Masked Linear Regression for Learning Local Receptive Fields for Facial Expression Synthesis

Nazar Khan   Arbish Akram   Arif Mahmood   Sania Ashraf
Kashif Murtaza

{nazarkhan, phdcsf18m002, sashraf, kashifmurtaza}@pucit.edu.pk
arif.mahmood@itu.edu.pk
http://faculty.pucit.edu.pk/nazarkhan/

Invited Talk at ITU, Lahore, Pakistan
15th October 2019

Facial expression synthesis (FES)

FES: Synthesis of a new expression on a given face.

How can we *learn* the transformation W?
Can we do it using a *linear* transformation (aka no deep learning)?

We present an FES method with the following contributions:

1. Convex optimization with closed-form solution of global minimum in a single iteration.
2. Extremely low spatial and computational complexity.
3. Trainable on very small datasets.
4. Intuitive interpretation of learned parameters can be exploited to improve results.
5. Good generalization over different types of images that state-of-the-art GANs find very challenging to synthesize.

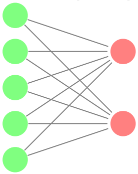
Related Work

- Basis learning (Blanz, Vetter, et al. 1999)
- Active appearance models (Cootes, Edwards, Taylor, et al. 2001)
- Deep belief nets (Susskind et al. 2008)
- Kernel regression (Huang and De la Torre 2010)
- GANs for image-to-image translation
  - Pix2Pix (Isola et al. 2017)
  - CycleGAN (Zhu et al. 2017)
  - StarGAN (Choi et al. 2018)
  - GANimation (Pumarola et al. 2019)

Regression

- Let $\mathbf{x} \in \mathbb{R}^D$ be a vectorized input image.
- Let $\mathbf{y} \in \mathbb{R}^K$ be a vectorized output image.
- Standard linear regression (ℓ_2) models output as $\mathbf{y} = W\mathbf{x}$ where $W \in \mathbb{R}^{K \times D}$ is a transformation matrix.
- This model corresponds to **global receptive fields**.
- Each output pixel is produced by *looking at* all input pixels.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} w_{11} & \dots & w_{1D} \\ w_{21} & \dots & w_{2D} \\ & \vdots & \\ w_{K1} & \dots & w_{KD} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$



Global

ℓ_2-regression – error formulation

$$E^{\mathrm{RR}}(W) = \frac{1}{2}||WX^T - T^T||_F^2 + \frac{\lambda_2}{2}||W||_F^2 \tag{1}$$

- $X \in \mathbb{R}^{N \times D}$ and $T \in \mathbb{R}^{N \times K}$ are the design matrices of vectorized input and target images respectively.
- Regularization parameter $\lambda_2 > 0$ controls over-fitting and $|| \cdot ||_F^2$ is the squared Frobenious norm of a matrix.
- This is a quadratic optimization problem with a global minimizer obtained in closed-form as

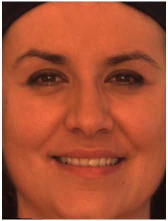
$$W^{\mathrm{RR}} = ((X^T X + \lambda_2 I)^{-1} X^T T)^T \tag{2}$$

Do all pixels determine expression?

Expression=?

Expression=?

Neutral



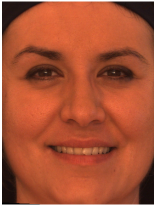
Happy

- Is there any benefit of *looking at* forehead pixels to generate smiling lips?
- Happy lips can be generated by looking at and transforming lips.
- Happy eyes can be generated by looking at and transforming eyes.
- So why carry so many parameters in W?

Expressions are local

- Transformation from one facial expression to another depends more on local information and less on global information.
- Facial expressions often constitute **sparsely distributed** and **locally correlated** changes.

Neutral

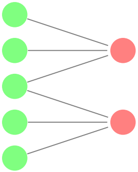


Happy

Masked Regression

- We propose a **Masked Regression (MR)**[1] model $\mathbf{y} = (W \circ M)\mathbf{x}$ where binary matrix M contains 1s only for locations that need to be looked at.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} w_{11}m_{11} & \dots & w_{1D}m_{1D} \\ w_{21}m_{21} & \dots & w_{2D}m_{2D} \\ & \vdots & \\ w_{K1}m_{K1} & \dots & w_{KD}m_{KD} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$



Local

- If $m_{ij} = 0$, then output pixel y_i is produced without looking at input pixel x_j.

[1] N. Khan et al. "Masked Linear Regression for Learning Local Receptive Fields for Facial Expression Synthesis". In: *International Journal of Computer Vision (IJCV)* (2019).

