DIVE INTO CODE

Learning from Simulated and Unsupervised Images through Adversarial Training

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, Russell Webb

Apple Inc

CVPR 2017, Best Paper Award

Method of improving the quality of synthetic images

Summary

By Yohei Kawakami

2019/01/16

DIVE INTO CODE

Learning from Simulated and Unsupervised Images through Adversarial Training

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, Russell Webb

Apple Inc

https://arxiv.org/abs/1612.07828

1, Summary
2, What is the simGAN…??
3, Experiments, conclusion, and discussion

Method of improving the quality of synthetic images

Summary

By Yohei Kawakami

2019/01/16

Summary

DIVE INTO CODE

Conclusion: simGANでは、教師学習のためのラベルを維持したまま、画像が生成できるため深層学習のデータ量不足問題が改善する。

Problem>>Farmer Work>>

(問題)>>深層学習などで学習させるには大規模なデータが必要
>> 教師学習ラベル(注釈)がついた多量の実データセットを得ることは難しい。
（金と時間の問題）
>> CGをつくってデータ量を増やせばよい。

(先行研究)>> 先行研究モデルGANsを使った合成された画像は現実の画像とは異なり、思ったように性能が向上しない。

(SimGAN)>> SimGANは、RefinerがCGを本物らしくし、識別器が本物の画像か、あるいはrefineされた画像かを識別する。
>>精緻化されたアウトプットによって、ニューラルネットワークを深く訓練することにより、人間による注釈の努力なしに、state-of-the-art(最新)な結果が出る。

What is this thesis for?

下記2点を同時に行うsimGANの提案。

1. CGを本物っぽくリファインする。
2. リファインされた画像かどうか識別する

Where is an important point compared to previous researches?

simGANモデルは現実の画像を使って、Refjner(合成画像のリアリティを強化するネットワーク)とDescriminater(成果を判定するネットワーク)の二つのニューラルネットワークを対抗させることで、GANモデルよりも、合成画像を改善する。

Where are the key points of technology and method?

下記の3点が、過去モデルGAN(*1)と技術的の相違点である。
①リファイナーへの入力はシミュレートされた人工画像
②敵対的損失(loss)に自己正則化(self-regularization)項を加える
③ピクセル単位で敵対的損失(adversarial loss)を求める
④過去のリファイナーの生成画像をバッチに混ぜる

How to verified whether it is valid?

実験①視線推定
視線推定の学習に使うデータセットは質が低い.
そこでSimGANでannotation付きデータを大量生成して学習させたところ,
sota達成。データセットに対するuser studyも行った.
50個の現実のデータと50個のSimGANによる生成データを
randomに提示してどちらが本物か答えさせたら正答率は51.7%だった.

実験②距離画像による手姿勢推定タスク
通常，現実の距離画像にはノイズが入っている.
そのノイズがシミュレータ画像にはないので，変換してノイズを再現.

Is there discussions?

将来的には、各合成画像に対して複数の精細画像を生成するための「ノイズ分布のモデル化」を検討し、単一画像ではなくビデオを調査する。

Which reserches should I read next?

UnityEyes[Wood et al.(2016)] : Simulater
Autoencoder[Zhang et al.(2015)]
MPIIGaze Dataset[Zhang et al.(2015)]
*1 Generative Adversarial Nets [Goodfellow et al.(2014)] : GANs
CG2Real[Johnson et al.(2011)]

What is the simGAN...??

DIVE INTO CODE

深層学習における画像データとsimGANモデルの構造

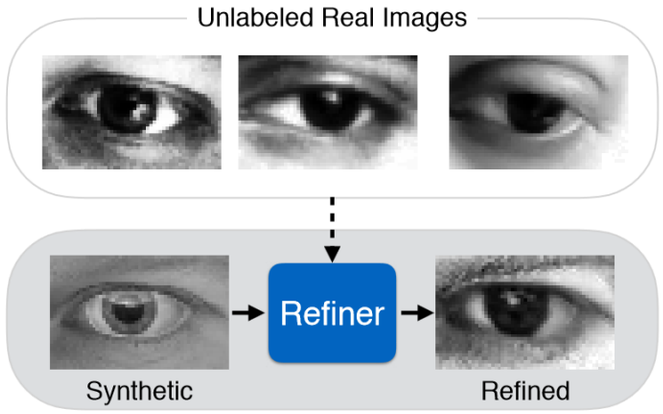


Figure 1. Simulated+Unsupervised (S+U) learning. The task is to learn a model that improves the realism of synthetic images from a simulator using unlabeled real data, while preserving the annotation information.

