コンピュータビジョンにおける無教師学習の進展とその課題

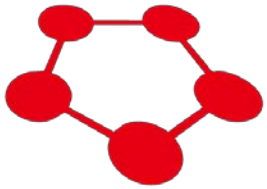
齋藤真樹

Preferred Networks, Inc.

2018/11/4

自己紹介 & 会社紹介

- 東北大学岡谷研 (CVの研究室) → 株式会社Preferred Networks
  - 博士過程の間はマルコフ確率場の最適化問題に関する研究に従事
  - 国際会議: CVPR, ICCV　　国内会議: MIRU, SSII
  - 現在: 製造業分野のCVの応用（特に外観検査），GANの研究
- 株式会社Preferred Networks
  - 自動運転，製造業，バイオに関する研究開発
  - 特に有名なものとしてはChainer, 研究成果だとSpectral Normalization



Chainer

Spectral Normalization for Generative Adversarial Networks

https://arxiv.org › cs ▾ このページを訳す

T Miyato 著 - 2018 - 被引用数: 103 - 関連記事

やはりメディア表現と画像という話でGANは避けられない

https://twitter.com/goodfellow_ian/status/1001965061806706691

やはりメディア表現と画像という話でGANは避けられない

例2: Crypko

GANで自分好みの
アニメキャラクターを生成

https://crypko.ai/

全体の流れ

やはりメディア表現と画像という話でGANは避けられない

- とはいえ，応用一辺倒では興味を持っていただくのは難しそう

1. 無教師学習の進展
   - 画像とGANを中心とした，無教師学習の主要なトレンドを俯瞰
2. 無教師学習の今後の課題
   - いくつかの実験における考察を通して

無教師学習の進展

(主にビジョン方面から)

(ビジョン分野での) 無教師学習の主観的勢力図

Generative model

Generative Adversarial Nets

DCGAN

SNGAN

SA-GAN

Pix2Pix

CycleGAN

Vid2Vid

WGAN

WGAN-GP

WGAN-GP

BigGAN

StackGAN++

Graphical model

Boltzmann Machine, RBM

Autoregressive model

PixelRNN, VideoRNN

Flow-based model

Flow, RealNVP, Glow

Variational Auto-Encoder

Generative Adversarial Network [Goodfellow2014]

- GeneratorとDiscriminatorの2つのネットワークを利用
  - Generator: ノイズzからサンプルxを生成
  - Discriminator: サンプルxがデータセットかGeneratorかを識別

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{z \sim p_z}[f(D_\psi(G_\theta(z))] + \mathbb{E}_{x \sim p_D}[f(-D_\psi(x))]$$

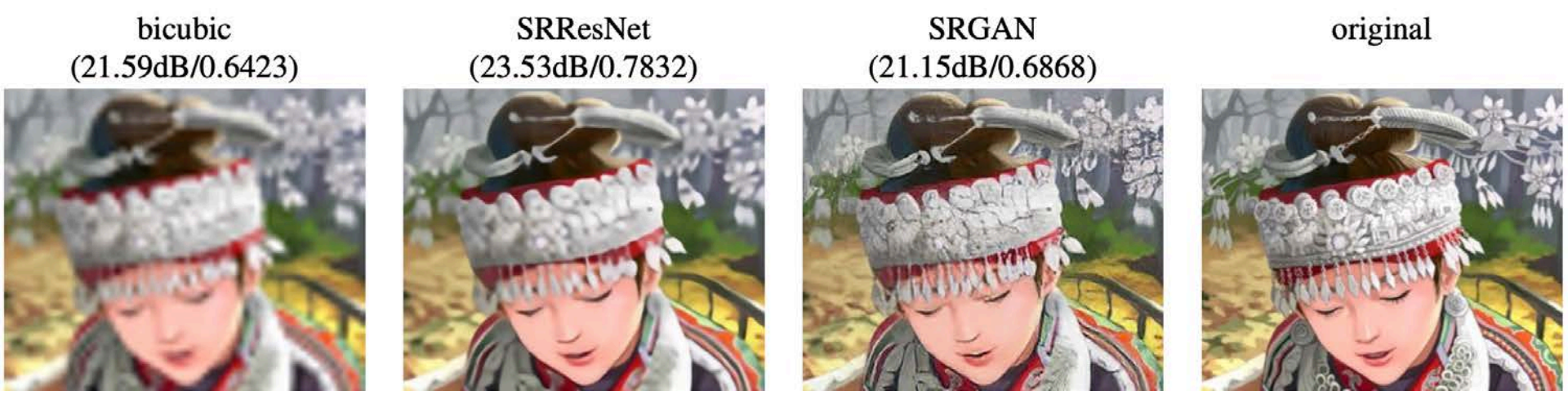
- なぜGANがここまで(CVの領域で)人気なのか？
  1. 画像生成において効率的なネットワークアーキテクチャの発見
  2. 画像空間で損失関数を定義しないアプローチ(Adversarial loss)の発明
  3. 実装が楽
  4. (綺麗な画像を生成するのは楽しい)

A. Radford et al., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, In ICLR, 2016

Generative Adversarial Network [Goodfellow2014]

- なぜGANがここまで(CVの領域で)人気なのか？
  1. 画像生成において効率的なネットワークアーキテクチャの発見
  2. 画像空間で損失関数を定義しないアプローチ(Adversarial loss)の発明
  3. 実装が楽

