GAN評価方法と
アニメーション生成

5/10/2019

酒井一徳

論文紹介

- 評価方法
  - IS
  - m-IS
  - FID
  - SWD
  - MMD
  - SSIM
- 動画像生成
  - vid2vid
  - everybody dance now
  - cartoonGAN

評価方法

望ましい Generative Model とは

- データセットに忠実かつ多様な生成。
- 人にわかりやすい特徴表現。
- 歪み、攻撃に敏感。
- 計算が少ない。

Inception Score (IS)

- ImageNet で学習済みの Inception Net を用いた指標。
  - 大きいほど良い。

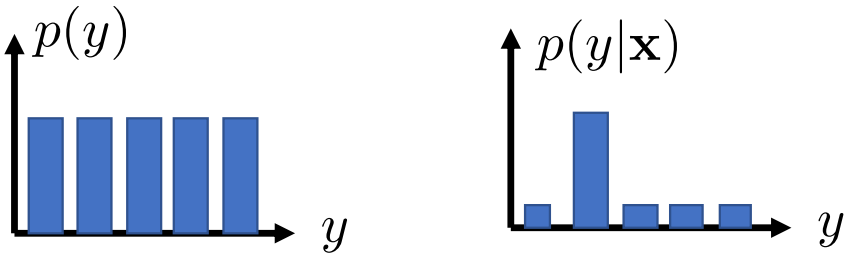
$$\exp\left(\mathbb{E}_{\mathbf{x}}[\mathbb{KL}(p(\mathrm{y}|\mathbf{x})\|p(\mathrm{y}))]\right) = \exp\left(H(y) - \mathbb{E}_{\mathbf{x}}[H(y|\mathbf{x})]\right)$$

Inception Net の出力

ラベル間の多様性

質の高さ

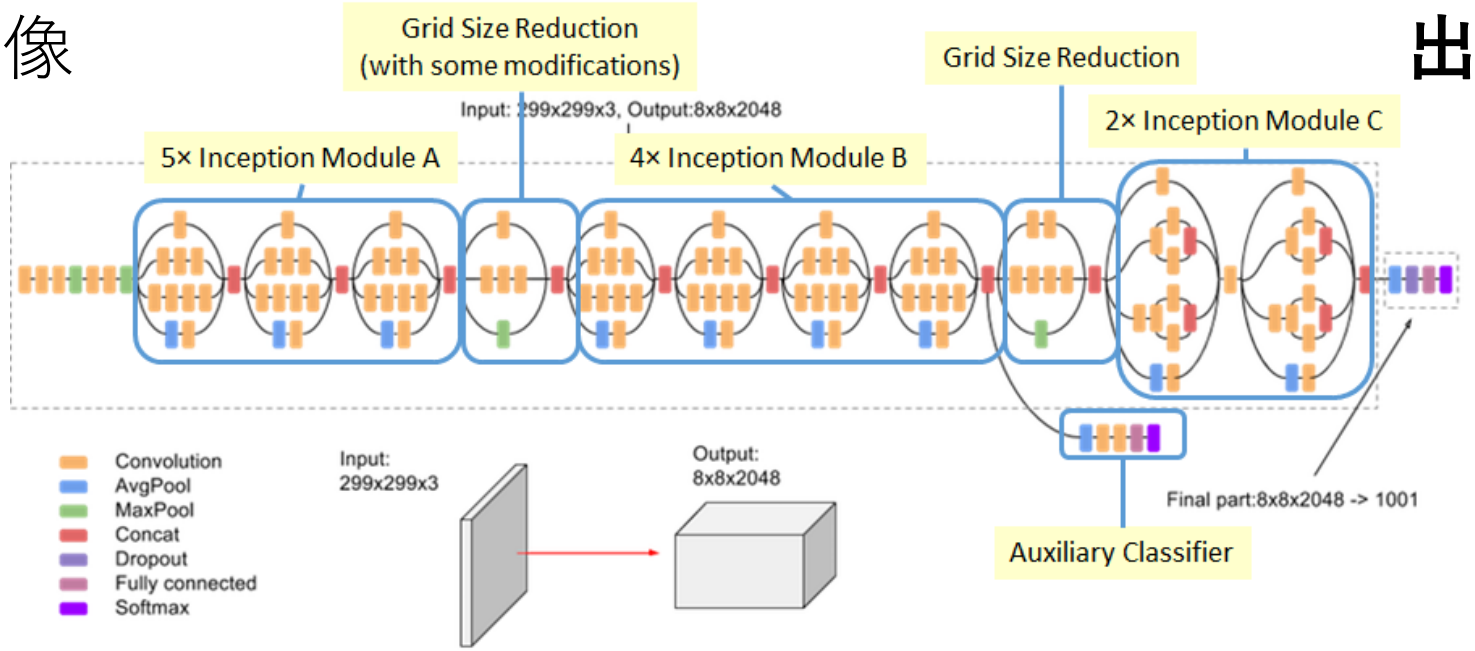
理想状態



Inception Net

入力: 画像

$\mathbf{x}$



出力: 確率値

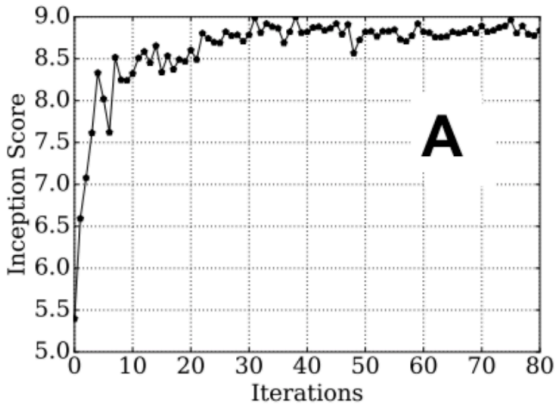
$p(y|\mathbf{x})$

IS の抱える問題

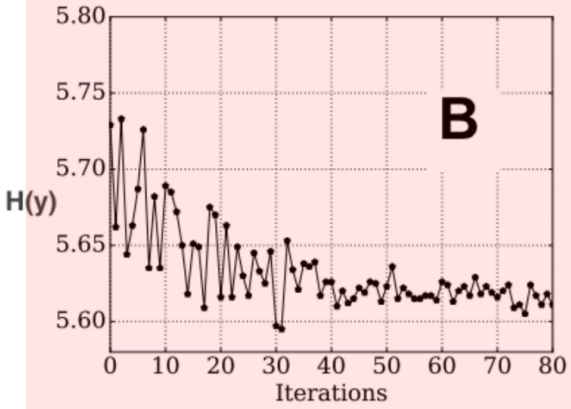
- 生成の多様性を測れない。
- mode collapse 発見できない。
- 画像の鮮明度が重要視される。
- 評価にNN使うのいいの？
  - 別ドメインのデータで学習したNNを評価に用いることの意義が不透明。
  - 画像以外のデータはこれだというNN(識別モデル)がない。

IS の抱える問題

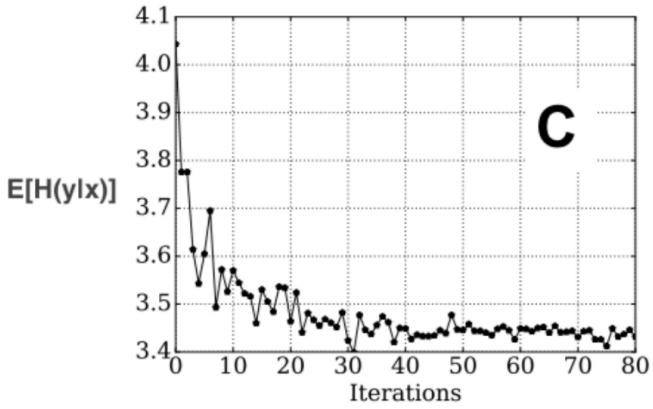
Inception Score



$H(y)$



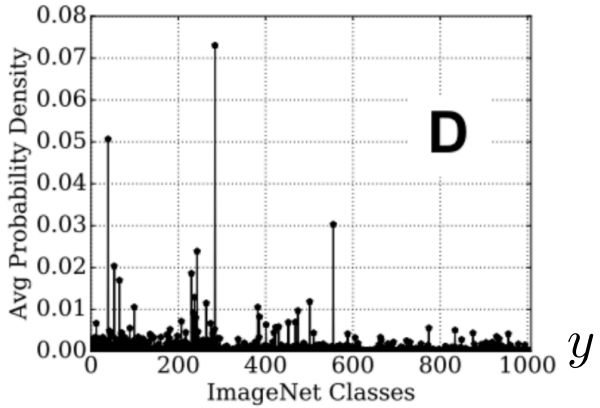
$H(y|\mathbf{x})$



なんで下がるの

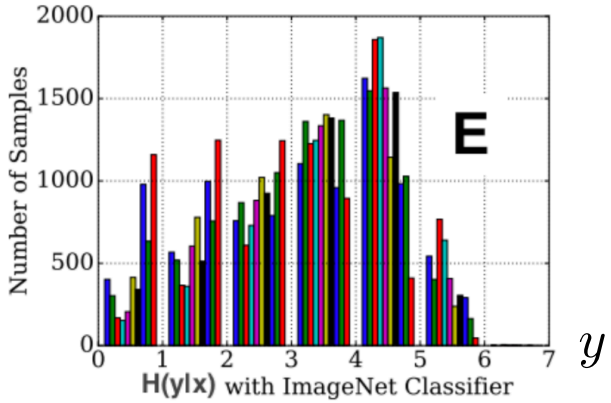
IS の抱える問題

- 評価に使うモデルを何で学習するかに超依存。

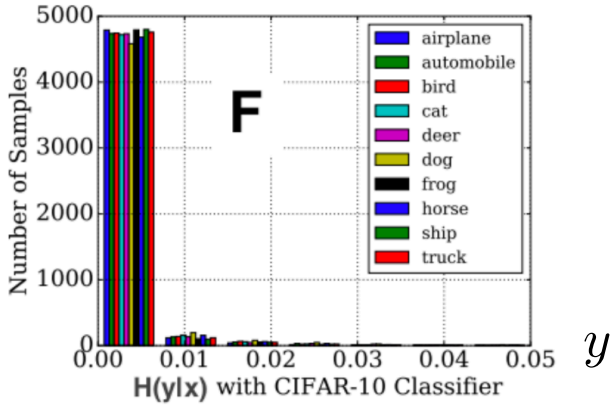


ImageNetで学習した
Inception Netに
CIFAR-10を入れたときの出力。

$p(y)$



ImageNetで学習した
Inception Net に
生成データを入れたもの



CIFAR-10 で学習した
Inception Net に
生成データを入れたもの

IS の抱える問題 (多様性、鮮明さ)

