

# CVPR2017速報

---

片岡 裕雄, 原健翔, 阿部香織, Yue Qiu  
鈴木亮太, 鈴木智之, 大喜周平, 張雨辰

<https://sites.google.com/site/cvpaperchallenge/>

# 概要

---

## CV分野のトップ会議CVPR2017の参加速報

- CVPR2016速報、ECCV2016速報をベースに作成  
※ SlideShare (<https://www.slideshare.net/HirokatsuKataoka/cvpr-2016>)より
- 速報性を重視したためメモ程度であることにご注意
- 全ての論文に目を通しているわけでは無いが、著者らができる限り聴講して議論を行った
- 事前の論文読み会も実施しています
- やはりDeep Neural Networks (DNN)の話が中心

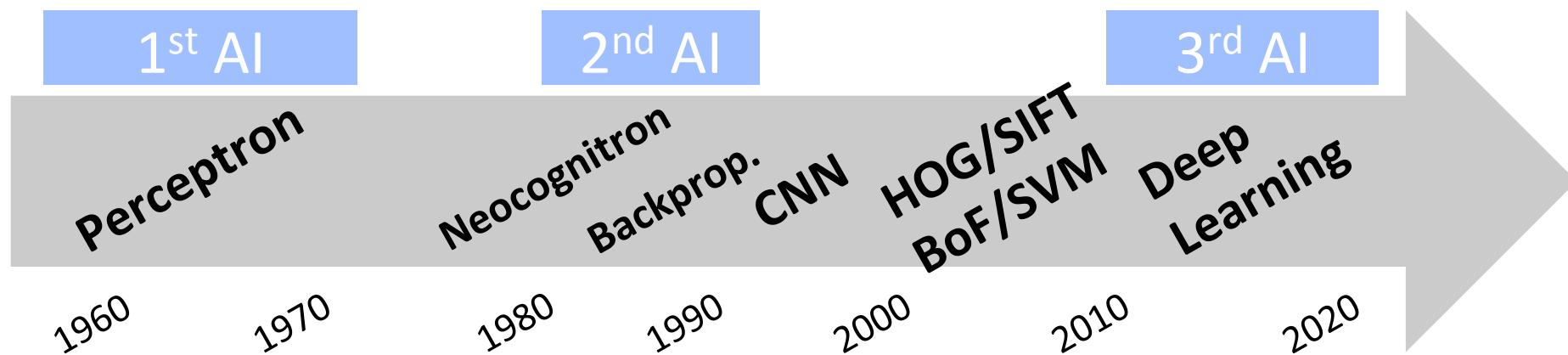
その前に（かなり大雑把な）DNNの概要

---

# DNNの動向 (1/8)

## DNN時代以前の動向

- Perceptron, MLP, Neocognitron, BackProp, CNN
- DNNが流行る前の画像認識では局所特徴が使用



F. Rosenblatt et al. "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms" in 1961.

Rumelhart et al. "Learning representations by back-propagating errors" in Nature 1986.

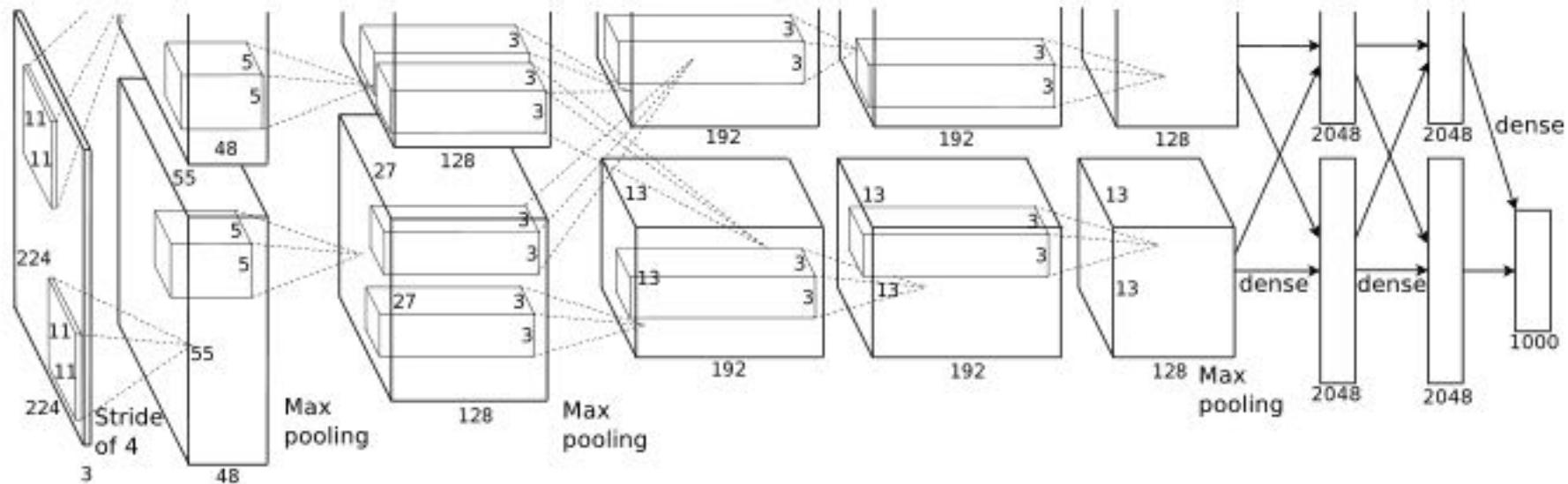
K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position", in 1980

Y. LeCun et al. "Gradient-based learning applied to document recognition" in IEEE 1998.

# DNNの動向 (2/8)

## ILSVRCを発端とする画像識別タスクへの応用

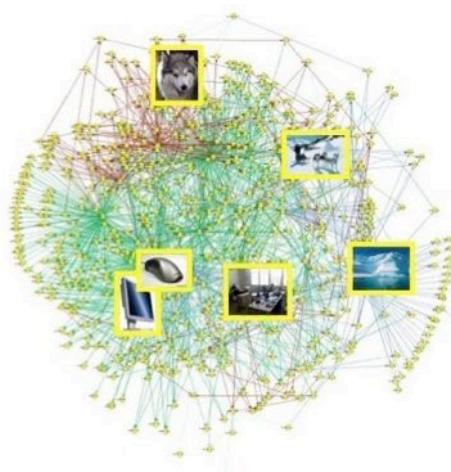
- AlexNet @画像認識コンペILSVRC2012
  - 第一著者Alexさんのネットワーク（仕掛け人のHintonNetになってたかも？）
- 背景にはBelief Propagation, ReLU, SGD, Dropoutなど構造をDEEPにする技術が揃ってきた



# DNNの動向 (3/8)

## DNNが勝てた背景

- ImageNet! (データが最も重要)
- NVIDIA! (圧倒的な計算力)



IMAGENET

<http://www.image-net.org/>



NVIDIA®

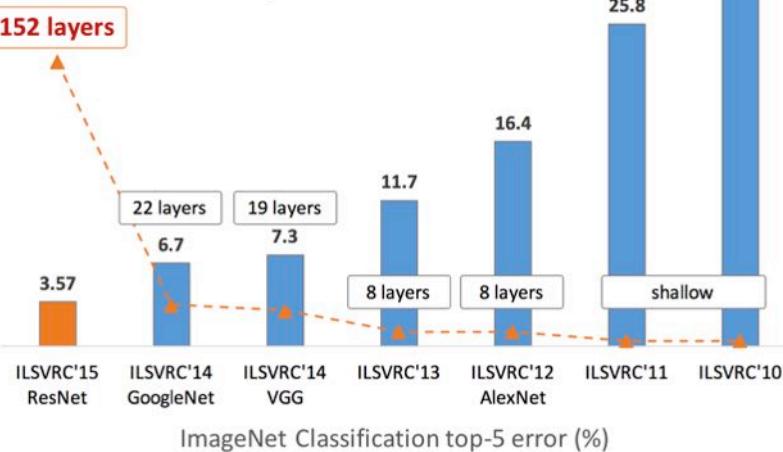
<http://cvpr2017.thecvf.com/>

# DNNの動向 (4/8)

## 構造の深化

- 2014年頃から「構造をより深くする」ための知見が整う
- 現在（主に画像識別で）主流なのはResidual Network

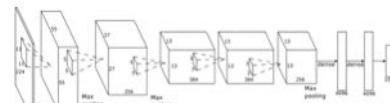
Revolution of Depth



ImageNet Classification top-5 error (%)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.

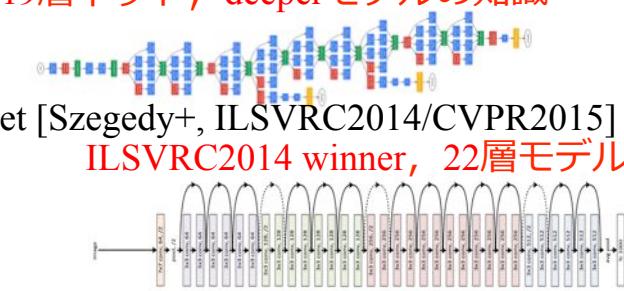
AlexNet [Krizhevsky+, ILSVRC2012]  
ILSVRC2012 winner, DLの火付け役



VGGNet [Simonyan+, ILSVRC2014]  
16/19層ネット, deeperモデルの知識



GoogLeNet [Szegedy+, ILSVRC2014/CVPR2015]  
ILSVRC2014 winner, 22層モデル

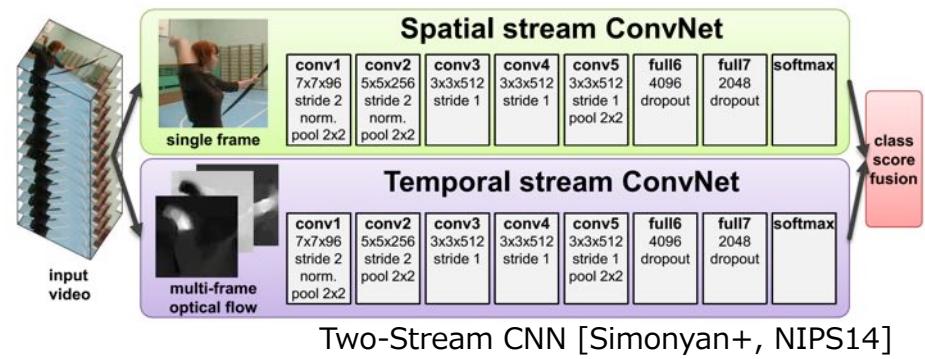
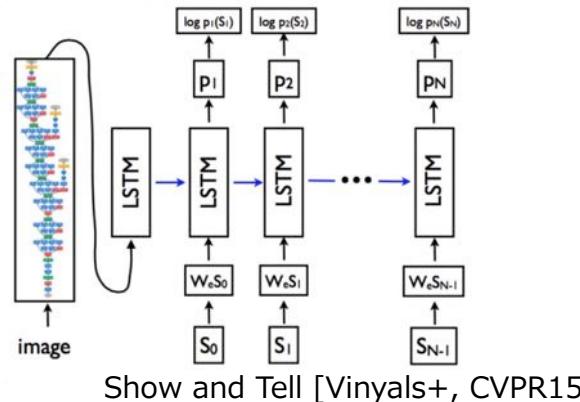
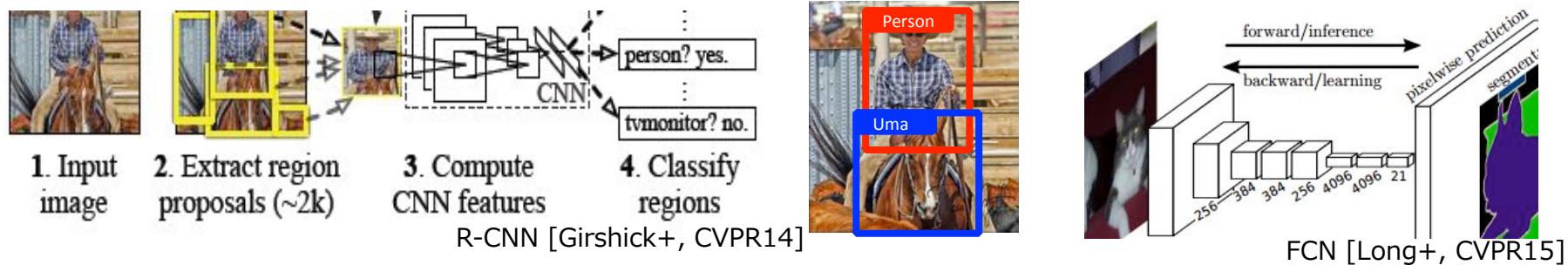


ResNet [He+, ILSVRC2015/CVPR2016]  
ILSVRC2015 winner, 152層！(実験では $10^3$ +層も)

# DNNの動向 (5/8)

## 他タスクへの応用 (画像認識・動画認識)

- R-CNN: 物体検出
- FCN: セマンティックセグメンテーション
- CNN+LSTM: 画像説明文
- Two-Stream CNN: 動画認識

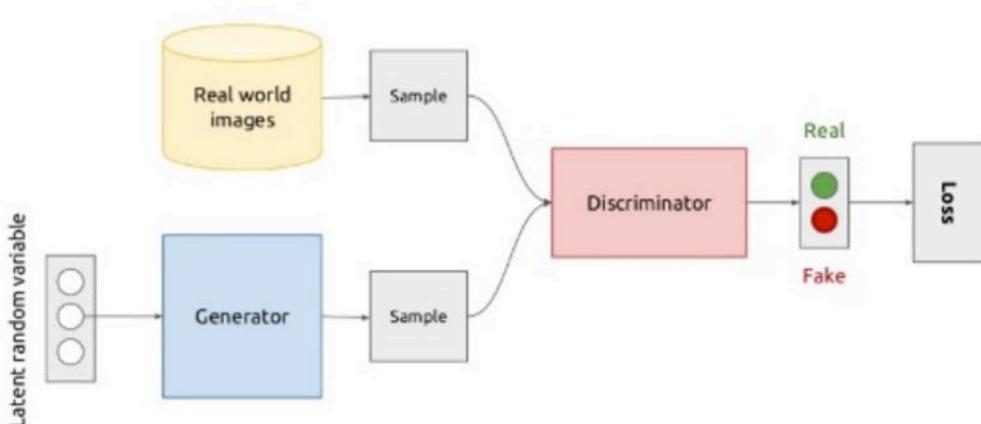


# DNNの動向 (6/8)

## 画像生成・強化学習への移行

- GAN: 敵対的学習, Generator (G) と Discriminator (D) が競い合いリアルな画像を生成
- DQN: 強化学習モデル

### Adversarial Learning



# DNNの動向 (7/8)

---

## DNNのフレームワークが次々にリリース

- Caffe/Caffe2, Theano, Chainer, TensorFlow, Keras, Torch/PyTorch, MatConvNet, Deeplearning4j, CNTK, MxNet, Lasagne  
(順不同, その他多数)
- 特に, Caffeが出てきてからCVにおけるDNNの研究は爆発的に広がった

# DNNの動向 (8/8)

---

現在も進化の一途を辿り、社会実装が進む

- 自動運転/ADAS
- ロボティクス
- ファッション
- 画像/動画検索
- 物流（ピッキング等）
- 等

研究者としては「こんなこともできる」を世に出したい

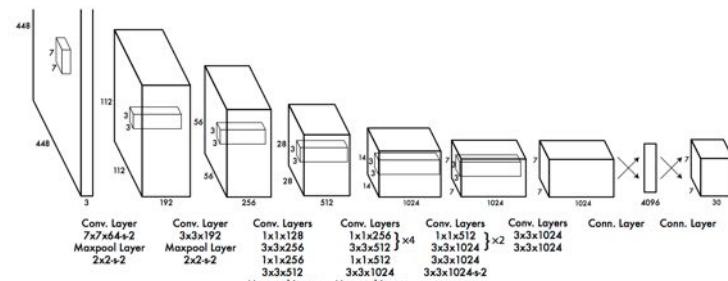
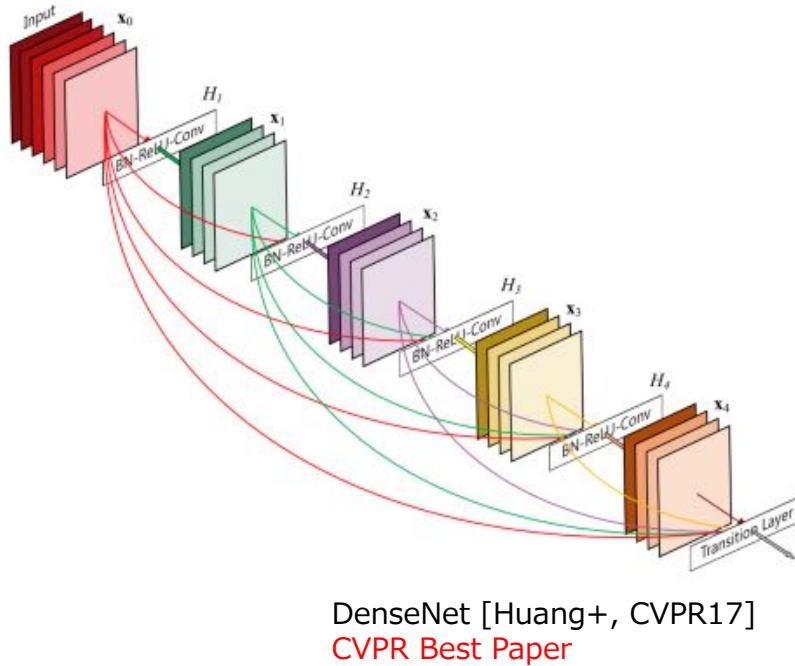
## CVPR 2017の動向・気付き

---

- 今回どんな研究が流行っていた？
- 海外の研究者は何をしている？
- 「動向」や「気付き」をまとめました

# CVPR2017の動向・気付き (1/19)

- (DNNは) タスク特化に加えて高精度化
  - 精度を改善するための取り組み
  - DNN初期からある画像識別/物体検出を例にすると…



	YOLO	YOLOv2							
batch norm?	✓	✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?		✓	✓	✓	✓	✓	✓	✓	✓
convolutional?		✓	✓	✓	✓	✓	✓	✓	✓
anchor boxes?		✓	✓						
new network?			✓	✓	✓	✓	✓	✓	✓
dimension priors?				✓	✓	✓	✓	✓	✓
location prediction?					✓	✓	✓	✓	✓
passthrough?						✓	✓	✓	✓
multi-scale?							✓	✓	✓
hi-res detector?								✓	✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	78.6

YOLO\_v2 [Redmon+, CVPR17]  
CVPR Honorable Mention Award

# CVPR2017の動向・気付き (2/19)

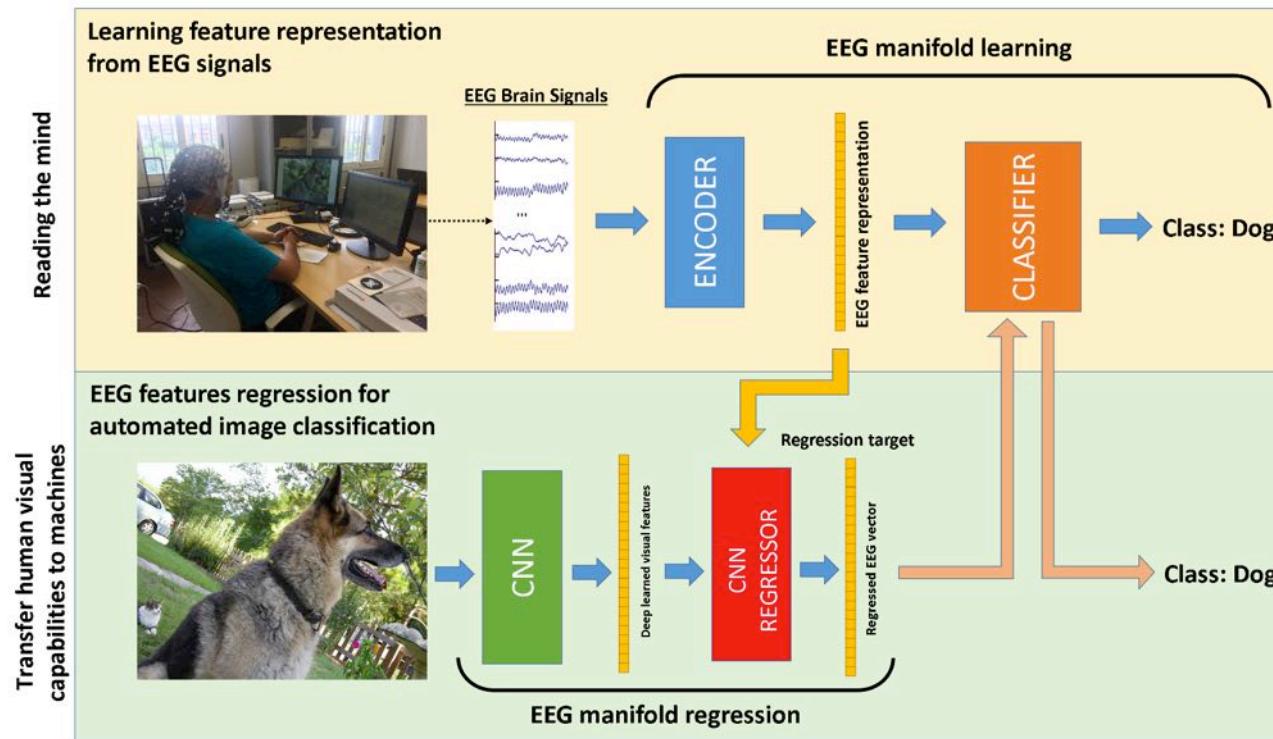
---

- データベースの大規模化はある程度収束
  - 「ただ大きい！」だけのデータは必要ない？
  - 意義のあるデータ/クリーンなデータが重要と分かってきた
  - 弱いラベル/ラベル無しデータでの手法を考案も評価高

# CVPR2017の動向・気付き (3/19)

## – DNNの解明に向けた研究

- 中間層の挙動解明 (e.g. Network Dissection)
- クラス間の境界を探る (e.g. Perturbation)
- 脳の信号から画像識別する研究も見られた (Deep Learning Human Mind; 下図)



# CVPR2017の動向・気付き (4/19)

---

## – 大規模な学習を実現するための研究

- Weakly Supervised Learningによる比較的容易に取得可能な教師データ（画像単位，クリックなど）からの学習（21件）
- CGなどから人工的な学習データを生成（静止画，動画問わず）
- より実データに近い人工データをGANなどにより生成
- ラベル無しデータにラベルを付与しながら学習（Self-supervised）

# CVPR2017の動向・気付き (5/19)

---

## - 従来のタスクをベースに複雑化

- 多人数を対象にリアルタイム処理
- より多様（姿勢，環境...）で実世界に近いデータを対象
- 複数のタスクを単一のフレームワークで処理

新しいデータセットやアノテーションを用意して  
少しでも独自の設定にしようという工夫

# CVPR2017の動向・気付き (6/19)

---

## - ボリュームデータへの移行

- ボリュームデータとは、画像xy空間から新しい次元を追加
- 奥行Zを含めた実空間 (XYZ) , 時間軸Tを含めた動画像 (XYT) 等
- 基礎的なデータや手法の整備をまずは進めている

# CVPR2017の動向・気付き (7/19)

---

## – GAN/DQNの効果的な利用

- Apple: 合成データ (Synthetic Data) をリアルにする仕組みを導入してリアル画像と見分けがつかない画像へと変換する (Best Paper)
- Google: ドメイン変換をGANで実現
- Snap&Google: 画像説明文のパラメータを強化学習で強くする
- アイディア次第

# CVPR2017の動向・気付き (8/19)

---

## – 根強く残る分野へ攻める

- Computational Photography, カメラ幾何等の純CV
- まだDeep Learningが入り込んでいない（というか入らなくても良い分野ももちろんある！）
- Machine Learningのセッションと比較して人数が少ない（= 査読を突破する可能性も上がる？）

今の時代、（逆に）DNNをやらない研究の方が通しやすい？

# CVPR2017の動向・気付き (9/19)

---

- 徹底的な比較論文は通る
  - Speed/Accuracy Trade-off (Google)
  - HPPatch: 局所特徴量のベンチマークと徹底比較

# CVPR2017の動向・気付き (10/19)

---

- ある意味万能・多機能なネットワークを構成
  - UNet, pix2pix

# CVPR2017の動向・気付き (11/19)

---

- マルチモーダル・マルチタスクを扱うDNNが増加
  - DNNが縦・横に伸びていく
    - ResNetのSkippingConnectionの成功を受けて？
    - GPUの高性能化（特にメモリ増大）の影響大？
  - データセットの多様化
    - 応用的課題に使えるデータセットの出現を受けた研究が出現
  - 異なる機能のモジュールを結合しても十分使える
    - タイトルを「Joint」で検索すると...
      - » CVPR2017 21件
      - » CVPR2016 16件
      - » CVPR2015 17件
      - » CVPR2014 15件
  - DNNの多分野利用の加速

# CVPR2017の動向・気付き (12/19)

---

- Oral, Spotlight, Posterのどれでも、十中八九、DNNの構造をいじってstate-of-the-artという新規性の主張
  - 手法的新規性の重視
  - 何やってもうまくいくので手法的新規性が出しやすいせい
  - DNNの構造が最大のハイパーパラメータ
- 既存技術の性質の調査もマイナーながらある
  - 問題設定、論文の書き方がうまい？ = ハードルが高い？

# CVPR2017の動向・気付き (13/19)

- 事前に論文の宣伝を行うことも重要？

- 每回通している人に聞いてもarXivに載せることが重要と話している（うまく活用している）
- 今回、YOLO9000やpix2pixの戦略が参考になった（真似できるかどうかは置いといて。。。）
  - 何れもコード公開，arXivへの論文投稿はもちろん，Twitterでも効果的に拡散させていた
  - 特に，YOLO9000の圧倒的な開発力は評価に値するし，ポスター（下図）はズルい



<https://pbs.twimg.com/media/DFnnqn0VwAExuiU.jpg>

# CVPR2017の動向・気付き (14/19)

---

- アイディアや研究テーマが似通ってきた?
  - 同データ (ImageNet/PascalVOC等) で同手法 (ResNets/R-CNN/FCN等) をベースにすると必然的にアイディアは重なる
  - (重要 1) 完成度を高めないと取り組み自体消される
  - (重要 2) 問題設定・手法ともにBrave Newなアイディアが重宝

# CVPR2017の動向・気付き (15/19)

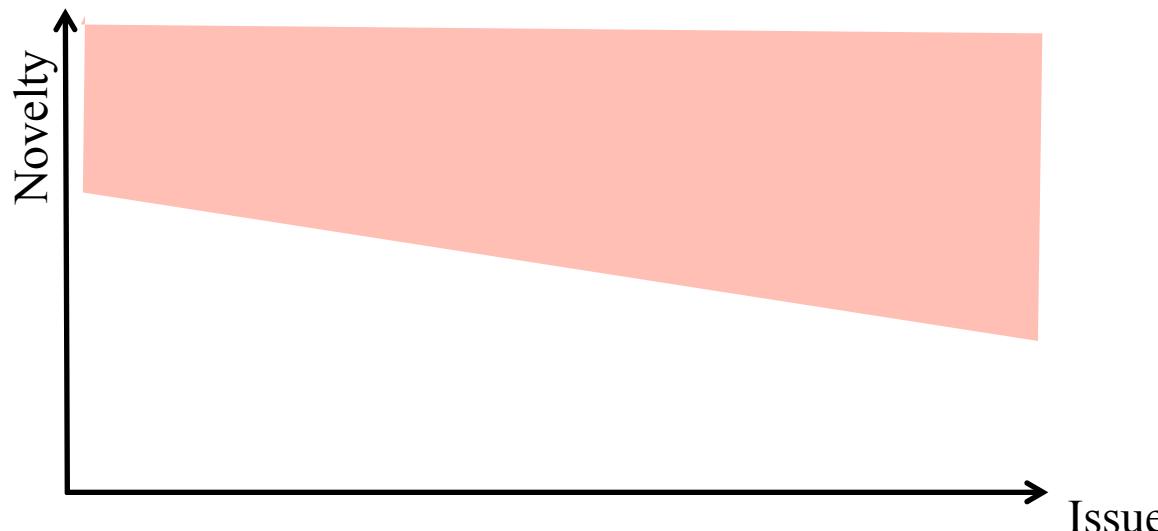
---

- 思いついたアイディアはその年のうちにやる
  - 次の年に同じような研究が出てくる（可能性が高い）
  - 我々のグループ内の（ブレストにより出てきた）アイディアが今年のCVPRでも散見された
  - 誰でも思いつきそうることは世界中で同時多発的に研究される

# CVPR2017の動向・気付き (16/19)

## - どんな論文が通っている？

- (当たり前だが) 手法として新規性のある手法が通る
- データセットだけを提案する論文はよほど面白みがないと通らない（らしい）
- 問題設定をIssue, 手法をNoveltyとすると（直感的に）下記エリア
  - Novelty重視



# CVPR2017の動向・気付き (17/19)

---

- 全部読むより現場に行こう！
  - 当グループは2015年, CVPR論文を全て読んだが,  
**「現場に来て情報収集」** する方が良い！
  - さらに言うと複数人で来て議論を行う
    - 情報収集を共有（複数人で来て自然に議論している研究グループは強い）
    - 論文の著者と直接話す
    - 欲を言えばその場で研究戦略やテーマを考案

# CVPR2017の動向・気付き (18/19)

---

- 分野間がシームレスになる傾向？

- これまでよりもアイディアの重なりが大きい (CVで戦うツラさ)
- DNNがツールとして整い他分野の研究者がCVPRに投稿  
=> 逆を言うとDNNを武器に他分野で発表するチャンス？

# CVPR2017の動向・気付き (19/19)

---

- そろそろ次の時代へ？

- 生の声を聞いているとDNNそろそろ飽きたという声もちらほら
  - パラメータ調整問題
  - 複雑ネットワーク構築問題
  - 大規模データセットアノテーション問題 等
- (去年くらいから) Post-DNNを探している研究グループもあるので  
は？

## これから引用されそう（流行りそう）な論文

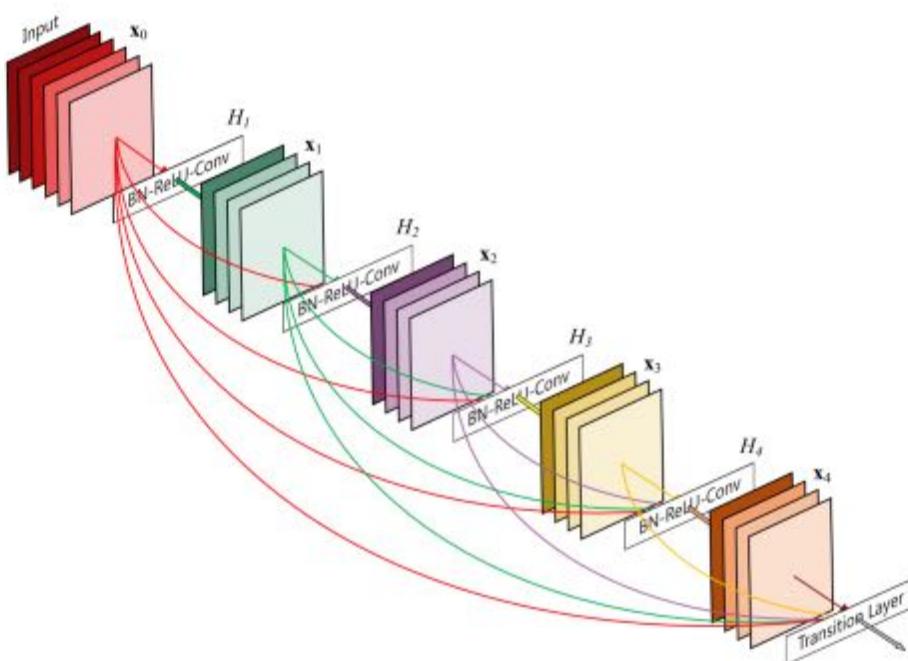
---

- すでに引用されている論文も含みます

# 引用されそうな論文 (1/18)

## DenseNets (CVPR Best Paper)

- ResNetのコネクションを密に結合して精度向上
- すでに多階層にして効率化する手法を提案
- オリジナルのDenseNetsは既にライブラリ導入済み
  - e.g. Pytorch: <https://github.com/pytorch/vision/blob/master/torchvision/models/densenet.py>

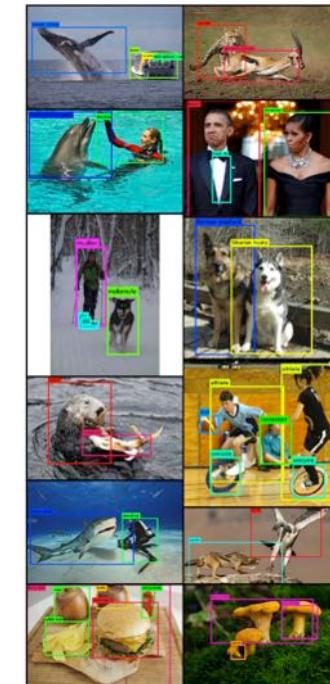


# 引用されそうな論文 (2/18)

## YOLO9000 (Honorable Mention Award)

- 物体検出手法YOLO [Redmon+, CVPR16]の改善
- 網羅的に下記の項目を検証
  - バッチ正規化, 高解像の入力, スキップコネクション 等

	YOLO									YOLOv2
batch norm?	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
hi-res classifier?		✓	✓	✓	✓	✓	✓	✓	✓	✓
convolutional?			✓	✓	✓	✓	✓	✓	✓	✓
anchor boxes?				✓	✓					
new network?					✓	✓	✓	✓	✓	✓
dimension priors?						✓	✓	✓	✓	✓
location prediction?							✓	✓	✓	✓
passthrough?								✓	✓	✓
multi-scale?									✓	✓
hi-res detector?										✓
VOC2007 mAP	63.4	65.8	69.5	69.2	69.6	74.4	75.4	76.8	78.6	

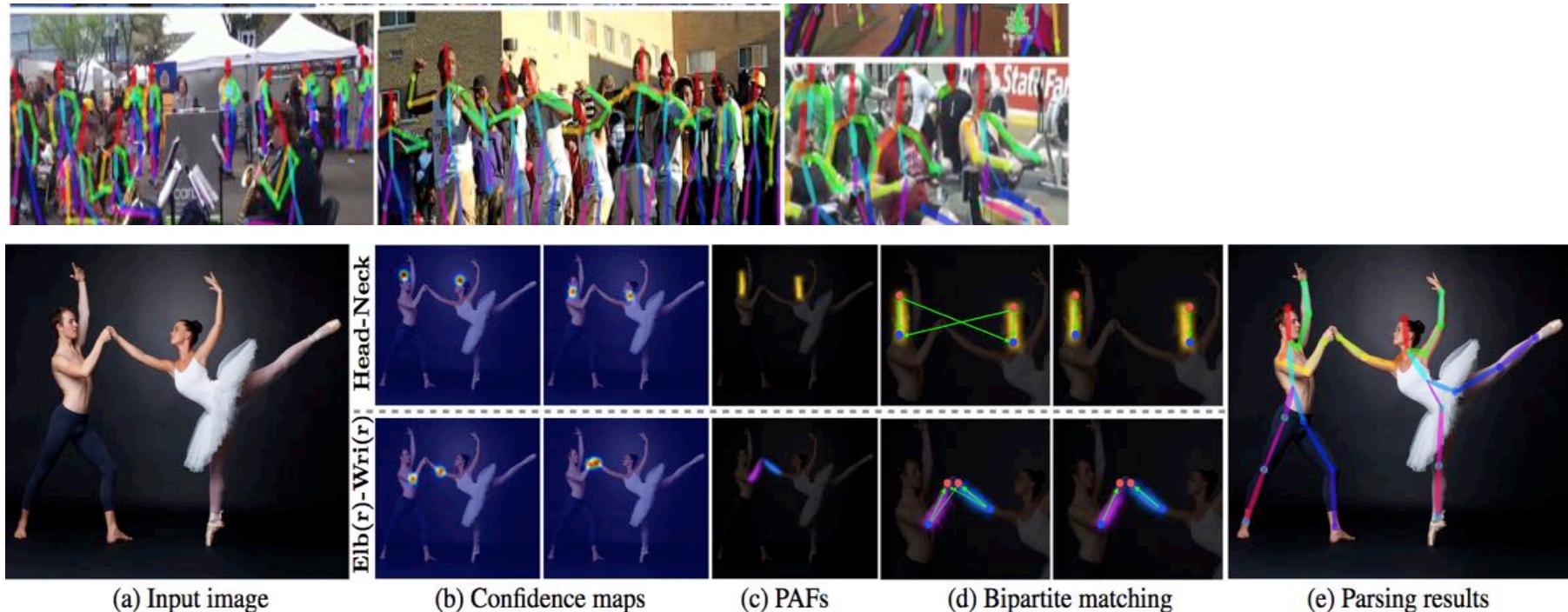


# 引用されそうな論文 (3/18)

## Part Affinity Fields (PAF)

- ベクトル場により画像と解剖学的な関節位置を対応付け
- 高精度かつ高速な姿勢推定を実現
- すでにOpenPoseというライブラリに実装済み

<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

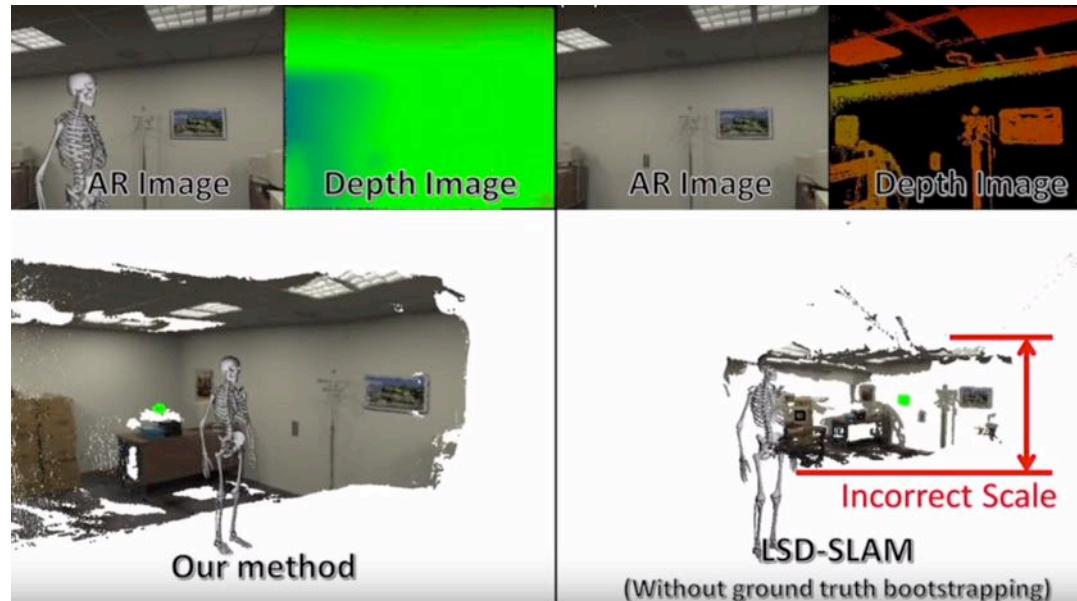


# 引用されそうな論文 (4/18)

## CNN-SLAM

- CNNの距離推定により、単眼RGBで高精度SLAMを実現
- スケール情報まで高精度に推定 & 屋外でも計測可能
- 動画参照：[https://www.youtube.com/watch?v=z\\_NJxbkQnBU](https://www.youtube.com/watch?v=z_NJxbkQnBU)

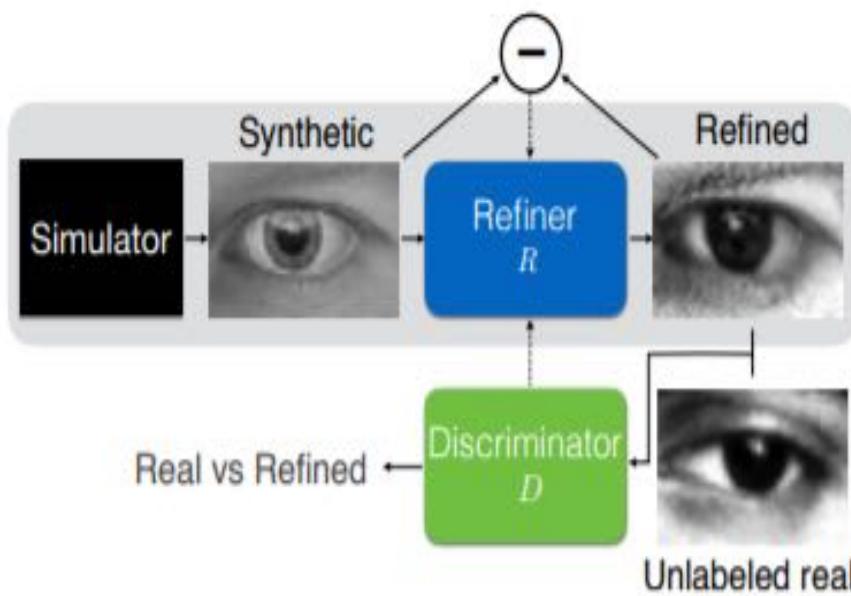
より汎用的に距離画像を推定する枠組みが重要



# 引用されそうな論文 (5/18)

## Learning from Simulated and Unsupervised Image through Adversarial Training (CVPR Best Paper)

- 合成画像を現実に近づけるRと実画像/合成画像を見分けるDの繰り返しにより合成画像をよりリアルにする
  - 学習データとして使用すると識別器の精度が向上
- 少量の実画像があれば大規模データの作成可能？



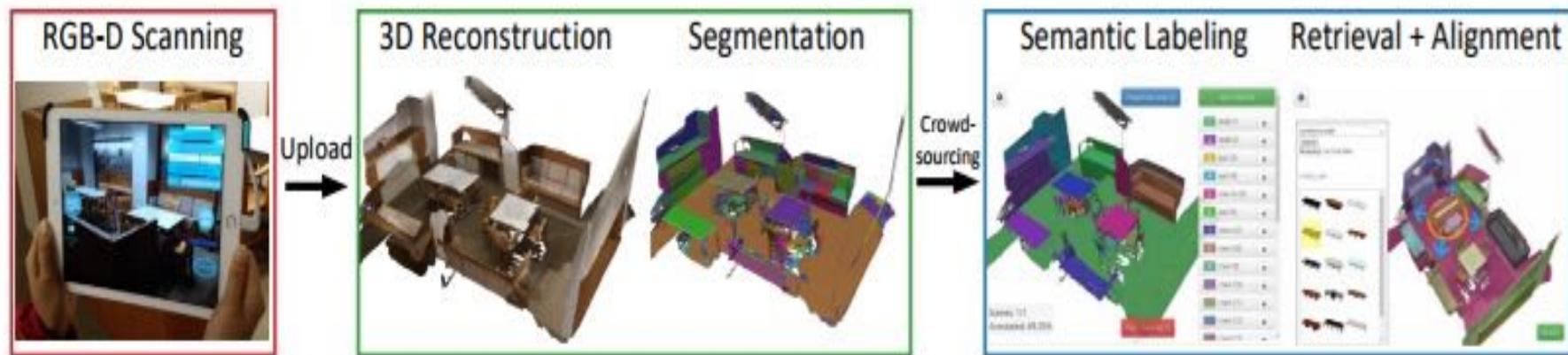
# 引用されそうな論文 (6/18)

## ScanNet (Spotlight)

- 大規模な (1,500シーン) RGB-Dデータセットを考案
- 3Dカメラ姿勢、表面リコンストラクション、セマンティッククラベル、CADモデル
- 動画参照：<https://www.youtube.com/watch?v=Olx4OnoZWQQ>



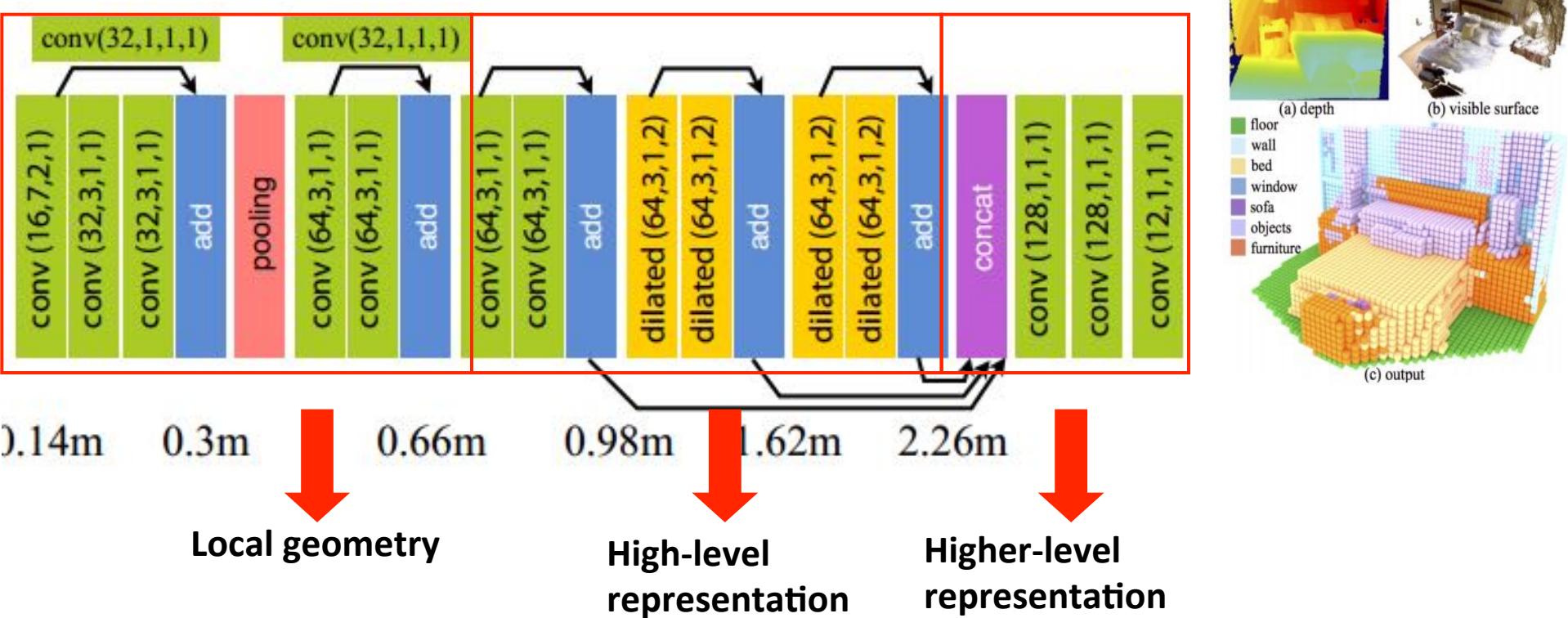
Figure 1. Example reconstructed spaces in ScanNet annotated with instance-level object category labels through our crowdsourced annotation framework.



