Continuous Emotion Transfer

Zoltán Szabó

joint work with

- Alex Lambert @ KU Leuven,
- Sanjeel Parekh, Florence d'Alché-Buc @ Télécom Paris.

LSE Statistics Research Day
June 15, 2022

Style transfer

- Goal: transfer an object according to a target style.

Numerous applications

- Computer vision [Ulyanov et al., 2016, Choi et al., 2018, Puy and Pérez, 2019, Yao et al., 2020], NLP [Fu et al., 2018], audio signal processing [Grinstein et al., 2018].
- Graphics: animating digital characters & avatars → body MOCAP [Aristidou et al., 2017, Aberman et al., 2020].

- Health & industry: digital twinning [Tao et al., 2019, Barricelli et al., 2019, Lim et al., 2020].

Style transfer

- Goal: transfer an object according to a target style.

Numerous applications

- Computer vision [Ulyanov et al., 2016, Choi et al., 2018, Puy and Pérez, 2019, Yao et al., 2020], NLP [Fu et al., 2018], audio signal processing [Grinstein et al., 2018].
- Graphics: animating digital characters & avatars → body MOCAP [Aristidou et al., 2017, Aberman et al., 2020].

- Health & industry: digital twinning [Tao et al., 2019, Barricelli et al., 2019, Lim et al., 2020].

Our aim: have a 'slider' (in $\mathbb{R}^p$; <u>continuum</u> of styles)

One-page summary

Focus: novel task

continuous style transfer $\xleftarrow{\text{spec.}}$ functional output regression.

- Framework: vv-RKHS.
- Ingredients: similarity on
  - objects: $k_{\mathcal{X}}$,
  - style: k_{Θ},
  - continuous style space: $\Theta \subset \mathbb{R}^p \leftrightarrow$ the slider.
- Running example: emotion transfer.

Emotion transfer

- Given: set of emotions.
- Goal: transform object representations of
  - faces [Choi et al., 2018], hands [Irimia et al., 2019], body movement [Aristidou et al., 2017], ...
  - repr: 2D images, 3D meshes, body skeletons, MOCAP sequences.

Emotion transfer

- Given: set of emotions.
- Goal: transform object representations of
  - faces [Choi et al., 2018], hands [Irimia et al., 2019], body movement [Aristidou et al., 2017], . . .
  - repr: 2D images, 3D meshes, body skeletons, MOCAP sequences.

Our example

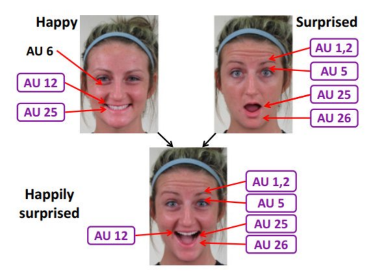
- style = emotion, object repr. = landmark locations.

Emotion transfer

Our example

- style = emotion, object repr. = landmark locations.

Why facial landmarks? ($\approx$ simplified FACS)



Problem formulation

- Object space: $\mathcal{X}$. Style space: Θ.
- Goal: (object, style) $\mapsto$ object, i.e. an

$$h : \mathcal{X} \times \Theta \mapsto \mathcal{X}, \text{ or } h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X}).$$

Problem formulation

- Object space: $\mathcal{X}$. Style space: Θ.
- Goal: $\boxed{\text{(object, style)} \mapsto \text{object}}$, i.e. an

$$h : \mathcal{X} \times \Theta \mapsto \mathcal{X}, \text{ or } h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X}).$$

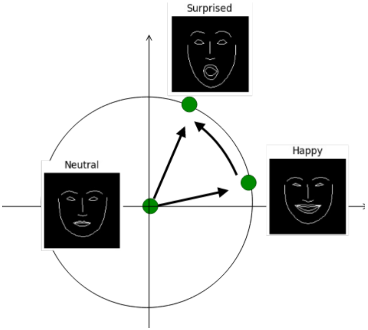
- In our case: landmarks $\mapsto \underbrace{(\text{emotion} \mapsto \text{landmarks})}_{\text{function-valued regression}}$.

Problem formulation

- Object space: $\mathcal{X}$. Style space: Θ.
- Goal: $\boxed{\text{(object, style)} \mapsto \text{object}}$, i.e. an

$$h : \mathcal{X} \times \Theta \mapsto \mathcal{X}, \text{ or } h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X}).$$

- In our case: landmarks $\mapsto \underbrace{(\text{emotion} \mapsto \text{landmarks})}_{\text{function-valued regression}}$.

