Portrait Expression Editing With Mobile Photo Sequence

Yiqin Zhao*
yzhao11@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Rohit Pandey
rohitpandey@google.com
Google
Mountain View, California
USA

Yinda Zhang
yindaz@google.com
Google
Mountain View, California
USA

Ruofei Du
ruofeidu@google.com
Google
Mountain View, California
USA

Feitong Tan
feitongtan@google.com
Google
Mountain View, California
USA

Chetan Ramaiah
cramaiah@google.com
Google
Mountain View, California
USA

Tian Guo
tian@wpi.edu
Worcester Polytechnic Institute
Worcester, MA, USA

Sean Fanello
seanfa@google.com
Google
Mountain View, California
USA

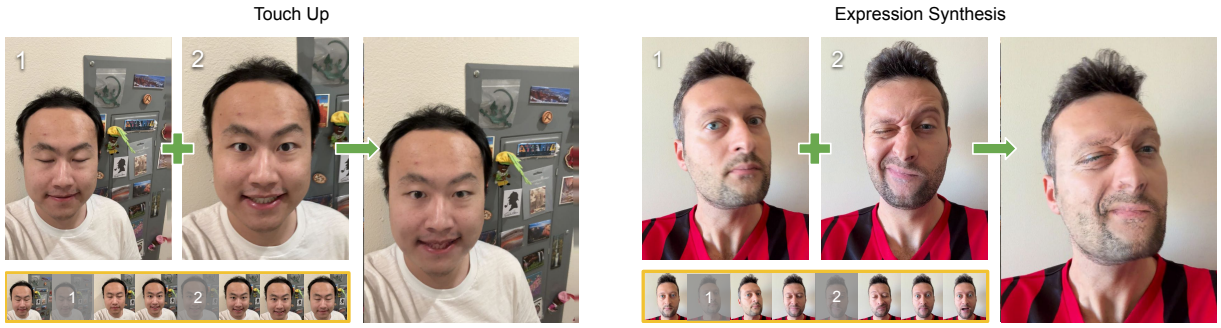


Figure 1: A result overview of ExShot, a high-fidelity portrait facial expression editing system. Given casually captured portrait photos, users can select a pair representing a target head pose and target expressions, ExShot synthesizes a high-quality photorealistic portrait with the desired head pose and expressions.

ABSTRACT

Mobile cameras have revolutionized content creation, allowing casual users to capture professional-looking photos. However, capturing the perfect moment can still be challenging, making post-capture editing desirable. In this work, we introduce ExShot, a mobile-oriented expression editing system that delivers high-quality, fast, and interactive editing experiences. Unlike existing methods that rely on learning expression priors, we leverage mobile photo sequences to extract expression information on demand. This design insight enables ExShot to address challenges related to diverse expressions, facial details, environment entanglement, and interactive editing. At the core lies ExprNet, a lightweight deep learning model that extracts and refines expression features. To train our model, we captured portrait images with diverse expressions, incorporating pre-processing and lighting augmentation techniques to ensure data quality. Our comprehensive evaluation results demonstrate that ExShot outperforms other editing approaches by up to 29.02% in PSNR. Ablation studies validate the effectiveness of our design choices, and user studies with 28 participants confirm the strong desire for expression editing and the superior synthesis quality of ExShot, while also identifying areas for further investigation.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Image manipulation**; **Rendering**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Portrait expression editing; Neural rendering; Mobile system

ACM Reference Format:
Yiqin Zhao, Rohit Pandey, Yinda Zhang, Ruofei Du, Feitong Tan, Chetan Ramaiah, Tian Guo, and Sean Fanello. 2023. Portrait Expression Editing With Mobile Photo Sequence. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610543.3626160

1 INTRODUCTION

Facial expressions are essential for human communication. However, it can be challenging for casual users to capture images with

*Work completed while the author was an intern at Google.