- モード崩壊に鈍感。
  - 各クラスのサンプルを一つ記憶すればISは高くなる。

$$\exp\left(\mathbb{E}_{\mathbf{x}}[\mathbb{KL}(p(\mathrm{y}|\mathbf{x})\|p(\mathrm{y}))]\right) = \exp\left(H(y) - \mathbb{E}_{\mathbf{x}}[H(y|\mathbf{x})]\right)$$

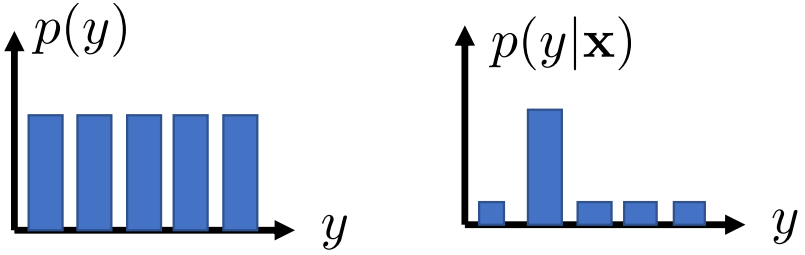
Inception Net の出力

ラベル間の多様性

質の高さ

多様性を鑑みる項はない

理想状態



Modified Inception Score (m-IS)

- クラス内の多様性を考慮したい。
  - 以下を最大化。

$$\exp\left(\mathbb{E}_{\mathbf{x}_i}\left[\mathbb{E}_{\mathbf{x}_j}\left[\left(\mathbb{KL}\left(P\left(y|\mathbf{x}_i\right)\|P\left(y|\mathbf{x}_j\right)\right)\right]\right]\right)$$

同じクラスの違うサンプル

- クロスエントロピーから着想。

Fréchet Inception Distance (FID)

- ImageNet で学習済みの Inception Net を用いた指標。
- 実データと生成データの分布間の Fréchet (Wasserstein-2) 距離。
  - 小さいほど良い。

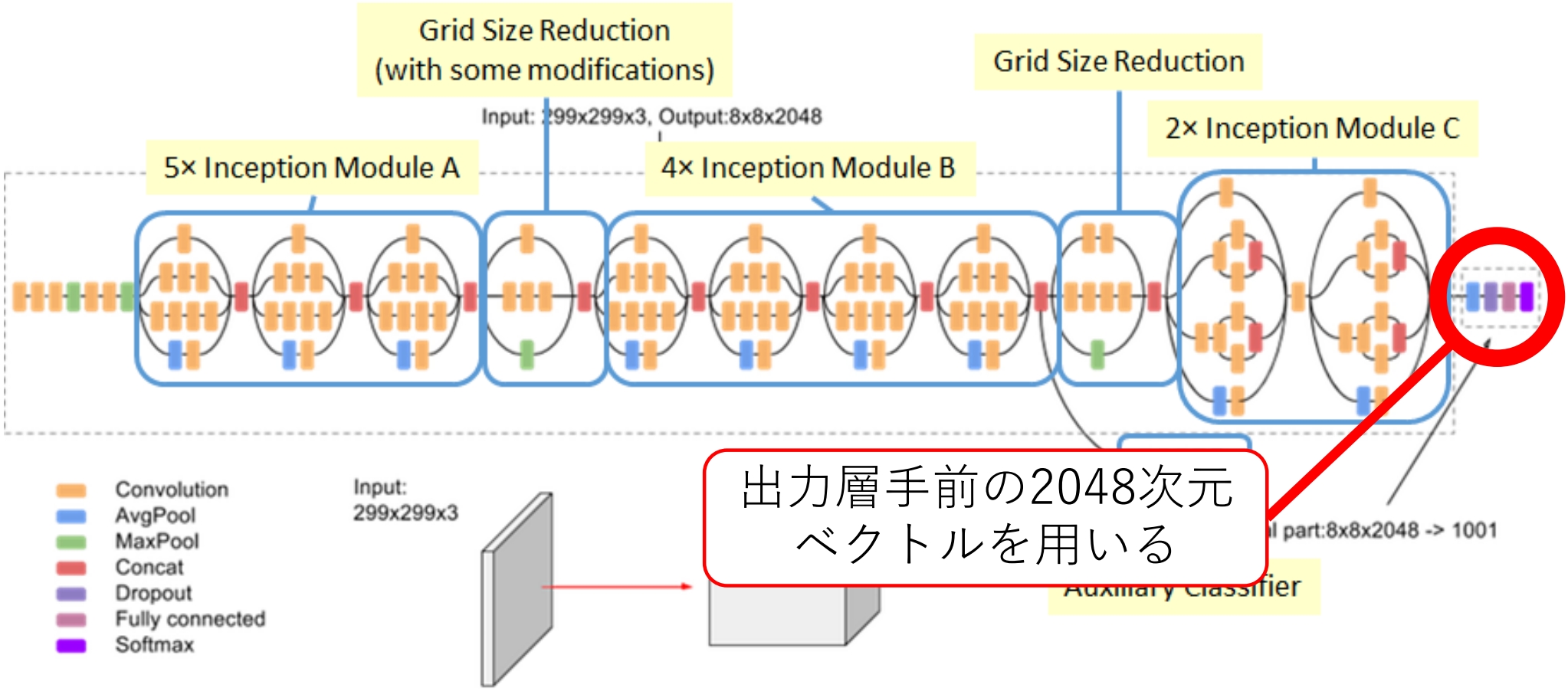
$$FID(r,g) = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{\frac{1}{2}}\right)$$

$$\text{where} \qquad \boldsymbol{\mu}_r = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_r^{(i)}$$

Inception Net の
中間層の出力(2048次元)

$$\boldsymbol{\Sigma}_r = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{h}_r^{(i)} - \boldsymbol{\mu})(\mathbf{h}_r^{(i)} - \boldsymbol{\mu})^{\mathrm{T}}$$