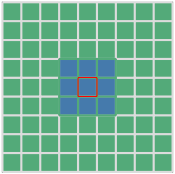
Linear Regression:

$$y_i = \sum_{j=1}^{D} w_{ij} x_j \tag{3}$$

Masked Regression:

$$y_i = \sum_{m_{ij}=1} w_{ij} x_j \tag{4}$$

If y_i is formed by looking at a 3×3 region in the input image, then the summation in MR is only over 9 pixels, irrespective of image size.



This corresponds to having **local receptive fields**.

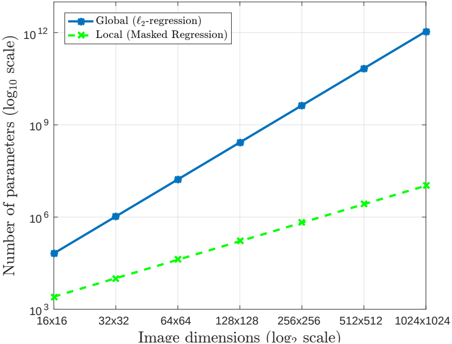
Input pixel index j in row-major order

Output pixel index i in row-major order

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1	1				1	1																		
2	1	1	1			1	1	1																	
3		1	1	1			1	1	1																
4			1	1	1			1	1	1															
5				1	1				1	1															
6	1	1				1	1				1	1													
7	1	1	1			1	1	1			1	1	1												
8		1	1	1			1	1	1			1	1	1											
9			1	1	1			1	1	1			1	1	1										
$\vdots$															$\vdots$										
19													1	1	1			1	1	1			1	1	1
20														1	1				1	1				1	1
21																1	1				1	1			
22																1	1	1			1	1	1		
23																	1	1	1			1	1	1	
24																		1	1	1			1	1	1
25																			1	1				1	1

Figure: Mask M corresponding to input image of size 5×5, output image of size 5×5 and receptive fields of size 3×3. For clarity, entries equal to 0 are left blank. If the entry at row i and column j is 1, then output pixel i has input pixel j in its receptive field.

Benefit of using mask



- Local receptive fields remain practical for larger image sizes.
- Regression with global receptive fields becomes impractical even for image sizes as small as 128×128 pixels.

Benefit of using mask

	Proposed	Pix2Pix	CycleGAN	StarGAN	GANimation
Size ($\times 10^4$)	1.68	4100	780	850	850
Time (msec)	2.70	320	710	580	507

- Comparison of MR with 4 state-of-the-art GAN architectures
- MR has more than two orders of magnitude fewer number of parameters than each of these GANs.
- MR is more than two orders of magnitude faster in synthesizing an expression.

Masked Regression – error formulation

- The error function for Masked Regression can be written as

$$E^{\text{MR}}(W) = \frac{1}{2}||(W \circ M)X^T - T^T||_F^2 + \frac{\lambda_M}{2}||W \circ M||_F^2 \quad (5)$$

- Only those weights are learned for which $m_{ij} = 1$. The rest are fixed to 0.
- Closed-form solution cannot be obtained due to the Hadamard product.

Masked Regression – error formulation

- Per-pixel decomposition

$$E^{\mathrm{MR}}(W) = \sum_{i=1}^{K} E^{\mathrm{MR}}(W^i) \tag{6}$$

where

$$E^{\mathrm{MR}}(W^i) = \frac{1}{2}||(W^i \circ M^i)X^T - T_i^T||_2^2 + \frac{\lambda_M}{2}||W^i \circ M^i||_2^2 \tag{7}$$

where W^i is the i-th row of W.

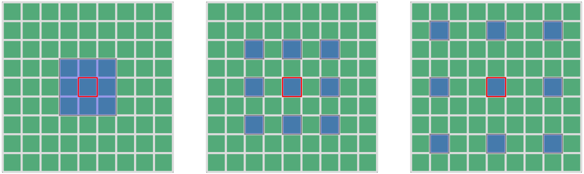
- Gradient and Hessian computations are a bit involved (refer to paper (Khan et al. 2019)).
- **Globally optimal W^i can now be computed in closed-form.**

Masked Regression – error formulation

- For receptive field of size $r \times r$, the i-th row of W can be computed independently by solving a linear system in r^2 unknowns.
- Linear regression would require solving a linear system in D^2 unknowns and $D^2 \gg r^2$.
- Example: for 128×128 images and 3×3 receptive fields, $D^2 = 128^4$ and $r^2 = 9$.

Dilated receptive fields

- The proposed method can be easily modified to have not-so-local receptive fields.
- We use dilated receptive fields to observe larger input regions using the same number of weights.
- This helps to avoid over-fitting by limiting the complexity of the model.



