Smile Like You Mean It: Driving Animatronic Robotic Face with Learned Models

Boyuan Chen* Yuhang Hu Lianfeng Li Sara Cummings Hod Lipson
Columbia University

***Abstract*— Ability to generate intelligent and generalizable facial expressions is essential for building human-like social robots. At present, progress in this field is hindered by the fact that each facial expression needs to be programmed by humans. In order to adapt robot behavior in real time to different situations that arise when interacting with human subjects, robots need to be able to train themselves without requiring human labels, as well as make fast action decisions and generalize the acquired knowledge to diverse and new contexts. We addressed this challenge by designing a physical animatronic robotic face with soft skin and by developing a vision-based self-supervised learning framework for facial mimicry. Our algorithm does not require any knowledge of the robot's kinematic model, camera calibration or predefined expression set. By decomposing the learning process into a generative model and an inverse model, our framework can be trained using a single motor babbling dataset. Comprehensive evaluations show that our method enables accurate and diverse face mimicry across diverse human subjects.**

I. INTRODUCTION

Facial expressions are an essential aspect of nonverbal communication. In our day-to-day lives, we rely on diverse facial expressions to convey our feelings and attitudes to others and interpret other people's emotions, desires and intentions [1]. Facial mimicry [2]–[5] is also recognized as a vital stepping stone towards the early development of social skills for infants. Therefore, building robots that can automatically mimic diverse human facial expressions [6]–[9] will facilitate more natural robotic social behaviors and further encourage stronger engagement in human-robot interactions. Mimicking human facial expressions is also the first step towards achieving adaptive facial reactions in robots. Despite the practical value of such systems, extant research in this domain mostly focuses on the hardware design and pre-programmed facial expressions, allowing robots to select one of the facial expressions from a predefined set. Generalizing across various human expressions has remained challenging.

Current robotic face systems cannot mimic human facial expressions adaptively. The key limitation is the lack of a general learning framework that can learn from limited human supervision. Some traditional methods [10]–[16] define a set of pre-specified facial expressions. Others generalize this process to search for closest match from a database [17] or by following an fitness function [17], [18]. However, as human expressions are highly diverse, these approaches have limited value in practical robot-human interactions.

We thank all the reviewers for their great comments and efforts. This research is supported by NSF NRI 1925157 and DARPA MTO grant HR0011-18-2-0020. *bchen@cs.columbia.edu

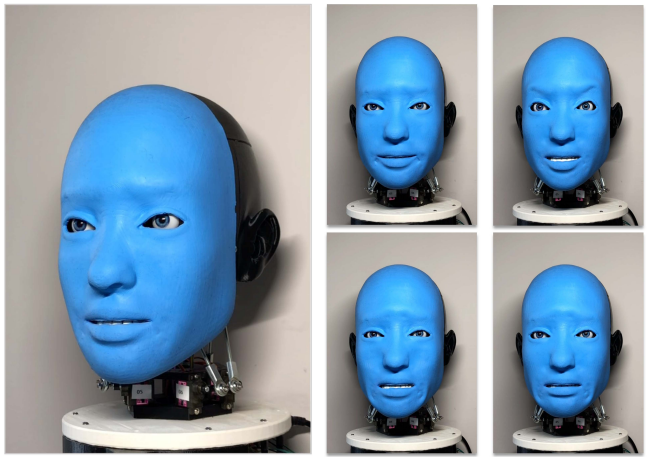

Fig. 1: **Eva 2.0** is a general animatronic robotic face for facial mimicry. The robot does so by learning the correspondence between facial landmarks and self-images as well as a learned inverse kinematic model. The entire learning process relies on the robot's motor babbling in a self-supervised manner. Our robot can mimic varieties of human expressions across many human subjects.

In this work, we present Eva 2.0 (Fig. 1) with significant upgrades to our previous Eva 1.0 [19] platform with more flexible and stable control. We further propose a general learning-based framework to learn facial mimicry from visual observations that can generalize well to different human subjects and diverse expressions. Importantly, our approach does not rely on human supervisions to provide ground-truth robot commands. Our key idea is to decompose the problem into two stages: (1) given normalized human facial landmarks, we first use a generative model to synthesize a corresponding robot self-image with the same facial expression and then (2) leverage an inverse network to output the set of motor commands from the synthesized image.

Our experiments suggest that our approach outperforms previous nearest-neighbor search-based algorithms and direct mapping from human face to action methods. Moreover, quantitative visualizations of our robot imitations demonstrate that, when presented with diverse human subjects, our method generates appropriate and accurate facial expression imitations.

Our primary contributions are threefold. First, we present an animatronic robotic face with soft skin and flexible control mechanisms. Second, we propose a vision-based learning framework for robot facial mimicry that can be trained in a

self-supervised manner. Third, we construct a human facial expression dataset from previous database, YouTube videos and real-world human subjects. Our approach enables strong generalization over 12 human subjects and nearly 400 salient natural expressions. Our robot has a response time within 0.18 s, indicating a strong capability of real-time reaction. Qualitative results and open-sourced platforms are in our supplementary video at https://rb.gy/9koqpg.

