Pixel-based Facial Expression Synthesis

Arbish Akram and Nazar Khan
Punjab University College of Information Technology
Lahore, Pakistan
Email: {arbishakram, nazarkhan}@pucit.edu.pk

***Abstract*—Facial expression synthesis has achieved remarkable advances with the advent of Generative Adversarial Networks (GANs). However, GAN-based approaches mostly generate photo-realistic results as long as the testing data distribution is close to the training data distribution. The quality of GAN results significantly degrades when testing images are from a slightly different distribution. Moreover, recent work has shown that facial expressions can be synthesized by changing localized face regions. In this work, we propose a pixel-based facial expression synthesis method in which each output pixel observes only one input pixel. The proposed method achieves good generalization capability by leveraging only a few hundred training images. Experimental results demonstrate that the proposed method performs comparably well against state-of-the-art GANs on in-dataset images and significantly better on out-of-dataset images. In addition, the proposed model is two orders of magnitude smaller which makes it suitable for deployment on resource-constrained devices.**

***Index Terms*—Facial expression synthesis, Image-to-image translation, Generative Adversarial Network, Pixel-based, Regression, Kernel Regression.**

I. Introduction

Facial expression synthesis (FES) has received considerable attention in recent years. FES aims to generate identity-preserving faces conditioned on a specific facial expression. It is a useful task for the animation of characters and avatars in video games and animated movies. It can be used to control interactive virtual assistants or to help in video surveillance across expressions. While humans can easily visualize/synthesize a person's face with a different facial expression, this task is captivating and challenging for computers.

Early methods for FES extract facial geometric features [1] or 3D meshes [2] from images to transfer facial expressions on new faces or animated avatars. Traditional statistical learning-based methods including Active Appearance models [3], Active Shape models [4], 3D warping-based methods [5], constrained local methods [6] and bilinear kernel-based methods [7] are also used for synthesizing realistic facial expressions. Since this synthesis task learns the mapping between high-dimensional input and output spaces, models such as regression or neural networks require a lot of data to learn the optimal parameters to approximate this mapping. The number of parameters of these models grows exponentially as the image size increases. Therefore, image synthesis in general and FES in particular, has received less attention as compared to recognition based tasks.

Recently, significant progress has been achieved in image synthesis related areas with the emergence of Generative Adversarial Networks (GANs) [8], [9]. GANs have proven effective for many image synthesis related tasks including image generation [8], [10], [11], [12], [13], [14], image-to-image translation [15], [16], image super-resolution [17], [18], text-based image synthesis [19] and many others [20], [21], [22]. GANs have also obtained remarkable progress in face-related tasks, such as facial attributes editing [11], [23], face aging [24] and facial expression synthesis [25], [26]. These GANs generate better results on in-dataset images as compared to out-of-dataset images. By in-dataset, we mean that training and testing images are taken from the same dataset that follows a certain distribution. In contrast, out-of-dataset images are unseen images that follow a significantly different distribution from the training distribution.

While capable of synthesizing photo-realistic facial expressions, existing GAN-based models are limited in two key aspects. First, these GAN-based methods generate photo-realistic results as long as the testing data distribution is similar to training data distribution or in-dataset images. Even the most successful expression synthesis methods such as StarGAN [25] and GANimation [26] produce artifacts on out-of-dataset images of humans, paintings and animal faces as can be seen in Figure 1. It seems GANs have been inherently biased toward the training data distribution, since they are designed to replicate the training data distribution. Second, these GAN-based methods require thousands of images during training. The quality of results generated by GANs significantly degrades when trained on smaller datasets [27]. So, they do not support generalization with a few hundred images. Finally, GANs require higher computational and storage resources at testing time. Due to these constraints, it becomes challenging to deploy these models on resource-constrained devices and embedded systems which often have limited memory and processing power [28].

In an attempt to address these limitations, we propose an alternative pixel-based approach to solve the facial expression synthesis problem. The key insight is derived from the work of Khan et al. [27] who showed that facial expressions mostly constitute localized changes of the input image. Motivated by this fact, our proposed method considers only one input pixel to produce an output pixel. This is in contrast to traditional linear regression or neural network based models that look at all input image pixels. The difference between traditional methods and our proposed method is illustrated in