# 引用されそうな論文 (7/18)

## Semantic Scene Completion (Oral)

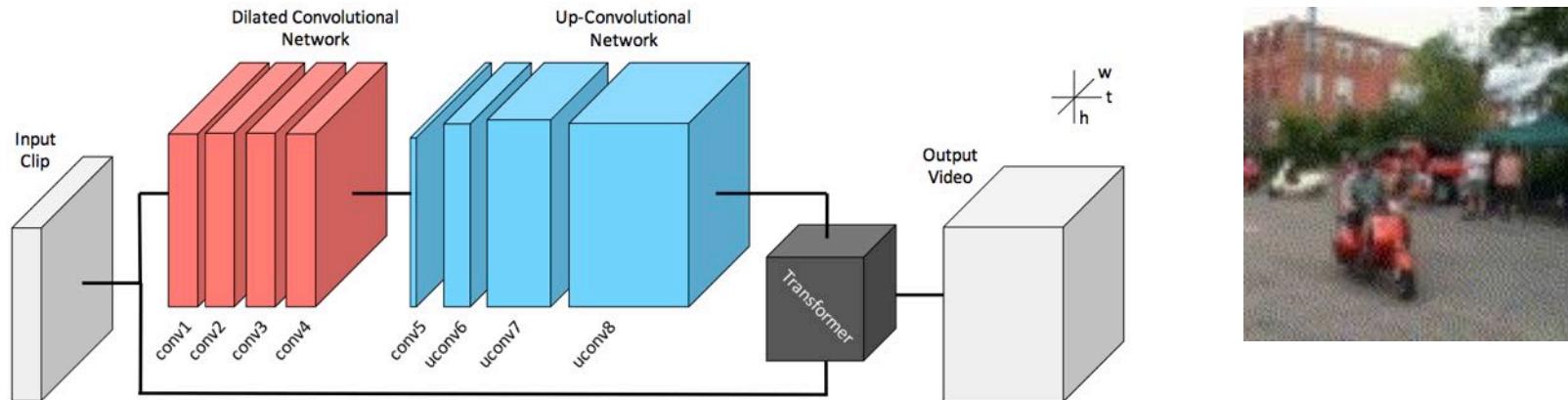
– 計測時に欠損を含む3次元シーンの補完をCNNで実装



# 引用されそうな論文 (8/18)

## Generating the Future with Adversarial Transformers

- Adversarial Learningを用いたビデオフレームの予測法
- 過去4frmから未来の12frmを推定し画素情報を保存しないモデル
- プロジェクト参照：<http://carlvondrick.com/transformer/>



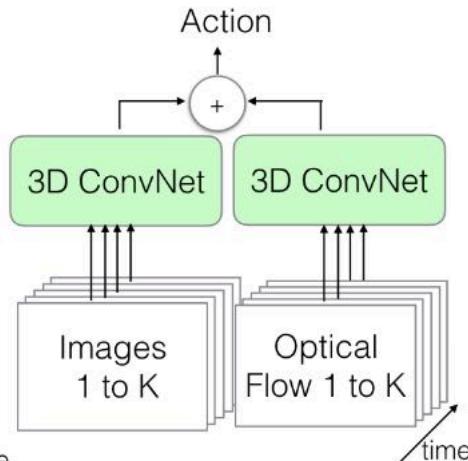
# 引用されそうな論文 (9/18)

## Quo Vadis, Action Recognition?

- 高品質な動画のKinetics Humanデータセットを提案
- データを用いて（これまで困難とされていた）3Dカーネルの最適化を高度に実行

データが最も重要！は動画認識にも言える

e) Two-Stream  
3D-ConvNet

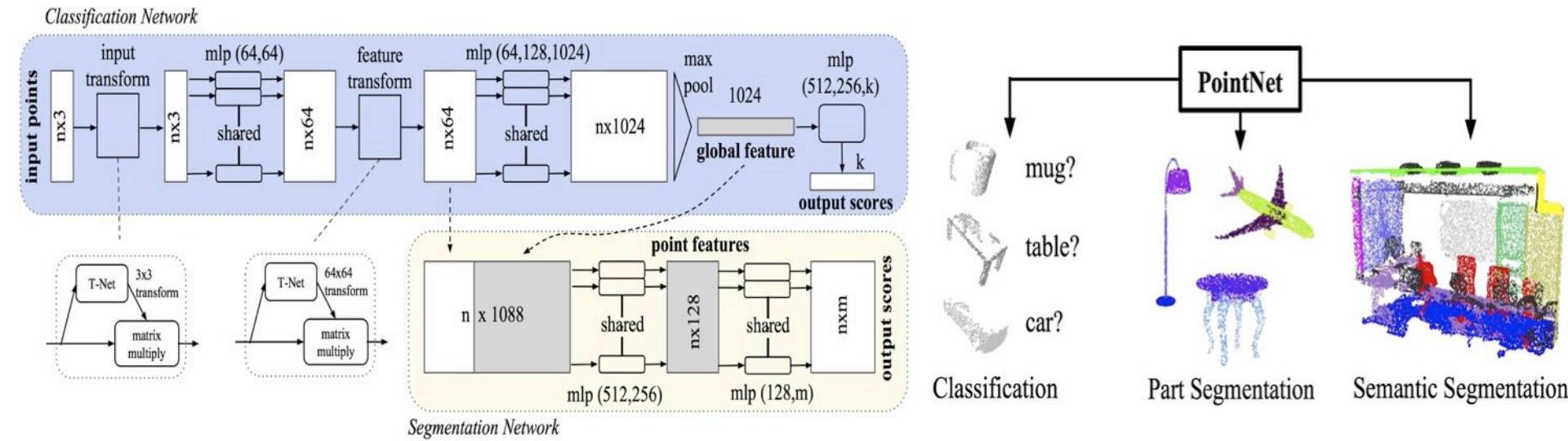


Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500
ActivityNet-200 [3]	2015	200	avg 141	28,108	19,994
Kinetics	2017	400	min 400	306,245	306,245

# 引用されそうな論文 (10/18)

## PointNet (Oral)

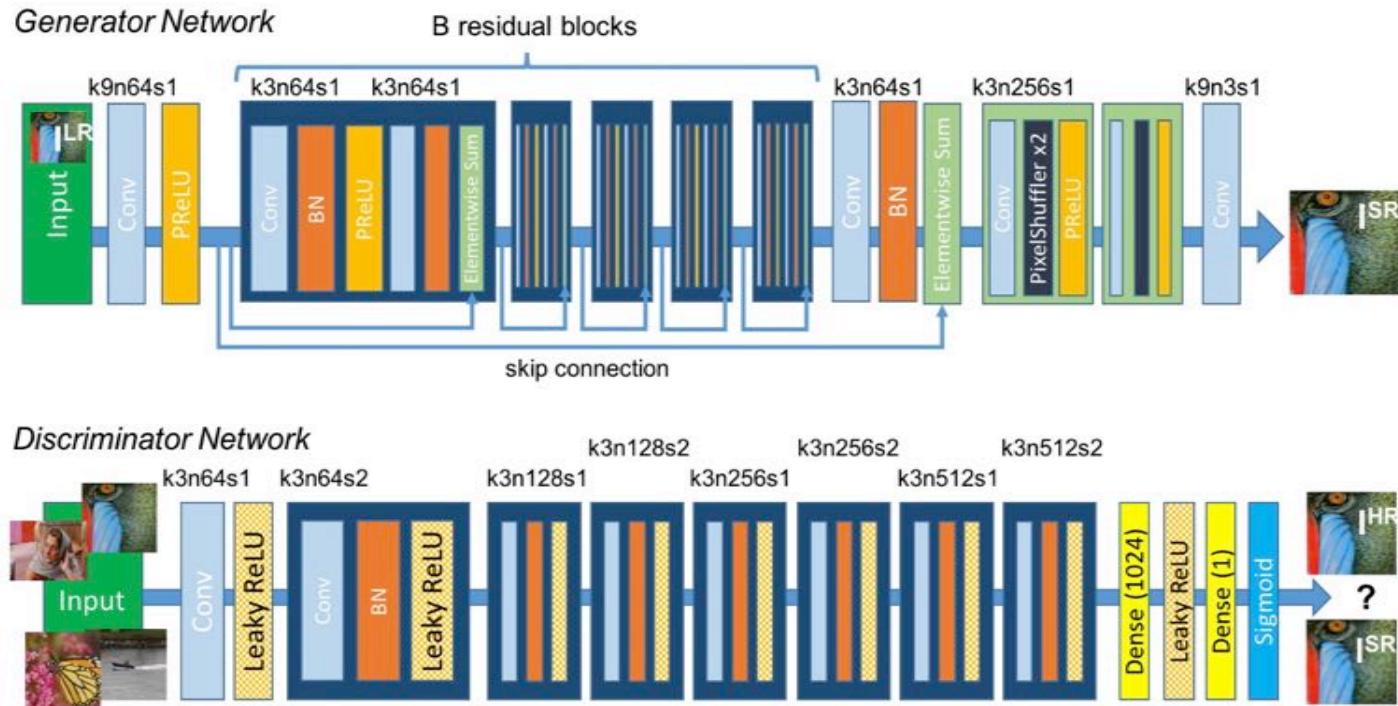
- 点群 (PointCloud) を直接置み込む
- 入力/特徴変換より Poolingにより効果的に点群を表現
- プロジェクト参照：<http://stanford.edu/~rqi/pointnet/>



# 引用されそうな論文 (11/18)

## Super-Resolution GAN (Oral)

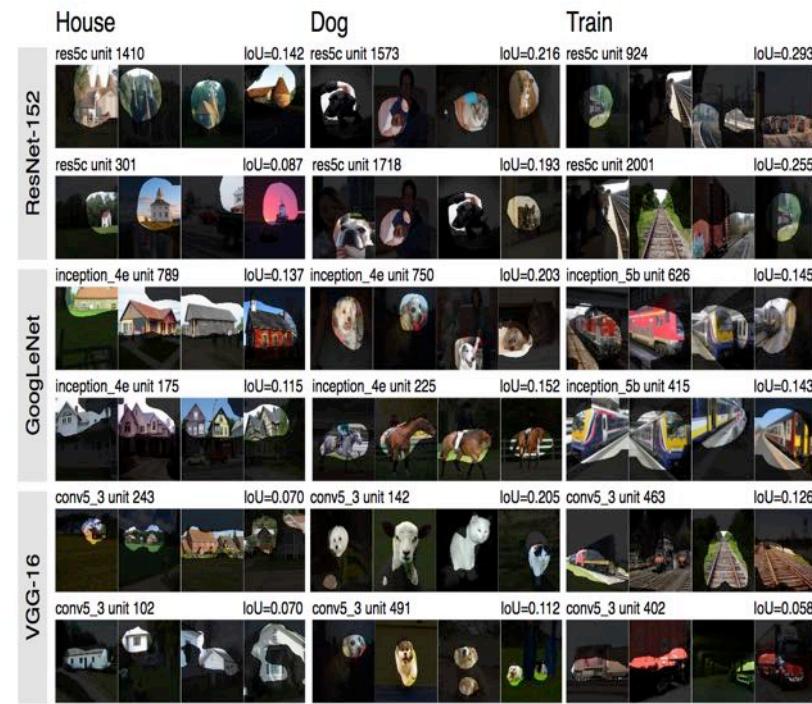
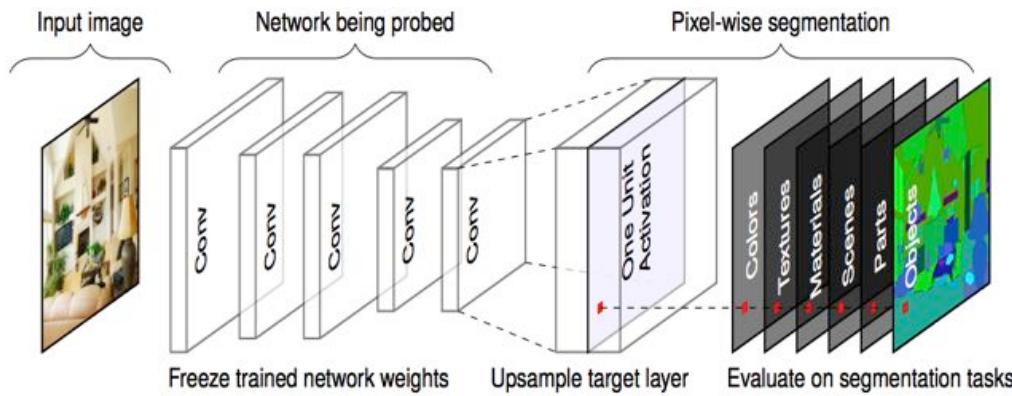
- 超解像のために生成モデルであるGANを用いた
- 高解像画像か復元画像か見分けがつかないようD/G学習
- Perceptual Lossによる超解像画像生成がキー



# 引用されそうな論文 (12/18)

## Network Dissection (Oral)

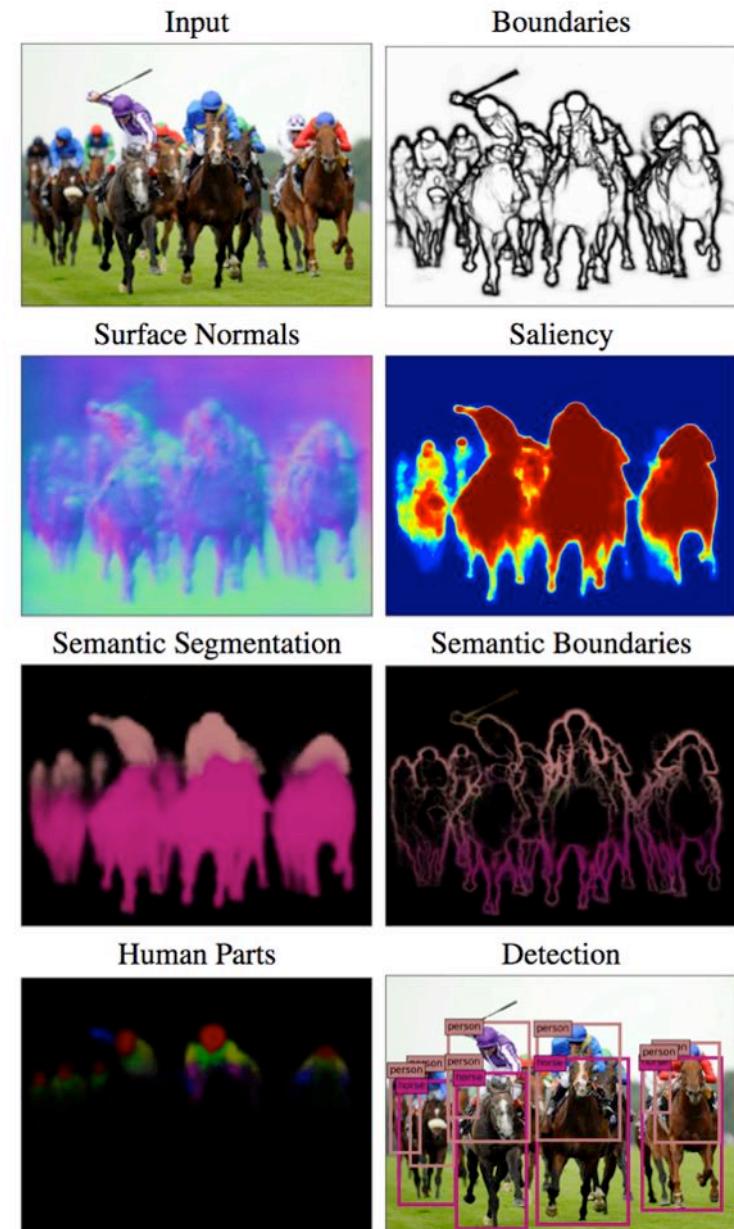
- CNNの隠れ層の特徴評価、意味的概念学習を検証
- 各DBからいかに概念 (e.g. objects, parts, scenes, textures, materials, colors) を学習したのかを知ることで深層学習の理解に挑戦



# 引用されそうな論文 (13/18)

## UberNet (Oral)

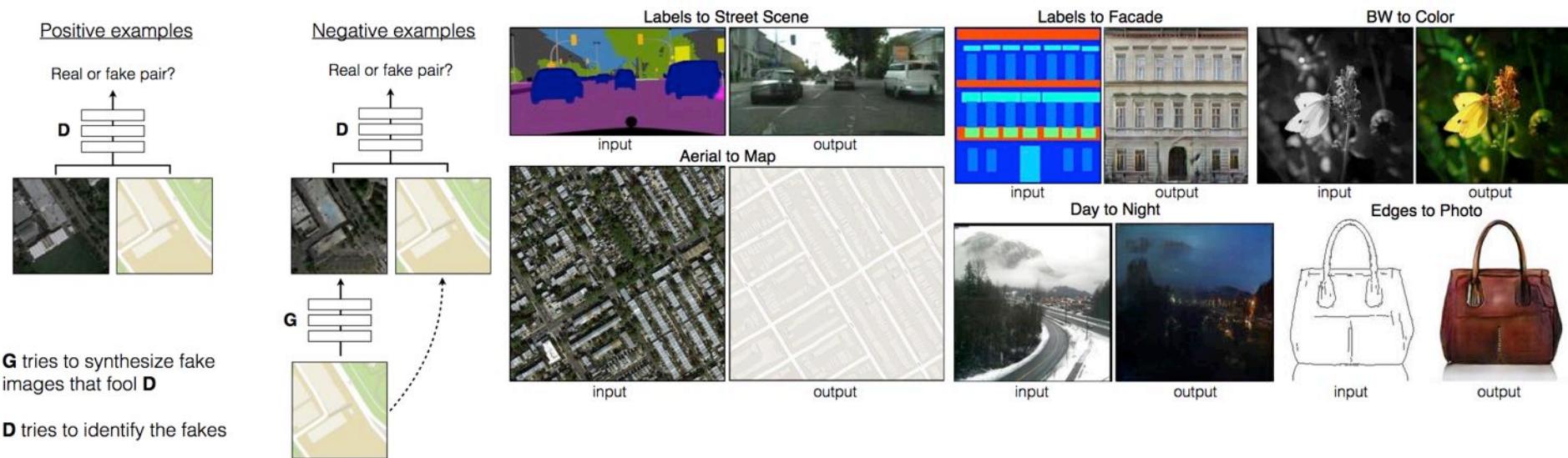
- 一つのネットワークで汎用的にタスクを解決 (右図)
- 多種類のマルチタスク学習の最適化についても言及



# 引用されそうな論文 (14/18)

## Image-to-Image (Pix2Pix)

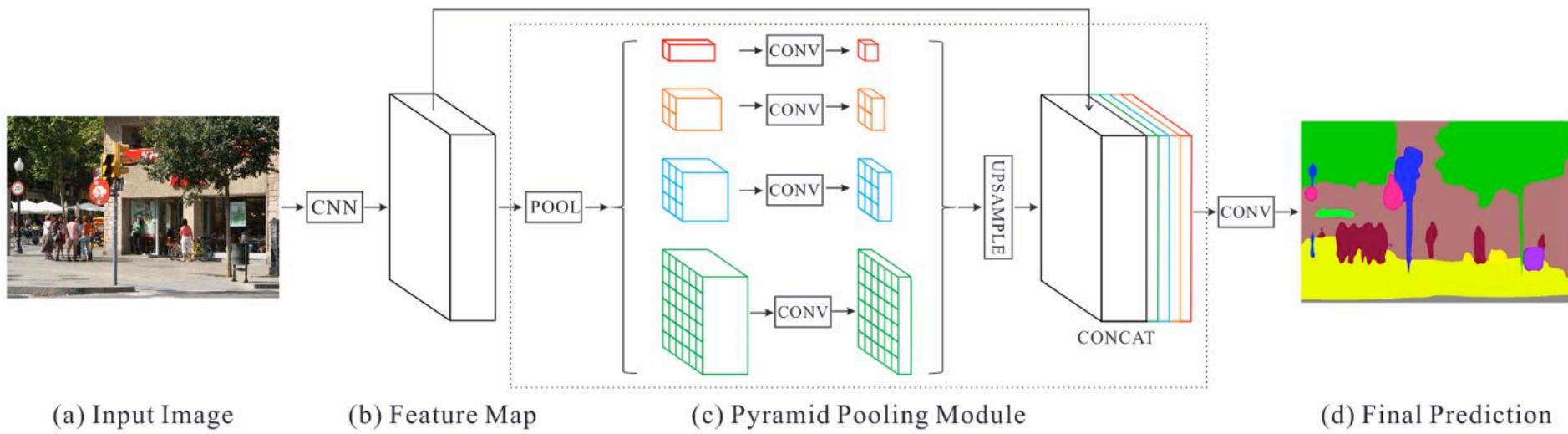
- ピクセル同士が対応する画像を相互変換
- 本物のペア？ / 変換画像とのペア？ を判定して学習
- デモあり：<https://phillipi.github.io/pix2pix/>



# 引用されそうな論文 (15/18)

## PSPNet

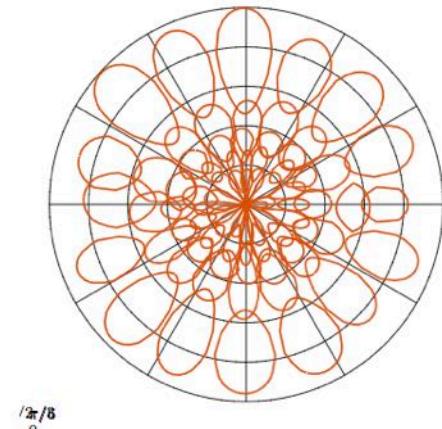
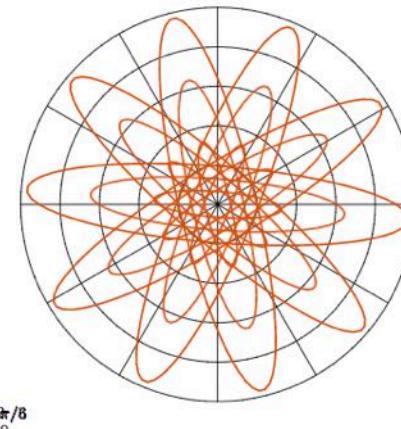
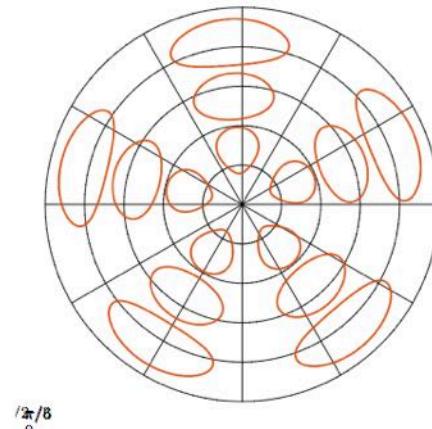
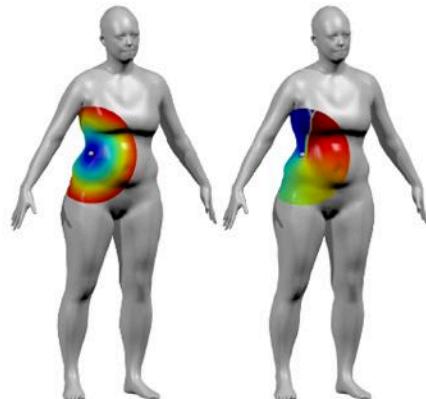
- CNNにて階層化されたマップを統合して意味情報を復元
- セマンティックセグメンテーションにおいて1位  
@ILSVRC2016
- 動画 : [https://www.youtube.com/watch?v=gdAVqJn\\_J2M](https://www.youtube.com/watch?v=gdAVqJn_J2M)



# 引用されそうな論文 (16/18)

## Geometric Deep Learning (Oral)

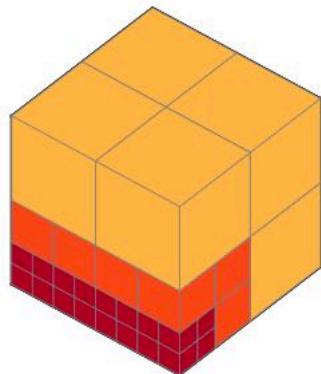
- 3次元的に置み込む際のつながりを置み込み可能
- 近傍との繋がりさえ定義できれば良いので3次元に限らず  
処理可能



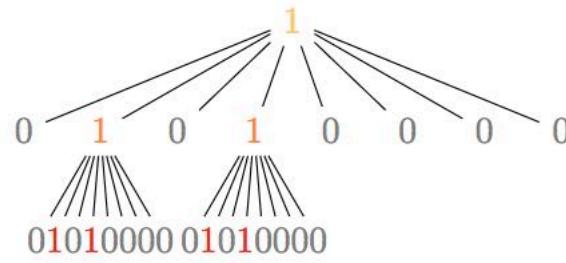
# 引用されそうな論文 (17/18)

## OctNet (Oral)

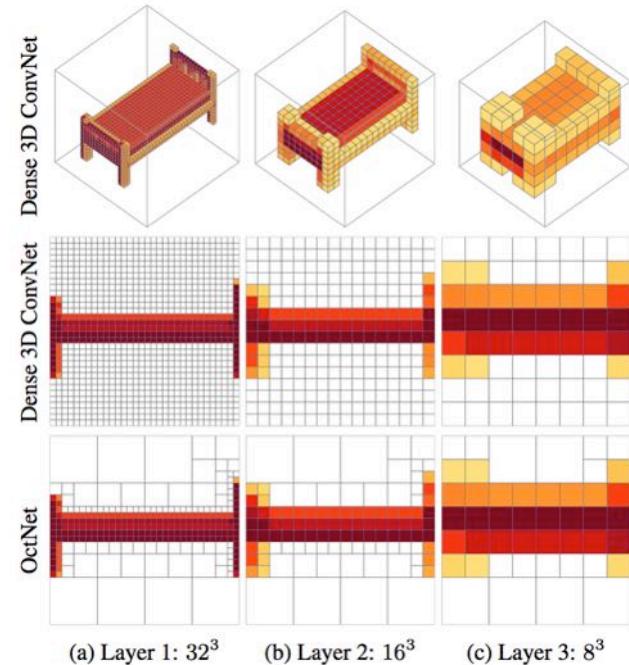
- 3次元カーネルにOctreeを導入：ボクセルがスパースな場合には手を抜いて置み込み
- Convolution/Upsampleも定義



(a) Shallow Octree



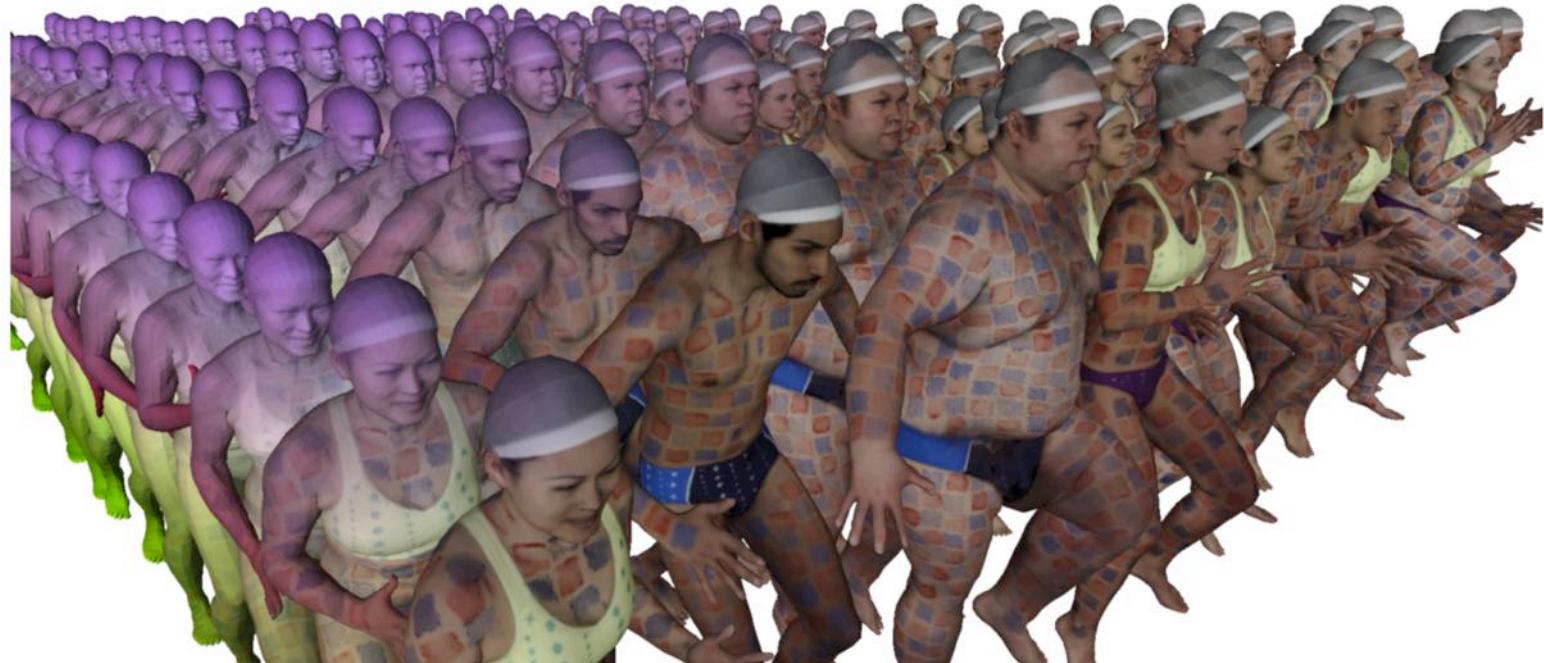
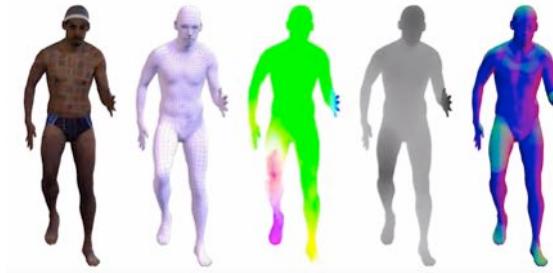
(b) Bit-Representation



# 引用されそうな論文 (18/18)

## Dynamic FAUST (Oral)

– 動きのある人体3Dモデルを提供



## 今後の方針

---

- では、どうすればよいか？

# 今後の方針（1/6）

---

## なんといっても問題設定

- こんな面白いことができる！と未来を見せる論文
- 今の技術を「使い、組み合わせ、洗練させ」ギリギリ実現可能な将来ビジョンを創り出す

## 今後の方針 (2/6)

---

### 他の追随を許さぬ強い手法を作る！

- 受賞論文や注目論文にあるように「極めて高速かつ高精度」を実現、コードをリリースして分野に貢献が理想
- 問題設定に対しての強い手法でも構わない
- 作れたらいいですね！ => e.g. DenseNets, YOLO9000, PAF, PSPNet...

# 今後の方針（3/6）

---

## 高品質論文でないと記録・記憶に残らない

- トップ会議の論文とて例外でない（CVPRは今年783本）
- 中途半端に分割した複数論文よりもパーフェクトな1本
- 動画やスライド公開・コード共有・DBリリースなども（できる限り）徹底して揃える

# 今後の方針 (4/6)

---

## 今まで以上にチームの力が重要

- 高品質論文 (前スライドより) には1人のパワーでは不十分？
- 当グループでは仕組みを再考
  - 通常の学生：1人1テーマ3年間継続（学部～修士を想定）
  - cvpaper.challenge：2～4人1テーマ1年でテーマ拡張/変更

# 今後の方針（5/6）

---

## CV系に限らず他分野へ投稿する

- (当然だが) CVに拘らない
- 分野が成熟してきたということは他の分野の手助けになる  
ということ

外の世界に出て行こう！

## □ストアアイディアを復活できないか

- Deep Learningにより(一時的に)消されたアイディアを復活
- もちろん、今風にアレンジ
- 可能性があるにも関わらず、消えたアイディアという意味

# 今後の方針？ (+a1)

---

## Beyond ImageNetのワークショップより

- 画像の総合的な理解へ
  - 画像識別/物体検出, 画像説明文にしても今は画像の一部分しか記述できていない
- Human-Levelの画像理解へ
  - より高次な理解へ (物体検出を組み合わせて複雑な理解にする)
- 今後ImageNetはKaggleにバトンタッチ
  - データ収集はより産業寄りに移行
  - CVの研究者はデータセットの構成や進むべき方向を定める
  - Kaggleがスポンサーになりデータサイエンティストやコンペティターにも分野を支えてもらう
  - という意思表示か

# 今後の方針？ (+a2)

## 企業展示より

- 自動運転ベンチャーAutoX
  - “AIの天才”と呼称されるJianxiong Xiao (Professor X) が起業
- 単眼カメラで自動運転のデモを実行
  - CV技術の発展によりLiDAR, ステレオなどはいらなくなる？
  - 天才が最もインパクトあるCVの応用と判断したのは自動運転？



<https://www.youtube.com/watch?v=bSUrn0lZinU>



# 今後の方針？ (+a3)

---

## YouTube-8Mワークショップより

- 大規模な動画像をどう扱うか
  - 現状Googleクラスの大企業クラスのマシンパワーが必要
  - その他はスパースなフレームレベルの特徴量しか扱えていない
- Spatio-temporal Localizationはまだ未解決
  - Classificationと比較して精度はかなり低い
  - データセットも少しずつ大きくなっている段階
- Video Classification, Detection, Captioningを越えて
  - Video Understanding（高次の理解）に発展させる
  - 高次になるに連れてアプリケーションとの関わりが強くなる
  - 何が要求されてそれをどう実現するか

## 以下、まとめ論文集

---

- 100本程度あります
- テンプレートが定まっていないものもありますがご了承ください

# (1) Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", in arXiv 1611.08050, 2016. (Oral)

Keywords: Pose Estimation, Part Affinity Fields

## 概要

・2D姿勢推定の決定版とも言える手法であり、ベクトル場により画像と解剖学的な関節位置を対応付け高精度かつ高速(10fps)な姿勢推定を実現した。提案モデルは関節位置やその関係性を記述するモデル

## 新規性・差分

・MS COCO 2016 keypoints challengeやMPII Multi-Person benchmarkにおいてstate-of-the-artなパフォーマンス（その上リアルタイム）

## Links

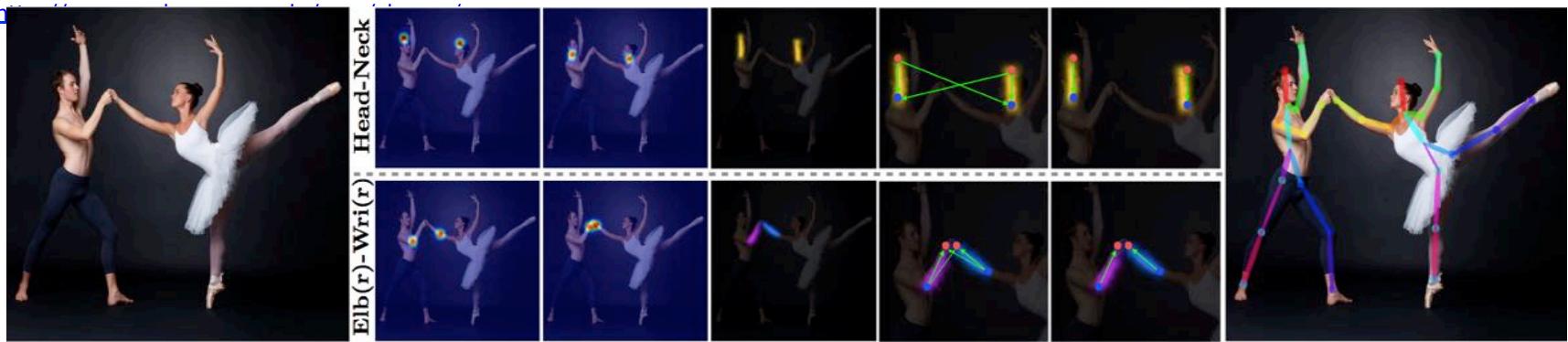
論文 <https://arxiv.org/pdf/1611.08050.pdf>

コード

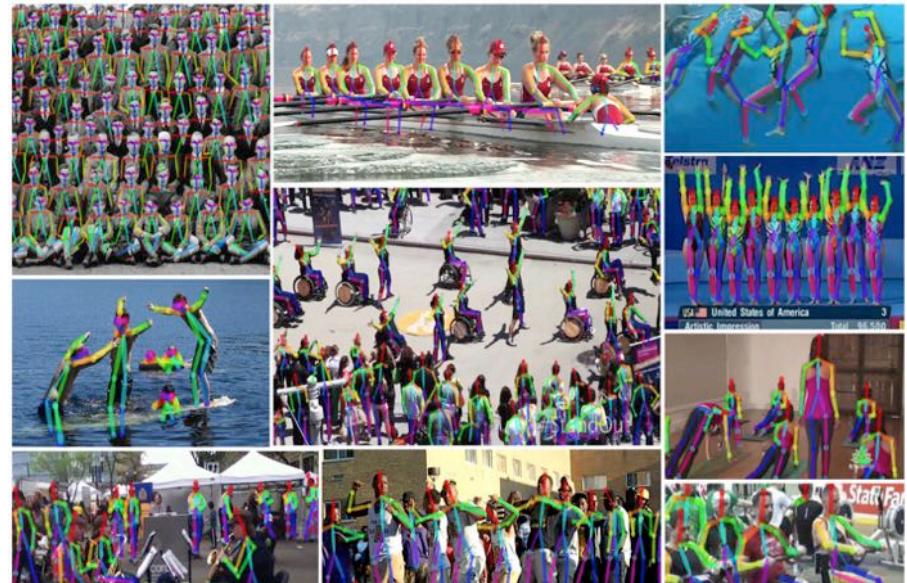
[https://github.com/ZheC/Realtime\\_Multi-Person\\_Pose\\_Estimation](https://github.com/ZheC/Realtime_Multi-Person_Pose_Estimation)

デモ <https://www.youtube.com/watch?v=pW6nZXeWlGM>

著者 <https://zhecao.com/>



・関節位置を推定 (confidence maps) し、PAFを計算。PAFに関しては相対的な位置関係やベクトル方向まで把握して最終的な姿勢位置や複数人のインタラクションを考慮した出力を行う。

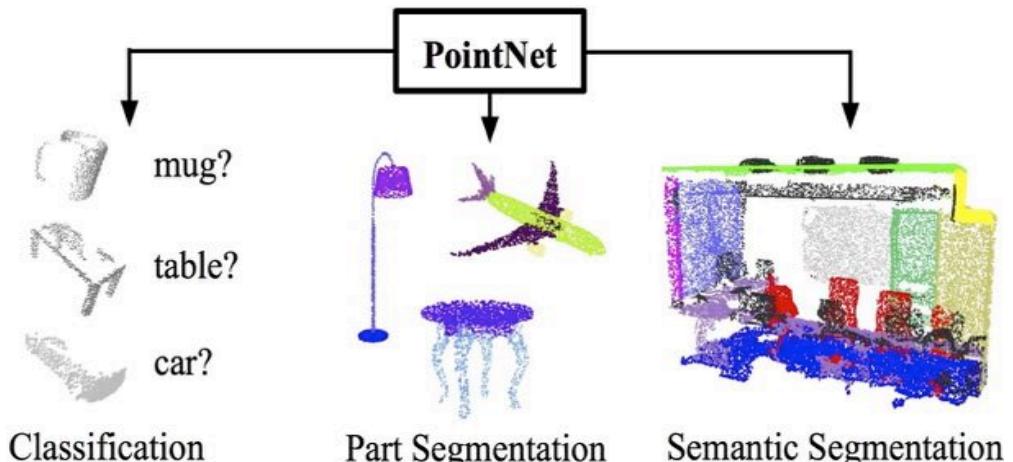


## [2] Charles R. Qi, Hao Su, Kaichun Mo, Lenidas J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation", in CVPR, 2017. (oral)

Keywords: PointCloud, PointNet

### 概要

- 点群 (PointCloud) を直接畳み込むことができる PointNetを提案。PointNetでは3次元認識、特に識別・パート分割・セマンティックセグメンテーションを行うことができる。(右下図) アーキテクチャのキーとなるのは MaxPoolにおけるSymmetric Functionであり、重要もしくは情報を多く含んでいる点群情報を選択して学習を行った。識別やセグメンテーションと用途に合わせてアーキテクチャの出力 (や途中処理) を変更した。Input/Feature Transformationを行い、MaxPoolingにより効果的に点群を表現する特徴を取得する。Multi-layer perception (mlp)の数字は層の数を表現している。全層のReLUには BatchNormが行われている。



### 新規性・差分

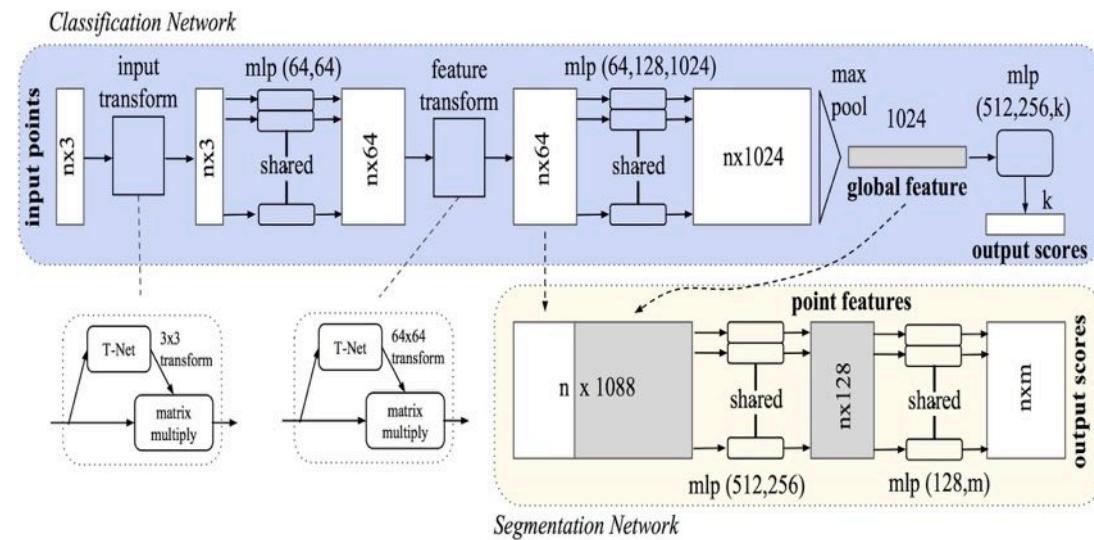
- 点群を直接的に処理可能なPointNetを提案し、識別・パート分割・セマンティックセグメンテーションに応用可能とした

### Links

論文 <https://arxiv.org/pdf/1612.00593.pdf>

プロジェクト <http://stanford.edu/~rqi/pointnet/>

コード <https://github.com/charlesq34/pointnet>



- (3) Iasonas Kokkinos, "UberNet: Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory", in arXiv 1609.02132, 2016 (CVPR2017 oral).**

Keywords: CNN, Object Detection, Semantic Segmentation

## 概要

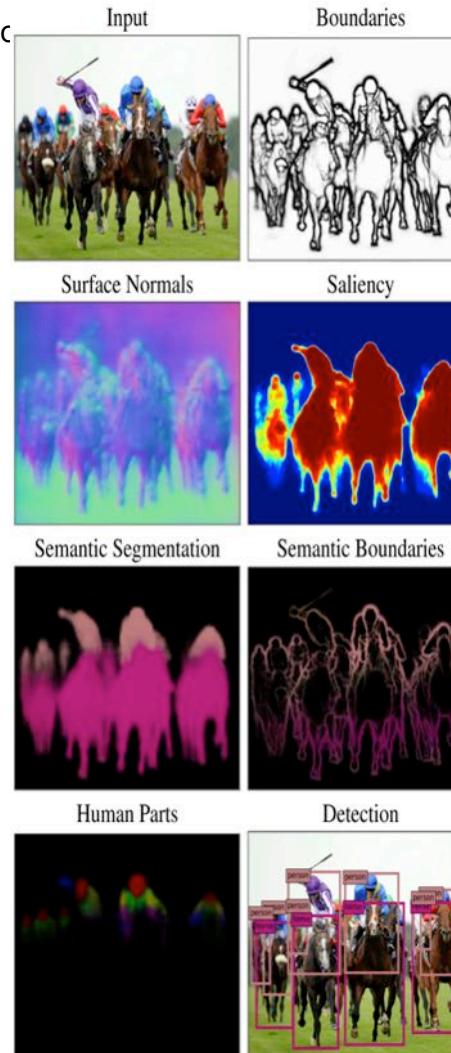
- 複数のタスクを単一のCNNで行うための手法を提案。すべてのタスクに必要なアノテーションを含んでいる単一のデータセットは存在しないため、各サンプルに付けられているアノテーションのみを利用するloss functionを設計
- 加えてタスクが増えると線形にメモリ使用量が増えていくため、タスク数に依存しないメモリ使用量のネットワークも提案。7タスクの設定で実験をした結果、タスクごとに学習した場合とほぼ同等の精度を達成した。

## 新規性・差分

- 複数のデータセットを使って学習するためのloss functionを提案
- タスク数に依存しないメモリ使用量のネットワークを提案

## Links

論文 <https://arxiv.org/pdf/1609.02132.pdf>



Method	mAP
F-RCNN, [32] VOC 2007++	73.2
F-RCNN, [32] MS-COCO + VOC 2007++	78.8
Ours, 1-Task	78.7
Ours, 2-Task	80.1
Ours, 7-Task	77.8

Table 5: Mean Average Precision performance (%) on the PASCAL VOC 2007 test set.

Method	mean IoU
Deeplab -COCO + CRF [78]	70.4
Deeplab Multi-Scale [49]	72.1
Deeplab Multi-Scale -CRF [49]	74.8
Ours, 1-Task	72.4
Ours, 2-Task	72.3
Ours, 7-Task	68.7

Table 6: Semantic segmentation - mean Intersection Over Union (IoU) accuracy on PASCAL VOC 2012 test.

Method	mAP	mMF
Semantic Contours [36]	20.7	28.0
Situational Boundary [100]	31.6	-
High-for-Low [7]	47.8	58.7
High-for-Low-CRF [7]	54.6	62.5
Ours, 1-Task	54.3	59.7
Ours, 7-Task	44.3	48.2

Table 8: Semantic Boundary Detection Results: we report mean Average Precision (AP) performance (%) and Mean Max F-Measure Score on the validation set of PASCAL VOC 2010, provided by [36].

# [4] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba, "Network Dissection: Quantifying Interpretability of Deep Visual Representations", in CVPR, 2017. (oral)

Keywords: Network Dissection, Latent Representation

## 概要

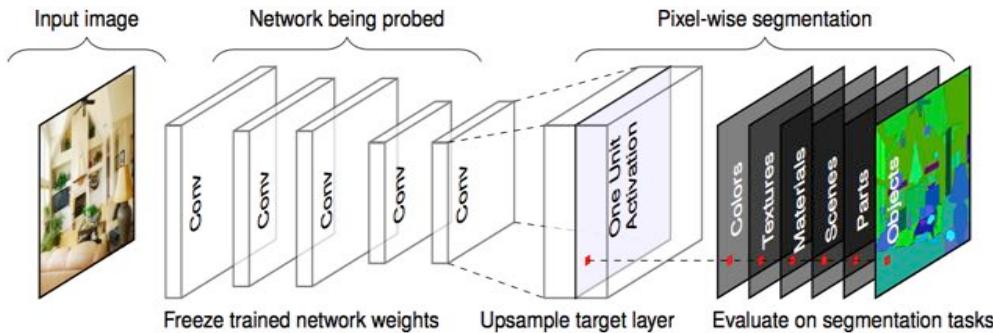
Network Dissectionと呼ばれる、CNNの隠れ層の特徴評価や意味的概念を学習するためのネットワークを提案する。それぞれ異なるデータセットからどのような概念 (e.g. objects, parts, scenes, textures, materials, colors) を学習したのかを知ることで深層学習を理解することに挑戦した。教師あり/なし学習、学習回数、初期値の違い、層の深さや幅、ドロップアウトやバッチ正規化などについて詳細に検討した。この疑問を明確にするためにBroadly and Densely Labeled Dataset (Broaden)を提案した。BroadenはADE, Open-Surfaces, Pascal-Context, Pascal-Partなどから構成される。テストを行ったネットワークは右の通りであり、学習なし/教師あり/教師なしによりテストを行った。

## 新規性・差分

- CNNの解釈可能性について評価した。
- 教師あり/なし学習、学習回数、初期値の違い、層の深さや幅、ドロップアウトやバッチ正規化などについて詳細な実験によりCNNが学習でどのような見えを学習するかを明らかにした。

## Links

論文 <http://netdissect.csail.mit.edu/final-network-dissection.pdf>  
プロジェクト <http://netdissect.csail.mit.edu/>  
コード <https://github.com/CSAILVision/NetDissect>



wheels in object net



people in video net

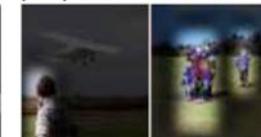
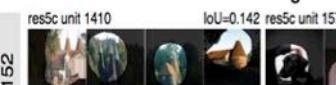


Table 2. Tested CNNs Models

Training	Network	Data set or task
none	AlexNet	random
Supervised	AlexNet	ImageNet, Places205, Places365, Hybrid.
	GoogLeNet	ImageNet, Places205, Places365.
	VGG-16	ImageNet, Places205, Places365, Hybrid.
Self	ResNet-152	ImageNet, Places365.
	AlexNet	context, puzzle, egomotion, tracking, moving, videoorder, audio, crosschannel, colorization, objectcentric.