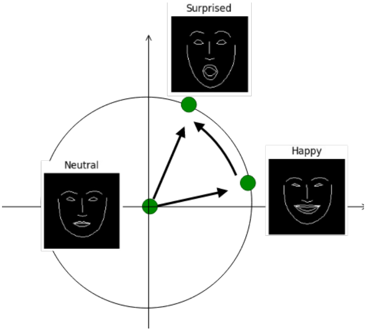


Problem formulation

- Object space: $\mathcal{X}$. Style space: Θ.
- Goal: $\boxed{\text{(object, style)} \mapsto \text{object}}$, i.e. an

$$h : \mathcal{X} \times \Theta \mapsto \mathcal{X}, \text{ or } h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X}).$$

- In our case: $\text{landmarks} \mapsto \underbrace{(\text{emotion} \mapsto \text{landmarks})}_{\text{function-valued regression}}$.



- $h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{Y})$ would work similarly [$\mathcal{Y}$ = avatars].

Cost function

- Training samples:
  - For each object $i \in [n]$: $|S_i|$ style transition pairs $\left\{(\theta_{i,j}^{\text{in}}, \theta_{i,j}^{\text{out}})\right\}_{j \in S_i}$.
  - $x_{i,j} \in \mathcal{X}$: object with input style $\theta_{i,j}^{\text{in}}$,
  - $y_{i,j} \in \mathcal{X}$: object with output style $\theta_{i,j}^{\text{out}}$.

Cost function

- Training samples:
  - For each object $i \in [n]$: $|S_i|$ style transition pairs $\left\{\left(\theta_{i,j}^{\text{in}}, \theta_{i,j}^{\text{out}}\right)\right\}_{j \in S_i}$.
  - $x_{i,j} \in \mathcal{X}$: object with input style $\theta_{i,j}^{\text{in}}$,
  - $y_{i,j} \in \mathcal{X}$: object with output style $\theta_{i,j}^{\text{out}}$.
- Cost (quality of the reconstruction) of h:

$$\mathcal{R}_{\mathcal{S}}(h) := \frac{1}{n} \sum_{i \in [n]} \frac{1}{|S_i|} \sum_{j \in S_i} \ell\Big(\overbrace{h(\underbrace{x_{i,j}}_{\text{input object}})(\underbrace{\theta_{i,j}^{\text{out}}}_{\text{output style}})}^{\text{predicted output object}}, \underbrace{y_{i,j}}_{\text{output object}} \Big).$$

Cost function

- Training samples:
  - For each object $i \in [n]$: $|S_i|$ style transition pairs $\left\{\left(\theta_{i,j}^{\text{in}}, \theta_{i,j}^{\text{out}}\right)\right\}_{j \in S_i}$.
  - $x_{i,j} \in \mathcal{X}$: object with input style $\theta_{i,j}^{\text{in}}$,
  - $y_{i,j} \in \mathcal{X}$: object with output style $\theta_{i,j}^{\text{out}}$.
- Cost (quality of the reconstruction) of h:

$$\mathcal{R}_{\mathcal{S}}(h) := \frac{1}{n} \sum_{i \in [n]} \frac{1}{|S_i|} \sum_{j \in S_i} \ell\Bigg(\overbrace{h(\underbrace{x_{i,j}}_{\substack{\text{input}\\\text{object}}})(\underbrace{\theta_{i,j}^{\text{out}}}_{\substack{\text{output}\\\text{style}}})}^{\text{predicted output object}}, \underbrace{y_{i,j}}_{\substack{\text{output}\\\text{object}}} \Bigg).$$

- Quadratic loss: $\ell = \frac{1}{2} \|\cdot\|_{\mathcal{X}}^2$.