Inception Net



FID のモチベーション

- 実データをどれだけ忠実に表現できてるか指標が欲しい。
- IS は実データを評価に使わない。

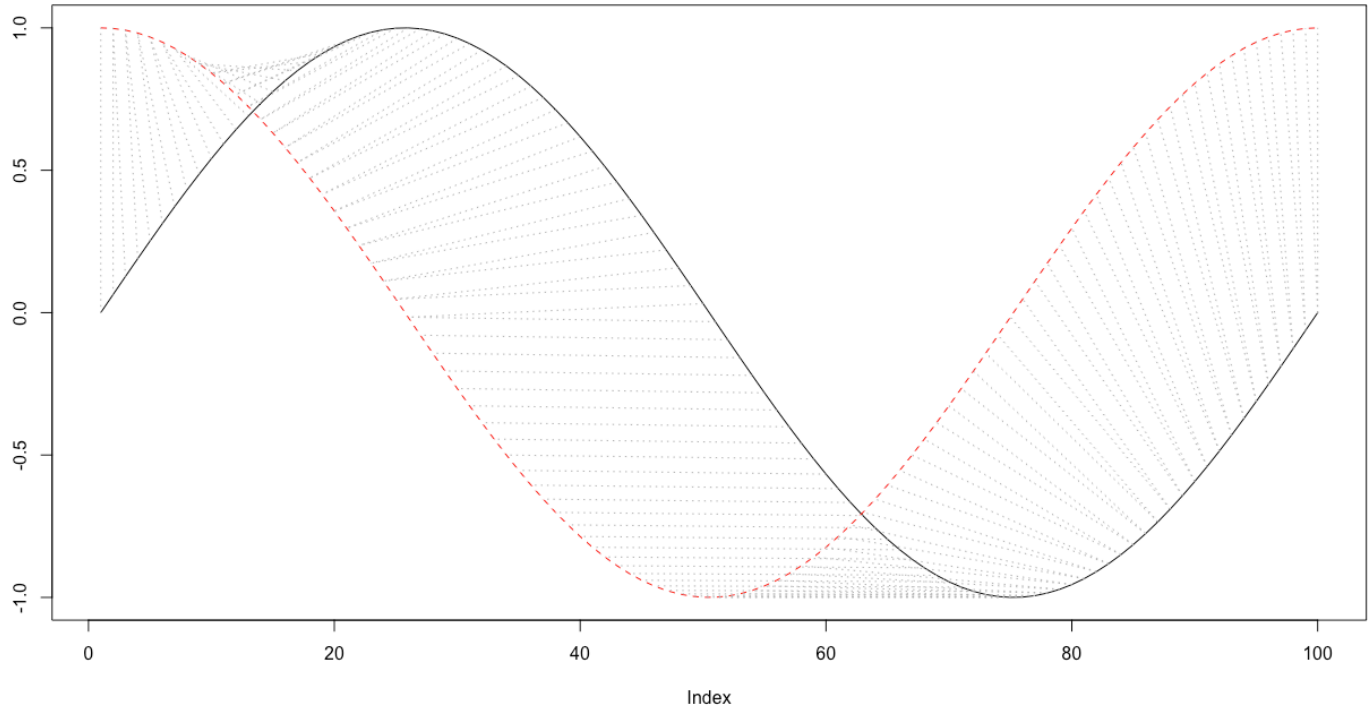
生成データ

$$\exp\left(\mathbb{E}_{\mathbf{x}}\left[\mathbb{KL}\left(p(\mathrm{y}|\mathbf{x})\|p(\mathrm{y})\right)\right]\right) = \exp\left(H(y) - \mathbb{E}_{\mathbf{x}}\left[H(y|\mathbf{x})\right]\right)$$

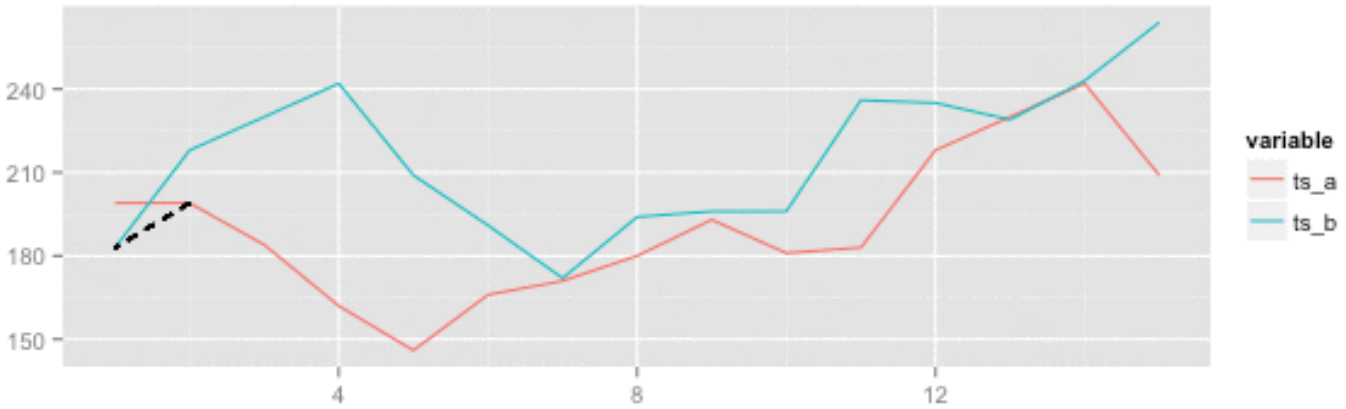
Inception Net
の出力

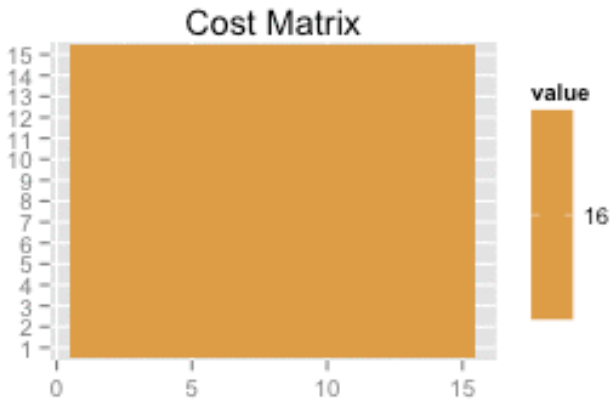
Fréchet (Wassestein-2) 距離とは？

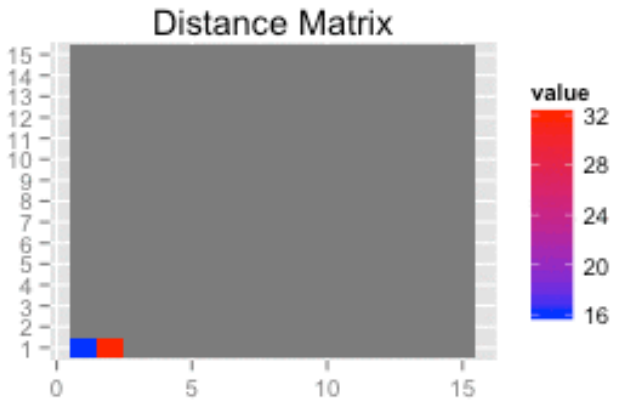
- 曲線間に定義される距離。
- 離散では DTW (Dynamic Time Warping)として知られる。
  - 2系列の各点の距離を総当たりし、その系列間の最短パスがDTW。



DTW







http://rtokei.tech/memo/memodynamic-time-warping/

FID

- Inception Net の出力がガウスに従うことを仮定している。
  - ガウス分布間の Fréchet 距離は以下で与えられる。

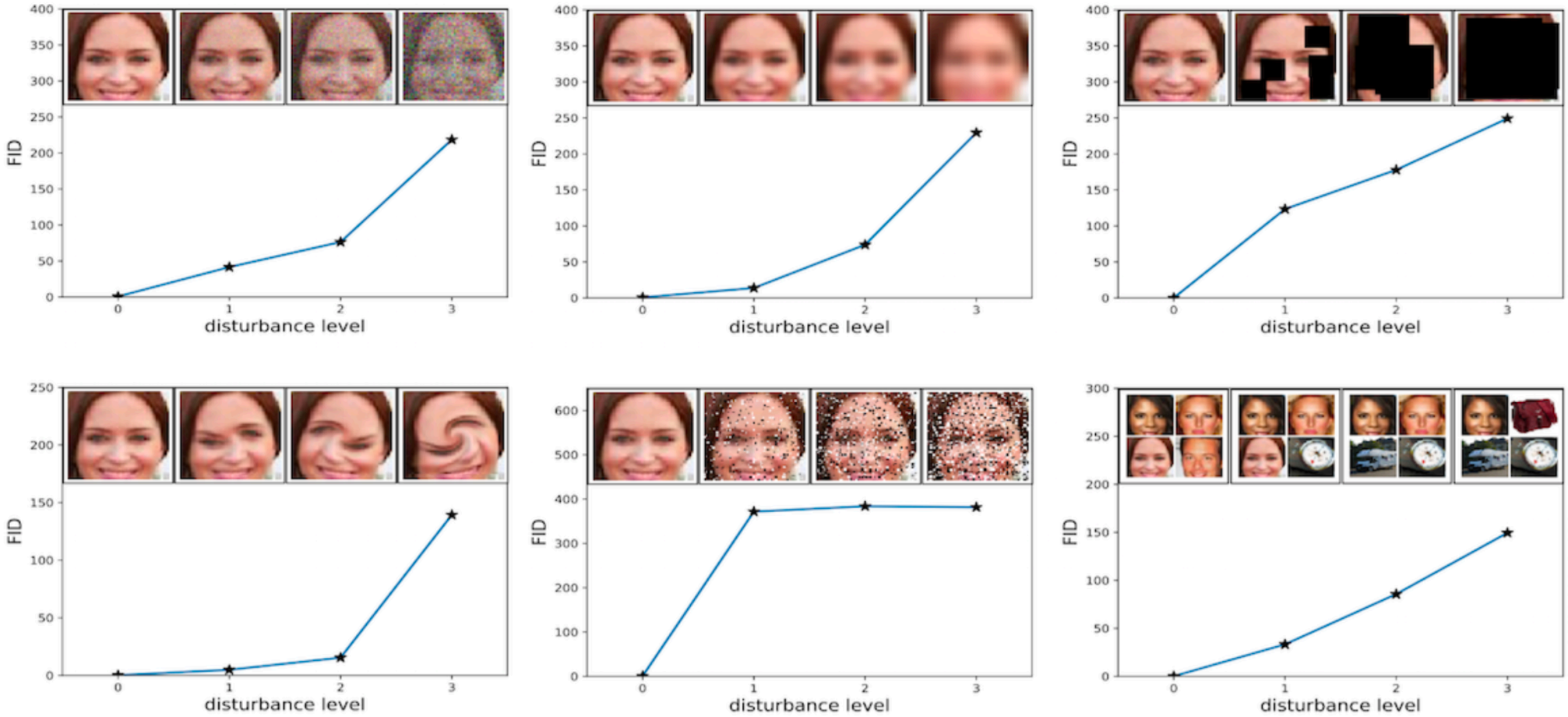
$$FID(r,g) = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{\frac{1}{2}}\right)$$

$$\text{where} \qquad \boldsymbol{\mu}_r = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_r^{(i)}$$