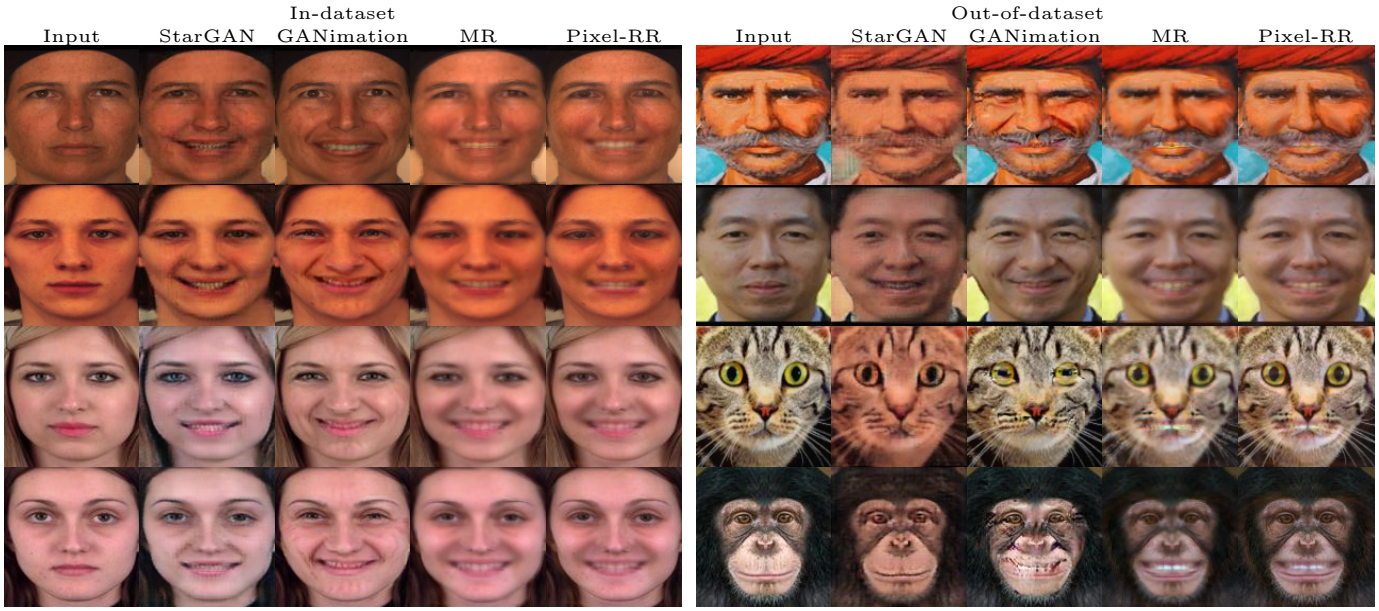

Fig. 1. Comparison of Pixel-based ridge regression (Pixel-RR) with state-of-the-art facial expression synthesis models, MR, StarGAN and GANimation. As shown in this figure, StarGAN is unable to induce plausible happy expression on in-dataset and out-of-dataset images (column 2 and 7, rows 1-4). GANimation successfully induces happy expression on in-dataset and out-of-dataset human faces. However, it also introduces noticeable artifacts around the eyes and mouth region (column 3, rows 2, 3 and 4; column 8, row 1). MR alleviates artifacts and generates plausible happy expression, but the synthesized images suffer from apparent blurring effects. Pixel-RR effectively transforms input expressions into the target expressions while preserving a person's identity and generates sharper images among all methods. Also, Pixel-RR preserves subtle details of the face. StarGAN and GANimation were not able to induce happy expression on animal faces while MR and our proposed method generalize well even on animal faces. Please zoom in for details.

Figure 2. This pixel-based mapping significantly decreases the computational and space complexity of the proposed model by reducing the number of parameters. To capture complex, non-linear characteristics of facial expression mappings, we also show how kernel regression [29] can be exploited. At the cost of increased model size, this step can help to achieve sharper and more plausible synthesis of expressions. Linear regression or vanilla neural-network based methods contain millions of parameters due to being fully connected. To avoid overfitting in such a network large datasets are required for training. However, for FES large paired datasets are not readily available. One solution of this problem is to minimize the number of parameters by employing sparsity. Sparsity ensures that each output pixel observes only a few input pixels. Sparse models contain fewer parameters, generalize well and can be trained using smaller datasets.

Facial expressions usually change localized regions of an image. This idea is exploited for masked regression (MR) [27], that employs local receptive fields to observe a local patch of the input image in order to obtain one pixel of the synthesized image. We extend their local receptive field idea to achieve maximal sparsity and ultimate locality. We contend that there is no need to look at all pixels of a local input patch. We call this approach *pixel-based regression* which looks at only one fixed pixel of the input image to produce a pixel of the output image. The idea of looking at only one input-pixel enforces maximum sparsity as well as locality. Our method takes input and target image pairs and learns to transform the input expression into the target expression by employing pixel-based ridge regression (Pixel-RR). Then we also show that this idea can be extended to capture non-linear characteristics of facial images by using pixel-based kernel regression (Pixel-KR). We show that our proposed maximally localized and sparse method can successfully synthesize facial expressions and improve generalization performance. Furthermore, we formulate a convex objective function that achieves a global minimum and demonstrates superior performance as compared to deep generative models. Moreover, it incurs a very low computational cost and can be easily embedded in mobile devices. To the best of our knowledge, this is the lightest facial expression synthesis model that can be deployed in mobile devices and embedded systems. Table I shows that our proposed model (Pixel-RR) contains two orders of magnitude fewer parameters compared to state-of-the-art GANs.