House



Dog



Train



ResNet-152



GoogLeNet



VGG-16



(5)

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Paschal Frossard, "Universal adversarial perturbations", in CVPR, 2017. (oral)

Keywords: Universal Perturbations

## 概要

・ニューラルネットは背景や物体の見え方による多少のノイズからPerturbation（混乱）が発生するという問題を緩和した。ここから、いかにしてネットワークを（効率的に）だますか、転じて騙されないようなネットワークにするのかという検討を行った。実験ではPerturbation Vector ( $v$ )を定義して入力 $x_i$ と合わせて $(x_i + v)$ ネットワークが騙される/騙されないという境界を探査した。右下は計算されたVGG/GoogLeNet/ResNetのUniversal Perturbationである。

## 新規性・差分

・ニューラルネットのノイズとカテゴリ誤りの原因のひとつであるPerturbationについての解析を行った  
・右下の表は各ネットワークに対して交差してUniversal Perturbationを適用した結果である。ここから、ネットワークに特化して「騙しやすい」パターンを学習できたといえる。

## Links

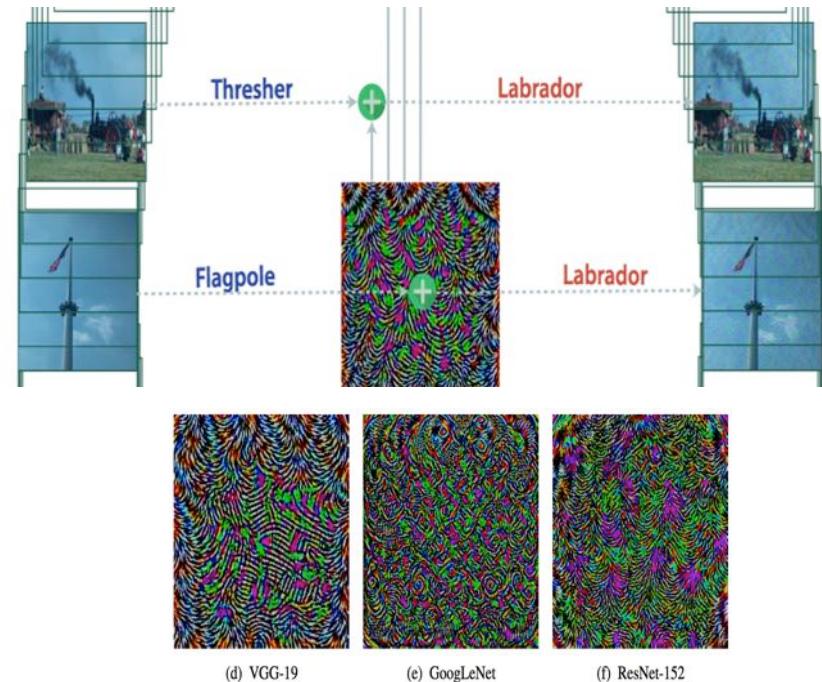
論文 <https://arxiv.org/pdf/1610.08401.pdf>

論文v2 <https://arxiv.org/pdf/1705.09554.pdf>

プロジェクト <https://github.com/ITS4/universal>

動画 <https://www.youtube.com/watch?v=jhOu5yhe0rc>

	VGG-F	CaffeNet	GoogLeNet	VGG-16	VGG-19	ResNet-152
VGG-F	<b>93.7%</b>	71.8%	48.4%	42.1%	42.1%	47.4 %
CaffeNet	74.0%	<b>93.3%</b>	47.7%	39.9%	39.9%	48.0%
GoogLeNet	46.2%	43.8%	<b>78.9%</b>	39.2%	39.8%	45.5%
VGG-16	63.4%	55.8%	56.5%	<b>78.3%</b>	73.1%	63.4%
VGG-19	64.0%	57.2%	53.6%	73.5%	<b>77.8%</b>	58.0%
ResNet-152	46.3%	46.3%	50.5%	47.0%	45.5%	<b>84.0%</b>



Algorithm 1 Computation of universal perturbations.

```
1: input: Data points  $X$ , classifier  $\hat{k}$ , desired  $\ell_p$  norm of
   the perturbation  $\xi$ , desired accuracy on perturbed samples  $\delta$ .
2: output: Universal perturbation vector  $v$ .
3: Initialize  $v \leftarrow 0$ .
4: while  $\text{Err}(X_v) \leq 1 - \delta$  do
5:   for each datapoint  $x_i \in X$  do
6:     if  $\hat{k}(x_i + v) = \hat{k}(x_i)$  then
7:       Compute the minimal perturbation that
          sends  $x_i + v$  to the decision boundary:
          
$$\Delta v_i \leftarrow \arg \min_r \|r\|_2 \text{ s.t. } \hat{k}(x_i + v + r) \neq \hat{k}(x_i).$$

8:       Update the perturbation:
          
$$v \leftarrow \mathcal{P}_{p,\xi}(v + \Delta v_i).$$

9:     end if
10:   end for
11: end while
```

# [6] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, Dilip Krishnan, "Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks", in CVPR , 2017. (oral)

Keywords: GAN, Domain Adaptation, Automatic Data Creation

## 概要

- ラベル付けされた完全教師あり学習のためのデータセットを作るための人的なコストが非常に高いことからこの作業を完全自動で代替しようとする試み。すでにラベルが豊富に手に入るドメインから転換してデータセットを自動でGAN（敵対的ネットワーク）により生成し、ラベルも自動で付与する。ドメインを変換するような生成モデル（GAN）を学習することが提案内容であり、結果的に画像と同時にラベルを対応づける。タスクとしてはMNIST（文字認識）、LineMod（3次元姿勢推定）に対して行った。

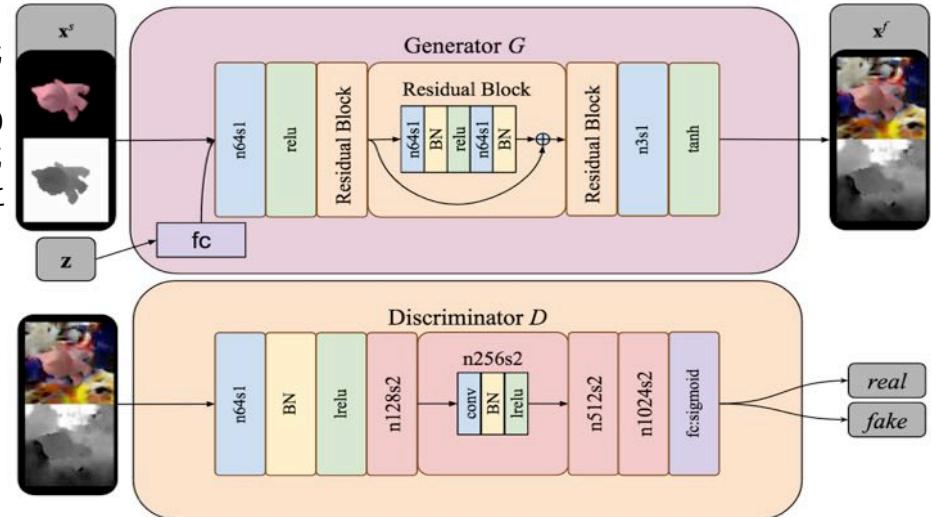
## 新規性・差分

- アノテーションデータを用いることなくドメイン変換を行うことでよりリアルなデータ（CADのようなSynthetic dataから背景を含むようなデータに変換）を生成し、生成画像のみでState-of-the-artを達成。
- 完全自動でアノテーションを含むようなデータまでを生成可能としただけでなく、機械学習を行った際にも精度が向上することを実証した。

## Links

論文 <https://arxiv.org/pdf/1612.05424.pdf>

右表はMNISTやLineModデータを用いた結果比較である。  
ドメイン変換なしの場合や  
ターゲットドメインに対する  
変換がありの場合、提案手法  
も含めて評価を行った。



Model	MNIST to USPS	MNIST to MNIST-M
Source Only	78.9	63.6 (56.6)
CORAL [41]	81.7	57.7
MMD [45, 31]	81.1	76.9
DANN [14]	85.1	77.4
DSN [5]	91.3	83.2
CoGAN [30]	91.2	62.0
Our model	<b>95.9</b>	<b>98.2</b>
Target-only	96.5	96.4 (95.9)

Model	Classification Accuracy	Mean Angle Error
Source-only	47.33%	89.2°
MMD [45, 31]	72.35%	70.62°
DANN [14]	99.90%	56.58°
DSN [5]	<b>100.00%</b>	53.27°
Our model	99.98%	<b>23.5°</b>
Target-only	100.00%	6.47°

[7]

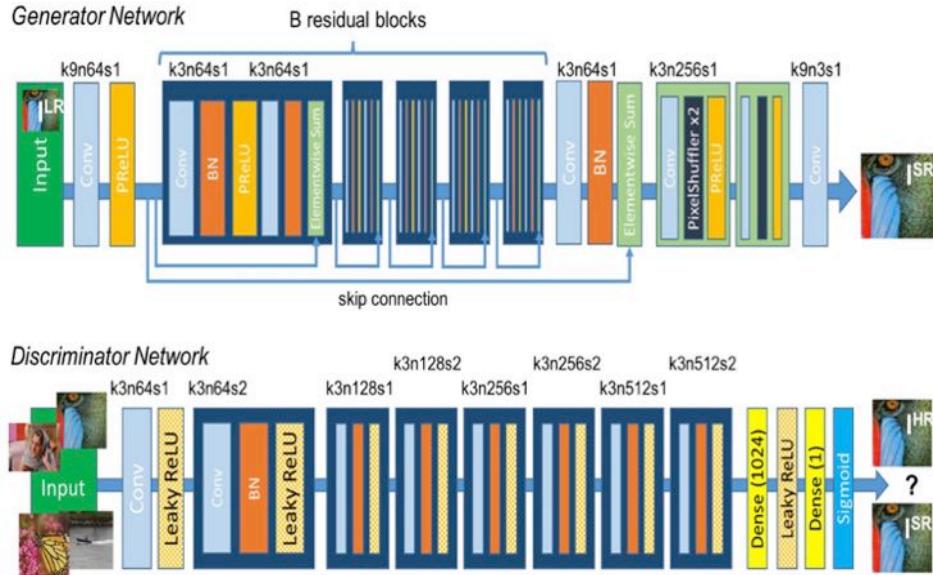
# Christian Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network", in CVPR, 2017. (oral)

Keywords: Super-Resolution GAN (SRGAN)

## 概要

- 超解像のために生成モデルであるGANを用い、Super-Resolution GANを提案。基本的なアイディアとしては、「偽物か本物か見分けがつかないように生成器と識別器を学習する」GANに対してSRGANでは「本物か生成した超解像か見分けがつかないように学習」する。ここではContent LossやAdversarial Lossを組み合わせた。

Perceptual Lossを提案して超解像画像の生成に寄与した。Deep Residual LearningやPixel Shufflerを用いたことも追加した項目である。



## 新規性・差分

- Perceptual Loss (Content/Adversarial Lossを統合)を超解像に適用してより鮮明な画像の生成に成功した
- 右表は各々PSNR, SSIM, MOSによる数値を、各手法に對して比較した結果である。PSNRやSSIMはSRResNetが良い結果であったが、MOSでは提案手法が勝った。

	nearest	bicubic	SRCCNN	SelfExSR	DRCN	ESPCN	SRResNet	SRGAN	HR
PSNR	26.26	28.43	30.07	30.33	31.52	30.76	<b>32.05</b>	29.40	$\infty$
SSIM	0.7552	0.8211	0.8627	0.872	0.8938	0.8784	<b>0.9019</b>	0.8472	1
MOS	1.28	1.97	2.57	2.65	3.26	2.89	3.37	<b>3.58</b>	4.32
<b>Set14</b>									
PSNR	24.64	25.99	27.18	27.45	28.02	27.66	<b>28.49</b>	26.02	$\infty$
SSIM	0.7100	0.7486	0.7861	0.7972	0.8074	0.8004	<b>0.8184</b>	0.7397	1
MOS	1.20	1.80	2.26	2.34	2.84	2.52	2.98	<b>3.72</b>	4.32
<b>BSD100</b>									
PSNR	25.02	25.94	26.68	26.83	27.21	27.02	<b>27.58</b>	25.16	$\infty$
SSIM	0.6606	0.6935	0.7291	0.7387	0.7493	0.7442	<b>0.7620</b>	0.6688	1
MOS	1.11	1.47	1.87	1.89	2.12	2.01	2.29	<b>3.56</b>	4.46

## Links

論文 [https://arxiv.org/pdf/1609\\_04802.pdf](https://arxiv.org/pdf/1609_04802.pdf)

GitHub (TensorFlow) <https://github.com/buriburisuri/SRGAN>

GitHub (Keras)

<https://github.com/titu1994/Super-Resolution-using-Generative-Adversarial-Networks>

GitHub (Chainer)

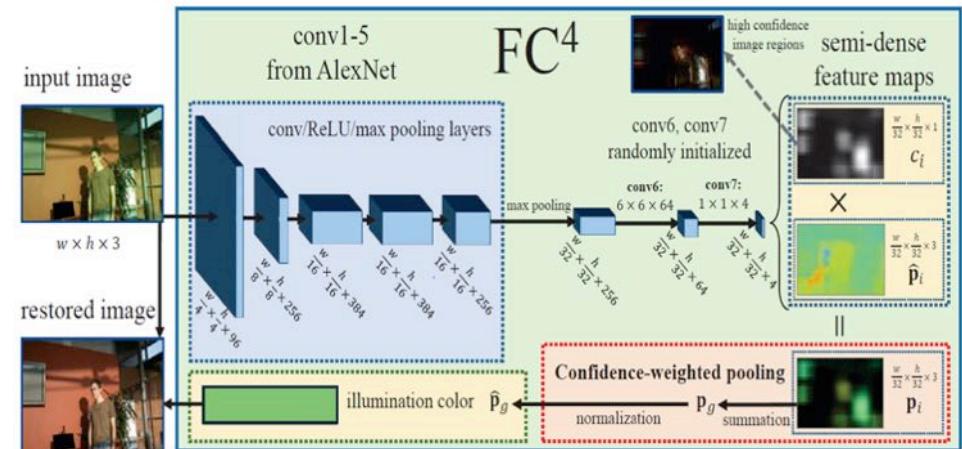
[https://github.com/Hi-king/superresolution\\_gan](https://github.com/Hi-king/superresolution_gan)

# [8] Yuanming Hu, Baoyuan Wang, Stephen Lin, "FC4: Fully Convolutional Color Constancy with Confidence-weighted Pooling", in CVPR, 2017.

Keywords: Color Constancy, Fully Convolutional Network, Weighted Pooling

## 概要

・画像中の照明色を推定して照明による変化をなくすための  
Color Constancyという課題に対して新しい手法を提案。従来は  
パッチベースで画像中の各パッチに対して推定を行いそれらを結合  
する手法が取られていたが、パッチによっては曖昧なものが存  
在するのが問題（黄色い壁があったときに黄色い照明で黄色く見  
えるのか元々黄色い壁なのか曖昧）。このような曖昧なパッチは  
多数存在し学習時でも推定時でも悪影響を与える。そこで、この  
論文では画像全体のパッチをFully Convolutional Network  
(FCN)ベースで同時に処理する手法を提案。加えて照明推定に  
有効な領域を推定してPoolingするWeighted Poolingを導入する  
ことで頑健な推定を実現。



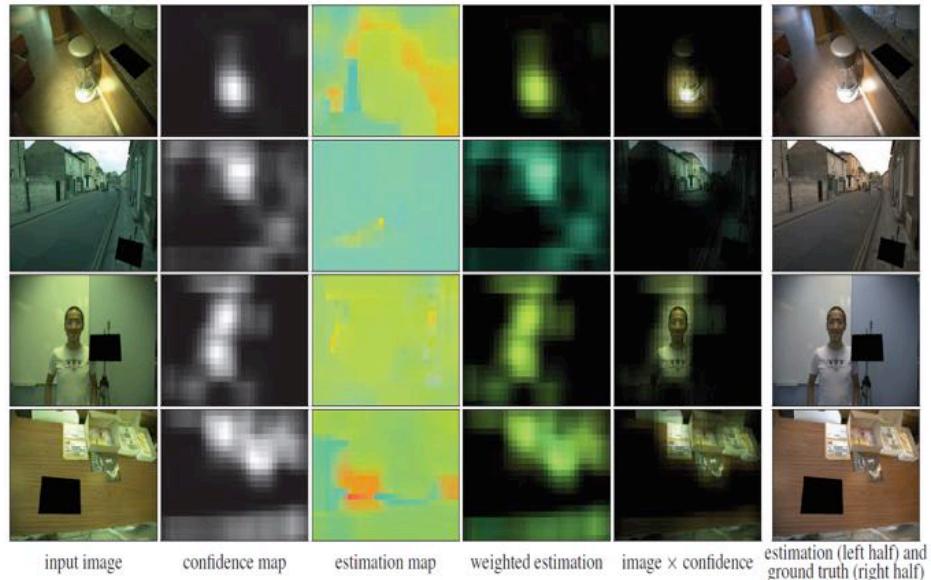
## 新規性・差分

- ・Color Constancyに対してFCNをベースとして画像全体の  
パッチを同時に処理する手法を提案
- ・Weighted Poolingという有用な部分を学習するLayerを提案

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Hu\\_FC4\\_Fully\\_Convolutional\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Hu_FC4_Fully_Convolutional_CVPR_2017_paper.pdf)



[9]

Mihoko Shimano, Hiroki Okawa, Yuta Asano, Ryoma Bise, Ko Nishino, Imari Sato, "Wetness and Color from A Single Multispectral Image", in CVPR, 2017.

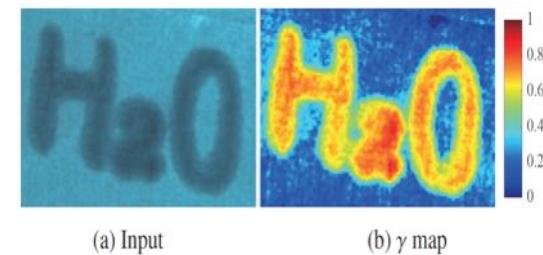
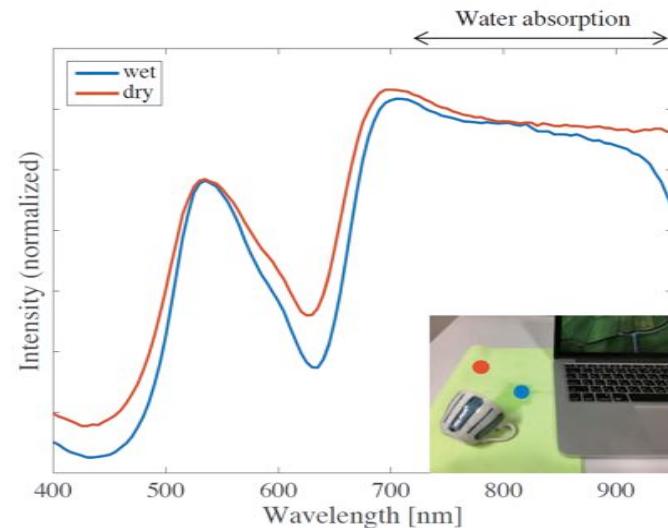
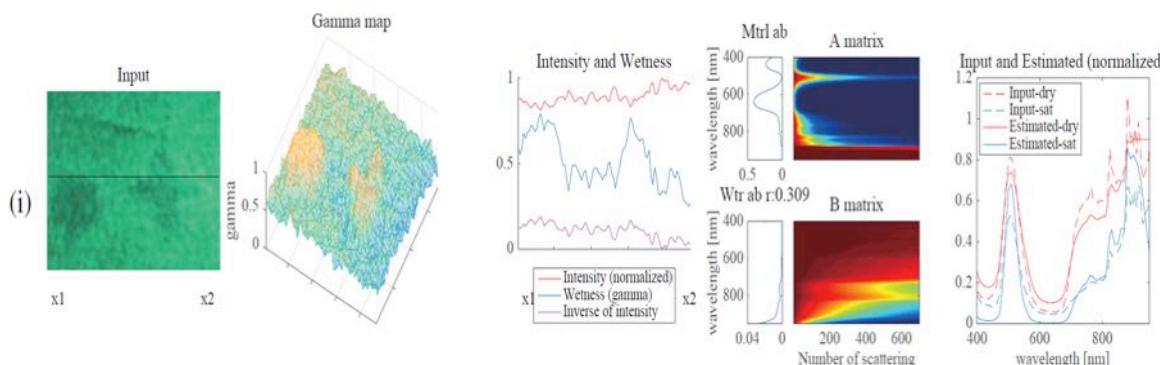
Keywords: Wetness, Multispectral Image

## 概要

- 対象の表面がどのくらい濡れているか (wetness) を推定するための手法を提案。マルチスペクトル画像を入力として wetness を推定。濡れによる色の変化やスペクトル分布の変化をモデル化することで、この推定を実現。提案手法は実世界の様々な素材に対して wetness を高精度に推定可能なことを実験的に示した。

## 新規性・差分

- Wetnessに関するSpectral Modelを導出



	cotton	gauze	felt	sand	leather	all
Ave	0.980	0.977	0.975	0.954	0.980	0.974
Std	0.006	0.024	0.010	0.032	0.002	0.016

Table 1: Correlation coefficients between weight and wetness for each material.

## Links

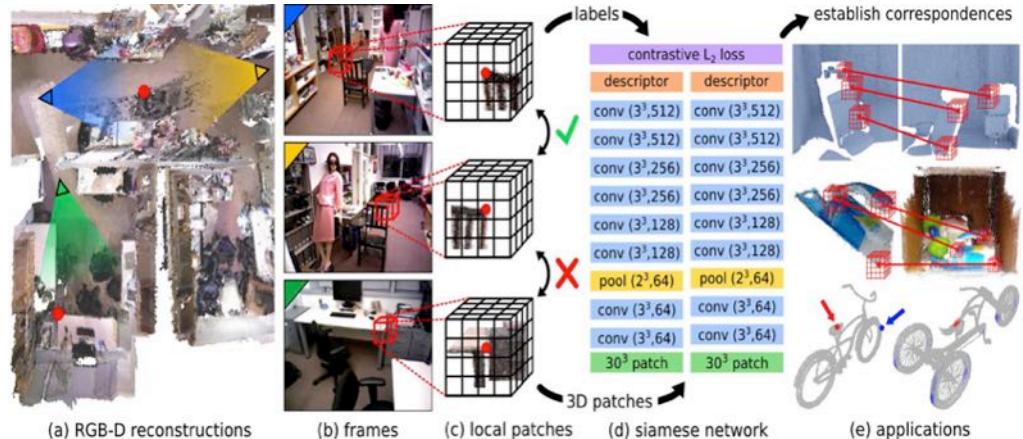
論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Shimano\\_Wetness\\_and\\_Color\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Shimano_Wetness_and_Color_CVPR_2017_paper.pdf)

# 【10】Andy Zeng, Shuran Song, Matthias Niessner, Jianxiong Xiao, Thomas Funkhouser, "3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions", in CVPR, 2017.

Keywords: 3D (Environment) Matching,

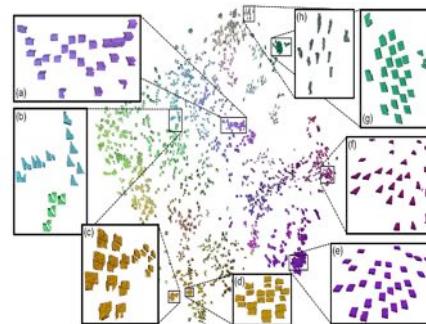
## 概要

- 3次元の局所特徴量でもうまくマッチングできない局所的な環境を、3D Convolutionにより畳み込み、環境自体の局所的なマッチングを行う手法を提案。大規模なRGB-Dデータから自己学習(self-supervised learning)により局所マッチングを成功させるための特徴を学習する。右に示すように $30^3$ の3Dパッチのペアを入力として、Contrastive L2 Lossにより最適化を行う。



## 新規性・差分

- XYZ3Dパッチペアの自己学習により、似たような形状を保有する部分の対応を取得することに成功した
- 右図に示すように似たような形状は同じような特徴を持つ(tSNE空間にて距離が近い)
- 同時にKeypoint Matching Benchmarkも提供してコミュニティに対して研究の問題設定を提起
- Amazon Picking Challengeのデータに対しても実験を行った



## Links

論文 <https://arxiv.org/pdf/1603.08182.pdf>

プロジェクト <http://3dmatch.cs.princeton.edu/>

ビデオ <https://www.youtube.com/watch?v=gZrsJ1tDvVA>

GitHub <https://github.com/andyzeng/3dmatch-toolbox>

Method	Error
Johnson <i>et al.</i> (Spin-Images) [19]	83.7
Rusu <i>et al.</i> (FPFH) [28]	61.3
2D ConvNet on Depth	38.5
Ours (3DMatch)	<b>35.3</b>

Fig 1. Keypoint matching task error (%) at 95% recall.

# 【11】 Mandar Dixit, Roland Kwitt, Marc Niethammer, Nuno Vasconcelos, "AGA: Attribute-Guided Augmentation", in CVPR, 2017.

Keywords: Data Augmentation, One-shot Recognition

## 概要

- Attributeに着目した新しいData Augmentation手法を提案
- 論文中ではAttributeとしてDepth, Poseを使って実験している。まずAttribute Regressor を学習してあるサンプルに対するAttributeの予測を可能にしておく。そしてあるサンプルを別のAttributeを持つサンプルに変換するためのEncoder-Decoder Networkを学習。このNetworkはAttribute Regressorの予測と設定した変換後のAttributeの誤差をLossとして学習している。これらの変換は画像空間ではなく特徴空間で行われるため計算コストが少ないことが利点。（多層のConvを入れなくていいから？）

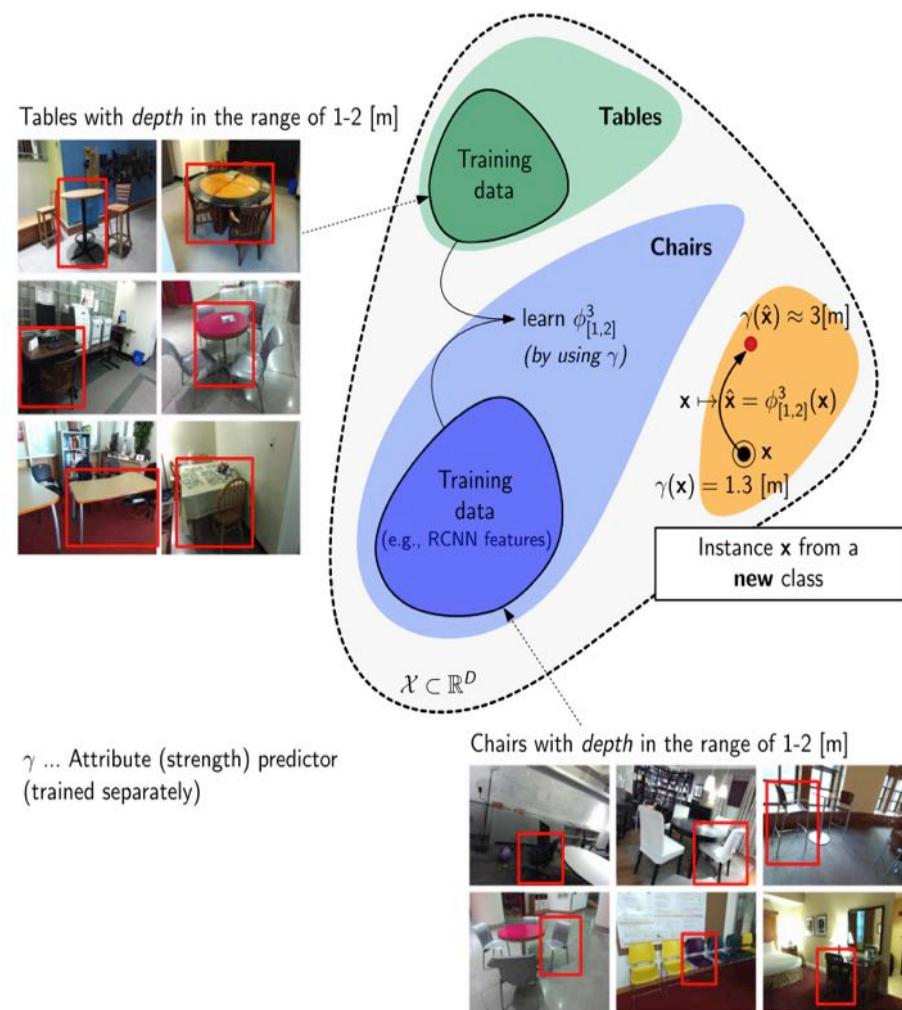
## 新規性・差分

- 従来のData Augmentationは基本的に画像空間でやるのでに対して特徴空間でのAugmentationを提案
- task-specificなものが多かった従来手法より汎用的に利用可能な手法を提案

## Links

論文

[http://www.svcl.ucsd.edu/publications/conference/2017/AGA/AGA\\_final.pdf](http://www.svcl.ucsd.edu/publications/conference/2017/AGA/AGA_final.pdf)

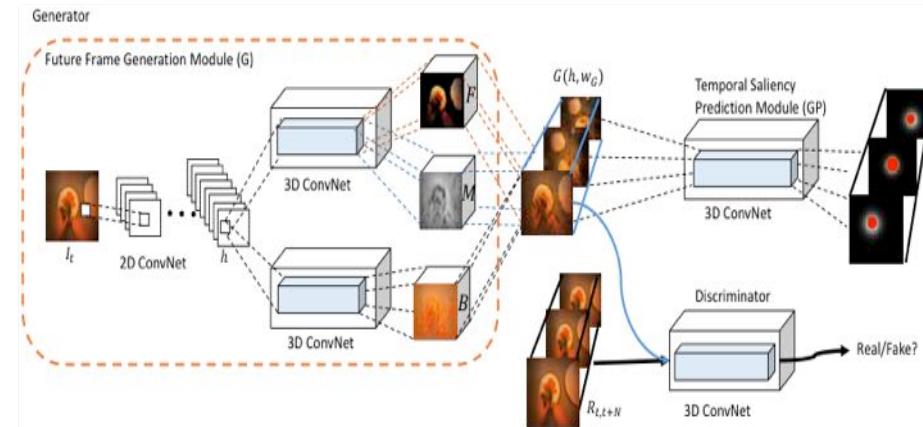


# (12) Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim , "Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks ", in CVPR(oral), 2017.

Keywords: gaze anticipation, egocentric videos, GAN

## 概要

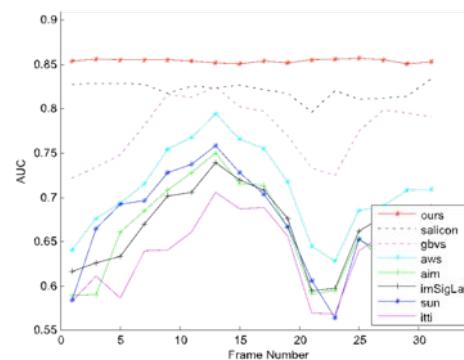
・先フレームの予測を用いた一人称視点における注視位置予測手法。GAN ベースで、あるフレームを条件として先フレーム画像を生成し、その画像から注視位置を推定している。学習はend to endで行われ、モデルの画像生成部分は生成の完成度を求めているというよりモデルに制約を与えることで、結果として注視位置予測の精度を上げているイメージ。生成部分では、単に先フレームを生成するのではなく、Fore ground, Back ground, Maskを生成し、Maskによってそれらの位置ごとの割合を決め、最終的な予測フレームを完成させる(式1)。



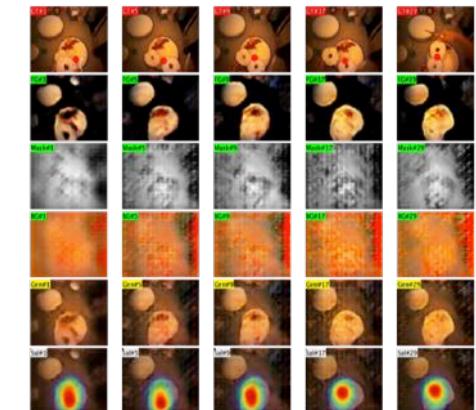
$$I_{t+1,t+N} = F(h(I_t)) \odot M(h(I_t)) \\ + (1 - M(h(I_t))) \odot B(h(I_t)), \quad (1)$$

## 新規性・差分

・GANによる先フレーム予測を一人称視点からの注視位置予測に利用  
・提案するデータセット含む3つのデータセットの予測タスクで従来手法よりもはるかに良い精度。また予測先フレームが遠くなっても安定した精度を保持できていた。



(c) Our OST Dataset (3.2 sec ahead)



## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Zhang\\_Deep\\_Future\\_Gaze\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhang_Deep_Future_Gaze_CVPR_2017_paper.pdf)

プロジェクト

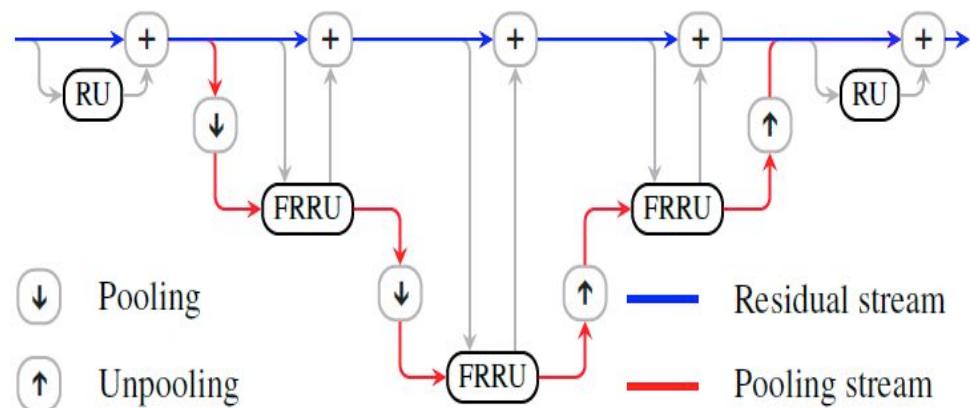
[https://github.com/Mengmi/deepfuturegaze\\_gan](https://github.com/Mengmi/deepfuturegaze_gan)

# 【13】Tobias Pohlen, Alexander Hermans, Markus Mathias, Bastian Leibe, “Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes”, in CVPR, 2017.

Keywords: Semantic Segmentation, Residual Units

## 概要

• Semantic Segmentationの新しい手法を提案。従来のstate-of-the-artはFully Convolutional Network (FCN)ベースでPoolingによって空間的な解像度を落としていく。これはConvolutionの受容野を広げたり変換への頑健性を実現したりするのに有効である反面、正確なLocalizationを阻害する要因となる。そこで、この論文では元の解像度を保ったResidual StreamとPoolingを行っていくPooling Streamの2つを組み合わせることで、両者の利点を合わせた手法を提案。後処理なしで正確なSegmentationを実現し、ImageNetでのPre-trainingを行うことなくstate-of-the-artな精度を達成。



Method	Subsample	Coarse	Pre-trained	$\ominus$	Mean
SegNet [2]	$\times 4$	✓			57.0
FRRN A	$\times 4$				63.0
ENet [44]	$\times 2$				58.3
DeepLab [43]	$\times 2$	✓	✓		64.8
FRRN B	$\times 2$			<b>71.8</b>	
Dilation [55]	$\times 1$	✓			67.1
Adelaide [34]	$\times 1$	✓			<b>71.6</b>
LRR [20]	$\times 1$	✓			69.7
LRR [20]	$\times 1$	✓	✓		<b>71.8</b>

Image      Ground Truth      Ours

## 新規性・差分

• 正確なSemantic Segmentationを行うための高解像度を保持したストリームと頑健さに寄与するPoolingストリームを組み合わせた新しいフレームワークを提案

## Links

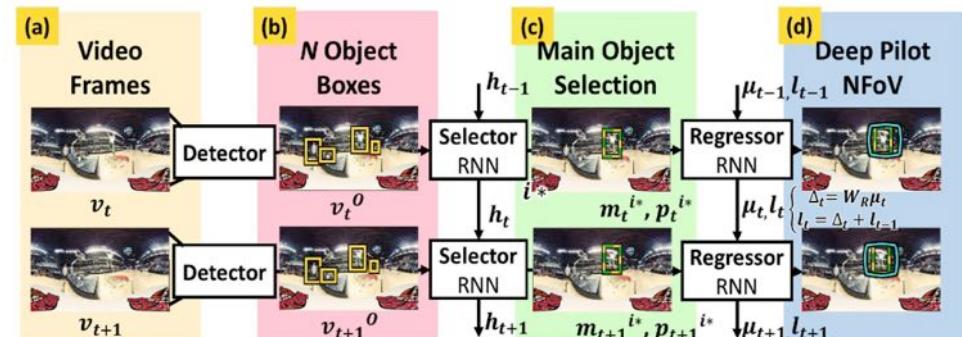
論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Pohlen\\_Full-Resolution\\_Residual\\_Networks\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Pohlen_Full-Resolution_Residual_Networks_CVPR_2017_paper.pdf)  
Github <https://github.com/TobyPDE/FRRN>

# 【14】Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, Min Sun, "Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Videos", in CVPR, 2017. (oral)

Keywords: Auto Camera System, Panoramic Video

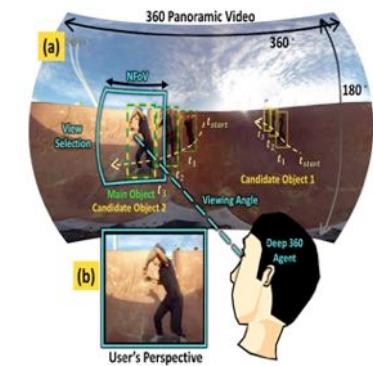
## 概要

- 360度のパノラマ画像からユーザの興味 (e.g. 注視領域) を学習してベストビューとなるようにビデオを切り抜き編集。ユーザの興味はマウス操作や実際に見ている部分を抽出。右図のフローチャートのように、ビデオの入力から物体検出、RNNによるメイン物体選択、最終的には領域を切り出すDeep Pilotによりベストビュービデオを選択する。



## 新規性・差分

- 全周囲のスポーツ映像から自動でベストビュー映像を切り出すDeep 360 Pilotを提案、同時にSports-360 Datasetも提案



## Links

論文

<https://arxiv.org/pdf/1705.01759.pdf>  
動画 <https://www.youtube.com/watch?v=8osw3FIPAvY>

Method	Skateboarding		Parkour		BMX		Dance		Basketball	
	MO	MVD	MO	MVD	MO	MVD	MO	MVD	MO	MVD
Ours w/o Regressor.	<b>0.71</b>	6.03	<b>0.74</b>	4.72	<b>0.71</b>	10.73	<b>0.79</b>	4.32	<b>0.67</b>	8.62
Ours	0.68	<b>3.06</b>	<b>0.74</b>	<b>4.41</b>	0.69	<b>8.36</b>	0.76	<b>2.45</b>	0.66	<b>6.50</b>
AUTOCAM [53]	0.56	0.25	0.56	0.71	0.47	0.55	0.73	0.15	0.51	0.66
RCNN+BMS.	0.25	37.5	0.2	30.8	0.22	32.4	0.24	40.5	0.2	25.27
RCNN+Motion.	0.56	34.8	0.47	26.2	0.42	25.2	0.72	31.4	0.54	25.2

# [15] Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, "Lip Reading Sentences in the Wild", in CVPR2017.(Oral)

Keywords: CNN, LSTM, Lip reading

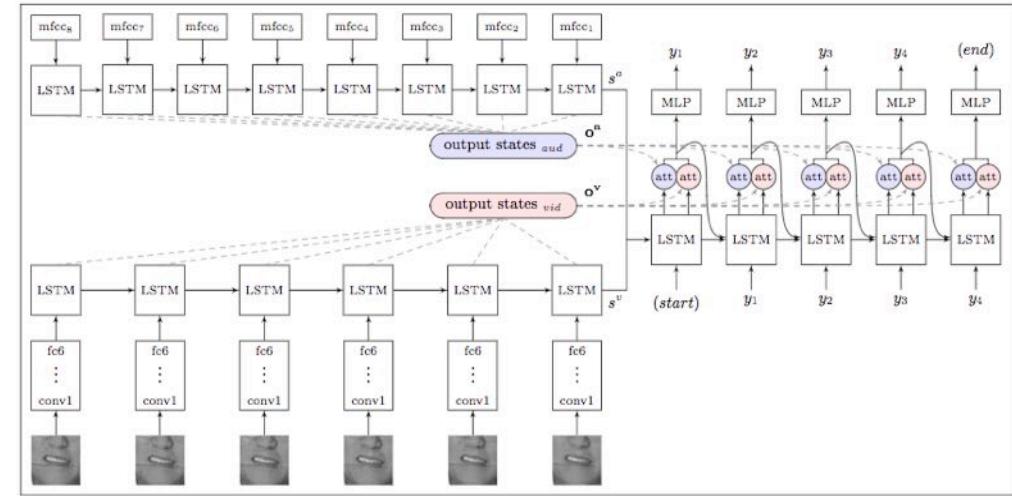
## 概要

- ・音声情報があっても話している顔（唇）から話している文章やフレーズを認識する取り組み
- (1) 独自のネットワークWLASの提案
- (2) 訓練の速度を早め、過剰適合を防ぐための学習戦略
- (3) 英国の大英百科全書から集めた、10万を超える文章からなるLRS/
- ・データセットの作成: LRSデータセットで訓練されたWLASモデルは、BBCテレビの実験において読唇のプロフェッショナルよりも優れたパフォーマンスを示している。

$$s^v, o^v = \text{Watch}(x^v) \quad (2)$$

$$s^a, o^a = \text{Listen}(x^a) \quad (3)$$

$$P(y|x^v, x^a) = \text{Spell}(s^v, s^a, o^v, o^a) \quad (4)$$



Method	SNR	CER	WER	BLEU <sup>†</sup>
Lips only				
Professional <sup>‡</sup>	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9
Audio only				
Google Speech API	clean	17.6%	22.6%	78.4
Kaldi SGMM+MMI*	clean	9.7%	16.8%	83.6
LAS+CL+SS+BS	clean	10.4%	17.7%	84.0
LAS+CL+SS+BS	10dB	26.2%	37.6%	66.4
LAS+CL+SS+BS	0dB	50.3%	62.9%	44.6
Audio and lips				
WLAS+CL+SS+BS	clean	7.9%	13.9%	87.4
WLAS+CL+SS+BS	10dB	17.6%	27.6%	75.3
WLAS+CL+SS+BS	0dB	29.8%	42.0%	63.1



## 新規性・差分

- ・これまで機械学習による読唇は限られた数の単語・文章に対してのみ有効だったのに対して、今回はよりオープンワールドな問題(in the Wild)に対応可能であるとしている

## Links

論文

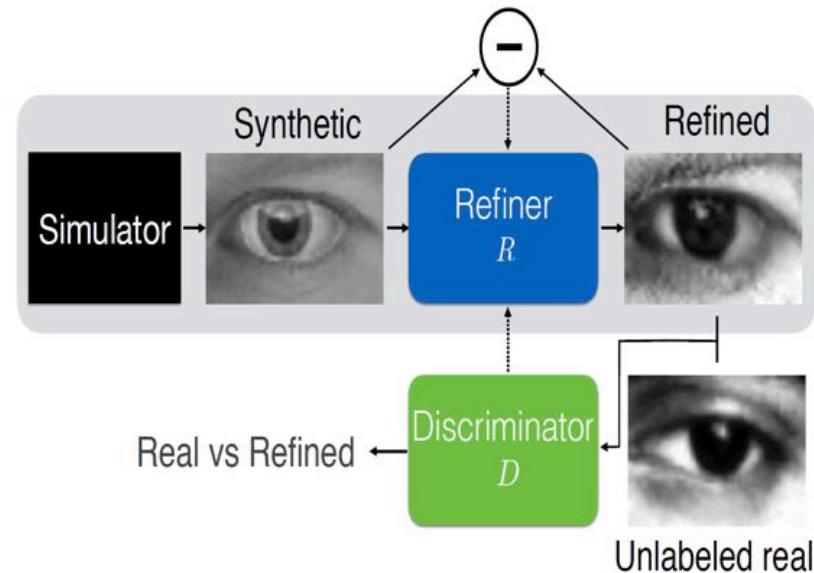
[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Chung\\_Lip\\_Reading\\_Sentences\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Chung_Lip_Reading_Sentences_CVPR_2017_paper.pdf)

- 【16】Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, Russ Webb, "Learning from Simulated and Unsupervised Images through Adversarial Training", in CVPR, 2017. (oral)

Keywords: Synthetic Data, Unsupervised, Adversarial Training

## 概要

- 人工的に生成した合成画像とラベルなしの実画像を使った学習アルゴリズムを提案。ラベル付き実画像を大量に用意するのは大変。合成画像は大量に用意可能だが実画像とのギャップから性能が低下する場合がある。そこで、実画像に近い画像を生成して学習するための手法を提案。基本的な枠組みはGANと同様のAdversarial Training。合成画像を実画像に近づけるように変換するRefinerと、Refineした画像か実画像かを識別するDiscriminatorを対立させ学習。これにより、合成画像をそのまま利用して学習するよりも高い性能を実現できることを示した。



## 新規性・差分

- 合成画像とラベルなし実画像を組み合わせる学習手法を提案 (Simulated+Unsupervised Learning)
- 安定した学習の実現やRefine過程にアーティファクトが出現しないようにするためにGANの枠組みを修正

Training data	% of images within $d$
Synthetic Data	62.3
Synthetic Data 4x	64.9
Refined Synthetic Data	69.4
Refined Synthetic Data 4x	<b>87.2</b>

## Links

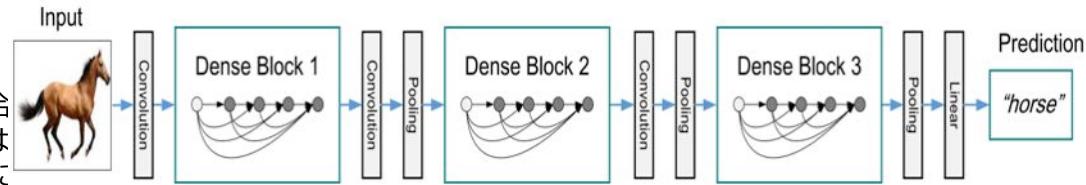
論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Shrivastava\\_Learning\\_From\\_Simulated\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Shrivastava_Learning_From_Simulated_CVPR_2017_paper.pdf)

# [17] Gao Huang, Zhuang Liu, Kilian Q. Weinberger Laurens van der Maaten, "Densely Connected Convolutional Networks", CVPR, 2017.

Keywords: CNN

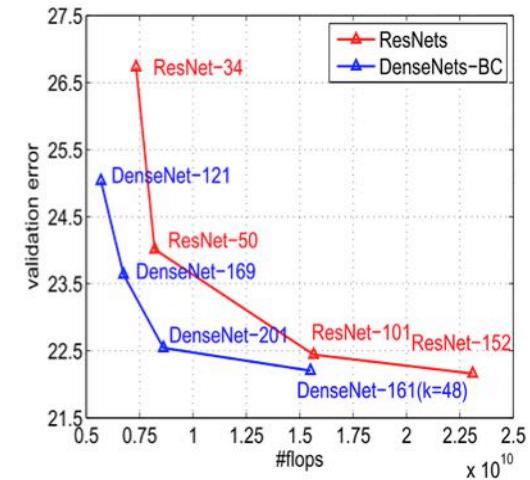
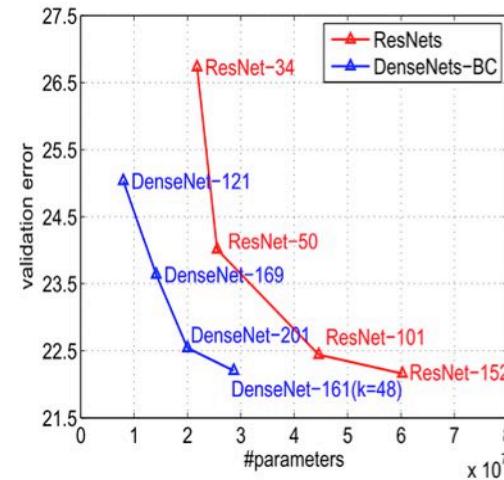
## 概要

- ResNetのような一部のskip connectionではなく、Dense connectionを導入するDenseNetを提案。ブロックごとに以前のすべての層の出力が入力される構造。結合はsumではなくconcatenation。各層での特徴マップ数は少なく設定しており、(k = 12など) 前層との密な結合により多数の特徴マップとなる。前層の再利用によりネットワーク全体のパラメータ数が少なく済むのが利点
- ResNetと比較して少ないパラメータ数でより高精度な結果を達成。



## 新規性・差分

- Dense connectionを採用した新しいネットワーク構造を提案
- 前の層の出力を再利用するためネットワークのパラメータ数が少なくなり省メモリ化、高速化を実現可能



## Links

論文 <https://arxiv.org/pdf/1608.06993.pdf>

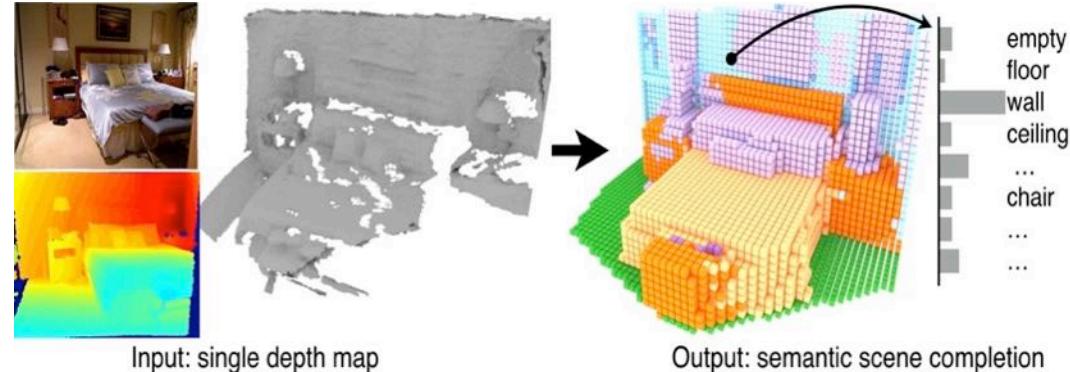
プロジェクト <https://github.com/liuzhuang13/DenseNet>

# 【18】Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, Thomas Funkhouser, "Semantic Scene Completion from a Single Depth Image", in CVPR, 2017. (oral)

Keywords: 3D Completion, Semantic Segmentation

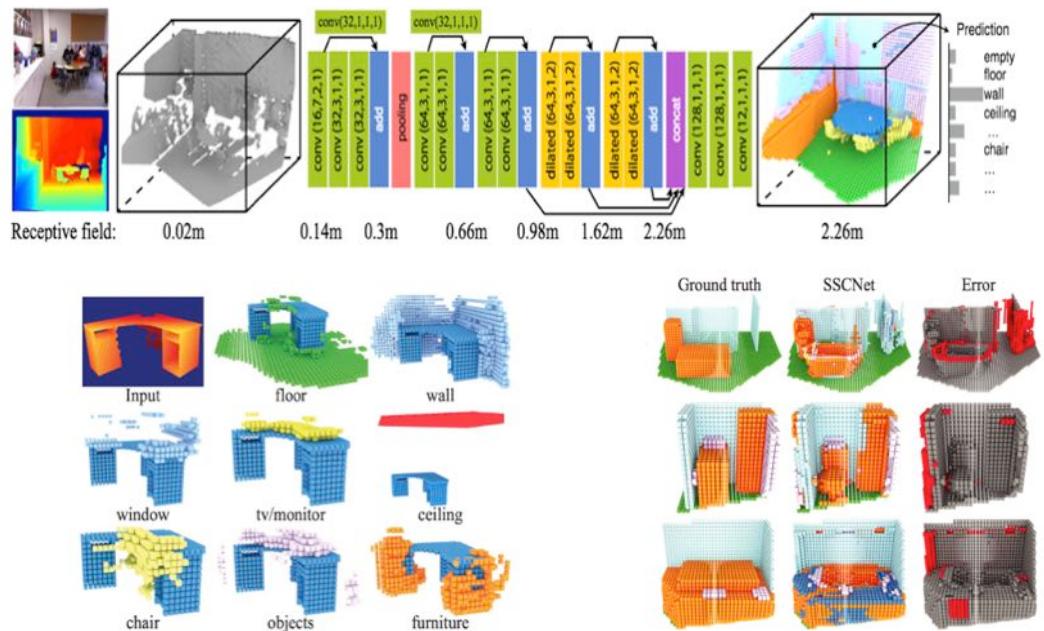
## 概要

• KinectのようなDepthセンサにより生じる（計測に對して物体の反対側の）欠損を、3Dのセマンティックセグメンテーションを用いて補完する研究。オクルージョン（e.g. 異物体同士、セルフオクルージョン）や計測による欠損が生じていても意味情報を使用してCNN（Semantic Scene Completion Network: SSCNet）により欠損を補うモデルとした。学習は意味情報・欠損を含む距離空間・（簡易的な）3次元モデルとした。



## 新規性・差分

- Depthの欠損を補完するネットワークであるSSCNetを提案
- 意味情報・欠損を含む距離空間と3次元モデルを対応付けて学習することにより欠損を補完するモデルとした
- 右下図ではGT・SSCNetによる推定・（GTとの）差分が示されている



## Links

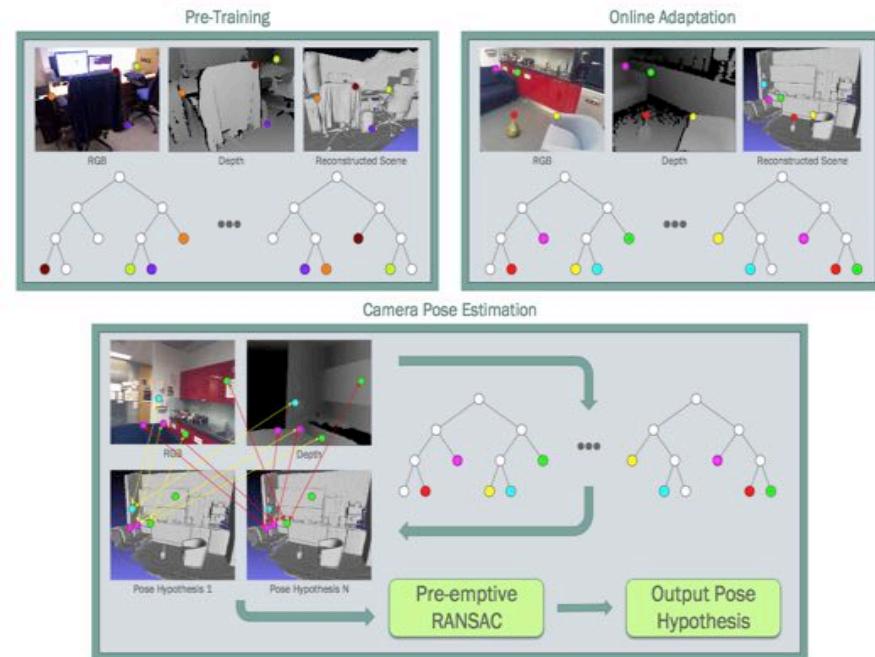
- 論文 <https://arxiv.org/pdf/1611.08974.pdf>  
プロジェクト <http://sscnet.cs.princeton.edu/>  
コード <https://github.com/shurans/sscnet>  
動画 <https://www.youtube.com/watch?v=Yjpmouaap6M>

# 【19】Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Luigi Di Stefano, Philip H. S. Torr, “On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation”, in CVPR , 2017. (oral)

Keywords: Camera Pose Estimation, Random Forests

## 概要

カメラ姿勢推定をランダムフォレスト (Regression Forests) にて行う研究である。対応点のマッチングは2D (RGB and D) から3Dの空間に対して行う。推定された2D-3Dの対応点はKabsch[17]やRANSAC[8]により3次元のカメラ姿勢推定を行う。あらかじめ学習したモデルが手に入ればカメラ姿勢推定は高速 (under 150 ms) に行われる。