II. RELATED WORKS

Animatronic Robotics Face Designing physical animatronic robot face capable of imitating human facial expressions is an important topic in human-robot interaction [20]–[27]. Early works [10]–[16], [28] solely focus on hardware design of the robot face and pre-program the facial expressions. Kismet [20], [29]–[34] generates diverse facial expressions by interpolating among predefined basis facial postures over a three-dimensional space. Albert HUBO [10] imitates specific human body movements, including facial expression. WE-4R [28] determines the expression with emotion vectors. Hashimoto et al. [12] maps the human face configuration to a robot through life-mask worn by a human subject. Asheber et al. [16] proposed a simplified design to enable easier programming and flexible movements. However, all these designs rely on a fixed set of pre-programmed facial expressions that cannot generalize to novel expressions and require extensive human efforts through trail-and-error.

Recent studies start to integrate more general motion control into the hardware designs. Affetto [35]–[37] adopts hysterical sigmoid functions to model motor displacement. However, the landmarks are tracked with OptiTrack sensors on the face. Our method does not require any hardware attachment on the robot face. XIN-REN [38] uses an AAM model to track the facial features for a specific person, which consequently cannot generalize well to novel subjects. They also calculate the robot solution with the robot kinematic equation, while ours learns the model automatically.

Hyung et al. [18] have proposed a genetic algorithms to search for the best facial expression. However, the algorithm is inefficient when the search space grows with the complexity of the robot. Therefore, the approach cannot perform online inference, while our robot can react to a novel facial expression within 0.18s in a fully online manner. Liu and Ren [17] classified the input emotion and search in a discrete predefined expression set which constrains possible outcomes.

Synthetic Video Generation and Animation Significant progresses have been made in synthetic video generation with a focus on motion re-targeting. The goal of video synthesis [39]–[42] is to learn the mapping from a source video to a target video domain to render photorealisitc videos. Our work does not aim to render a realistic video or image. Rather, we aim to map human facial expressions to a physical robot platform.

On the other hand, some works in character face animation [43]–[47] targets at driving a digital face with human motions. Cao et al. [43] learns to recognize human facial landmarks and map them to a 3D avatar. Their method requires full 3D knowledge of the digital face, making it hard to generalize to real robots. Our approach learns a kinematic model of the physical robot through self-supervised motor babbling without any prior knowledge of the control mechanisms.

The most relevant works pertain to motion re-targeting [48]–[51]. Chan et al. [52] proposes to transfer dance motion to a different human subject in a rendered video. Similarly, X2Face [53] and Face2Face [54] learns to synthesize novel face movements videos. However, the aim of these works is realistic video rendering, rather than a physical robot realization. Moreover, most prior works assume a strong correlation of kinematic configurations between the source and target domain. However, as most robotic platforms, our robot does not share same control mechanisms as humans. Therefore, existing approaches cannot be directly applied in real-life contexts.

Imitation Learning Our work also shares the same high-level goal as imitation learning. However, most imitation learning research focus on the manipulation [55]–[59], locomotion [58], [60], [61] or navigation [62]–[64], as discussed in detail in comprehensive review articles [65]–[67]. Our work studies facial mimicry. With our learning framework, we can directly train our models with the robotic face on an offline dataset and generalize on different human face configurations.

III. DESIGN

Our robot design — based on our previous Eva 1.0 [19] with significant upgrades — consists of two sub-assembly modules: facial movement module and neck movement module. We use micro servo motors (MG90S) to actuate all the components of our robotic face. All parts in our design are based on off-the-shelf hardware components that can be easily purchased online or 3D printed. To accelerate research, we open-source the design and step-by-step assembly process on our website. An overview of our hardware design is shown in Fig. 2.

Facial Movement Module Our facial movement module can be further divided into skull frame, eye module, muscle module and jaw module. Compared to Eva 1.0, we redesign the 3D shape of the skull frame to enlarge the maneuvering space and thus allow for more flexible movements. The new skull frame also facilitates tighter connection to the skin with smoother and more natural looks on the face surface.

By adding a pair of ball joints and three pairs of parallelogram mechanisms, we now also ensure that the 6DoF eyeball module can move freely within a $\pm 20^\circ$ range both horizontally and vertically, making the movements more stable. The rotation of each eyeball is controlled by two motors and three ball joint linkages. We also upgraded the eyelid design based on human face.