The main contributions of this work can be summarized as follows.

- We have introduced the first pixel-based ridge regression method to solve the facial expression synthesis problem and have presented a kernel-based extension as well.
- Compared to state-of-the-art GAN-based models, the proposed method generalizes much better for a variety of out-of-dataset images.
- The proposed model is two orders of magnitude smaller than GAN-based models that makes it suitable for mobile devices and embedded systems.

TABLE I

COMPARISON OF DIFFERENT FES MODELS SIZES. STARGAN AND GANIMATION CANNOT BE UTILIZED IN RESOURCE-CONSTRAINED DEVICES DUE TO A LARGE NUMBER OF PARAMETERS. THE PROPOSED PIXEL-RR MODEL IS TWO ORDERS OF MAGNITUDE SMALLER THAN DEEP GANS. IT IS ALSO SMALLER THAN THE RECENT LINEAR REGRESSION BASED MR MODEL WHILE SYNTHESIZING SHARPER FACIAL DETAILS (SEE FIGURE 4).

Parameters	StarGAN [25]	GANimation [26]	MR [27]	Pixel-KR	Pixel-RR
x10^4	850	850	16.2	655	**3.28**

II. RELATED WORK

In the past few years, GANs have emerged as a powerful class of generative models for image-to-image translation problem. GANs exhibit an astonishing capability to generate photo-realistic and sharp images as opposed to blurry images generated via other traditional loss functions such as mean-squared error. GAN [8] consists of two neural networks that compete against each other to learn a minimax objective function. The first network, called generator (G), tries to learn the data distribution while the other network, called discriminator (D), tries to discriminate between real (coming from real data) and fake (coming from generator G) images. GAN extension to generate images on some condition called conditional Generative Adversarial Network (cGAN) [9]. Due to the development of cGAN, image-to-image translation has achieved compelling results [16], [30], [25], [14]. The Pix2pix [16] framework uses cGAN and ℓ_1 loss function to learn the mapping between paired input and output images. For unpaired image-to-image translation, cycleGAN and other frameworks [31], [32] have been proposed to learn the mapping in an unsupervised way. Conditional GANs have also been exploited to synthesize photo-realistic facial expressions. The most successful expression synthesis networks proposed in recent years are StarGAN [25] and GANimation [26]. StarGAN, a multi-domain image-to-image translation network, has been proposed for facial expression synthesis among multiple domains using discrete labels. GANimation [26] is a facial action transfer method that employs GAN to synthesize facial expressions conditioned on action units. Although these GANs generate photo-realistic images, they also employ deep networks that entail large computational and storage overheads. Moreover, the performance of these GAN-based methods is not satisfactory when trained on small datasets. Recently, Khan et al. [27] proposed a linear regression-based method that addresses the above defined limitations of GANs. They observed that expressions mostly constitute local changes. With this motivation, they introduced masked regression to consider only local and sparse regions of images. However, the results of masked regression suffer from blurring effects. We extend their idea to learn maximally localized and maximally sparse receptive fields. This maximally localized and sparse receptive field idea reduces blurriness and enhances facial local details of the image. To improve the quality of results, we also employed kernel regression to capture non-linear characteristics of the face images.

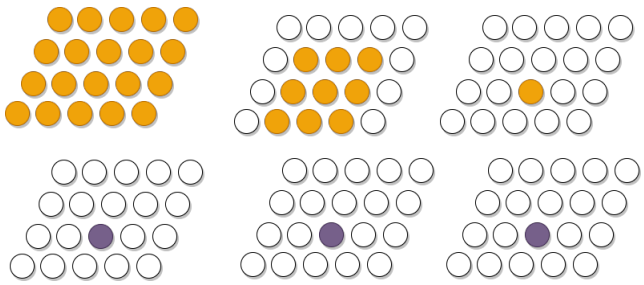

Fig. 2. (Left to Right) Comparison between regression, masked regression and our proposed, pixel regression. **Left:** One output pixel looks at all input pixels (Regression). **Middle:** Recently proposed MR [27] considers a local set of input pixels to make one output pixel (Masked Regression). The selection of pixels is fixed. **Right:** Pixel Regression observes only one input pixel (Ours).