## 新規性・差分

- ・2Dから3Dのカメラ姿勢推定をランダムフォレストにより高速 (150ms) に行った
- ・右側の表は7-Scenes dataset[34]の結果である

Training Scene		Relocalisation Performance on Test Scene								Average (all scenes)
		Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs		
Chess	Reloc	99.8%	95.7%	95.5%	91.7%	82.8%	77.9%	25.8%	81.3%	84.9%
	+ ICP	99.9%	97.8%	99.5%	94.1%	91.3%	83.3%	28.4%	84.9%	
Fire	Reloc	98.4%	96.9%	98.2%	89.7%	80.5%	71.9%	28.6%	80.6%	84.6%
	+ ICP	99.1%	99.2%	99.9%	92.1%	89.1%	81.7%	31.0%	84.6%	
Heads	Reloc	98.0%	91.7%	100%	73.1%	77.5%	67.1%	21.8%	75.6%	82.0%
	+ ICP	99.3%	92.3%	100%	81.1%	87.7%	82.0%	31.9%	82.0%	
Office	Reloc	99.2%	96.5%	99.7%	97.6%	84.0%	81.7%	33.6%	84.6%	87.1%
	+ ICP	99.4%	99.0%	100%	98.2%	91.2%	87.0%	35.0%	87.1%	
Pumpkin	Reloc	97.5%	94.9%	96.9%	82.7%	83.5%	70.4%	30.7%	75.5%	84.1%
	+ ICP	98.9%	97.6%	99.4%	86.9%	91.2%	82.3%	32.4%	84.1%	
Kitchen	Reloc	99.9%	95.4%	98.0%	93.3%	83.2%	86.0%	28.2%	83.4%	86.1%
	+ ICP	99.9%	98.2%	100%	94.5%	90.4%	88.1%	31.3%	86.1%	
Stairs	Reloc	97.3%	95.4%	97.9%	90.8%	80.6%	74.5%	45.7%	83.2%	86.3%
	+ ICP	98.0%	97.4%	99.8%	92.1%	89.5%	81.0%	46.6%	86.3%	
Ours (Author's Desk)	Reloc	97.3%	95.7%	97.3%	83.7%	85.3%	71.8%	24.3%	79.3%	84.2%
	+ ICP	99.2%	97.7%	100%	88.2%	90.6%	82.6%	31.0%	84.2%	
Average	Reloc	98.4%	95.3%	97.9%	87.8%	82.2%	75.2%	29.8%	80.9%	84.9%
	+ ICP	99.2%	97.4%	99.8%	90.9%	90.1%	83.5%	33.5%	84.9%	

## Links

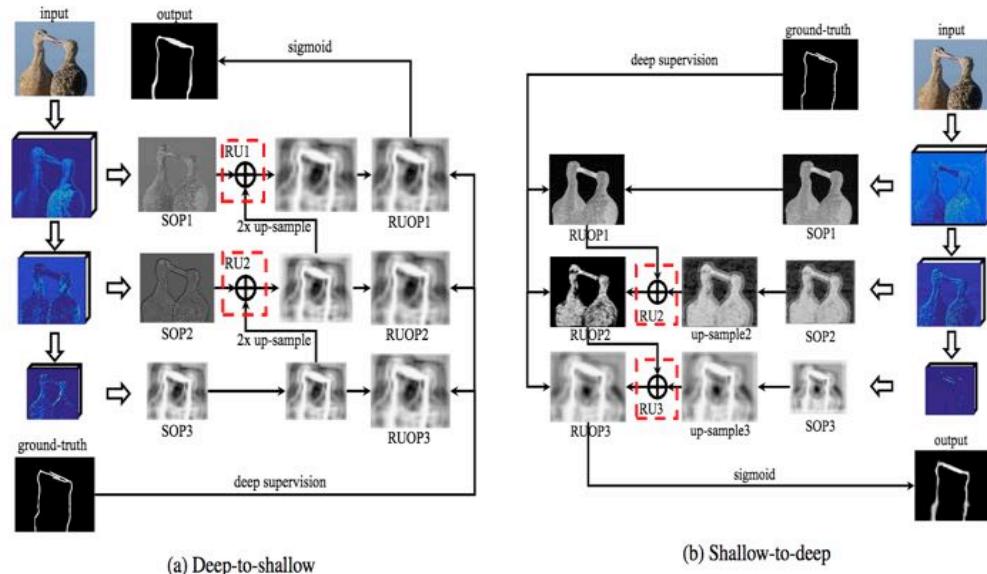
論文 <https://arxiv.org/pdf/1702.02779.pdf>

# 【20】Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, Qixiang Ye, "SRN: Side-output Residual Network for Object Symmetry Detection in the Wild", in CVPR, 2017. (oral)

Keywords: Residual Network, Object Symmetry

## 概要

- ・物体の概要線（物体の姿勢や状態などの概要を表す中心線のこと）を推定するためにSym-Pascal BenchmarkとSide-out Residual Networks (SRN) を提案。ベンチマークはPascalVOCのセマンティックセグメンテーションの正解値から[20]のスケルトン生成器を用いることで概要線としての正解値を生成。カテゴリなどの取捨選択、クロスチェックなどを経て同ベンチマークとして提案。右下図はフローチャートであり、各層に対する出力と正解値の教示を行い、階層的にResidual Unitをスタッキングして概要線の推定を行う。



## 新規性・差分

- ・物体の概要線を抽出するために必要なアノテーションを含んだベンチマークを作成した
- ・Residual構造による畳み込みをスタッキングしたSide-output Residual Network (SRN) を提案した。SRNは複雑背景下で対称性を的確に把握して概要線を推定することができる

## Links

論文 <https://arxiv.org/pdf/1703.02243.pdf>  
GitHub <https://github.com/KevinKecc/SRN>

SYMMAX



Methods	F-measure	Runtime(s)
Partical Filter [27]	0.129	25.30
Levinshtein [11]	0.134	183.87
Lee [9]	0.135	658.94
Lindeberg [12]	0.138	5.79
MIL [26]	0.174	80.35
HED (baseline) [28]	0.369	<b>0.10†</b>
FSDS [22]	0.418	0.12†
FasterRCNN [17]+FSDS [22]	0.343	0.33†
YOLO [16]+FSDS [22]	0.354	0.12†
FCN [15]+[20]	0.386	0.76†
SRN (ours)	<b>0.443</b>	0.12†

[21]

Ning Xu, Brian Price, Scott Cohen, Thomas Huang, "Deep Image Matting", in CVPR, 2017. (oral)

Keywords: Image Matting, Encoder-Decoder

## 概要

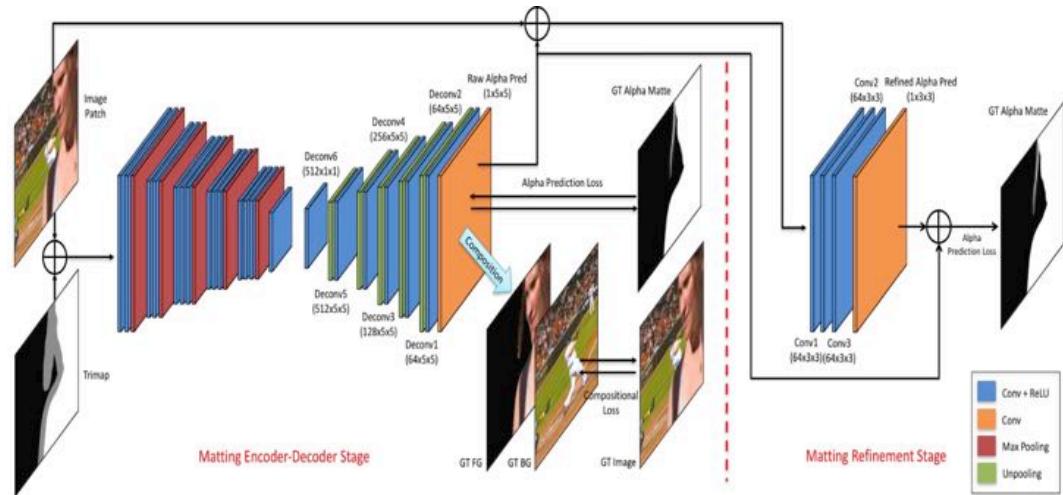
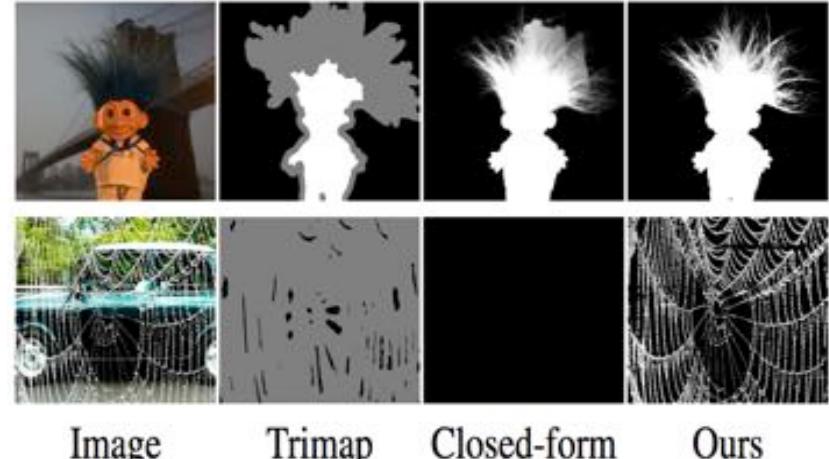
・Image Matting（静止画に対する背景除去）の従来の問題点としては前景と背景が似ている際に切り抜きができなかつたが、本論文では（コンテキストなど）より高次な情報を用いてこの問題を解決する。DNNのモデルではEncoder-Decoderモデルや、alpha matte（重みを調整して背景除去？）を実現する small DNNを採用。Trimap (foreground/ background/ unknown) を用いた二段階学習により推定とリファインを行う。

## 新規性・差分

・静止画における背景除去において、今までで最高精度を達成。背景・前景・unknownな領域を示したTrimapから詳細な領域除去を行うことに成功した

Links 論文 <https://arxiv.org/pdf/1703.03872.pdf>

Methods	SAD	MSE	Gradient	Connectivity
Shared Matting [13]	128.9	0.091	126.5	135.3
Learning Based Matting [34]	113.9	0.048	91.6	122.2
Comprehensive Sampling [28]	143.8	0.071	102.2	142.7
Global Matting [16]	133.6	0.068	97.6	133.3
Closed-Form Matting [22]	168.1	0.091	126.9	167.9
KNN Matting [5]	175.4	0.103	124.1	176.4
DCNN Matting [8]	161.4	0.087	115.1	161.9
<i>Encoder-Decoder network (single alpha prediction loss)</i>	59.6	0.019	40.5	59.3
<i>Encoder-Decoder network</i>	54.6	0.017	36.7	55.3
<i>Encoder-Decoder network + Guided filter [17]</i>	52.2	0.016	<b>30.0</b>	52.6
<i>Encoder-Decoder network + Refinement network</i>	<b>50.4</b>	<b>0.014</b>	31.0	<b>50.8</b>

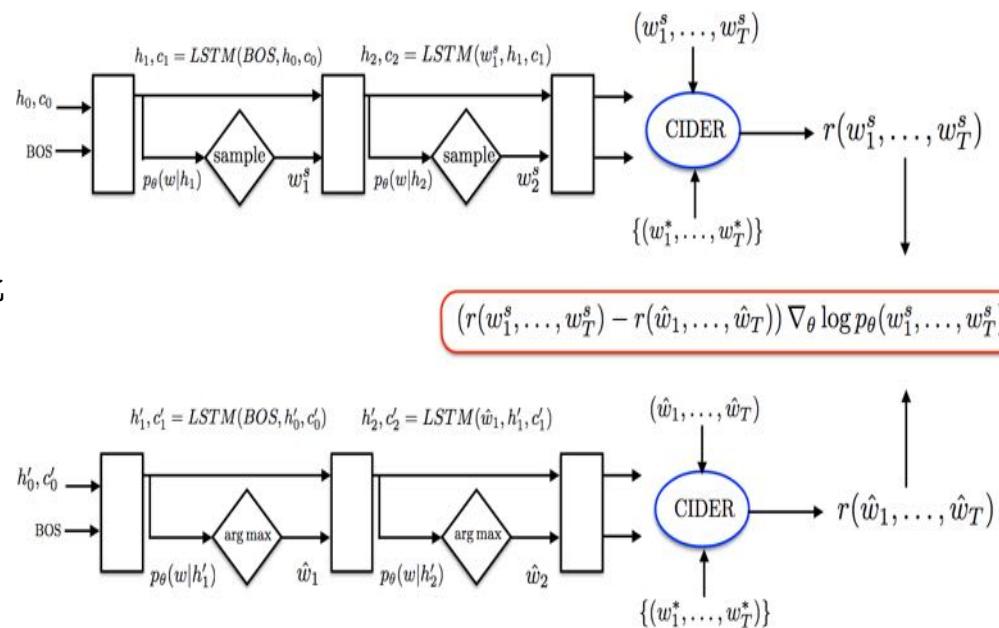


## 【22】Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, Vaibhava Goel, "Self-critical Sequence Training for Image Captioning", in CVPR, 2017. (oral)

Keywords: Reinforcement Learning, Image Captioning

### 概要

- ・強化学習を用いた画像説明文の研究である。強化学習にて獲得する報酬などパラメータは不安定というところから、注意深く最適化、MSCOCOの評価値を用いて正規化。正規化にはSelf-Critical Sequence Training (SCST)を導入して、強化学習の不安定さを緩和し、画像説明文を生成する。画像説明文を推定する基本モデルは主にAttentionから特徴を抽出するCNN-LSTMモデル (Show, attend and tell) である。



- ・学習段階で獲得したパラメータに対して強化を行う。サンプルのセンテンスと推定のセンテンスの差分から報酬を発生させ、テスト時により表現力豊かになるよう（正規化を行いつつ）パラメータを強化。

### 新規性・差分

- ・本手法SCSTを用いた強化学習におけるパラメータ安定の正規化により、画像説明文の評価値を104.9から112.3(CIDEr)に引き上げた@MSCOCO
- ・いわゆる（単純な）強化学習のように報酬が発生した場面を覚えるのではなく、パラメータ自体を正規化することにより柔軟な”強化（Reinforce）”ができる
- ・実はIBMのWatsonのグループ

### Links

論文 <https://arxiv.org/pdf/1612.00563.pdf>

プロジェクト

<https://github.com/ruotianluo/self-critical.pytorch>

Ensemble SCST models	Evaluation Metric			
	CIDEr	BLEU4	ROUGEL	METEOR
Ens. 4 (Att2in)	<b>112.3</b>	<b>34.4</b>	<b>55.9</b>	<b>26.8</b>
Previous best	104.9	34.3	55.2	26.6

【23】Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, Shuicheng Yan, "Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach", in arXiv 1703.08448, 2017 (CVPR2017).

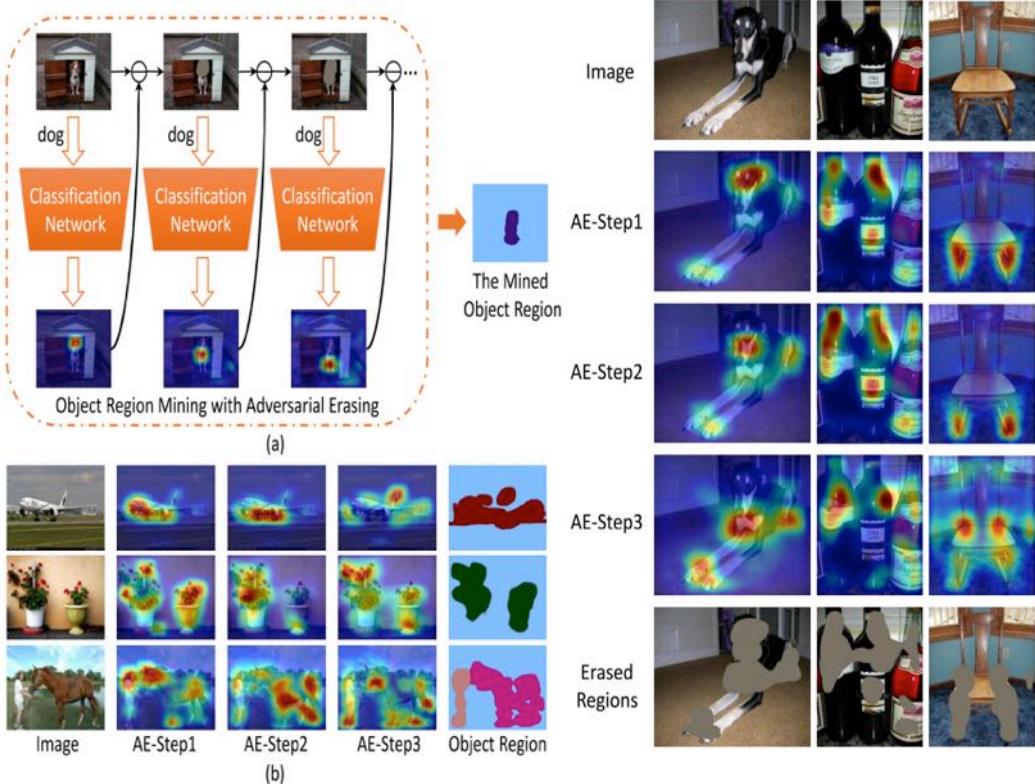
Keywords: Semantic Segmentation, Weakly Supervised

## 概要

- Weakly Supervised Semantic Segmentationのための新しい手法を提案。Classification Networkを予め学習しておき各物体を認識する際に画像中のどこが使われているかを計算できるようにしておく。認識に使われた領域を入力画像から消して再び認識する、ということを繰り返すと異なる領域が次々に選択されていくことになる。使われた領域は対象の物体の領域ということで、この領域をGround TruthとしてSemantic Segmentationを学習する手法を提案。従来のWeakly Supervised手法よりも高い精度を達成した。

## 新規性・差分

- 画像レベルのラベルから物体領域をマイニングするための新しい手法を提案
- 今回の手法でマイニングしたGround Truthから頑健に学習するための新しいOnline Prohibitive Segmentation Learningも提案



## Links

論文 <https://arxiv.org/pdf/1703.08448.pdf>

Table 3. Comparison of segmentation mIoU scores using object regions from different AE steps on VOC 2012 val set.

AE Steps	bkg	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
AE-step1	82.6	63.0	27.5	45.9	38.3	43.6	61.3	29.2	60.0	13.6	52.0	32.6	52.4	49.8	47.9	43.7	32.6	61.4	29.4	35.1	41.9	44.9
AE-step2	82.2	69.3	29.7	60.9	40.8	52.4	59.3	44.2	65.3	13.0	58.9	32.2	60.0	56.6	49.1	43.0	34.2	69.7	32.1	42.8	43.2	49.5
AE-step3	78.5	71.8	29.2	64.1	39.9	57.8	58.5	54.5	63.0	10.3	60.5	36.0	61.6	56.1	62.6	42.9	36.5	64.5	31.5	49.5	38.7	50.9
AE-step4	74.4	65.5	28.2	59.7	38.5	57.8	57.5	59.0	57.2	9.6	54.9	39.2	56.5	52.6	65.0	43.2	34.9	55.9	30.4	47.9	36.8	48.8

# 【24】Huazhe Xu, Yang Gao, Fisher Yu, Trevor Darrell, "End-to-End Learning of Driving Models from Large-scale Video Datasets", in CVPR, 2017.

Keywords: End-to-End Self-driving

## 概要

- End-to-End学習で自動運転の操作を学習するネットワークを実装した。ビデオの入力からDilated FCNによるセマンティックセグメンテーションを行う。セマンティックセグメンテーションの場合にはひとまず誤差を計算してパラメータを調整。その後、セマンティック画像と過去のモーションをLSTMへの入力として自動運転のためのモーションを学習する。

## 新規性・差分

- End-to-Endで自動運転のためのアクションを学習できる仕組みを提案した。比較対象としてCNN-LSTMやTCNNよりも提案のFCN-LSTMの方が良好な精度を出すことを示し、他の学習手法よりも提案手法(Privileged Training Approach)の方が良いことを示した。

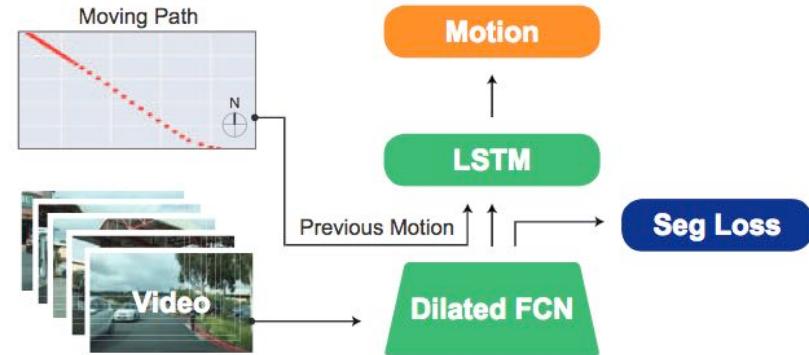
## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Xu\\_End-To-End\\_Learning\\_of\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Xu_End-To-End_Learning_of_CVPR_2017_paper.pdf)

プロジェクト

method	perplexity	accuracy
Motion Reflex Approach	0.718	71.31%
Mediated Perception Approach	0.8887	61.66
Privileged Training Approach	<b>0.697</b>	<b>72.4%</b>

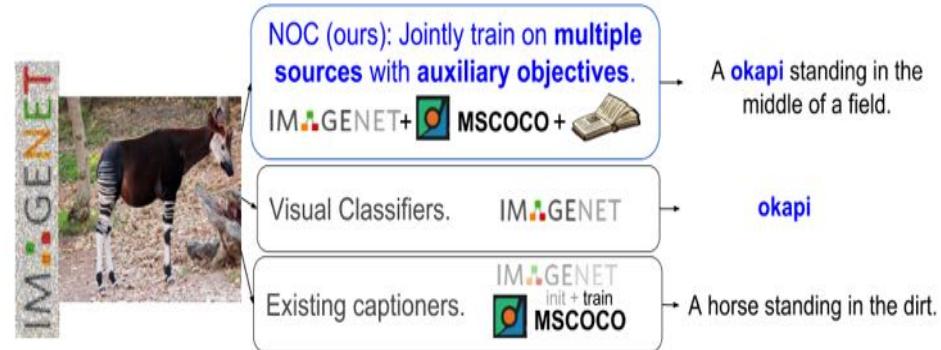


# 【25】Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, “Captioning Images with Diverse Objects”, in arXiv 1606.07770, 2016 (CVPR2017).

Keywords: CNN, LSTM, Image Captioning, Novel Object

## 概要

- 複数のデータセットを合わせて学習することで、画像
- ・キャプションペアのアノテーションが与えられていない対象に対するCaptioningを実現。CaptioningにはVisual Recognition Network (CNN) とLanguage Model (LSTM) を使う。普通のCaptioningだとそれをPre-Trainingしてから画像・キャプションペアを使ってCaptioniongのFine-tuningをする。しかし、それだとこのペアに含まれていない物体のことを忘れていくつていまう。そこでこの研究では3つを同時に使う手法を提案した。

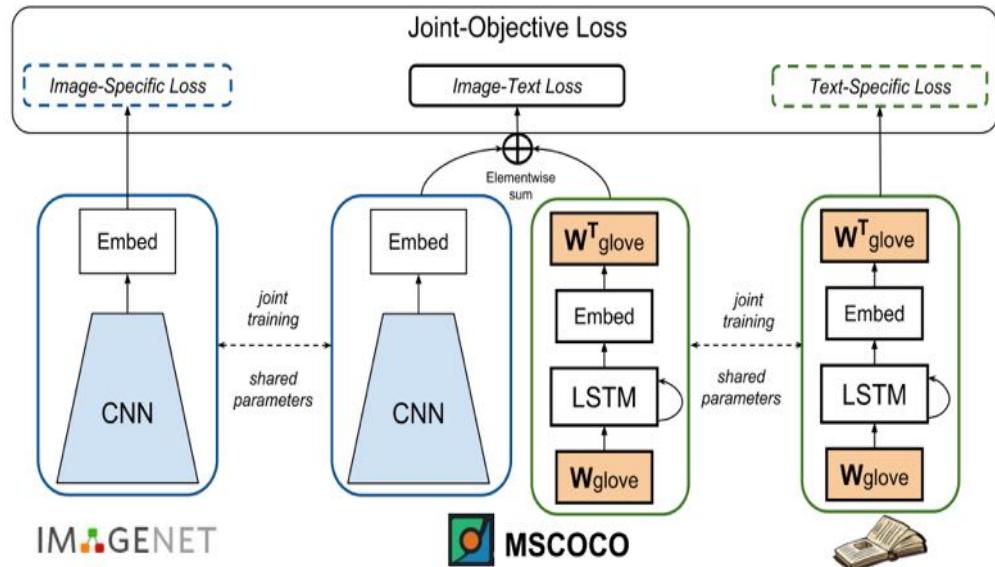


## 新規性・差分

- 画像・キャプションペアのデータセットに含まれていない対象のCaptioningを実現するためのJoint-Objective Lossを提案

## Links

論文 <https://arxiv.org/pdf/1606.07770.pdf>

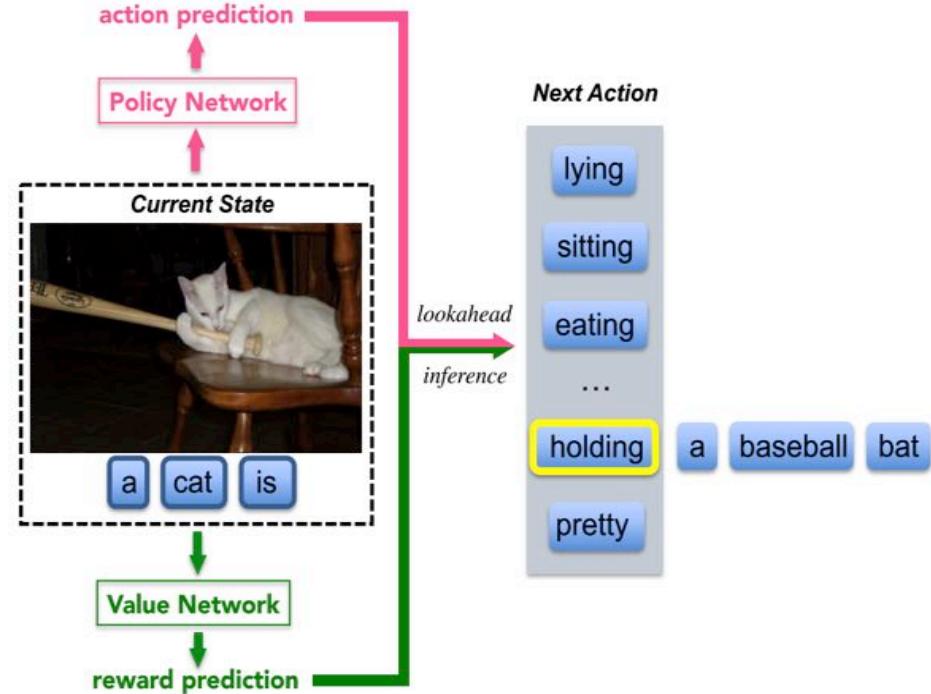


【26】Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, Li-Jia Li, “Deep Reinforcement Learning-based Image Captioning with Embedding Reward”, in arXiv 1704.03899, 2017 (CVPR2017).

Keywords: CNN, LSTM, Image Captioning, Reinforcement Learning

## 概要

- 強化学習を用いてCaptioningを行う手法を提案。従来手法は基本的にCNN & LSTMでCaptioningを行うが、LSTMは現在の状態に基づいてgreedyに次の単語を決定していくためそのときに確率が低い単語は選択できないのが問題。この論文の手法はPolicy Network (CNN & LSTM)による次の単語の予測 (local) とValue Networkによるrewardの予測 (global) を組み合わせることで確率が低い単語でも最終的なrewardが高そうな単語であれば選択されるので良い。強化学習のためのrewardとしてはVisual-Semantic Embeddingによる画像とテキストの類似度を用いる手法を提案。state-of-the-artを超える精度を達成した。



## 新規性・差分

- 強化学習によるImage Captioningを提案
- Visual-Semantic EmbeddingによるRewardを設計

## Links

論文 <https://arxiv.org/pdf/1606.07770.pdf>

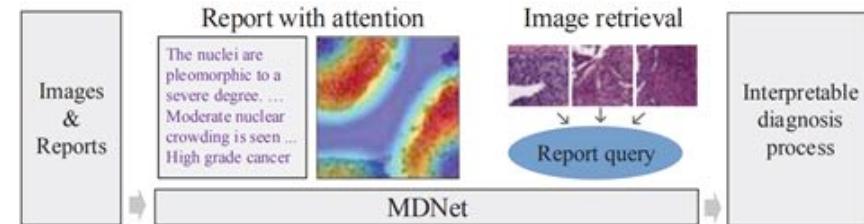
Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	Rouge-L	CIDEr
Google NIC [44]	0.666	0.461	0.329	0.246	—	—	—
m-RNN [30]	0.67	0.49	0.35	0.25	—	—	—
BRNN [17]	0.642	0.451	0.304	0.203	—	—	—
LRCN [7]	0.628	0.442	0.304	0.21	—	—	—
MSR/CMU [3]	—	—	—	0.19	0.204	—	—
Spatial ATT [46]	<b>0.718</b>	0.504	0.357	0.25	0.23	—	—
gLSTM [15]	0.67	0.491	0.358	0.264	0.227	—	0.813
MIXER [35]	—	—	—	0.29	—	—	—
Semantic ATT [48] *	0.709	0.537	0.402	<b>0.304</b>	0.243	—	—
DCC [13] *	0.644	—	—	—	0.21	—	—
Ours	0.713	<b>0.539</b>	<b>0.403</b>	<b>0.304</b>	<b>0.251</b>	<b>0.525</b>	<b>0.937</b>

**[27] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, Lin Yang, "MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network", in CVPR Oral, 2017.**

Keywords: computer-aided diagnosis, attention image generation, diagnosis report generation, image retrieval, ensemble connection

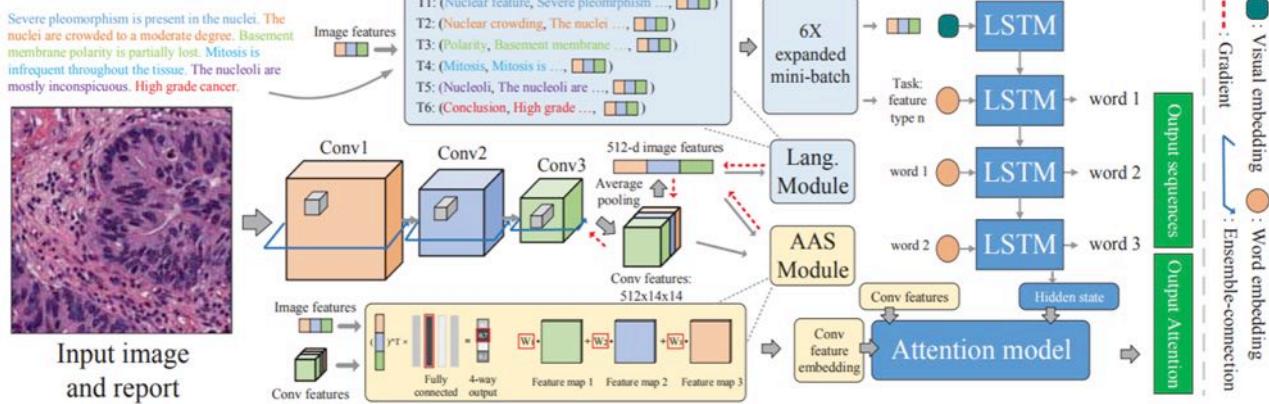
## 概要

- 医療における疾病診断レポートと医療画像の間の直接マッピングを実現するMDNetの提案。画像モデルと言語モデルの併用による。画像モデルではResNetを拡張し、SkipConnectionをEnsenbleConnectionとして再構成する。言語モデルではLSTMを用いる。さらに、学習データに領域レベルのラベルを必要としない枠組みのために、AAS(auxiliary attention sharpening)モジュールを提案する。画像モデル、言語モデル、AASモジュールのパラメータを最急降下法により最適化する。



## 新規性・差分

- EnsenbleConnectionによるネットワークによる画像認識タスクで高い性能
- 診断レポート生成、関連画像検索においてstate-of-the-artの性能
- 学習データに領域ラベリング不要
- 注目領域推定の、診断医による定性評価



## Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Zhang\\_MDNet\\_A\\_Semantically\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhang_MDNet_A_Semantically_CVPR_2017_paper.pdf)

# 【28】Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, Jitendra Malik, "Multi-view Supervision for Single-view Reconstruction via Differentiable Ray Consistency", in CVPR, 2017. (oral)

Keywords: Self-supervised 2d image from 3d model

## 概要

- ・2D-3Dの対応関係により、3次元モデルを自己学習する。Differentiable Ray Consistency (DRC)によりビュー普遍性を定式化する。右図は提案手法の概念を示しており、推定された3DモデルからDRCにより2次元画像を生成してこれを2D-3D対応の教師データとする。図の右側は2D画像を入力とした際の、3Dモデルを生成した結果である。



## 新規性・差分

- ・2D-3Dの自己学習のための枠組みであるDRCを提案。推定された3Dモデルから2Dモデルを生成してこれを応付けて学習を行い、2D画像から3Dモデルを推定。
- ・右図はアプリケーションの例である。

Multi-view Mask/Depth Supervision (ShapeNet)



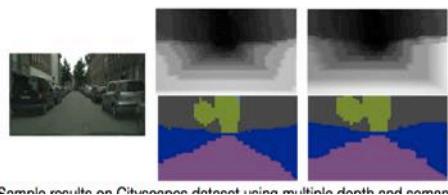
Sample results on ShapeNet dataset using multiple depth images as supervision for training. a) Input image. b,c) Predicted 3D shape.

Single-view Mask Supervision (PASCAL VOC)



Sample results on PASCAL VOC dataset using pose and foreground masks as supervision for training. a) Input image. b,c) Predicted 3D shape.

Driving Sequences as Multi-view Supervision (Cityscapes)



Sample results on Cityscapes dataset using multi-view depth and semantic

Multi-view Color Images as Supervision (ShapeNet)



## Links

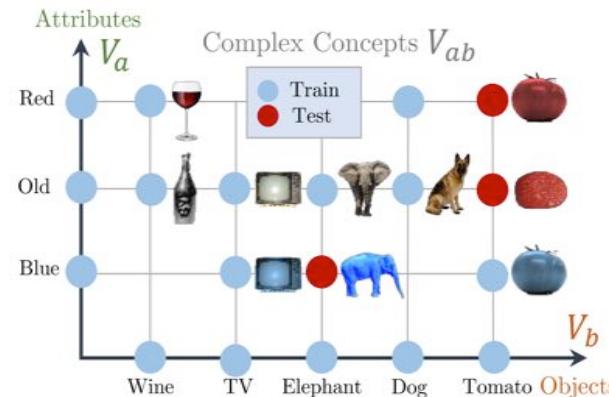
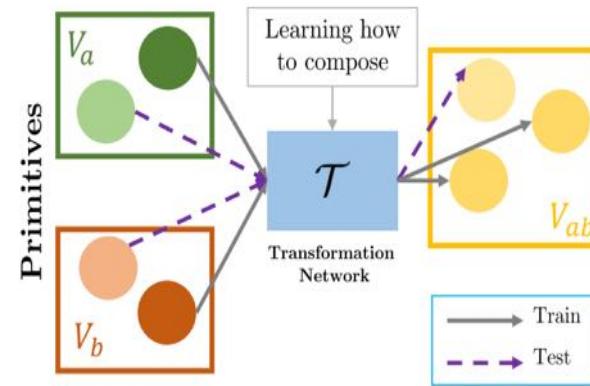
論文 <https://arxiv.org/pdf/1704.06254.pdf>  
プロジェクト <https://shubhtuls.github.io/drc/>  
GitHub <https://github.com/shubhtuls/drc>

# (29) Ishan Misra Abhinav Gupta Martial Hebert, "From Red Wine to Red Tomato: Composition with Context", in CVPR, 2017.

Keywords: Attributes, Adjective, Visual Concept

## 概要

- Attribute (形容詞) を含む表現 (Complex Concepts) を学習するための手法を提案。基本的にはAttributesとObjectsに分けてそれぞれを学習し、それらのモデルの組み合わせで対象を表現。AttributeのClassifierの重みとObjectsのClassifierの重みを入力とし、それらの組み合わせの表現を出力するTransformation Networkを提案。モデル (Classifier weight) 空間では似ている物体は近く、異なる物体は遠いということを仮定しており、異なる物体 (TomatoとWine) よりも近い物体 (TomatoとBerry) と同様の組み合わせ方を行うことで未知の物体に対しても適切な組み合わせ表現を実現。（"Red" Tomatoと"Red" Wineは同じ"Red"というAttributeを持つが、色としては異なる）



## 新規性・差分

- Transformation Networkにより未知のObject-Attributeの組み合わせの表現を実現

## Links

論文

[http://www.cs.cmu.edu/~imisra/data/composing\\_cvpr17.pdf](http://www.cs.cmu.edu/~imisra/data/composing_cvpr17.pdf)

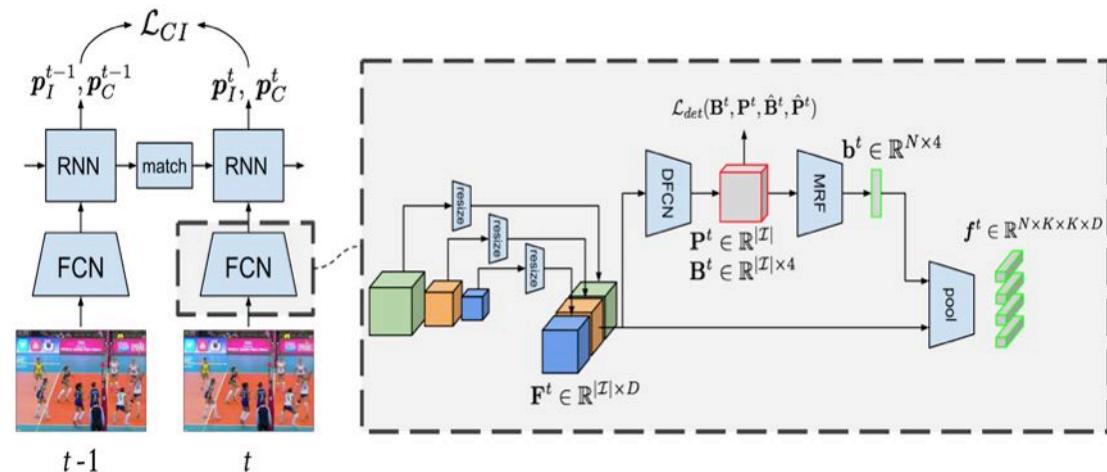


【30】Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, Silvio Savarese, "Social Scene Understanding: End-to-End Multi-Person Action Localization and Collective Activity Recognition", in CVPR, 2017.

Keywords: Social Action, Collective Activity

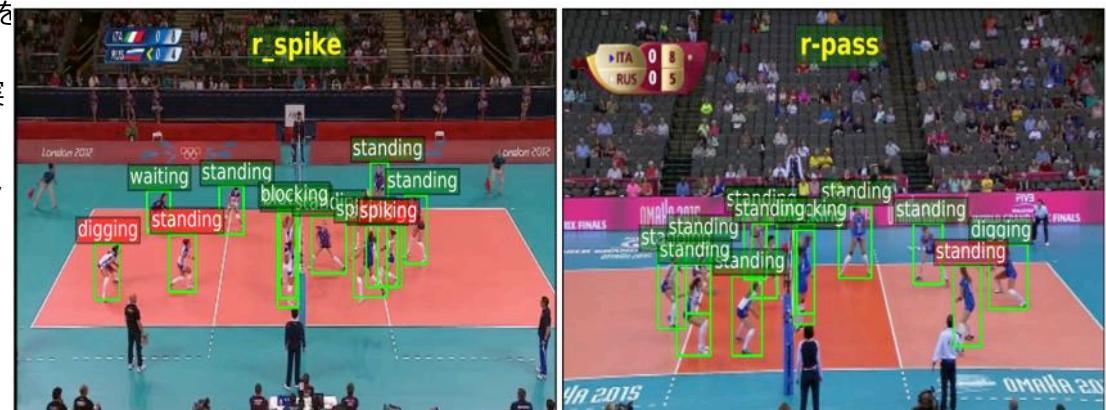
## 概要

- Raw Videoを入力として複数人の検出と個々の行動の認識、集団としての行動の認識を行う手法を提案
- 各フレームごとにCNN (FCN) に入力し、N個の Bounding Boxに対する特徴ベクトルを算出、それを RNNに入力してTemporal Reasoningすることで洗練化。これら全体をEnd-to-Endで学習するためのフレームワークを実現。



## 新規性・差分

- 事前の人検出などを一切することなく Raw Videoを入力として Individual Action, Collective Activity の認識と、そのEnd-to-Endの学習フレームワークを実現
- 従来の手法（人物領域のGround Truthを利用）した手法よりも高い精度を実現



## Links

論文

<http://www.idiap.ch/~fleuret/papers/bagautdinov-et-al-cvpr2017.pdf>

Github

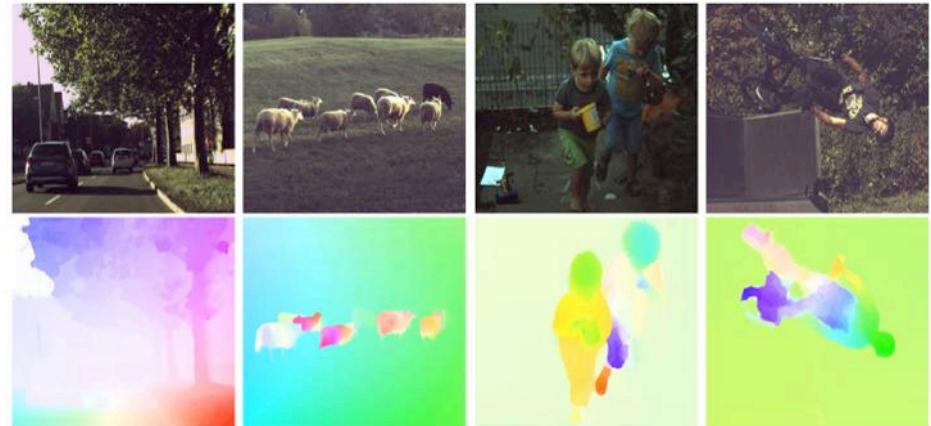
<https://github.com/cvlab-epfl/social-scene-understanding>

【31】Joel Janai, Fatma Guney, Jonas Wulff, Michael Black, Andreas Geiger,  
“Slow Flow: Exploiting High-Speed Cameras for Accurate and Diverse  
Optical Flow Reference Data”, in CVPR, 2017.

Keywords: Optical Flow

## 概要

- Optical Flowの学習に必要な大規模データセットを構築するためのフレームワークを提案。ハイスピードカメラを用いて撮影する高解像 (2560x1440) かつ高フレームレート (> 200 fps) の動画をこの手法では利用。これにより従来の撮影環境が制限されていたり、レーザスキャナなどの機器を利用する手法よりも容易にデータの取得が可能。このような大規模な時空間に対して密な画素追跡を行う手法は従来にはなかったため、それを行うための手法 (Slow Flow) も提案。従来よりも高い精度のOptical Flow推定を実現し、Reference Dataとして活用可能なデータが生成できることを示した。



## 新規性・差分

- ハイスピードカメラを利用したデータ生成手法を提案
- 高解像 (Space-Time) な動画に対しても密な画素追跡を行うことができる手法を提案

## Links

論文 <http://www.cvlabs.net/publications/Janai2017CVPR.pdf>  
プロジェクト [http://www.cvlabs.net/projects/slow\\_flow/](http://www.cvlabs.net/projects/slow_flow/)

Methods	All (Edges)	Visible (E.)	Occluded (E.)
Epic Flow (24fps)	5.53 (16.23)	2.45 (10.10)	16.54 (20.68)
Epic Flow (Accu. 144fps)	4.73 (12.76)	1.04 (4.41)	17.09 (18.44)
Slow Flow (Accu. 144fps)	4.03 (12.03)	0.78 (4.43)	15.24 (17.28)
Slow Flow (Accu. 1008fps)	5.38 (11.78)	1.35 ( <b>2.60</b> )	19.18 (17.93)
Slow Flow (Full Model)	<b>2.58 (10.06)</b>	<b>0.87</b> (4.65)	<b>9.45 (14.28)</b>

(a) EPE on MPI Sintel

# [32] Amir R. Zamir et al., "Feedback Networks", in arXiv:1612.09508, 2016.

Keywords: LSTM, CNN, cifar 100

## 概要

- 通常CNNではfeedforwardのネットワーク構成だが、この論文ではLSTMの構造をCNNに応用することでfeedbackによる推定が可能なネットワーク構成を提案する。cifar 100とstanford cars datasetを用いて分類精度実験、MPII Human Pose estimation benchmarkにて姿勢推定の精度実験を行っている。

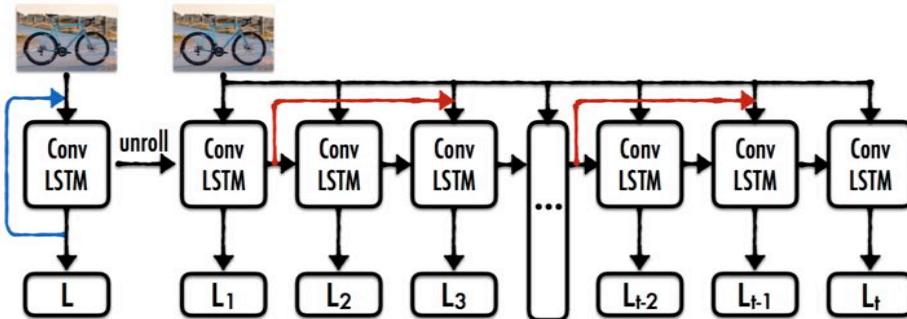
## 新規性・差分

- LSTMをCNNに応用したConvolutinal LSTMの提案。また、skip connectionという時間方向にスキップするコネクションを用いることで精度の向上が見込める。ConvLSTMは基本的に通常のLSTMの内積計算を畳み込みに置き換えたものになっている。ConvLSTM内に複数の畳み込み層やBatchNormalization層を加えたモデルも提案している。

## Links

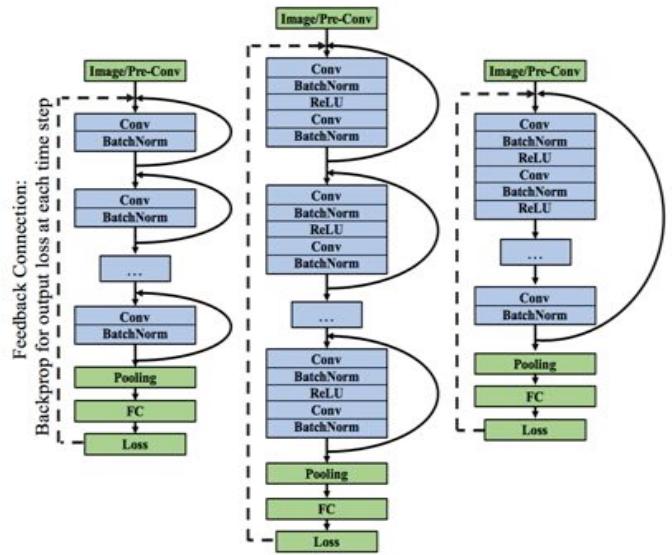
論文 <https://arxiv.org/pdf/1612.09508v2.pdf>

プロジェクト <http://feedbacknet.stanford.edu/>



Model	CL	Top1(%) - Fine	Top1(%) - Coarse
Feedback Net	N	68.21	79.7
	Y	<b>69.57(+1.34%)</b>	<b>80.81(+1.11%)</b>
Feedforward	N	69.36	80.29
	Y	69.24(-0.12%)	80.20(-0.09%)
ResNet w/ Aux loss	N	69.36	80.29
	Y	65.69(-3.67%)	76.94(-3.35%)
Feedforward	N	63.56	75.32
	Y	64.62(+1.06%)	77.18(+1.86%)
VGG w/ Aux loss	N	63.56	75.32
	Y	63.2(-0.36%)	74.97(-0.35%)

Table 5. Evaluation of the impact of Curriculum Learning (CL) on CIFAR100. The CL column denotes if curriculum learning was used. The difference made by curriculum for each method is shown in parentheses.



# 【33】Peiyun Hu, Deva Ramanan, "Finding Tiny Faces", in arXiv pre-print 1612.04402, 2016.

Keywords: Face Detection, Tiny Faces, WIDER Face, FDDB

## 概要

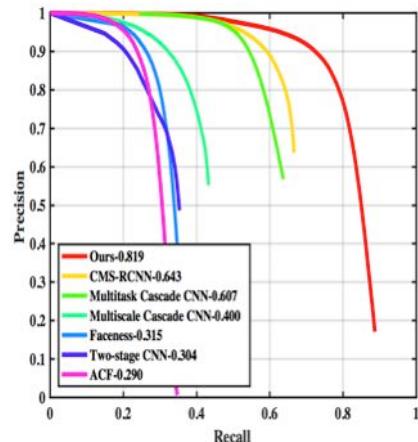
- ・物体検出の試みが2014年からの数年で進んでいるが、ひとつの課題が小さな物体への対応である。本論文では3つの側面（スケール普遍性、画像解像度、コンテキストを考慮した意味づけ）から小さな顔を検出する手法を模索した。極めて小さい顔(e.g. 3pxl tall)は通常の顔とは異なるため、異なる学習を実行した。効率化を図るために、完全に分離した学習ではなく、Multi-task学習を行う。通常の顔に対してはStraight-forwardな学習を行うが、小さな物体に対してはコンテキストが効果的である。FDDBやWIDER FACEなど巨大なデータベースにおいてState-of-the-artな精度を達成した。
- ・小さな顔ほど、画像パッチに対するTight-fittingが重要である。面白いことに小さな顔は小さな受容野(Receptive Field)が効果的であった（小さな顔はオクルージョンがなく、全て見えていることが多い）。

## 新規性・差分

- ・複数解像度や複数のデータセット拡張やMulti-task学習により極めて小さな顔を検出した
- ・FDDBやWIDER FACEなど巨大なDBに対してState-of-the-artな精度（右図）

## Links

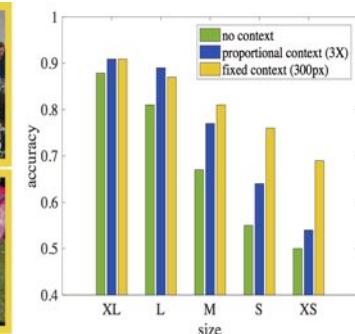
論文 <https://arxiv.org/pdf/1612.04402.pdf>  
プロジェクト  
<https://www.cs.cmu.edu/~peiyunh/tiny/>  
FDDB <http://vis-www.cs.umass.edu/fddb/>  
WIDER FACE  
<http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/>  
コード <https://github.com/peiyunh/tiny>



・小さな顔ほど、コンテキストが驚くほど効いている。右表はサイズとコンテキストの関係性である。



・密に解像度を変更して、ピラミッド画像を構成する。異なるスケールには異なる検出器を割り当てる。上の図に示すようにあらゆる画像変換を行い学習を実行する。小さな顔は通常学習画像に含まれていないが、積極的に生成する。



[34]

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia  
 "Pyramid Scene Parsing Network", in ILSVRC, 2016.