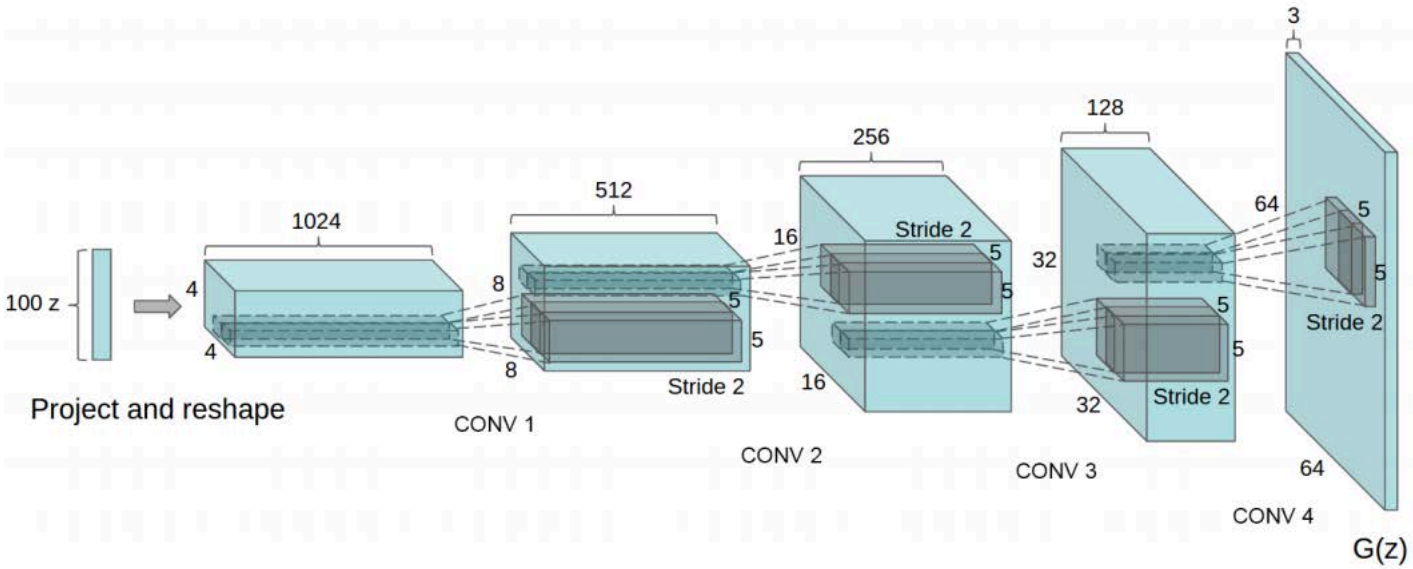
$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR}))$$



C. Ledig et al., Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, In CVPR, 2017

DCGANの台頭 [Radford2016]

Convolutional/Deconvolutional layersとGANの組み合わせが
画像生成において極めて有効であることを実証



それ以降，GANを用いた応用研究が次々と発表されることに

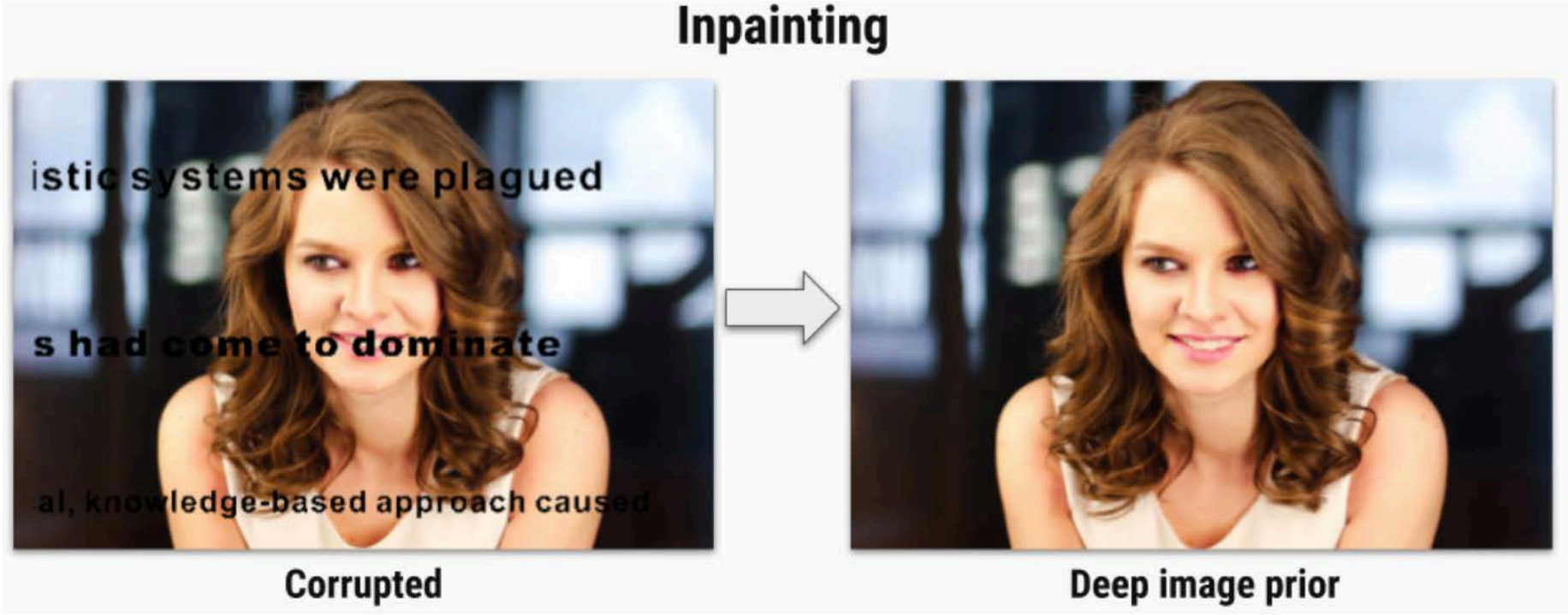
A. Radford et al., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, In ICLR, 2016

なぜDCGANが有効なのか？ [Ulyanov2018]

CNNのネットワーク構造自体が自然画像の空間をよく捉えている

(Deep Image Prior)