III. PIXEL-BASED REGRESSION

The FES problem can be modeled as ridge regression (RR) whereby output faces are compared with target faces. Let $X \in \mathcal{R}^{N\times D}$ and $T \in \mathcal{R}^{N\times K}$ be the design and response matrices of N input faces (vectorized in $\mathcal{R}^D$) and target faces (vectorized in $\mathcal{R}^K$) respectively. The objective function to learn this mapping via ridge regression can be written as

$$E(W) = \frac{1}{2}\|WX^T - T^T\|_F^2 + \frac{\lambda}{2}\|W\|_F^2 \qquad (1)$$

where $W \in \mathcal{R}^{K\times D}$ is a transformation weight matrix and $\lambda > 0$ is a regularization hyper-parameter that controls overfitting. The unique global minimizer for Eq. 1 can be computed analytically as

$$W = \left[(X^TX + \lambda I)^{-1}X^TT\right]^T \qquad (2)$$

Facial expressions usually constitute local instead of global changes in the input image. For example, the transformation from neutral to happy mostly affects the regions around the eyes, nose and mouth to induce happy expression. It is therefore unnecessary to observe all input pixels while synthesizing a single output pixel. It should be sufficient to observe a relevant portion of the input image. We propose to observe only one input pixel for producing an output pixel. Let the input and target image pixels located at position p in the n-th training pair be denoted as $x_p^n \in \mathcal{R}$ and $t_p^n \in \mathcal{R}$, respectively. Taking into account the pixel position, the entire training set can be divided into K groups $\{\mathbf{x}_p, \mathbf{t}_p\}_{p=1}^K$, where $\mathbf{x}_p = [x_p^1, x_p^2, \ldots, x_p^N] \in \mathcal{R}^{1\times N}$ and $\mathbf{t}_p = [t_p^1, t_p^2, \ldots, t_p^N] \in \mathcal{R}^{1\times N}$ represent the corresponding input and target training pixel values at position p for all N training pairs. The objective function for pixel-based ridge regression (Pixel-RR) can be written as,

$$E(w_p, b_p) = \frac{1}{2}\|w_p\mathbf{x}_p + b_p\mathbf{1} - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}(w_p^2 + b_p^2) \quad (3)$$

where scalars w_p and b_p are learnable weight and bias values that transform input pixels at position p into output pixels and $\mathbf{1}$ is a vector of ones of size $1 \times N$. . The partial derivatives of Eq. 3 can be written as,

$$\frac{\partial E(w_p, b_p)}{\partial w_p} = (w_p\mathbf{x}_p + b_p\mathbf{1} - \mathbf{t}_p)\,\mathbf{x}_p^T + \lambda w_p = 0 \quad (4)$$

$$\frac{\partial E(w_p, b_p)}{\partial b_p} = (w_p\mathbf{x}_p + b_p\mathbf{1} - \mathbf{t}_p)\,\mathbf{1}^T + \lambda b_p = 0 \quad (5)$$

from which the unique global minimizers for Eq. 3 can be computed analytically as

$$\begin{bmatrix} w_p \\ b_p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_p\mathbf{x}_p^T + \lambda & \mathbf{1}\mathbf{x}_p^T \\ \mathbf{1}\mathbf{x}_p^T & N + \lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{t}_p\mathbf{x}_p^T \\ \mathbf{t}_p\mathbf{1}^T \end{bmatrix} \quad (6)$$

Since linear regression cannot effectively capture complex relationships between inputs and outputs. The proposed pixel-based mapping idea can be extended by using kernel regression. We can project input pixels $\mathbf{x}_p$ into a higher-dimensional Reduced Kernel Hilbert Space (RKHS) as

$$\phi(\mathbf{x}_p) = \begin{bmatrix} \phi(x_p^1) & \phi(x_p^2) & \dots & \phi(x_p^N) \end{bmatrix} \in \mathcal{R}^{d \times N} \quad (7)$$

The objective function pixel-based regression from this new RKHS to target values can be written as

$$E(\mathbf{w}_p^\phi) = \frac{1}{2}\|\mathbf{w}_p^\phi\phi(\mathbf{x}_p) - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}_p^\phi\|^2 \quad (8)$$

where $\mathbf{w}_p^\phi \in \mathcal{R}^{1 \times d}$ represents a learnable mapping from the RKHS to targets.