Keywords: Scene Parsing、Semantic Segmentation、ImageNet Scene Parsing Challenge 2016

## 概要

- ・シーンの解析で用いられるネットワークPSPNetを提案した。  
 2016のImageNet Scene Parsing Challenge 2016に優勝した。  
 右の図で示しているのはPSPNetのネットワーク構造です。

## 新規性・差分

・従来のFCN ( fully convolutional network) をベースとした手法は1、mismatched relationship ; 2、confusion categories ; 3、Inconspicuous Classesの三つの問題点がある。これらの問題の解決するために、部分的だけではなく、適切なグローバルシーンレベルの情報が必要となる。PSPNetはピラミッドプーリングモジュールによって、4つの異なるレベルの特徴を得ることができる。結果としては従来手法の問題点を有効に解決した。

・代表的なデータセットADE20K、PASCAL VOC 2012、  
 Cityscapesの三つを用いて実験を行って、三つのデータセットとも  
 提案モデルの優位性が証明された。

## Links

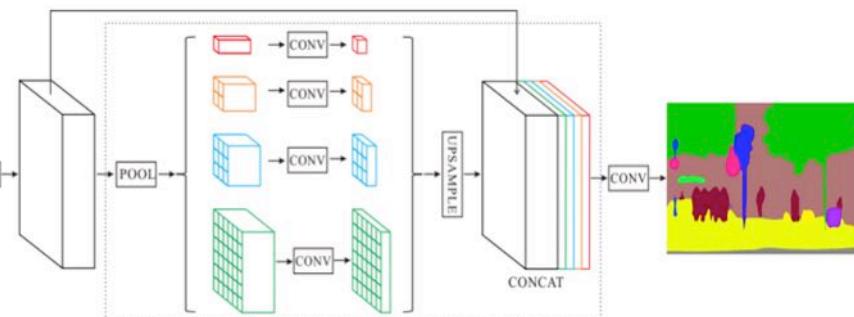
論文 <https://arxiv.org/pdf/1612.01105v1.pdf>

プロジェクト

<http://appsrv.cse.cuhk.edu.hk/~hszhao/projects/pspnet/index.html>



(a) Input Image



(b) Feature Map

(c) Pyramid Pooling Module

(d) Final Prediction

Rank	Team Name	Final Score (%)
1	Ours	57.21
2	Adelaide	56.74
3	360+MCG-ICT-CAS_SP - (our single model)	55.56 (55.38)
4	SegModel	54.65
5	CASIA_IVA	54.33
-	DilatedNet [40]	45.67
-	FCN [26]	44.80
-	SegNet [2]	40.79

# 【35】Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, "Semantic Understanding of Scenes through the ADE20K Dataset ", in, arXiv:1608.05442v1[cs.CV], 2016.

Keywords: Scene parsing ,ADE20K dataset,ADE20K benchmark,segn

## 概要

- 今のシーン解析で用いられるデータセットがカテゴリが少ない、解析できるシーンが制限されている、オブジェクトパートの情報がなしなどの問題点があります。そういう問題を踏まえてシーンの解析で用いられるデータセットADE20K（左下図：データセットのデータの例）を紹介した。そして、Cascadeというセグメンテーションのモジュールを提案した。Cascadeを用いたら、オブジェクトのパートまでの解析が可能になる。

## 新規性・差分

- ADE20Kデータセットは従来のデータセットのCOCOやPASCAL VOCなどと比べたら、平均的1枚の画像あたりオブジェクトクラス数が多い。異なるセグメンテーション手法で検証を行ったら、オブジェクトに対しての解釈の一貫性が高い（統合的82.4%程度）。
- シーンの解析はロングテール性（例えば道路、床などがよく出てくる、石鹼箱などがめったにない）、空間レイアウトには関係がある（壁に貼ってある絵が壁の一部分など）ため、Cascadeセグメンテーションモジュール（右上図：フレームワーク）を提案した。このモジュールを用いて、実験を行ったら、右下の図で示しているように、セグメンテーションの表現が向上してきた。

Table 2. Performance on the validation set of SceneParse150.

Networks	Pixel Acc.	Mean Acc.	Mean IoU	Weighted IoU
FCN-8s	71.32%	40.32%	0.2939	0.5733
SegNet	71.00%	31.14%	0.2164	0.5384
DilatedNet	73.55%	44.59%	0.3231	0.6014
Cascade-SegNet	71.83%	37.90%	0.2751	0.5805
Cascade-DilatedNet	<b>74.52%</b>	<b>45.38%</b>	<b>0.3490</b>	<b>0.6108</b>

Table 3. Performance of stuff and discrete object segmentation.

Networks	35 stuff		115 discrete objects	
	Mean Acc.	Mean IoU	Mean Acc.	Mean IoU
FCN-8s	46.74%	0.3344	38.36%	0.2816
SegNet	43.17%	0.3051	27.48%	0.1894
DilatedNet	49.03%	0.3729	43.24%	0.3080
Cascade-SegNet	40.46%	0.3245	37.12%	0.2600
Cascade-DilatedNet	<b>49.80%</b>	<b>0.3779</b>	<b>44.04%</b>	<b>0.3401</b>

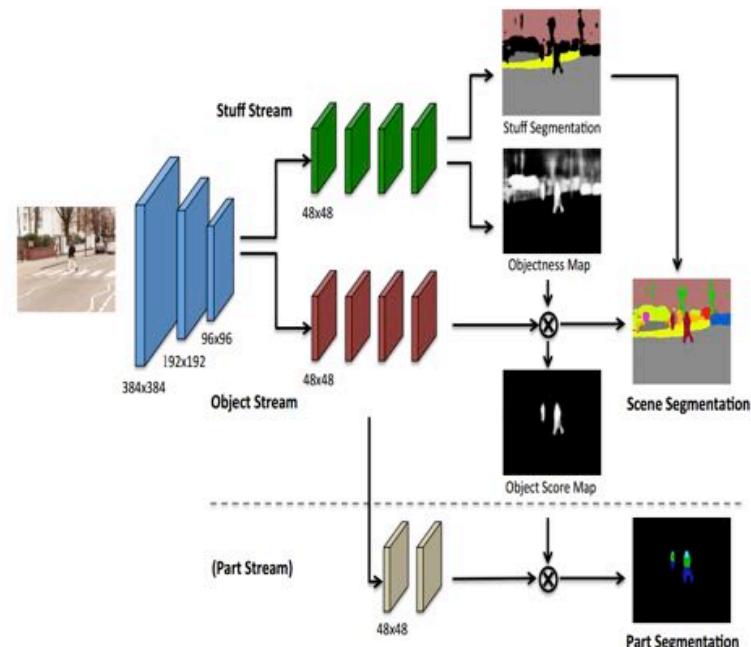


Figure 6. The framework of Cascade Segmentation Module for scene parsing. *Stuff stream* generates the stuff segmentation and objectness map from the shared feature activation. The *object stream* then generates object segmentation by integrating the objectness map from the *stuff stream*. Finally the full scene segmentation is generated by merging the object segmentation and the stuff segmentation. Similarly, *part stream* takes object score map from *object stream* to further generate object-part segmentation. Since not all objects have part annotation, the *part stream* is optional. Feature sizes are based on the Cascade-dilatedNet, the Cascade-SegNet has different but similar structures.



Figure 1. Images in ADE20K dataset are densely annotated in detail with objects and parts. The first row shows the sample images, the second row shows the annotation of objects, and the third row shows the annotation of object parts.

## Links

論文

<https://www.researchgate.net/publication>

[306357649\\_Semantic\\_Understanding\\_of\\_Scenes\\_through\\_the\\_ADE20K\\_Dataset](https://306357649_Semantic_Understanding_of_Scenes_through_the_ADE20K_Dataset)

[36]

# Samarth Brahmbhatt, James Hays, "DeepNav: Learning to Navigate Large Cities", in arXiv 1701.09135, 2017.

Keywords: StreetView, CNN, DeepNavigation

## 概要

- ストリートビューから画像を大量に採取して道案内のためのCNNを構築した。提案のデータセットには10都市から100万画像を超えるストリートビューの画像が含まれる。ナビゲーションの課題において3種類の教師あり学習を提案し、A\*-searchがどの程度学習の生成に有効かを検証した。データセットは完全にストリートビューのタグを使用した。CNNアーキテクチャに関してはDeepNav- $\{\text{distance}, \text{direction}, \text{pair}\}$ を提案。-distanceはVGG16の特徴から目的地までの距離を返却 (fc8が距離を算出)、-directionは同特徴から方向を返却 (fc8が各目的地・各方向のスコアを算出)、-pairはSiameseNetにより構成され、画像のペアから方向を算出。

## 新規性・差分

- 米国の10都市から100万枚のストリートビューの画像を収集してデータセットを構成した。本データセットには5種類の目的地 (Bank of America, Church, Gas Station, High School and McDonald's) を含む。

- 手法としては3つのCNNアーキテクチャを提案した。従来手法[Khosla+, CVPR14]と比較の結果、有効性を示した。

- A\*-searchにより経路生成を行い、これをGNNの教師と設定した

## Links

論文

<https://arxiv.org/pdf/1701.09135.pdf>

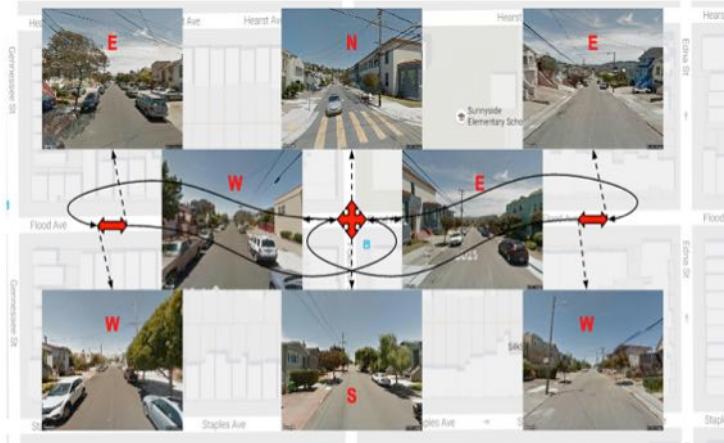
著者 <https://samarth-robo.github.io/>

City	Images	BofA	church	gas station	high school	McDonald's
Atlanta	78,808	10	32	32	7	7
Boston	105,000	40	40	39	20	20
Chicago	105,001	22	33	10	15	32
Dallas	105,000	7	25	35	9	13
Houston	117,297	8	19	30	4	14
Los Angeles	80,701	9	15	30	6	13
New York	105,148	30	20	21	27	31
Philadelphia	105,000	14	42	35	30	19
Phoenix	101,419	4	23	29	18	15
San Francisco	101,741	35	50	45	22	12
Total	1,005,115	179	299	306	158	176

DB中に含まれるデータ数と各統計

Method	Expected number of steps		
	$d_s=470\text{m}$	$d_s=690\text{m}$	$d_s=970\text{m}$
Random walk	733.99	854.8913	911.85
A*	18.73	27.3204	39.57
HOG+SVR [19]	588.66	705.31	791.93
DeepNav-distance	580.69	<b>684.22</b>	773.02
DeepNav-direction	626.28	697.26	780.53
DeepNav-pair	<b>553.39</b>	689.04	<b>766.32</b>

各アルゴリズムの平均ステップ数



サンフランシスコの経路（ノード）と目的地（ピン）

# 【37】Yevhen Kuznetsov, Jorg Stuckler, Bastian Leibe, "Semi-Supervised Deep Learning for Monocular Depth Map Prediction", in arXiv 1702.02706, 2017.

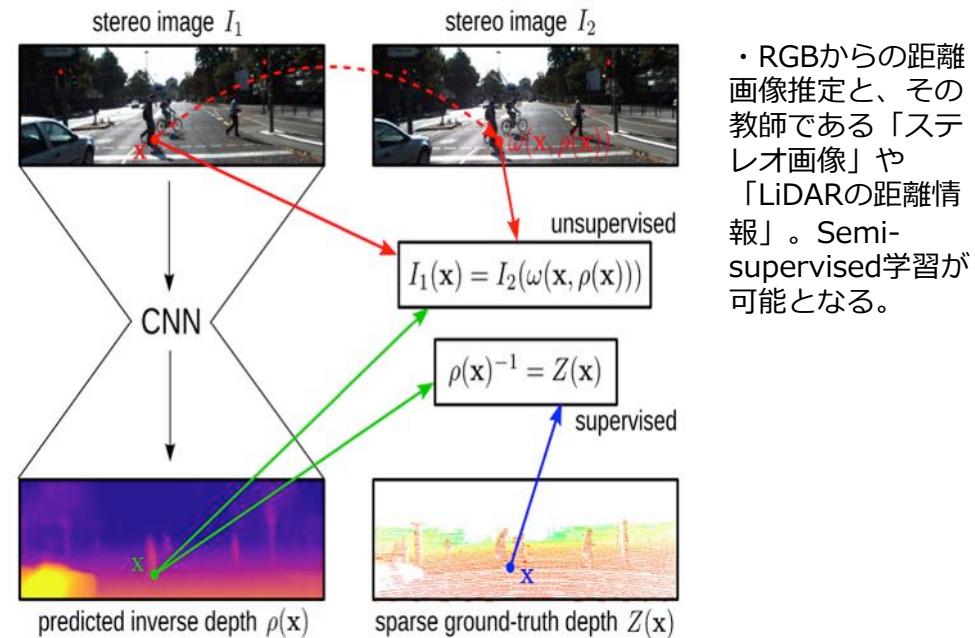
Keywords: Semi-supervised learning, depth prediction, ResNet

## 概要

- 半教師あり学習による車載画像のデプスマップ推定をDeep Learningにより行う。学習時にはステレオ画像やLiDARなどの3次元情報を抽出・教師としてRGB画像と対応づけることで、テスト時に高精度な距離画像を生成することができる。
- 教師あり/教師なしが混ざっていてもシームレスに学習可能な誤差関数を提案した。式(5)は誤差関数を示し、トレードオフパラメータは教師あり（第一項）、教師なし（第二項）、正規化項（第三項）に存在。 $\theta$ はCNNのパラメータであり、 $I_l$ ,  $I_r$ は左右ステレオ画像、 $Z_l$ ,  $Z_r$ は左右に対応するLiDARの3次元情報である。ネットワーク構造はResNet-50をベースとしてエンコーダとして用いるが、さらにUpsampling層も追加してデコーダとして動作する。

## 新規性・差分

- 教師あり・教師なし学習により距離画像を推定する枠組みを提案し、さらには誤差関数を提案した。
- ResNet-50をベースとしたエンコーダ・デコーダによるアーキテクチャにより距離画像推定においてもっとも良い値を出した。



$$\mathcal{L}_{\theta}(I_l, I_r, Z_l, Z_r) = \lambda_t \mathcal{L}_{\theta}^S(I_l, I_r, Z_l, Z_r) + \gamma \mathcal{L}_{\theta}^U(I_l, I_r) + \mathcal{L}_{\theta}^R(I_l, I_r), \quad (5)$$

## Links

論文 <https://arxiv.org/pdf/1702.02706.pdf>

- RGBからの距離画像推定と、その教師である「ステレオ画像」や「LiDARの距離情報」。Semi-supervised学習が可能となる。

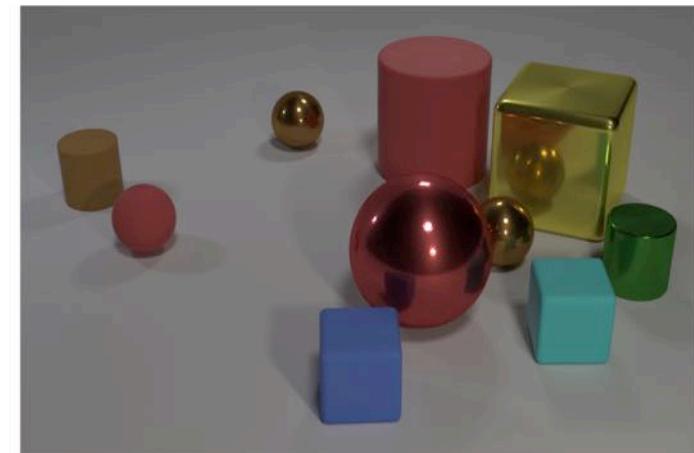
Approach	RMSE	RMSE (log)	lower is better			higher is better		
			$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Our full approach	<b>4.627</b>	<b>0.189</b>	<b>0.856</b>	<b>0.960</b>	<b>0.986</b>			
Our full approach*	4.679	0.192	0.854	0.959	0.985			
L2-norm instead of BerHu-norm in supervised loss	4.659	0.195	0.841	0.958	<b>0.986</b>			
No long skip connections	4.762	0.194	0.853	0.958	0.985			
No Gaussian smoothing in unsupervised loss	4.752	0.193	0.854	0.958	<b>0.986</b>			
Supervised training only	4.862	0.197	0.839	0.956	<b>0.986</b>			
Unsupervised training only (50 m cap)	6.930	0.330	0.745	0.903	0.952			
Only half of laser points used	4.808	0.192	0.852	0.958	<b>0.986</b>			

[38] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning", in arXiv 1612.06890, 2016.

Keywords: VQA

## 概要

- 画像による質問回答 (VQA; Visual Question Answering) のための新しい取り組みで、回答に対する診断(Reasoning)を詳細にできるようにし、さらにデータセットも提案する。データセットである CLEVR は多側面からの質問が用意されており、属性・カウント・比較・空間的な位置関係・論理的な操作がそれにあたる。質問は自然言語により構成され、画像解析により質問への回答を生成する。



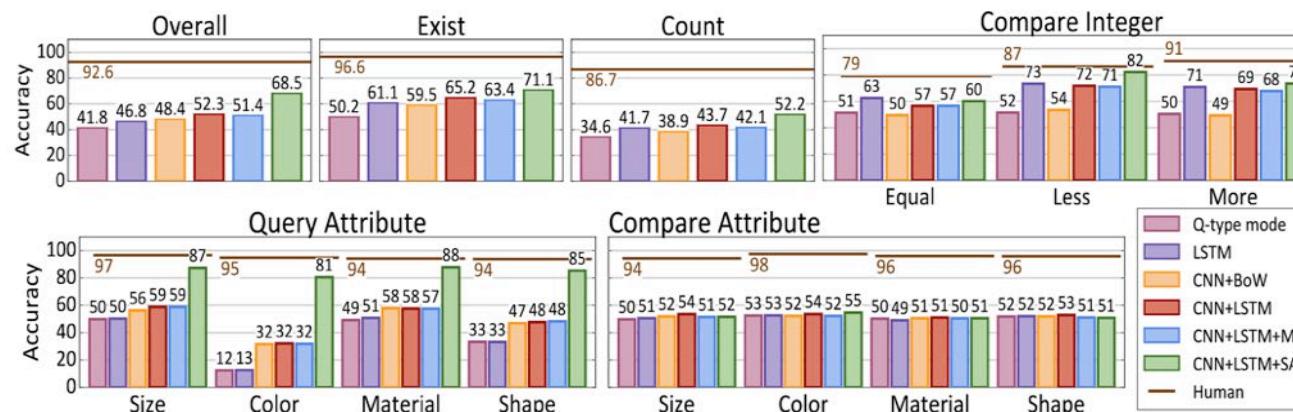
## 新規性・差分

- 従来のVQAでは画像やテキストの学習を行い、画像の意味を理解していないまま回答していたが、本論文のデータセットでは詳細な理由付けまで行えるようにした

- この問題を設定するために、Compositional Language and Elementary Visual Reasoning diagnostics dataset (CLEVR)を提供。同データは100Kの画像や1M(うち853Kはオリジナル)の自動生成された質問文を含んでいる。

## Links

論文 <https://arxiv.org/pdf/1612.06890v1.pdf>  
プロジェクト (データセットあり)



← 精度比較のグラフでは、LSTMやCNN+BoW, CNN+LSTM (+MCB, +SA)による比較を行っている。

(39)

Roy Jevnisek, Shai Avidan, "Co-Occurrence Filter", in CVPR, 2017.

Keywords: Co-Occurrence Filter, Bilateral Filter

## 概要

- バイラテラルフィルタをベースとしているが、ガウシアンによる重み付けの代わりに共起行列（Co-occurrence Matrix）を用いることでエッジを保存しつつ物体境界を強調する。二つのパッチのヒストグラムにより共起行列を計算する（式6）。 $C(a, b)$ は二つのヒストグラムのco-occ. valueのカウントである。

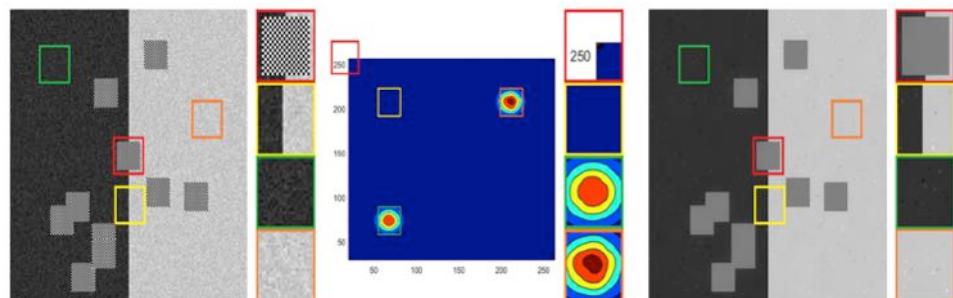
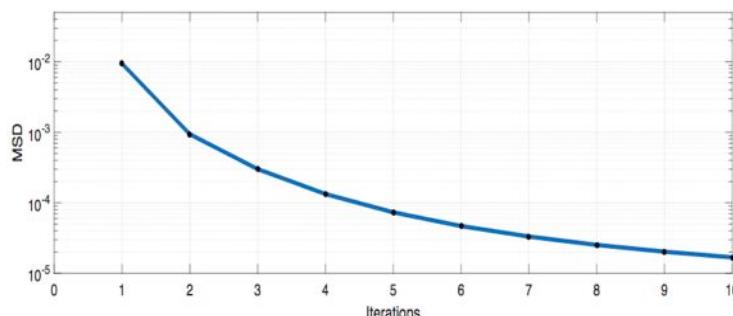
## 新規性・差分

$$M(a, b) = \frac{C(a, b)}{h(a)h(b)}. \quad (6)$$

- バイラテラルフィルタの性質であったエッジ部分を保存してノイズを除去する効果に止まらず、共起性に着目した弱いエッジなどに着目した境界部分も抽出することに成功した。
- 評価はIterationごとのMean-Squared-Difference (MSD)

## Links

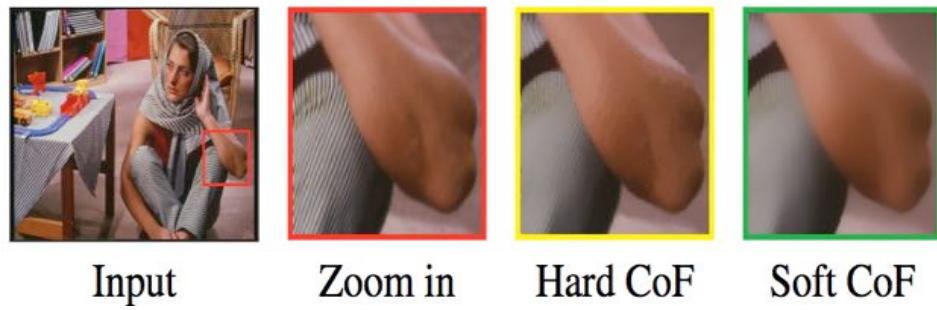
論文 <https://arxiv.org/pdf/1703.04111.pdf>



- CoFの例。左側が入力で、右側が出力結果をグレースケールで示したもの。エッジのみでなく、テクスチャの共起関係により重み付けして出力。

$$J_p = \frac{\sum_{q \in N(p)} G(p, q) M_T(T_p, T_q) \cdot I_p}{\sum_{q \in N(p)} G(p, q) M_T(T_p, T_q)} \quad (11)$$

- CoFをカラー画像に適用すると $256^3 \times 256^3$ の行列になり、実利用上、非常に大きな行列になってしまふ。そこで、(11)のTをk-meansにより $k \times k$ の行列としてピクセル値の階層を削減した。



← 繰り返すごとにテクスチャが滑らかに、エッジが保存されていることがわかる

# 【40】Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, Ling Shao, "Deep Sketch Hashing: Fast Free-hand Sketch-based Image Retrieval", in CVPR, 2017. (spotlight oral)

Keywords: Sketch-based Image Retrieval (SBIR), Deep Sketch Hashing (DSH)

## 概要

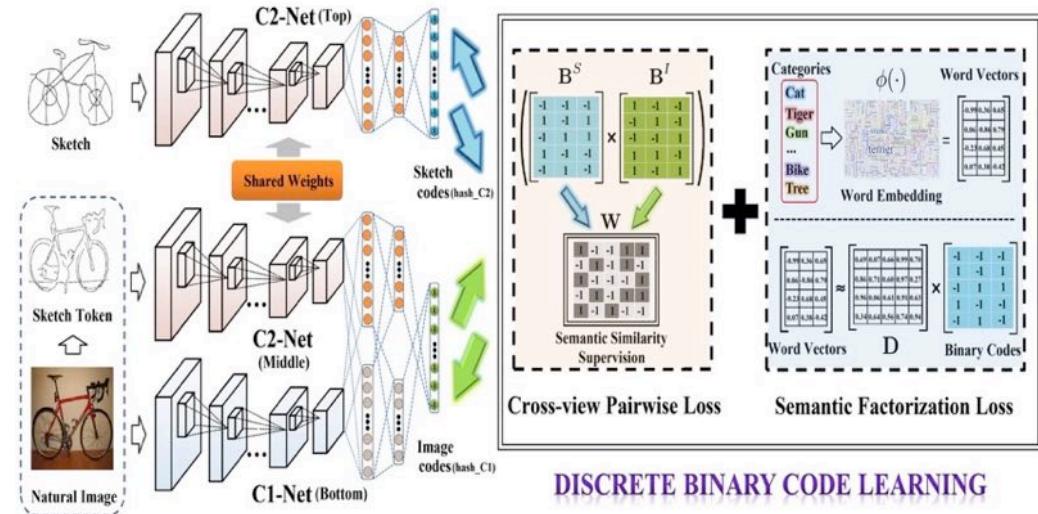
- スケッチによる画像検索をバイナリハッシング、特に Deep Sketch Hashing (DSH)により高速化する。自然画像とそのSketch-Token（エッジ部分を取り出したもの）、手書きスケッチとの重みを共有して学習することで、手書き—Sketch-Token—自然画像とドメインを共有することができる。DNNの出力にはHash値が割り当てられ、その後Pairwise LossやFactorization Lossを計算してスケッチから画像を検索できるようにする。

## 新規性・差分

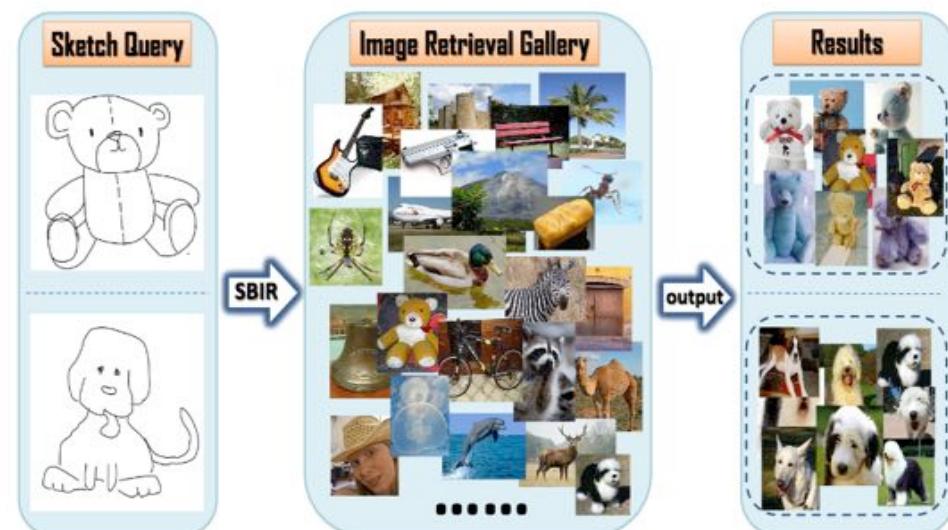
- 初めてカテゴリレベル、End-to-Endの深層学習を実現
- TU-BerlinやSketchyデータセットにおいてstate-of-the-artな性能を実現

## Links

論文 <https://arxiv.org/pdf/1703.05605.pdf>  
GitHub <https://github.com/ymcidence/DeepSketchHashing>



DISCRETE BINARY CODE LEARNING



# 【41】Limin Wang, Yuanjun Xiong, Dahua Lin, Luc Van Gool, "UntrimmedNets for Weakly Supervised Action Recognition and Detection", in arXiv 1703.03329, 2017. (accepted by CVPR2017)

Keywords: Weakly Supervised Learning, CNN, Action Recognition, Action Detection

## 概要

・時間的なアノテーションが与えられていない状態で行動認識モデルを学習するWeakly Supervised LearningをするためのUntrimmedNetsを提案した。Untrimmed Videoから複数のClipを抽出し、Clipごとに特徴抽出する。ClipごとにClassification Moduleで各クラスの識別スコアを算出した後、Selection Moduleで識別スコアに基いて行動インスタンスを含むClipを選択することで映像全体のラベルを決定する。Strong Supervisedな手法と比較しても高い精度を達成した。

## 新規性・差分

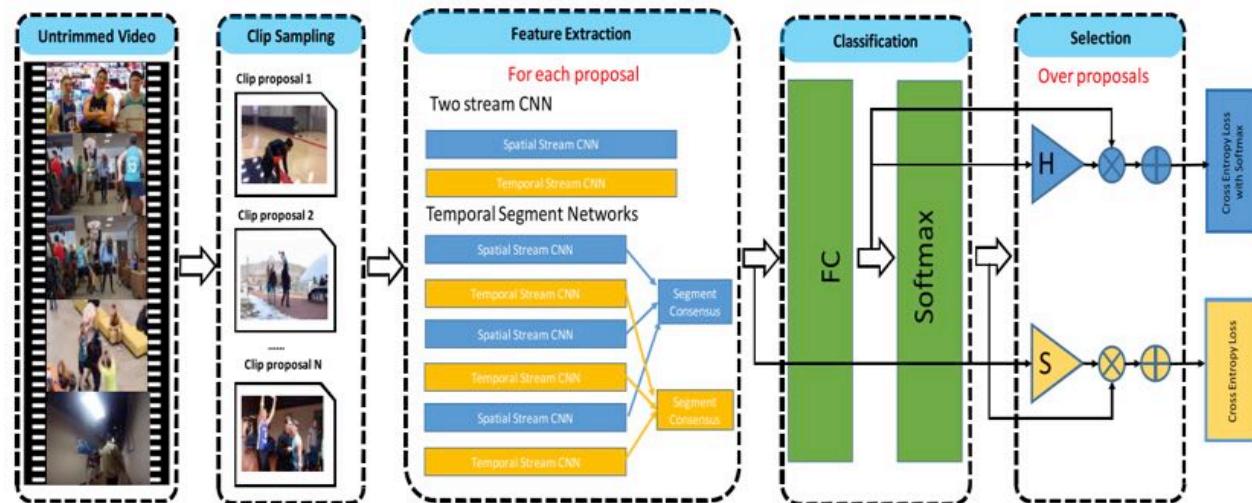
・時間情報に関するラベルを一切用いないWeakly SupervisedなAction Recognitionという問題設定を提案。そして、それをUntrimmedNetsによりEnd-to-Endな学習を可能にしている点に新規性がある。Weakly Supervisedだと精度は劣るケースが多いが、この論文では高い精度も実現している点でインパクトが大きい。

## Links

論文 <https://arxiv.org/pdf/1703.03329.pdf>

THUMOS14	ActivityNet	←認識精度	
iDT+FV [43]	63.1%	iDT+FV [43]	66.5%*
Two Stream [38]	66.1%	Two Stream [38]	71.9%*
EMV+RGB [54]	61.5%	C3D [40]	74.1%*
Objects+Motion [17]	71.6%	Depth2Action [55]	78.1%*
TSN [48]	78.5%	TSN [48]	88.8%*
UntrimmedNet (hard)	81.2%	UntrimmedNet (hard)	91.3%
UntrimmedNet (soft)	<b>82.2%</b>	UntrimmedNet (soft)	90.9%

	IoU = 0.5	IoU = 0.4	IoU = 0.3	IoU = 0.2	IoU = 0.1
Oneata <i>et al.</i> [31]*	14.4	20.8	27.0	33.6	36.6
Richard <i>et al.</i> [33]*	15.2	23.2	30.0	35.7	39.7
Shou <i>et al.</i> [37]*	19.0	28.7	36.3	43.5	47.7
Yeung <i>et al.</i> [52]*	17.1	26.4	36.0	44.0	48.9
Yuan <i>et al.</i> [53]*	18.8	26.1	33.6	42.6	51.4
UntrimmedNet (soft)	13.7	21.1	28.2	37.7	44.4

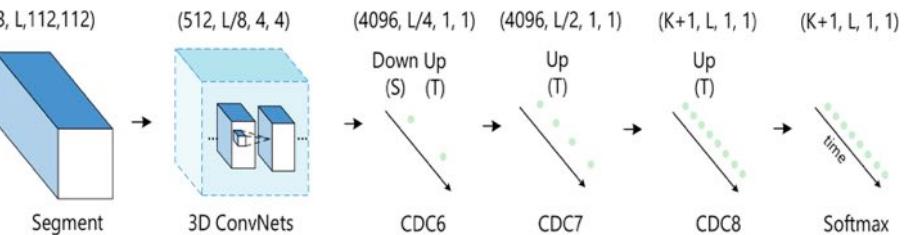


[42] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, Shih-Fu Chang,  
 “CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos”, in arXiv 1703.01515, 2017.

Keywords: Temporal Action Localization, CNN

## 概要

・従来のTemporal Action Localizationではセグメント単位で行動を識別することでLocalizationをしていた。それに対してこの研究ではframeごとに各クラスのスコアを求めてLocalizationする手法を提案する。そのために、Spatial ConvとTemporal Deconvを同時に行うCDCフィルタを導入する。C3DベースのCNNで、C3DでFC層があるところにCDCを入れる。Deconvにより出力のフレーム数は入力映像と同じになるためフレームごとのスコアが推定可能となる。これにより高精細なTemporal Localizationを実現した。



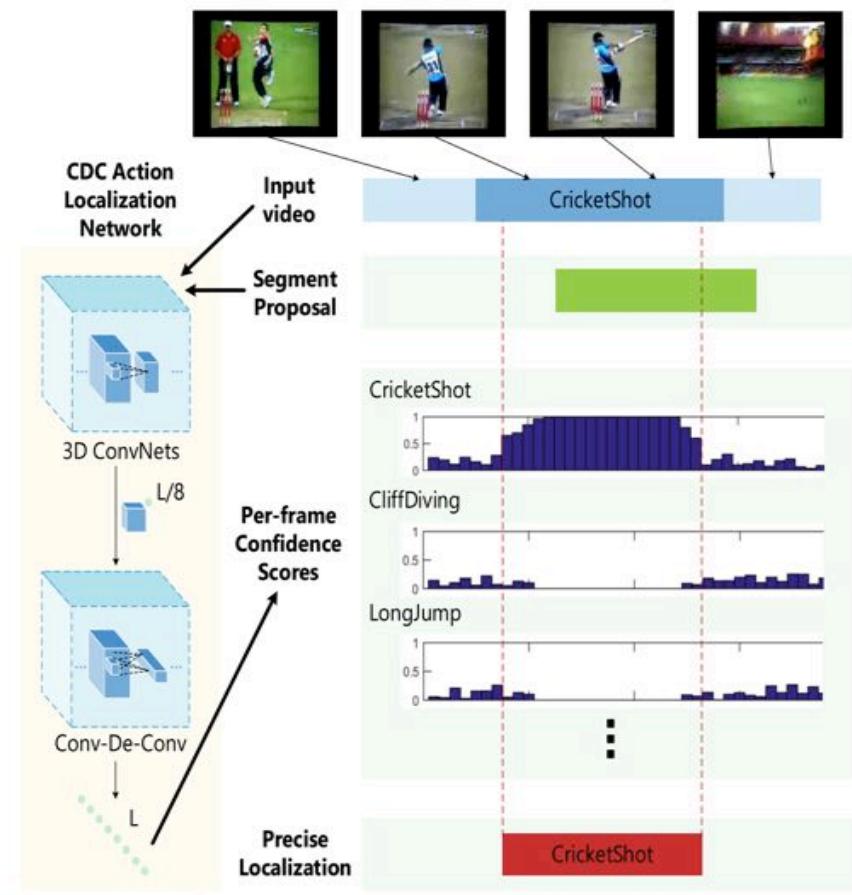
## 新規性・差分

- ・Deconvによりフレームごとのスコアを推定するネットワーク構造を新規に提案
- ・Spatial ConvとTemporal Deconvを同時に行うCDCを新規に提案  
(Spatial Conv → Temporal Deconvとするよりも精度が向上している)

## Links

論文 <https://arxiv.org/abs/1703.01515v1>

IoU threshold	0.3	0.4	0.5	0.6	0.7
Karaman <i>et al.</i> [25]	0.5	0.3	0.2	0.2	0.1
Wang <i>et al.</i> [64]	14.6	12.1	8.5	4.7	1.5
Heilbron <i>et al.</i> [17]	-	-	13.5	-	-
Escorcia <i>et al.</i> [9]	-	-	13.9	-	-
Oneata <i>et al.</i> [38]	28.8	21.8	15.0	8.5	3.2
Richard and Gall [42]	30.0	23.2	15.2	-	-
Yeung <i>et al.</i> [72]	-	-	17.1	-	-
Yuan <i>et al.</i> [75]	33.6	26.1	18.8	-	-
S-CNN [46]	36.3	28.7	19.0	10.3	5.3
C3D + LinearInterp	36.0	26.4	19.6	11.1	6.6
Conv & De-conv	38.6	28.2	22.4	12.0	7.5
CDC (fix 3D ConvNets)	36.9	26.2	20.4	11.3	6.8
<b>CDC</b>	<b>40.1</b>	<b>29.4</b>	<b>23.3</b>	<b>13.1</b>	<b>7.9</b>



(43)

Hang Yan, Yebin Liu, Yasutaka Furukawa, "Turning an Urban Scene Video into a Cinemagraph", in CVPR, 2017.

Keywords: Cinemagraph

## 概要

- 自動車など並進運動があるビデオから（マスクにより）ある部分のみが変化するビデオ（これをCinemagraphと呼ぶ）を自動生成するための技術を提案する。アプリケーションとしては例えば、Google Street Viewなどで特定の箇所だけ動きをつけたいときに用いる。入力動画からはStructure from Motion (SfM)、Multi-View Stereo、Image Warpingにより再レンダリングを行いWarped Videoを生成これによりエゴモーションではない動領域を抽出（マスクを抽出）できる。結果的にマスク領域のみが動いている動画像を生成する。

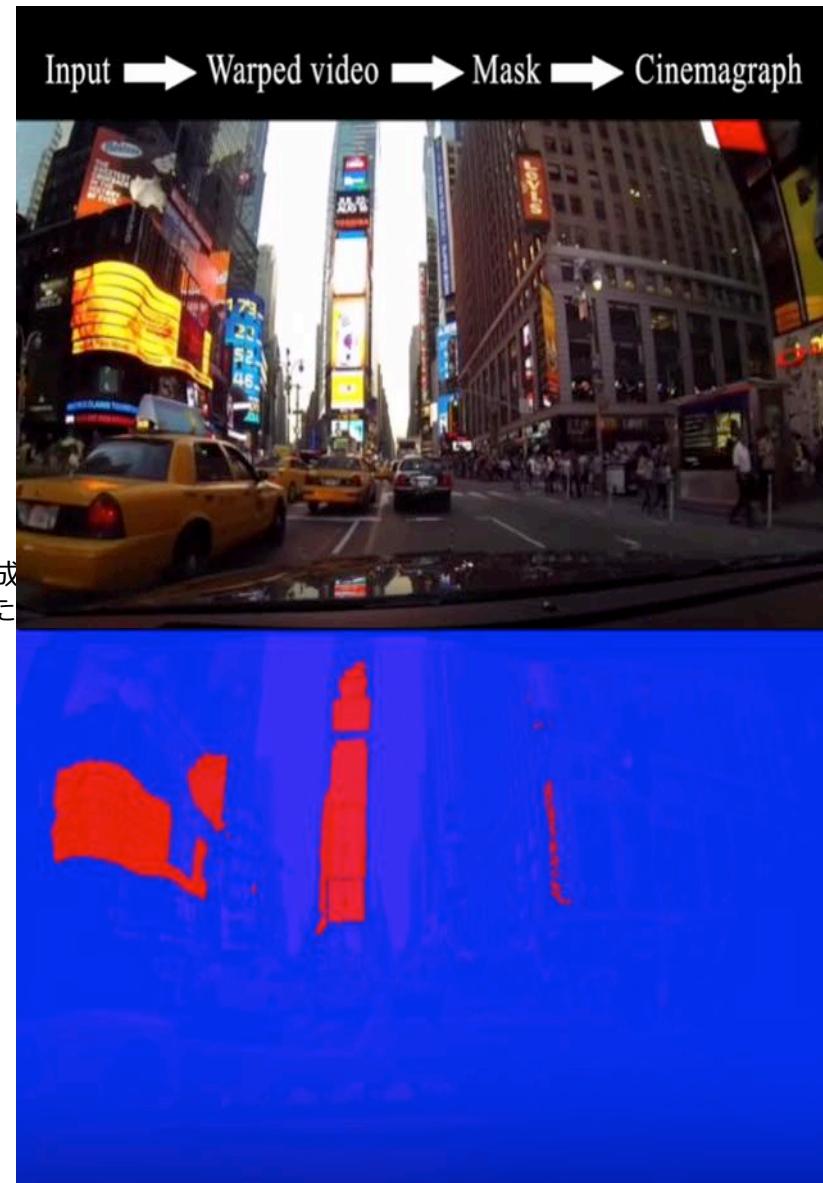
## 新規性・差分

- 自動車などエゴモーション(並進運動) がある動画像からCinemagraphを生成
- Video Stabilization, 空間的・時間的な正規化を行い、より鮮明な映像とした
- 詳細にはビデオを参照

## Links

論文 <https://arxiv.org/abs/1612.01235>

ビデオ <https://www.youtube.com/watch?v=r3yyL6qrVX4>



# 【44】Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, Jitendra Malik, "Cognitive Mapping and Planning for Visual Navigation", in CVPR, 2017.

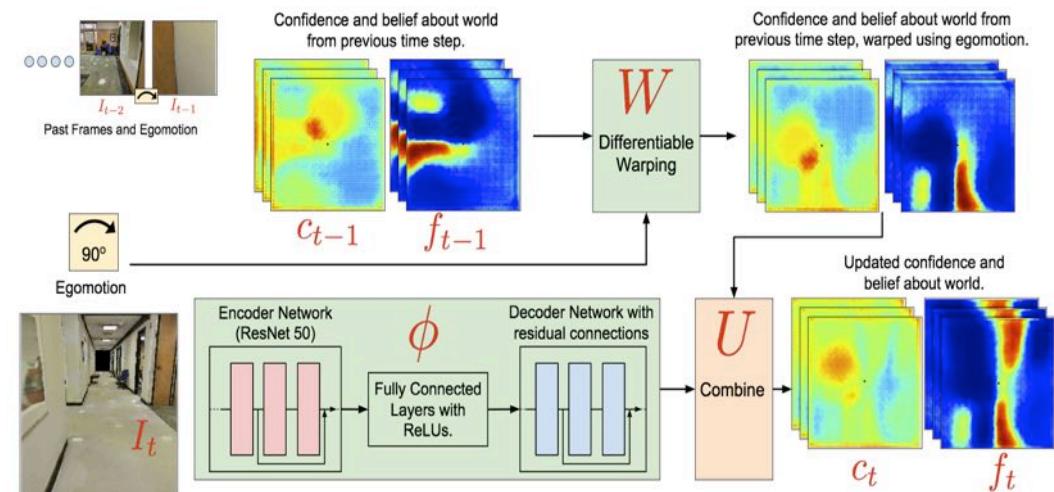
Keywords: Mapping, Navigation

## 概要

・ロボット視点からの画像から、マッピングとナビゲーションを同時にを行い、ゴールまでの経路を推測する研究である。このCognitive Mapping and Planner (CMP)は(1)ナビゲーションのためのマッピングとプランニングを同時に行うモデルであり、(2)空間メモリを用いることで、完成されていない部分的な入力からゴールを常時推測

## 新規性・差分

・画像の入力からEnd-to-Endでロボットのアクション(e.g. 直進、左右のターン)を推定するモデルを考案した  
・RGBとDepthの入力によるナビゲーションを比較したところ、Depthの方が良好な結果であり、成功率を計算したところ、78%であった



Method	Mean		75 <sup>th</sup> %ile		Success %age	
	RGB	Depth	RGB	Depth	RGB	Depth
<b>Geometric Task</b>						
Initial	25.3	25.3	30	30	0.7	0.7
No Image LSTM	20.8	20.8	28	28	6.2	6.2
Reactive (1 frame)	20.9	17.0	28	26	8.2	21.9
Reactive (4 frames)	14.4	8.8	25	18	31.4	56.9
LSTM	10.3	5.9	21	5	53.0	71.8
Our (CMP)	<b>7.7</b>	<b>4.8</b>	<b>14</b>	<b>1</b>	<b>62.5</b>	<b>78.3</b>
<b>Semantic Task (Aggregate)</b>						
Initial	16.2	16.2	25	25	11.3	11.3
Reactive	14.2	14.2	22	23	23.4	22.3
LSTM	13.5	13.4	20	23	23.5	27.2
Our (CMP)	<b>11.3</b>	<b>11.0</b>	<b>18</b>	<b>19</b>	<b>34.2</b>	<b>40.0</b>

↑  $W$ は直前のエゴモーション $e_t$ による前ステップの状態 $f_{t-1}$ から計算した変換行列、 $U$ は前のすべての状態の蓄積である

## Links

論文 <https://arxiv.org/pdf/1702.03920.pdf>

プロジェクト

<https://sites.google.com/view/cognitive-mapping-and-planning/>

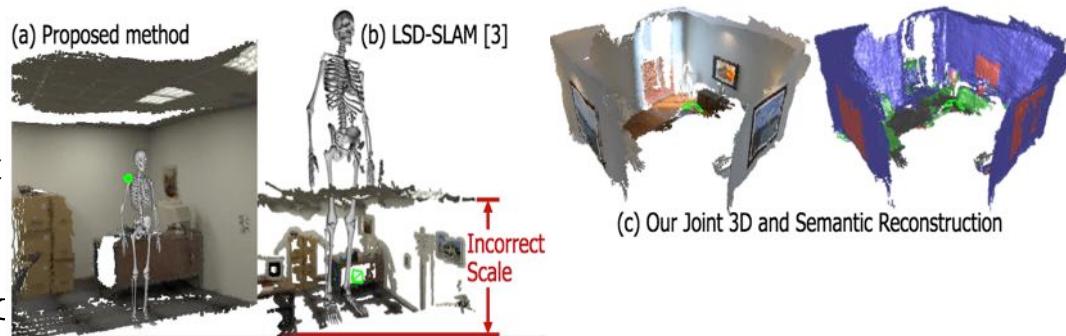
動画 <https://www.youtube.com/watch?v=BNmz3xBtcJ8>

# 【45】Keisuke Tateno, Federico Tombari, Iro Laina, Nassir Navab, "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction", in CVPR, 2017.

Keywords: CNN-SLAM, Depth Prediction

## 概要

SLAMにおいて、CNNによる距離画像推定とダイレクトに推定したMono-SLAM（単眼によるビジュアルSLAM）の結果を統合していくかに高精度化するかを検討した論文。 Mono-SLAMにおける問題点であったスケールの問題を、 CNNによる距離画像推定を用いて正規化する実験についても行う。最終的にはセマンティックラベルも用いることで、さらなる高精度化を図る。



## 新規性・差分

- ・ビジュアルSLAMとCNNにより推定した距離画像を統合することで、正規化されたスケールの点群データを出力とする（正規化された様子は動画を参照）
- ・セマンティックラベルを用いることで、SLAMがさらに高精度化することを示した

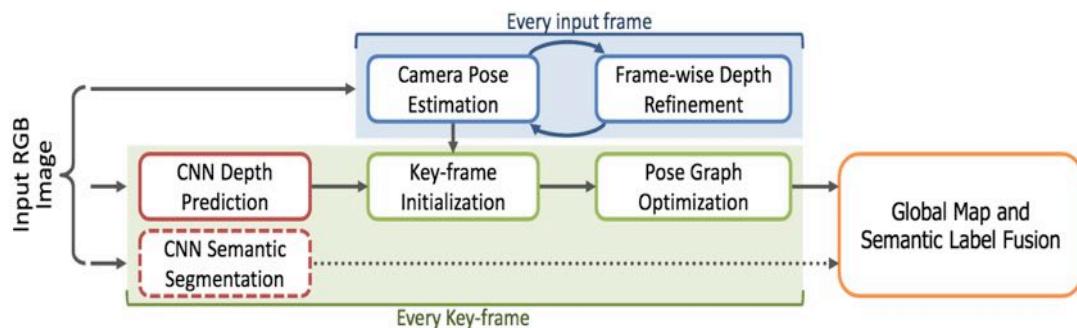


Figure 2. CNN-SLAM Overview.



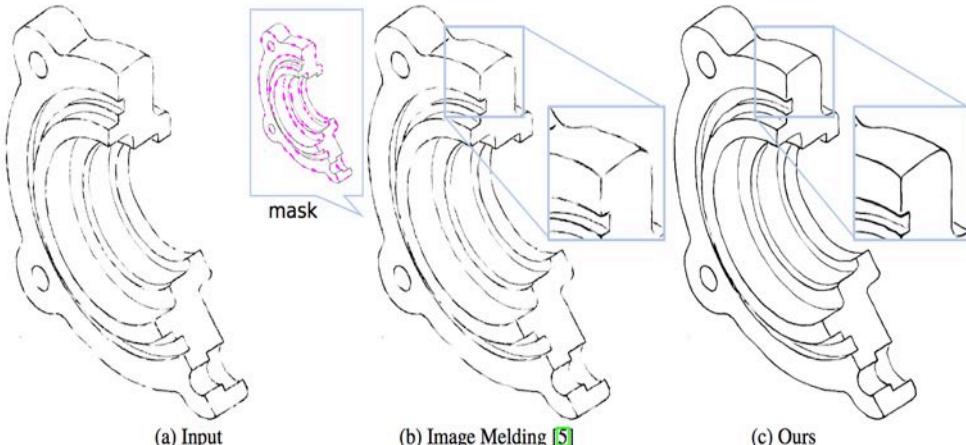
上図：提案のCNN-SLAMのフレームワーク  
左図：LSD-SLAMと提案手法の比較

# 【46】Kazuma Sasaki, Satoshi Iizuka, Edgar Simo-Serra, Hiroshi Ishikawa, "Joint Gap Detection and Inpainting of Line Drawing", in CVPR, 2017.