Inception Net の
中間層の出力(2048次元)

$$\boldsymbol{\Sigma}_r = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{h}_r^{(i)} - \boldsymbol{\mu})(\mathbf{h}_r^{(i)} - \boldsymbol{\mu})^{\mathrm{T}}$$

FID の挙動



FID の欠点

- Inception Net の中間層がガウスに従うことの仮定。
- IS と同様に NN を使うことの意義は不透明。

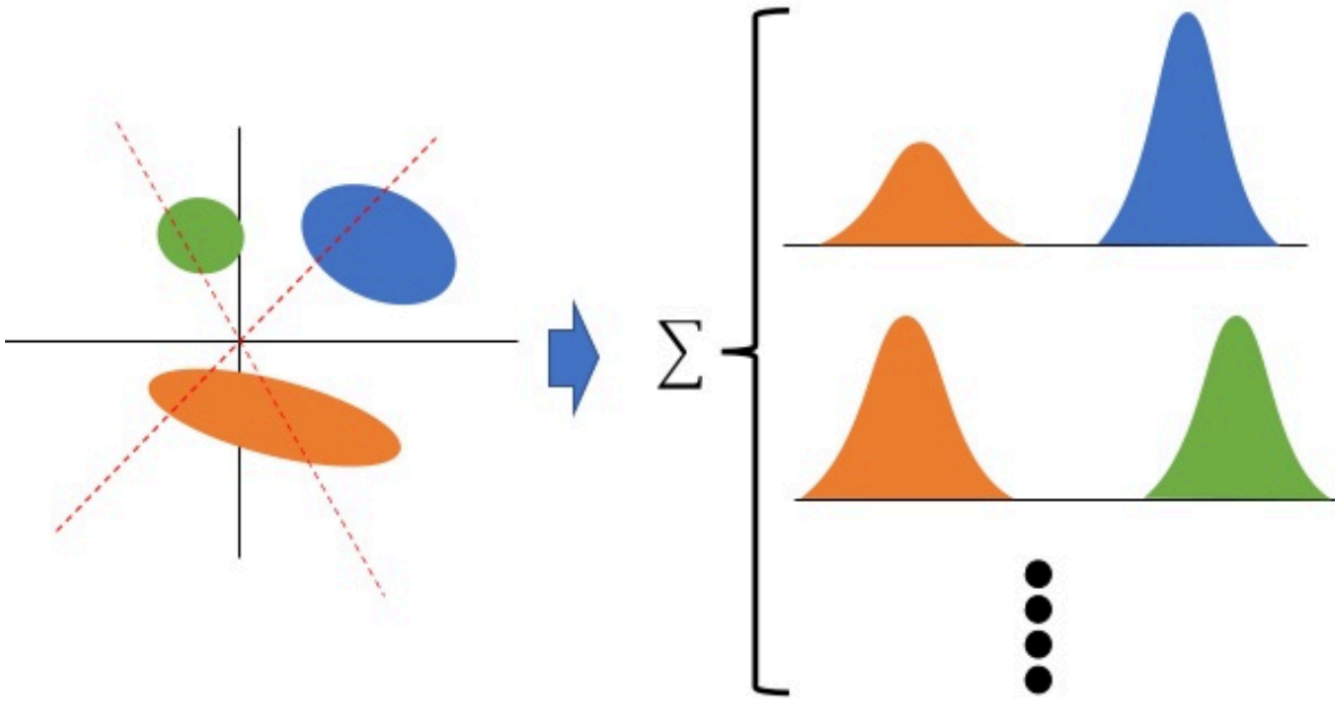
Sliced Wasserstein Distance

- Wasserstein Distance
  - 高次元では計算量増大の上、近似になる。

$$W\left(P_r, P_g\right) \propto \max_f \mathbb{E}_{\mathbf{x}\sim P_r}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_g}[f(\mathbf{x})]$$

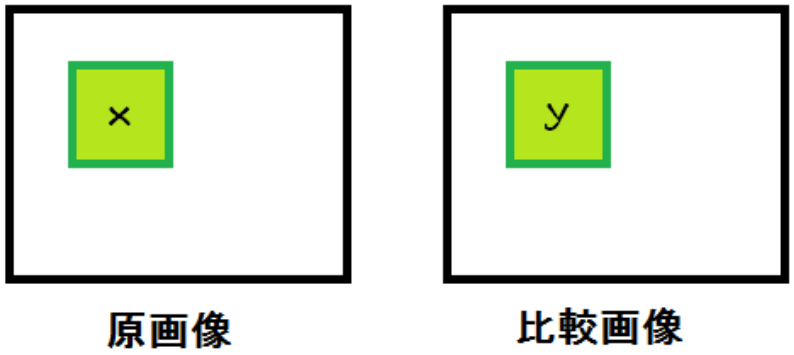
$$\hat{W}\left(\mathbf{x}_{test}, \mathbf{x}_g\right) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}\left(\mathbf{x}_{test}[i]\right) - \frac{1}{N}\sum_{i=1}^{N} \hat{f}\left(\mathbf{x}_g[i]\right)$$

  - 1次元ではclosed formで計算が楽。
    - Sliceしよう。


Σ

SSIM(Structural Similarity)

- 画質の評価基準の一つ。
  - MSEなどは画像全体が少し違う場合と、局所的に大きく違う場合を見分けられない。



画像の局所領域の平均と分散で計算。

SSIM

輝度

$$I(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

コントラスト

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

構造

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

$$\mathrm{SSIM}(x,y) = I(x,y)^{\alpha} C(x,y)^{\beta} S(x,y)^{\gamma}$$

結局何を使うのさ

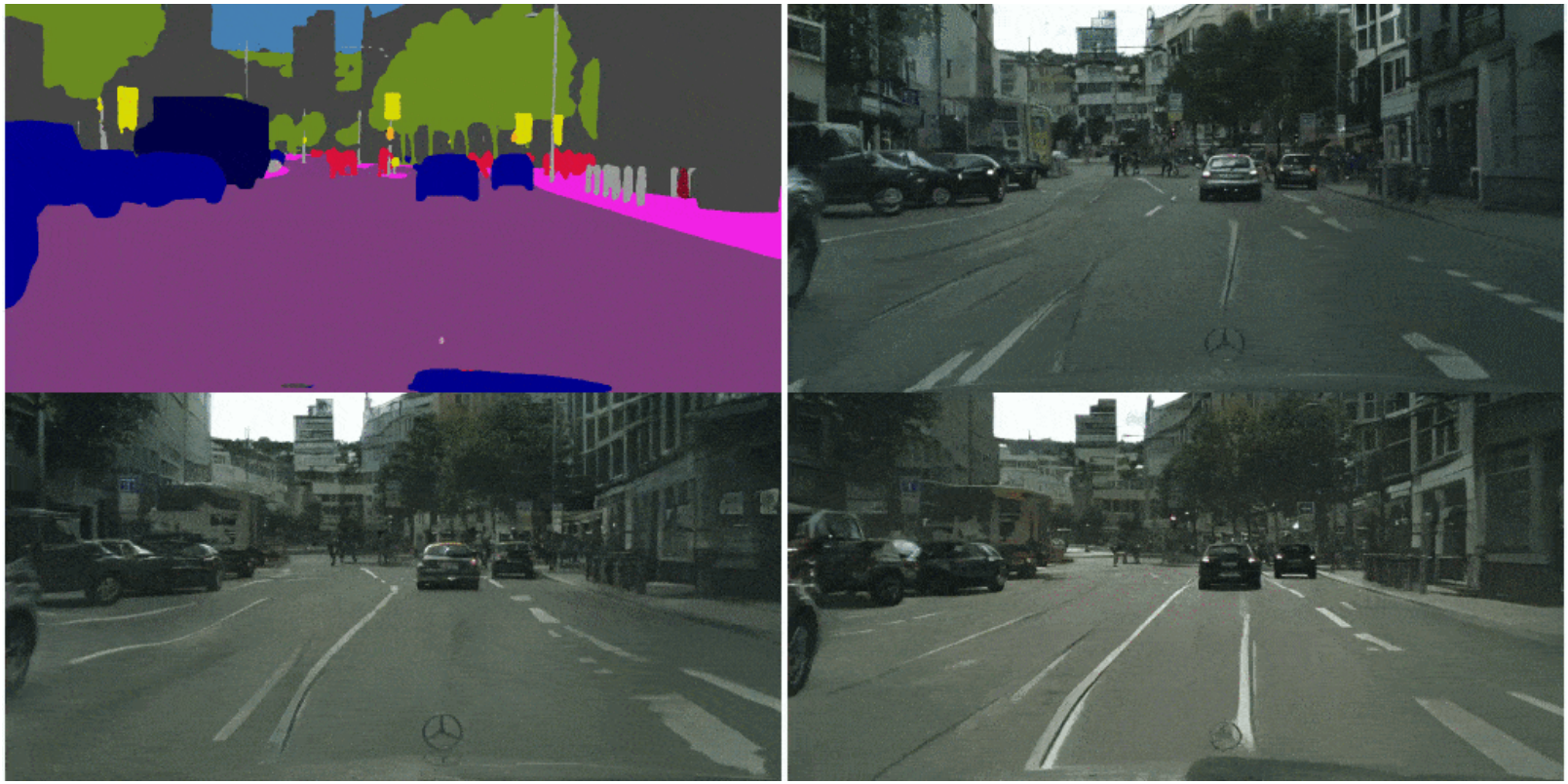
- 画像ならISかFID。
  - でもNNを評価に使う意義はやっぱ不透明。
  - 評価するNNの学習まで気を使う必要がある。