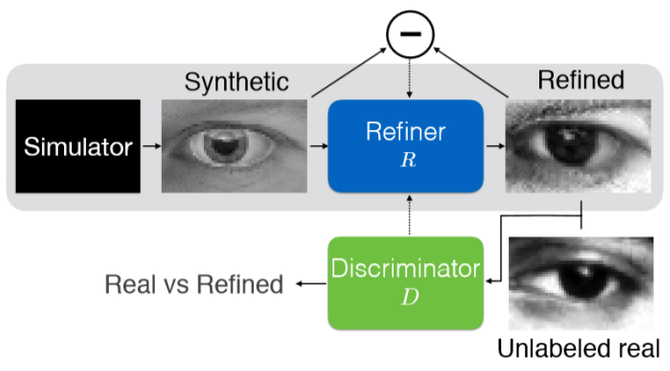


Figure 2. Overview of SimGAN. We refine the output of the simulator with a refiner neural network, R, that minimizes the combination of a local adversarial loss and a 'self-regularization' term. The adversarial loss 'fools' a discriminator network, D, that classifies an image as real or refined. The self-regularization term minimizes the image difference between the synthetic and the refined images. The refiner network and the discriminator network are updated alternately.

Figure 1) 導入

(問題)

深層学習などで学習させるには大規模なデータが必要

>> 教師学習ラベル(注釈)がついた多量の実データセットを得ることは難しい。
（金と時間の問題)

>> CGをつくってデータ量を増やせばよい。

(先行研究)

>> 先行研究モデルGANsを使った合成された画像は現実の画像とは異なり、思ったように性能が向上しない。

Figure 2) simGANの構造

1, シミュレータ(Simulator)で合成した画像（Synthetic）をリファイナーネットワーク（Refiner） に入力して本物らしい画像（Refiened)を生成する。

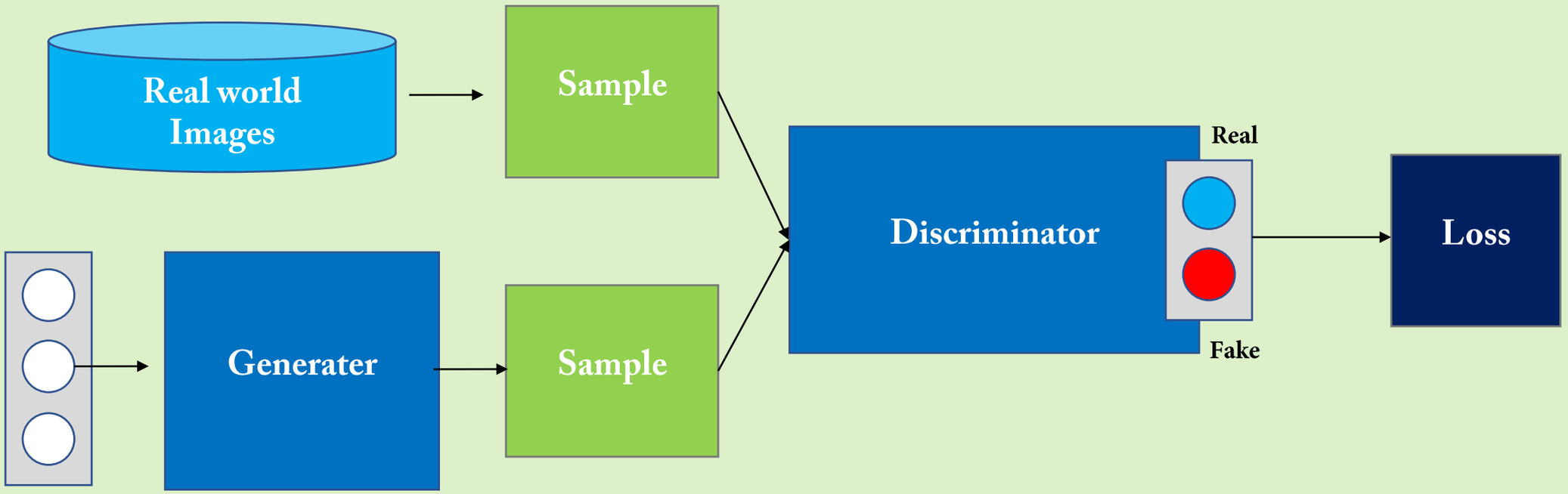
2, 識別器(Discriminator)にはrefineした画像（Refiened）と本物の画像（Unlabeled real）とで作ったミニバッチを入力し、識別させる。

3, Discriminatorではピクセルごとに交差エントロピーを求め学習させる

4, RefinerはGAN誤差と自己正則化項とで敵対的損失(loss)を求め、学習させる

DIVE INTO CODE

比較：先行研究モデルGAN(敵対的生成ネットワーク)とは…？



Generative Adversarial Nets [Goodfellow et al.(2014)]

Rf) GAN(敵対的生成ネットワーク)

