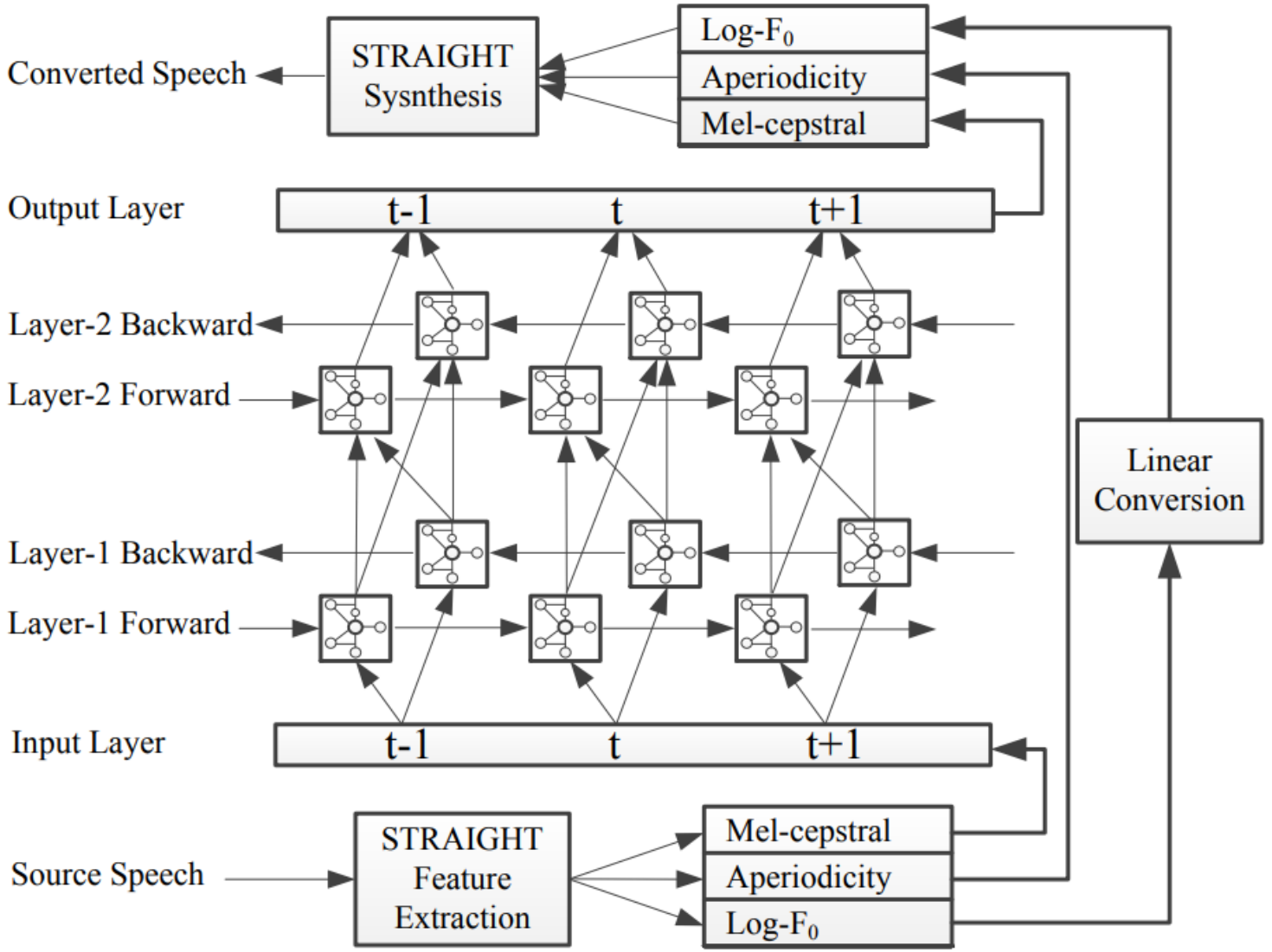
GANs for Voice Conversion

Karan Taneja

Indian Institute of Technology Bombay, Mumbai

VC using deep BLSTM based RNN

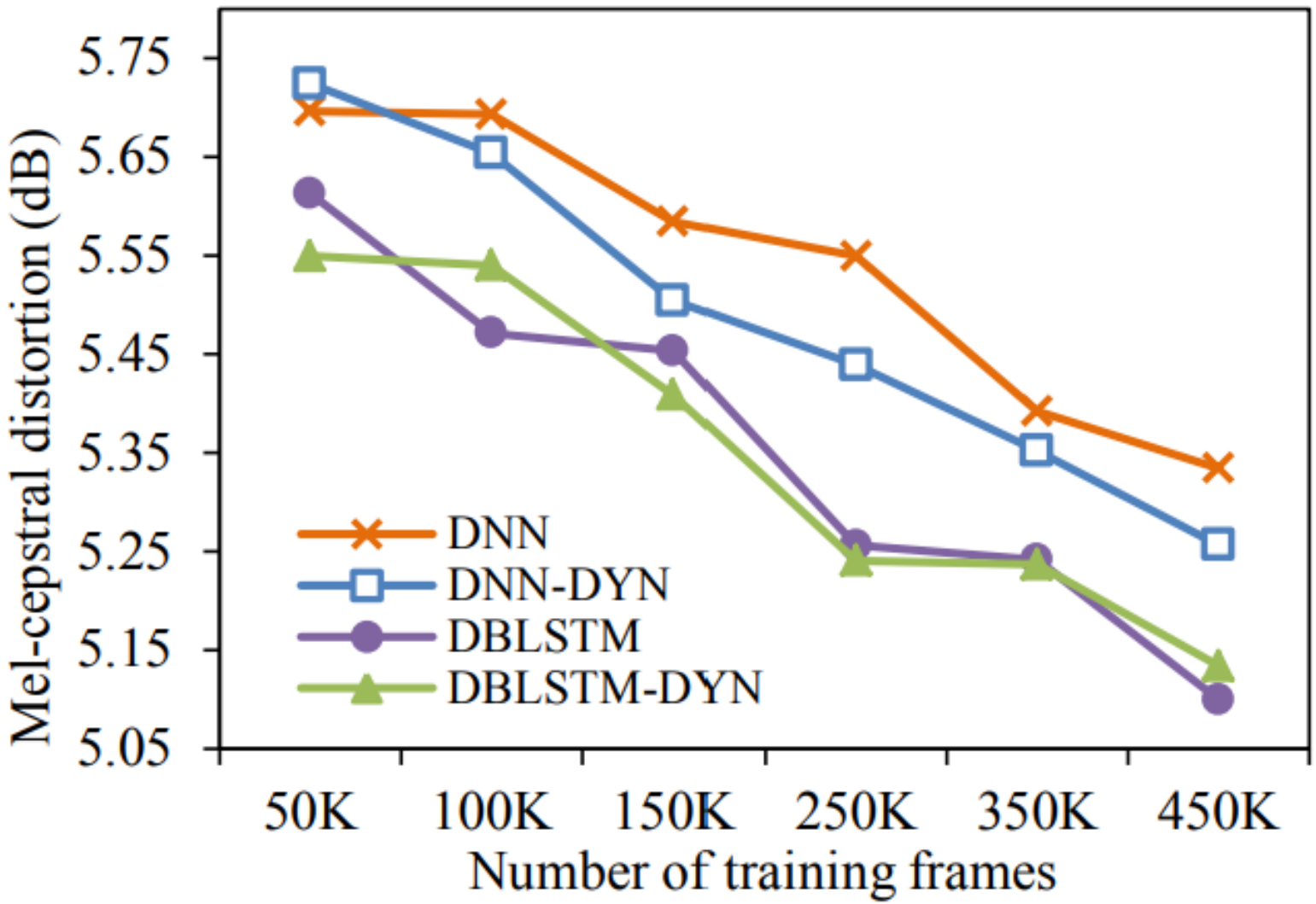
Lifa Sun et al., ICASSP 2015

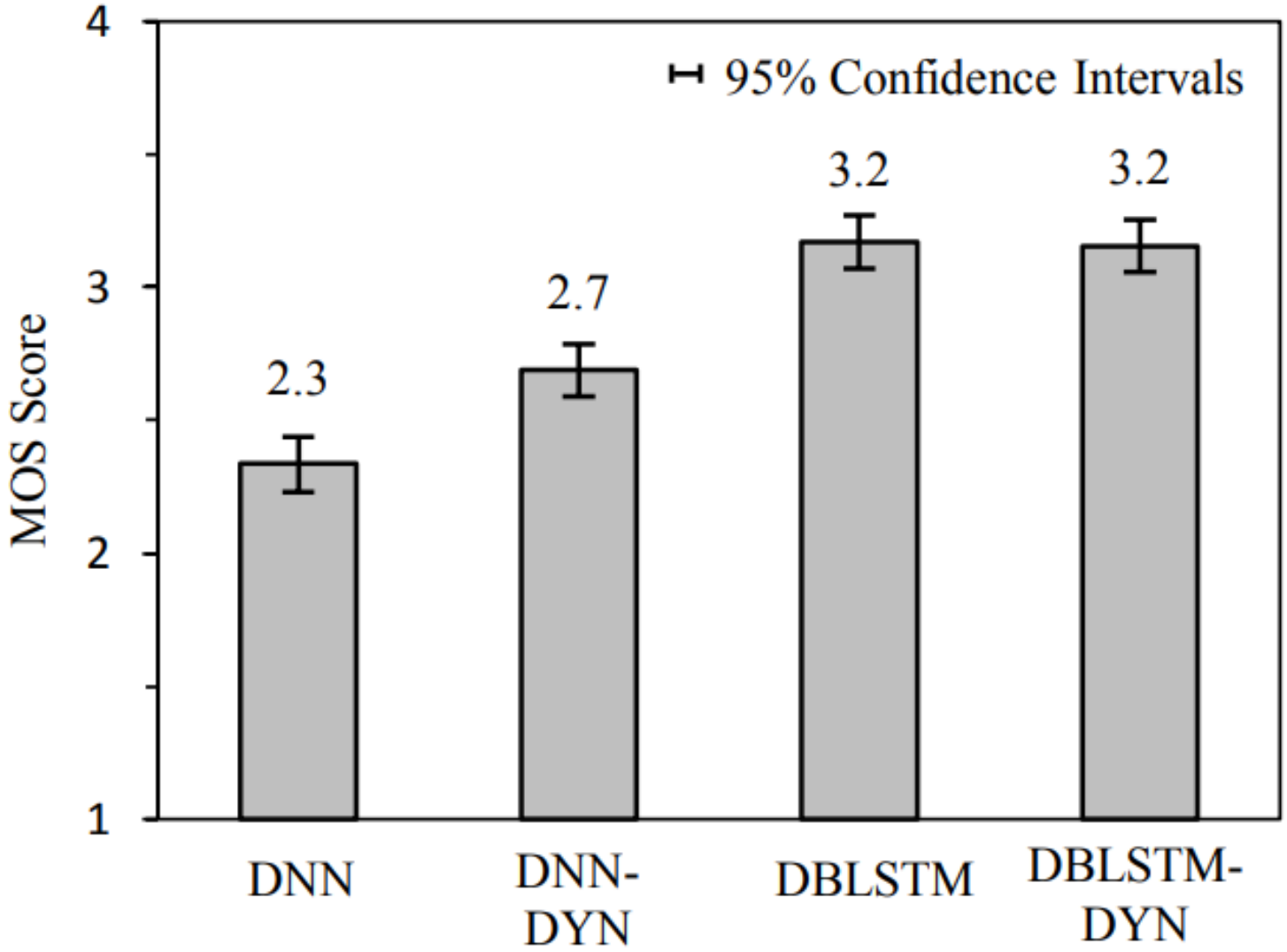


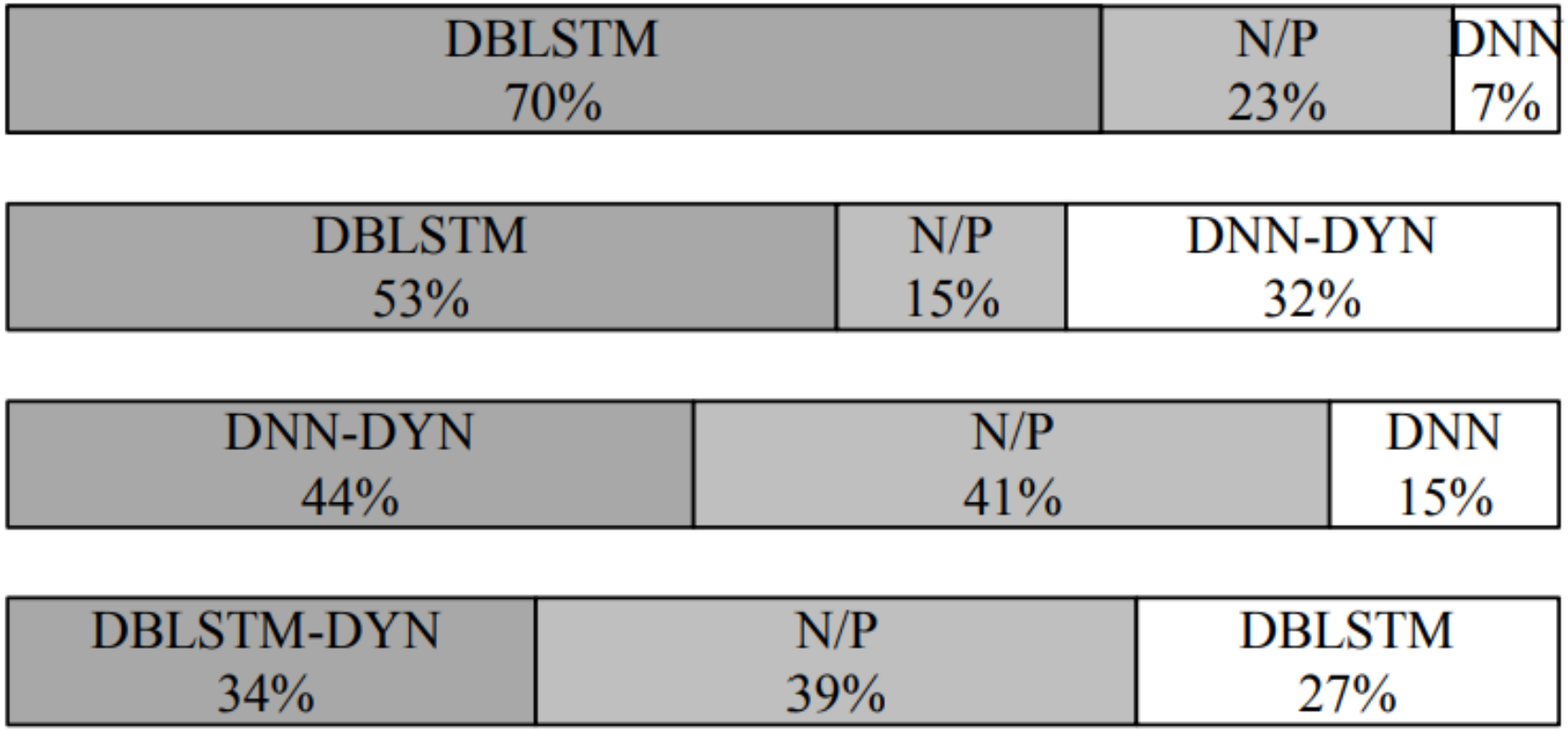
https://ieeexplore.ieee.org/document/7178896

- To get the parallel utterances, the dynamic time warping (DTW) algorithm is used to align the features sequences of the source and target speaker.