3×3 5×5 7×7

Local vs Sparse

- Local receptive fields can be viewed as extremely sparse receptive fields with manually designed and fixed localizations.
- Alternative: learn sparse receptive fields.
- **Will a sparsely learned topology also converge to our local receptive fields?**

Local vs Sparse

- To answer this question we learn the receptive field W^i for each output pixel by minimizing the ℓ_1-regularized sum of squared errors

$$\min_{W^i} \frac{1}{2}||XW^i - T_i||_2^2 + \lambda_1||W^i||_1 \tag{8}$$

using the LASSO algorithm[2].

- We also learn by minimizing the ℓ_0-regularized sum of squared errors

$$\min_{W^i} \frac{1}{2}||XW^i - T_i||_2^2 \text{ s.t. } ||W^i||_0 \leq \lambda_0 \tag{9}$$

using the OMP algorithm[3].

[2] Tibshirani 1996.

[3] Pati, Rezaiifar, and Krishnaprasad 1993.

Experiments

- We combine three datasets
  - KDEF (Lundqvist, Flykt, and Öhman 1998)
  - Bosphorous (Savran et al. 2008)
  - JAFFE (Lyons et al. 1998)
- Expressions include neutral, afraid, angry, disgusted, happy, sad and surprised.
- Total 1116 facial expression images.
- Approximately only 200 images per-mapping.
- $80\%, 10\%, 10\%$ split into training, validation and testing sets.
- All faces are aligned with respect to a reference face image.
- Pixel values normalized between 0 and 1.
- Hyperparameters were cross-validated.

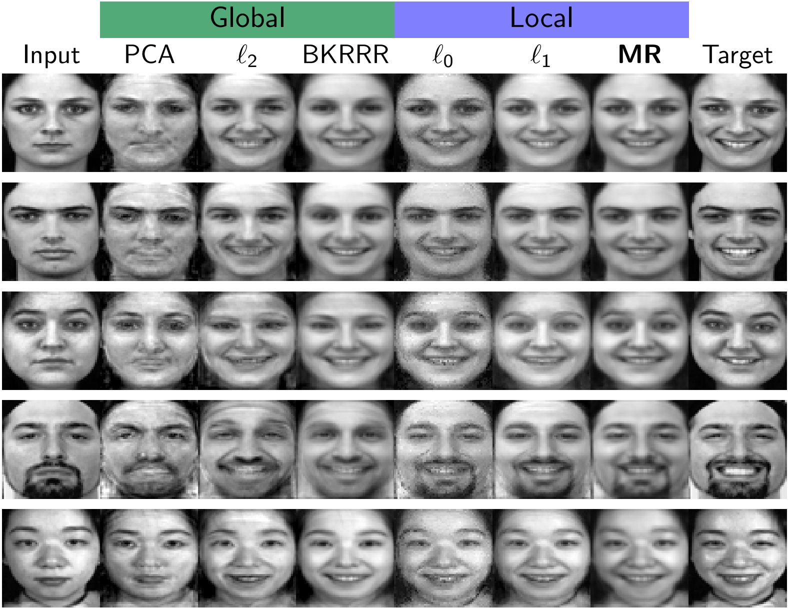


Figure: MR sucessfully sntesized a happy expression while preserving identity and retaining facial details the most.

Figure: For each neutral input, MR effectively transformed into 6 different expressions while preserving identities and facial details.

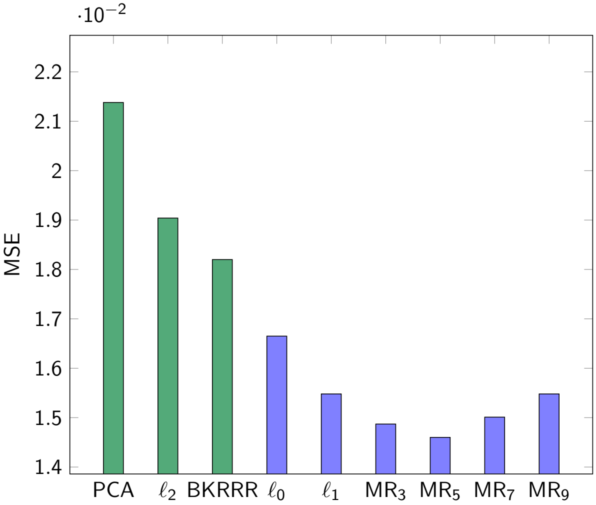


Figure: MSE for different methods averaged over 12 expression mappings. Employing too large a receptive field increased the MSE since long-range receptive fields fail to capture the local nature of facial expressions.