Cost function

- Training samples:
  - For each object $i \in [n]$: $|S_i|$ style transition pairs $\left\{\left(\theta_{i,j}^{\text{in}}, \theta_{i,j}^{\text{out}}\right)\right\}_{j \in S_i}$.
  - $x_{i,j} \in \mathcal{X}$: object with input style $\theta_{i,j}^{\text{in}}$,
  - $y_{i,j} \in \mathcal{X}$: object with output style $\theta_{i,j}^{\text{out}}$.
- Cost (quality of the reconstruction) of h:

$$\mathcal{R}_{\mathcal{S}}(h) := \frac{1}{n} \sum_{i \in [n]} \frac{1}{|S_i|} \sum_{j \in S_i} \ell\Big(\overbrace{h(\underbrace{x_{i,j}}_{\substack{\text{input}\\ \text{object}}})(\underbrace{\theta_{i,j}^{\text{out}}}_{\substack{\text{output}\\ \text{style}}})}^{\text{predicted output object}}, \underbrace{y_{i,j}}_{\substack{\text{output}\\ \text{object}}} \Big).$$

- Quadratic loss: $\ell = \frac{1}{2} \|\cdot\|_{\mathcal{X}}^2$.
- Hypothesis class (vv-RKHS): $h : \underbrace{\mathcal{X} \mapsto \underbrace{(\Theta \mapsto \mathcal{X})}_{\in \mathcal{F} := \mathcal{H}_G}}_{\in \mathcal{H} := \mathcal{H}_K}$.

Emotion representation: $\Theta \subset \mathbb{R}^p$

- Classical categorical description:

  'happy', 'sad', 'angry', 'surprised', 'disgusted', 'fearful'.

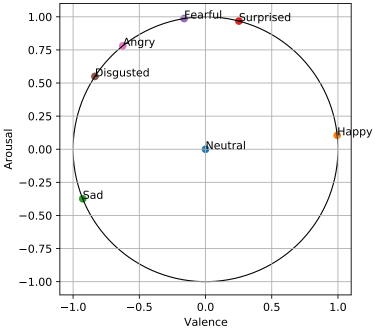
Emotion representation: $\Theta \subset \mathbb{R}^p$

- Classical categorical description:

  'happy', 'sad', 'angry', 'surprised', 'disgusted', 'fearful'.

- Valence-arousal model [Russell, 1980]: $\Theta \subset \mathbb{R}^2$,
  - valence: pleasure to displeasure,
  - arousal: high to low.

  Normalized demo:

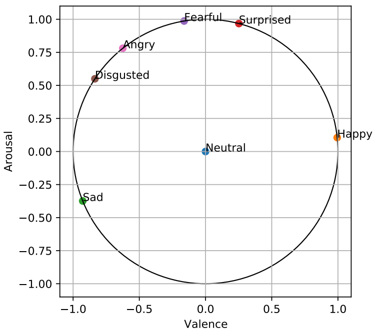


Emotion representation: $\Theta \subset \mathbb{R}^p$

- Classical categorical description:

  'happy', 'sad', 'angry', 'surprised', 'disgusted', 'fearful'.
- Valence-arousal model [Russell, 1980]: $\Theta \subset \mathbb{R}^2$,
  - valence: pleasure to displeasure,
  - arousal: high to low.

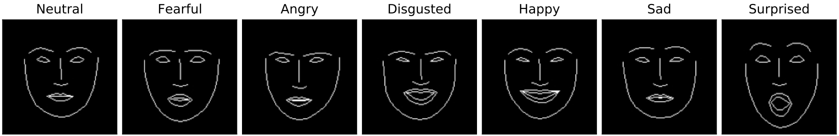
  Normalized demo:



- HigherD ($\Theta \subset \mathbb{R}^p$, $p \geq 2$) [Vemulapalli and Agarwala, 2019].

Object representation: $\mathcal{X} \subset \mathbb{R}^d$

- Face: landmarks points.
- Example: corners of the eyes, that of the mouth, ...



- Typically $M \approx 50 - 100 \Rightarrow$
  - $\mathcal{X} \subset \mathbb{R}^{d:=2M}$,
  - compact description (few trees: saved).

Hypothesis class: $\mathcal{H}$

- Recall: $h : \mathcal{X} \mapsto \underbrace{(\underbrace{\boxed{\Theta \mapsto \mathcal{X}}}_{\in \mathcal{F} := \mathcal{H}_G})}_{\in \mathcal{H} := \mathcal{H}_K}$.
- Model: vv-RKHSs,

$$G(\theta, \theta') = k_\Theta(\theta, \theta')\mathbf{A}, \qquad K(x, x') = k_\mathcal{X}(x, x')\mathrm{Id}_{\mathcal{H}_G},$$