- DYN: dynamic features i.e. delta and double-delta features.
- Training data: 42 minutes, validation: 9 minutes
- Mel-cepstral distortion (MCD):

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{N}(c_d - c_d^{converted})^2}$$


Mel-cepstral distortion (dB)
5.75
5.65
5.55
5.45
5.35
5.25
5.15
5.05
DNN
DNN-DYN
DBLSTM
DBLSTM-DYN
50K
100K
150K
250K
350K
450K
Number of training frames

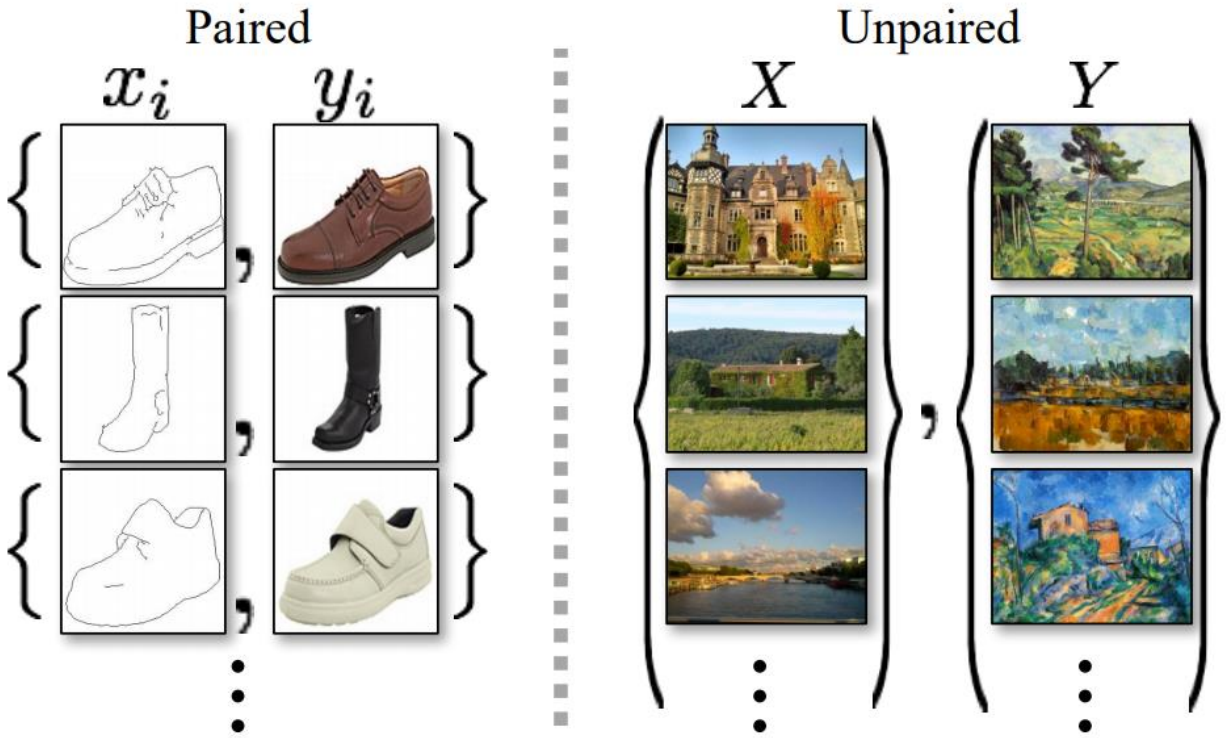

4
3
2
1
MOS Score
95% Confidence Intervals
2.3
2.7
3.2
3.2
DNN
DNN-DYN
DBLSTM
DBLSTM-DYN