According to the representer theorem, the kernel regression weights $\mathbf{w}_p^\phi$ can be represented by a finite linear combination of the input vector $\phi(\mathbf{x}_p)$ as

$$\mathbf{w}_p^\phi = \mathbf{c}_p\phi(\mathbf{x}_p)^T \quad (9)$$

where $\mathbf{c}_p \in \mathcal{R}^{1 \times N}$ is the projection vector containing the coefficients of the linear combination. Thus, the optimization in Eq. 8 can be reformulated to compute the projection matrix $\mathbf{c}_p$. The objective function for pixel-based kernel regression (Pixel-KR) can be written in terms of $\mathbf{c}_p$ as

$$E(\mathbf{c}_p) = \frac{1}{2}\|\mathbf{c}_p\phi(\mathbf{x}_p)^T\phi(\mathbf{x}_p) - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}_p\phi(\mathbf{x}_p)^T\|_2^2 \quad (10)$$

$$= \frac{1}{2}\|\mathbf{c}_pK_p - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}\mathbf{c}_pK_p\mathbf{c}_p^T \quad (11)$$

where $K_p = \phi(\mathbf{x}_p)^T\phi(\mathbf{x}_p) \in \mathcal{R}^{N \times N}$ is the kernel matrix that computes the non-linear similarity between x_p^i and x_p^j via the kernel function $k(x_p^i, x_p^j) = \phi(x_p^i)^T\phi(x_p^j)$ in the RKHS. We have used the Guassian kernel

$$k(x_p^i, x_p^j) = \exp\left(\frac{-(x_p^i - x_p^j)^2}{2\sigma^2}\right) \quad (12)$$

The gradient of Eq. 11 w.r.t $\mathbf{c}_p$ can be written as:

$$\frac{\partial E(\mathbf{c}_p)}{\partial \mathbf{c}_p} = (\mathbf{c}_pK_p - \mathbf{t}_p)\,K_p + \lambda\mathbf{c}_pK_p = \mathbf{0} \quad (13)$$

where we have used the fact that the kernel matrix K_p is symmetric. The optimal projection matrix $\mathbf{c}_p$ can then be computed as

$$\mathbf{c}_p = \mathbf{t}_p(K_p + \lambda I)^{-1} \quad (14)$$

Given an input pixel x at testing time, the corresponding output pixel y can be computed as

$$\begin{aligned} y &= w_p^\phi\phi(x) \\ &= \mathbf{c}_p\phi(\mathbf{x}_p)^T\phi(x) \\ &= \mathbf{t}_p(K_p + \lambda I)^{-1}\kappa(x) \end{aligned} \quad (15)$$

where the kernel vector $\kappa(x)$ is constructed as

$$\kappa(x) = \begin{bmatrix} k(x_p^1, x) & k(x_p^2, x) & \dots & k(x_p^N, x) \end{bmatrix}^T \quad (16)$$

IV. Experiments and Results

The proposed method is evaluated on four well-known facial expression synthesis datasets (KDEF, Bosphorous, Radboud and CFEE). The Karolinska Directed Emotional Faces (KDEF) [33] dataset consists of 70 participants (35 Male, 35 Female) facial expressions face images. Bosphorous [34] dataset consists of 105 participants facial expressions. Radboud faces dataset (RafD) [35] consists of 67 participants facial expressions from five different angles and three different gaze directions (left, right and frontal). We used only frontal images with frontal gaze directions in our experiments. The Compound Facial Expressions of Emotion (CFEE) [36] database consists of 230 participants face expressions. To evaluate the generalization performance of our proposed method, some arbitrary celebrities and well-known people images, portraits and avatars are downloaded from the internet. These images, called out-of-dataset images, have significant variability in terms of background and pose. We crop the face images into 128×128 size with faces in the center. The regularization parameter λ in Equation 6 and 14 is set to 0.4 after cross-validation. The parameter σ in the Gaussian function in Eq. 12 is cross-validated and set to 3 for neutral-to-happy mapping while for other expressions the sigma value is also cross-validated and set to 10. The pre-trained GANimation model

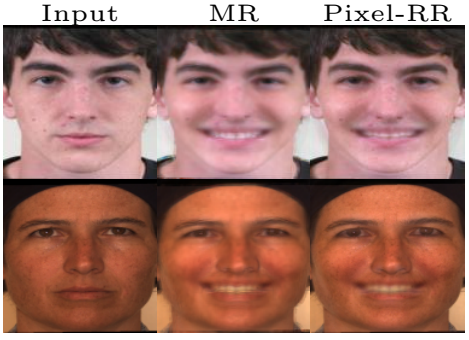

Fig. 3. Visual comparison of MR and Pixel-RR. These results demonstrate that the Pixel-RR obtains sharper images with subtle facial details as opposed to MR. The results of MR suffer from smoothness artifacts. Pixel-RR suppresses the blurring effects, generates realistic images and retains local details of the face such as freckles, sparkle in the eyes and contours of the face image.