互いに
例)偽札を作る偽造者と、見破る警察のような関係。偽造者は、本物と見間違える巧妙な偽造紙幣を作ること（Generator:生成器）、警察は極めて本物に見える偽造紙幣と本物の紙幣を見分ける能力を獲得すること（Discriminator:識別器）を目指す。警察は本物と偽造紙幣を見分けられるように学習、偽造者は巧妙に偽造紙幣を作るというイタチゴッコを繰り返す。
・ このように、Generator（生成器）とDiscriminator（識別器）の両者を繰り返し競わせて学習する仕組みを取り入れたのがGANの特徴

simGANとGANとの相違点①: 敵対的損失の計算式

$$\mathcal{L}_D(\phi) = -\sum_i \log(D_{\boldsymbol{\phi}}(\tilde{\mathbf{x}}_i)) - \sum_j \log(1 - D_{\boldsymbol{\phi}}(\mathbf{y}_j)). \tag{2}$$

Formula 2)識別器の敵対的損失の計算式

- $\tilde{x}_i$はrefineされた画像、yjはCGで生成していない本物の画像
- 識別器のlossは通常のGANと同様だが、交差エントロピーはピクセル単位で求め、それを合計する。
- Dのパラメータϕは,ミニバッチごとにSGDで更新する(確率的勾配降下法)

$$\mathcal{L}_R(\boldsymbol{\theta}) = -\sum_i \log(1 - D_{\boldsymbol{\phi}}(R_{\boldsymbol{\theta}}(\mathbf{x}_i))) + \lambda \|\psi(R_{\boldsymbol{\theta}}(\mathbf{x}_i)) - \psi(\mathbf{x}_i)\|_1. \tag{4}$$

Formula 4)リファイナーネットワークの敵対的損失の計算式

- １項目はGANのGeneratorと同様の誤差
- ２項目は自己正則化項を導入(L1ノルム)→注釈情報を保持するため
- 通常のGANは、それっぽい画像であればどのようなものでも生成すればいいが、simGANはラベルに沿った画像を生成する必要
- 例えば、視線推定で利用する目の画像セットを考えると、右向き、左向き、正面向きなどのラベルに対して、その画像が存在する。右向きの合成画像をリファイナーに入れて左向きの画像が出てくると困る。
  よってなんらかの制約が必要となる。これを自己正則化項で実現

- まず、合成された画像に対しニューラルネットワーク等で特徴空間へ変換
- 一方で、精緻化された画像も特徴空間へ変換
- この両者のL1ノルムをピクセル単位で求める

DIVE INTO CODE

simGANとGANとの相違点②: 領域単位の分解と過去画像からの学習

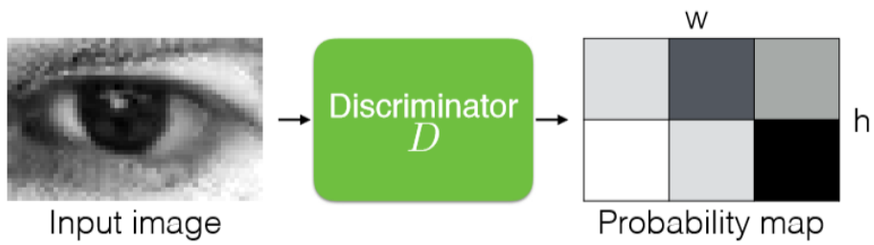


Figure 3. Illustration of local adversarial loss. The discriminator network outputs a $w \times h$ probability map. The adversarial loss function is the sum of the cross-entropy losses over the local patches.

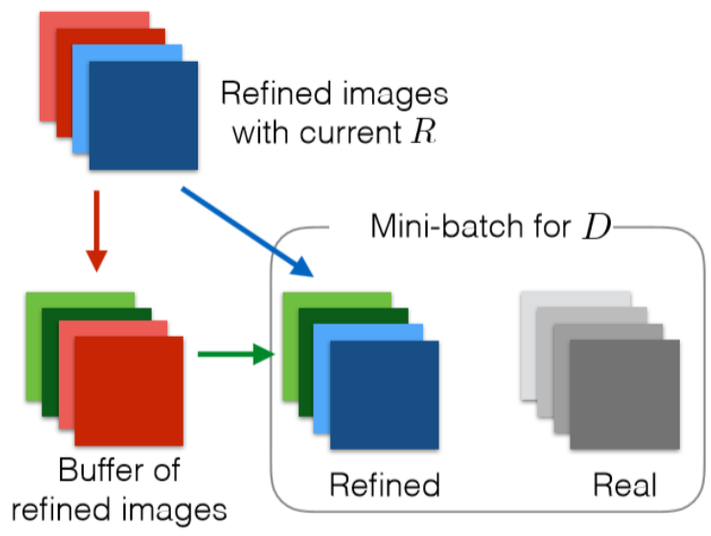


Figure 4. Illustration of using a history of refined images. See text for details.