ABX preference test results. The p-values of the four pairs are 3.9×10^{-24}, 3.0×10^{-3} , 2.4×10^{-4} and 0.24 respectively

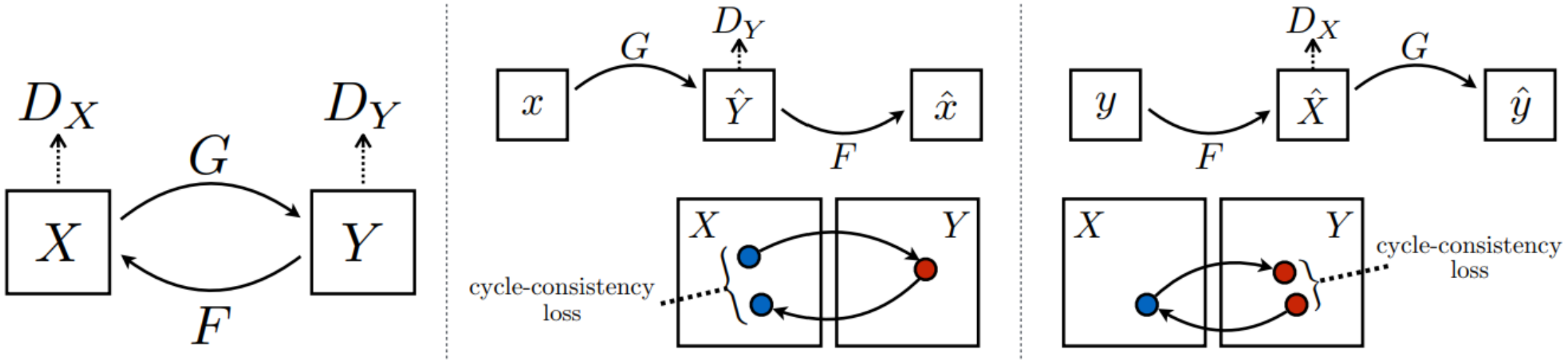
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu et al. at BAIR UCB, ICCV 2017



Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu et al. at BAIR UCB, ICCV 2017



Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

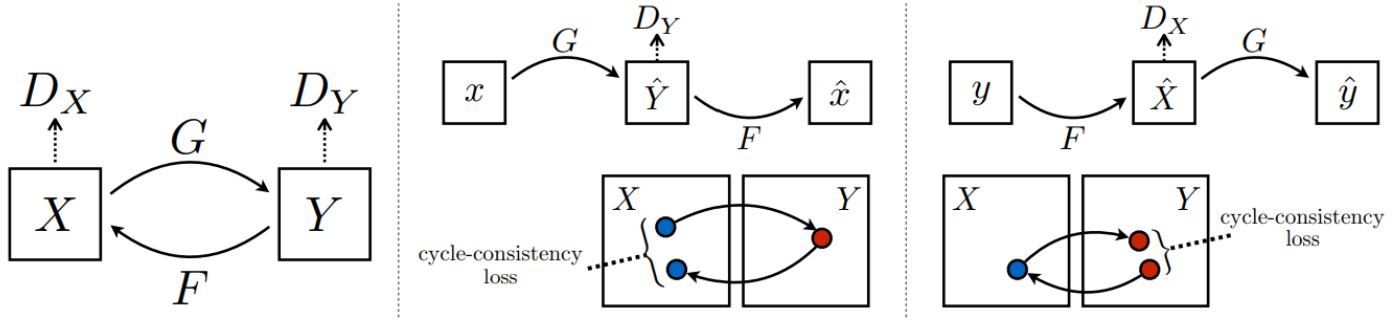
Jun-Yan Zhu et al. at BAIR UCB, ICCV 2017

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]$$

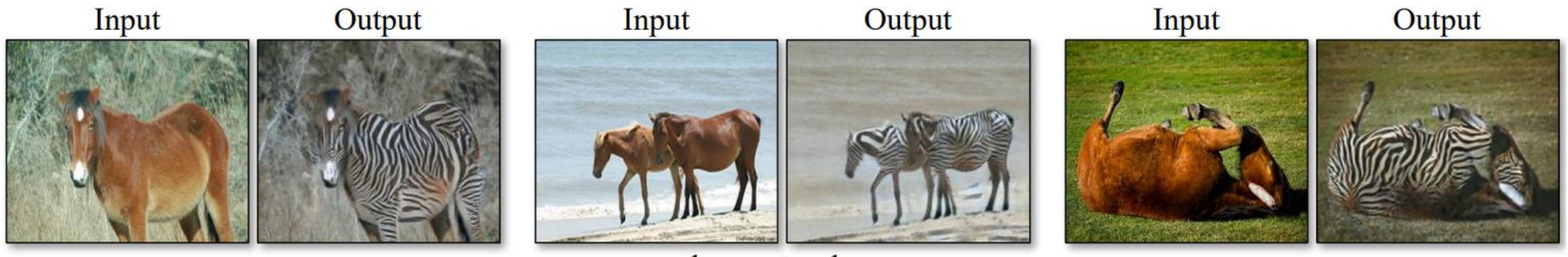
$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]$$