Training times and Sparsity

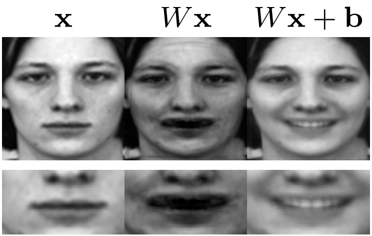
Table: Comparison of training times in seconds averaged over 12 different expression mappings.

MR	ℓ_1	ℓ_0	ℓ_2
0.010	16.782	0.237	0.115

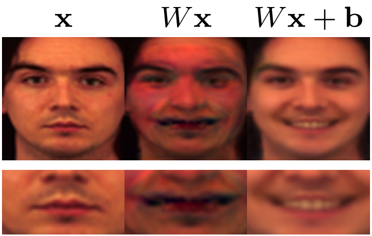
MR models were *twice as sparse* as the best cross-validated ℓ_1-regression models.

Learned Biases

Role of weights and biases



Grayscale synthesis

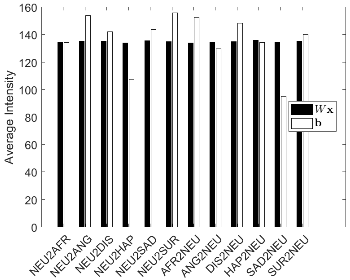


Color synthesis

- Weights are predominantly used to transform the visible parts of the input expression into the target.
- Biases are used to insert hidden information such as teeth for a happy expression.

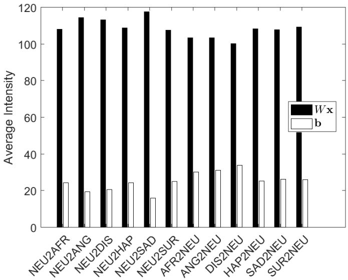
Over 12 expression mappings, we compare the average absolute intensity of the transformation produced by the weights with the additive transformation learned as biases.

ℓ_2-regression



Bias often dominated the weights.
Leads to loss of identity.

Masked regression



Weights $\sim$5 times as important as bias.
Leads to better identity preservation.

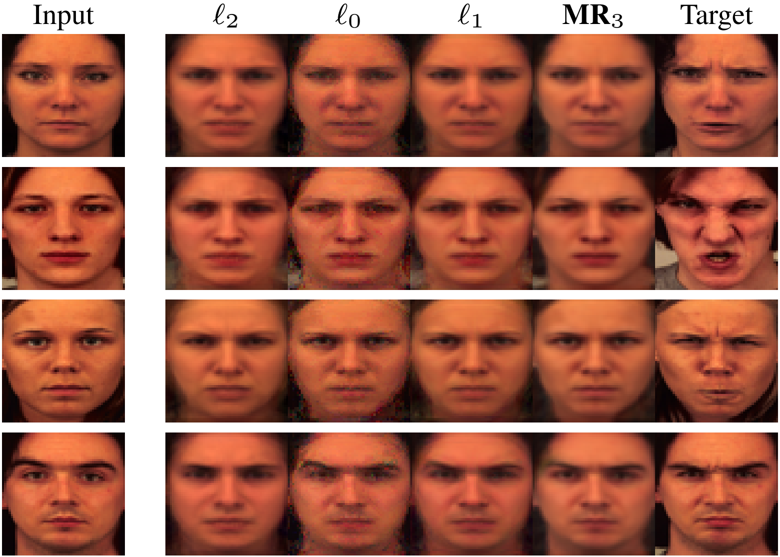


Figure: MR preserves background and other details unrelated to the desired expression.

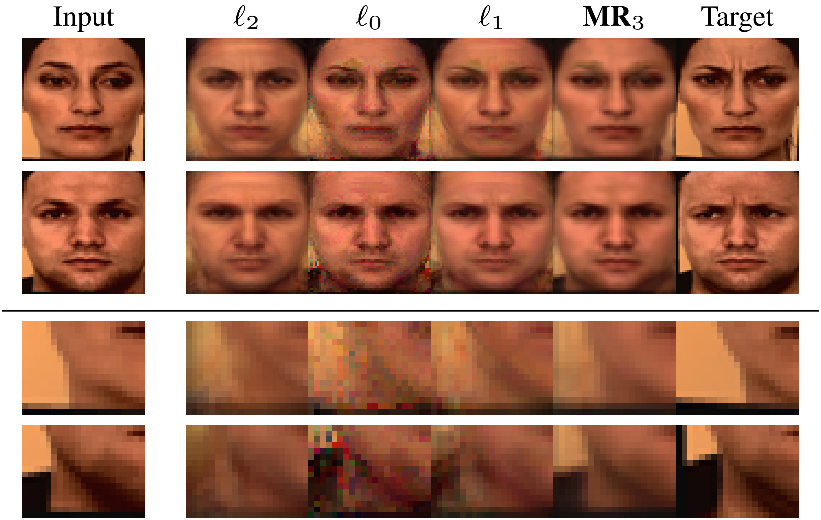


Figure: MR preserves background and other details unrelated to the desired expression.