Figure 3) 局所的敵対的損失(local adversarial loss)

・学習時に識別器へ入力画像をすべて入れるのではなく、ある領域単位に分割する

・各パッチで,現実データである確率を求め,損失関数では全領域分の交差エントロピー損失の和

Figure 4) 過去の精緻化された画像による識別器の学習

・学習されているリファイナーから得られる精緻化された画像だけでミニバッチを形成すると、損失が発散するなどの問題が発生する
・学習を安定させるため、下図のように過去に精緻化された画像を溜め込み、これを含めたミニバッチを形成する
(1)バッファBの内のb/2枚と,現在のR精製b/2枚でバッチ作成
(2)イテレーション中にバッファB中のb/2枚を現在のR精製b/2枚を交換。バッファが更新

DIVE INTO CODE

おまけ：simGANのトレーニング手順

Algorithm 1: Adversarial training of refiner network $R_{\boldsymbol{\theta}}$

Input: Sets of synthetic images $\mathbf{x}_i \in \mathcal{X}$, and real images $\mathbf{y}_j \in \mathcal{Y}$, max number of steps ($T$), number of discriminator network updates per step (K_d), number of generative network updates per step (K_g).

Output: ConvNet model $R_{\boldsymbol{\theta}}$.

for $t = 1, \ldots, T$ **do**

  for $k = 1, \ldots, K_g$ **do**

    1. Sample a mini-batch of synthetic images $\mathbf{x}_i$.
    2. Update $\boldsymbol{\theta}$ by taking a SGD step on mini-batch loss $\mathcal{L}_R(\boldsymbol{\theta})$ in (4) .

  end

  for $k = 1, \ldots, K_d$ **do**

    1. Sample a mini-batch of synthetic images $\mathbf{x}_i$, and real images $\mathbf{y}_j$.
    2. Compute $\tilde{\mathbf{x}}_i = R_{\boldsymbol{\theta}}(\mathbf{x}_i)$ with current $\boldsymbol{\theta}$.
    3. Update $\boldsymbol{\phi}$ by taking a SGD step on mini-batch loss $\mathcal{L}_D(\boldsymbol{\phi})$ in (2).

  end

end

Algorithm 1) リファイナーネットワークRθの敵対的トレーニング

[計算の構造]
・リファイナーの重みを更新する
→識別器の重みを更新する

・片方を固定して、もう片方の重みを更新する。

[計算の構造]

Input: 人工画像$x_i \in X$と現実画像$y_j \in y$のセット、最大ステップ数(T), ステップごとの識別器ネットワーク更新数(Kd),1ステップあたりの生成ネットワークの更新回数(Kg)

Output: 畳み込みニューラルネットワーク Rθ

For t = 1,…..,T do

  for k =1, … ,Kg do

    1. 人工画像X_i のミニバッチをサンプルする
    2. (4)式のミニバッチ損失関数に対し、確率的勾配降下法(SGD)を用いてθを更新する。

  end

  for k =1, … , Kd do

    1. 人工画像X_iと現実画像y_jのミニバッチをサンプルする。
    2. 現在のθで$x = R\theta(x_i)$を計算する
    3.(2)式のミニバッチ損失関数に対し確率的勾配降下法(SGD)を用いて、Φを更新する。

  end

end

Two Experiments, Conclusion, Discussion

DIVE INTO CODE

Experiments1:

実験①［視線推定］

・以下データセットに対してSimGANを使用し,評価

MPIIGaze dataset [Zhang (2015)]

UnityEyes[Wood (2016)]

・人工画像をRefineするSimGANのネットワークと目の視線方向を出力する視線推定ネットワークで実験
・MPIIGazeデータセットのラベルは未使用
・Refiner 入力データサイズ (55 x 32)