Our robot skin is attached to the front skull. The muscles driving facial expressions are attached to the inner side of the skin with a piece of fiber fabric, a strand of nylon cord,

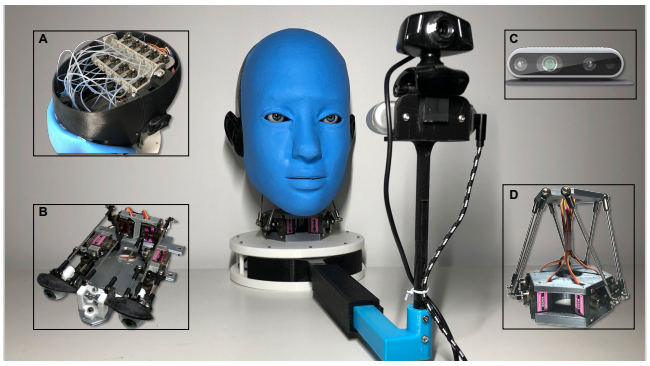

Fig. 2: **Mechanical Design:** our robot is actuated by the motor servo module (A) controlled by a Raspberry Pi 4 located at the bottom. The soft skin is connected to 10 motors via nylon cord. Our 6DoF eye module (B) is decoupled from the front skull. The RGB camera (C) is only used for random data collection of robot self-images but not for testing. The 6DoF neck module (D) follows Steward platform.

and a servo actuator inside the back of the skull. By pulling different nylon cords, a specific skin region can be deformed. We use the same region selection as Eva 1.0. To reduce the noise for muscle control, we run each nylon cord through a transparent vinyl tube linking the skull front and the motor. Our design ensures that the deformation in each muscle is proportional to the motor's rotation angle. Our design enables a large possible space of facial expressions by manipulating 10 pairs of symmetrical muscles in different proportions.

The jaw module is responsible for deforming the two muscles around the mouth. This is similar to the aforementioned skin movement design, but benefits from two additional coupled motors that control the movement, allowing the jaw to open up to 20°.

Neck Movement Module Our 6DoF neck module design is inspired by Stewart platform. Six motors are arranged in 3 pairs evenly distributed on 3 sides of the hexagon base. Each pair comprises two motors in a mirrored arrangement that are connected to a ball joint linkage.

IV. MODELS

We propose a learning-based framework for controlling the animatronic robotic face to mimic varieties of human facial expressions. An overview of our learning algorithms is shown in Fig. 3. We consider the following problem setup: given an image of human face displaying a natural facial expression, the model needs to output the motor commands to actuate the robot to imitate the given facial expression.

Without human pre-programming for different expressions, the problem poses several key challenges and desired properties. First, we hope that the learning algorithm can generalize to diverse unseen human faces. We leveraged the recent advances in human facial landmark detection to obtain abstract representations from high-dimensional image frames (Section IV-A) which can be shared among different human subjects.

The second challenge in attempting to achieve facial mimicry stems from the lack of ground-truth pairs of human expressions and robot motor commands. Without hard-coding and extensive trail-and-error, obtaining such a pair is not practical. In this paper, we overcome this issue by adopting a two stage learning-based method: (1) a generative model which first synthesizes a robot self-image from facial landmarks processed by a proposed normalization algorithm (Section IV-C and Section IV-B), (2) and an inverse model that is trained to produce desired motor commands from the generated robot image (Section IV-D). As we will show, the ground-truth labels for both models can be acquired with one round of self-supervised data collection without any human input.

A. Representation of Facial Expression

We capture facial expressions via facial landmarks, as this has been shown as an effective means of representing the underlying emotions. More importantly, facial landmarks provide a unified abstraction from diverse high-dimensional human images under different lighting, background and poses.

Specifically, we extract facial landmarks with OpenPose [68], [69] software. The output is a vector of 53×3 size representing the spatial position of 53 landmarks on human faces, with the last dimension being confidence scores. We also extracted the head pose for direct neck movement control. As ground-truth pairs for generating motor commands directly from the extracted human landmarks are not available, we adopt a different strategy, as discussed below.

B. Landmark Normalization

Before we send the landmarks from human faces for robot learning, we need to normalize the spatial locations of the landmark vector to the robot domain. This is necessary due to potential variations in the scale or landmark arrangements in the captured human faces. Formally we normalize each landmark coordinate from human space L_H to robot space L_R with:

$$L_R = \frac{(L_H - H_{min})(R_{max} - R_{min})}{H_{max} - H_{min}} + R_{min}$$

where $H_{min}, H_{max}, R_{min}, R_{max}$ represent the value ranges of the spatial location per corresponding landmark in the sampled human and robot image frames.

C. Generative Model

The generative model takes in the normalized human landmarks and generates a synthetic RGB image allowing the robot to conceive the same facial expression, which we denote as robot's self-image. We parametrize the generative model G with a deep neural network with parameter θ.

A key challenge here is to map the coordinate vector to a high-dimensional image. This could be accomplished by encoding the input vectors with a fully-connected network, which is thus used by the robot to learn the entire spatial mapping. However, this simple approach requires a network