Keywords: Completing Line Drawing

## 概要

Convolutional Neural Networks (CNN)を用いた線画の自動補完（イラストに対するインペインティング）に関する研究。ここでは線画の太さや曲線までカバーするモデルを提供し、（ユーザとのインタラクションではなく）全てデータセットからの学習により補完を行う。CNNの構造はFCN [Long+, CVPR15]等を参考にして全結合層を含まない、Conv層やUpsampling層を含むアーキテクチャに設計されている。データセットについては深層学習用に何百万というペア画像を作成することは困難であるため、少量のデータから自動生成を試みた。



## 新規性・差分

従来では不足部分に関する学習が必要であったが、本論文ではCNNが自動で判断して補完できるように改善した（おそらく初めて完全自動で欠損を補完した）

## Links

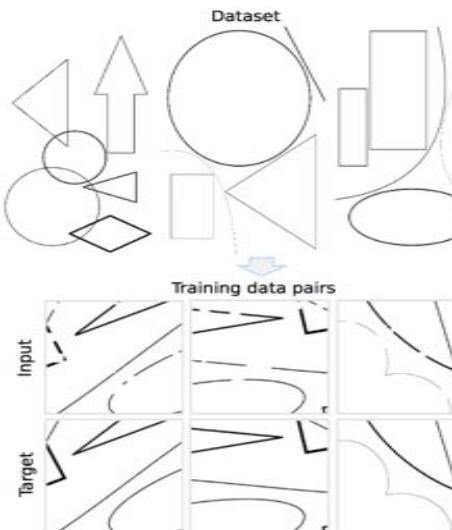
論文

[http://hi.cs.waseda.ac.jp/~iizuka/projects/inpainting/data/inpainting\\_cvpr2017.pdf](http://hi.cs.waseda.ac.jp/~iizuka/projects/inpainting/data/inpainting_cvpr2017.pdf)

プロジェクト <http://hi.cs.waseda.ac.jp/~iizuka/projects/inpainting/ja/>

vs	[1]	[5]	Ours	Ours (Masked)
[1]	-	26.4	6.1	1.4
[5]	73.6	-	4.2	5.3
Ours	93.9	95.8	-	53.9
Ours (Masked)	98.6	94.7	46.1	-

・従来法との比較。[1]PatchMatch[5]Image Meldingを引用



・データセットは少量（60枚）の画像から自動生成。クロップアウトするパッチをランダム選択、さらに回転・リバース・ダウンスケール・欠損作成もランダムで実行した。

# [47] Joao Carreira, Andrew Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset", in CVPR, 2017.

Keywords: Kinetics Human Dataset

## 概要

Kinetics Human Datasetを用いた行動認識の研究。同時にRGBやOpticalFlowの2つのモダリティを入力とするTwo-Stream C3Dを提案して、xytの3Dカーネルの学習もより高いレベルで実現させている。  
2Dカーネルから3Dカーネルへの膨張（Inflated Inception-v1）、Two-Streamへの拡張（e, Two-stream 3D-convnet）は右図に示されている。

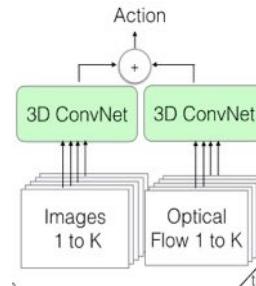
## 新規性・差分

- 3DConvolutionの学習を成功させるためにImageNetの2Dカーネルのパラメータを適用、Two-StreamCNNのモデルを採用した
- Kinetics Datasetを用いた学習済みモデルは転移学習にも有効であることが判明した

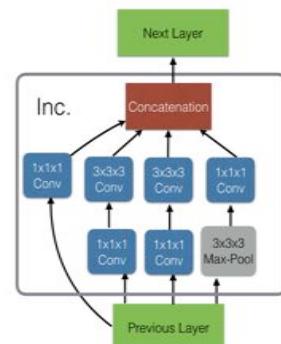
## Links

論文 <https://arxiv.org/pdf/1705.07750.pdf>

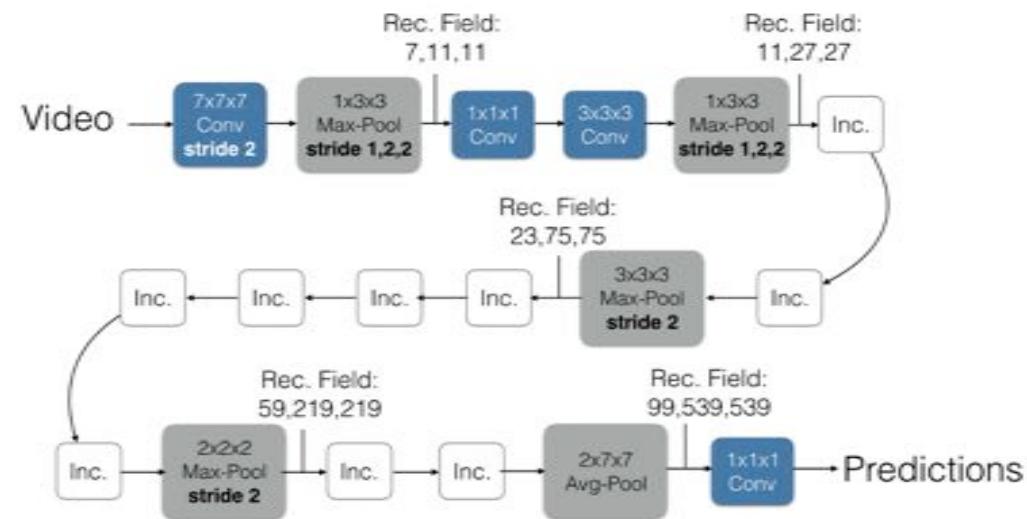
e) Two-Stream  
3D-ConvNet



### Inception Module (Inc.)



### Inflated Inception-V1



Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	—	—	36.0	—	—	69.9	—	—
(b) 3D-ConvNet	51.6	—	—	24.3	—	—	60.0	—	—
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	<b>84.5</b>	<b>90.6</b>	<b>93.4</b>	<b>49.8</b>	<b>61.9</b>	<b>66.4</b>	<b>74.1</b>	<b>69.6</b>	<b>78.7</b>

Model	UCF-101	HMDB-51
RGB-I3D, miniKinetics pre-training	91.8	66.4
RGB-I3D, Kinetics pre-training	95.6	74.8
Flow-I3D, miniKinetics pre-training	94.7	72.4
Flow-I3D, Kinetics pre-training	96.7	77.1
Two-Stream I3D, miniKinetics pre-training	96.9	76.3
Two-Stream I3D, Kinetics pre-training	<b>98.0</b>	<b>80.7</b>

【48】

Johann, Abhilash, "Weakly Supervised Affordance Detection", in CVPR, 2017.

Keywords: weakly supervised, object affordance

## 概要

- object affordance推定のデータセットと弱教師あり学習による推定手法の提案
- 周囲とのコンテキストを考慮したaffordanceの教師を得るためにCAD120データセットの一部の画像にピクセルごとのアノテーションをし、新たなデータセットを提案
- そのデータセットにおいてキーポイント(画像内のある1ピクセル)のみのアノテーションを用いて画像全体の1ピクセルごとのマルチクラスなアフォーダンス推定を行う

## 新規性・差分

- CAD120における9916の物体についてアフォーダンスを付与したデータセットの提案
- 最初にキーポイントによって学習したCNNの訓練データに対する出力にGrab-cutを施してよりrefineし、それを訓練データとして再度用いて学習を行う
- F値, IoUの評価によって、Grab-cutを用いたデータによる再学習を行った方が良い結果となった。

## Links

論文

[http://pages.iai.uni-bonn.de/qall\\_juergen/download/jgall\\_affordancedetection\\_cvpr17.pdf](http://pages.iai.uni-bonn.de/qall_juergen/download/jgall_affordancedetection_cvpr17.pdf)

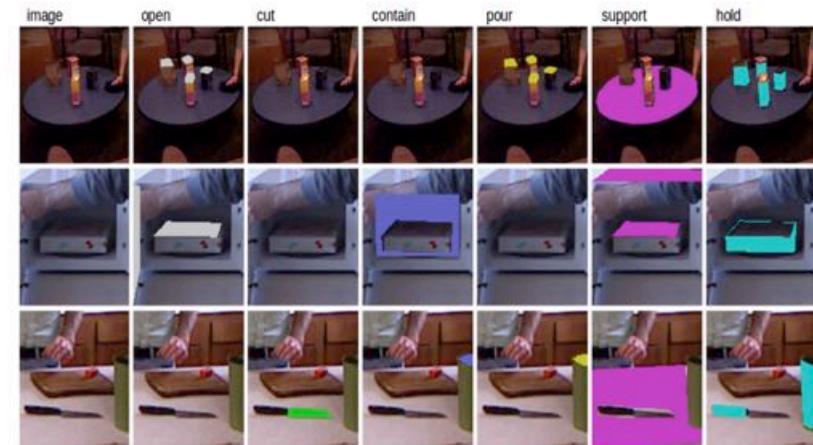
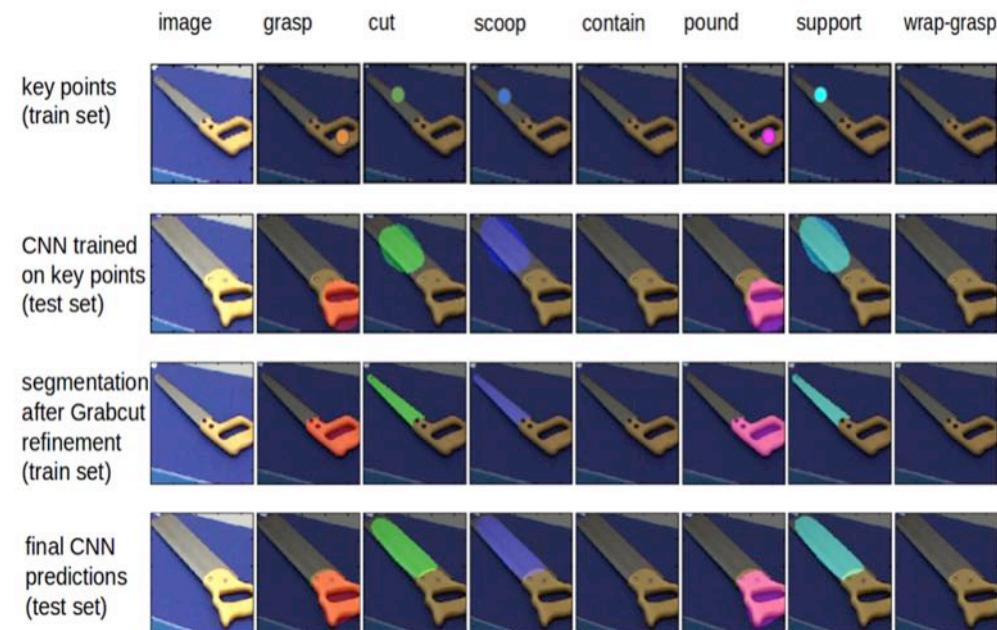


Figure 2. Example images with annotations from the proposed CAD120 affordance dataset. Pixels that do not belong to any affordance are considered as background. Best viewed in color.



[49]

# Weifeng Ge Yizhou Yu , "Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-Tuning ", in CVPR, 2017.

Keywords: CNN, Fine-tuning, Multi-Task Learning,

## 概要

- より効率的に画像の機械学習をするために、多くのオリジナルソースの画像から「低水準の」特徴を抽出しターゲットなる画像を選別する。低水準の特徴とは、ソース画像の細かな情報は学習させず、色合いなどをベースにターゲット画像に近いものだけを選別する。今までの傾向とは逆に低水準の情報を学習させることにより、効率よくfine-tuningができる、さらに少ないラベル設定を可能にできた。

## 新規性・差分

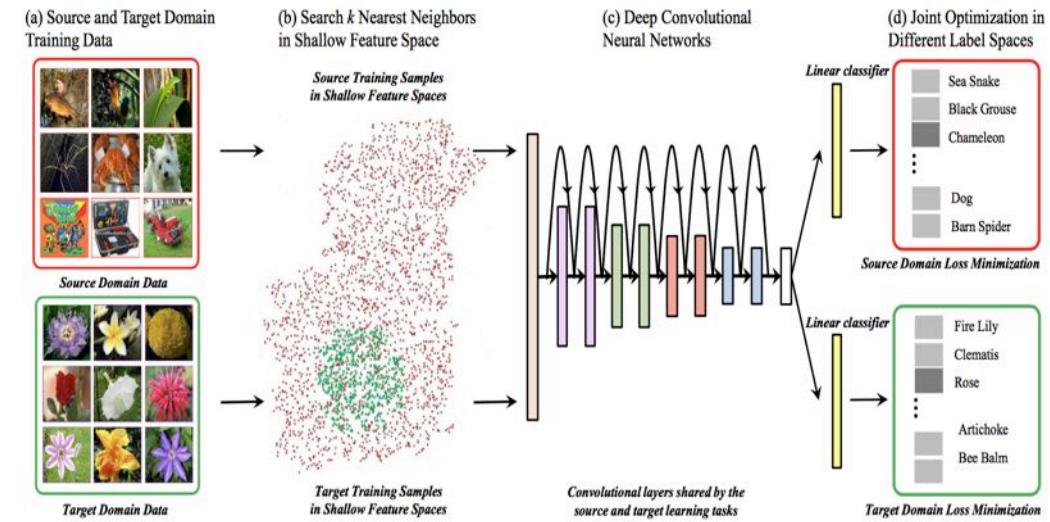
- 従来とは異なり、低水準での特徴学習によるDeep Convolutional Neural Networksで効率が高く高精度なselective joint fine-tuning を提案
- 特定のターゲット学習にどのようにしてもっとも正確なソースを見つけるかが今後の課題（画像学習のための画像学習？）

## Links

論文 <https://arxiv.org/pdf/1702.08690.pdf>

プロジェクト

GitHub <https://github.com/ZYYSzj>Selective-Joint-Fine-tuning>



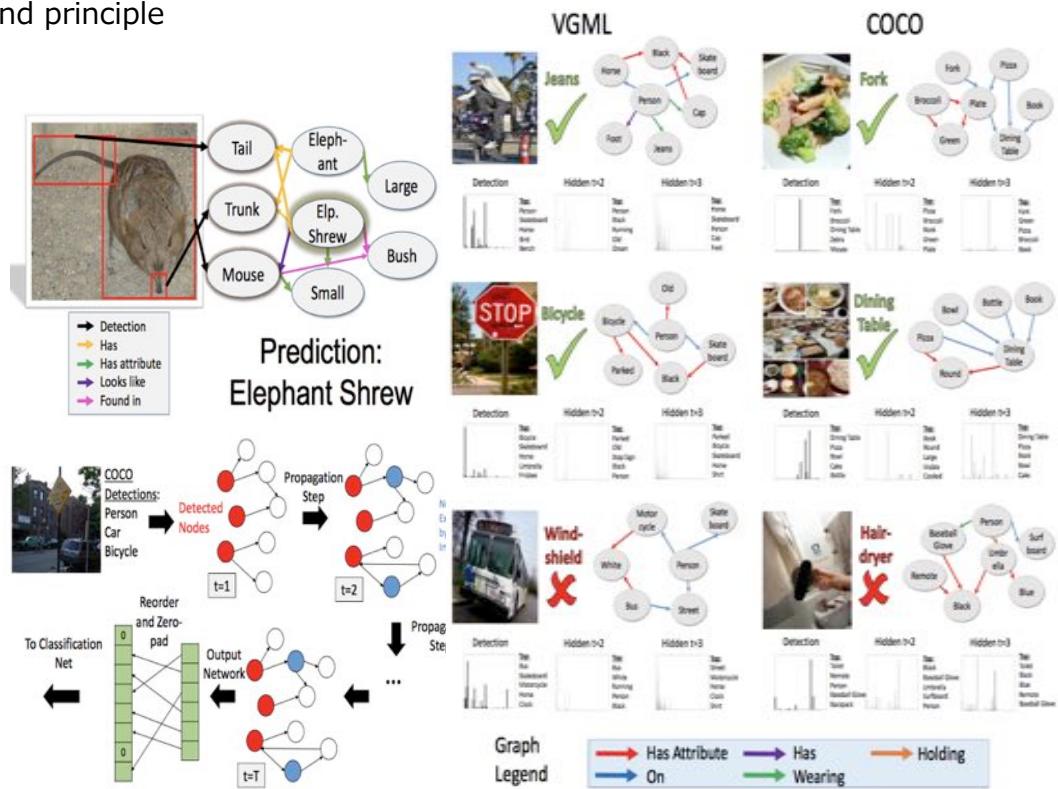
Method	mean Acc(%)
MPP [47]	91.3
Multi-model Feature Concat [1]	91.3
MagNet [38]	91.4
VGG-19 + GoogleNet + AlexNet [20]	94.5
Training from scratch using target domain only	58.2
Selective joint training from scratch	80.6
Fine-tuning w/o source domain	92.3
Joint fine-tuning with all source samples	93.4
Selective joint FT with random source samples	93.2
Selective joint FT w/o iterative NN retrieval	94.2
Selective joint FT with Gabor filter bank	93.8
Selective joint fine-tuning	94.7
Selective joint FT with model fusion	95.8
VGG-19 + Part Constellation Model [38]	95.3
Selective joint FT with val set	97.0

# (50) Kenneth Marino, Ruslan Salakhutdinov, Abhinav Gupta , "The More You Know: Using Knowledge Graphs for Image Classification ", in CVPR, 2017.

Keywords: Graph Network, Nural Network, End-to-end principle

## 概要

- 人間の認識方法である連想的ネットワークをコンピューター画像認識に適用。このときの処理をEnd-to-endにすることで簡潔化。この時に使ったデータでknowledge bankを作り、物と物、事象と事象の関連性を与え、グラフの繋がり方の傾向を把握。

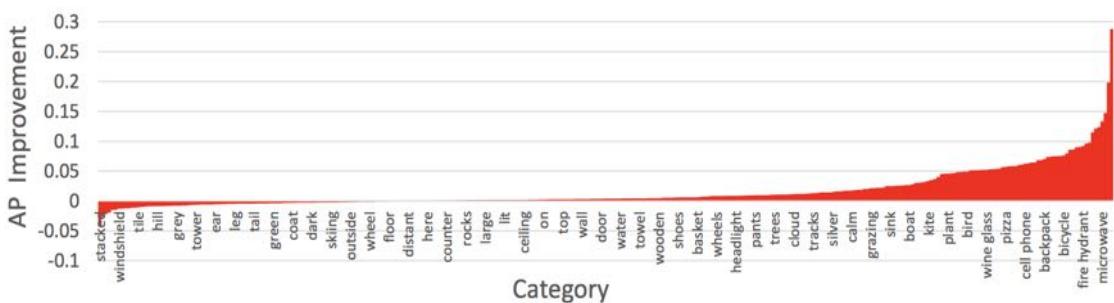


## 新規性・差分

- グラフ的なネットワークとニューラルネットワークを組み合わせたGraph Search Neural Network (GSNN)を提案
- GSNNは汎用性が高く様々なシーンでの活躍が期待できる。

## Links

論文 <https://arxiv.org/pdf/1612.04844.pdf>



# 【51】Pichao Wang, et al., “Scene Flow to Action Map: A New Representation for RGB-D based Action Recognition with Convolutional Neural Networks”, in CVPR, 2017.

Keywords: 3D Action Recognition

## 概要

- RGB-Dの入力から3Dの行動認識を行う新しい表現方法であるScene Flow to Action Map (SFAM)を提案。コンテキストを考慮したConvNetsとなっている。RGB-Dを用いて、シーン、フローなどの情報をチャンネル変換カーネルにより相似のRGB空間へ射影する。この操作はImageNet Pre-trainのCNNよりも精度が高いことが示された。



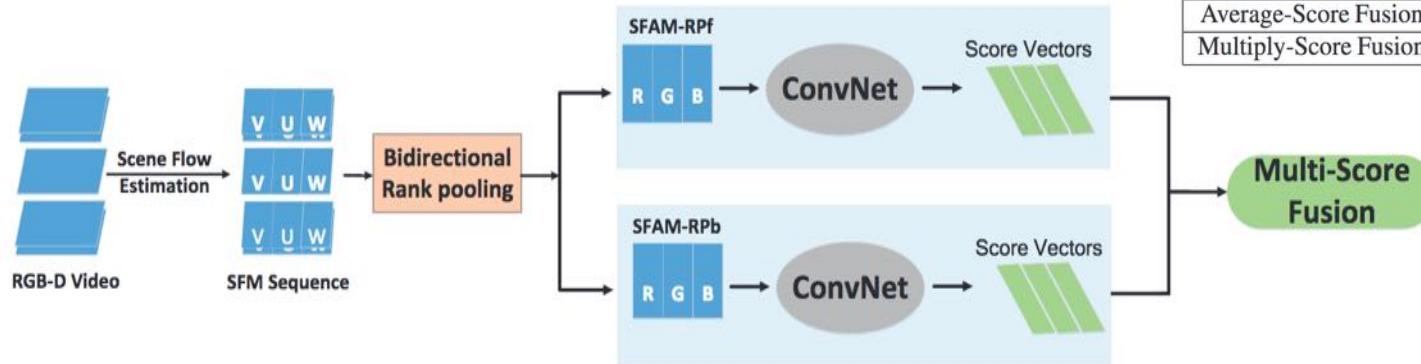
## 新規性・差分

- RGB-Dを用いた行動認識において、新しい表現であるSFAMを提案
- シーンフローを用いた、self-calibration方法を提案
- 様々な特性を持ったSFAM（右上図）を提案、認識の際に相補的に情報を補完する

## Links

論文 <https://arxiv.org/pdf/1702.08652.pdf>

Method	Accuracy	
	SV → FV	FV → SV
iDT-Tra [21]	43.3%	39.2%
iDT-COM [21]	70.2%	67.7%
iDT-HOG+MBH [21]	75.8%	71.8%
iDT-HOG+HOF [21]	78.2%	72.1%
SFAM-D	66.7%	65.2%
SFAM-S	68.2%	60.2%
SFAM-RP	71.6%	65.2%
SFAM-AMRP	77.7%	66.7%
SFAM-LABRP	76.9%	65.9%
Max-Score Fusion All	84.7%	73.8%
Average-Score Fusion All	85.3%	75.3%
Multiply-Score Fusion All	<b>87.6%</b>	<b>76.5%</b>



# 【52】Lluis Gomez, Yash Patel, Marcal Rusinol, Dimosthenis Karatzas, C. V. Jawahar, "Self-supervised learning of visual features through embedding images into text topic spaces", in CVPR, 2017.

Keywords: Self-Supervised Learning

## 概要

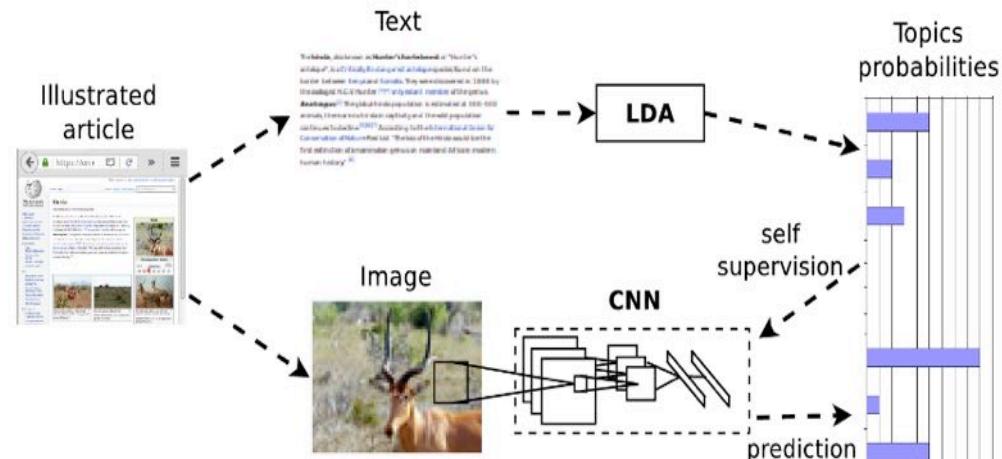
- ドキュメントを教師として画像認識用のCNNを学習することで自己学習（Self-Supervised Learning）を実現。具体的には右図のようにTextからトピックモデルのLDAにより教師を作成、画像認識のCNNに対する正解値として学習。テキスト・画像のペアはWikipediaから抽出してデータベースを作成。この自己学習により、画像識別、物体検出、マルチモーダル検出のタスクにて良好な精度を実現。

## 新規性・差分

- トピックモデルによりテキスト解析を行い、画像認識に対する教師データとし、自己学習を実現した
- STL-10データセットにてState-of-the-artな精度を実現（右表）、左表はPascal VOC 2007の結果である

## Links

論文 <https://arxiv.org/pdf/1705.08631.pdf>  
GitHub <https://github.com/lluisgomez/TextTopicNet>



Method	max5	pool5	fc6	fc7
TextTopicNet (Wiki)	-	<b>47.4</b>	<b>48.1</b>	<b>48.5</b>
Sound [25]	<b>39.4</b>	46.7	47.1	47.4
Texton-CNN	28.9	37.5	35.3	32.5
K-means [20]	27.5	34.8	33.9	32.1
Tracking [41]	33.5	42.2	42.4	40.2
Patch pos. [6]	26.8	46.1	-	-
Egomotion [1]	22.7	31.1	-	-
TextTopicNet (COCO)	-	<b>50.7</b>	<b>53.1</b>	<b>55.4</b>
ImageNet [21]	<b>63.6</b>	<b>65.6</b>	<b>69.6</b>	<b>73.6</b>
Places [46]	59.0	63.2	65.3	66.2

Table 2: PASCAL VOC2007 %mAP image classification.

Method	Acc.
TextTopicNet (Wiki) - CNN-finetuning *	<b>76.51%</b>
TextTopicNet (Wiki) - fc7+SVM *	66.00%
Semi-supervised auto-encoder [44]	<b>74.33%</b>
Convolutional k-means [8]	74.10%
CNN with Target Coding [43]	73.15%
Exemplar convnets [7]	72.80%
Unsupervised pre-training [26]	70.20%
Swersky <i>et al.</i> [34] *	70.10%
C-SVDDNet [37]	68.23%
K-means (Single layer net) [4]	51.50%
Raw pixels	31.80%

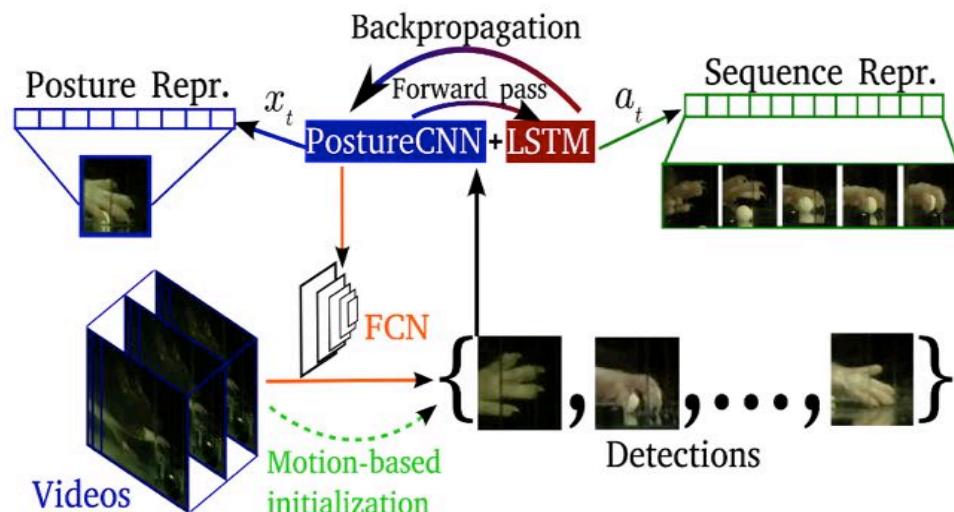
Table 4: STL-10 classification accuracy. Methods with an asterisk mark make use of external (unlabeled) data.

# [53] Biagio Brattoli, Uta Buchler, Anna-Sophia Wahl, Martin E. Schwab, Bjorn Ommer, "LSTM Self-Supervision for Detailed Behavior Analysis", in CVPR, 2017.

Keywords: Self-Supervision, PostureCNN, LSTM

## 概要

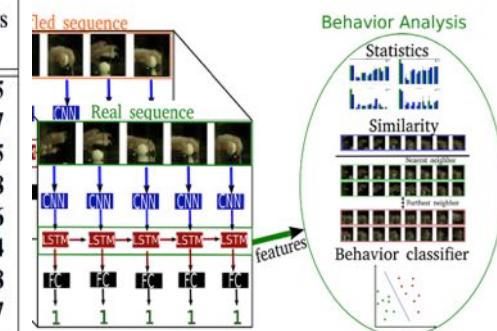
- ラベルづけされていない大量のデータから動作認識を行うために姿勢情報に対してLSTMのゲーティングを用いて自己学習（Self-Supervision）を行う。ここでは、（1）関節検出をFCNにより行い、動作認識のための初期値とする、（2）時系列の順序をLSTMにより学習、（3）時系列姿勢との差分により誤差逆伝播を行い、パラメータを強化することで自己学習を実現。



## 新規性・差分

- CNNとLSTMにより姿勢ベースの教示を行い、自己学習を成功させた
- 他のUnsupervised Deep Learningの手法[8, 13]やAlexNetよりも高い精度を出した（右表はOlympic Sports Dataset）

Category	HOG-LDA [19]	Ex. SVM [29]	Ex. CNN [13]	Alex net [24]	Clique CNN [8]	Ours
Basketball	0.51	0.63	0.58	0.55	0.70	<b>0.75</b>
Bowling	0.57	0.63	0.58	0.55	0.85	<b>0.87</b>
Clean&Jerk	0.61	0.71	0.58	0.62	0.81	<b>0.85</b>
Discus Thr.	0.42	0.76	0.56	0.59	0.65	<b>0.68</b>
Diving 10m	0.42	0.54	0.51	0.57	0.70	<b>0.76</b>
Diving 3m	0.50	0.57	0.52	0.66	0.76	<b>0.84</b>
HammerThr.	0.62	0.64	0.51	0.66	0.82	<b>0.88</b>
High Jump	0.64	0.76	0.59	0.62	0.82	<b>0.87</b>
Javelin Thr.	0.71	0.72	0.57	0.74	<b>0.85</b>	<b>0.85</b>
Long Jump	0.60	0.69	0.57	0.71	0.78	<b>0.85</b>
Pole Vault	0.59	0.64	0.60	0.64	0.81	<b>0.83</b>
Shot Put	0.51	0.67	0.52	0.70	0.75	<b>0.76</b>
Snatch	0.64	0.76	0.59	0.67	0.84	<b>0.89</b>
TennisServe	0.70	0.75	0.64	0.71	0.84	<b>0.87</b>
Triple Jump	0.63	0.65	0.58	0.65	0.80	<b>0.83</b>
Vault	0.59	0.71	0.63	0.68	0.81	<b>0.86</b>
Mean	0.58	0.67	0.56	0.65	0.79	<b>0.83</b>



Models	Accuracy(%)
Max frame similarity	74.1
Avg frame similarity	75.9
DTW[10]	76.8
ClusterLSTM	64.0
<b>CNN-LSTM<sub>2</sub></b>	<b>80.5</b>

## Links

論文

[https://hci.iwr.uni-heidelberg.de/sites/default/files/publications/files/1911875248/brattoli\\_buechler\\_cvpr17.pdf](https://hci.iwr.uni-heidelberg.de/sites/default/files/publications/files/1911875248/brattoli_buechler_cvpr17.pdf)

プロジェクト

<https://hci.iwr.uni-heidelberg.de/node/6137>

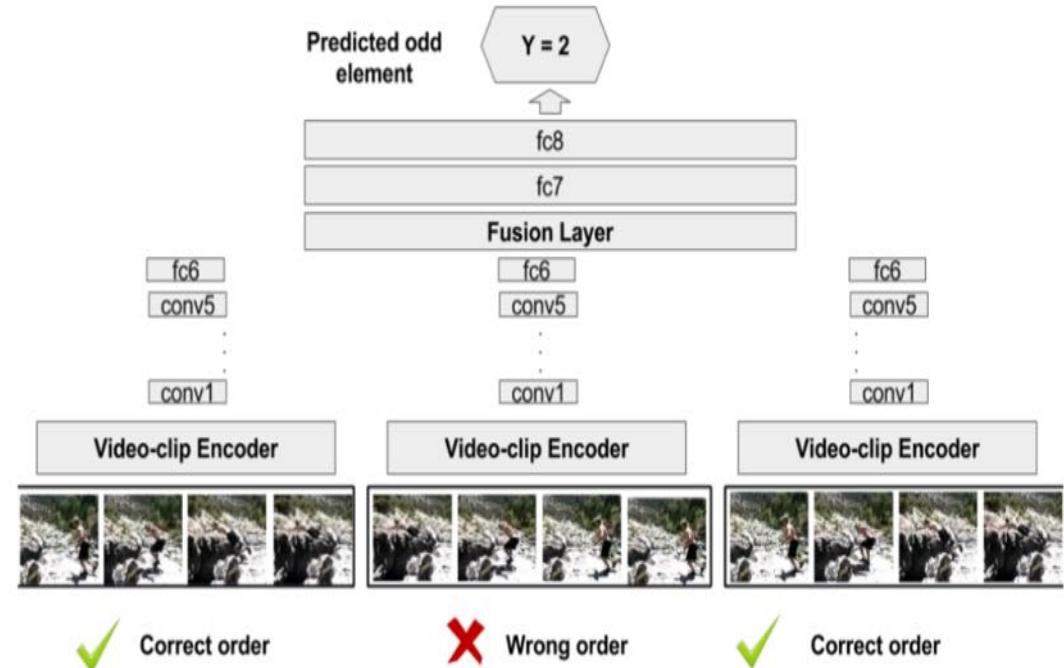
GitHub <https://github.com/bbrattoli/PermutationLSTM>

# 【54】Basura Fernando, Hakan Bilen, Efstratios Gavves, Stephen Gould, "Self-Supervised Video Representation Learning with Odd-One-Out Networks", in CVPR, 2017.

Keywords: Self-supervised CNN Pre-training, Odd-One-Out Learning

## 概要

- ・CNN Pre-trainをベースとして自己学習を行うための実験Odd-One-Out Learningを提案。本論文の手法では、関係もしくはおかしな要素をサンプルから見つけ出すことを基本戦略としている。右図は誤った（他とは異なる）1列の動画を探し出すことに成功している。



## 新規性・差分

- ・動画のオーダーを他の動画と比較することで自己学習実現するOdd-One-Out Networksを提案
- ・同學習に適したネットワーク構造を提案した
- ・動画セグメンテーションに対して有効であることを示した
- ・右下の表はUCF101, HMDB51に対する動画識別の結果である。

Method	UCF101-split1	HMDB51-split1
DrLim [17]	45.7	16.3
TempCoh [32]	45.4	15.9
Obj. Patch [44]	40.7	15.6
Seq. Ver. [31]	50.9	19.8
Our - Stack-of-Diff.	<b>60.3</b>	<b>32.5</b>
Rand weights - Stack-of-Diff.	51.3	28.3
ImageNet weights - Stack-of-Diff.	70.1	40.8

## Links

論文

<https://www.robots.ox.ac.uk/~vgg/publications/2017/Fernando17/fernando17.pdf>

# (55) Ke Gong, Xiaodan Liang, Xiaohui Shen, Liang Lin, "Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing", in CVPR, 2017.

Keywords: Look into Person Dataset (LIP), Self-Supervised Learning

## 概要

・多様な人物のアピランス分析を行うLook into Person (LIP) datasetを提案した。同データセットには50,000枚の画像データが含まれており、19の部位ごとのセマンティックラベル（セマンティックセグメンテーション向け）、16のキーポイント（関節推定）を含む（右図）。右下図はSelf-Supervised Structure-Sensitive Learningの簡易的なフローチャートである。学習済みのモデルからセグメンテーションや関節を推定し、正解値と比較し、それぞれの掛け合わせであるStructure-Sensitive Lossを計算してパラメータを学習する。

## 新規性・差分

- ・人物のセマンティックセグメンテーションと関節推定を同時に解析するLIPデータセットを提案
- ・同課題を自己学習により解決するためのStructure-Sensitive Lossを提案

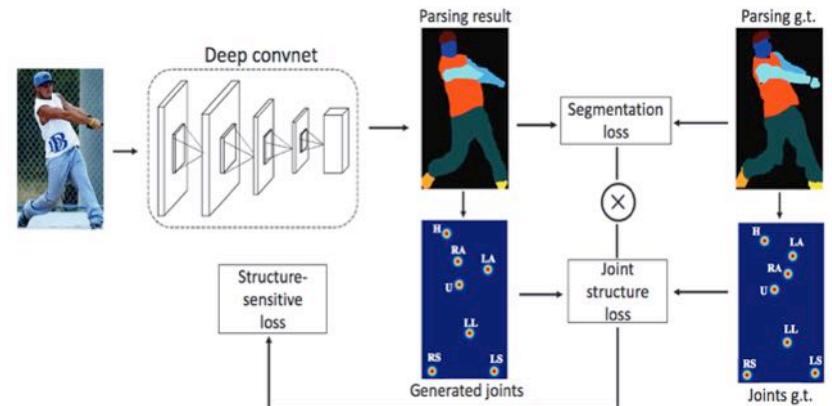
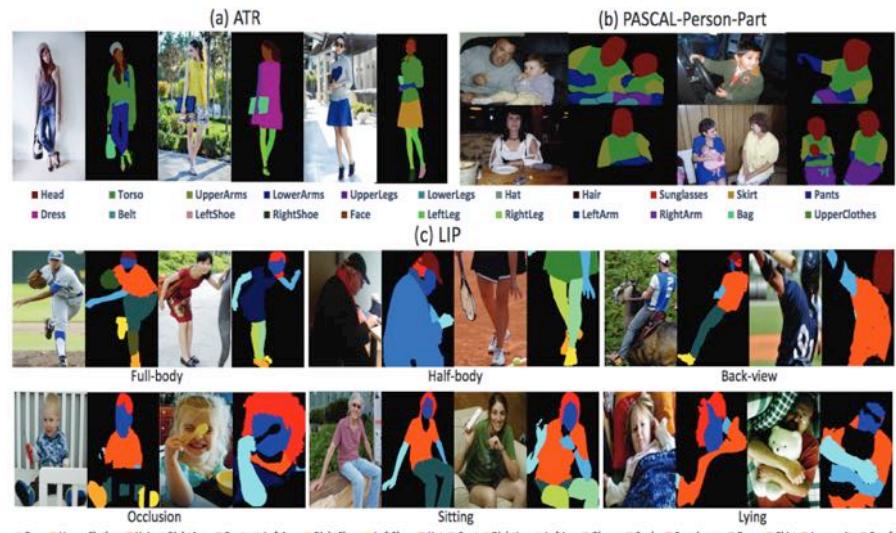
## Links

論文 <https://arxiv.org/abs/1703.05446>

プロジェクト <http://hcp.sysu.edu.cn/lip/>

GitHub [https://github.com/Engineering-Course/LIP\\_SSL](https://github.com/Engineering-Course/LIP_SSL)

Method	ATR	LIP	small	medium	large	153	321	513
SegNet [3]	15.79	21.79	16.53	18.58	18.18	16.92	18.37	16.44
FCN-8s [22]	34.44	32.28	22.37	29.41	28.09	14.52	15.55	16.25
DeepLabV2 [4]	48.64	43.97	28.77	40.74	43.02	36.49	37.59	37.28
Attention [5]	49.35	45.38	31.71	41.61	44.90	-	-	-
DeepLab + SSL	49.92	44.81	30.05	41.50	44.10	<b>38.27</b>	<b>38.97</b>	<b>39.84</b>
Attention + SSL	<b>52.69</b>	<b>46.85</b>	<b>33.48</b>	<b>43.12</b>	<b>46.73</b>	-	-	-



【56】

Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe, Elisa Ricci,  
“Viraliency: Pooling Local Virality”, in CVPR Poster, 2017.

Keywords: analyzing subjective attributes, virality, global pooling layer

## 概要

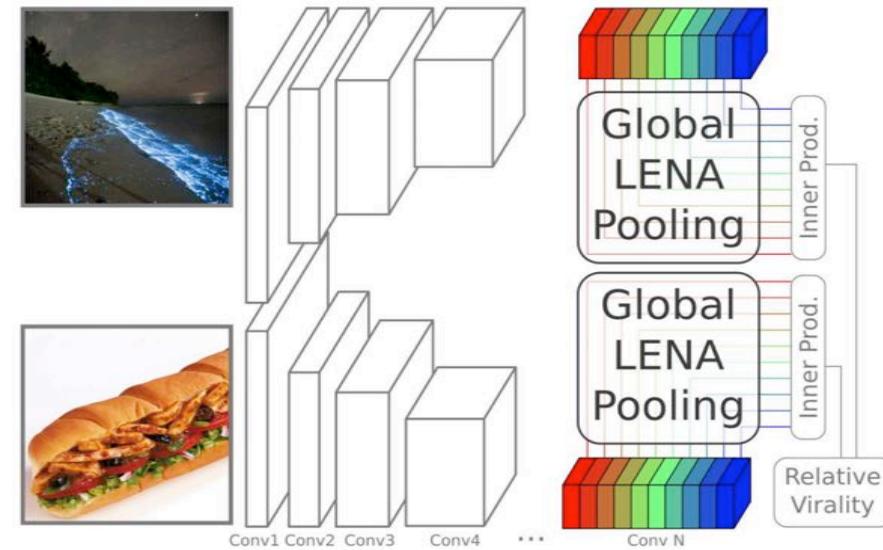
・画像において、Memorability, Popularity , Virality, Emotional contentなどのsubjective attributeの解析に取り組む研究が行われている。

このうち、Virality (SNS上での急速な拡散性；炎上、バズり？) に注目した。

Global Pooling戦略において、画像中のあるクラスのインスタンスを含んでいると思わしき領域をハイライトするactivation map推定が行われているが、これにのっとり、新たなプーリングレイヤLENA(The Learned top-N Average pooling)を提案する。

LENAではプーリングの各領域のサイズを学習できる。

ViralityのPredictingタスク、Localizingタスクによりパフォーマンスを評価し、state-of-the-artな性能を達成していることを示す。



## 新規性・差分

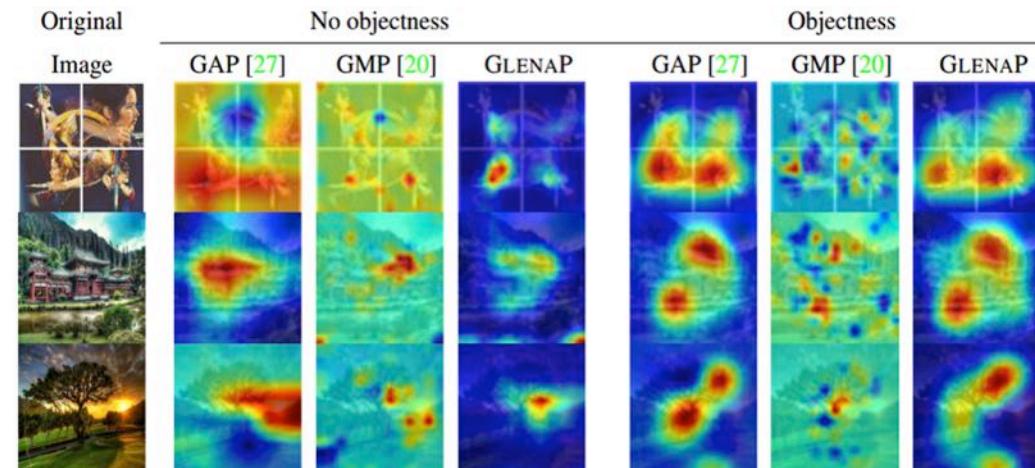
・社会現象の画像への表れに関する解析という社会的要請に対する取り組み、新しい手法 (CNNに対する新しいレイヤ) の提案、GAP,GNAPと比較して良いPrediction。

・LocalizingについてはViraliencyMapの定性的な比較評価を行い、画像に表れるViralityの性質について議論している。

## Links

論文 :

<http://xavirema.eu/wp-content/papercite-data/pdf/Alameda-CVPR-2017.pdf>



[57]

Felix Juefei-Xu, Vishnu Naresh Boddeti, Marios Savvides, "Local Binary Convolutional Neural Networks", in CVPR , 2017.

Keywords: CNN,LBP

## 概要

- ・畳み込み込みニューラルネットワーク(CNN)にローカルバイナリパターン(LBP)記述子を使用したLBCNN(Local Binary Convolutional Neural Networks)を提案する。従来のCNNの学習は畳み込みフィルタの重み付けを学習する必要があったが、あらかじめ定義された非可変スパースバイナリフィルタを組み合わせることにより、CNNと比較して学習する必要のあるパラメータが大幅に削減され、モデルサイズの節約が可能である。

## 新規性・差分

- ・LBPを用いたLBCレイヤの提案
- ・学習可能なパラメータ数の削減
- ・モデルサイズが最大で169倍も節約される

## Links

論文 <https://arxiv.org/pdf/1608.06049.pdf>

プロジェクト <http://vishnu.boddeti.net/projects/lbcnn.html>

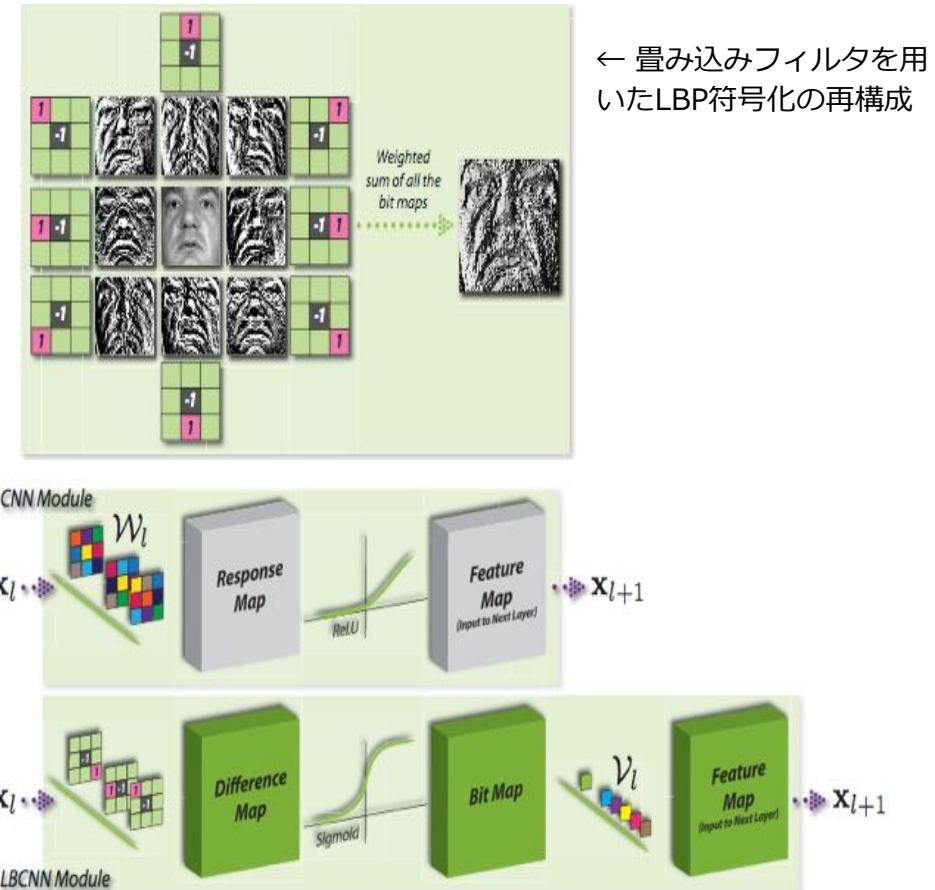


Figure 3: Basic module in CNN and LBCNN.  $W_l$  and  $V_l$  are the learnable weights for each module.

	LBCNN	Baseline	BinaryConnect [6]	BNN [5, 14]	ResNet [12]	Maxout [9]	NIN [23]
MNIST	99.51	99.48	98.99	98.60	/	99.55	99.53
SVHN	94.50	95.21	97.85	97.49	/	97.53	97.65
CIFAR-10	92.99 (93.66 NetEverest)	92.95	91.73	89.85	93.57	90.65	91.19

Table 2: Classification accuracy (%). LBCNN column only shows the best performing model and the Baseline column shows the particular CNN counterpart.

# 【58】De-An Huang, Joseph J. Lim, Li Fei-Fei, Juan Carlos Niebles, "Unsupervised Visual-Linguistic Reference Resolution in Instructional Videos", in CVPR, 2017.