結果

最先端(state-of-the-art)の性能を達成した。

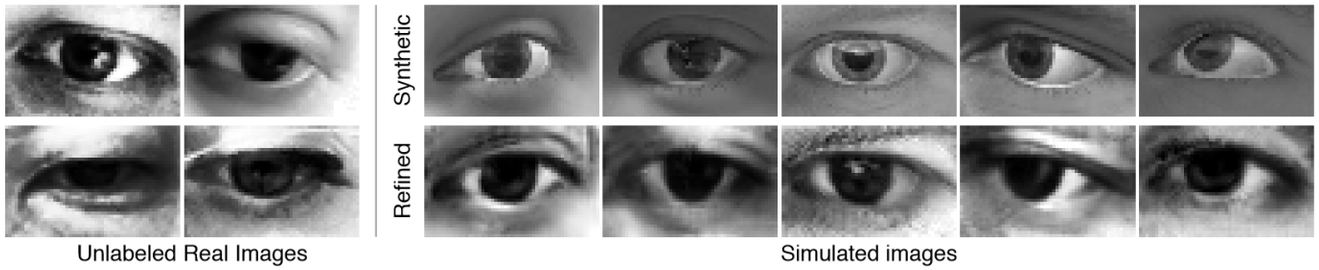


Figure 5. Example output of SimGAN for the UnityEyes gaze estimation dataset [43]. (Left) real images from MPIIGaze [47]. Our refiner network does not use any label information from MPIIGaze dataset at training time. (Right) refinement results on UnityEye. The skin texture and the iris region in the refined synthetic images are qualitatively significantly more similar to the real images than to the synthetic images. More examples are included in the supplementary material.

［白黒画像］

左側が本物の画像、右側上段が合成された画像。
右側下段が精緻化された画像

・Refineしたとき視線方向が保持
・ノイズ,皮膚テクスチャ,虹彩がより現実画像に近い
・無機的で過度に鮮明な合成画像が精緻化されることにより、肌の質感や自然なノイズを得ている。また瞳の虹彩もそれっぽくなっている

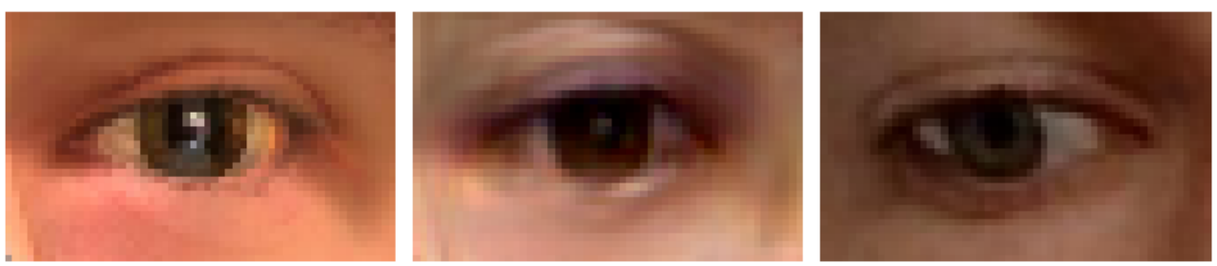


Figure 6. Self-regularization in feature space for color images.

［カラー画像］

・特徴空間の自己正則化
・下図の左から順に、人工、精製、現実
・カラー画像における精製画像
→　人工画像と現実画像の分布には、明らかな差がみられた

DIVE INTO CODE

Experiments1:

実験①［視線推定］

・以下データセットに対してSimGANを使用し,評価

MPIIGaze dataset [Zhang (2015)]

UnityEyes[Wood (2016)]

・人工画像をRefineするSimGANのネットワークと目の視線方向を出力する視線推定ネットワークで実験
・MPIIGazeデータセットのラベルは未使用
・Refiner 入力データサイズ (55 x 32)

結果

最先端(state-of-the-art)の性能を達成した。

［ビジュアルチューリングテスト(Vsiaul Turing Test)］

1, 画像を「**現実(real)画像**」か「**精製(refined)画像**」か分類させる実験
・被験者10名 現実画像50枚 精製画像50枚 計100枚の画像セットを
1枚ずつ{現実,人工}でラベル付け
→1000回試行

Accuracy = 0.517 (p=0.148)
→人間は51.7%しか現実画像と精製画像を区別できなかった

2, 画像を「**現実画像**」か「**人口画像**」かを分類させる実験
・それに対して,現実画像10枚、人工画像10枚を被験者数10人について実験
→200回試行

・Accuracy = 0.81 (p≤ 108)
→ 人間は、81.0%現実画像と人工画像を区別した

DIVE INTO CODE

Experiments1:

実験①［視線推定］

・以下データセットに対してSimGANを使用し,評価

MPIIGaze dataset [Zhang (2015)]

UnityEyes[Wood (2016)]

・人工画像をRefineするSimGANのネットワークと目の視線方向を出力する視線推定ネットワークで実験
・MPIIGazeデータセットのラベルは未使用
・Refiner 入力データサイズ (55 x 32)