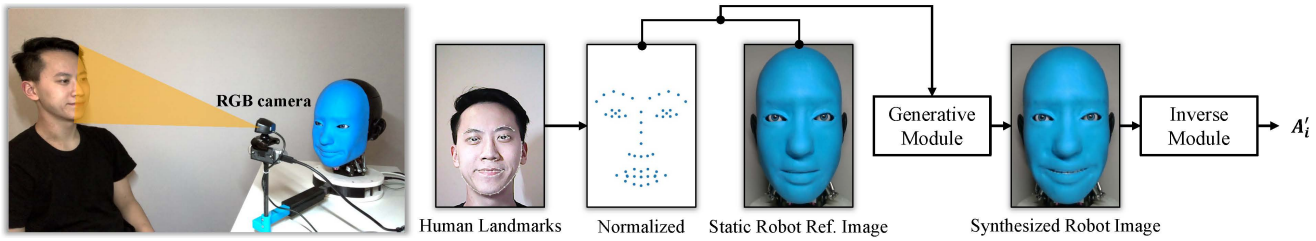

Fig. 3: **Model Overview:** our two-stage framework consists of two major modules: a generative network and an inverse network. Given an image captured by a regular RGB camera, we first extract facial landmarks with OpenPose. We then normalize the human landmarks to the robot scale and embed it on an image. Together with a reference static robot self-image, these two images are concatenated to the generative network to synthesize a robot self-image as if the robot makes the same expression. The inverse model takes the synthetic robot self-image to output the final motor commands for execution.

of strong capacity and optimization algorithm and does not perform well in practice [70].

To this end, we propose to encode the spatial coordinates of the landmarks to a two-channel image mask $\mathbf{M_i^{w\times h\times 2}}$ [52]. The first channel has the value of 1 if there is a landmark at the particular location and 0 otherwise. The second channel is a greyscale image indicating the confidence score returned by the landmark detection algorithm. This encoding matches the size of the robot image, which helps with ensuring correct spatial correspondence.

Furthermore, instead of relying on the network to directly regress the absolute value in the output image, it only needs to output the "change" in the image introduced by landmark displacements. In practice, we achieve this by conditioning the network with a static robot self-image $\mathbf{I_s^{w\times h\times 3}}$ whereby the two images are concatenated along the depth channel. Our generative model can be expressed as: $\mathbf{I_i} \leftarrow G(\mathbf{M_i}, \mathbf{I_s})$.

Implementation Details We use a fully convolutional encoder-decoder architecture [71], [72] where the resolution of the decoder network is enhanced by several hierarchical feature refinement convolutional layers [73]. Since the network is fully convolutional, we can preserve all the spatial information with high-quality outputs. Our network is optimized with a simple pixel-wise mean-squared error loss using Adam [74] optimizer and a learning rate of 0.001. We train the network for 200 epochs with batch size 196 until convergence on a validation dataset. We include more details in the supplementary. The formal objective function that we minimize is:

$$\mathcal{L}_G = \text{MSE}(G(\mathbf{M}, \mathbf{I_s}), \mathbf{I})$$

D. Inverse Model

The inverse model F maps the synthetic robot self-image to motor commands in order to learn an inverse mapping from the goal image to actions.

Action representation Our robot utilizes N motors to actuate the face muscles. Given the robot face image, it is straightforward to frame this problem as continuous value regression. That is, given an input image, the network needs to output N values for N motor encoders. However, this framing can quickly become impractical in complex real-world scenarios and may lead to over-engineering. In practice, we observe that the motors have certain angle threshold to produce salient and stable motion. As a result, we can achieve salient motions with a discrete parameterized angle encoder without loss of accuracy.

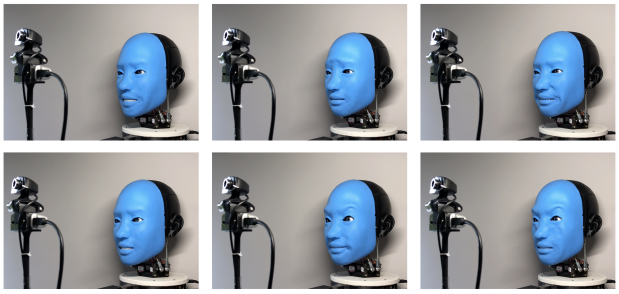

Fig. 4: **Training Data Collection:** the whole training process of our generative model and the inverse model rely on a single robot dataset without human supervision. We collect our training data with random motor babbling in a self-supervised manner whereby the camera facing the robot is used solely for gathering the training data, *i.e.*, it is not used during evaluation

We normalized and discretized the motor values to the $[0, 1]$ range with 0.25 step size, resulting in 5 values per motor. This discretization converts the original problem to a multi-class classification problem. Given an input robot image, the inverse model will output $5 \times N$ numbers, where every five numbers represents the probability of choosing each motor angle for one actuator. Our inverse model can be expressed as: $\mathbf{A_i} \leftarrow F(\mathbf{I_i})$.

Implementation Details Our architecture has 6 convolutional layers followed by several fully-connected layers to adapt the output dimension to be $5 \times N$. We train our network with a multi-class Cross-Entropy loss with Adam optimizer and a learning rate of 0.00005 for 58 epochs. We have $14,000$ pairs for training and $1,000$ pairs for validation and testing respectively. The formal objective function is:

$$\mathcal{L}_F = \Sigma_{n=0}^{N-1}\text{CE}(F_n(\mathbf{I}), \mathbf{A_n})$$

E. Training Data Collection

Both models can be trained separately with the same data collected via self-supervised motor babbling (Fig. 4). We

randomized the angles of the 10 motors ranging from 0 to 1 with an interval of 0.25 for 16,000 steps. We used Intel RealSense D435i to capture RGB images and cropped the image to 480×320 to center the robot head. For each step i, we recorded the motor command values $\mathbf{A_i}$, the corresponding robot images $\mathbf{I_i}$, as well as the extracted landmark $\mathbf{L_{R,i}}$ from the robot with OpenPose. We thus obtain the training pairs of the generative model as ($\mathbf{L_{R,i}}$, $\mathbf{I_i}$) and the training pairs of the inverse model as ($\mathbf{I_i}$, $\mathbf{A_i}$). Since the data collection is purely random, the process does not require any human labeling.