SA Technical Communications '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0314-0/23/12…$15.00
https://doi.org/10.1145/3610543.3626160

the desired facial expression and other portrait features, e.g., head pose, camera pose and background. Providing the portrait expression editing feature to mobile devices could liberate and enhance the content creation process with mobile cameras. With the ability to edit expressions, users can convey a wider range of emotions and messages, and tailor the emotional content of digital portraits to their specific needs or preferences.

Traditional image editing methods typically rely on using control points or face 3D geometries to warp images [Rao et al. 2020]. While these methods could work reasonably for simple touchup tasks, they often fall short in synthesizing expression details that are not already present in the current photo. Recent works on portrait expression editing, such as learning expression priors from large portrait datasets [Doukas et al. 2021; Siarohin et al. 2019; Yin et al. 2022] or personalized portrait captures [Bai et al. 2023; Grassal et al. 2022], can produce good synthesized results. However, their editing qualities are heavily depend on the quality and diversity of the training dataset [Yin et al. 2022]; and their editing performances can be subject to the time-consuming per-personal model training and heavyweight inference [Bai et al. 2023].

Achieving high-quality portrait expression editing for mobile devices faces three key challenges. The first is obtaining accurate *expression priors* is vital for achieving high-quality results in expression editing. Current methods rely on learning from extensive datasets, which may not be suitable for editing real-world data due to biases and privacy concerns. The second challenge lies in complex environment entanglement. Facial expressions in casual portrait photos can be entangled with undesirable head poses and illumination conditions. Finally, users prefer interactive expression editing, involving multiple consecutive changes to the same image. To provide a seamless editing experience, portrait editing systems must offer low editing latency.

To address these challenges, we design ExShot, a system that uses a lightweight deep learning model and an efficient image processing pipeline for expressive portrait expression editing. In summary, we make the following key contributions:

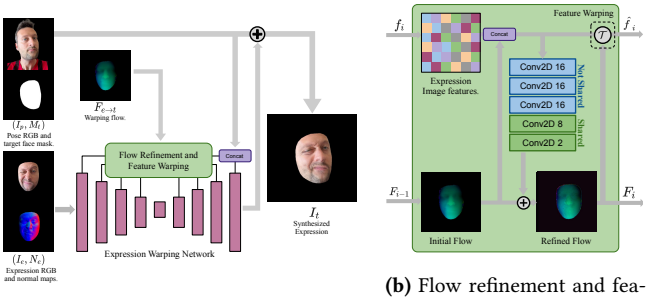
- We design a novel deep learning model, ExprNet, which can synthesize high-fidelity facial expression images. By explicitly modeling the expression transformation, our model can generalize to unseen identities and expressions. The resulting images are of high resolution, ensuring that the details of the facial expressions are preserved.
- We perform an in-depth evaluation that shows the effectiveness of ExShot. Quantitatively, ExShot achieves 29.02% higher PSNR than a recent system StyleHEAT [Yin et al. 2022], indicating accurate and faithful expression synthesis. Our user study also confirms the high expression synthesis quality with 152 out of 224 responses scored the synthesized expressions with 5 points and above in a 7-point Likert scale.

2 METHOD

ExShot leverages ExprNet, a deep model to generate photorealistic expressions on selected head poses, even in the presence of visibility and illumination changes. Our expression synthesizing process involves transforming the head pose and illumination of existing expression observations. This transformation allows our model to avoid the entanglement of learned expression or identity information that can lead to over-fitting, a common issue in many generative models.

(a) Samples from different cameras. **(b)** Sample training triplet image.

Figure 2: Visualization of our professionally captured raw data and pre-processed data. We use the synchronized camera array to capture human subjects with different head poses (a), which we further process to generate the training triplet data, i.e., pose, expression, and ground truth (b).



(a) ExprNet architecture. **(b)** Flow refinement and feature warping (level i=0).

Figure 3: ExprNet architecture. ExprNet is a variant of residual UNet with flow-based warping applied at each feature level. ExprNet takes the input of pose RGB, expression RGB, estimated expression normal map, target face mask, and 3D face mesh generated warping flow field. By jointly training the pipeline, our model overcomes challenges such as head pose differences and lighting variations to generate high-quality facial expression images.