- 音楽生成では人へのアンケートが基本っぽい。
  - 画像でもアンケートは基本してる。

動画像生成

vid2vid

- セマンティックセグメンテーション等がなされたソース系列から目的の系列を生成。(pix2pixの動画版)



vid2vid の問題設定

ソース系列: $\mathbf{s}_1^T \equiv \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T\}$

ターゲット系列: $\mathbf{x}_1^T \equiv \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$

生成系列: $\tilde{\mathbf{x}}_1^T \equiv \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_T\}$

- 以下の分布の獲得が目標

$$p\left(\tilde{\mathbf{x}}_1^T | \mathbf{s}_1^T\right) = p\left(\mathbf{x}_1^T | \mathbf{s}_1^T\right)$$

- 目的関数

$$\max_D \min_G E_{\left(\mathbf{x}_1^T, \mathbf{s}_1^T\right)} \left[\log D\left(\mathbf{x}_1^T, \mathbf{s}_1^T\right)\right] + E_{\mathbf{s}_1^T} \left[\log\left(1 - D\left(G\left(\mathbf{s}_1^T\right), \mathbf{s}_1^T\right)\right)\right]$$

vid2vid の工夫 (Generator)

- マルコフ性を仮定(論文ではL=2)。

$$p\left(\tilde{\mathbf{x}}_1^T|\mathbf{s}_1^T\right)=\prod_{t=1}^{T}p\left(\tilde{\mathbf{x}}_t|\tilde{\mathbf{x}}_{t-L}^{t-1},\mathbf{s}_{t-L}^{t}\right)$$

過去L時刻までの生成系列
＋現在も含んだL+1時刻のソース系列

- 以下のようなGenerator を定義

$$\tilde{\mathbf{x}}_t=F\left(\tilde{\mathbf{x}}_{t-L}^{t-1},\mathbf{s}_{t-L}^{t}\right)$$

$$F\left(\tilde{\mathbf{x}}_{t-L}^{t-1},\mathbf{s}_{t-L}^{t}\right)=(\mathbf{1}-\tilde{\mathbf{m}}_t)\odot\tilde{\mathbf{w}}_{t-1}\left(\tilde{\mathbf{x}}_{t-1}\right)+\tilde{\mathbf{m}}_t\odot\tilde{\mathbf{h}}_t$$

vid2vid の工夫 (Generator)

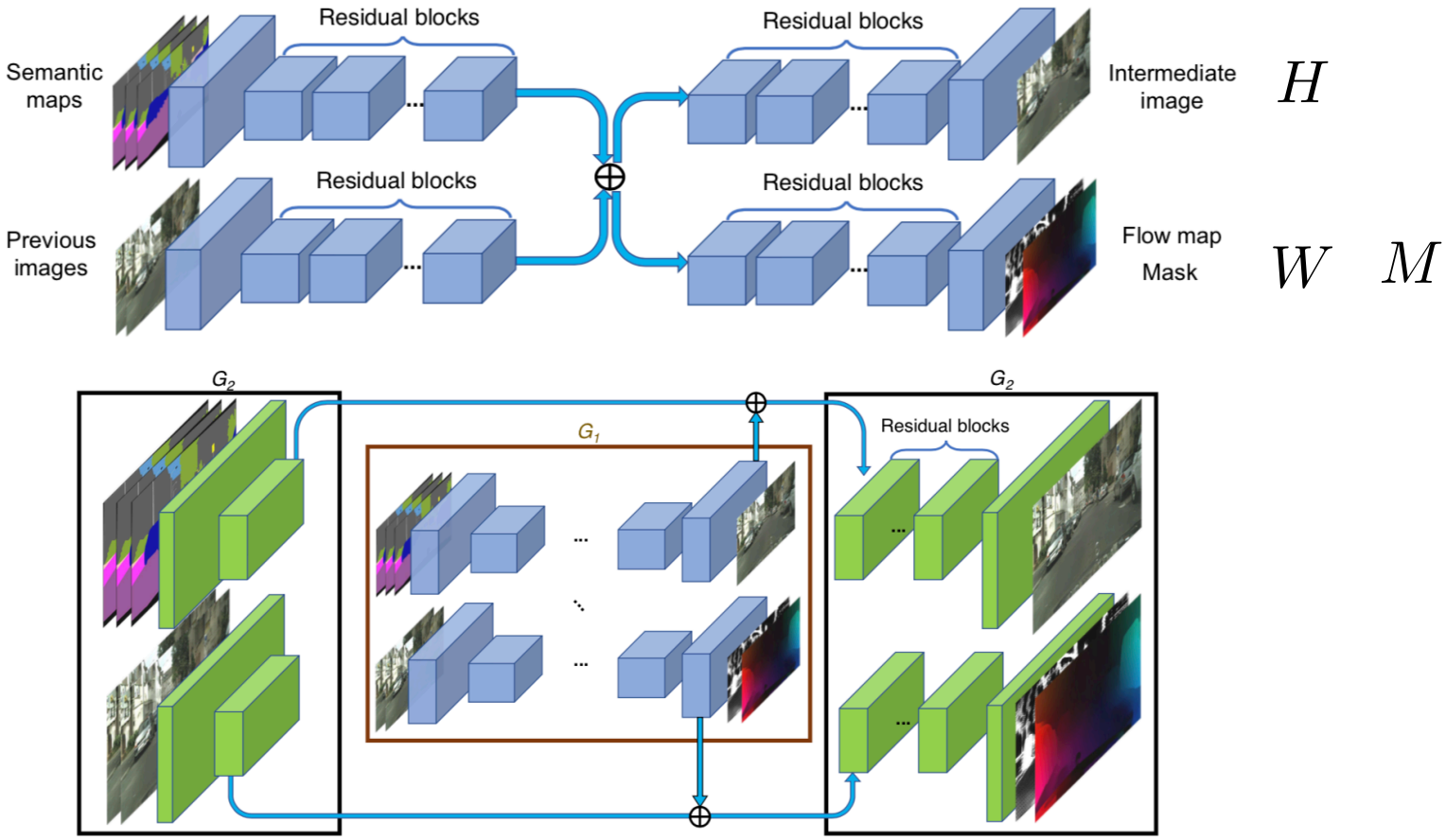
$$F\left(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}\right) = (\mathbf{1} - \tilde{\mathbf{m}}_t) \odot \tilde{\mathbf{w}}_{t-1}\left(\tilde{\mathbf{x}}_{t-1}\right) + \tilde{\mathbf{m}}_t \odot \tilde{\mathbf{h}}_t$$

- $\tilde{\mathbf{w}}_{t-1} = W\left(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}\right)$: estimated optical flow $(\mathbf{x}_{t-1} \to \mathbf{x}_t)$
- $\tilde{\mathbf{h}}_t = H\left(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}\right)$: hallucinated image
- $\tilde{\mathbf{m}}_t = M\left(\tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}\right)$: occlusion mask

全部NN(ResNet)で構成。
動画の冗長性を考慮する。

vid2vid の工夫 (Generator)

- 低解像度のGと高解像度のGを設計し、組み合わせる。



vid2vid の工夫 (Discriminator)

- 2 種類の判別器を用いる。
  - 画像の判別器 D_Iと、動画の判別器 D_V 。

- 最終的な目的関数:

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F)$$

(論文では$\lambda_W = 10$)

$\mathcal{L}_W(F)$: 生成系列と真の系列のVGG net で得られる中間層の特徴量の L1 loss。(学習速度を上げる目的)