Keywords: Unsupervised Learning, Visual-Linguistic

## 概要

- (料理などの) インストラクションビデオにおいて言語的な対応関係と合わせてビデオを解析することで、教師なしで物体や動詞の対応関係をマッチング (onion -- cuttingなど) する。視覚モデルと言語モデルをつなぐAction Graph (G)により名詞と動詞の対応関係を明らかにする(右上図)。Gにはビデオの視覚的・言語的な履歴を含んでおり、視覚と言語の処理は独立に行われ、EMアルゴリズムにより最適化される。

## 新規性・差分

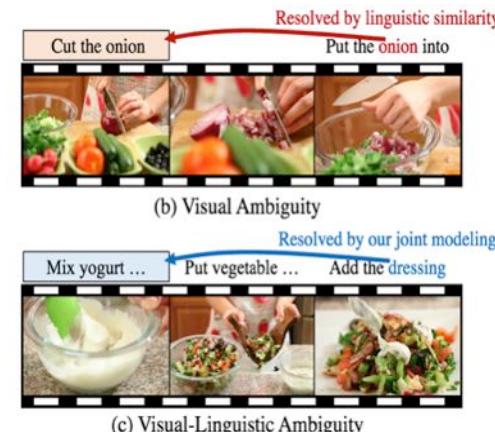
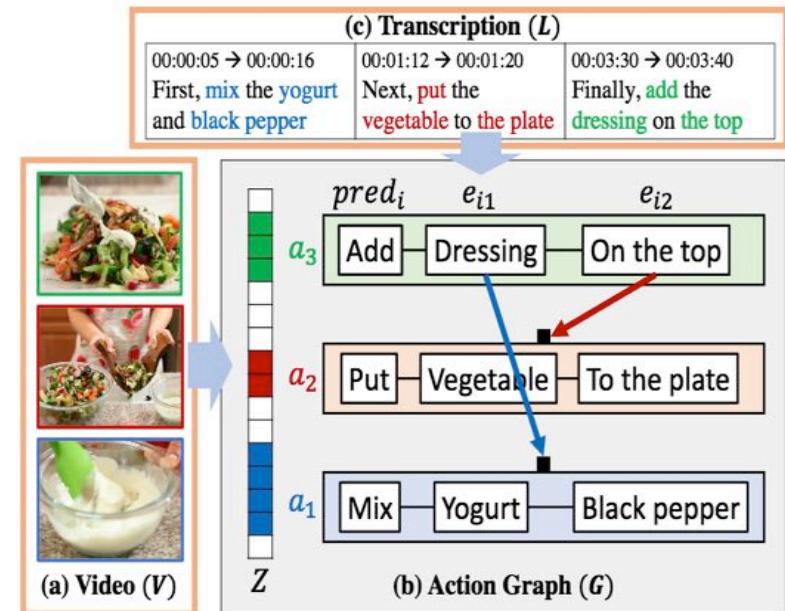
- 視覚的・言語的な処理により品詞間の対応関係を把握した
- 教師なし学習の枠組みにより、高精度な動画像解析(特に、インストラクションビデオ解析)を実現

## Links

論文 <https://arxiv.org/pdf/1703.02521.pdf>

プロジェクト <http://ai.stanford.edu/~dahuang/projects/vlrr/>

Methods	P	R	F1
Sequential Initialization	0.483	0.478	0.480
Random Perturbation	0.399	0.386	0.397
Our Visual Model Only	0.294	0.292	0.293
Our Linguistic Model Only [23]	0.621	0.615	0.618
RFES + Linguistic w/o Align	0.424	0.422	0.423
FES + Linguistic w/o Align	0.547	0.543	0.545
Our Visual + Linguistic w/o Align	0.691	0.686	0.688
Our Visual + Linguistic (Our Full)	<b>0.710</b>	<b>0.704</b>	<b>0.707</b>



# 【59】Alexander Richard, Hilde Kuehne, Juergen Gall, "Weakly Supervised Action Learning with RNN based Fine-to-Coarse Modeling", in CVPR, 2017. (oral)

Keywords: Temporal Action Detection, Weakly Supervised Learning

## 概要

弱教師付き学習により行動の開始終了を特定するTemporal Action Detectionの課題を解決する。ここでは開始終了のラベルがない状態で行動間の切れ目を推定することをもって弱教師付き学習と言っている。(1) 識別しやすいサブアクションをRNNにより推定(Fine) (2) 長い時系列ビデオの中から粗く確率分布を推定する識別器(coarse)、の統合によりラベルなしの状態で開始終了を推定する。明確な教師を用いるわけではないため本手法で用いるRNNの学習には時間を要するが、収束するまで繰り返し学習する。

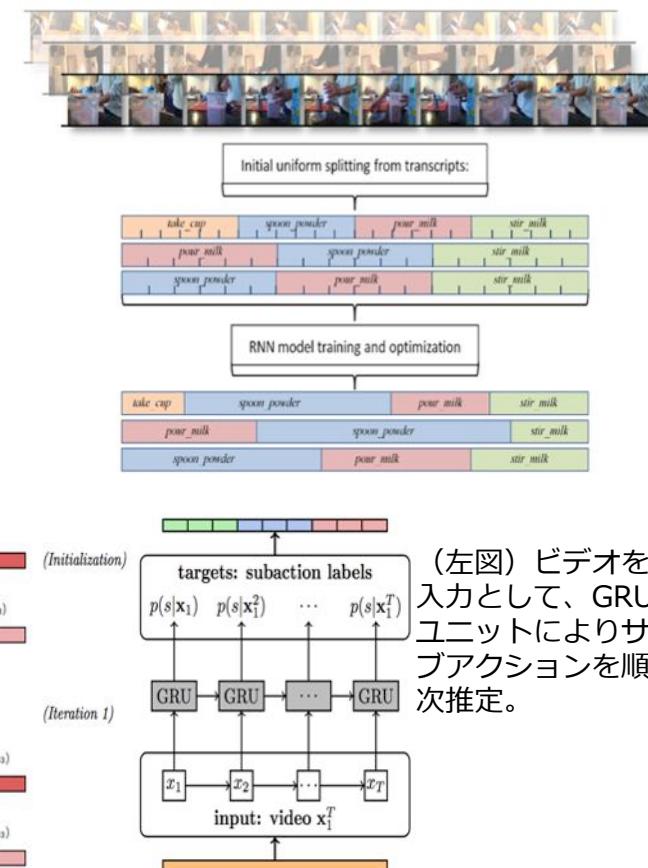
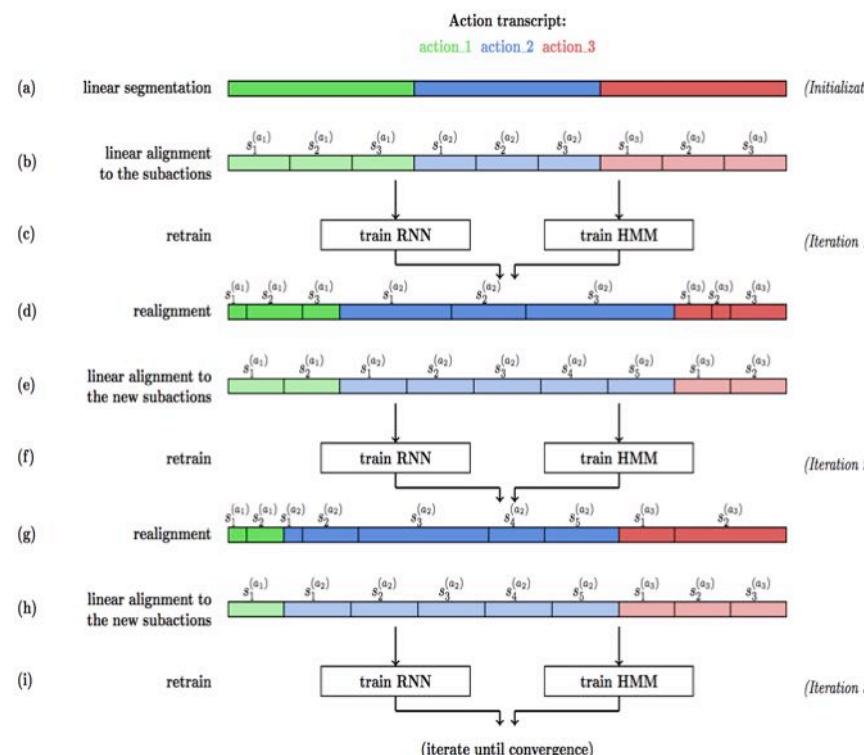
## 新規性・差分

- RNNを用いた弱教師付き学習による行動の開始終了推定
- Fine-to-Coarseな戦略(サブアクションへの分割と再学習)にて収束させる

## Links

論文

[http://pages.iai.uni-bonn.de/  
kuehne\\_hilde/projects/  
rnnActionModeling/paper.pdf](http://pages.iai.uni-bonn.de/kuehne_hilde/projects/rnnActionModeling/paper.pdf)



(左図) 弱教師付き学習の流れ。初期値ではサブアクションの数は等分される。学習が進んでいくごとにアクションの時間調整やサブアクションの分割数が増えていく。RNNとともにHMMも学習して推定値を合わせていく。

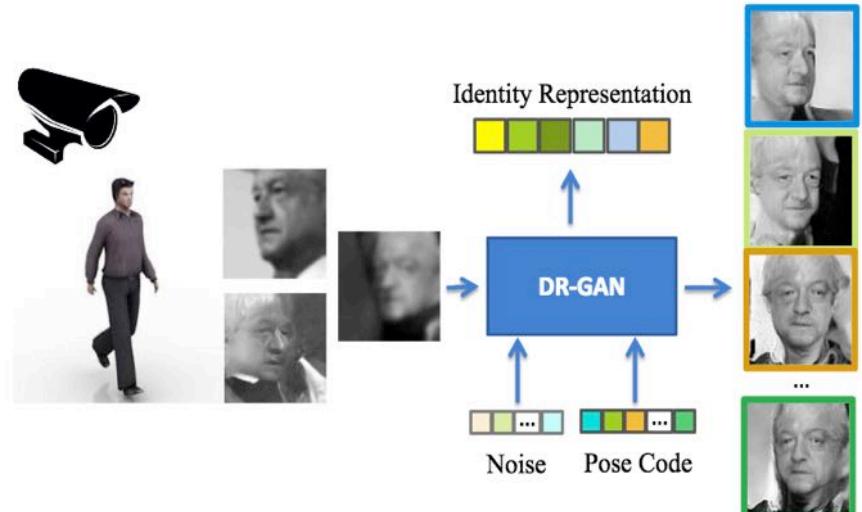
[60]

Luan Tran, Xi Yin, Xiaoming Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition", in CVPR, 2017. (oral)

Keywords: GAN, Pose-Invariant

## 概要

- GANによる画像生成の仕組みを利用して、姿勢不变の（角度が異なる）顔画像を生成しようとする試み。例えば横顔を正面顔に変換する、オクルージョンがある場合にオクルージョンを外すと言ったことができる。このDisentangled Representation GAN (DR-GAN) は顔画像と固定長のノイズ、姿勢コードを入力することで異なるビューポイントの顔画像を出力する。DR-GANの特徴としては、(1) Encoder-Decoderモデルにより、識別的にも生成的にもすぐれたモデルを構築 (2) 表情やビューポイントが異なる場合にも推定可能 (3) 同一人物の複数画像入力からより精度の高い生成結果を得ることができる。



## 新規性・差分

- Encoder-Decoderの仕組みによるGANにより、異なる角度の顔画像を推定できる。
- 表情や姿勢の変化に頑健な、なおかつ複数画像の入力からより精度の高い生成結果を得られる



- 上の結果は(a) 入力画像、(b) 正面顔の生成結果、(c) 正面顔の正解値である。

## Links

論文

[http://cvlab.cse.msu.edu/pdfs/Tran\\_Yin\\_Liu\\_CVPR2017.pdf](http://cvlab.cse.msu.edu/pdfs/Tran_Yin_Liu_CVPR2017.pdf)

プロジェクト <http://cvlab.cse.msu.edu/project-dr-gan.html>

# 【61】Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, Bernt Schiele, "ArtTrack: Articulated Multi-person Tracking in the Wild", in CVPR, 2017. (oral)

Keywords: Pose Tracking

## 概要

- ・人体姿勢トラッキングの手法。推定された姿勢に対してBody-Part Relationship Graphを用いることで時系列方向に対して高速かつ高精度に姿勢を対応づける。複数人を時系列的に対応づけることにより、Identificationや関節の外れ値を推定してしまうことが減少する。

## 新規性・差分

- ・関節を追跡するフレームワークを提案し、高速化（まいフレームの姿勢推定と比較すると24倍高速）や高精度化に寄与した
- ・MOTA, MPII PoseなどのデータセットにおいてState-of-the-artな精度を実現した

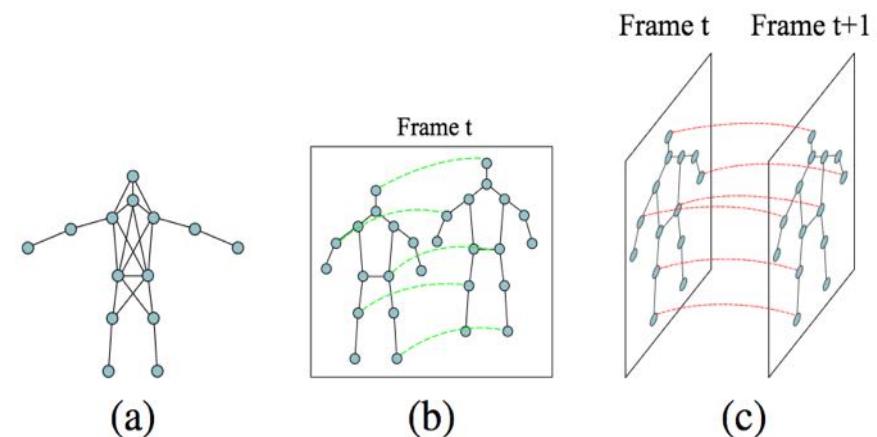
## Links

論文 <https://arxiv.org/abs/1612.01465>

ビデオ <https://www.youtube.com/watch?v=eYtn13fzGG0>



Figure 1. Example articulated tracking results of our approach.



Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP
<i>BU-sparse</i>	84.5	84.0	71.8	59.5	74.4	68.1	59.2	71.6
+ <i>det-distance</i>	84.8	84.3	72.9	61.8	74.1	67.4	59.1	72.1
+ <i>deepmatch</i>	85.5	83.9	73.0	62.0	74.0	68.0	59.5	72.3
+ <i>det-distance</i>	85.1	83.6	72.2	61.5	<b>74.4</b>	68.8	62.2	72.5
+ <i>sift-distance</i>	<b>85.6</b>	<b>84.5</b>	<b>73.4</b>	<b>62.1</b>	73.9	<b>68.9</b>	<b>63.1</b>	<b>73.1</b>

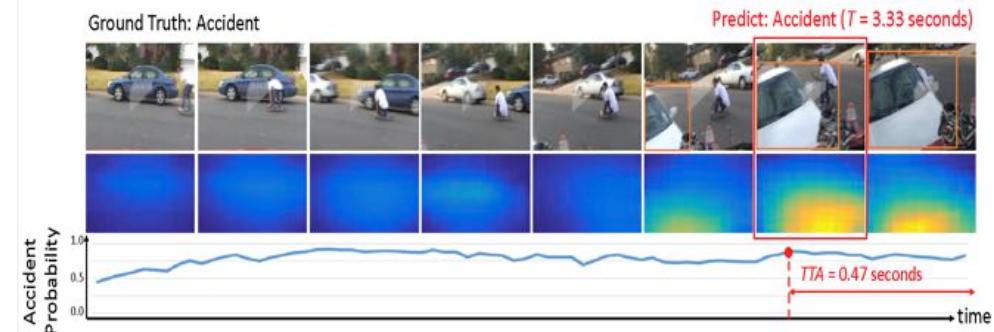
Method	Head	Sho	Elb	Wri	Total
<i>TD/BU</i>	<b>97.5</b>	<b>86.2</b>	<b>82.1</b>	<b>85.2</b>	<b>87.7</b>
<i>DeeperCut</i> [14]	92.6	81.1	75.7	78.8	82.0
<i>DeepCut</i> [23]	76.6	80.8	73.7	73.6	76.2
<i>Chen&amp;Yuille</i> [7]	83.3	56.1	46.3	35.5	55.3

# 【62】Kuo-Hao Zeng, Shih-Han Chou Fu-Hsiang Chan Juan Carlos Niebles, Min Sun , "Agent-Centric Risk Assessment:Anticipation and Risky Region Localization ", in CVPR, 2017.

Keywords: risk estimation, risk anticipation

## 概要

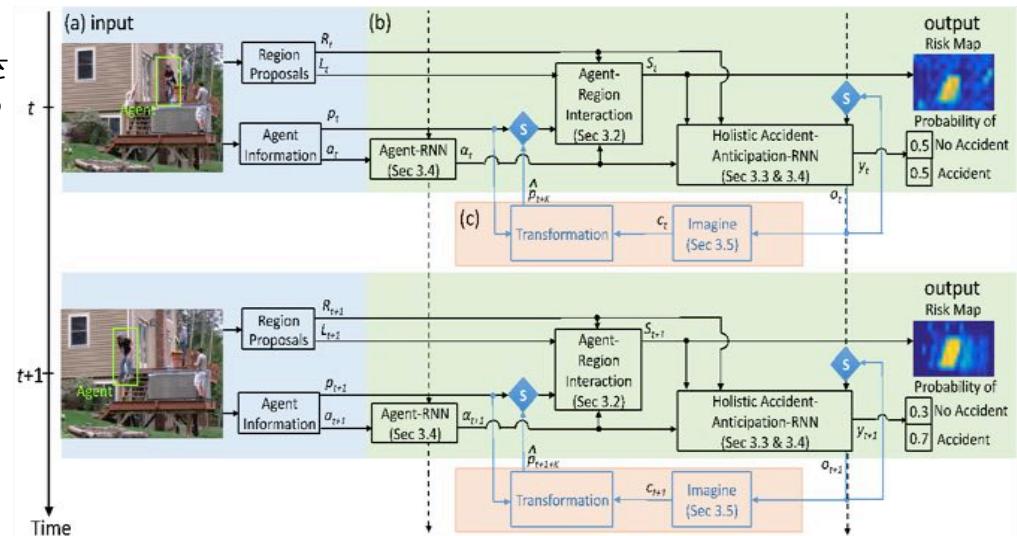
- agent-centricの危険度推定, 危険因子の領域推定, 及び危険度予測の論文。agentと危険因子候補領域のapperance特徴, 位置関係を元に位置危険度を表すRisk mapを作成。Risk mapと各apperanceを入力にagentごとの現時刻の危険度を推定。agentのapperanceと危険度を推定する直前の機構にLSTMを用いて時系列情報を考慮する。さらに, imagine機構によって, 次時刻のagent位置予測し, そこから先のフレームの危険度の推定が可能であり, 再帰的に任意の先フレームでの危険度推定が可能である。最終的に予測した各フレームの危険度の重み付け平均をFuture risk prediction とする。



- agent, 危険因子候補領域はFaster-RCNNによって検出しており, apperanceはその中間層の出力をGlobal Average Pooling することで取得している。実験では, 作成したEpic Fail (EF) dataset を用いて危険度の予測指標Time to accident にて評価を行った。学習はend to endで行うことができる。

## 新規性・差分

- agent と危険因子候補領域のを時系列的に関連付け, さらにimagine機構を用いて次時刻のagent位置推定することで, Street Accident (SA) dataset , Epic Fail (EF) dataset において危険度予測, 危険候補領域推定で最良
- Epic Fail (EF) dataset の提案



## Links

論文 <https://arxiv.org/abs/1705.06560>

# [63] Martin Simonovsky, Nikos Komodakis , "Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs ", in CVPR, 2017.

Keywords: CNN, Point cloud classification

## 概要

- ・グラフの畳み込みをpoint cloud classificationに応用して、CNNでのグラフ生成より優れたedge-conditioned convolution (ECC) を提案。フィルターの重みがedge labelsで調節され,そのまま特定のインプットを生成する空間ドメインを用いる。これによりedge labelが正しく選ばれた時、standard convolutionがグラフに生成される。

## 新規性・差分

- ・CNNで扱われていない、edge labelに着眼。情報チャネルを利用し、graph classificationの性能をあげる畳み込み演算を提案。
- ・benchmark NCI1 で優れたgraph classificationを実現。

## Links

論文 <https://arxiv.org/pdf/1704.02901.pdf>

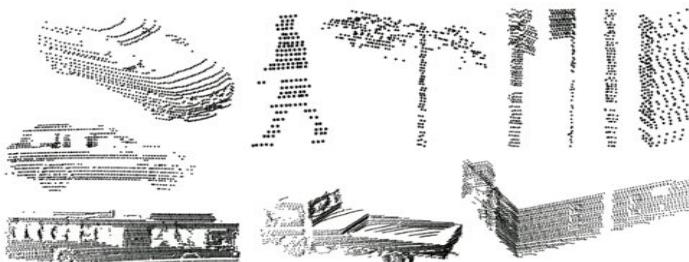
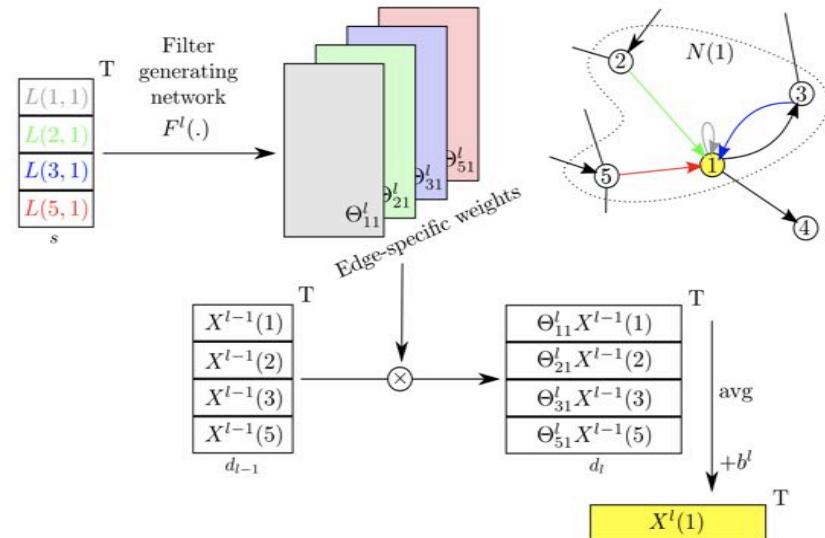
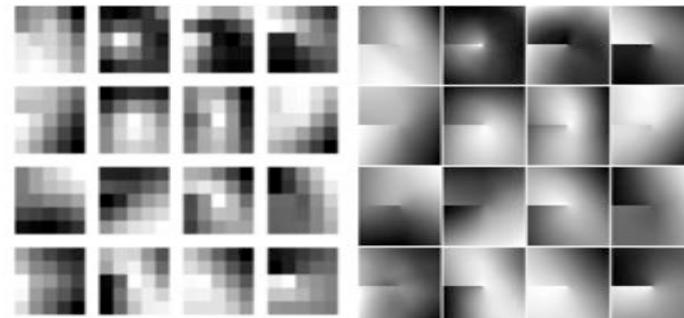


Figure 4. Illustrative samples of the majority of classes in Sydney Urban Objects dataset, reproduced from [9].



Model	Train accuracy	Test accuracy
ECC	99.12	99.14
ECC (sparse input)	99.36	99.14
ECC (one-hot)	99.53	99.37

Table 5. Accuracy on MNIST dataset [23].

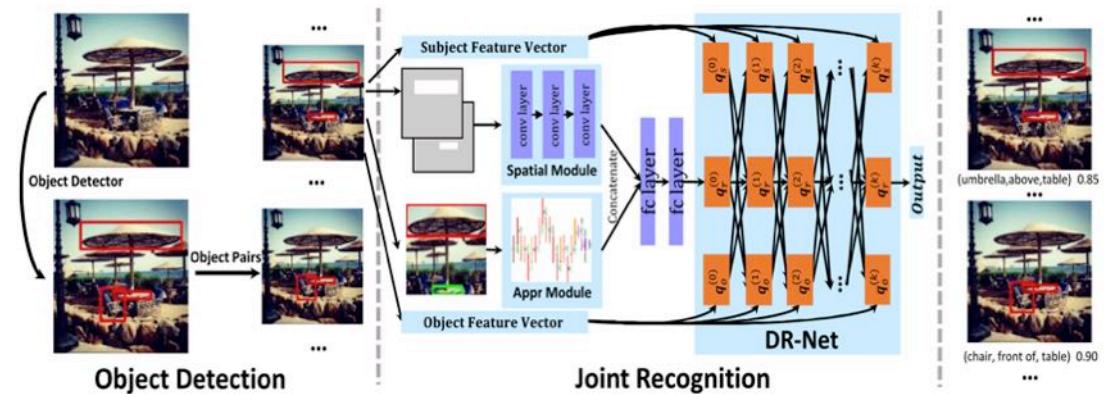


# [64] Bo Dai, Yuqi Zhang, Dahua Lin, "Detecting Visual Relationships with Deep Relational Networks", in CVPR, 2017.

Keywords: Visual Relationship Detection, Object Detection, DNN

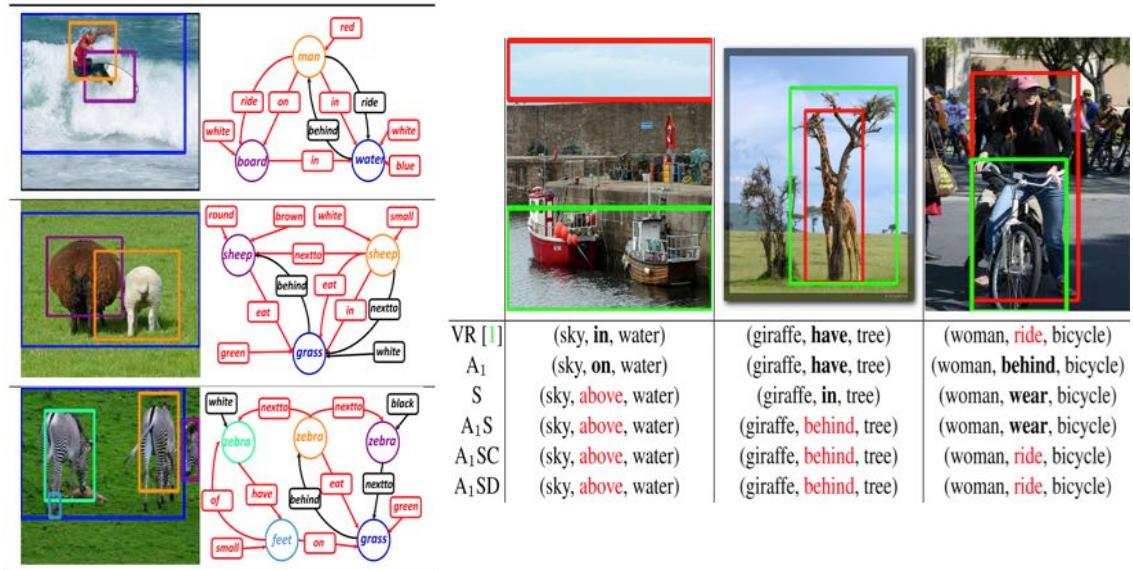
## 概要

- Visual Relationship Detectionを行うDNNを提案  
画像中に存在する物体の検出と2物体間の関係  
(sitなどのactionやaboveなどの位置, taller thanなどの比較などを含む一般的なもの) を記述する課題. (source, relation, object)というtripletを推定する. tripletごとに1クラスと扱う (visual phrase) と2物体の組み合わせが非常に多いので, この研究では物体クラスとrelationクラスを別々に扱い, 3要素のJoint Recognitionにより認識. 手法の流れとしては, まず物体検出を行い, 画像中からObject Pairsを生成する. そしてtripletを推定するDeep Relation NetworkにPairを入力することでVisual Relationship Detectionを行う.



## 新規性・差分

- 従来やされていたCRFなどによるモデリングではなく, statistical inferenceを組み込んだDNNによりVisual Relationshipを認識



## Links

論文 <https://arxiv.org/pdf/1704.03114.pdf>  
Github <https://github.com/doubledaibo/drnet>

【65】

Pengfei Dou, Shishir K. Shah, Ioannis A. Kakadiaris, "End-to-end 3D face reconstruction with deep neural networks", in CVPR, 2017. (poster)

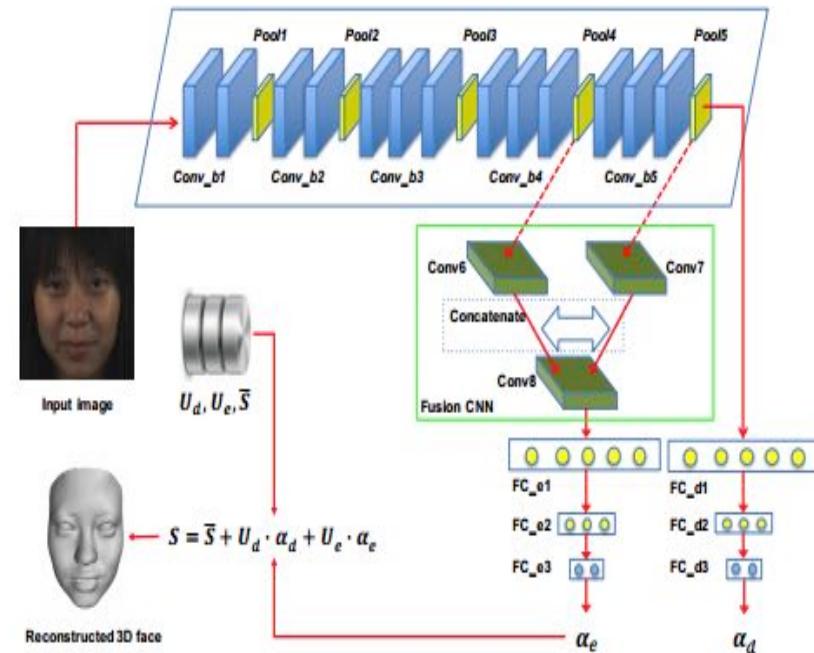
Keywords: 3D face reconstruction, DNN, end-to-end

## 概要

- DNN-based手法、一枚の顔画像から顔の3Dモデル生成する。end-to-end手法なのでrendering processが不要。3Dカタチはidentityとexpressionで構成。
- Reconstructionを二つのsub-taskに分ける、neutral 3D facial shapeとexpressive 3D facial shape。
- 一つのDNNモデルで違う種類のneural layersをtraining。実験結果によって、reconstructionの精度向上を示した。

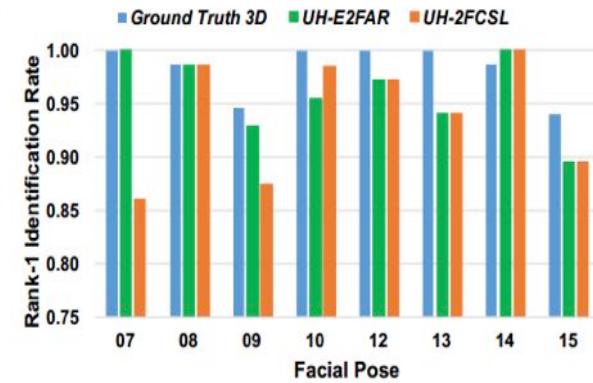
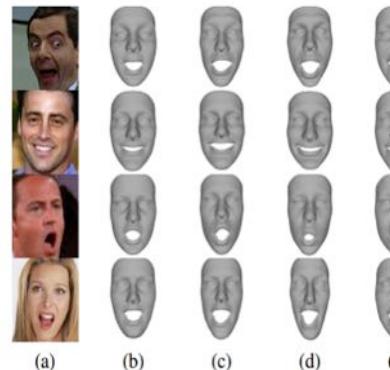
## 新規性・差分

- 従来の手法では、顔画像以外、geometry imageや初期顔モデルなどの付加情報が必要である。この研究はただ一枚の顔画像で3dモデルを生成する手法を提案する。
- フレームワークを簡易化した。DNNモデルはEnd-to-endなので、rendering processが不要。



## Links

論文 <https://arxiv.org/pdf/1704.05020.pdf>



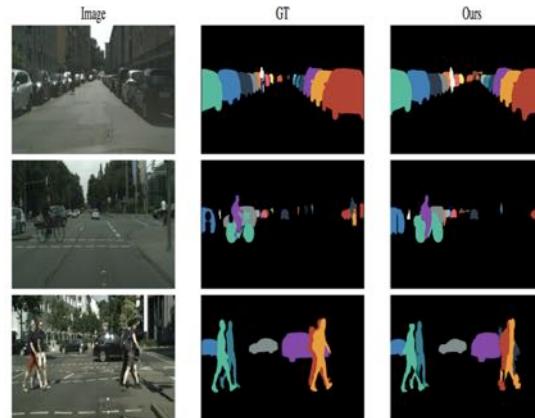
[66]

Mengye Ren, Richard S. Zemel, "End-to-End Instance Segmentation with Recurrent Attention", in CVPR, 2017.

Keywords: Instance segmentation, iterative procedure,

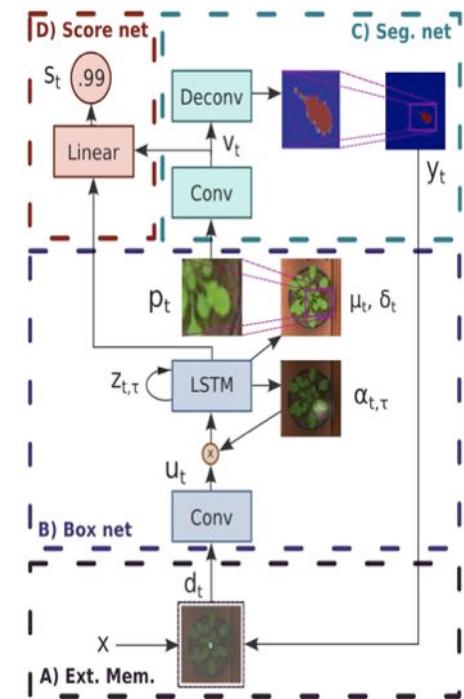
## 概要

- 人間が物を数えながら区別する認識法を真似た新たなEnd-to-endのRNNの手法を提案。
- 4つのセクションに分けられる; External Memory (常にsegmentedしたオブジェクトをトラッキングする)、Box proposal network(オブジェクトのローカライジング)、Segmentation network(boxの中で画像ピクセルをsegmenting)、Scoring network(セグメントの評価をする)



## 新規性・差分

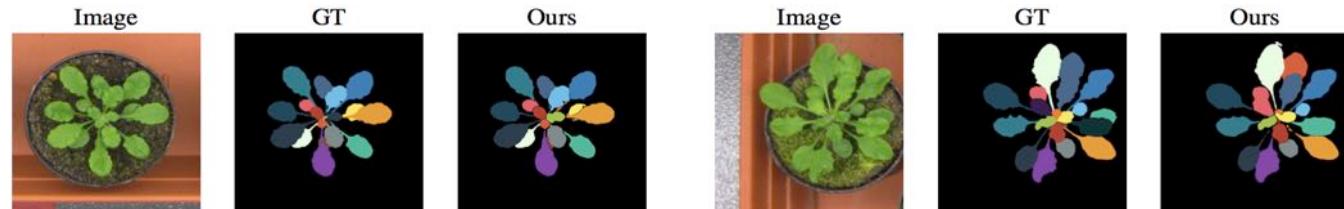
- 正解値を用いた学習をさせて徐々にそれを予測値に置き換えるBootstrap training、徐々に正解値へのrelianceを取り除き最終的に前のステップで出たアウトプット値を見るように学習させるScheduled samplingを提案。
- End-to-endな反復方法のため、かなり少ないパラメータでsegmentationができる。



## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Ren\\_End-To-End\\_Instance\\_Segmentation\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Ren_End-To-End_Instance_Segmentation_CVPR_2017_paper.pdf)

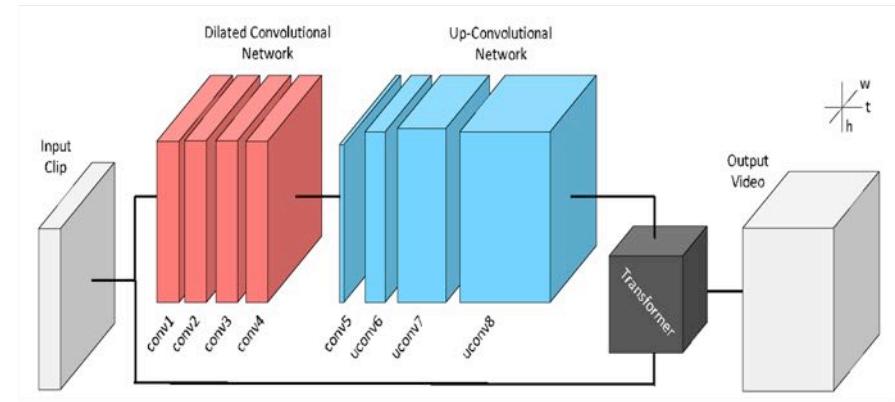


# 【67】Carl Vondrick and Antonio Torralba , “Generating the Future with Adversarial Transformers”, in CVPR, 2017.

Keywords: GAN, generate future

## 概要

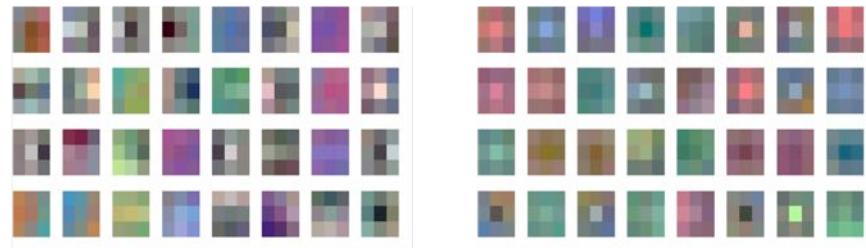
- Adversarial Learning を用いたビデオフレームの予測手法。  
画素値の値をそのまま予測するのではなく、現フレームから次フレームへの変換を推定する事で、ネットワークに画素情報の保存を強いて予測を行う。変換は出力画像のある画素に対して、その近傍の入力画素を重み付けしたものとして定義し, end to endで学習が可能である。  
実験では生成フレームの尤もらしさの評価と、予測を学習したモデルを初期値としたモデルでfine tuningを行った場合の識別タスクでの評価を行っている。



全体図

## 新規性・差分

- 画素を直接生成するモデル、Adversarial Learningではなくregressionで推定するモデルと比較して最も尤もらしい予測フレームが生成できた。またrandom initializeしたものよりもfine tuningした際にPascal VOCの識別タスクにおいても良い結果となった。フィルタを可視化した結果、変換を生成する場合は、モデル内で画素置を保存する必要がないので汎用的なエッジ特徴などに着目していることがわかった。



フィルタの可視化

## Links

論文 <http://carlvondrick.com/transformer.pdf>

		Not Preferred					Method	2007 mAP	2012 mAP
		Adv+Tra	Reg+Tra	Adv+Int	Reg+Int	Real			
Preferred	Adv+Tra	-	<b>55.6</b>	<b>61.2</b>	<b>55.1</b>	30.6	Chance	7.3	7.2
	Reg+Tra	44.4	-	<b>60.8</b>	<b>54.1</b>	36.4	Random Initialization	26.7	30.6
	Adv+Int	38.8	39.2	-	39.6	37.3	Regression + No Transform	30.0	33.6
	Reg+Int	44.9	45.9	<b>60.4</b>	-	38.0	Adversary + No Transform	29.7	33.3
Real		<b>69.4</b>	<b>63.6</b>	<b>62.7</b>	<b>62.0</b>	-	Regression + Transform	32.6	38.8
							Adversary + Transform	32.0	38.1

# 【68】George Trigeorgis, Patrick Snap, “Face Normals “in-the-wild” using Fully Convolutional Networks”, in CVPR, 2017.

Keywords: Estimating surface normals, FCN

## 概要

- ・1枚のintensity画像から表面法線を推定するデータ駆動型アプローチを提案した。また、顔画像をフォーカス。様々な顔表情と姿勢から正確に法線を生成できるFully畳みこみネットワークを学習した

## 新規性・差分

- ・従来手法と比べてより正確的でリアルな表面法線を生成が可能
- ・FCNを用いているので、アライメントステップが必要なし
- ・現在利用可能な顔データベースをデータセット構築に利用

## Links

論文

[https://ibug.doc.ic.ac.uk/media/uploads/documents/normal\\_estimation\\_cvpr\\_2017 -4.pdf](https://ibug.doc.ic.ac.uk/media/uploads/documents/normal_estimation_cvpr_2017 -4.pdf)

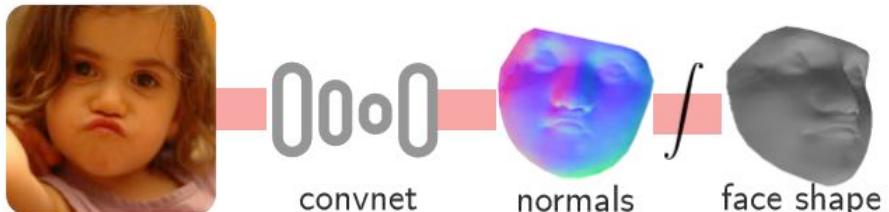
## 提案手法との定量的な比較

Name	Mean $\pm$ Std	$< 20^\circ$	$< 25^\circ$	$< 30^\circ$
PS w/o Light	$42.9 \pm 15.2$	1.1%	13.1%	35.8%
IMM [29, 58]	$24.2 \pm 5.4$	23.5	64.6%	88.3%
3DMM	$26.3 \pm 10.2$	4.3%	56.05%	89.4%
Marr Rev. [2]	$28.3 \pm 10.1$	31.8%	36.5%	44.4%
UberNet [33]	$29.1 \pm 11.5$	30.8%	35.5%	55.2%
Proposed	$22.0 \pm 6.3$	36.63%	59.8%	79.6%

Loss	Mean $\pm$ Std	$< 20^\circ$	$< 25^\circ$	$< 30^\circ$
Cosine Loss	$21.5 \pm 6.9$	29.9%	55.9%	81.5%
Smooth $\ell_1$ Loss	$22.0 \pm 6.3$	36.63%	59.8%	79.6%

Table 2: Angular error for the different loss functions.

## 提案手法の流れ



1. 顔画像から表面法線生成するFCNをトレーン；
2. 従来手法を用い表面法線から3次元形状復元

## 提案手法との直感的な比較

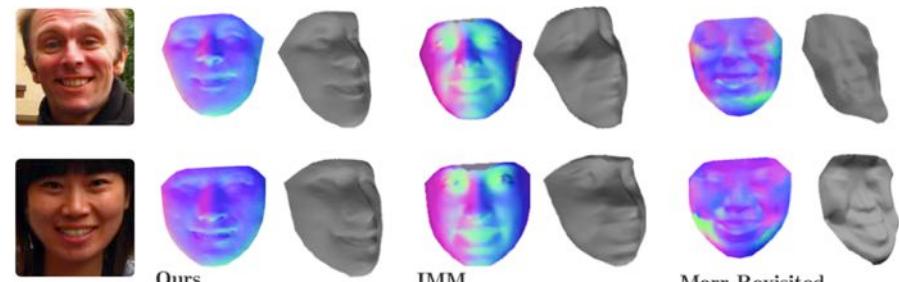


Figure 6: Example facial normal estimation and surface reconstruction from the Helen Dataset.

Architecture	Mean $\pm$ Std	$< 20^\circ$	$< 25^\circ$	$< 30^\circ$
Resnet + Cosine	$21.5 \pm 6.9$	29.9%	55.9%	81.5%
Pixelnet + Cosine	$23.5 \pm 6.3$	35.17%	58.0%	78.2%

Table 3: Angular error for the different architectures.

# [69] Dan Xu, Elisa Ricci, "Multi-Scale Continuous CRFs as Sequential Deep Networks for Monocular Depth Estimation", in CVPR2017.

Keywords: Continuous CRFs, Depth Estimation

## 概要

- 1枚の静止のRGB画像からデプス推定を行う。複数のCNNの側出力から得られる補充情報を融合するディープモデルを提案した。融合する際、連続CRFsを用いた。

## 新規性・差分

- 連続CRFsを用いてCNNのマルチスケールな側出力を統合するフレームワークを提案。
- 提案フレームワークを用いていくつかの同じCNN構造の統合が可能
- End-to-Endトレーニングに適応できる。

## Links

論文 <https://arxiv.org/pdf/1704.02157.pdf>

## 異なるCNN構造を適応した比較結果

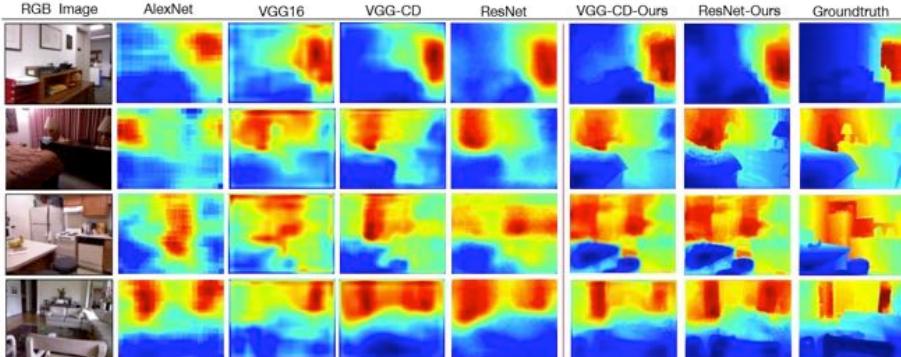


Figure 5. Examples of depth prediction results on the NYU v2 dataset. Different network architectures are compared.

## ネットワーク構造

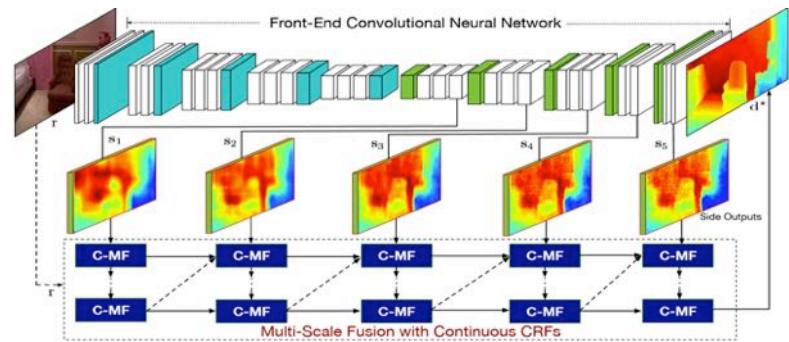


Figure 2. Overview of the proposed deep architecture. Our model is composed of two main components: a front-end CNN and a fusion module. The fusion module uses continuous CRFs to integrate multiple side output maps of the front-end CNN. We consider two different CRFs-based multi-scale models and implement them as sequential deep networks by stacking several elementary blocks, the C-MF blocks.

- front-end CNN + 融合モジュール

## 実験

- NYU Depth V2, Make3Dデータセットで検証

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Karsch <i>et al.</i> [29]	0.349	-	1.214	0.447	0.745	0.897
Ladicky <i>et al.</i> [14]	0.35	0.131	1.20	-	-	-
Liu <i>et al.</i> [21]	0.335	0.127	1.06	-	-	-
Ladicky <i>et al.</i> [17]	-	-	-	0.542	0.829	0.941
Zhuo <i>et al.</i> [37]	0.305	0.122	1.04	0.525	0.838	0.962
Liu <i>et al.</i> [20]	0.230	0.095	0.824	0.614	0.883	0.975
Wang <i>et al.</i> [32]	0.220	0.094	0.745	0.605	0.890	0.970
Eigen <i>et al.</i> [9]	0.215	-	0.907	0.611	0.887	0.971
Roi and Todorovic [26]	0.187	0.078	0.744	-	-	-
Eigen and Fergus [8]	0.158	-	0.641	0.769	0.950	0.988
Laina <i>et al.</i> [18]	0.129	0.056	<b>0.583</b>	0.801	0.950	0.986
Ours (ResNet50-4.7K)	0.143	0.065	0.613	0.789	0.946	0.984
Ours (ResNet50-95K)	<b>0.121</b>	<b>0.052</b>	0.586	<b>0.811</b>	<b>0.954</b>	<b>0.987</b>

Table 4. NYU Depth V2 dataset: comparison with state of the art.

# (70) Frank Michel, Alexander Kirillov, "Global Hypothesis Generation for 6D Object Pose Estimation", in CVPR, 2017.

Keywords: 6D Object Pose Estimation

## 概要

- 一枚のRGB-D画像から既知の3Dオブジェクトの6Dポーズを推定するタスクに取り組んでいます。特に、ローカルな特徴量を計算することにフォーカスしている。

## 新規性・差分

- 既存な手法はRANSAC,Hough投票などを用い局所的な推定を統合している方がいい。
- 提案手法は全結合CRFを用いてより数が少ない姿勢を推定している
- グローバルな仮説統合により“一部遮蔽されている物体”に対して、従来手法よりいい結果を出している

## 結果

Method	Our method	Hinterstoisser et al.[10]	Krull et al.[18]	Brachmann et al.[3]
Object	Scores			
Ape	80.7%	<b>81.4%</b>	68.0%	53.1%
Can	88.5%	<b>94.7%</b>	87.9%	79.9%
Cat	<b>57.8%</b>	55.2%	50.6%	28.2%
Driller	<b>94.7%</b>	86.0%	91.2%	82.2%
Duck	74.4%	<b>79.7%</b>	64.7%	64.3%
Eggbox	47.6%	<b>65.5%*</b>	41.5%	9.0%
Glue	<b>73.8%</b>	52.1%	65.3%	44.5%
Hole Puncher	<b>96.3%</b>	95.5%	92.9%	91.6%
Average	<b>76.7%</b>	76.2%	70.3%	56.6%

## 提案手法の流れ

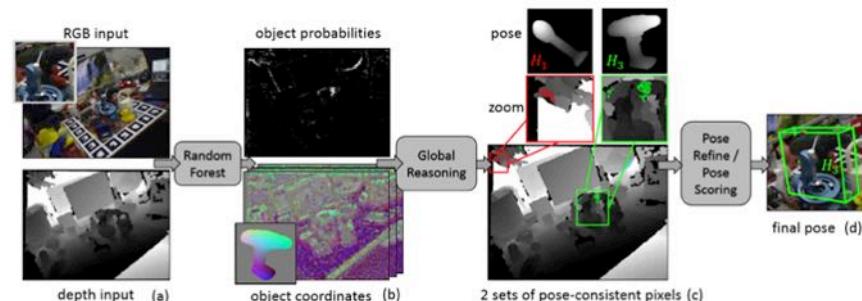


Figure 2. Our pipeline: Given an RGB-D image (a) a random forest provides two predictions: object probabilities and object coordinates (b). In a second stage our novel, fully-connected CRF infers pose-consistent pixel-sets (see zoom) (c). In the last stage, pose hypotheses given by pose-consistent pixels of the CRF are refined and scored by an ICP-variant. The pose with the lowest score is given as output (d).

- ・ランダムフォレストでオブジェクトリスト可能な座標を推定・全結合CRFにより姿勢推定・ICPで姿勢最適化

- ・“一部遮蔽されている物体”的データセットを用いた評価実験でトップな平均推定を達している

## Links

論文 <https://arxiv.org/pdf/1612.02287.pdf>

# [71] Yvain Quéau, Tao Wu, "A Non-Convex Variational Approach to Photometric Stereo Under Inaccurate Lighting", in CVPR, 2017.