MR for color images

$$E^{\mathrm{CMR}}(W) = \frac{1}{2}\sum_{c=1}^{C} ||(W \circ M)X_c^T - T_c^T||_F^2 + \frac{\lambda_M}{2}||W \circ M||_F^2 \quad (10)$$

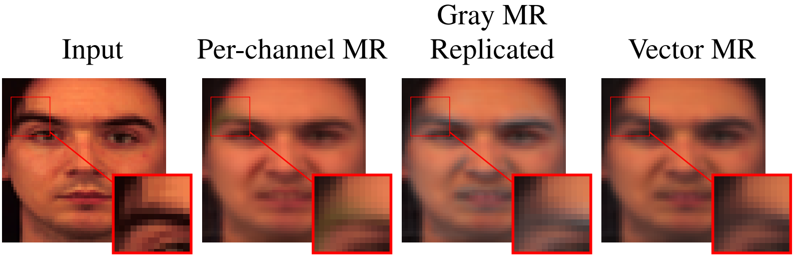


Figure: Regression on color tuples leads to lesser color leakage compared to separate regressions on each channel.

MR on non-frontal faces

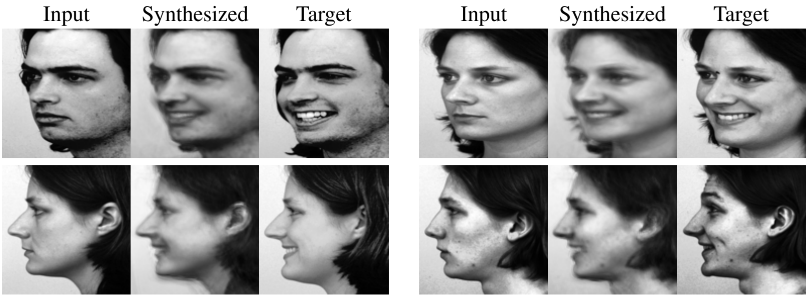
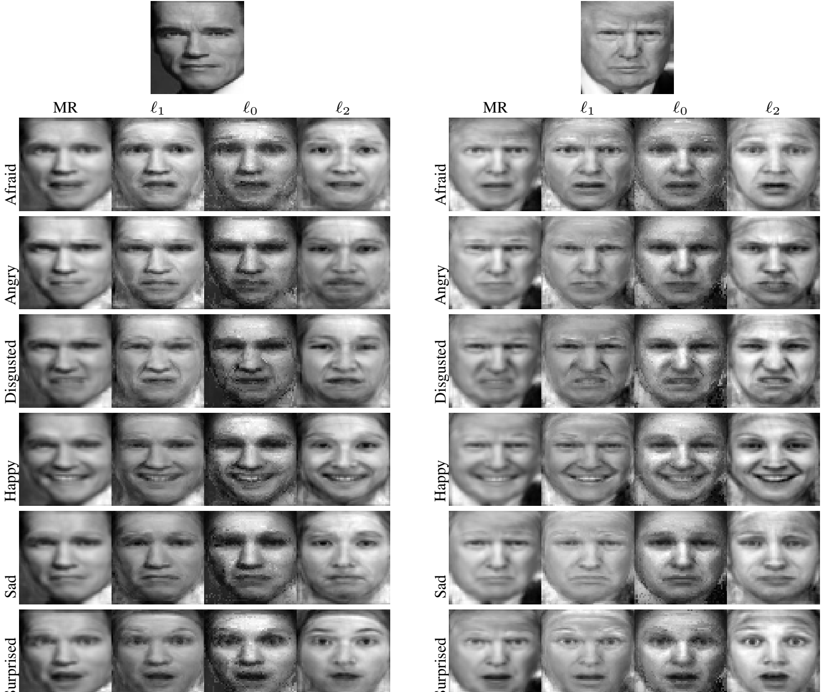