F. Inference

Once the generative model and the inverse model are trained, we can use them jointly to perform inference. Given the input human image captured by an RGB camera, we first extract the landmark coordinates. After our normalization procedure, we can send it to our generative model which then outputs a synthetic robot self-image. This synthetic robot image serves as input for the inverse model which outputs the motor commands. All the network training and testing can be accomplished on a single NVIDIA 1080Ti GPU, whereas the motor commands on the robot are executed via WiFi.

V. EXPERIMENTS

A. Evaluation Dataset

Even though the training of our models do not require any human face dataset, we constructed a human facial dataset with salient facial expressions to perform extensive evaluations. Our human expression dataset is a combination of MMI Facial Expression database [75] and a set of online videos featuring eight subjects. We uniformly down-sampled the original videos to obtain 380 salient frames covering a variety of facial expressions and human subjects.

B. Baselines

We are interested in evaluating the effectiveness of our approach for the generative model (GM), the inverse model (IM), and finally the entire pipeline. To this end, we design our baseline and ablation studies for each of them. For the generative model, we compared our method with a randomly sampled (RS) image from a collection of real images. This comparison aims to ascertain whether the synthetic image can outperform the real image and whether our model can predict reasonable motion on the static robot template.

Even though our method successfully convert the inverse model to a classification problem, a purely random baseline only has a 20% success rate. We also compared our model with another two less random simple baselines — a randomly initialized network (RI) and a model trained for about 100 iterations (RI-100).

For the entire pipeline, we evaluated different combinations of the above baselines. For example, we obtained one baseline by combining our generative model with an inverse model that has been trained for 100 iterations. Additionally, we present another three baselines. The first one is to perform a nearest neighbor retrieval (NN) with the landmarks extracted from the output of our generative model. We can directly search within our robot dataset. This baseline replaces our inverse model with a NN model. Similarly, we can perform such NN operation with human landmarks as a direct input. Lastly, we assessed whether we can directly generate the motor commands without our two-stage algorithm by training a network to output motor commands from our embedded input landmarks.

We used the same architecture for all the above baselines while varying only the number of channels in the first layer for different input dimensions. We trained all the models with three random seeds and report the mean and standard errors.

C. Evaluation Metrics

Our evaluation metrics cover both the pixel-wise accuracy of synthetic image and the accuracy of the output commands. We also extracted landmarks from our synthetic image to measure if the model successfully learns the facial expression than simply copying the static image. Moreover, we provide qualitative visualizations for the final pipeline. We also extracted landmarks from our entire pipeline execution to compare against the input human landmarks. We use L2 metric for both the image and landmark distances.

D. Results

Generative Model Tab. I shows the quantitative evaluation for our generative model. Our generative model outperforms the random baseline by a large margin. Note that the image distance is normalized by the total number of pixel values ($480 \times 320 \times 3$) which range from 0 and 1, whereas the landmark distance is normalized by the total number of landmarks (53).

TABLE I: Accuracy of the Generative Model

Method	Image Distance ↓ ($\times 10^{-5}$)	Landmark Distance ↓
GM (ours)	$\mathbf{3.47 \pm 0.009}$	$\mathbf{0.46 \pm 0.002}$
RS	6.47 ± 0.039	0.84 ± 0.007

Inverse Model Tab. II shows the evaluation result for our inverse model. Compared with the random initialized model, the model trained for 100 iterations as well as the purely random baselines (20% accuracy), our inverse model produces much more accurate motor commands.

TABLE II: Accuracy of the Inverse Model

Method	Command Distance ↓	Command Accuracy ↑ (%)
IM (ours)	$\mathbf{0.53 \pm 0.009}$	$\mathbf{75.86 \pm 0.042}$
RI	1.39 ± 0.009	30.34 ± 0.038
RI-100	0.90 ± 0.010	56.40 ± 0.041

Two-Stage Pipeline By combining the generative and inverse model, we can use the synthetic output from the generative model as the input for the inverse model, allowing us to evaluate the entire two-stage pipeline. As shown in Tab. III, our two-stage strategy outperforms all baselines, but is