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F)$$

$$G^*, F^* = \arg \min_{G,F} \max_{D_x, D_Y} \mathcal{L}(G, F, D_X, D_Y)$$

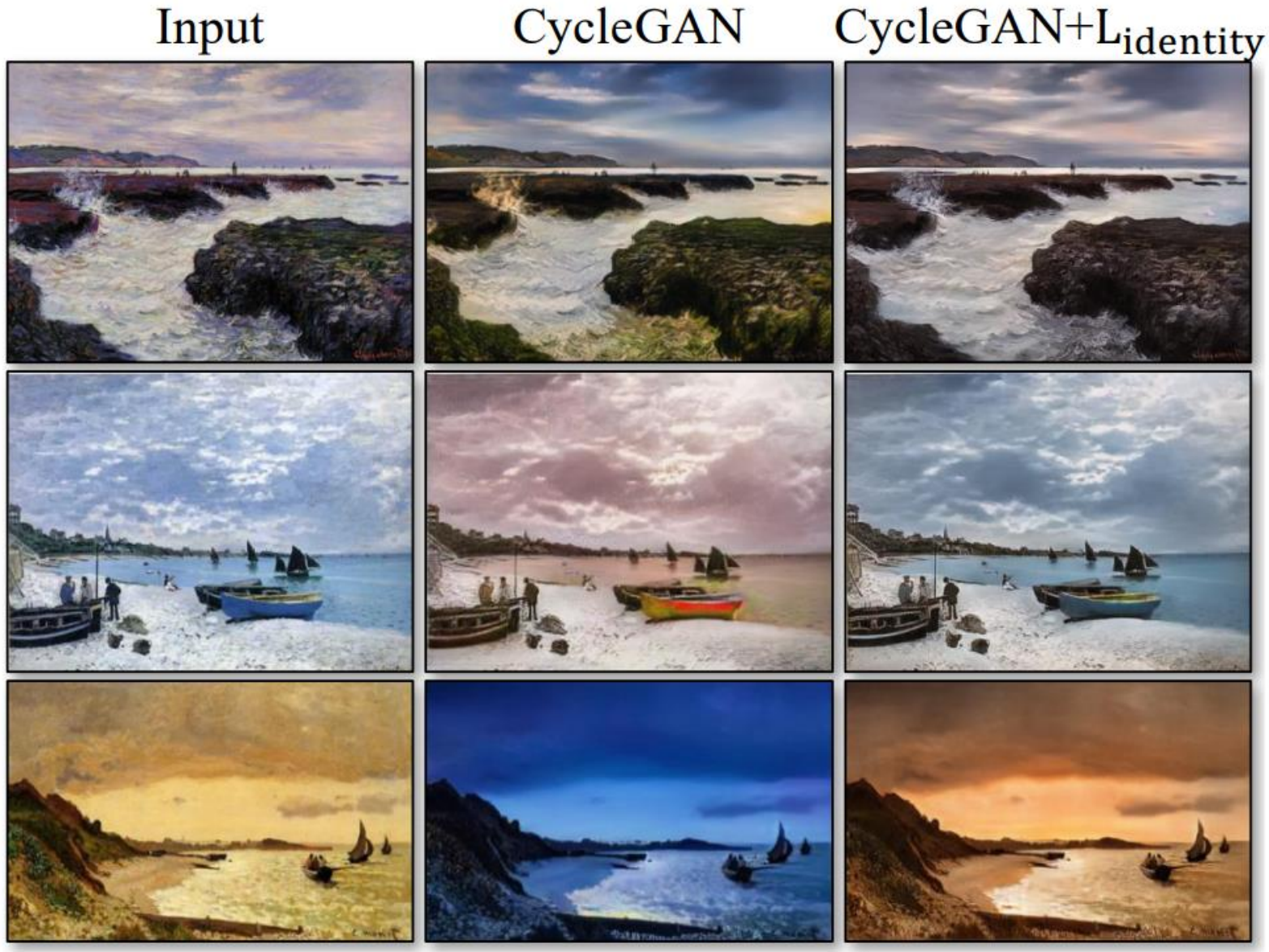



Input
Monet
Van Gogh
Cezanne
Ukiyo-e


Input
Output
Input
Output
Input
Output
horse → zebra
zebra → horse

winter Yosemite → summer Yosemite

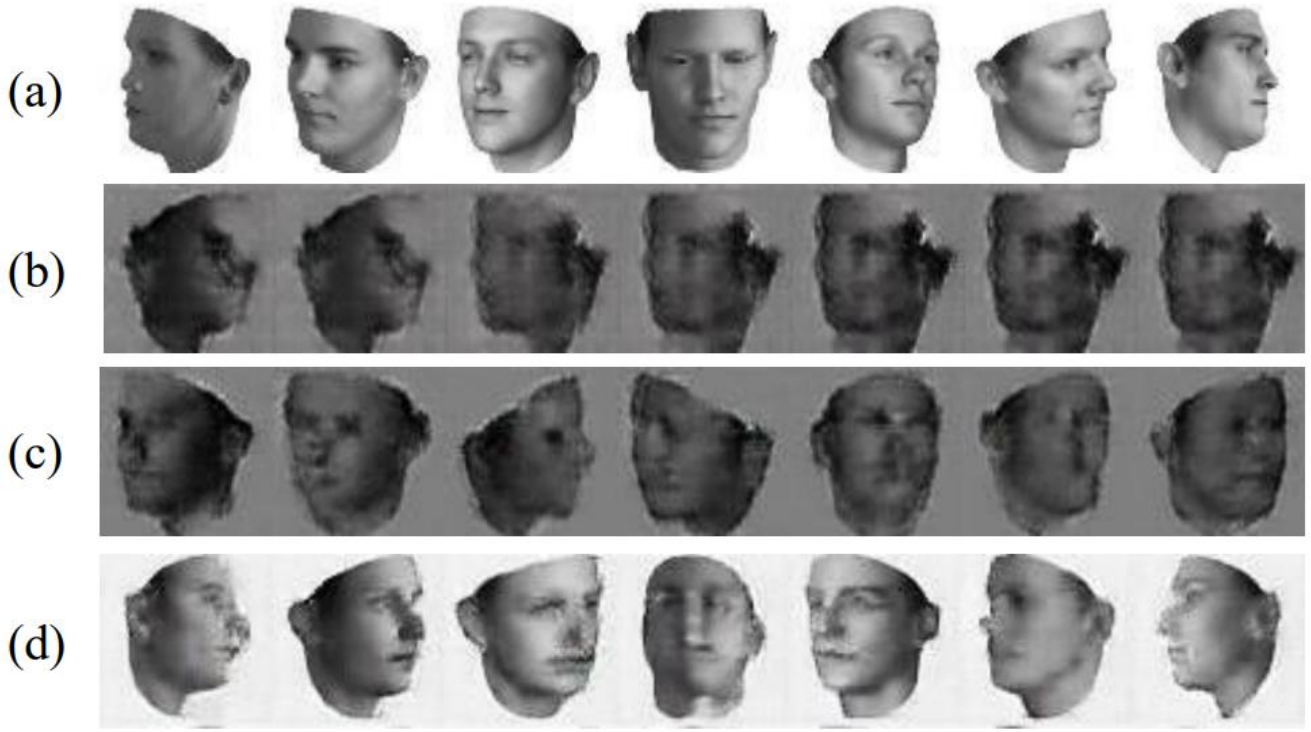
summer Yosemite → winter Yosemite



$$\mathcal{L}_{\text{identity}}(G, F) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(y) - y\|_1] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(x) - x\|_1]$$

Learning to Discover Cross-Domain Relations with GANs

Taeksoo Kim et al., ICC PMLR 2017



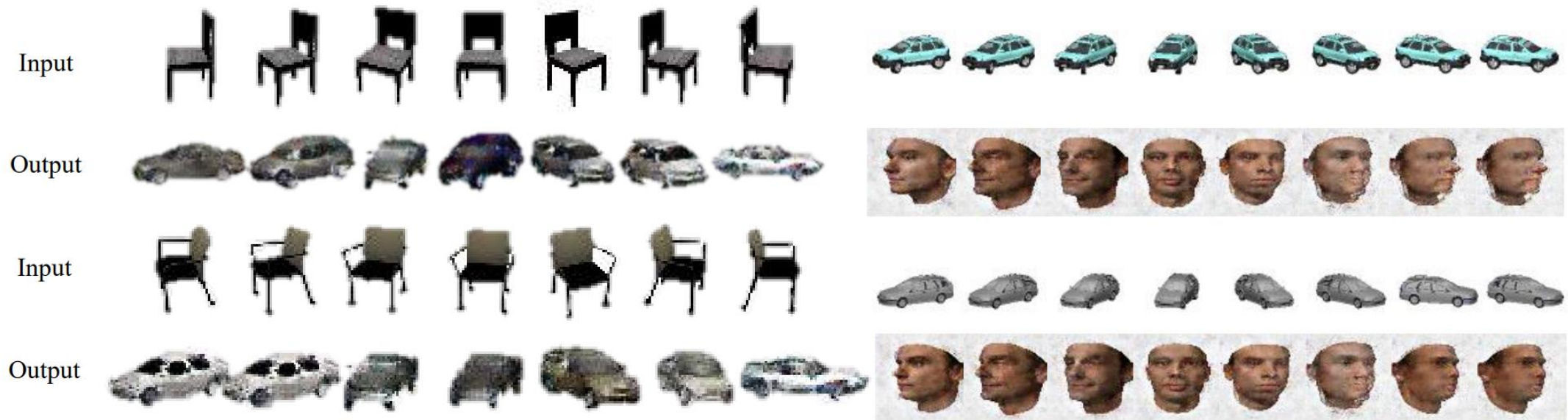
Face to Face translation experiment.

(a) input face images from -90◦ to +90◦

(b) results from a standard GAN

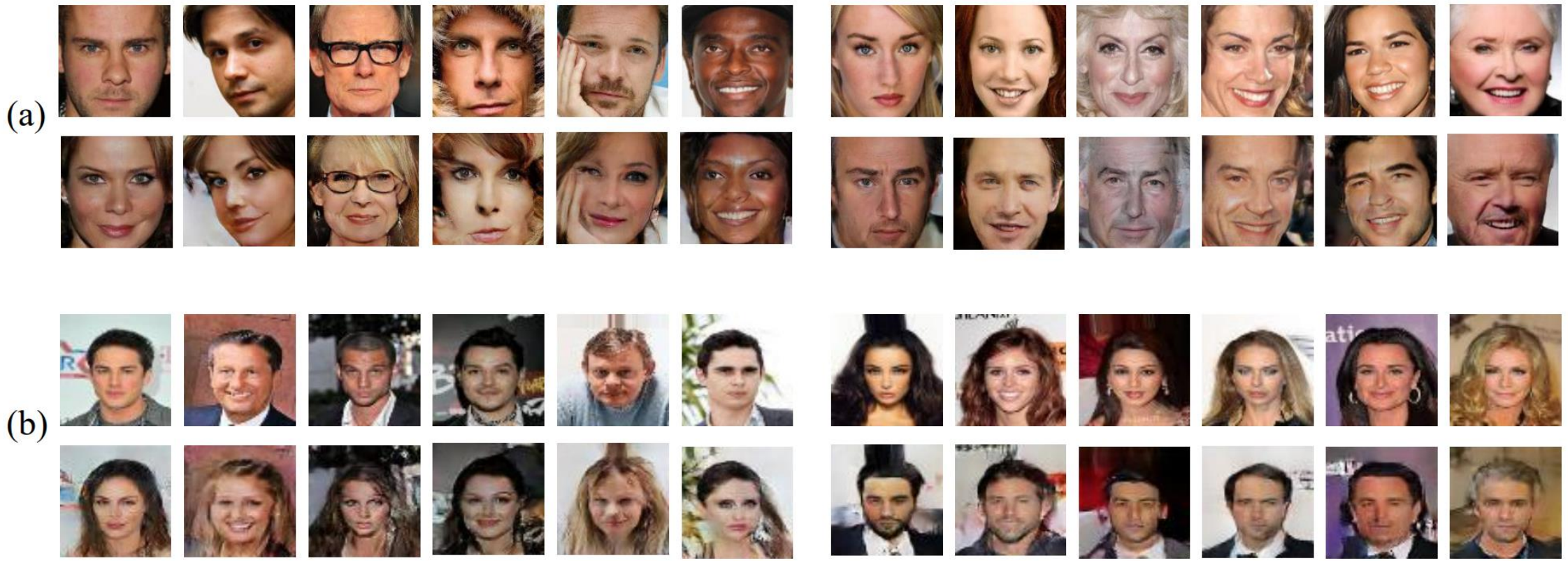
(c) results from GAN with a reconstruction loss

(d) results from our DiscoGAN. Here model generated images in the opposite range, from +90◦ to -90◦ .