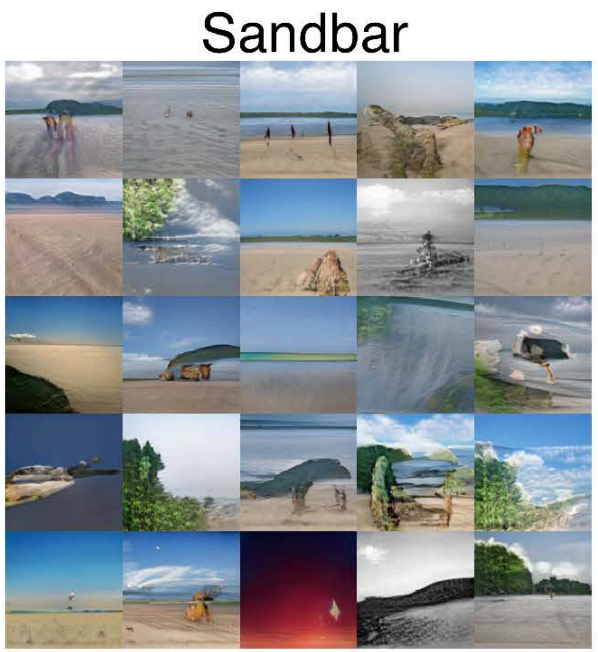
$$\min_{\theta} E(f_\theta(z); x_0)$$



D. Ulyanov et al., Deep Image Prior, In CVPR, 2018

Spectral Normalization (a.k.a. SNGAN) [Miyato2018]

- Discriminatorからの勾配を安定してGeneratorに伝えることが重要
  - Discriminatorが入力に対してK-Lipschitz連続である制約を付与
  - いろんな抑え方がある(e.g. WGAN-GP, SVC)が，その中の一つとして
    すべてのウェイトのスペクトラルノルムを一定数以下にするものがある



T. Miyato et al., Spectral Normalization for Generative Adversarial Networks, In ICLR, 2018

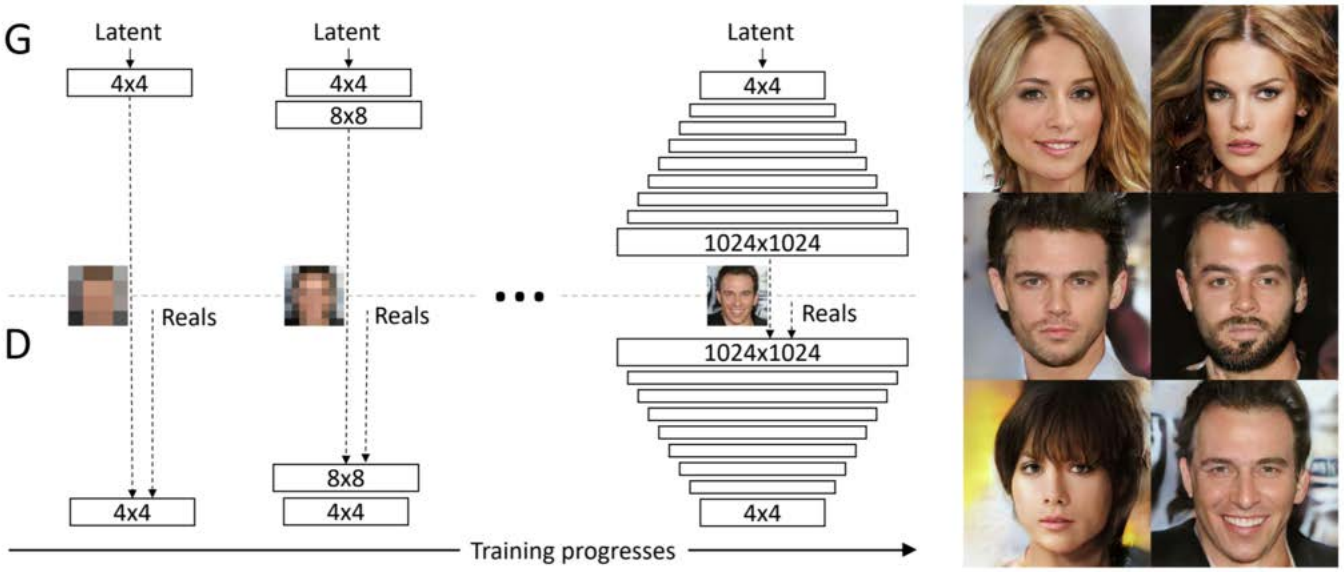
SelfAttention-GAN [Zhang2018]

- 発展版．SNGANにSelf Attentionを加えるなどのトリックを加えることで画質の向上に寄与
- (ただSAを自分の研究で試そうとしたが今の所失敗している)

H. Zhang et al., Self-Attention Generative Adversarial Networks, arXiv, 2018

ProgressiveGAN [Karras2018]

- GeneratorとDiscriminatorの解像度を学習が進むにつれて徐々に上げることで，高詳細な画像を得る.
  - 多重解像度はCVの領域で頻繁に用いられるアプローチの一つ
  - Pros: 1024pxレベルの画像を効率的に生成
  - Cons: アニーリングライクな手法は調整が難しい



T. Karras et al., Progressive Growing of GANs for Improved Quality, Stability, and Variation, In ICLR, 2018

BigGAN [Brock2018]

- いくつか改良を加えているが，その中でもバッチサイズとチャネル数の増加が重要と指摘
- 実際にある程度動かしてみての感想
  - バッチ数とチャネル数は確かに重要（レイヤ数増やすのはだめ）
  - ProgressiveGANライクなアプローチは最近あまり使われていないのは同意

Batch	Ch.	Param (M)	Shared	Hier.	Ortho.	Itr $\times 10^3$	FID	IS
256	64	81.5	SA-GAN Baseline			1000	18.65	52.52
512	64	81.5	✗	✗	✗	1000	15.30	58.77(±1.18)
1024	64	81.5	✗	✗	✗	1000	14.88	63.03(±1.42)
2048	64	81.5	✗	✗	✗	732	12.39	76.85(±3.83)
2048	96	173.5	✗	✗	✗	295(±18)	9.54(±0.62)	92.98(±4.27)
2048	96	160.6	✓	✗	✗	185(±11)	9.18(±0.13)	94.94(±1.32)
2048	96	158.3	✓	✓	✗	152(±7)	8.73(±0.45)	98.76(±2.84)
2048	96	158.3	✓	✓	✓	165(±13)	8.51(±0.32)	99.31(±2.10)
2048	64	71.3	✓	✓	✓	371(±7)	10.48(±0.10)	86.90(±0.61)