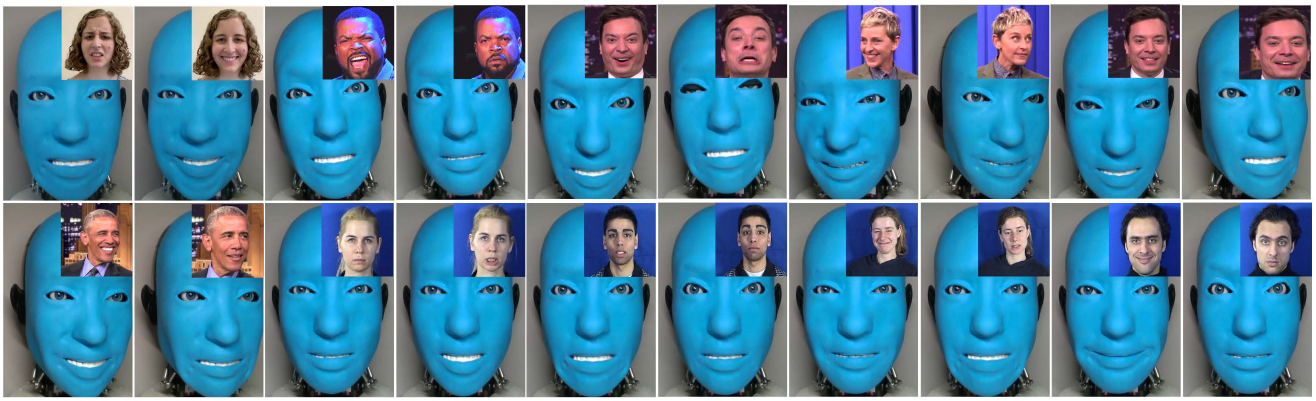

Fig. 5: **Robot Visualizations:** we executed the output motor commands on our physical robot to demonstrate that our method supports accurate facial mimicry of a variety of human expressions across multiple human subjects.

inferior to the performance of its component models. This is to be expected, as the inverse model takes the synthetic output from the generative model as its input when these are evaluated jointly, whereas individual evaluations are based on real images from our robot dataset. Nonetheless, our method still achieves the best predictions and is particularly advantageous compared to two single model baselines, indicating that decomposing the problem into two sub-steps is a prudent design choice.

TABLE III: Accuracy of the Entire Two-Stage Pipeline

Method		Command Distance ↓	Command Accuracy ↑ (%)
GM	**IM**	**0.83 ±0.008**	**54.57 ±0.048**
GM	NN	1.06 ±0.009	45.74 ±0.045
GM	RI	1.39 ±0.009	30.34 ±0.039
GM	RI-100	0.98 ±0.009	53.48 ±0.041
Landmark to Motor		1.03 ±0.011	52.81 ±0.048
Landmark NN		0.98 ±0.095	49.3 ±0.045

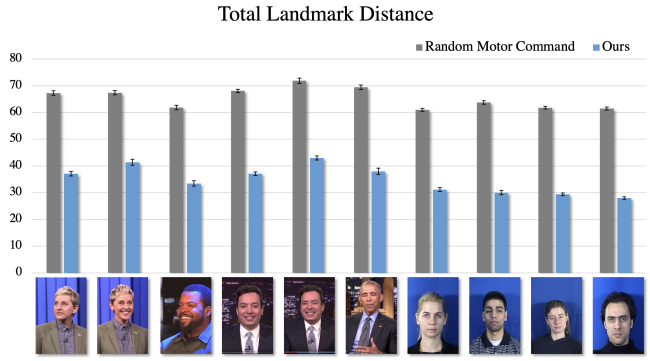

Fig. 6: **Pipeline Execution:** we executed the motor commands output by our entire two-stage pipeline and extracted the landmarks from the resulting physical robot face to compare against the ground-truth human landmarks as well as a random baseline to benchmark the task difficulty. The results show that our robot can imitate different human expressions accurately.

Pipeline Execution We executed the output motor commands from the above two-stage pipeline on our physical robot and computed the landmark distance extracted from the resulted robot face with the ground-truth normalized landmarks. To provide qualitative evaluations, we visualized the final robot face together with the input human face in Fig. 5.

Our evaluation was conducted on 380 salient frames which covers 10 video clips from 8 subjects. We show the quantitative results compared with a random baseline in Fig. 6. Our approach demonstrates a flexible learning-based framework to mimic human facial expression. Our method also generalizes across various human subjects without any human supervision.

VI. CONCLUSIONS AND FUTURE WORK

We present a new animatronic robot face design with soft skin and visual perception system. We also introduce a two-stage self-supervised learning framework for general face mimicry. Our experiments demonstrate that the two-stage algorithm improves the accuracy and diversity of imitating human facial expressions under various conditions. Our method enables real-time planning and opens new opportunities for practical applications.

Although we show that robots based on our design are capable of mimicking diverse human facial expressions from visual observations, it is necessary to study this problem with other modalities. For example, when several humans are involved in a conversation, the context will also influence their facial expressions. Thus, for future investigations, it would be beneficial to incorporate speech signal into the decision-making process while also allowing robots to reciprocate verbally.

Furthermore, even though imitation is an important step towards imbuing robots with more complex skills, being able to generate appropriate reaction expressions will be essential for interactive social robots. More broadly, building interactive robot face will require higher-level understanding of other's emotion, desires and intentions. Hence, an interesting direction is to explore higher-order thinking for robots.