(a) Chair to Car

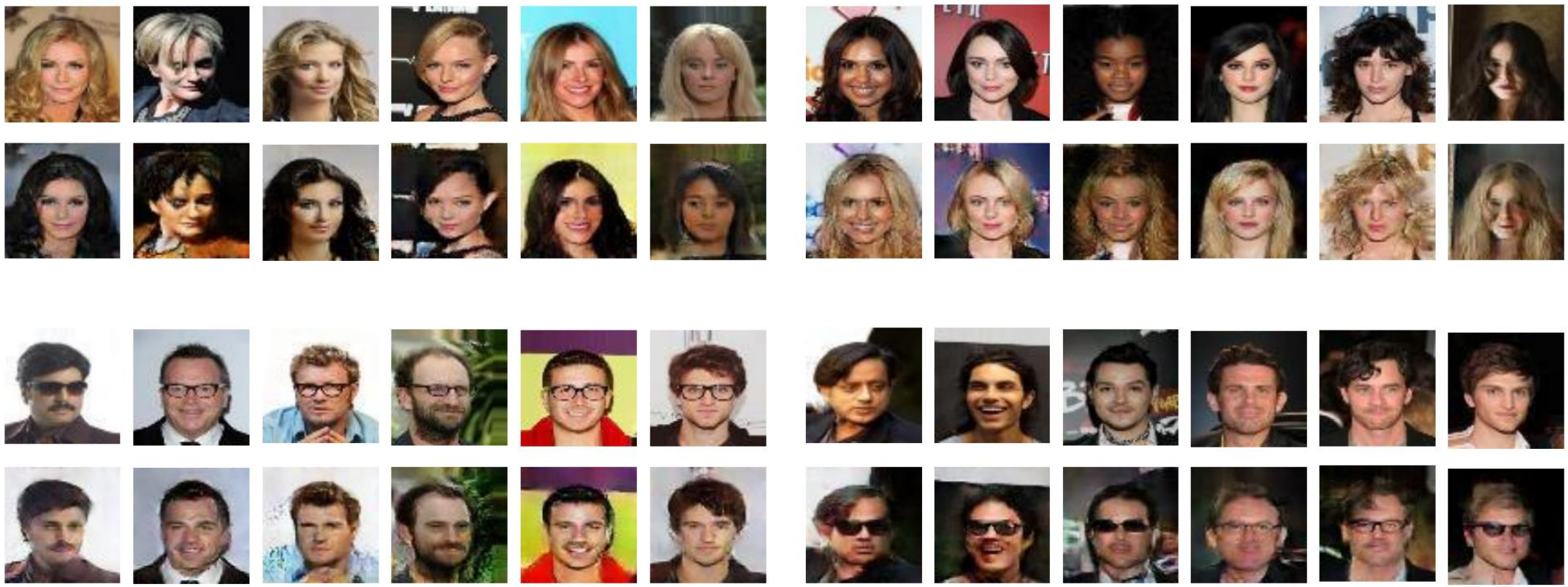
(b) Car to Face



Translation of gender in Facescrub dataset and CelebA dataset.

(c)

(d)



(c) Blond to black and black to blond hair color conversion in CelebA dataset.
(d) Wearing eyeglasses conversion in CelebA dataset

Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks

Kaneko and Kameoka at NTT, arXiv (Dec. 2017)

$$\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})$$

Gated CNN:

$$\boldsymbol{H}_{l+1} = (\boldsymbol{H}_l * \boldsymbol{W}_l + \boldsymbol{b}_l) \otimes \sigma(\boldsymbol{H}_l * \boldsymbol{V}_l + \boldsymbol{c}_l)$$

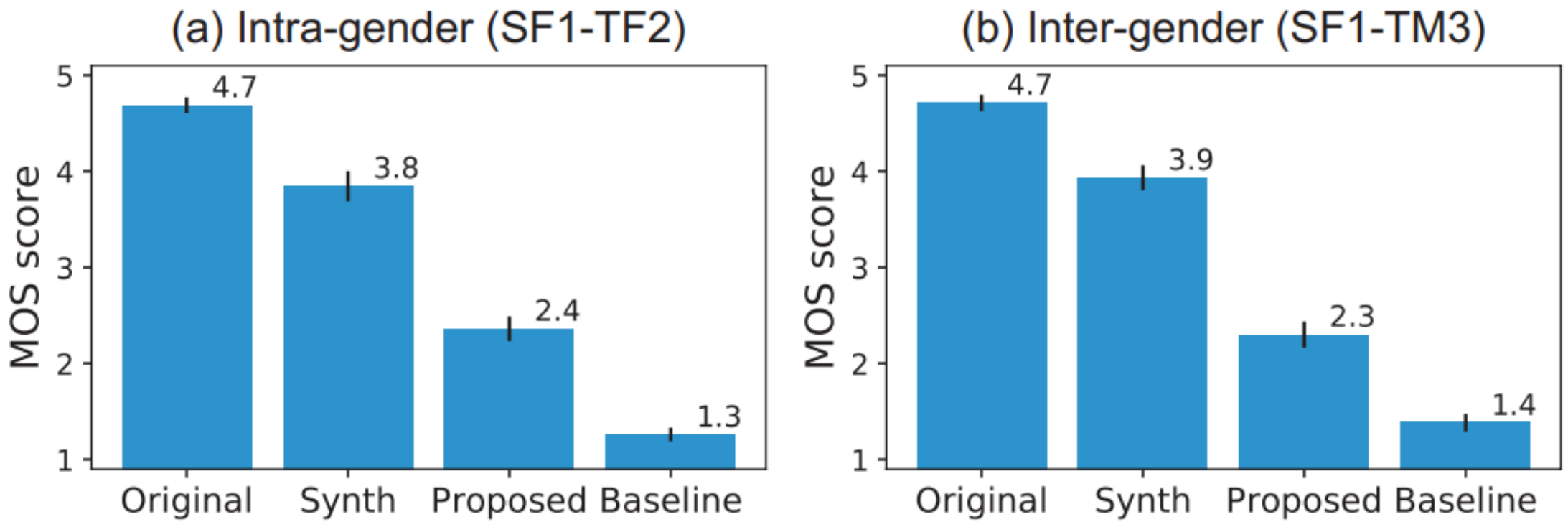
Identity-mapping loss:

$$\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) = \mathbb{E}_{y\sim P_{\mathrm{Data}}(y)}[||G_{X\to Y}(y) - y||_1] + \mathbb{E}_{x\sim P_{\mathrm{Data}}(x)}[||G_{Y\to X}(x) - x||_1]$$

Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks

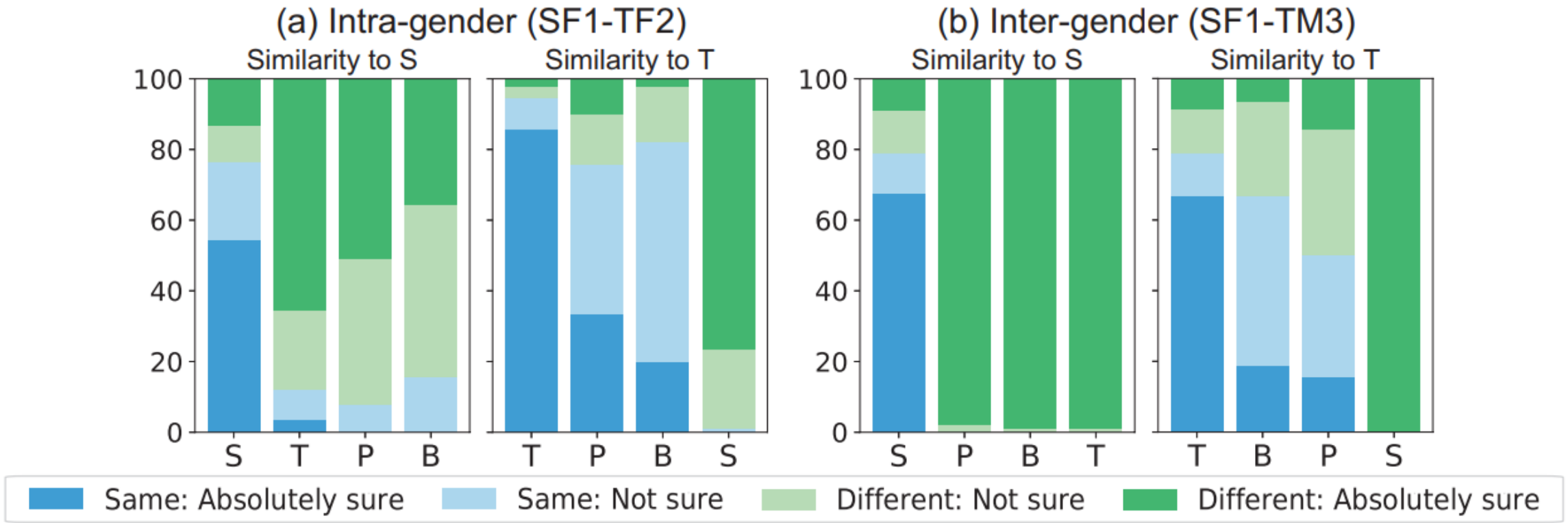
Kaneko and Kameoka at NTT, arXiv (Dec. 2017)

- For features, they used 24 Mel-cepstrcum coefficients, log of fundamental freqeuncy and aperiodicities every 5ms using WORLD analysis system.
- It should be noted that their baseline, which is a GMM-based method, uses parallel data and twice the amount of data that they use.