結果

最先端(state-of-the-art)の性能を達成した。
（データセットを増やせば大きな改善をみせる）

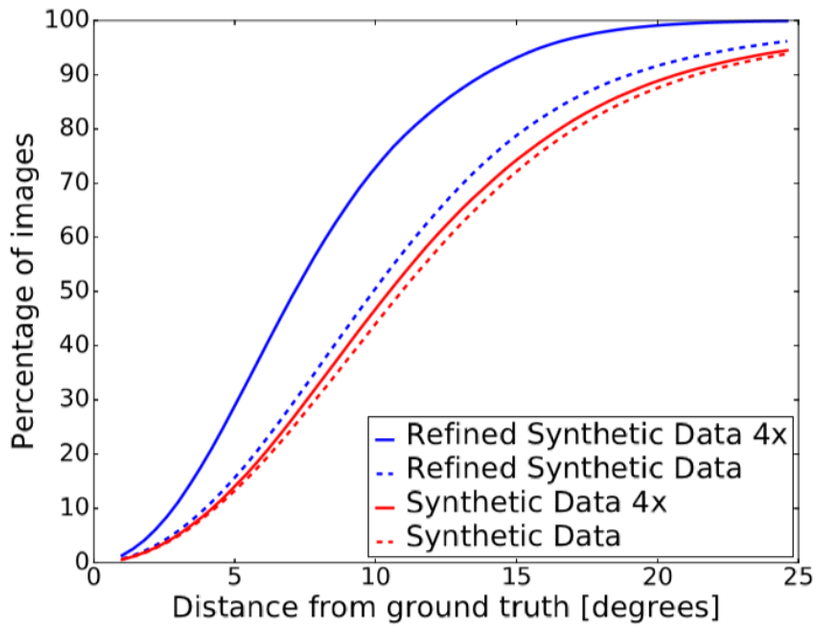


Figure 7. Quantitative results for appearance-based gaze estimation on the MPIIGaze dataset with real eye images. The plot shows cumulative curves as a function of degree error as compared to the ground truth eye gaze direction, for different numbers of training examples of data.

	Selected as real	Selected as synt
Ground truth real	224	276
Ground truth synt	207	293

Table 1. Results of the 'Visual Turing test' user study for classifying real vs refined images. The average human classification accuracy was 51.7% (chance = 50%).

Training data	% of images within d
Synthetic Data	62.3
Synthetic Data 4x	64.9
Refined Synthetic Data	69.4
Refined Synthetic Data 4x	**87.2**

Table 2. Comparison of a gaze estimator trained on synthetic data and the output of SimGAN. The results are at distance $d = 7$ degrees from ground truth. Training on the output of SimGAN outperforms training on synthetic data by 22.3%.

DIVE INTO CODE

Experiments2:

実験②［手の姿勢推定］

・NYU hand poseトレーニングセット
Stacked Hourglass Net[Yangら,2016]

と似たCNNを学習→14の手関節を学習

・NYU hand poseテストセットで評価

結果

最先端(state-of-the-art)の性能を達成した。

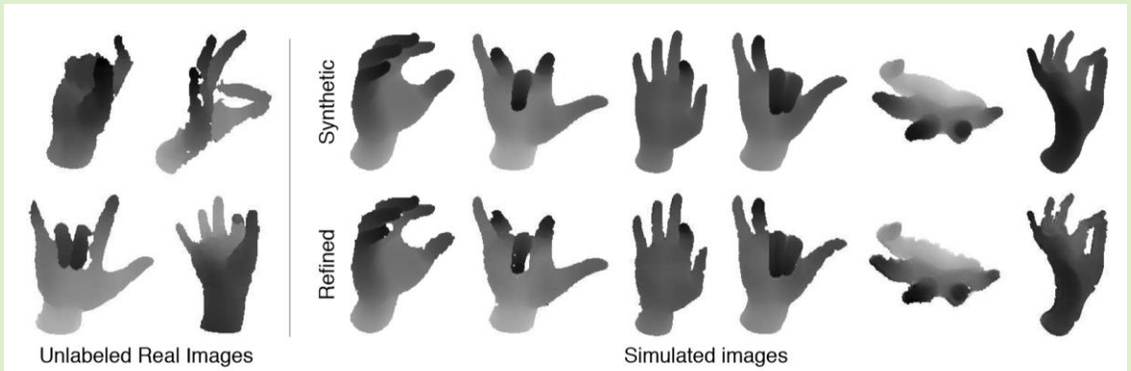


Figure 8. Example refined test images for the NYU hand pose dataset [38]. (Left) real images, (right) synthetic images and the corresponding refined output images from the refiner network. The major source of noise in the real images is the non-smooth depth boundaries that the refiner networks learns to model.

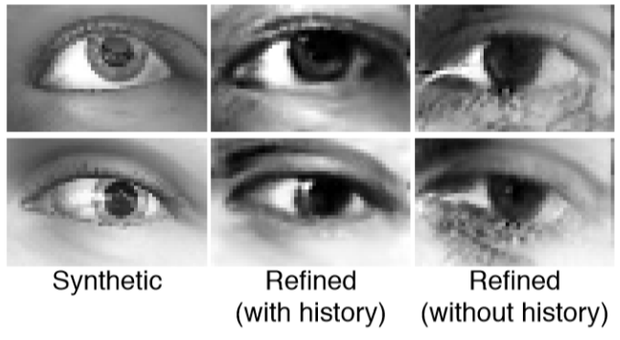


Figure 10. Using a history of refined images for updating the discriminator. (Left) synthetic images; (middle) result of using the history of refined images; (right) result without using a history of refined images (instead using only the most recent refined images). We observe obvious unrealistic artifacts, especially around the corners of the eyes.