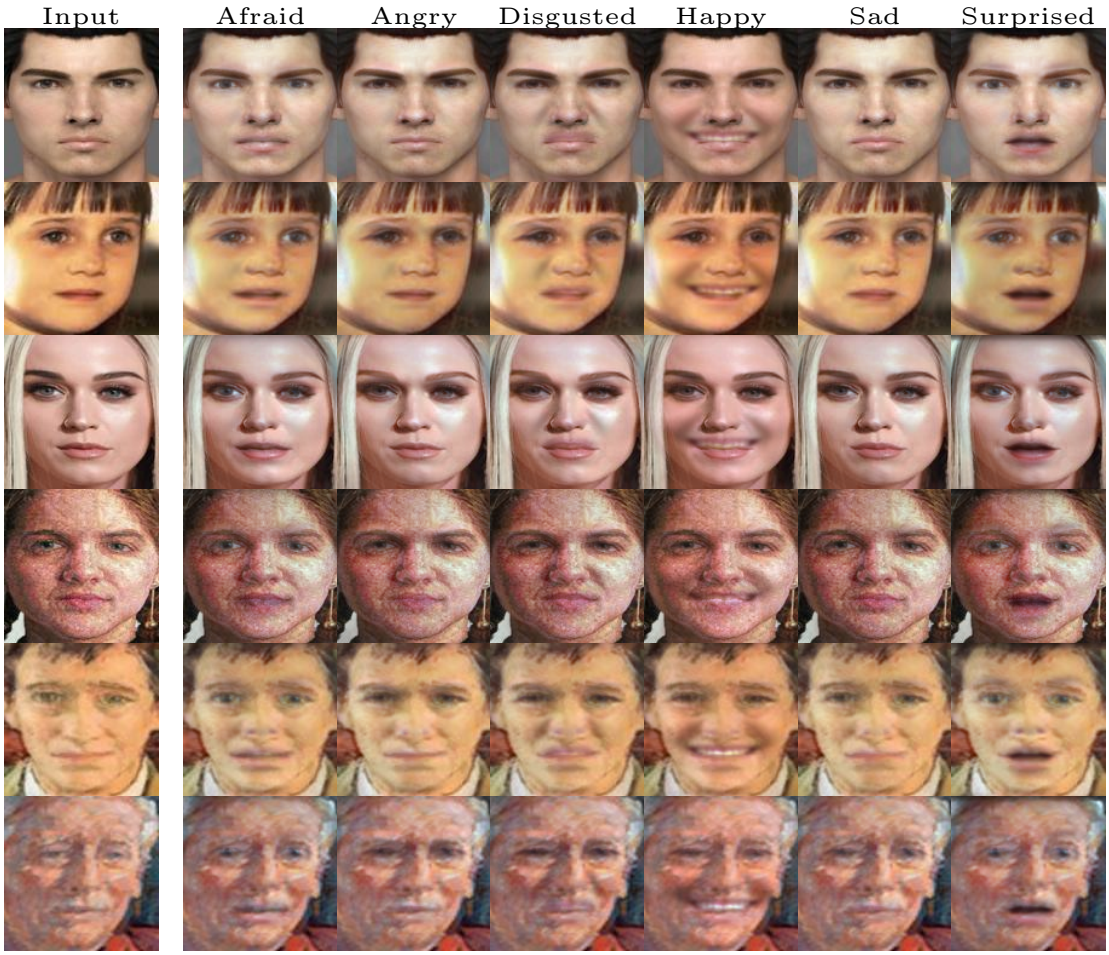

Fig. 4. Qualitative results of Pixel-RR for facial expression synthesis on out-of-dataset images. The proposed method successfully synthesized facial expressions on these out-of-dataset images. It can be seen that the proposed method performs well on portrait (last three rows) and non-frontal images (second and last row).

is used to generate GANimation results for comparison while StarGAN, MR, Pixel-RR and Pixel-KR are trained on the 400 images from all these datasets.