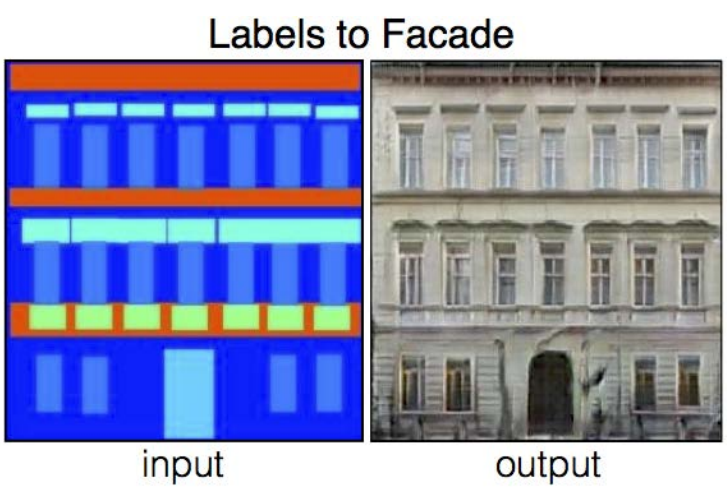
A. Brock et al., Large Scale GAN Training for High Fidelity Natural Image Synthesis, arXiv, 2018

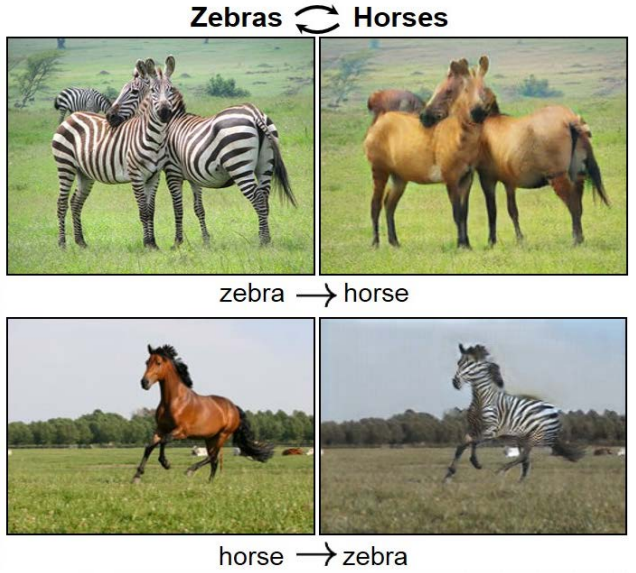
BigGAN [Brock2018]

A. Brock et al., Large Scale GAN Training for High Fidelity Natural Image Synthesis, arXiv, 2018

バッチサイズの不思議 (Pix2Pix, CycleGAN)

- じゃあGANの学習は全部バッチサイズを上げればいいのかというと…
- 入力ドメインを別ドメインに変換する手法は1から4程度のサイズで十分
- 出力に必要な多様性によって，要求されるバッチサイズは変化する？
  - 参考: image classification, semantic segmentation

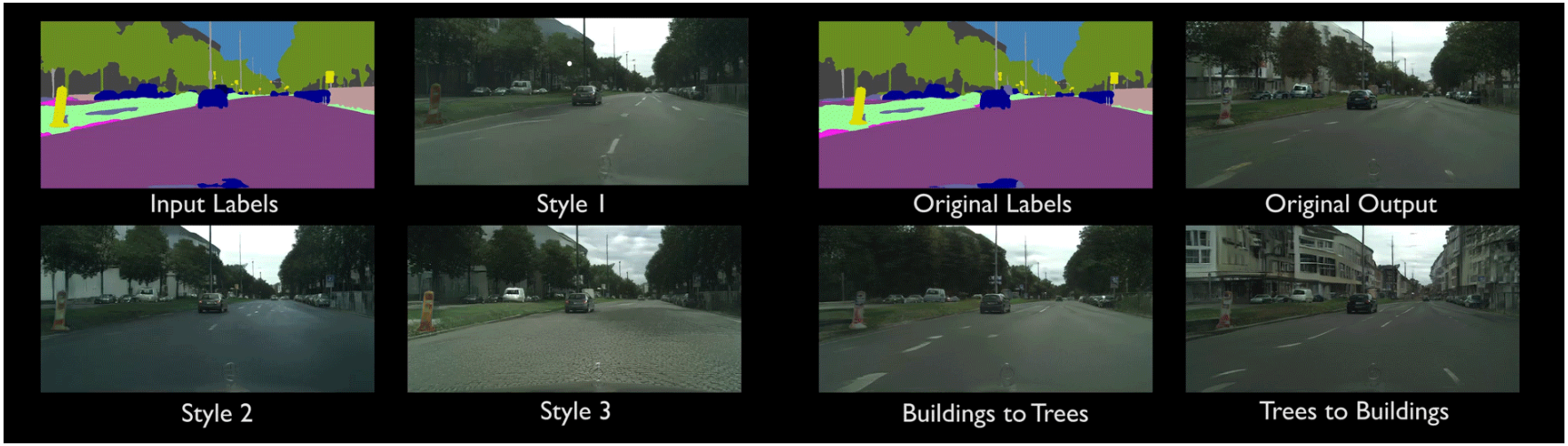




P. Isola et al., Image-to-Image Translation with Conditional Adversarial Networks, In CVPR, 2017
J. Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, In ICCV, 2017
N. S. Keskar et al., On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, In ICLR, 2017