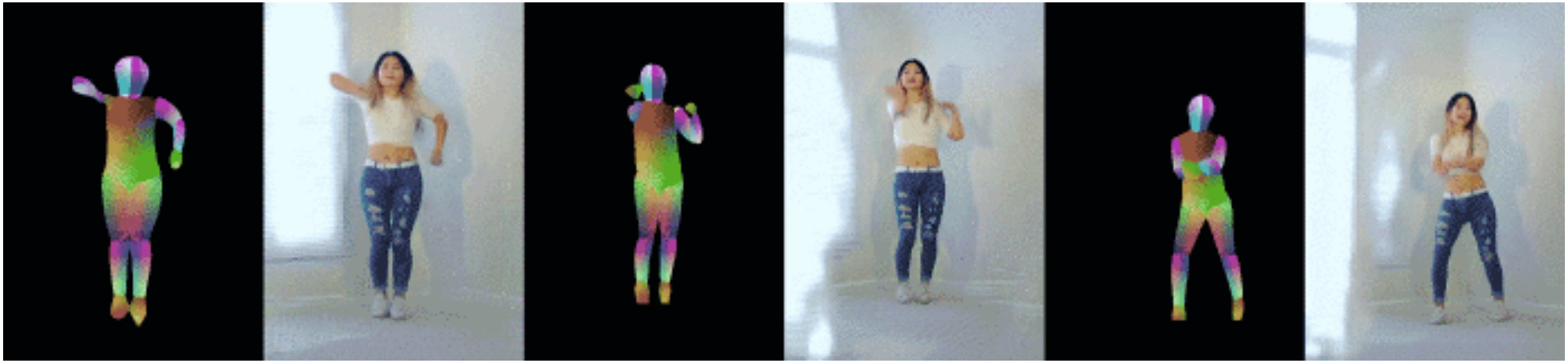
vid2vid の生成物

- labelの操作。

- スケッチをソースに。

vid2vid の生成物

- poseをソースに。



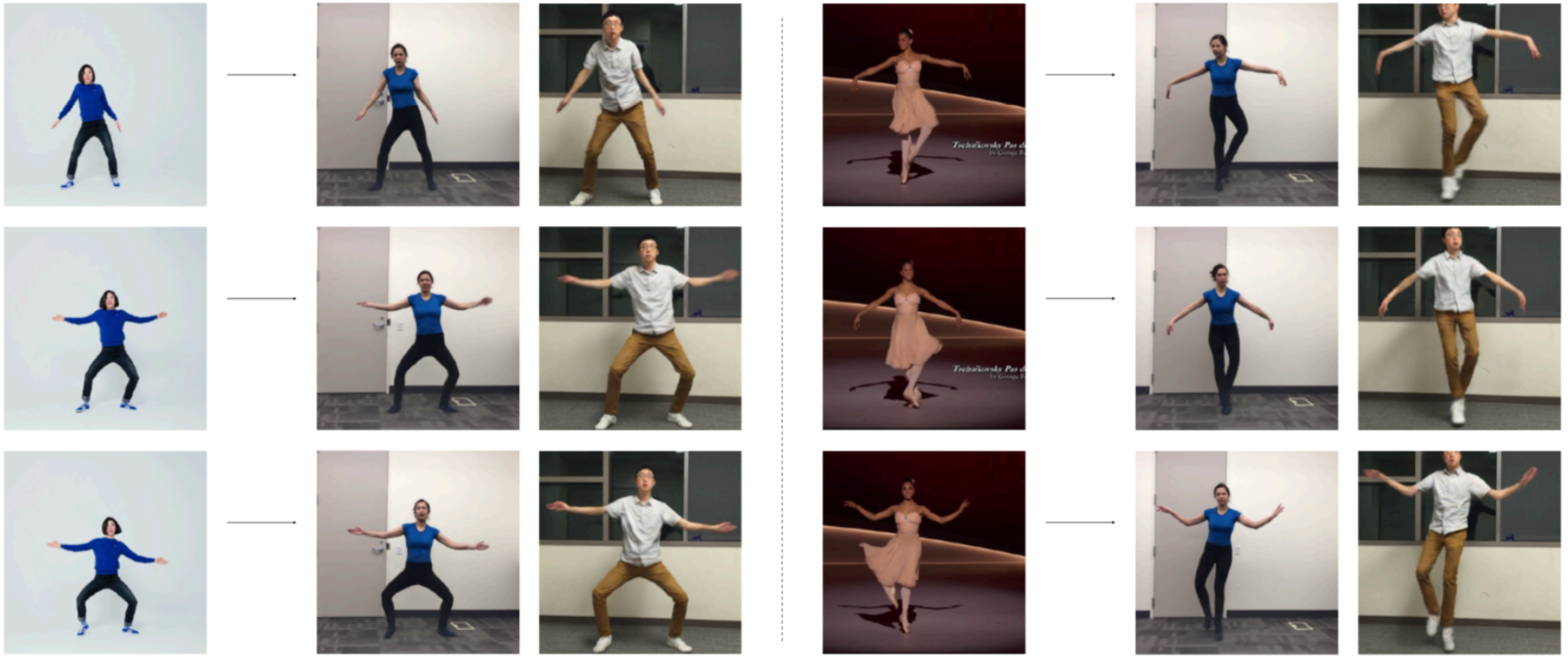
- 評価にはFID、生成物の評価はアンケートで。

vid2vid の限界

- 深さ方向の情報はないため、方向転換する車などの生成は難しい(セマンティックセグメンテーションの場合)。

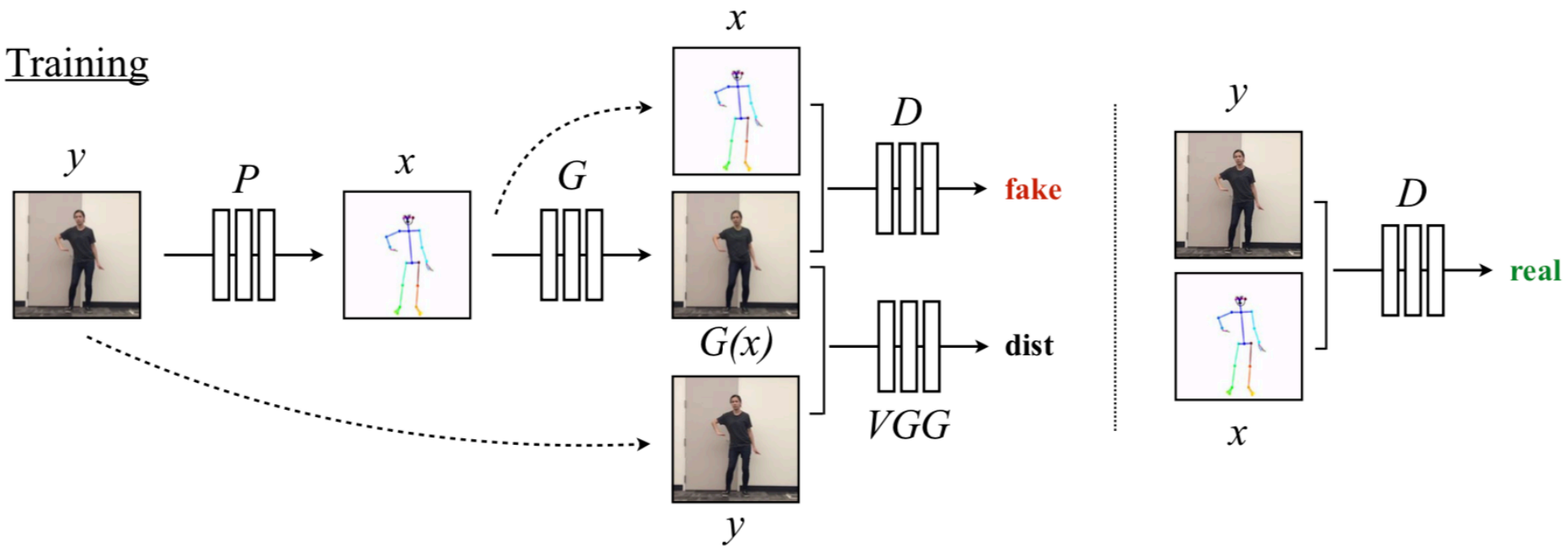
- 色や質感の一貫性は保証できない(形状も上記の場合などでだめ)。

everybody dance now



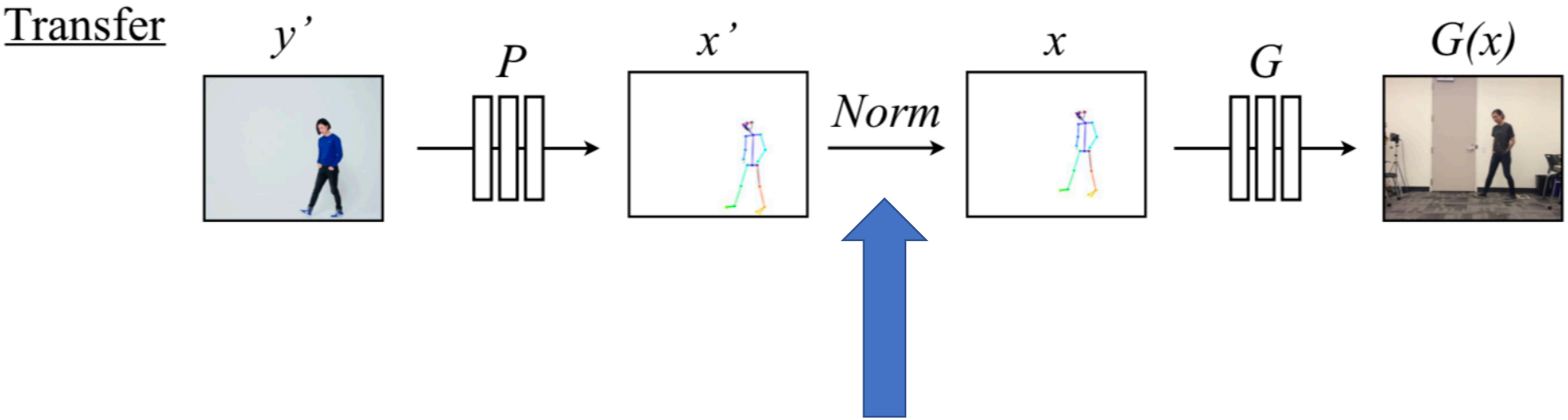
base idea - training

openpose と pix2pix



base idea - transfer

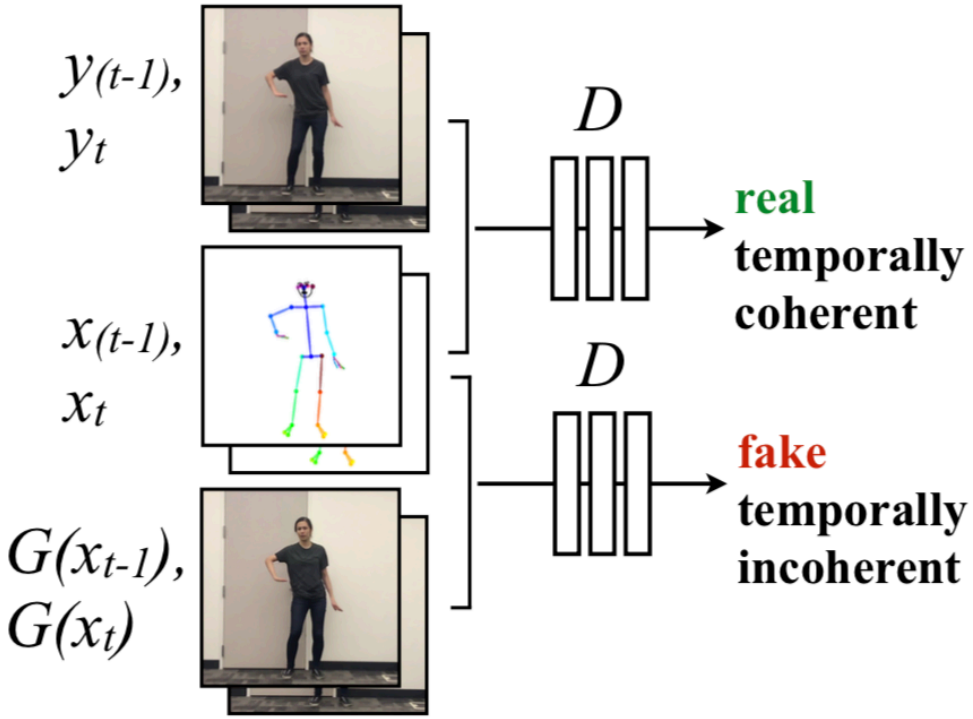
sourceと同じ姿勢のtargetを生成。



骨格や、床の位置、スケールを補正する。

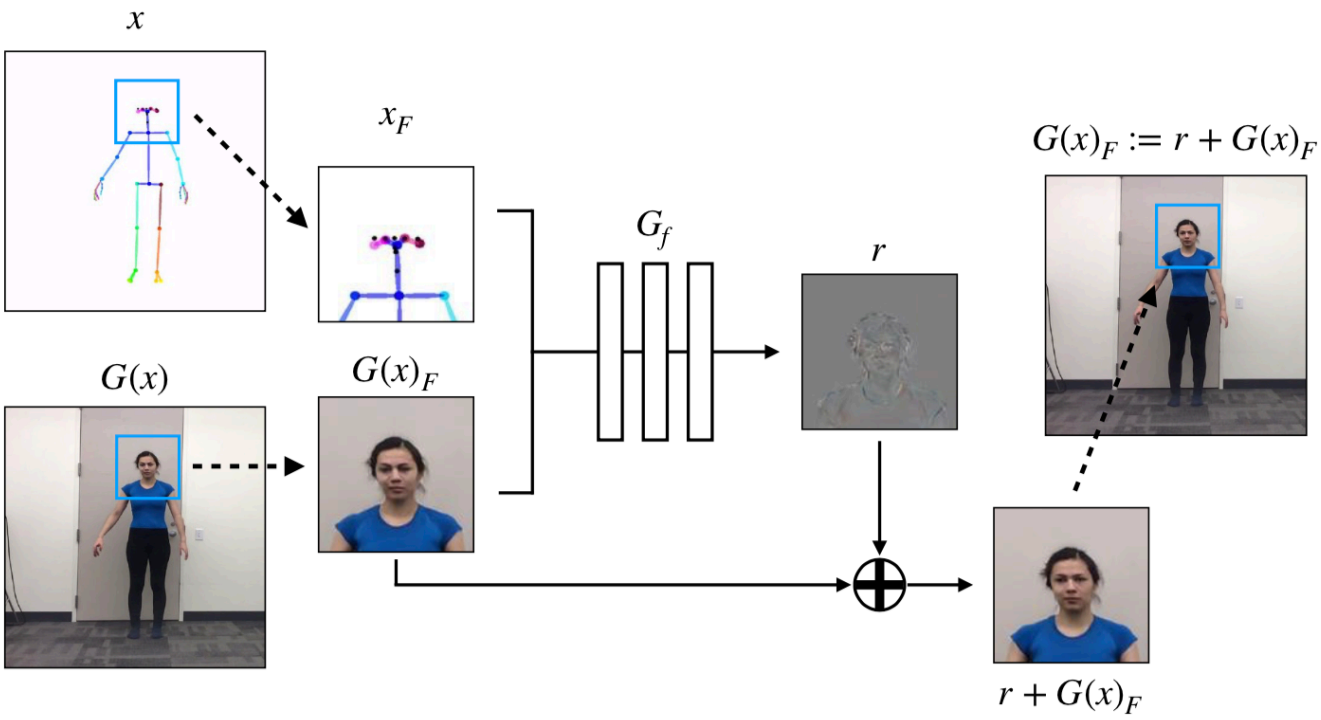
Temporal Smoothing

- 連続するフレームが自然になるように。
  - pix2pixを連続する2枚組で行う。



Face GAN

- 顔部分の生成性能の向上のために。
  - 別のGANで差分を生成。



Face GAN の効果

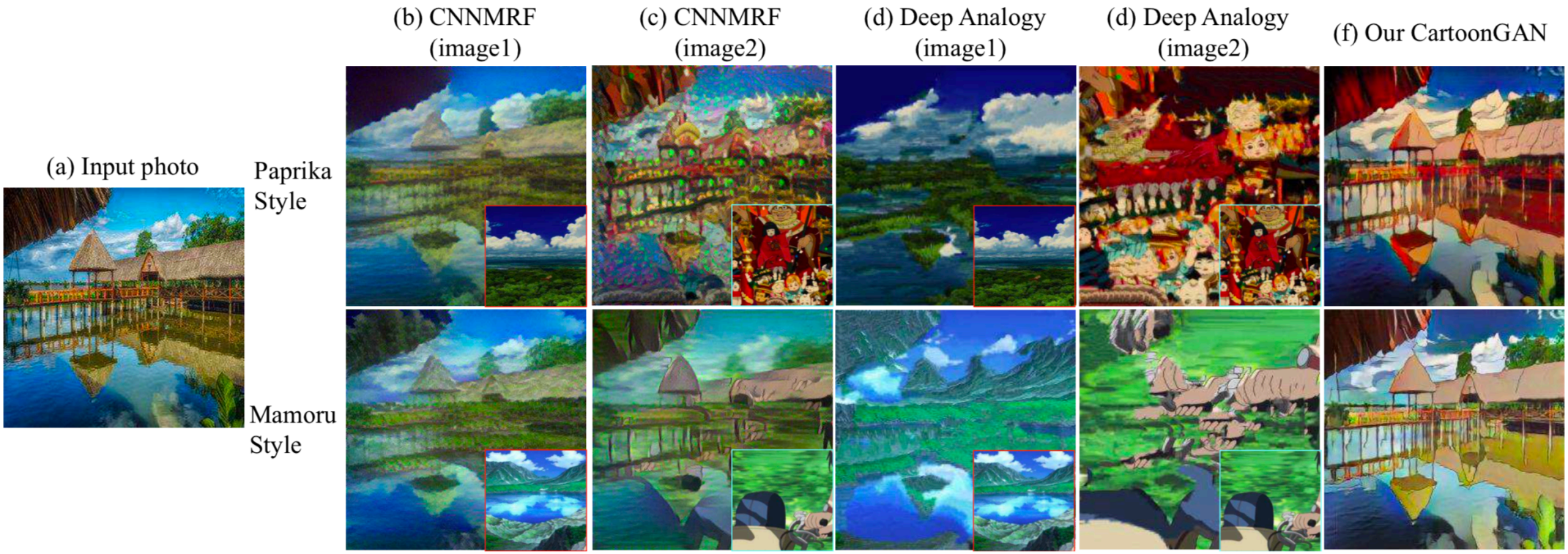
CartoonGAN

(a) Original scene

(b) Our result

実写のアニメ調化とは

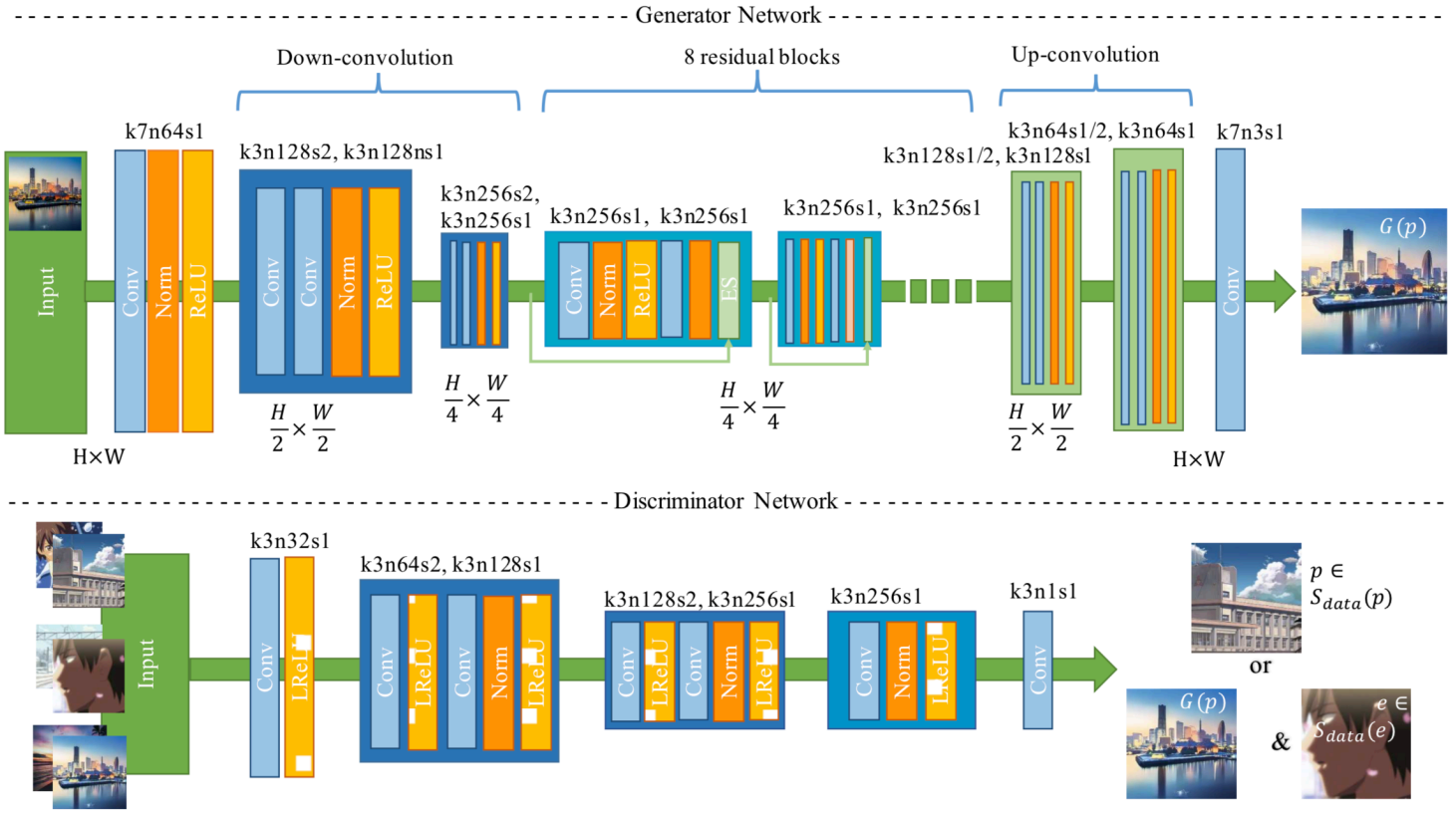