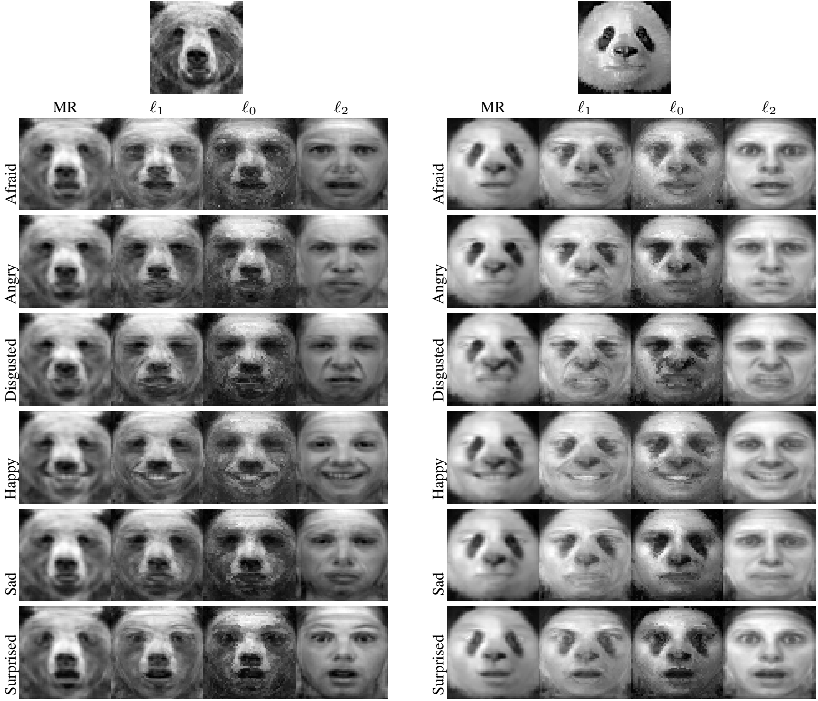
Figure: FES on non-frontal faces.

Out-of-dataset generalization

- MR is most effective for *out-of dataset* images.
- Images not belonging to any of the datasets used for training, validation and testing.
- Such images belong to significantly different distributions compared to training distribution.
- Three categories
  1. People
  2. Sketch drawings
  3. Animals


MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised
MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised


MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised
MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised


MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised
MR
ℓ_1
ℓ_0
ℓ_2
Afraid
Angry
Disgusted
Happy
Sad
Surprised

Refinement

- MR adjusts weights according to whether a particular pixel is relevant for a particular expression.
- For generating happy expressions, an output pixel *looking at* the mouth might have a greater role than a pixel looking at the forehead.
- We compute the *ego* of each output pixel as follows
  1. Compute ℓ_1-norm of each pixel's receptive field, i.e. $|W^i|_1$.
  2. Compute mean and standard deviation of these norms.
  3. Standardize to obtain z-scores. High score $\implies$ atypical field.
  4. Dilate with disk to expand influence of atypical receptive fields.
  5. Post-process and scale between 0 and 1.

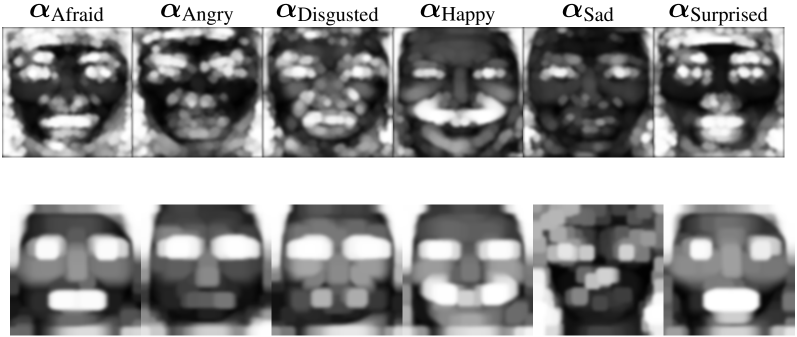
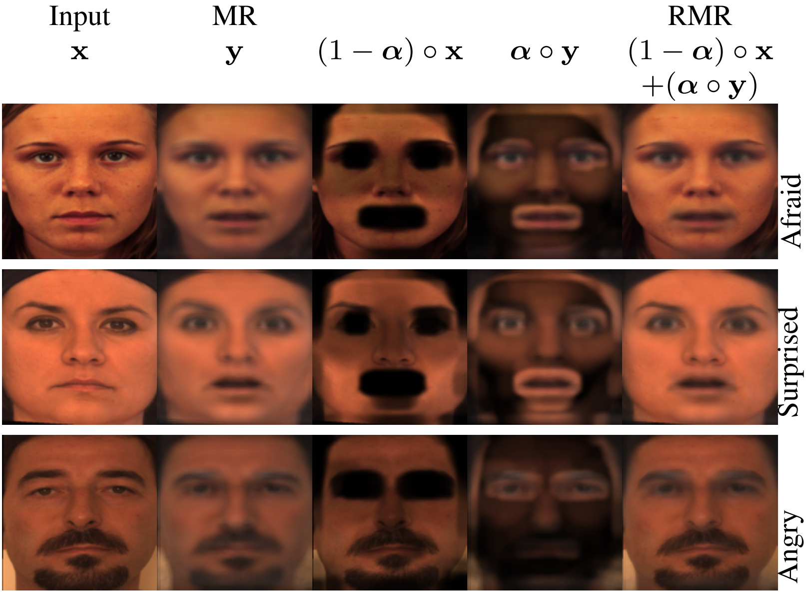