動画問題への応用

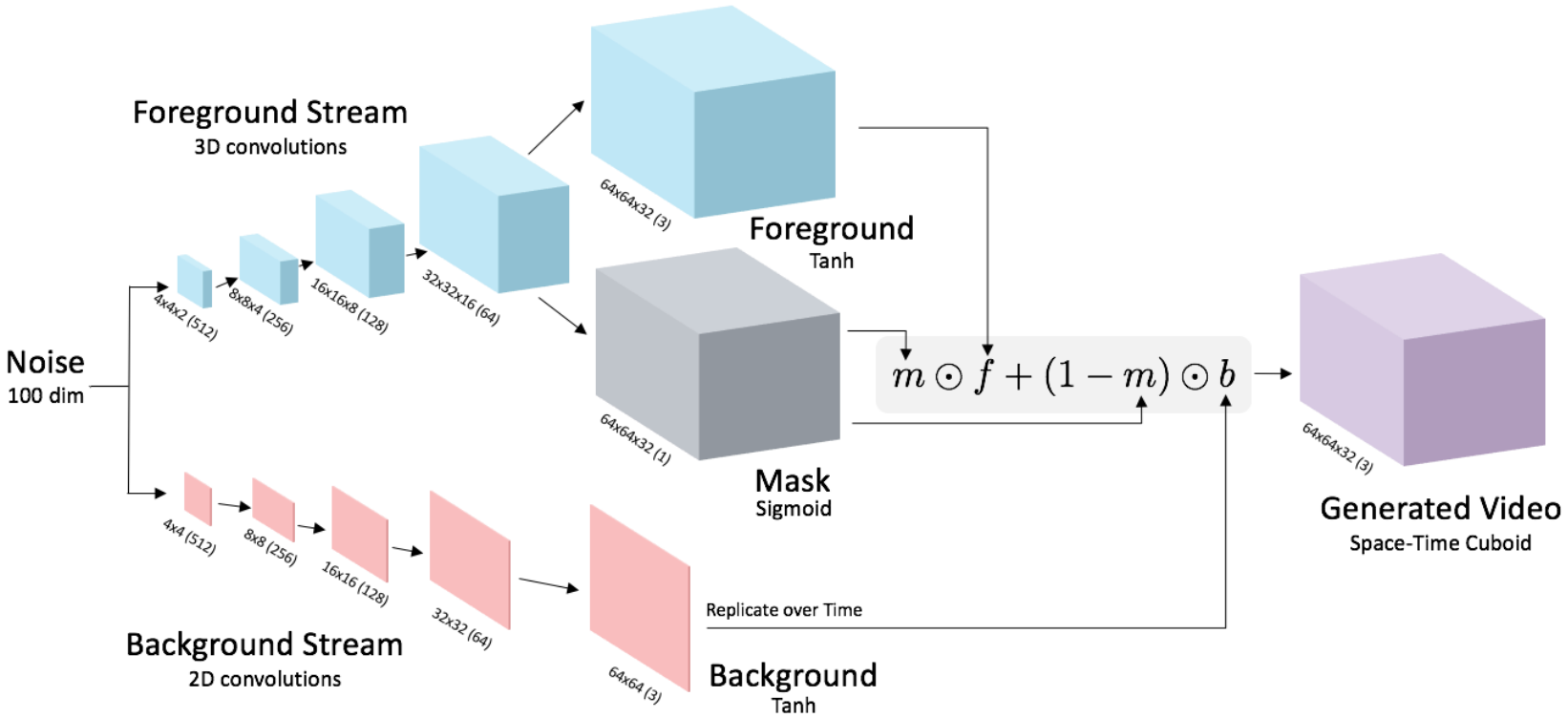
- 基本的にGANの応用は画像が一番盛り上がっている
  - アプローチとネットワーク構造が安定してきたため
- もう一つのCVが扱う主要なドメインとして動画が挙げられる
  - 時系列を扱うのは比較的難しい (安定した構造がない)
  - その中で上手く行っているのはPix2Pix系統の応用



T. Wang et al., Video-to-Video Synthesis, arXiv, 2018

Scene Dynamics (a.k.a. VGAN) [Vondrick2016]

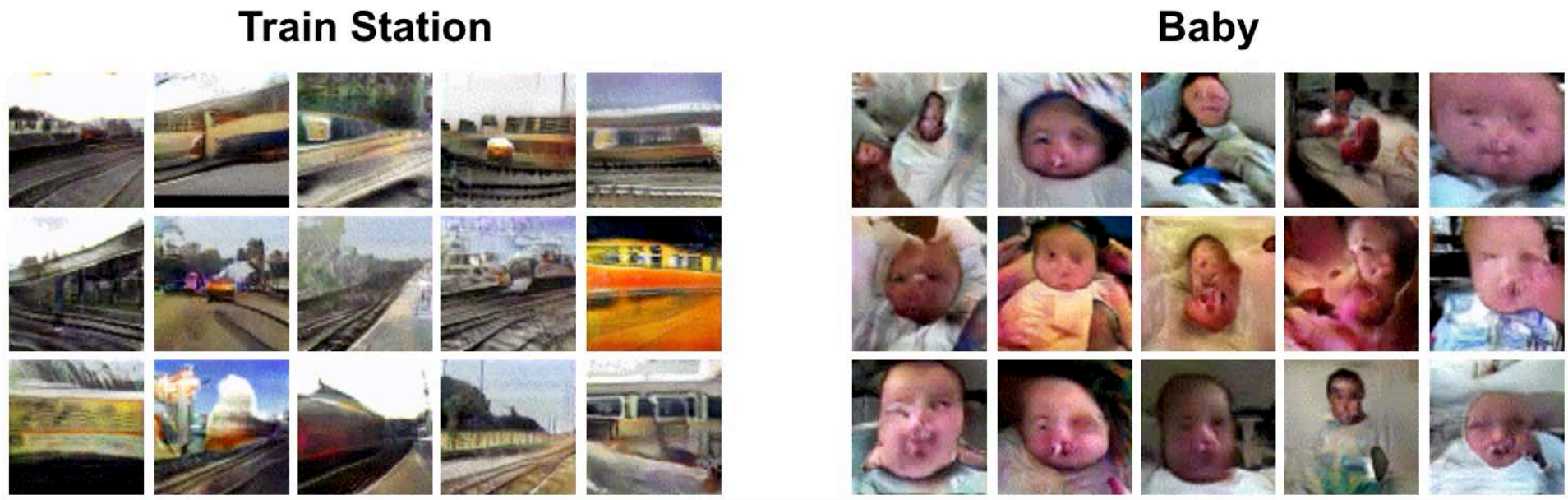
- 画像以外のドメインでは（少なくともビジョンの領域では）GANの成功例はあまり多くはない
- DCGANのような方向を決定付けるモデルの不在？



C. Vondrick et al., Generating Videos with Scene Dynamics, In NIPS, 2016

Scene Dynamics (a.k.a. VGAN) [Vondrick2016]