MOS for naturalness with 95% confidence intervals

Demo at http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/

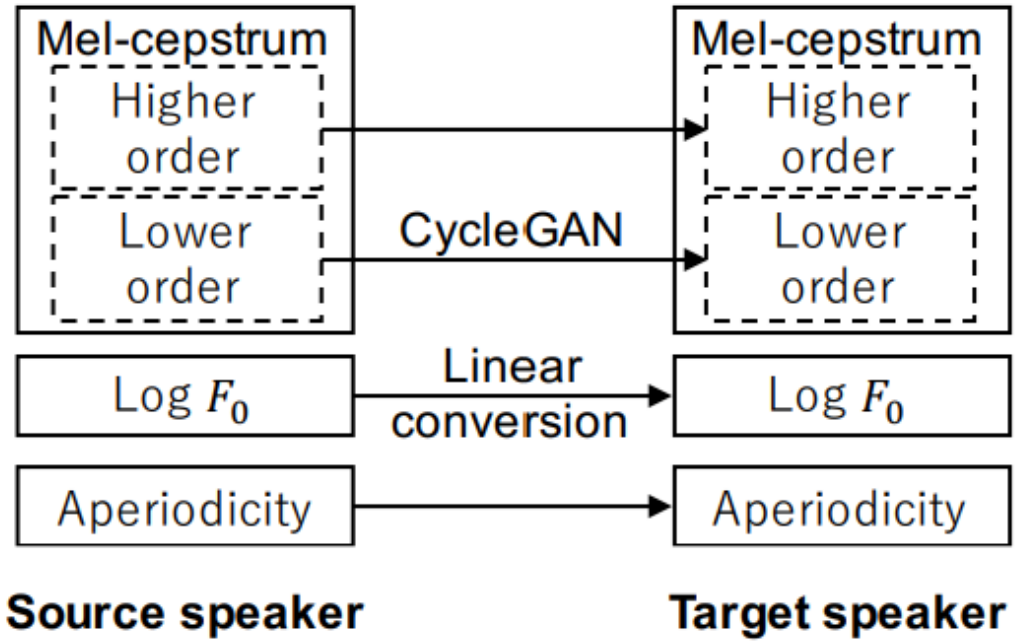


Similarity to source speaker and to target speaker (S: Source, T: Target, P: Proposed, and B: Baseline)

Demo at http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/

Another very similar paper

- High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network, Fuming Fang et al., ICASSP 2018
  - No contraints to preserve linguistic information though.



METHOD		QUALITY	SIMILARITY
CycleGAN (Nonparallel VC)	F→M	2.89	2.53
	M→F	2.50	1.76
	AVG.	**2.69**	**2.15**
GAN (Parallel VC)	F→M	2.20	2.26
	M→F	2.51	1.79
	AVG.	**2.36**	**2.02**
Merlin-based baseline (Parallel VC)	F→M	1.37	1.52
	M→F	1.52	1.38
	AVG.	**1.45**	**1.45**

Perceptual evaluation (MOS scale) for speech quality and speaker similarity. “CycleGAN” denotes proposed nonparallel VC method and “GAN” and “Merlin-based baseline” (DNN based) denote the two baseline methods based on parallel VC.

StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks

Hirokazu Kameoka et al. at NTT, IEEE SLT 2018

Adversarial Loss:

$$\mathcal{L}_{\text{adv}}^{D}(D) = -\mathbb{E}_{c\sim p(c),\mathbf{y}\sim p(\mathbf{y}|c)}[\log D(\mathbf{y},c)] - \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),c\sim p(c)}[\log(1-D(G(\mathbf{x},c),c))]$$

$$\mathcal{L}_{\text{adv}}^{G}(G) = -\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),c\sim p(c)}[\log D(G(\mathbf{x},c),c)]$$

Cycle Consistency Loss:

$$\mathcal{L}_{\text{cyc}}(G) = \mathbb{E}_{c'\sim p(c),\mathbf{x}\sim p(\mathbf{x}|c'),c\sim p(c)}[\|G(G(\mathbf{x},c),c')-\mathbf{x}\|_{\rho}]$$

StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks

Hirokazu Kameoka et al. at NTT, IEEE SLT 2018

Domain Classification Loss:

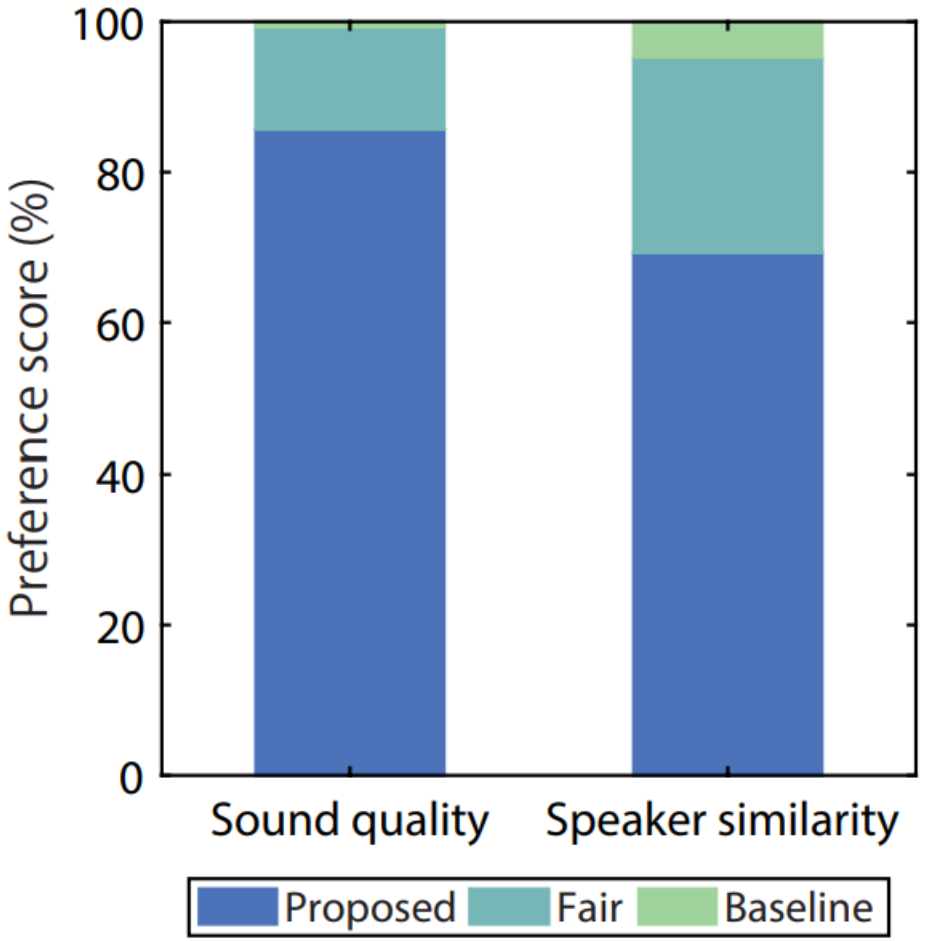
$$\mathcal{L}_{\mathrm{cls}}^{C}(C) = -\,\mathbb{E}_{c\sim p(c),\mathbf{y}\sim p(\mathbf{y}|c)}[\log p_C(c|\mathbf{y})]$$
$$\mathcal{L}_{\mathrm{cls}}^{G}(G) = -\,\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),c\sim p(c)}[\log p_C(c|G(\mathbf{x},c))]$$

full objectives of StarGAN-VC

$$\begin{aligned}\mathcal{I}_G(G) =& \mathcal{L}_{\mathrm{adv}}^{G}(G) + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}^{G}(G) \\ &+ \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}}(G) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}(G)\\ \mathcal{I}_D(D) =& \mathcal{L}_{\mathrm{adv}}^{D}(D)\\ \mathcal{I}_C(C) =& \mathcal{L}_{\mathrm{cls}}^{C}(C)\end{aligned}$$

Details

- Generator architecture inspired from gated CNN architectures for voice conversion and audio source separation, for which effectiveness has already been confirmed.
- Used PatchGANs to devise a real/fake discriminator D, which classifies whether local segments of an input feature sequence are real or fake. The final output is given by the product of all these probabilities.
- Also devise a domain classifier C using a gated CNN, which takes an acoustic feature sequence y and produces a sequence of class probability distributions. The final output is given by the product of all these distributions.



Results of the AB test for sound quality and the ABX test for speaker similarity.

Demo at http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/

Another similar paper

- Voice Impersonation Using Generative Adversarial Networks, Yang Gao et al. at CMU ECE, ICASSP 2018
  - Only major difference is that they have used gender as their only domain.

Other work using Cycle-Consistent GANs in Speech Processing

- Cycle-Consistent Speech Enhancement, Zhong Meng et al. at MS Redmond and GaTech, Interspeech 2018
  - Speech enhancement, two generators remove and add noise.
  - Good thing: no need of parallel noisy and clean utterances