Hypothesis class: $\mathcal{H}$

- Recall: $h : \mathcal{X} \mapsto \underbrace{\left(\underbrace{\boxed{\Theta \mapsto \mathcal{X}}}_{\in \mathcal{F} := \mathcal{H}_G}\right)}_{\in \mathcal{H} := \mathcal{H}_K}$.
- Model: vv-RKHSs,

$$G(\theta, \theta') = k_\Theta(\theta, \theta')\mathbf{A}, \qquad K(x, x') = k_\mathcal{X}(x, x')\mathrm{Id}_{\mathcal{H}_G},$$

- Ingredients:
  - smoothness: Gaussian kernel
    - $k_\Theta(\theta, \theta') = e^{-\gamma\|\theta - \theta'\|_2^2}$ $(\gamma > 0)$,
    - $k_\mathcal{X}$: similarly on $\mathcal{X}$.
  - dependency among output coordinates: $\mathbf{A} \succcurlyeq \mathbf{0}$.

Optimization

- vv-RKHS $\Rightarrow$ rich still tractable; natural regularization.
- Task:

$$\min_{h \in \mathcal{H}_K} \mathcal{R}_\lambda(h) := \underbrace{\mathcal{R}_{\mathcal{S}}(h)}_{\text{data fitting}} + \frac{\lambda}{2} \underbrace{\|h\|^2_{\mathcal{H}_K}}_{\text{smoothness}}, \ \lambda > 0.$$

Optimization

- vv-RKHS $\Rightarrow$ rich still tractable; natural regularization.
- Task:

$$\min_{h\in\mathcal{H}_K} \mathcal{R}_\lambda(h) := \underbrace{\mathcal{R}_{\mathcal{S}}(h)}_{\text{data fitting}} + \frac{\lambda}{2} \underbrace{\|h\|^2_{\mathcal{H}_K}}_{\text{smoothness}}, \quad \lambda > 0.$$

- Representer lemma:

$$\hat{h}(x)(\theta) = \sum_{i=1}^{t}\sum_{j=1}^{m} k_{\mathcal{X}}(x,x_i)k_{\Theta}(\theta,\theta_{i,j})\mathbf{A}\hat{\mathbf{c}}_{i,j}, \ \{\hat{\mathbf{c}}_{i,j}\}_{i\in[t],j\in[m]} \subset \mathbb{R}^d.$$

- vv-RKHS $\Rightarrow$ rich still tractable; natural regularization.
- Task:

$$\min_{h\in\mathcal{H}_K} \mathcal{R}_\lambda(h) := \underbrace{\mathcal{R}_{\mathcal{S}}(h)}_{\text{data fitting}} + \frac{\lambda}{2}\underbrace{\|h\|^2_{\mathcal{H}_K}}_{\text{smoothness}}, \quad \lambda > 0.$$

- Representer lemma:

$$\hat{h}(x)(\theta) = \sum_{i=1}^{t}\sum_{j=1}^{m} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_{i,j})\mathbf{A}\hat{\mathbf{c}}_{i,j}, \quad \{\hat{\mathbf{c}}_{i,j}\}_{i\in[t], j\in[m]} \subset \mathbb{R}^d.$$

- With quadratic loss (ℓ): linear equation to $\hat{\mathbf{c}}_{i,j}$-s.

Towards demos: two problem families

- Single emotional input:
  - input emotion: identical & fixed for everyone (θ_0).
  - output emotion: same m number.

$\Rightarrow$ I-O emotion pairs: $\{(\theta_0, \theta_{i,j})\}_{j\in[m]}$, $|S_i| = m$ $\forall i$.

Towards demos: two problem families

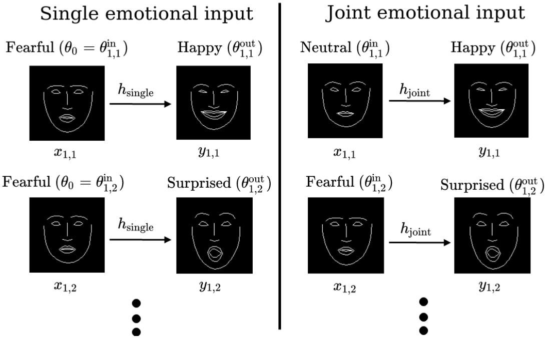
- Single emotional input:
  - input emotion: identical & fixed for everyone (θ_0).
  - output emotion: same m number.

  $\Rightarrow$ I-O emotion pairs: $\{(\theta_0, \theta_{i,j})\}_{j \in [m]}$, $|S_i| = m$ $\forall i$.
- Joint emotional input:
  - m emotions for each person: $\{\theta_{i,a}\}_{a \in [m]}$, with all combinations,

  $\Rightarrow$ I-O emotion pairs: $\{(\theta_{i,a}, \theta_{i,b})\}_{a,b \in [m]}$, $|S_i| = m^2$ $\forall i$.