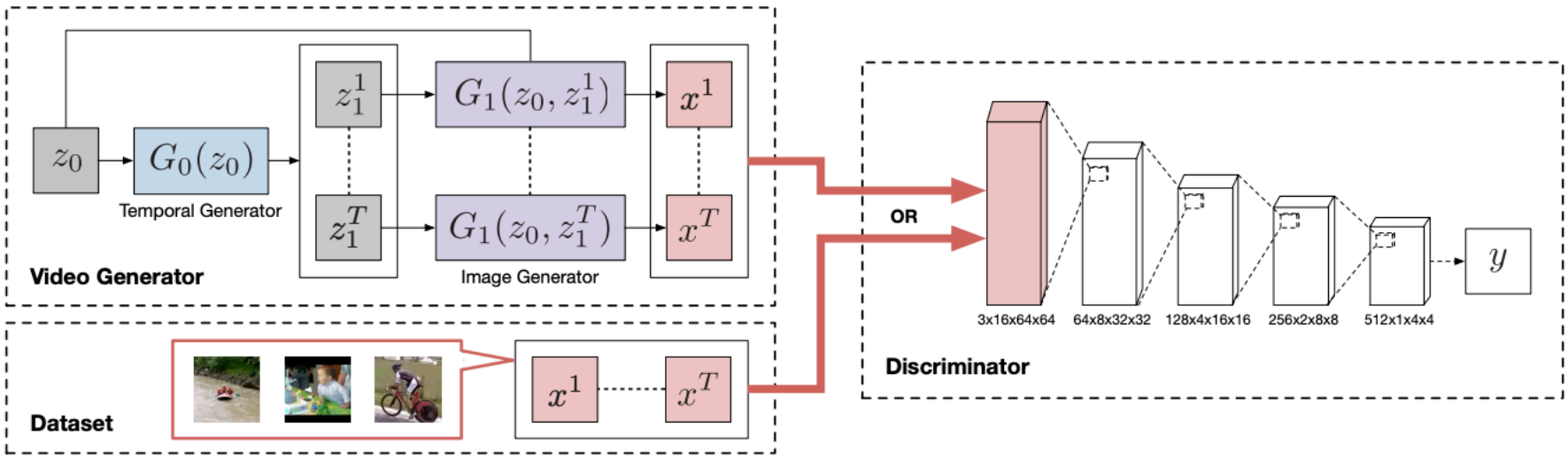
- 画像以外のドメインでは（少なくともビジョンの領域では）GANの成功例はあまり多くはない
- DCGANのような方向を決定付けるモデルの不在？



T. Wang et al., Video-to-Video Synthesis, arXiv, 2018

Temporal GAN [Saito & Matsumoto2017]

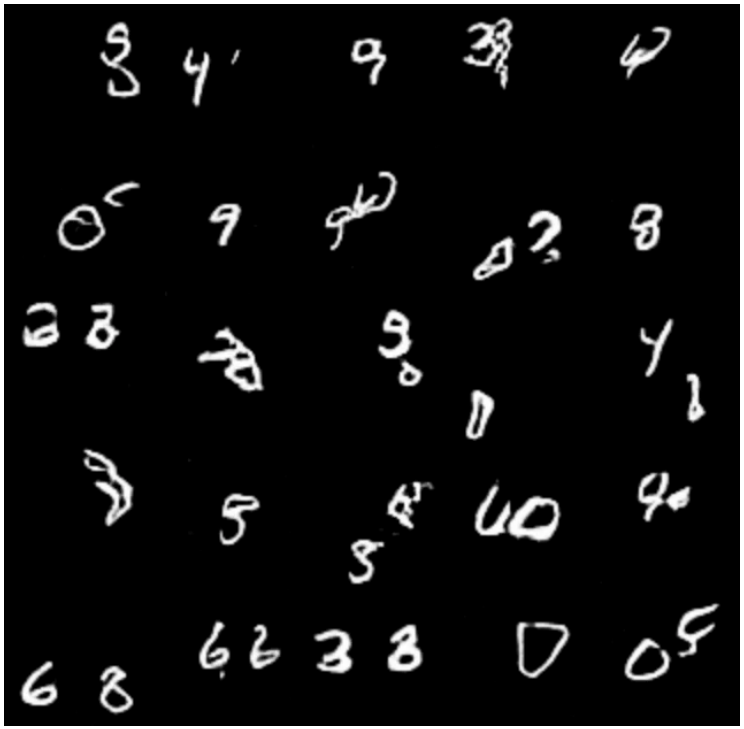
- 3次元空間を明示的に時系列と空間方向に分離して学習
  - 動画分類の領域では，時空間を明示的に分けるアプローチと
    3次元CNNで一気にやるアプローチが拮抗している印象
- Spectral Normalizationと類似のアプローチ: Singular Value Clipping
  - Spectral Normalizationのほうがいいのでそちら使うのがベター



M.Saito et al., Temporal Generative Adversarial Nets with Singular Value Clipping, In ICCV, 2017

Temporal GAN [Saito & Matsumoto2017]

- VGANよりは性能は良いがそれでもまだ大分厳しい
  - 生成に関しては時系列を3次元で表現するのは難しいのではないか



M.Saito et al., Temporal Generative Adversarial Nets with Singular Value Clipping, In ICCV, 2017

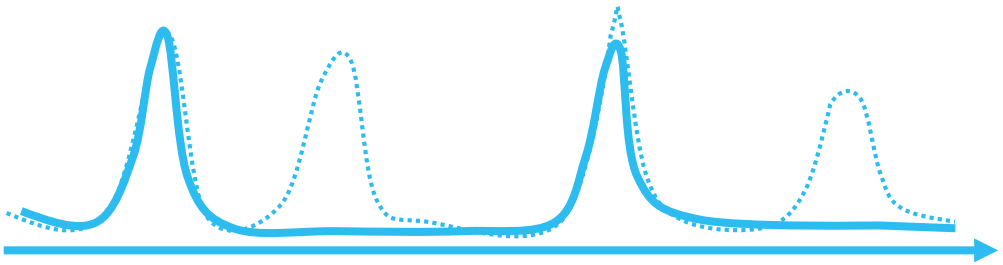
現状のまとめ

- 主観的な印象では，CVで主に盛り上がっている無教師学習はGAN
- その理由は主にDCGANの発明が大きい
- 画像の他にも動画や三次元モデルもあるものの，あまり進展していない