- 鮮明な輪郭、なめらかな色調、少ないテクスチャ
- 高度な抽象化、単純化



CartoonGANの工夫

- アニメに合わせた2つの損失関数
- 実写とアニメのデータセットを関連付けずに学習可能
- 収束を早めるための初期化

NNの機構



- Discriminatorへの入力は部分画像(パラメータ削減)

損失関数

$$\mathcal{L}(G, D) = \mathcal{L}_{adv}(G, D) + \omega \mathcal{L}_{con}(G, D)$$

鮮明さ　　変換後の内容の保持

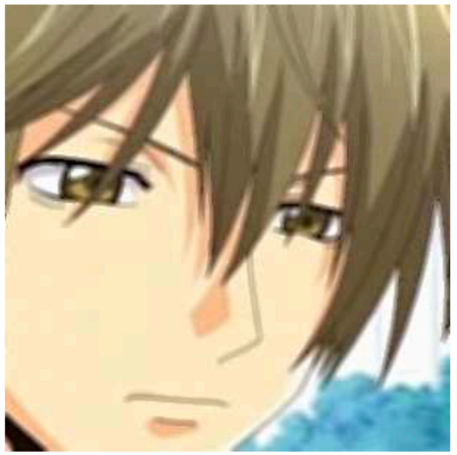
(論文では $w = 10$)

損失関数1 (鮮明な輪郭)

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{c_i \sim S_{\mathrm{data}}(c)} \left[\log D\left(c_i\right)\right] \quad \text{(アニメ画像)}$$
$$+ \mathbb{E}_{e_j \sim S_{\mathrm{data}}(c)} \left[\log\left(1 - D\left(e_j\right)\right)\right] \quad \text{(ぼけ画像)}$$
$$+ \mathbb{E}_{p_k \sim S_{\mathrm{data}}(p)} \left[\log\left(1 - D\left(G\left(p_k\right)\right)\right)\right] \quad \text{(変換画像)}$$

正例

負例



損失関数1 (変換後の内容の保持)

$$\mathcal{L}_{con}(G, D) = \mathbb{E}_{p_i \sim S_{data}(p)} \left[\| VGG_l \left(G \left(p_i \right) \right) - VGG_l \left(p_i \right) \|_1 \right]$$