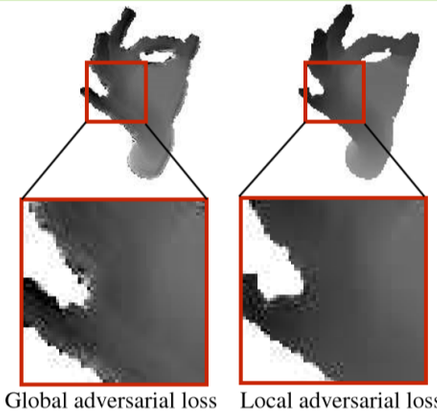


Figure 11. Importance of using a local adversarial loss. (Left) an example image that has been generated with a standard 'global' adversarial loss on the whole image. The noise around the edge of the hand contains obvious unrealistic depth boundary artifacts. (Right) the same image generated with a local adversarial loss that looks significantly more realistic.

Figure 8, 10, 11) 距離のある画像にあるノイズを再現する

通常，現実の距離画像にはノイズが入っているが、シミュレータ画像にはないので，変換してノイズを再現.

Experiments2:

実験②［手の姿勢推定］

・NYU hand poseトレーニングセット
Stacked Hourglass Net[Yangら,2016]

14の手関節を学習

・NYU hand poseテストセットで評価

結果

最先端(state-of-the-art)の性能を達成した。
(21%精度が向上した)

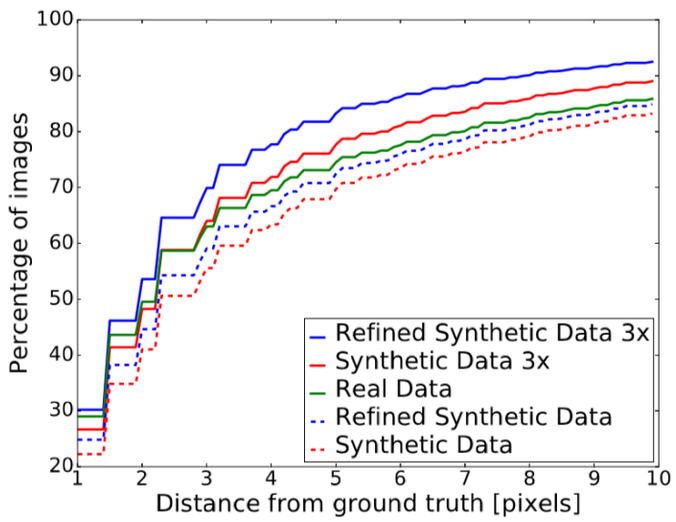


Figure 9. Quantitative results for hand pose estimation on the NYU hand pose test set of real depth images [38]. The plot shows cumulative curves as a function of distance from ground truth keypoint locations, for different numbers of training examples of synthetic and refined images.

Training data	% of images within d
Synthetic Data	69.7
Refined Synthetic Data	72.4
Real Data	74.5
Synthetic Data 3x	77.7
Refined Synthetic Data 3x	**83.3**

Table 4. Comparison of a hand pose estimator trained on synthetic data, real data, and the output of SimGAN. The results are at distance $d = 5$ pixels from ground truth.

Method	R/S	Error
Support Vector Regression (SVR) [33]	R	16.5
Adaptive Linear Regression ALR) [23]	R	16.4
Random Forest (RF) [36]	R	15.4
kNN with UT Multiview [47]	R	16.2
CNN with UT Multiview [47]	R	13.9
k-NN with UnityEyes [43]	S	9.9
CNN with UnityEyes Synthetic Images	S	11.2
CNN with UnityEyes Refined Images	S	**7.8**

Table 3. Comparison of SimGAN to the state-of-the-art on the MPIIGaze dataset of real eyes. The second column indicates whether the methods are trained on Real/Synthetic data. The error the is mean eye gaze estimation error in degrees. Training on refined images results in a 2.1 degree improvement, a relative 21% improvement compared to the state-of-the-art.