無教師学習のこれからの進展

モード崩壊の問題

- GANが形成するモデル分布は，自然画像が形成するデータセットの分布を完全に捉えられていない

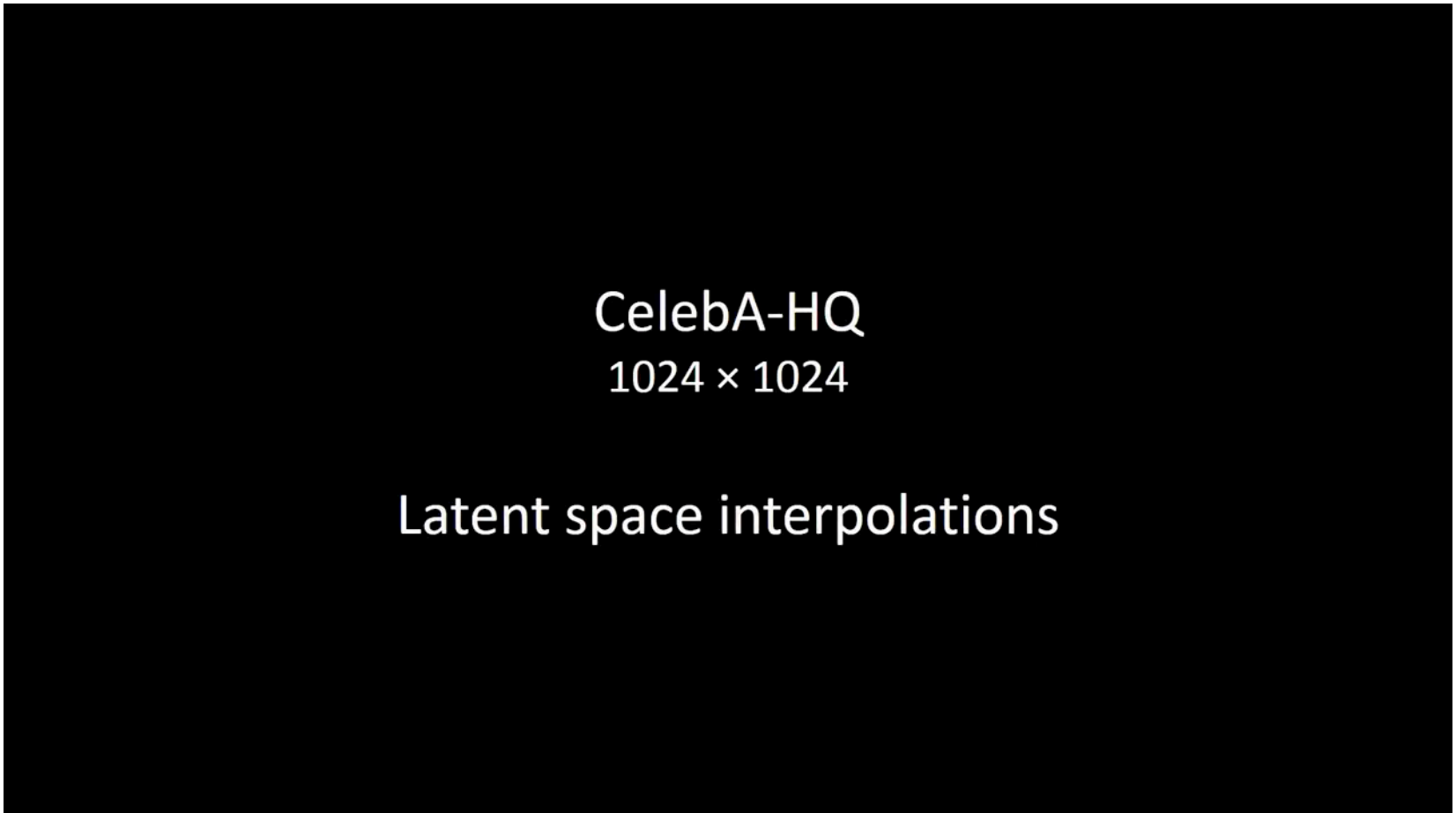


- 複数の問題が考えられる:

1. GANの手法に起因するもの (minibatch features, unrolled GAN)
2. ネットワーク構造に起因するもの
    - 特に，記憶容量について

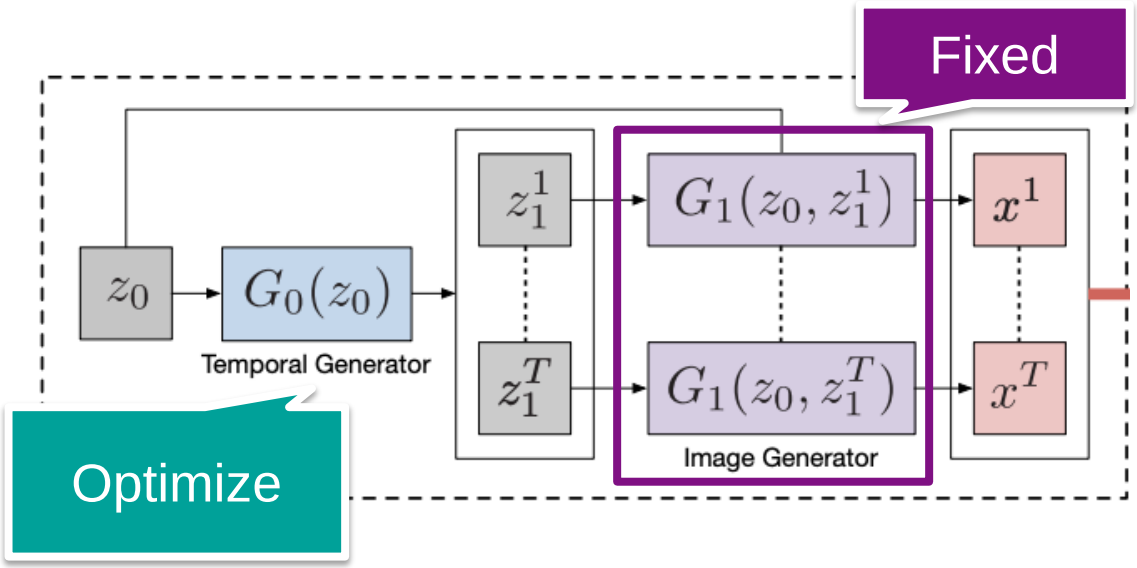
モード崩壊の問題: 画像生成の観点から

- 隠れ変数を変遷させた際のGeneratorの変遷動画は，まるで複数の静止画をモーフィングで繋いだような動画が得られる
  - 3次元回転などのセマンティックな構造は得られていない？