Similar to [Li et al. 2022; Pandey et al. 2021], we generated the training data for our model by using a lab-based professional portrait photo capturing system. This setup includes 5 synchronized cameras placed around the subject to capture multi-view expression images. We recorded subjects displaying 12 diverse facial expressions, covering a range of emotions, and captured the same expressions at different head poses using the synchronized camera array. Figure 2a shows a sample of our synchronized portrait image captures from different camera angles. We preprocess the captured data and generate 7920 instances of training data triplets consisting of target expression, target pose, and ground truth images. Figure 2b shows an example training data.

Figure 3a illustrates the overall model architecture. ExprNet generates facial expressions from input data, namely an input RGB image pair (I_p, I_e), the estimated expression image normal map N_e, the target face mask M_t, and the forward expression warping flow $F_{e\to t}$. We synthesize a target facial expression image I_t as follows:

$$I_t = \text{ExprNet}(I_p, I_e, N_e, M_t, F_{e\to t}) \quad (1)$$

ExprNet aims to combine a desired facial expression from I_e with the preferred head pose from I_p to create I_t. However, the

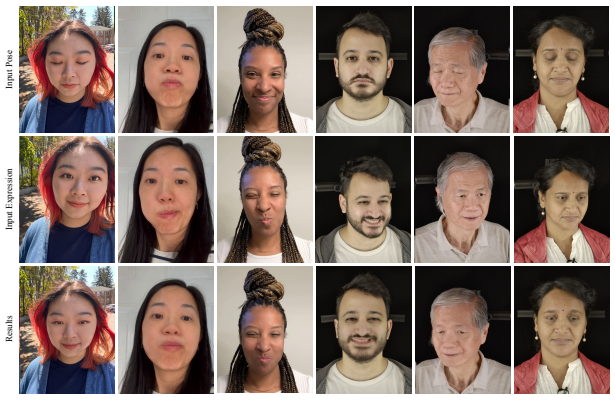


Figure 4: Facial expression editing results by ExShot. We show the synthesized expression results on in-the-wild captured portraits (columns 1 to 3) and our professionally captured data (columns 4 to 6). ExShot can synthesize high-quality facial portraits with photorealistic and sharp details, even for challenging facial details like wrinkles (column 2) and facial hairs (column 4).

process encounters two challenges: differing head poses and illuminations in I_e and I_p, and accurately modelling head pose movement. To overcome these challenges, ExprNet uses two components: a UNet-based network [Ronneberger et al. 2015] for adapting to head pose-induced illumination change, a flow refinement network to improve warping flow field $F_{e\to t}$. Figure 3b shows a details visualization of the flow refinement and expression feature warping process. Finally, ExprNet uses an image synthesis network to generate the final output.

To train the model, we utilize a blend of L1, VGG perceptual, and adversarial losses. These loss functions are applied on both synthesized and warped expression images for complete supervision and direct supervision of the flow refinement module, respectively. To improve ExprNet's generalization ability under various lighting conditions, we introduce an augmented dataset with synthetic lighting. We use 30 real-world HDRI environment lighting maps to apply lighting augmentation using a pre-trained portrait relighting framework [Pandey et al. 2021]. During inference, our system provides an efficient image processing pipeline for facial expression editing consists of three key components including pre-processing, synthesizing and post-processing.