Figure: **Top**: Role of each pixel for expression generation. Higher intensity implies greater role. The role of the *i*-th pixel is computed entirely from its learned receptive field W^i. **Bottom**: Using different dilation and post-processing parameters.

$$\mathbf{y}' = (1 - \boldsymbol{\alpha}) \circ \mathbf{x} + \boldsymbol{\alpha} \circ \mathbf{y} \tag{11}$$



Input
$\mathbf{x}$
MR
$\mathbf{y}$
$(1-\boldsymbol{\alpha}) \circ \mathbf{x}$
$\boldsymbol{\alpha} \circ \mathbf{y}$
RMR
$(1-\boldsymbol{\alpha}) \circ \mathbf{x} + (\boldsymbol{\alpha} \circ \mathbf{y})$
Afraid
Surprised
Angry

Input	MR			RMR
$\mathbf{x}$	$\mathbf{y}$	$(1-\boldsymbol{\alpha})\circ\mathbf{x}$	$\boldsymbol{\alpha}\circ\mathbf{y}$	$(1-\boldsymbol{\alpha})\circ\mathbf{x}$ $+(\boldsymbol{\alpha}\circ\mathbf{y})$

Input	MR			RMR
$\mathbf{x}$	$\mathbf{y}$	$(1-\boldsymbol{\alpha})\circ\mathbf{x}$	$\boldsymbol{\alpha}\circ\mathbf{y}$	$(1-\boldsymbol{\alpha})\circ\mathbf{x}$ $+(\boldsymbol{\alpha}\circ\mathbf{y})$

Happy

Happy

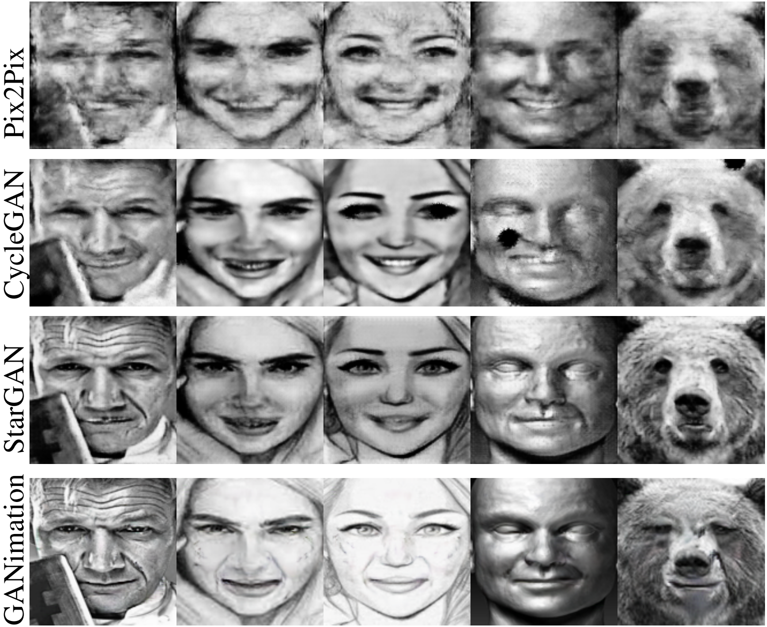
GAN comparisons

- We compare with 4 state-of-the-art GANs used for image-to-image translation tasks.
- GANs produce sharp photo-realistic results.
- Good results **as long as test image belongs to the same distribution as training images.**
- Poor out-of-dataset generalization.