REFERENCES

[1] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[2] N. Reissland, "Neonatal imitation in the first hour of life: Observations in rural nepal." *Developmental Psychology*, vol. 24, no. 4, p. 464, 1988.

[3] A. N. Meltzoff and M. K. Moore, "Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms." *Developmental psychology*, vol. 25, no. 6, p. 954, 1989.

[4] R. Van Baaren, L. Janssen, T. L. Chartrand, and A. Dijksterhuis, "Where is the love? the social aspects of mimicry," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1528, pp. 2381–2389, 2009.

[5] J. Piaget, *Play, dreams and imitation in childhood*. Routledge, 2013, vol. 25.

[6] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[7] M. Blow, K. Dautenhahn, A. Appleby, C. L. Nehaniv, and D. Lee, "The art of designing robot faces: Dimensions for human-robot interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 331–332.

[8] C. Breazeal, A. Takanishi, and T. Kobayashi, "Social robots that interact with people." 2008.

[9] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social human–robot interaction," *International Journal of Social Robotics*, vol. 11, no. 4, pp. 575–608, 2019.

[10] J.-H. Oh, D. Hanson, W.-S. Kim, Y. Han, J.-Y. Kim, and I.-W. Park, "Design of android type humanoid robot albert hubo," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 1428–1433.

[11] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, "Development of the face robot saya for rich facial expressions," in *2006 SICE-ICASE International Joint Conference*. IEEE, 2006, pp. 5423–5428.

[12] T. Hashimoto, S. Hiramatsu, and H. Kobayashi, "Dynamic display of facial expressions on the face robot made by using a life mask," in *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2008, pp. 521–526.

[13] H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, and H. Lee, "Designing of android head system by applying facial muscle mechanism of humans," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. IEEE, 2012, pp. 799–804.

[14] D. Loza, S. Marcos Pablos, E. Zalama Casanova, J. Gómez García-Bermejo, and J. L. González, "Application of the facs in the design and construction of a mechatronic head with realistic appearance," 2013.

[15] C.-Y. Lin, C.-C. Huang, and L.-C. Cheng, "An expressional simplified mechanism in anthropomorphic face robot design," *Robotica*, vol. 34, no. 3, p. 652, 2016.

[16] W. T. Asheber, C.-Y. Lin, and S. H. Yen, "Humanoid head face mechanism with expandable facial expressions," *International Journal of Advanced Robotic Systems*, vol. 13, no. 1, p. 29, 2016.

[17] N. Liu and F. Ren, "Emotion classification using a cnn_lstm-based model for smooth emotional synchronization of the humanoid robot ren-xin," *PloS one*, vol. 14, no. 5, p. e0215216, 2019.

[18] H.-J. Hyung, H. U. Yoon, D. Choi, D.-Y. Lee, and D.-W. Lee, "Optimizing android facial expressions using genetic algorithms," *Applied Sciences*, vol. 9, no. 16, p. 3379, 2019.

[19] Z. Faraj, M. Selamet, C. Morales, P. Torres, M. Hossain, B. Chen, and H. Lipson, "Facially expressive humanoid robotic face," *HardwareX*, vol. 9, p. e00117, 2021.

[20] C. Breazeal, "Emotion and sociable humanoid robots," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[21] M. A. Goodrich and A. C. Schultz, *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[22] H. Yan, M. H. Ang, and A. N. Poo, "A survey on perception methods for human–robot interaction in social robots," *International Journal of Social Robotics*, vol. 6, no. 1, pp. 85–119, 2014.

[23] A. Kalegina, G. Schroeder, A. Allchin, K. Berlin, and M. Cakmak, "Characterizing the design space of rendered robot faces," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 96–104.

[24] V. Dimitrievska and N. Ackovska, "Behavior models of emotion-featured robots: A survey," *Journal of Intelligent & Robotic Systems*, pp. 1–23, 2020.

[25] M. Song, D. Tao, Z. Liu, X. Li, and M. Zhou, "Image ratio features for facial expression recognition application," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 3, pp. 779–788, 2009.

[26] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao, "A facial expression emotion recognition based human-robot interaction system," 2017.

[27] J. Gu, H. Hu, and H. Li, "Local robust sparse representation for face recognition with single sample per person," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 2, pp. 547–554, 2017.

[28] K. Itoh, H. Miwa, M. Zecca, H. Takanobu, S. Roccella, M. C. Carrozza, P. Dario, and A. Takanishi, "Mechanical design of emotion expression humanoid robot we-4rii," in *Romansy 16*. Springer, 2006, pp. 255–262.

[29] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The cog project: Building a humanoid robot," in *International Workshop on Computation for Metaphors, Analogy, and Agents*. Springer, 1998, pp. 52–87.

[30] C. Breazeal and J. Velásquez, "Toward teaching a robot 'infant'using emotive communication acts," in *Proceedings of the 1998 Simulated Adaptive Behavior Workshop on Socially Situated Intelligence*. Citeseer, 1998, pp. 25–40.