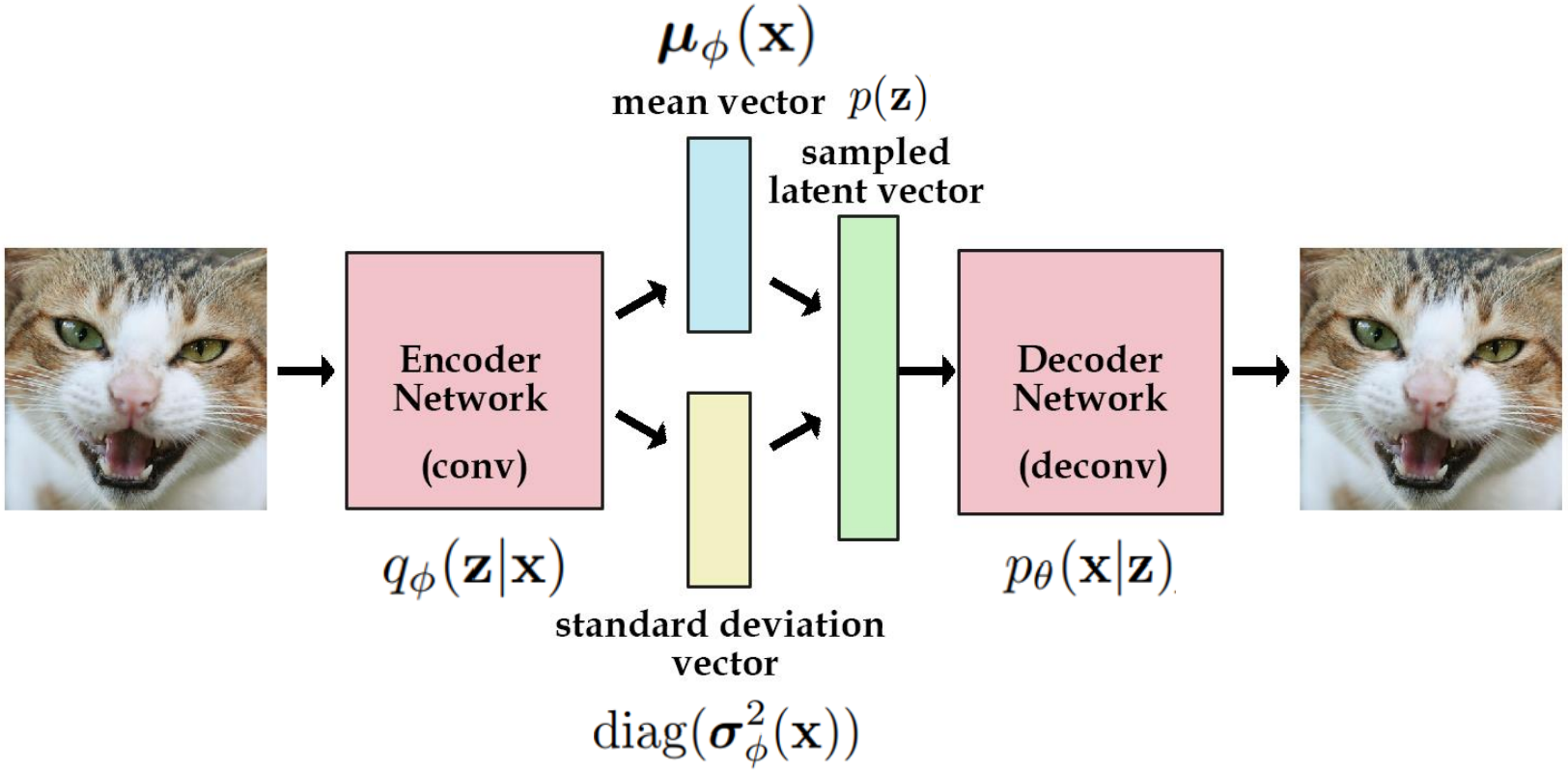
- Previous work by same authors: Adversarial Feature-Mapping for Speech Enhancement, Zhong Meng et al. at MS Redmond and GaTech, Interspeech 2018
  - Here they used simple GAN based model for parallel speech enhancement

Other work using Cycle-Consistent GANs in Speech Processing

- WaveCycleGAN: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks, Kou Tanaka et al. at NTT, arXiv (Sept. 2018)
  - A general additional module to be used after VOCODER in speech generating systems to make the speech sound more natural.

ACVAE-VC: Non-parallel Many-to-many VC with Auxiliary Classifier VAE

Hirokazu Kameoka et al. at NTT, arXiv (Aug. 2018)



ACVAE-VC: Non-parallel Many-to-many VC with Auxiliary Classifier VAE

Hirokazu Kameoka et al. at NTT, arXiv (Aug. 2018)

$$\log p_\theta(\mathbf{x}) = \log \int q_\phi(\mathbf{z}|\mathbf{x}) \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

$$= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})]$$

$\mathrm{KL}[q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x})]$ is minimized when $q_\phi(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{z}|\mathbf{x})$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x})))$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{z}), \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z})))$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

ACVAE-VC: Non-parallel Many-to-many VC with Auxiliary Classifier VAE

Hirokazu Kameoka et al. at NTT, arXiv (Aug. 2018)

$$q_\phi(\mathbf{z}|\mathbf{x}, c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}, c), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x}, c)))$$

$$p_\theta(\mathbf{x}|\mathbf{z}, c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\theta(\mathbf{z}, c), \mathrm{diag}(\boldsymbol{\sigma}^2_\theta(\mathbf{z}, c)))$$

$$\mathcal{J}(\phi, \theta) = \mathbb{E}_{(\mathbf{x},c)\sim p_\mathrm{D}(\mathbf{x},c)}\left[\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x},c)}[\log p(\mathbf{x}|\mathbf{z}, c)] - \mathrm{KL}[q(\mathbf{z}|\mathbf{x}, c)\|p(\mathbf{z})]\right]$$

ACVAE-VC: Non-parallel Many-to-many VC with Auxiliary Classifier VAE

Hirokazu Kameoka et al. at NTT, arXiv (Aug. 2018)

- How to ensure that model doesn't ignore 'c'?
  - Penalize the mutual information between output and 'c'.

$$I(c, \mathbf{x}|\mathbf{z}) = \mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c),c'\sim p(c|\mathbf{x})}[\log p(c'|\mathbf{x})] + H(c)$$

$$\begin{aligned}
&\mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c),c'\sim p(c|\mathbf{x})}[\log p(c'|\mathbf{x})] \\
=&\mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c),c'\sim p(c|\mathbf{x})}\left[\log \frac{r(c'|\mathbf{x})p(c'|\mathbf{x})}{r(c'|\mathbf{x})}\right] \\
\geq&\mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c),c'\sim p(c|\mathbf{x})}[\log r(c'|\mathbf{x})] \\
=&\mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c)}[\log r(c|\mathbf{x})].
\end{aligned}$$

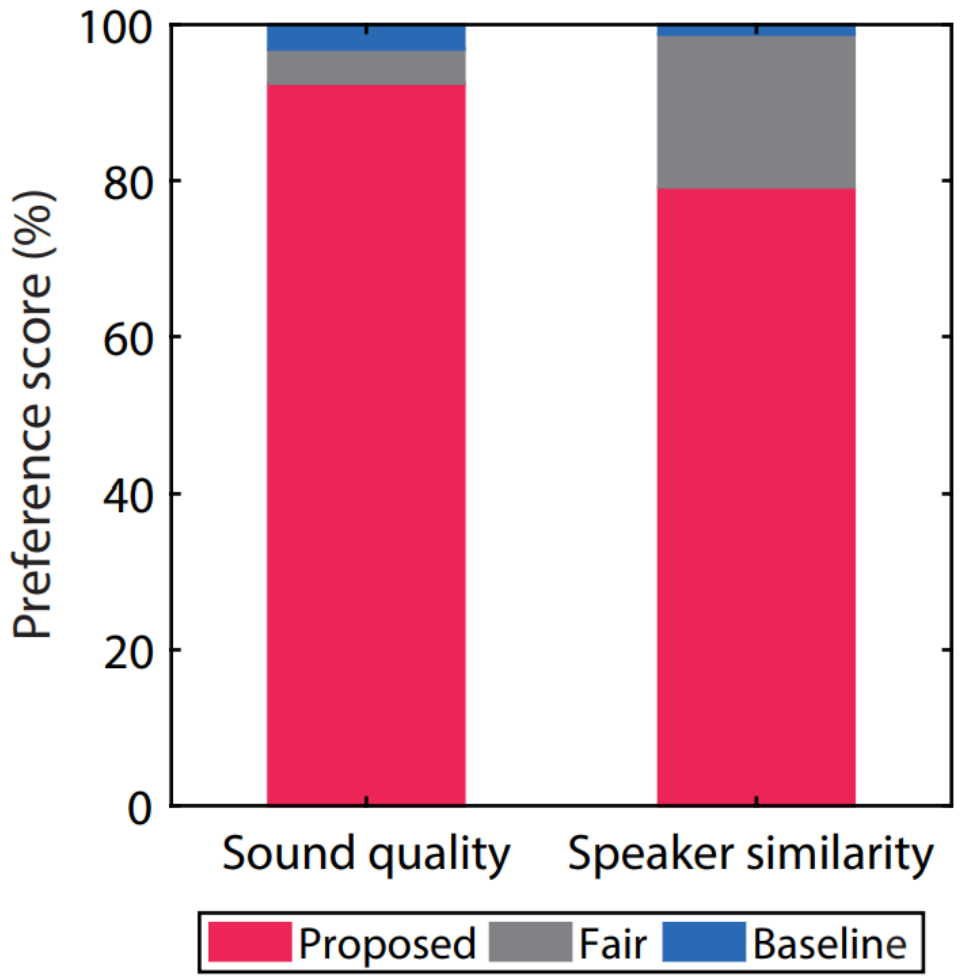
ACVAE-VC: Non-parallel Many-to-many VC with Auxiliary Classifier VAE

Hirokazu Kameoka et al. at NTT, arXiv (Aug. 2018)

$$\mathcal{L}(\phi,\theta,\psi) = \mathbb{E}_{(\tilde{c},\tilde{x})\sim p_D(\tilde{\mathbf{x}},\tilde{c}),q_\phi(\mathbf{z}|\tilde{\mathbf{x}},\tilde{c})}\left[\mathbb{E}_{c\sim p(c),\mathbf{x}\sim p_\theta(\mathbf{x}|\mathbf{z},c)}[\log r_\psi(c|\mathbf{x})]\right]$$

We can also use the training examples $\mathcal{I}(\psi) = \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{c})\sim p_D(\tilde{\mathbf{x}},\tilde{c})}[\log r_\psi(\tilde{c}|\tilde{\mathbf{x}})]$

The entire training criterion is $\mathcal{J}(\phi,\theta) + \lambda_{\mathcal{L}}\mathcal{L}(\phi,\theta,\psi) + \lambda_{\mathcal{I}}\mathcal{I}(\psi)$