- VGGの中間層の特徴量のL1。
- L1使うのが著者曰く工夫ポイント。理由不明。

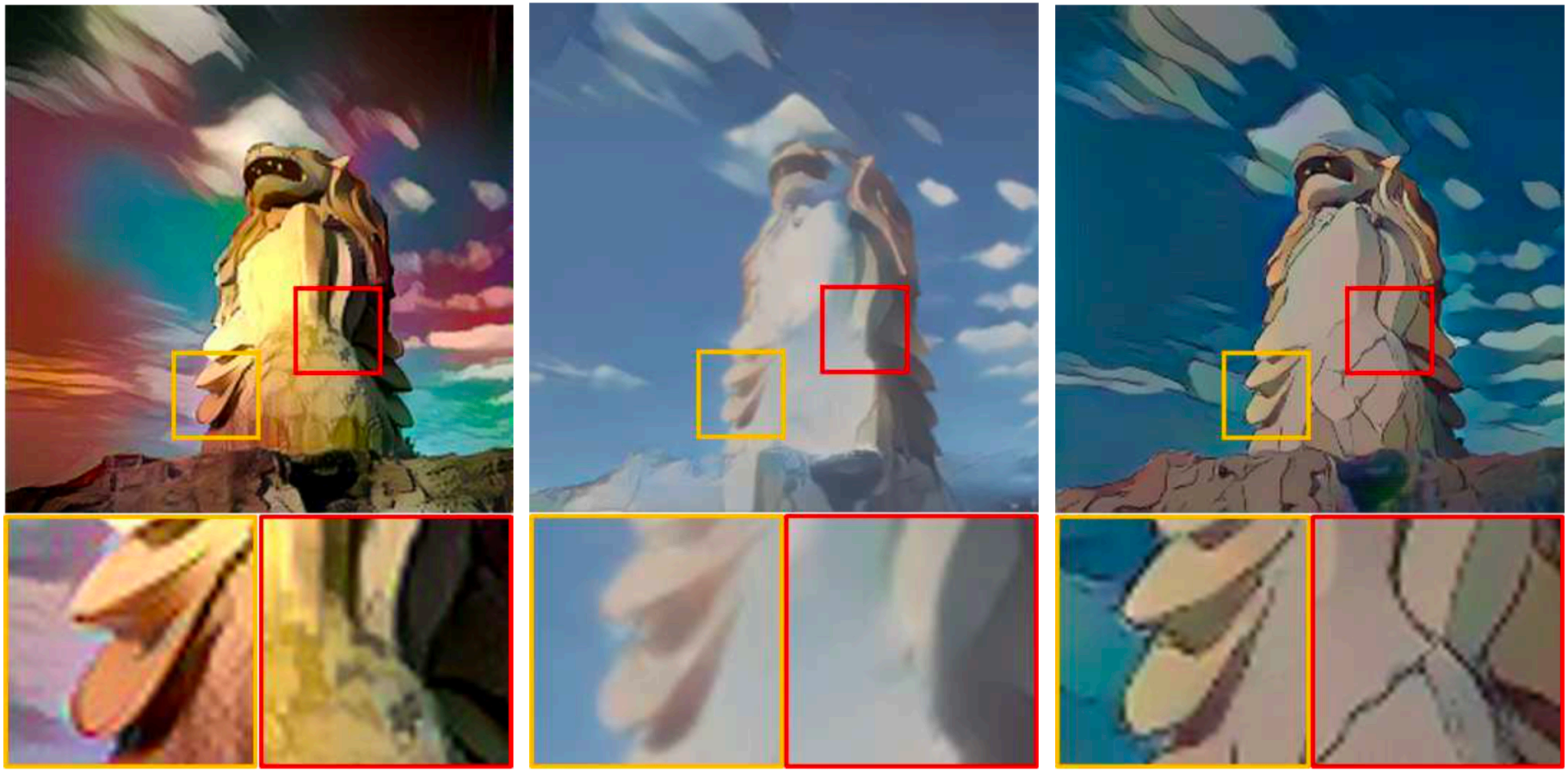
画像の初期化

- 学習初期は $\mathcal{L}_{con}(G, D)$ のみで学習 (ただの復元誤差)

(a) Original photo

(b) Image after initialization

生成物



(a) NST (b) CycleGAN (c) CartoonGAN

生成物

※これは画像ごとに適用しただけ

https://qiita.com/t-Asai/items/9e199db5fd0574f2ff8b

本家 slideshare

https://www.slideshare.net/hamadakoichi/anime-generation

参考文献

- Shmelkov, K., Schmid, C., & Alahari, K. (2018). How good is my GAN?. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 213-229).
- Wang, T. C., Liu, M. Y., Zhu, J. Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*.
- Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5933-5942).
- Chen, Y., Lai, Y. K., & Liu, Y. J. (2018). Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9465-9474).