A. Results on in-dataset and out-of-dataset images

We compare our proposed method against state-of-the-art models including StarGAN [25], GANimation [26] and MR [27]. Qualitative results on in-dataset and out-of-dataset images can be seen in Figure 1 and 4. It can be seen that StarGAN and GANimation generate quite convincing happy expression on these images. However, these methods also induce artifacts around the eyes and mouth regions. It must be noted that GANimation sometimes also fails to preserve a person's identity as can be seen in the left panel of Figure 1 (column 3, rows 1, 3 and 4). MR synthesizes good happy expressions as compared to state-of-the-art GANs but the synthesized images also suffer from blurriness. Pixel-RR produces sharper results and induces plausible expressions on the input images. The right panel in Figure 1 presents results on out-of-dataset images. Results demonstrate that GANs are not able to synthesize happy expressions on animal faces while regression-based methods MR and Pixel-RR successfully induce happy expressions on animal faces as well. Results for the proposed Pixel-RR method are sharper than results for MR. This can also be seen from Figure 3 which also demonstrates that Pixel-RR retains finer facial details such as the sparkle in the eyes. Figure 4 demonstrates that Pixel-RR generalized well for out-of-dataset images such as portraits and faces with varying head pose.

B. Effects of σ on the results of Pixel-KR

As discussed and shown before, the idea of pixel-based mapping can be extended to capture the non-linear characteristics of facial images by using kernel-regression. However, Pixel-KR results are highly sensitive to choosing the optimal value of σ. To study the impact of σ in the facial expression synthesis problem, we perform experiments by varying σ from 0.1 to 5. Figure 5 shows that values smaller than 1 introduce artifacts on images whereas values greater than 1 produce blurry images. It can be seen that Pixel-KR produces results similar to Pixel-RR when $\sigma = 1$. However, the number of

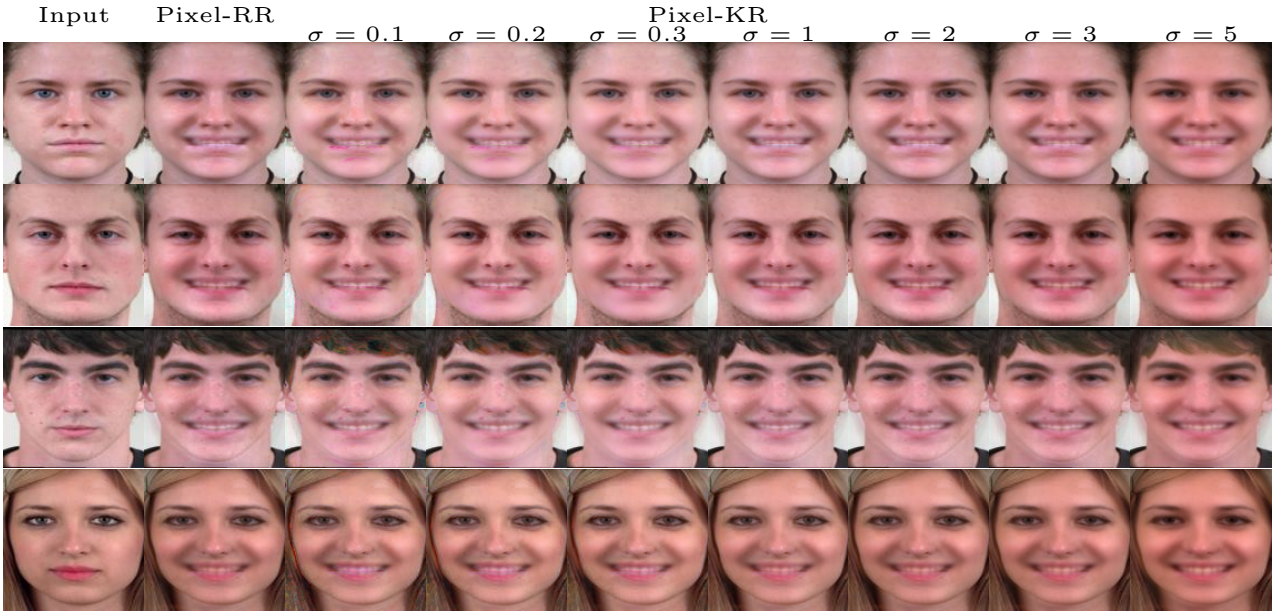

Fig. 5. Results of Pixel-KR for varying σ. It can be observed that the results of Pixel-KR are highly sensitive to the optimal value of σ.

parameters in Pixel-RR is much smaller as compared to Pixel-KR.

C. Quantitative Evaluation Results

In this subsection, we evaluate the quality of generated faces quantitatively.

TABLE II
USER STUDY TO EVALUATE EXPRESSIONS AND VISUAL QUALITY OF GENERATED IMAGES.

Model	Neutral $\rightarrow$ Happy
GANimation	0.26
MR	0.17
Pixel-RR	**0.57**

1) User Study: A user study is conducted to evaluate the quality of the visual results of the proposed method. Ten input images are selected for evaluation. The human evaluators are asked to choose the best synthesized happy face image based on these three attributes: perceptual quality, expression realism and identity preservation. This study has been performed by 80 evaluators. The results in Table II reinforce that our proposed regression-based method performs best among state-of-the-art facial expression synthesis methods.