[31] B. Scassellati, "Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot," in *International Workshop on Computation for Metaphors, Analogy, and Agents*. Springer, 1998, pp. 176–195.

[32] C. Breazeal and B. Scassellati, "Infant-like social interactions between a robot and a human caregiver," *Adaptive Behavior*, vol. 8, no. 1, pp. 49–74, 2000.

[33] C. Breazeal, "Regulation and entrainment in human—robot interaction," *The International Journal of Robotics Research*, vol. 21, no. 10-11, pp. 883–902, 2002.

[34] C. L. Breazeal, *Designing sociable robots*. MIT press, 2004.

[35] H. Ishihara, Y. Yoshikawa, and M. Asada, "Realistic child robot "affetto" for understanding the caregiver-child attachment relationship that guides the child development," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2. IEEE, 2011, pp. 1–5.

[36] H. Ishihara and M. Asada, "Design of 22-dof pneumatically actuated upper body for child android 'affetto'," *Advanced Robotics*, vol. 29, no. 18, pp. 1151–1163, 2015.

[37] H. Ishihara, B. Wu, and M. Asada, "Identification and evaluation of the face system of a child android robot affetto for surface motion design," *Frontiers in Robotics and AI*, vol. 5, p. 119, 2018.

[38] F. Ren and Z. Huang, "Automatic facial expression learning method based on humanoid robot xin-ren," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 6, pp. 810–821, 2016.

[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[40] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," *arXiv preprint arXiv:1910.12713*, 2019.

[41] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2377–2386.

[42] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4570–4580.

[43] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on graphics (TOG)*, vol. 33, no. 4, pp. 1–10, 2014.

[44] H. Bülthoff, C. Wallraven, and M. A. Giese, "Perceptual robotics," in *Springer Handbook of Robotics*. Springer, 2016, pp. 2095–2114.

[45] M. Seymour, K. Riemer, and J. Kay, "Interactive realistic digital avatars-revisiting the uncanny valley," 2017.

[46] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, and H. Li, "pagan: real-time avatars using dynamic textures," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–12, 2018.

[47] S.-E. Wei, J. Saragih, T. Simon, A. W. Harley, S. Lombardi, M. Perdoch, A. Hypes, D. Wang, H. Badino, and Y. Sheikh, "Vr facial animation via multiview image translation," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–16, 2019.

[48] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment." *ACM Trans. Graph.*, vol. 34, no. 6, pp. 183–1, 2015.

[49] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. Berg, "Dance dance generation: Motion transfer for internet videos," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[50] O. Gafni, L. Wolf, and Y. Taigman, "Vid2game: Controllable characters extracted from real-world videos," *arXiv preprint arXiv:1904.08379*, 2019.

[51] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems*, 2019, pp. 7137–7147.

[52] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.

[53] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.

[54] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[55] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann, "Imitation learning of dual-arm manipulation tasks in humanoid robots," *International Journal of Humanoid Robotics*, vol. 5, no. 02, pp. 183–202, 2008.

[56] L. Rozo, P. Jiménez, and C. Torras, "A robot learning from demonstration framework to perform force-based manipulation tasks," *Intelligent service robotics*, vol. 6, no. 1, pp. 33–51, 2013.

[57] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[58] N. Ratliff, J. A. Bagnell, and S. S. Srinivasa, "Imitation learning for locomotion and manipulation," in *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2007, pp. 392–397.

[59] M. Mühlig, M. Gienger, and J. J. Steil, "Interactive imitation learning of object movement skills," *Autonomous Robots*, vol. 32, no. 2, pp. 97–114, 2012.

[60] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato, "Learning from demonstration and adaptation of biped locomotion," *Robotics and autonomous systems*, vol. 47, no. 2-3, pp. 79–91, 2004.

[61] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, "Learning agile robotic locomotion skills by imitating animals," in *Robotics: Science and Systems*, 07 2020.

[62] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4693–4700.

[63] F. Codevilla, M. Müller, A. Dosovitskiy, A. M. López, and V. Koltun, "End-to-end driving via conditional imitation learning," *CoRR*, vol. abs/1710.02410, 2017. [Online]. Available: http://arxiv.org/abs/1710.02410

[64] Y. Pan, C. Cheng, K. Saigol, K. Lee, X. Yan, E. A. Theodorou, and B. Boots, "Agile off-road autonomous driving using end-to-end deep imitation learning," *CoRR*, vol. abs/1709.07174, 2017. [Online]. Available: http://arxiv.org/abs/1709.07174

[65] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[66] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *arXiv preprint arXiv:1811.06711*, 2018.

[67] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Comput. Surv.*, vol. 50, no. 2, Apr. 2017. [Online]. Available: https://doi.org/10.1145/3054912

[68] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[69] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[70] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Advances in Neural Information Processing Systems*, 2018, pp. 9605–9616.

[71] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[72] B. Chen, C. Vondrick, and H. Lipson, "Visual behavior modelling for robotic theory of mind," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[73] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[75] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.