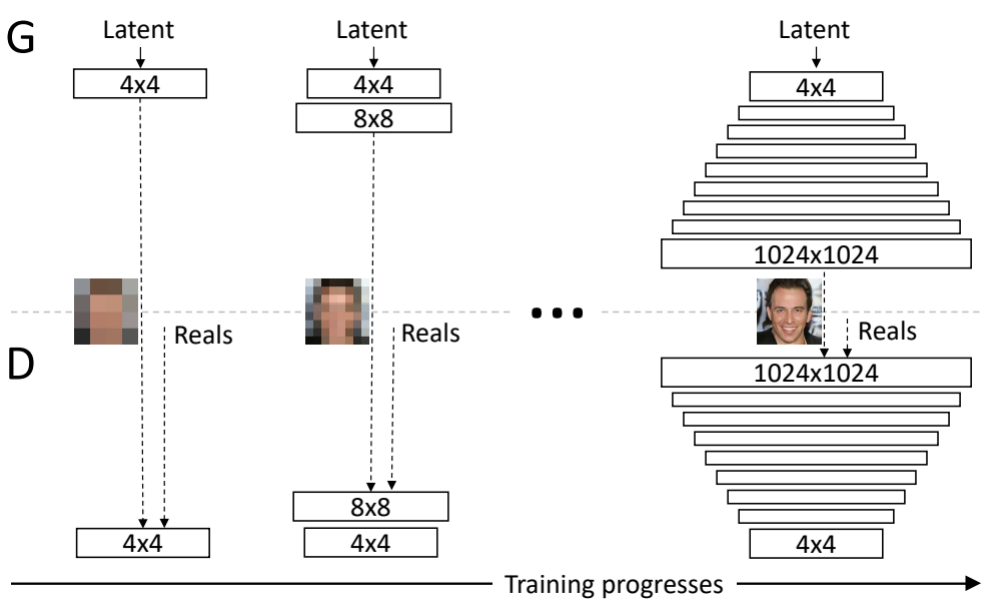


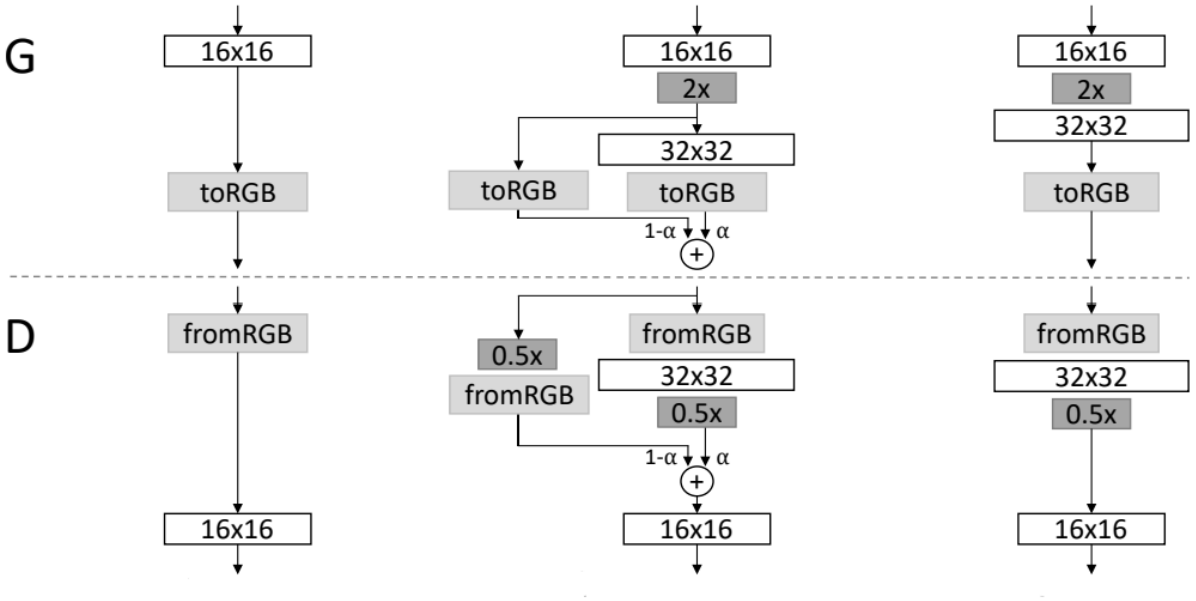
Results of the AB test for sound quality and the ABX test for speaker similarity.

Demo at http://www.kecl.ntt.co.jp/people/kameoka.hirokazu/Demos/

Some other interesting papers on GANs

- Progressive Growing of GANs for Improved Quality, Stability and Variation, Tero Karras et al., ICLR 2018
  - Increase the layers over time and improve the resolution as you increase layers.
  - Fade in the new layers, don't let them affect the model suddenly!





Some other interesting papers on GANs

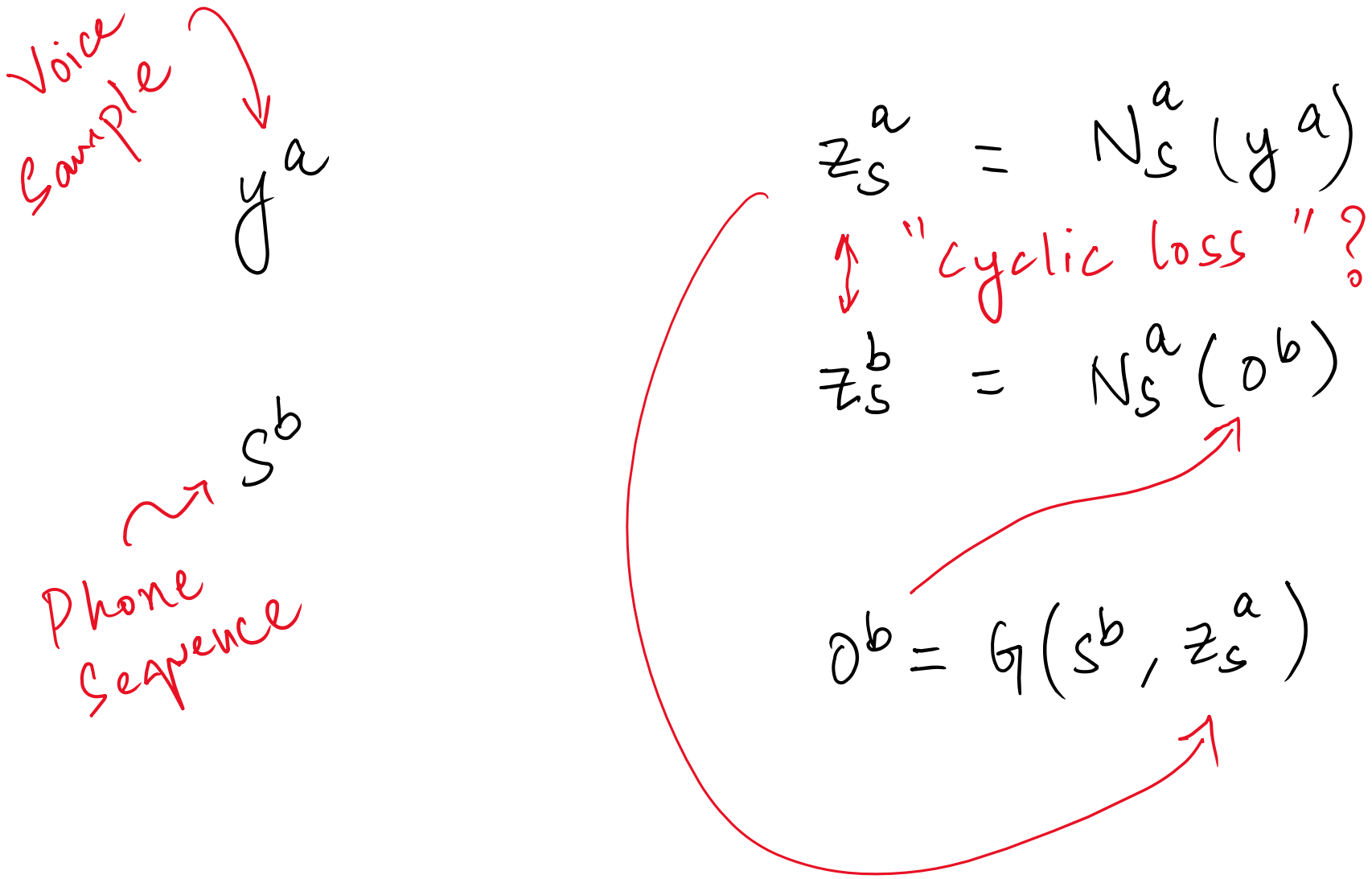
- Least Squares Generative Adversarial Networks, Xudong Mao et al., ICCV 2017
  - Original GAN paper proposes cross-entropy function for training. Minimizing the generator loss minimizes the Jensen-Shannon divergence between the output distribution from generator and real distribution of data.
  - Instead, this paper suggests using squared error on discriminator output. This is equivalent to minimizing the Pearson χ^2 divergence between $p_d + p_g$ and $2p_g$.

$$\min_D V_{\mathrm{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x})-1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}))+1)^2\right]$$

$$\min_G V_{\mathrm{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z})))^2\right]$$

Some other interesting papers on GANs

- Least Squares Generative Adversarial Networks, Xudong Mao et al., ICCV 2017
  - They show that LSGANs generate higher quality images than regular GANs on several scene datasets, and on a handwritten Chinese character dataset.
  - Besides, they demonstrate that LSGANs are more stable than regular GANs during the learning process.


Voice Sample
y^a
s^b
Phone Sequence
$z_s^a = N_s^a(y^a)$
"cyclic loss"?
$z_s^b = N_s^a(o^b)$
$o^b = G(s^b, z_s^a)$

The End