained with real and stable diffusion image-guided generation images, and tested on real and IM or IE images

Dataset	TNR ⇑	TPR ⇑	Accuracy ⇑	AUC ⇑
IM $s = 3$	0.001	0.999	0.500	0.366
IM $s = 6$	0.005	0.998	0.501	0.437
IM $s = 9$	0.007	0.995	0.501	0.437
IM $s = 12$	1.000	0.000	0.500	0.368
IE v_1,pos	0.851	0.815	0.833	0.914
IE v_1,neg	0.851	0.820	0.835	0.913
IE v_2,pos	0.872	0.846	0.859	0.935
IE v_2,neg	0.874	0.826	0.850	0.925

Discussions

- Image editing and image morphing generate visually plausible synthetic images and reasonable quality metrics.
- These newly generated datasets may reveal new challenges to DeepFake detection methods, e.g., low performance on IM.
- Future work may conduct a comprehensive analysis to understand the classification of real vs. generated distributions.

References

1. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.
2. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216.
3. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," arXiv preprint arXiv:2006.07397, 2020.
4. Z. Chen, K. Sun, Z. Zhou, X. Lin, X. Sun, L. Cao, and R. Ji, "Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis," arXiv preprint arXiv:2403.18471, 2024.
5. Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of Riemannian geometry," Advances in Neural Information Processing Systems, vol. 36, pp. 24 129–24 142, 2023.
6. K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan, "Diffmorpher: Un-leashing the capability of diffusion models for image morphing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7912–7921.
7. U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24 480–24 489.
8. T. Kynk¨a¨anniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," Advances in Neural Information Processing Systems, vol. 32, 2019.

Acknowledgement: LF and JR are supported in part by the National Science Foundation under grant CNS-2027114. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.