3 EVALUATION

3.1 Evaluation Setup

To quantitatively evaluate our expression synthesizing visual qualities, we create a testing dataset from our professionally captured dataset. We employ several commonly used metrics [LeGendre et al. 2019, 2020; Pandey et al. 2021] to evaluate the quality of synthesized image. These metrics include: *(i)* the mean absolute error (MAE), *(ii)* the mean squared error (MSE), *(iii)* the structural similarity index measure (SSIM), *(iv)* and the Learned Perceptual Image Patch Similarity metric (LPIPS [Zhang et al. 2018]). We choose two recent expression editing methods, StyleHEAT [Yin et al. 2022] and MonoAvatar [Bai et al. 2023], that represent the state-of-the-art

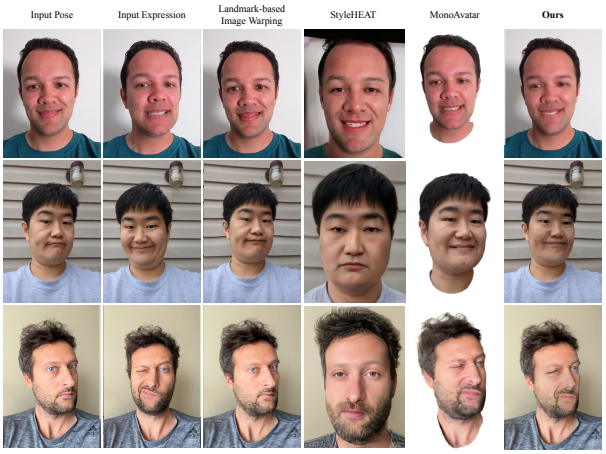


Figure 5: Visual results comparisons. Column 4 and 5 shows the results of StyleHEAT [Yin et al. 2022] and MonoAvatar [Bai et al. 2023], respectively, showcasing the use of general and personalized expression priors. Our method, shown in Column 6, achieves comparable or superior visual results while generating images at higher resolutions. Note that for our method, the shown results do not include changes in edited gaze directions as we intend to showcase the original synthesized visual results of ExprNet.

expression editing with general expression priors and personalized expression priors. We also include a baseline method that uses landmark-based image warping.

3.2 ExShot Result Overview

Figure 4 presents the expression editing results on six subjects. We use both in-the-wild photos (captured by commodity mobile cameras) and professionally captured photos. Despite the challenging editing details, we observe that ExShot can synthesize photorealistic images for all three in-the-wild subjects.

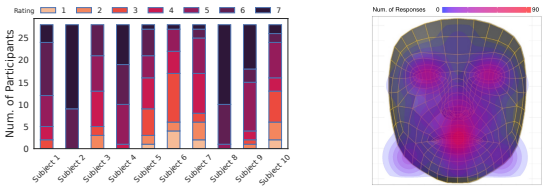
3.3 Comparisons to Other Methods

Figure 5 shows the visual comparisons. Compare to StyleHEAT [Yin et al. 2022] and MonoAvatar [Bai et al. 2023], our method can synthesize better expression results on challenging details, such as facial hairs and wrinkles. We also observed that the editing results generated by StyleHEAT [Yin et al. 2022] were shown in shifted head poses, which is likely due to it's incorrect GAN inversion results. In contrast, our method generates sharp and clear expression with plausible details.

Table 1 reports the quality comparisons between baseline methods and the variation of our models. We see that ExShot outperforms both baselines in all quality metrics. For example, when comparing to StyleHEAT, ExShot achieves higher 8.15 dB PSNR respectively. Our quantitative evaluation results also show that our design on flow refinement, perceptual loss and data augmentation provides important impacts on our models' synthesizing quality.

Table 1: Quality comparisons. We calculate the image quality metric on our testing dataset. Our method outperforms others and achieves the best perceptual quality.

Model	MAE ↓	LPIPS ↓	PSNR(db) ↑	SSIM ↑
Landmark-based warping	0.0171	0.0322	25.15	0.8134
StyleHeat [Yin et al. 2022]	0.0162	0.0299	28.08	0.8331
ExShot (Ours)	**0.0061**	**0.0115**	**36.23**	**0.9598**
ExprNet with RGB warping	0.0110	0.0214	29.83	0.8993
ExprNet w/o flow refinement	0.0091	0.0170	32.21	0.9207
ExprNet w/o perceptual loss	0.0103	0.0185	31.34	0.9052
ExprNet w/o data augmentation	0.0077	0.0136	34.21	0.9412



(a) Quality ratings. **(b)** Reported issues.