Towards demos: two problem families

- Single emotional input:
  $\Rightarrow$ I-O emotion pairs: $\{(\theta_0, \theta_{i,j})\}_{j\in[m]}$, $|S_i| = m \; \forall i$.
- Joint emotional input:
  $\Rightarrow$ I-O emotion pairs: $\{(\theta_{i,a}, \theta_{i,b})\}_{a,b\in[m]}$, $|S_i| = m^2 \; \forall i$.



Quantitative illustration: setting

- 2 popular facial benchmarks: KDEF, RaFD.
- 68 2D landmarks: $M = 68, \mathbf{x} \in \mathbb{R}^{136=2\times 68}$.
- emotion representation: 2-dimensional valence-arousal ($\boldsymbol{\theta} \in \mathbb{R}^2$).
- $k_{\mathcal{X}}$, k_{Θ}: Gaussian kernels.

Quantitative illustration: setting

- 2 popular facial benchmarks: KDEF, RaFD.
- 68 2D landmarks: $M = 68, \mathbf{x} \in \mathbb{R}^{136=2\times 68}$.
- emotion representation: 2-dimensional valence-arousal ($\boldsymbol{\theta} \in \mathbb{R}^2$).
- $k_{\mathcal{X}}$, k_{Θ}: Gaussian kernels.
- Performance metrics:
  - test MSE – direct measure,
  - classification accuracy – indirect evaluation (ResNet-18 classifier; cross).

Quantitative illustration: setting

- 2 popular facial benchmarks: KDEF, RaFD.
- 68 2D landmarks: $M = 68, \mathbf{x} \in \mathbb{R}^{136=2\times 68}$.
- emotion representation: 2-dimensional valence-arousal ($\boldsymbol{\theta} \in \mathbb{R}^2$).
- $k_{\mathcal{X}}$, k_{Θ}: Gaussian kernels.
- Performance metrics:
  - test MSE – direct measure,
  - classification accuracy – indirect evaluation (ResNet-18 classifier; cross).
- Baseline: StarGAN
  - discrete emotion labels,
  - modified to handle landmarks (=: Landmark-StarGAN),
  - unstable training (as usual).

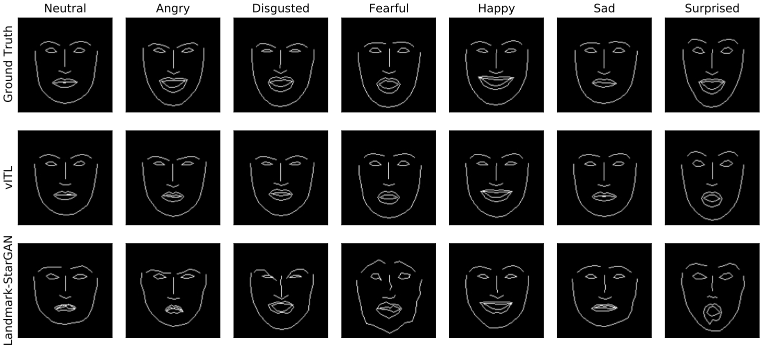
Quantitative illustration: mean $\pm$ std

Methods	MSE Error $\downarrow$		Emotion Classification Acc. $\uparrow$	
	KDEF frontal	RaFD frontal	KDEF frontal	RaFD frontal
vITL: θ_0 = neutral	0.010 ± 0.001	0.009 ± 0.004	76.12 ± 4.57	79.76 ± 7.88
vITL: θ_0 = fearful	0.010 ± 0.001	0.010 ± 0.005	76.22 ± 4.91	78.81 ± 8.36
vITL: θ_0 = angry	0.012 ± 0.002	0.010 ± 0.005	74.49 ± 2.31	78.10 ± 7.51
vITL: θ_0 = disgusted	0.012 ± 0.001	0.010 ± 0.004	74.18 ± 4.22	78.33 ± 4.12
vITL: θ_0 = happy	0.011 ± 0.001	0.010 ± 0.004	73.57 ± 2.74	80.48 ± 5.70
vITL: θ_0 = sad	0.011 ± 0.001	0.009 ± 0.004	75.82 ± 4.11	77.62 ± 5.17
vITL: θ_0 = surprised	0.010 ± 0.001	0.011 ± 0.006	74.69 ± 2.25	80.71 ± 5.99
vITL: Joint	$\mathbf{0.011} \pm 0.001$	$\mathbf{0.007} \pm 0.001$	$\mathbf{74.81} \pm 3.10$	$\mathbf{77.11} \pm 3.97$
Landmark-StarGAN	0.029 ± 0.003	0.024 ± 0.007	70.69 ± 8.46	65.88 ± 8.92

Both MSE and classification accuracy improve.