2) Expression Classification Accuracy: To quantitatively evaluate the out-of-dataset generalization performance of the proposed method, we use the pre-trained expression[1] classifier to classify the synthesized images. A comparison of expression classification accuracy on the synthesized faces by our proposed method and other state-of-the-art facial expression synthesis methods is reported in Table III. These results also support our claim that the proposed approach generalizes well even without deep networks and large training datasets.

[1]https://github.com/thoughtworksarts/EmoPy

TABLE III
COMPARISON OF EXPRESSION CLASSIFICATION ACCURACY ON FACES GENERATED BY OUR PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS.

Model	Accuracy
GANimation	68%
MR	84%
Pixel-RR	**85%**

D. Performance of FES models on limited datasets

GAN based FES models such as StarGAN and GANimation are trained on huge datasets. StarGAN employs CelebA dataset to synthesize facial images with different attributes while GANimation utilizes EmotionNet [37] dataset for training. These datasets contain millions of images. However, such huge datasets are not readily available in all kinds of domains. Both pixel-based solutions perform well on in-dataset and out-of-dataset images while StarGAN is unable to produce realistic results (shown in Figure 1). It seems that StarGAN does not recover input image details and is not able to perform well with only few-hundred training images. This poor out-of-dataset generalization of GANs has also been discussed in [27]. In contrast, our proposed method performs significantly well even with such a small dataset. It generates satisfactory results on in-dataset and out-of-dataset images.

V. CONCLUSION

We have presented a novel and simple pixel-based ridge regression model for facial expression synthesis that considers only one input pixel to produce an output pixel. Qualitative and quantitative results confirm that Pixel-RR produces promising results on in-dataset and out-of-dataset images as compared to state-of-the-art GAN models while requiring two

orders of magnitude fewer parameters. Due to the smaller number of parameters, Pixel-RR can be easily deployed in mobile devices and embedded systems to synthesize realistic facial expressions. We have also shown that the pixel-based idea can be extended to capture non-linear characteristics of facial expression mappings by using kernel-regression. Results demonstrate that the proposed Pixel-RR and Pixel-KR methods effectively transform an input expression into a target expression while preserving identities and facial details. However, using the kernel-based method comes at the cost of increased number of parameters.

REFERENCES

[1] Q. Zhang, Z. Liu, G. Quo, D. Terzopoulos, and H.-Y. Shum, "Geometry-driven photorealistic facial expression synthesis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 1, pp. 48–60, 2005.

[2] R. B. Queiroz, A. Braun, and S. R. Musse, "A framework for generic facial expression transfer," *Entertainment Computing*, vol. 18, pp. 125–141, 2017.

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[4] T. F. Cootes and C. J. Taylor, "Active shape models—smart snakes," in *BMVC92*. Springer, 1992, pp. 266–275.

[5] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[6] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *BMVC*, vol. 1, no. 2. Citeseer, 2006, p. 3.

[7] D. Huang and F. De la Torre, "Bilinear kernel reduced rank regression for facial expression synthesis," in *ECCV*. Springer, 2010, pp. 364–377.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[11] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.

[12] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

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

[15] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[17] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[18] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 723–731.

[19] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[20] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang, "Towards the automatic anime characters creation with generative adversarial networks," *arXiv preprint arXiv:1708.05509*, 2017.

[21] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *Advances in neural information processing systems*, 2016, pp. 82–90.

[22] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," 2019. [Online]. Available: https://openreview.net/pdf?id=H1xQVn09FX

[23] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4030–4038.

[24] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5810–5818.

[25] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[26] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: One-shot anatomically consistent facial animation," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 698–713, 2020.

[27] N. Khan, A. Akram, A. Mahmood, S. Ashraf, and K. Murtaza, "Masked linear regression for learning local receptive fields for facial expression synthesis," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1433–1454, 2020.

[28] M. G. S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, and F. Hussain, "Machine learning at the network edge: A survey," *arXiv preprint arXiv:1908.00080*, 2020.

[29] J. Shi and G. Zhao, "Face hallucination via coarse-to-fine recursive kernel regression structure," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2223–2236, 2019.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international Conference on Computer Vision*, 2017, pp. 2223–2232.

[31] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1857–1865.

[32] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.

[33] D. Lundqvist and J. Litton, "The averaged karolinska directed emotional faces-akdef, cd rom from department of clinical neuroscience, psychology section, karolinska institutet," ISBN 91-630-7164-9, Tech. Rep., 1998.

[34] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," *Biometrics and identity management*, pp. 47–56, 2008.

[35] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[36] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[37] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.