Figure 6: Results for the quality assessment study. We analyze the responses from 28 participants and show (a) the distribution of Likert rating scores for 10 subjects; and (b) a visualization of the aggregated face regions of quality issues reported by participants.

3.4 User Study

Our study, constructed as an online survey, recruited 28 participants (18 male and 10 female) from our organization. Participants were required to answer all questions, and each study took about 10 minutes. Participants were presented with 10 groups of questions in the same order, each containing three portrait photos. They were asked to rate the quality of the resulting image on a 7-point Likert scale, provide open feedback, and identify areas for improvement. To put the ratings in context, we included a control group where it's third photo is not synthesized (subject 2). We also included a case with challenging facial geometries which we know ExShot cannot effectively handle (subject 6).

Across the 224 assessment results, excluding two control group subjects (subject 2 and subject 6), ExShot received an average rating of 5.02 (out of 7) from 28 participants. Specifically, we received the highest average rating on subject 8 of 6.61, and the lowest rating on subject 10 of 3.89. In comparison, the average rating on our our control group with real expression (subject 2) is 6.67, less than 1% higher than that on subject 8. Additionally, as shown in Figure 6a, more than 50% participants rated the "result" image of subjects 1, 3, 4, 8, and 9, with 5 points and above. Lastly, the "result" image of subject 6, a known challenging case for ExShot, received an average rating number of 3.32. Other subjects that show similar rating distribution, i.e., the median rating is <= 4, is subject 10. This indicates that in certain cases, ExShot is capable of synthesizing highly realistic results even for human perception, and in most cases, can generate expression editing results that are satisfactory to the majority.

Additionally, our study identified the most problematic face regions reported by the participants are the eyes and mouth regions. Figure 6b presents a generic face overlaid with the reported issues that are grouped by provided regions of the forehead, temporal, eye, cheek, nose, chin, jaw, and mouth. The region heat map is generated based on the number of responses. Red regions indicate more reported issues.

4 CONCLUSION

ExShot represents a step forward in the field of portrait expression editing. Unlike traditional approaches, ExShot can work with a wider range of users and expression types without the need for personalized data capture or retraining. Future work may involve exploring additional ways to enhance face geometry understanding, e.g., focusing on problematic face regions identified in our user study, to further improve the realism and accuracy of synthesized expressions.

ACKNOWLEDGMENTS

This work was supported in part by NSF Grants #2236987.

REFERENCES

Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, et al. 2023. Learning Personalized High Quality Volumetric Head Avatars from Monocular RGB Videos. (2023).

Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. 2021. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14398–14407.

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. 2019. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5918–5928.

Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. 2020. Learning Illumination from Diverse Portraits. In *SIGGRAPH Asia 2020 Technical Communications* (Virtual Event, Republic of Korea) *(SA '20)*. Association for Computing Machinery, New York, NY, USA, Article 7, 4 pages. https://doi.org/10.1145/3410700.3425432

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: A Hybrid Representation for Photorealistic Synthesis, Animation and Relighting of Human Eyes. *ACM Trans. Graph.* 41, 4, Article 166 (jul 2022), 16 pages. https://doi.org/10.1145/3528223.3530130

Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–21.

Srinivas Rao, Rodrigo Ortiz-Cayon, Matteo Munaro, Aidas Liaudanskas, Krunal Chande, Tobias Bertel, Christian Richardt, Alexander J. B., Stefan Holzer, and Abhishek Kar. 2020. Free-Viewpoint Facial Re-Enactment from a Casual Capture. In *SIGGRAPH Asia 2020 Posters* (Virtual Event, Republic of Korea) *(SA '20)*. Association for Computing Machinery, New York, NY, USA, Article 37, 2 pages. https://doi.org/10.1145/3415264.3425453

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 234–241.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. StyleHEAT: One-shot high-resolution editable talking face generation via pre-trained StyleGAN. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 85–101.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.