Summary

DIVE INTO CODE

Conclusion: simGANでは、教師学習のためのラベルを維持したまま、画像が生成できるため深層学習のデータ量不足問題が改善する。

Problem>>Farmer Work>>

(問題)>>深層学習などで学習させるには大規模なデータが必要
>> 教師学習ラベル(注釈)がついた多量の実データセットを得ることは難しい。
（金と時間の問題）
>> CGをつくってデータ量を増やせばよい。

(先行研究)>> 先行研究モデルGANsを使った合成された画像は現実の画像とは異なり、思ったように性能が向上しない。

(SimGAN)>> SimGANは、RefinerがCGを本物らしくし、識別器が本物の画像か、あるいはrefineされた画像かを識別する。
>>精緻化されたアウトプットによって、ニューラルネットワークを深く訓練することにより、人間による注釈の努力なしに、state-of-the-art(最新)な結果が出る。

What is this thesis for?

下記2点を同時に行うsimGANの提案。

1. CGを本物っぽくリファインする。
2. リファインされた画像かどうか識別する

Where is an important point compared to previous researches?

simGANモデルは現実の画像を使って、Refjner(合成画像のリアリティを強化するネットワーク)とDescriminater(成果を判定するネットワーク)の二つのニューラルネットワークを対抗させることで、GANモデルよりも、合成画像を改善する。

Where are the key points of technology and method?

下記の3点が、過去モデルGAN(*1)と技術的の相違点である。
①リファイナーへの入力はシミュレートされた人工画像
②敵対的損失(loss)に自己正則化(self-regularization)項を加える
③ピクセル単位で敵対的損失(adversarial loss)を求める
④過去のリファイナーの生成画像をバッチに混ぜる

How to verified whether it is valid?

実験①視線推定
視線推定の学習に使うデータセットは質が低い.
そこでSimGANでannotation付きデータを大量生成して学習させたところ,
sota達成。データセットに対するuser studyも行った.
50個の現実のデータと50個のSimGANによる生成データを
randomに提示してどちらが本物か答えさせたら正答率は51.7%だった.

実験②距離画像による手姿勢推定タスク
通常，現実の距離画像にはノイズが入っている.
そのノイズがシミュレータ画像にはないので，変換してノイズを再現.

Is there discussions?

将来的には、各合成画像に対して複数の精細画像を生成するための「ノイズ分布のモデル化」を検討し、単一画像ではなくビデオを調査する。

Which reserches should I read next?

UnityEyes[Wood et al.(2016)] : Simulater
Autoencoder[Zhang et al.(2015)]
MPIIGaze Dataset[Zhang et al.(2015)]
*1 Generative Adversarial Nets [Goodfellow et al.(2014)] : GANs
CG2Real[Johnson et al.(2011)]

DIVE INTO CODE

Learning from Simulated and Unsupervised Images through Adversarial Training

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, Russell Webb

Apple Inc

CVPR 2017, Best Paper Award

Method of improving the quality of synthetic images

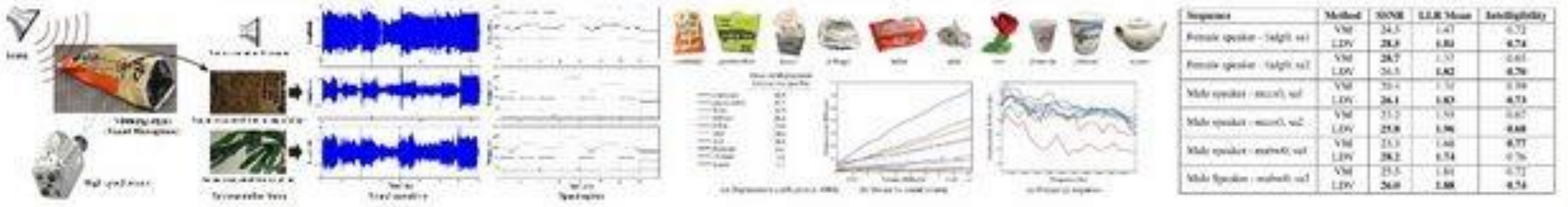
Summary

By Yohei Kawakami

2019/01/16

The Visual Microphone: Passive Recovery of Sound from Video

Abe Davis[1] Michael Rubinstein[2,1] Neal Wadhwa[1] Gautham J. Mysore[3] Frédo Durand[1] William T. Freeman[1]



どんなもの？

高速カメラの映像からその場所でかかっていた音を復元する．一眼レフを使った例も検証した．

先行研究と比べてどこがすごい？

レーザー手法（レーザードップラー）は80年代からあったが，本手法ではレーザーを当てなくてもハイスピード動画から音を復元できる．

技術や手法のキモはどこ？

動画から微細な変化を検出する手法
[Wu et al 2012]や他を音声に応用

どうやって有効だと検証した？

レーザードップラーと比べても，有用なデータが出た．統計的誤差を比較した．
スピーチを復元してみて周波数分布を比べてみた．
音響解析してみて振動範囲を調べた．

議論はある？

軽くて硬いものは精度よく復元できる．
光を当てないでよい分レーザーより便利かもしれない．

次に読むべき論文は？

Wu, et al 2012, かな