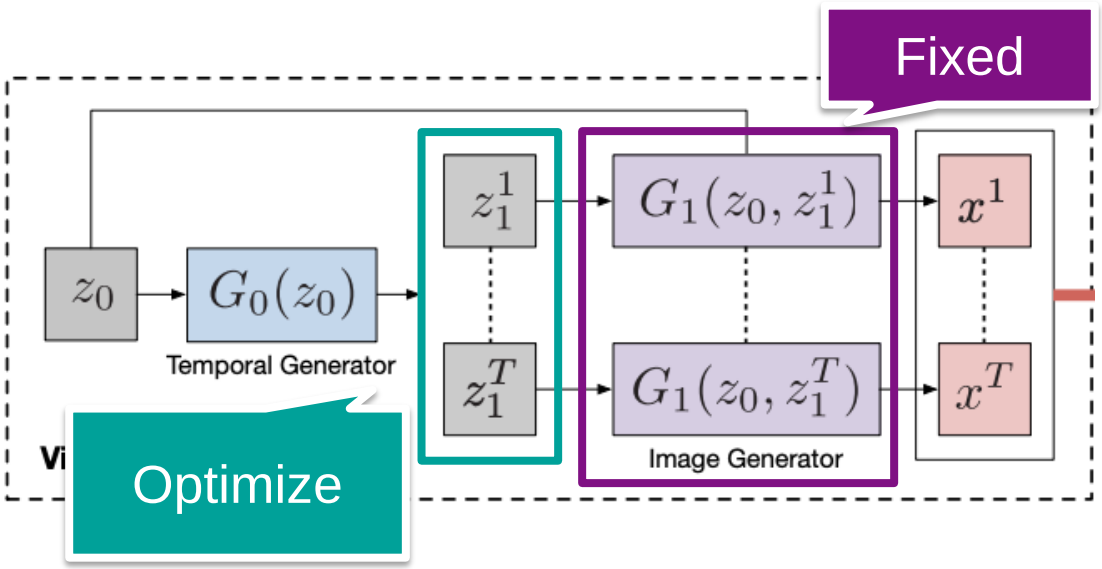
モード崩壊の問題: 動画生成の観点から(1)

- ImageNetで学習された学習済みモデルを用いて動画を生成しようと試みたが，生成された動画は悲惨なものだった
  - 画像生成のためのGeneratorは，動画を生成できるほど十分な空間を学習できていない？



モード崩壊の問題: 動画生成の観点から(2)

- 動画用のデータセットで学習されたTGANで類似のアプローチを試みた



- 結果としてはかなりきれいに復元できた
  - 高次の表現力に難がある？
    （≒ 高次情報を表現するパラメータが不足している？）
    （記憶容量仮説）

学習不安定性の問題

- （少なくとも実応用の領域では）GANの学習が不安定であるという問題に遭遇することはあまりなくなった
  - 学習を安定化させるためのトリック(WGAN-GP, SN, etc.)の発見
  - 特に，Zero-centered gradient penalty termは有効

$$R_1(\psi) := \frac{\gamma}{2} \mathrm{E}_{p_{\mathcal{D}}(x)} \left[\|\nabla D_\psi(x)\|^2 \right].$$

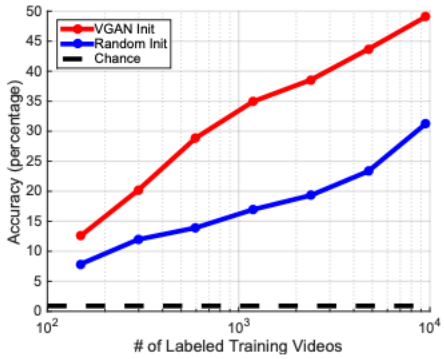
L. Mescheder et al., Which Training Methods for GANs do actually Converge?, In ICML, 2018

Discriminatorが獲得する内部表現とは？

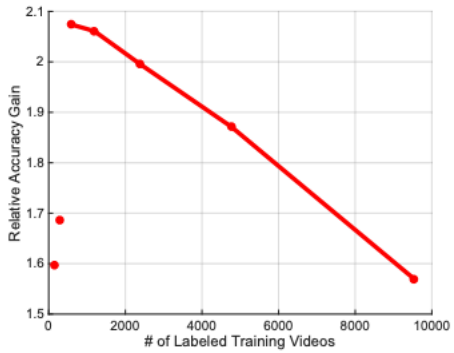
- CNNの解析に比べて，GANの特徴解析はあまり進んでいない
- Discriminatorの最終層の特徴ベクトルは他の用途にも転用できる
  - Discriminatorを利用した転移学習など
  - ただ，その精度は専用の既存手法よりは低い

Method	Accuracy
Chance	0.9%
STIP Features [36]	43.9%
Temporal Coherence [10]	45.4%
Shuffle and Learn [25]	50.2%
VGAN + Random Init	36.7%
VGAN + Logistic Reg	49.3%
VGAN + Fine Tune	**52.1%**
ImageNet Supervision [47]	91.4%

(a) Accuracy with Unsupervised Methods



(b) Performance vs # Data



(c) Relative Gain vs # Data

C. Vondrick et al., Generating Videos with Scene Dynamics, In NIPS, 2016

Discriminatorが獲得する内部表現とは？

- 「本物らしさ」の尺度を使った有意義な応用はできなかった
  - Discriminatorのスコアは人間の直感とは異なる

まとめ

- DCGAN, SNGAN, SA-GAN, BigGANの変遷について紹介
- 一方で，より複雑なドメインではあまりうまく行っていない
  - うまくいかない理由の一つに初期ブロックの表現能力の不足がある
  - バッチサイズを増やすことで部分的には改善されることが期待
  - 表現力を高める工夫，あるいは大規模モデルを効率的に学習するための手法が必要
- どのような特徴が学習されているのかの可視化の研究