In-dataset images


Pix2Pix
CycleGAN
StarGAN
GANimation

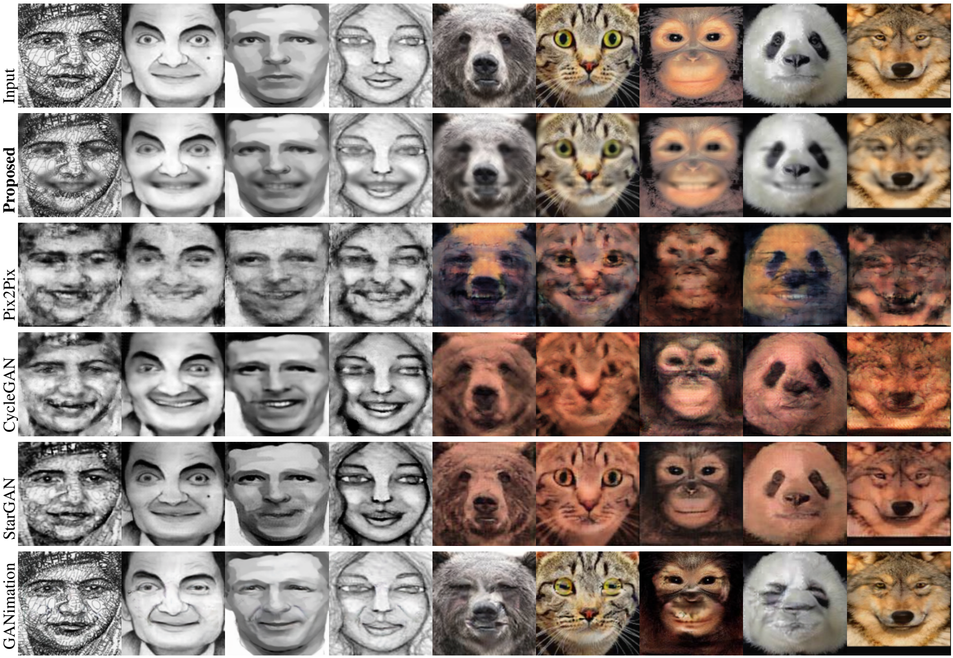
Out-of-dataset images (sketches and animals)


Pix2Pix
CycleGAN
StarGAN
GANimation

Out-of-dataset images (people)


Pix2Pix
CycleGAN
StarGAN
GANimation


Input
Afraid
Angry
Disgusted
Happy
Sad
Surprised
RMR
GANimation
RMR
GANimation
RMR
GANimation


Input
Proposed
Pix2Pix
CycleGAN
StarGAN
GANimation

Quantitative comparison with GANs

Table: Drop in expression recognition accuracy (in percentage points) when changing from test set images to out-of-dataset images.

Pix2Pix	CycleGAN	StarGAN	GANimation	MR
35.72	16.39	21.43	20.74	12.39

Conclusion

- Constrained version of ridge regression for local receptive fields.
- Efficient closed-form solution of global minimum.
- Excellent learning ability on very small datasets despite simplicity.
- Easy implementation and extremely fast training.
- Better generalization despite using small training datasets.
- Extremely small model size.
- Intuitive interpretation of receptive fields exploited to refine results.
- Better out-of-dataset generalization compared to state-of-the-art GANs.

References I

Volker Blanz, Thomas Vetter, et al. "A morphable model for the synthesis of 3D faces.". In: *SIGGRAPH*. Vol. 99. 1999. 1999, pp. 187–194.

Timothy F Cootes, Gareth J Edwards, Christopher J Taylor, et al. "Active appearance models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), pp. 681–685.

Yunjey Choi et al. "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.

References II

Dong Huang and Fernando De la Torre. "Bilinear kernel reduced rank regression for facial expression synthesis". In: *European Conference on Computer Vision*. Springer. 2010, pp. 364–377.

Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1125–1134.

N. Khan et al. "Masked Linear Regression for Learning Local Receptive Fields for Facial Expression Synthesis". In: *International Journal of Computer Vision (IJCV)* (2019).

References III

D. Lundqvist, A. Flykt, and A. Öhman. *The Karolinska Directed Emotional Faces - KDEF, CD ROM*. Stockholm, Sweden: Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, 1998. ISBN: 91-630-7164-9.

Michael Lyons et al. "Coding facial expressions with Gabor wavelets". In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE. 1998, pp. 200–205.

References IV

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition". In: *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. Nov. 1993, 40–44 vol.1.

A. Pumarola et al. "GANimation: One-Shot Anatomically Consistent Facial Animation". In: *International Journal of Computer Vision (IJCV)* (2019).

Arman Savran et al. "Bosphorus database for 3D face analysis". In: *European Workshop on Biometrics and Identity Management*. Springer. 2008, pp. 47–56.

References V

Joshua M Susskind et al. "Generating facial expressions with deep belief nets". In: *Affective Computing*. Ed. by Jimmy Or. InTech, 2008. Chap. 10, pp. 421–440.

Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2223–2232.

Questions?