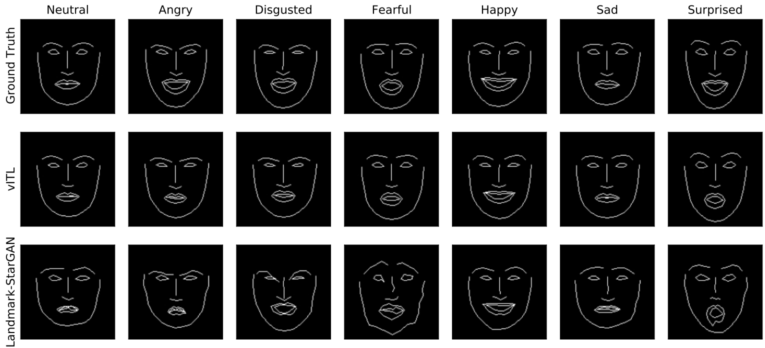
Qualitative illustration

Vs Landmark-StarGAN:

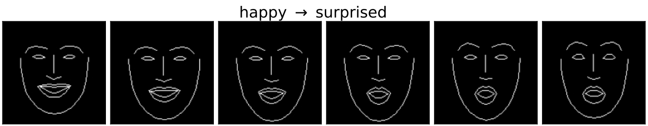


Qualitative illustration

Vs Landmark-StarGAN:



Continuous traversal by $\hat{h}$:



Summary

- We considered a new task: continuous style transfer.
- Formulation: functional output regression.
- Model:
  - vv-RKHS framework: $\mathcal{X} \mapsto (\Theta \mapsto \mathcal{Y})$,
  - general umbrella: similarity on object/style space.
- Application: emotion transfer.

Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., and Chen, B. (2020).
Unpaired motion style transfer from video to animation.
ACM Transactions on Graphics (TOG), 39(4):64.

Aristidou, A., Zeng, Q., Stavrakis, E., Yin, K., Cohen-Or, D., Chrysanthou, Y., and Chen, B. (2017).
Emotion control of unstructured dance movements.
In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–10.

Barricelli, B. R., Casiraghi, E., and Fogli, D. (2019).
A survey on digital twin: Definitions, characteristics, applications, and design implications.
IEEE Access, 7:167653–167671.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018).
StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation.

In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018).
Style transfer in text: Exploration and evaluation.
In *Conference on Artificial Intelligence (AAAI)*, pages 663–670.

Grinstein, E., Duong, N. Q., Ozerov, A., and Pérez, P. (2018).
Audio style transfer.
In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590.

Irimia, A.-S., Chan, J. C., Mistry, K., Wei, W., and Ho, E. S. (2019).
Emotion transfer for hand animation.
In *Motion, Interaction and Games*, pages 1–2.

Lim, K. Y. H., Zheng, P., and Che, C.-H. (2020).
A state-of-the-art survey of digital twin: techniques, engineering product lifecycle management and business innovation perspectives.

Journal of Intelligent Manufacturing, 31:1313–1337.

Puy, G. and Pérez, P. (2019).
A flexible convolutional solver for fast style transfers.
In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8963–8972.

Russell, J. A. (1980).
A circumplex model of affect.
Journal of Personality and Social Psychology, 39(6):1161–1178.

Tao, F., Zhang, H., Liu, A., and Nee, A. Y. C. (2019).
Digital twin in industry: State-of-the-art.
IEEE Transactions on Industrial Informatics, 15(4):2405 – 2415.

Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. (2016).
Texture networks: Feed-forward synthesis of textures and stylized images.

In *International Conference on Machine Learning (ICML)*, pages 1349–1357.

Vemulapalli, R. and Agarwala, A. (2019).
A compact embedding for facial expression similarity.
In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5683–5692.

Yao, X., Puy, G., Newson, A., Gousseau, Y., and Hellier, P. (2020).
High resolution face age editing.
In *International Conference on Pattern Recognition (ICPR)*.