Keywords: photometric stereo

## 概要

- ・キャリブレーション済みまたはキャリブレーションされていない測光ステレオ法のいずれかによって得られた、照明が正確ではない場合の測光ステレオ問題をフォーカス。
- ・self-shadow、cast-shadowまたは鏡面などの問題を処理できる。

## 新規性・差分

- ・照明が正確じゃないデータからまずノイズと外れ値を精密にモデリング
- ・提案手法を用いて証明の強度と方向両方を正しく修正できる。
- ・キャリブレーションされていないデータも解決できる。

## Links

[https://pdfs.semanticscholar.org/c741/0792ab6e52ead1586d24cd7f0137f0cbff61.pdf?  
\\_ga=2.20267960.1668793166.1499924432-705425954.1485736521](https://pdfs.semanticscholar.org/c741/0792ab6e52ead1586d24cd7f0137f0cbff61.pdf?_ga=2.20267960.1668793166.1499924432-705425954.1485736521)

## 照明修正結果

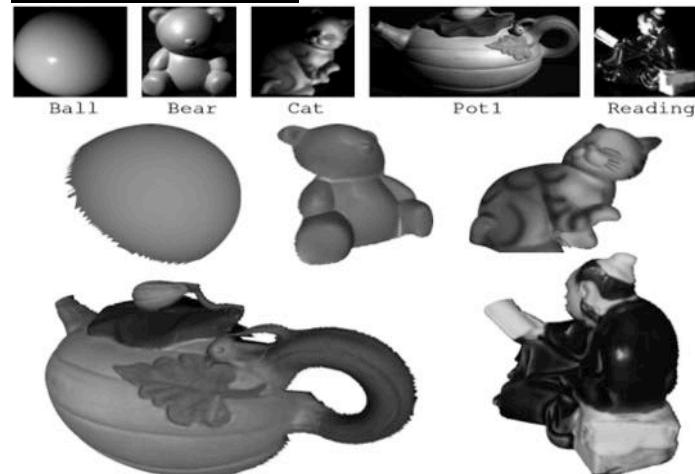


Figure 1. Top: five real-world PS datasets [31], containing self-shadows (all), cast-shadows (all except Ball), specular spikes (Ball and Reading), or broader specular lobes (Bear). Bottom: 3D-models estimated by the proposed method, taking as initial lighting the calibration from [31]. Qualitatively similar results are obtained using uncalibrated PS as initialization, see Figure 7.

## 定量的結果

	Ball	Bear	Cat	Pot1	Reading
ME [1]	6.56	15.29	19.85	16.49	82.37
DM [25]	5.04	9.20	10.62	10.27	24.49
LR [36] + DM [25]	2.66	7.69	8.89	8.94	42.23
Proposed	<b>1.40</b>	<b>6.66</b>	<b>7.59</b>	<b>8.46</b>	<b>20.16</b>

- ・平均角度誤差
- ・提案手法を用いて照明修正を行ったら3次元復元に利用

## 【72】Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, Lizhen Qu, "Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach", in CVPR, 2017.

Keywords: Label Noise, Robustness, Machine Learning

### 概要

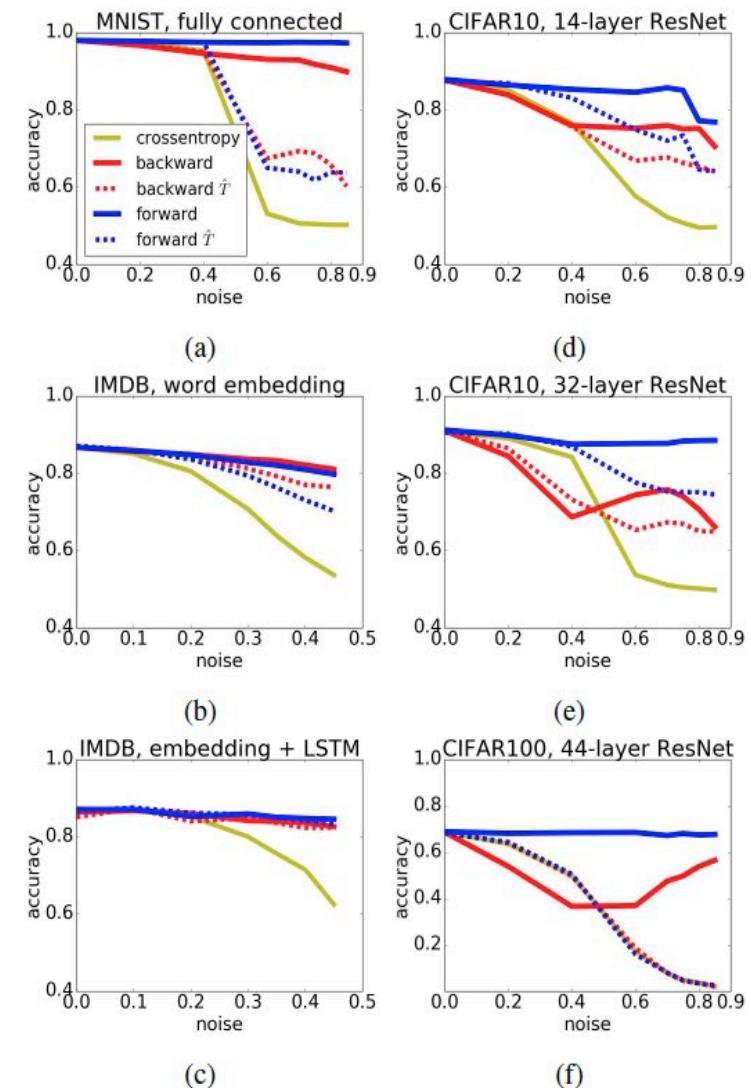
- ラベルにノイズが混ざっている場合でも頑健に学習するための手法を提案。  
データセットが大規模になるに連れて質の高いラベリングを行うことは困難になるため、このような頑健化は重要。従来ノイズに頑健なloss (corrected loss) を計算する手法はあったが、ノイズ発生確率が既知であるという条件があつたため実用的ではなかった。そこで、この研究ではノイズ発生確率を推定する手法とノイズに頑健なlossを組み合わせることで、実用的にノイズに頑健な学習を行う手法を提案。理論的な有効性の証明に加えて、実験的に提案手法の有用性を示した。実験結果では、ノイズが少ない状態での精度を低下させることなく、ノイズが多いときに精度を向上させることができることが示されている。

### 新規性・差分

- ノイズ推定とcorrected lossの組み合わせによりノイズに頑健な学習を実現
- lossを訂正するための2つの手法を提案（従来手法の多クラス拡張など）
- ノイズ推定の手法を多クラスの設定に適用できるように拡張

### Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Patrini\\_Making\\_Deep\\_Neural\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Patrini_Making_Deep_Neural_CVPR_2017_paper.pdf)

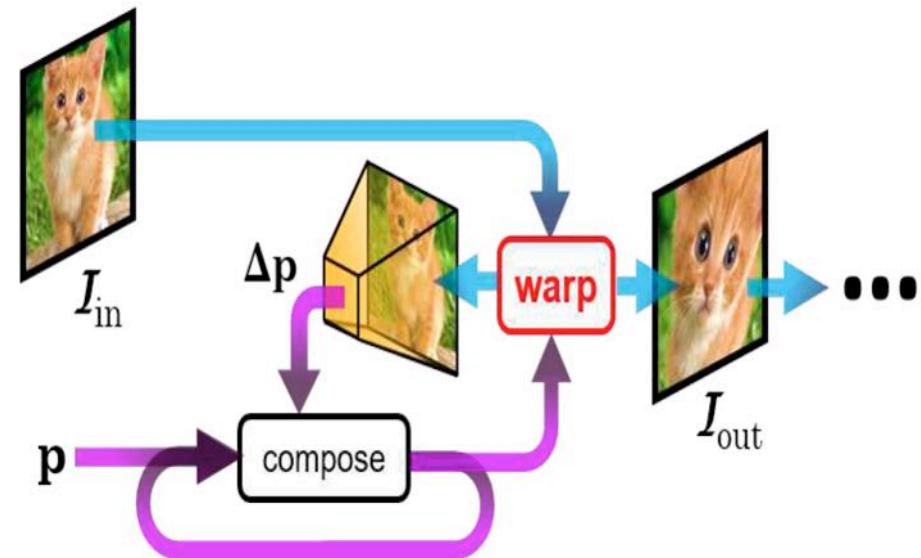


# 【73】Chen-Hsuan Lin, Simon Lucey, "Inverse Compositional Spatial Transformer Networks", in CVPR, 2017.

Keywords: Lucas-Kanade, Spatial Transformer Netowrk, Spatial Alignment

## 概要

・画像をAlignmentしながら対象を識別するSpatial Transformer Networks (STN) を拡張した論文。 STNはCNN内部で画像を warpしながらAlignmentしていくが、毎回直接画像をwarpしてしまうのでboundaryの問題が発生する。この手法で提案するIC-STNでは、画像ではなくgeometric parameterをwarpさせるためこの問題が発生しない(Lucas-Kanade法から着想)。加えて、 LK法を効率的に計算するためのInverse Compositional (IC) アルゴリズムのアイデアも導入している。MNIST, GTSRBデータセットでの実験から有効性を確認。



## 新規性・差分

・STNのアイデアとLKのアイデアをつなげ、 STNの問題を解決

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Lin\\_Inverse\\_Compositional\\_Spatial\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Lin_Inverse_Compositional_Spatial_CVPR_2017_paper.pdf)



# 【74】Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, Hongkai Wen, "VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization", in CVPR Poster, 2017.

Keywords: Camera re-localization , Bidirectional RNN

## 概要

・CNNやRegressionForestsにおいて、単眼画像において6DOFの定位が実現の見込みが示されている(Posenetなど)が、実ケースでは画像のシーケンスが利用可能である。また、そのような学習ベースの既存手法では時間的平滑さを考慮していないため、フレームごとの定位誤差がカメラモーションより大きくなる状況につながる。

この論文では、単眼ビデオクリップの6DOFの定位を実行するためのRecurrentモデルを提案する。20フレームという短いシーケンスだけを考慮することで、姿勢推定は平滑化され、定位誤差は劇的に減る。また、確率的姿勢推定を得る手段について講じる。

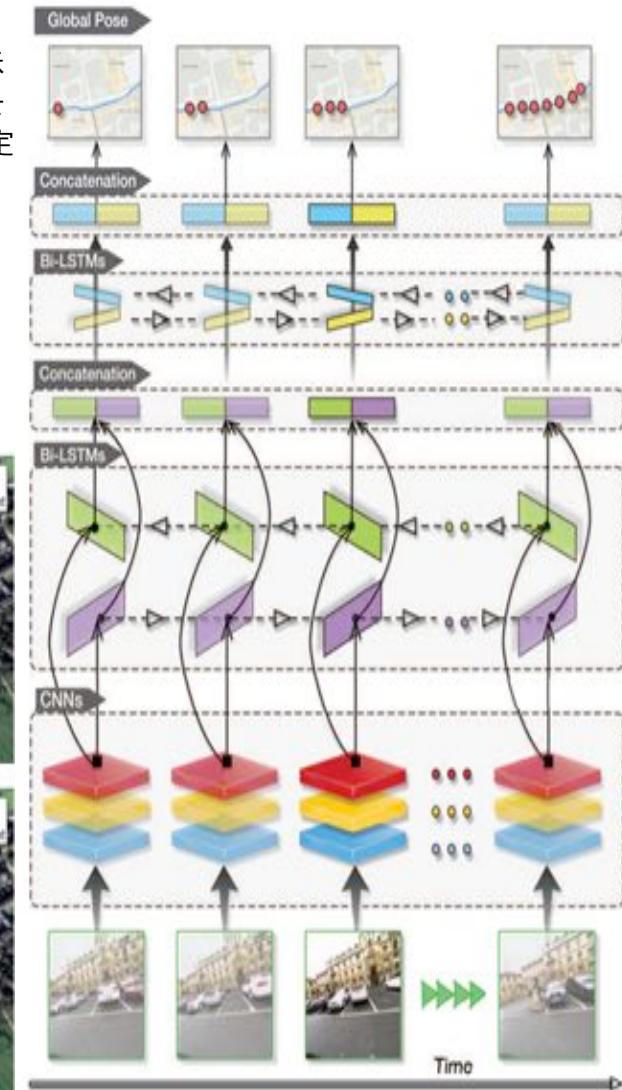
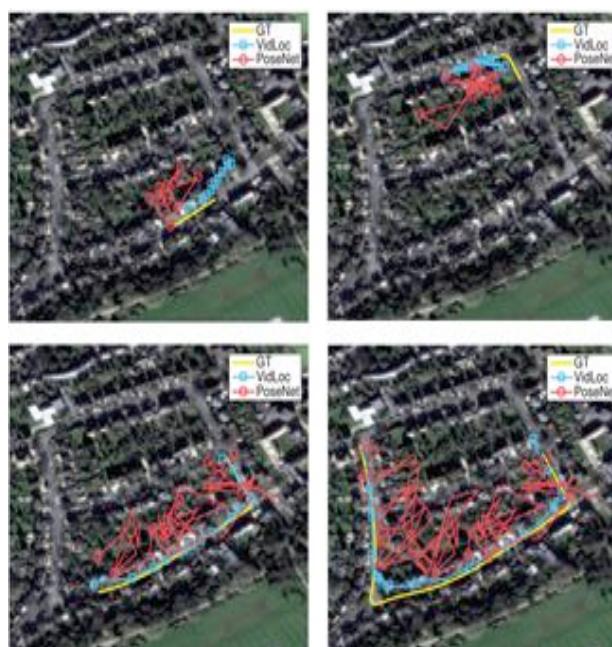
## 新規性・差分

- ・単眼画像からのCNNによるカメラ定位
- ・姿勢推定の共分散の取得
- ・Smoothing Baselineとの比較による、RNN構造による時間情報の理解に対する優位性の確認

## Links

論文

[http://openaccess.thecvf.com/  
content\\_cvpr\\_2017/papers/  
Clark\\_VidLoc\\_A\\_Deep\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Clark_VidLoc_A_Deep_CVPR_2017_paper.pdf)



# [75] Yuliang Liu, Lianwen Jin, "Deep Matching Prior Network: Toward Tighter Multi-Oriented Text Detection", in CVPR Spotlight, 2017.

Keywords: Incidental text area localization, quadrangle window, shared Monte-Carlo, relative regression

## 概要

・あらゆる姿勢、射影歪み、テキストサイズ、色を含む、incidental(付帯的)なシーンテキストの検出は挑戦的タスクである。旧来からの研究では、テキスト領域の定位に矩形だと水平スライディングウィンドウのみ取り扱われ、結果として背景ノイズに弱い、不必要なオーバーラップ、情報欠如といった問題が起きていた。この問題に対し、DMPNet(Deep Matching Prior Network)を提案する。これは、よりタイトなQuadrangle (広義の?四角形)によりテキストを検出するものである。

・まず初めに、広めにオーバーラッピングする領域をラフに覚えておくため、四角形スライディングウィンドウをいくつかの特定の中間Conv層に使う。それから、共有Monte-Carlo法で高速にかつ精度よく多角形領域を得る。次に、テキスト領域のにコンパクトな四角形を確実に推定するため、テキスト位置のRelativeRegressionをする逐次処理を行う。さらに、テキスト位置のRegressionに緩和平滑L<sub>n</sub>ロスを導入する。これは頑健性や安定性の面でL<sub>2</sub>ロスとか平滑L<sub>1</sub>ロスより全体的に良いパ

フォーマンスを示す。

## 新規性・差分

・BoundingBoxの表現力を高めつつ高精度な領域推定を行っている  
・F値70.64%。state-of-the-artは63.76%。

## Links

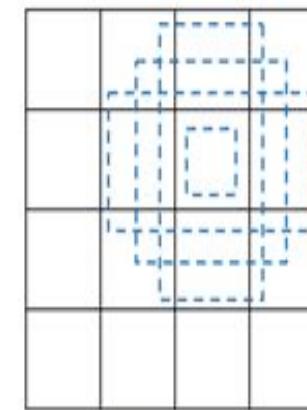
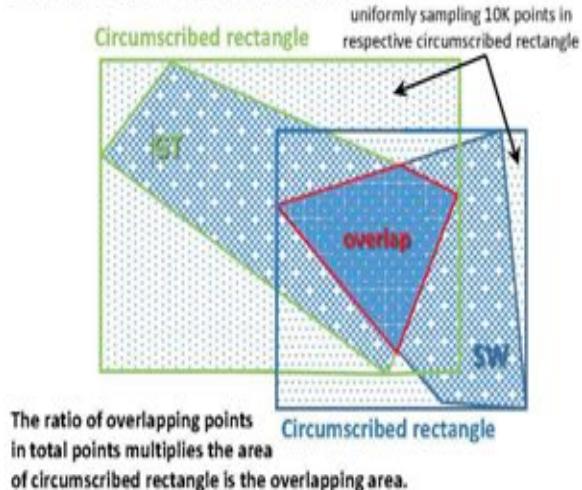
論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Liu\\_Deep\\_Matching\\_Prior\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Liu_Deep_Matching_Prior_CVPR_2017_paper.pdf)

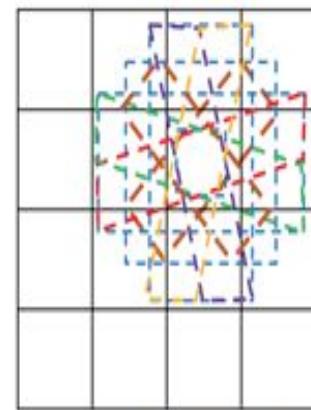


(a) Comparison of recalling scene text.

## Our shared Monte-Carlo method



(b) Horizontal sliding windows.



(c) Proposed quadrilateral sliding windows.

# 【76】Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun, "EAST: An Efficient and Accurate Scene Text Detector", in CVPR, 2017.

Keywords: Scene text detection, Word or text-line level predictions, Multi-channel FCN

## 概要

- 新しい自然画像シーンから文字列（単語やtext line）の位置を検出するフレームワークを提案

## 新規性・差分

- まず文字領域を抽出し、単語領域を統合する従来手法とは違い、End2endで文字列の位置を検出する。検出の手順が少なくなる。
- 高精度且つ高速な検出を実現  
720pの場合では13.2fpsの検出スピードを達成
- 検出した領域の信頼度（Score map）及び二種類の位置（Rotation boxとQuadrangle box）に対してそれぞれLoss関数を設置

$$L = L_s + \lambda_g L_g$$

$L_s$ : Loss for score map  
 $L_g$ : Loss for geometries

## Links

論文: <https://arxiv.org/abs/1704.03155>

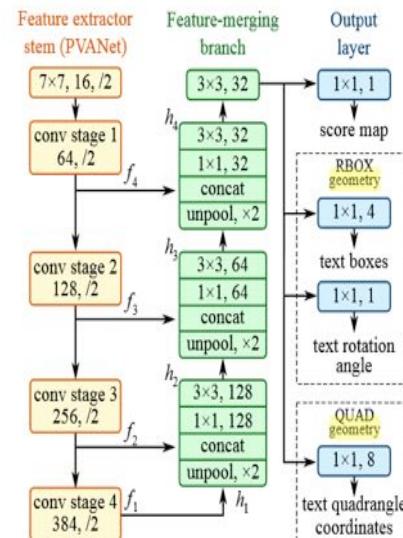
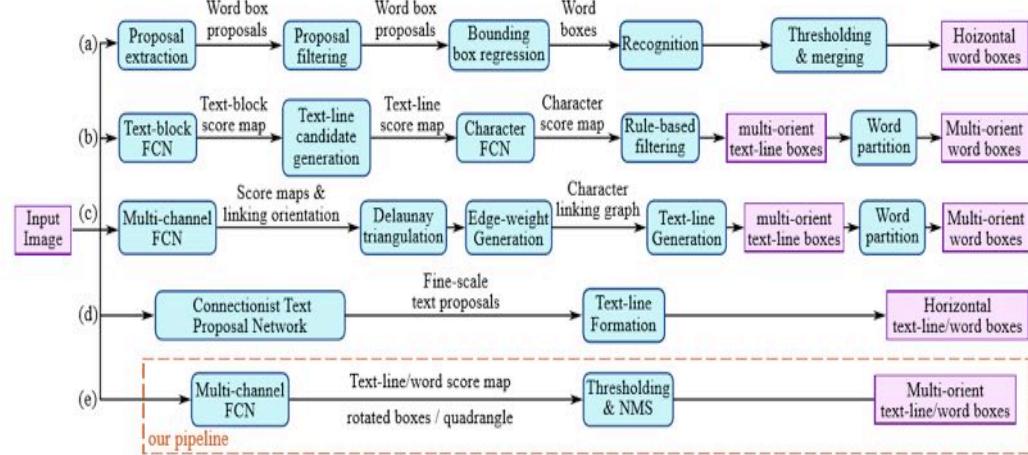


Figure 3. Structure of our text detection FCN.

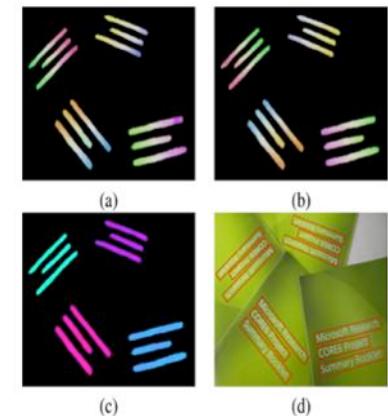


Figure 6. Intermediate results of the proposed algorithm. (a) Estimated geometry map for  $d_1$  and  $d_4$ . (b) Estimated geometry map for  $d_2$  and  $d_3$ . (c) Estimated angle map for text instances. (d) Predicted rotated rectangles of text instances. Maps in (a), (b) and (c) are color-coded to represent variance (for  $d_1, d_2, d_3$  and  $d_4$ ) and invariance (for angle) in a pixel-wise manner. Note that in the geometry maps only the values of foreground pixels are valid.

[77]

Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, Serge Belongie,  
 "Stacked Generative Adversarial Network", in CVPR, 2017

Keywords: GAN, top-down, bottom-up representation discriminator

## 概要

- Stackの構造を用いて高質な画像を自動的に生成するGANを提案

## 新規性・差分

- Top-down stackの構造のCNNにHigh-levelの記述を入力し、中間層の出力を用いて強力なDiscriminatorを訓練する。また、これらのDiscriminatorを用いて、Generatorを各階層のGeneratorを調整できる
  - 条件付きのLossを設置
  - 単純にノイズを入力してすべてのバリエーションを表すGANとは異なり、SGANはそれらのバリエーションを複数のレベルに分解し、Top-downの生成プロセスを用いて不確実性を徐々に解決します、

## Links

論文: <https://arxiv.org/pdf/1612.04357.pdf>

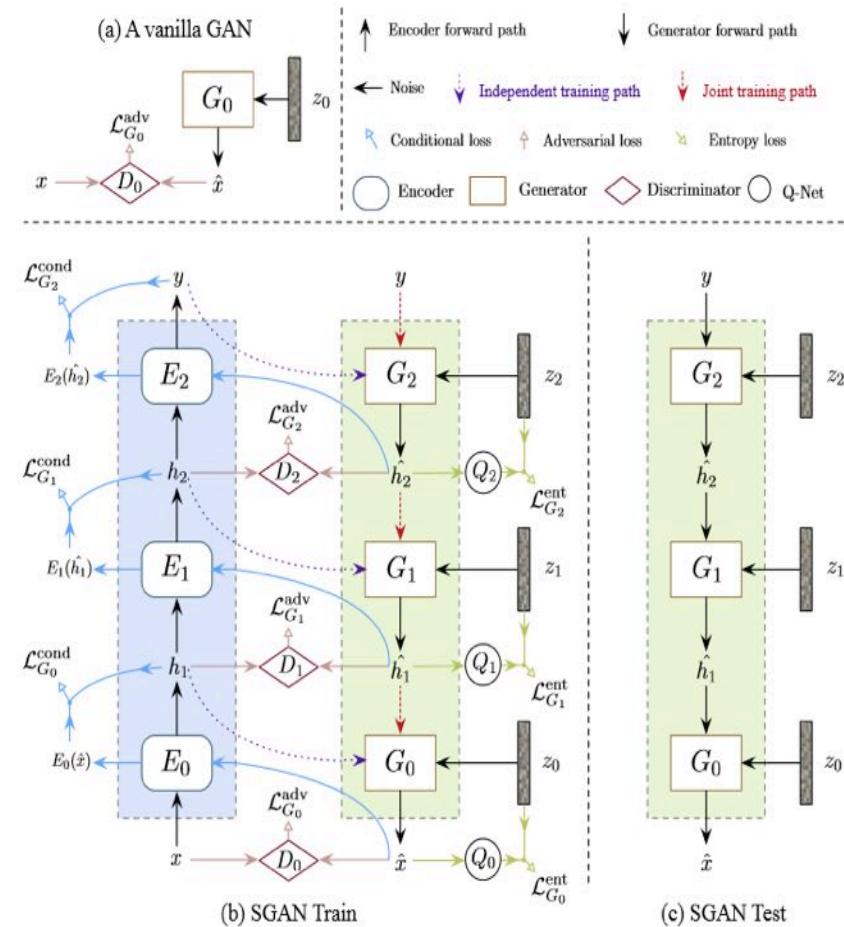


Figure 1: An overview of SGAN. (a) The original GAN in [17]. (b) The workflow of training SGAN, where each generator  $G_i$  tries to generate plausible features that can fool the corresponding representation discriminator  $D_i$ . Each generator receives conditional input from encoders in the independent training stage, and from the upper generators in the joint training stage. (c) New images can be sampled from SGAN (during test time) by feeding random noise to each generator  $G_i$ .

# 【78】Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang, "Generative Face Completion", in CVPR, 2017

Keywords: Face completion, Deep generative model, Reconstruction loss, Adversarial loss

## 概要

・人の顔が遮断される部分を自動的に補完するDeep generative modelを提案。大量な顔画像を処理できる同時に、高質に補完することもできる。

## 新規性・差分

- ・ニューラルネットワークを用いて直接欠損部分を予測して補完する。
- ・欠損部分だけ、全画像またはSegmentation画像に対して、三つのDiscriminatorをトレーニングする。
- ・三種類のLoss (reconstruction loss, two adversarial losses, semantic parsing loss) を用いてモデルをトレーニングする。

## Links

論文: <https://arxiv.org/pdf/1704.05838.pdf>

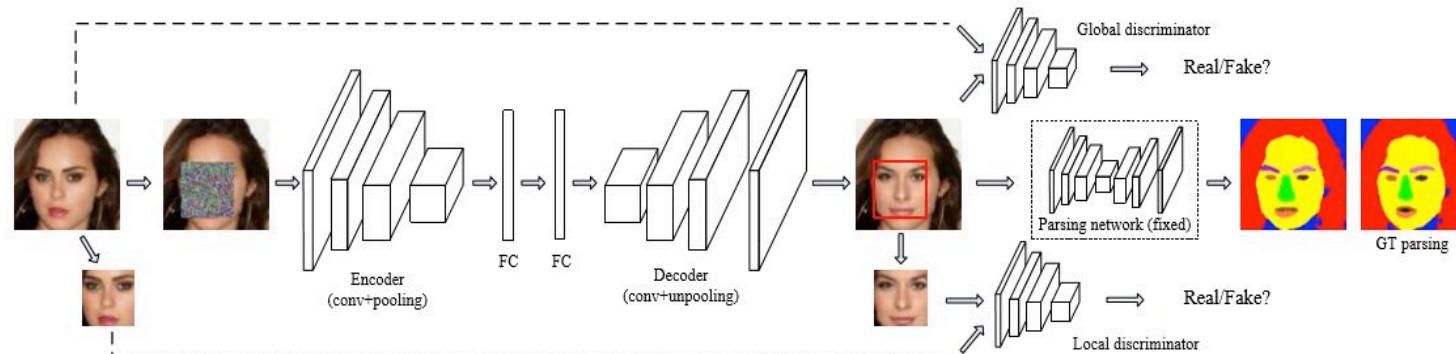


Figure 2. Network architecture. It consists of one generator, two discriminators and a parsing network. The generator takes the masked image as input and outputs the generated image. We replace pixels in the non-mask region of the generated image with original pixels. Two discriminators are learned to distinguish the synthesized contents in the mask and whole generated image as real and fake. The parsing network, which is a pretrained model and remains fixed, is to further ensure the new generated contents more photo-realistic and encourage consistency between new and old pixels. Note that only the generator is needed during the testing.

# 【79】Wei Li, Farnaz Abtahi, Zhigang Zhu, "Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing", in CVPR, 2017

Keywords: Action Unit, LSTM-based temporal fusing, Multi-label learning

## 概要

・顔での些細な変化を検出するために、CNN+LSTMの構造を提案した。人の顔にAction Unitを表記して、顔が動いているときにどのUnitが動いているかを検出できる

## 新規性・差分

- ・ROI Netsを用いて、各Unitの局所領域をそれぞれ特徴を抽出することが可能
- ・Multi-label learningを用いて、各局所領域の間の関係や局所領域と全体の関係を学習することが可能
- ・多層のLSTMを構築し、時系列の情報を考えて、より高い検出精度を達成できます。

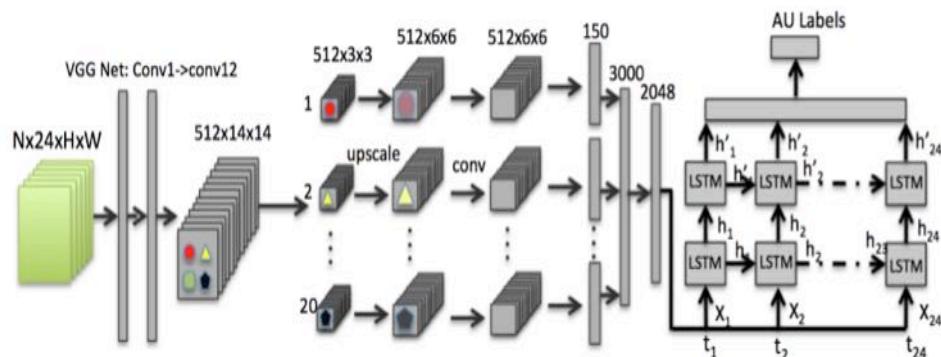


Figure 1. Framework of the proposed neural network with VGG Net, ROI Nets and LSTM Net

## Links

論文: <https://arxiv.org/pdf/1704.05838.pdf>

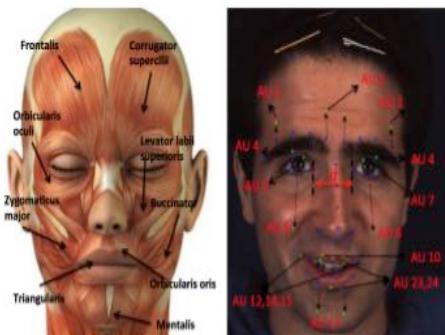


Figure 2. ROI center selection based on muscles and landmarks on one BP4D



Figure 6. Comparison of single and multi-label learning on BP4D

- [80] Xaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers, "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases", in CVPR Spotlight, 2017.

Keywords: Large Scale Database, chest X-ray image, natural language processing

## 概要

- 8種の疾患に対する108,948枚の胸部X線画像からなる病院規模の大規模データベースを作成した。医者のカルテから診断名と症状を自然言語処理により抽出し、アノテーションした。通常の自然言語処理では捉えられなかった、カルテ中の疾患の“否定・不確実”をいくつかの工夫により捉えられるようにした。この工夫により、既存データベースとの比較によりFalse-Positiveが劇的に改善したことを確認した。また、BoundingBox付きのサブセット（各疾患200、計1600インスタンス）を用意した。そして、統合DCNNによる弱教師付き疾患部位定位を行い、その効果を検証した。

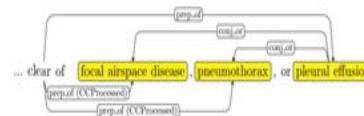


Figure 3. The dependency graph of text: "clear of focal airspace disease, pneumothorax, or pleural effusion".

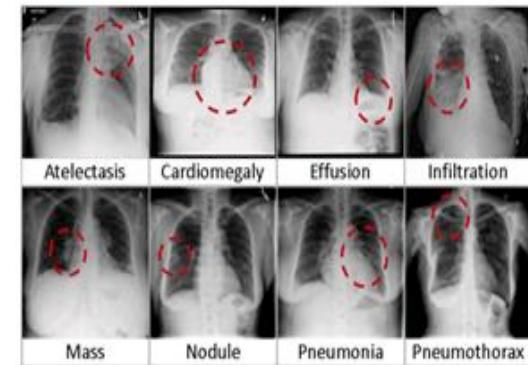


Figure 1. Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis.

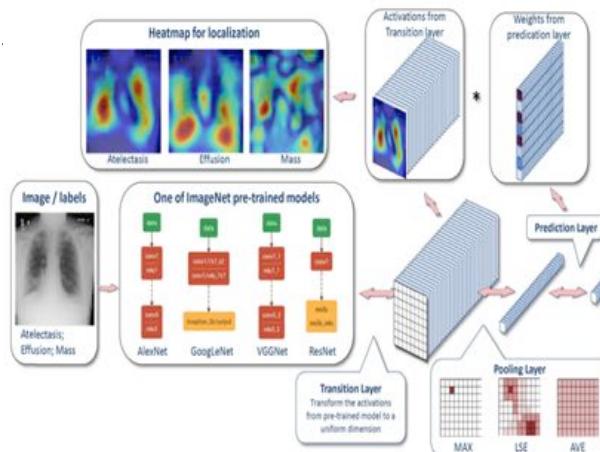
## 新規性・差分

- 今までにない大規模データベースの作成
- ラベル付けをきれいにすると性能が上がる
- 実際にDNNで効果を検証

## Links

論文

[http://openaccess.thecvf.com/  
content\\_cvpr\\_2017/papers/Wang\\_ChestX-ray8\\_Hospital-Scale\\_Chest\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf)



Disease	MetaMap			Our Method		
	P /	R /	F	P /	R /	F
Atelectasis	0.95	/ 0.95	/ 0.95	0.99	/ 0.85	/ 0.91
Cardiomegaly	0.99	/ 0.83	/ 0.90	1.00	/ 0.79	/ 0.88
Effusion	0.74	/ 0.90	/ 0.81	0.93	/ 0.82	/ 0.87
Infiltration	0.25	/ 0.98	/ 0.39	0.74	/ 0.87	/ 0.80
Mass	0.59	/ 0.67	/ 0.62	0.75	/ 0.40	/ 0.52
Nodule	0.95	/ 0.65	/ 0.77	0.96	/ 0.62	/ 0.75
Normal	0.93	/ 0.90	/ 0.91	0.87	/ 0.99	/ 0.93
Pneumonia	0.58	/ 0.93	/ 0.71	0.66	/ 0.93	/ 0.77
Pneumothorax	0.32	/ 0.82	/ 0.46	0.90	/ 0.82	/ 0.86
Total	0.84	/ 0.88	/ 0.86	0.90	/ 0.91	/ 0.90

Table 2. Evaluation of image labeling results on OpenI dataset. Performance is reported using P, R, F1-score.

[81]

Pavel Tokmakov, Karteek Alahari, Cordelia Shmid, "Learning Motion Patterns in Videos", in CVPR, 2017.

Keywords: Moving object, Segmentation

## 概要

- 動的な物体をシーン中から切り出す問題。直感的にはオプティカルフロー空間において学習の結果、セグメンテーションを実行する。Synthetic VideoからFully Convolutional Network (FCN)によりモーションセグメンテーションを実行。このEncoder-Decorder構造によるネットワークは初期の段階ではCoarseなオプティカルフロー画像を獲得し、繰り返しにより高解像なモーションを学習することができる。

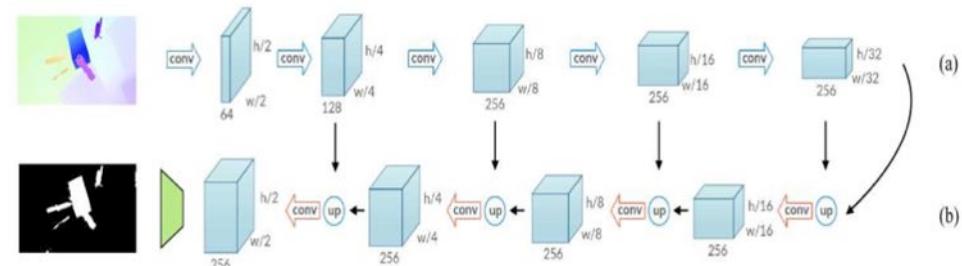
## 新規性・差分

- 提案手法は動的シーンからセグメンテーションを行うタスクであるDAVIS Datasetにて今までのもっとも高い精度を出し、従来手法と比較すると5.6%も精度を向上させた。同時にBekeley Motion Segmentation DatabaseにおいてもState-of-the-artを達成。

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Tokmakov\\_Learning\\_Motion\\_Patterns\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Tokmakov_Learning_Motion_Patterns_CVPR_2017_paper.pdf)



Measure	NLC [9]	CVOS [34]	TRC [10]	MSG [5]	KEY [20]	SAL [39]	FST [27]	PCM [3]	Ours
$\mathcal{J}$	Mean	64.1	51.4	50.1	54.3	56.9	42.6	57.5	45.5
	Recall	73.1	58.1	56.0	63.6	67.1	38.6	65.2	44.3
	Decay	8.6	12.7	5.0	<b>2.8</b>	7.5	8.4	4.4	11.8
$\mathcal{F}$	Mean	59.3	49.0	47.8	52.5	50.3	38.3	53.6	46.1
	Recall	65.8	57.8	51.9	61.3	53.4	26.4	57.9	43.7
	Decay	8.6	13.8	6.6	<b>5.7</b>	7.9	7.2	6.5	10.7
$\mathcal{T}$	Mean	35.6	24.3	32.7	25.0	<b>19.0</b>	60.0	27.6	51.3
									68.6

Table 3. Comparison to state-of-the-art methods on DAVIS with intersection over union ( $\mathcal{J}$ ), F-measure ( $\mathcal{F}$ ), and temporal stability ( $\mathcal{T}$ ).

Measure	CUT [17]	FST [27]	TRC [10]	MTM [42]	CMS [24]	PCM [3]	MP+Obj	MP+Obj + FST [27]
$\mathcal{F}$	73.0	64.1	72.8	66.0	62.5	<b>78.2</b>	71.8	<b>78.1</b>

Table 4. Comparison to state-of-the-art methods on the subset of BMS-26 used in [3] with F-measure. 'MP+Obj' is MP-Net with objectness.

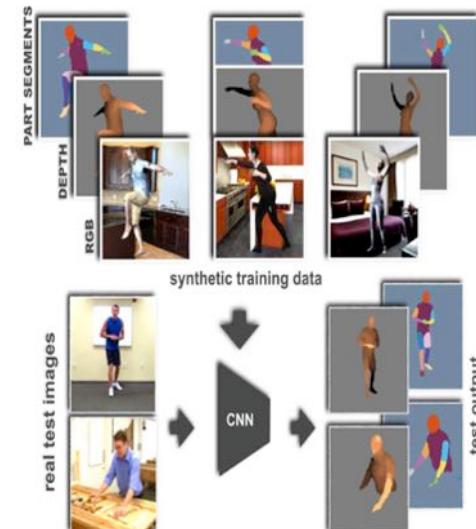
[82]

Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, "Learning from Synthetic Humans", in CVPR, 2017.

Keywords: Synthetic Data, Pose Estimation, SURREAL dataset

## 概要

- Synthetic Data（合成データ）を用いた人物姿勢推定の研究。2Dの人物姿勢推定では大規模な画像データと対応するアノテーションデータが手に入ったが、3Dにおいてはあらかじめ合成データを生成して2Dデータと対応付ける方がデータ確保の面からも有利であるということが判明した。データはモーションキャプチャから取得した3D姿勢、部位ごとに分割したセグメント画像、デプス画像、静的な背景（LSUN datasetから40K枚ランダム抽出）に埋め込んだRGBの人物画像などから構成される。このSURREAL datasetは6,536,752枚の画像やその3次元情報がペアとして付与されている。



## 新規性・差分

- 合成データ（提案）、リアルデータを混ぜたデータセットで学習したモデルがHuman3.6M datasetにてもっとも高い精度を叩き出した。これは人物のデプス推定タスクにおいても効果的であることが判明した。

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Varol\\_Learning\\_From\\_Synthetic\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Varol_Learning_From_Synthetic_CVPR_2017_paper.pdf)

プロジェクト <https://github.com/gulvarol/surreal>

動画 <https://www.youtube.com/watch?v=SJ0vw6CzS7U>



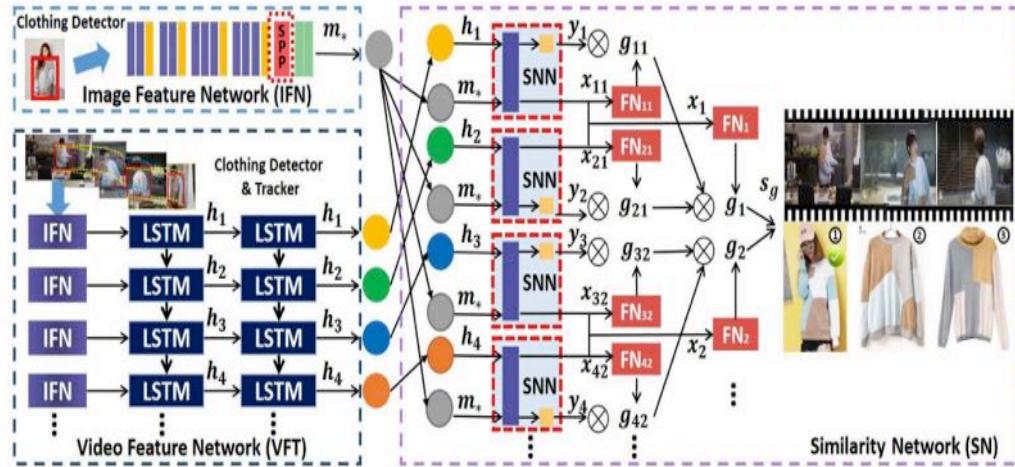
Training data	Head IOU	Torso IOU	Legs <sub>up</sub> IOU	mean IOU	mean Acc.
Real+Pascal[24]	-	-	-	64.10	81.78
Real	58.44	24.92	30.15	28.77	38.02
Synth	73.20	65.55	39.41	40.10	51.88
Synth+Real	72.88	80.76	65.41	59.58	78.14
Synth+Real+up	<b>85.09</b>	<b>87.91</b>	<b>77.00</b>	<b>68.84</b>	<b>83.37</b>

# 【83】Zhi-Qi Cheng , Xiao Wu , Yang Liu , Xian-Sheng Hua, "Video2Shop: Exact Matching Clothes in Videos to Online Shopping Images", in CVPR, 2017.poster

Keywords: online shopping、video、Clothing retrieval、matching

## 概要

- ・ビデオに出てくる服装を検出し、同じようなものをオンラインショッピングで探す手法を提案する。
- ・Image Feature Network (IFN) を用いて、任意サイズの服装画像から特徴を抽出する。
- ・各フレームから服装の特徴を抽出した上で、lstmで時系列データをモデリングする。
- ・ビデオ服装×オンラインショッピング服装画像の matching。



## 新規性・差分

- ・従来のstreet-to-shop服装matchingに対して、この論文はビデオに出てくる服装とオンラインショッピングの服装のmatchingを行いました。video-to-shop
- ・cross-domain sources 画像×ビデオ

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Cheng\\_Video2Shop\\_Exact\\_Matching\\_CVPR\\_2017.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Cheng_Video2Shop_Exact_Matching_CVPR_2017.pdf)

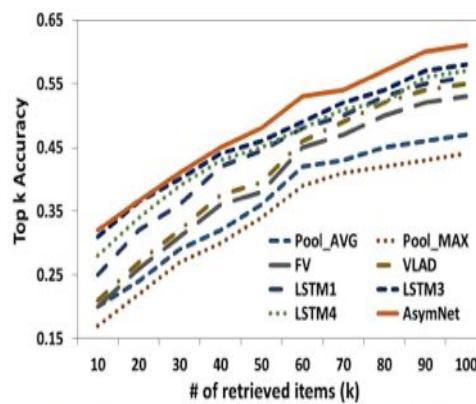


Figure 3. Performance Comparison of Representation Networks



# 【84】 Jianwen Xie, Song-Chun Zhu, Ying Nian Wu, "Synthesizing Dynamic Patterns by Spatial-Temporal Generative ConvNet", in CVPR, 2017.

Keywords: Spatio-Temporal Generative ConvNet

## 概要

- ・ショートクリップレベルの動画像の生成をConvNetにて行うための空間、時間フィルタの最適化。下記(1)式は再帰的畳み込みを示し、 $F * I$ は時系列フィルタと画像列の畳み込みを示す。(2)式では画像列  $I = (I(x, t))$  から  $q(I)$  により画像を生成する(?)モデルである。

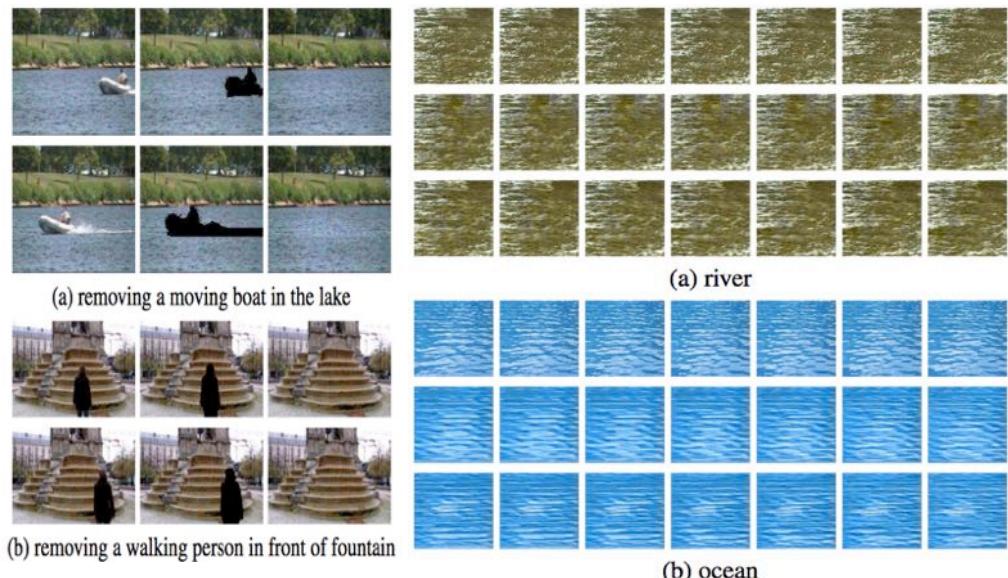
## 新規性・差分

- ・画像列から時系列フィルタを学習し、ショートクリップを生成する研究である。
- ・静止画によるテクスチャから動的な画像生成、(炎など)定常性がない動作も生成、動的なシーンからのインペインティングも可能とした

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Xie\\_Synthesizing\\_Dynamic\\_Patterns\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Xie_Synthesizing_Dynamic_Patterns_CVPR_2017_paper.pdf)



$$\begin{aligned} [F_k^{(l)} * \mathbf{I}](x, t) &= h \left( \sum_{i=1}^{N_l-1} \sum_{(y,s) \in \mathcal{S}_l} w_{i,y,s}^{(l,k)} \right. \\ &\quad \times \left. [F_i^{(l-1)} * \mathbf{I}](x+y, t+s) + b_{l,k} \right), \end{aligned} \quad (1)$$

$$p(\mathbf{I}; w) = \frac{1}{Z(w)} \exp [f(\mathbf{I}; w)] q(\mathbf{I}), \quad (2)$$

# 【85】Siavash Gorji, James J. Clark, "Attentional Push: A Deep Convolutional Network for Augmenting Image Saliency with Shared Attention Modeling in Social Scenes", in CVPR Spotlight, 2017.

Keywords: Shared attention, augmented saliency

## 概要

・視覚的顕著性が低い場所でも注目を集める場所があるが、何かの高次的な特徴が含まれているものと考えられる。その理解には、人間の注目を集める（Pull attention）場所だけでなく、人が注視する（Push attention）場所にも目を向けることが重要であると考えられる。この論文では、共同注視モデルに基づく新たな視覚的注意の追跡手法を提案する。撮影者と画像中の人物が共同作業を行っていると考え、画像中の人物の情報から生成されるAttentionalPushMapを統合し、共同注視の拡張SaliencyMapを生成する。

・具体的には、AttentionalPushMapを生成するCNNを構築する。その後、その事前学習されたCNNを内部に組み込み、Saliencyと統合してAugmentedSaliencyを出力するCNNを構築する。

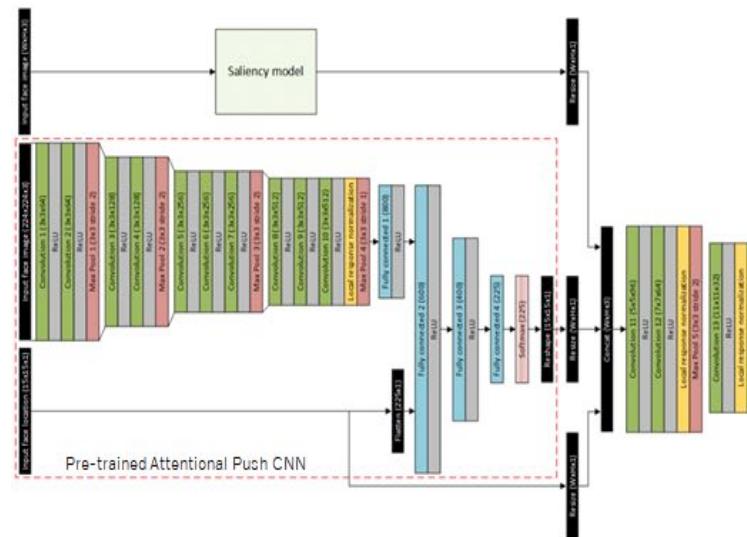
## 新規性・差分

・社会性の考慮という前衛的課題に対する新提案

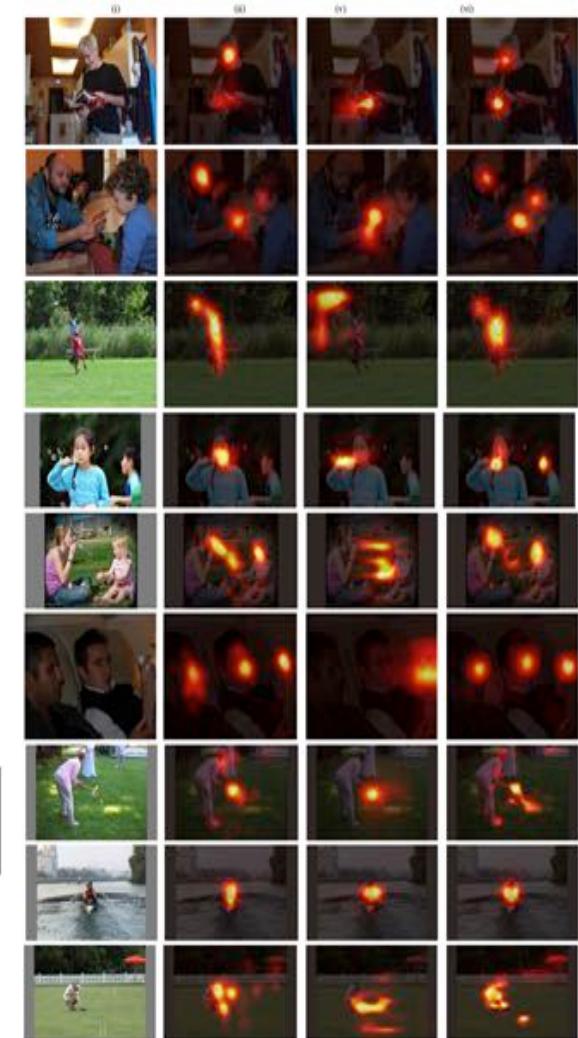
## Links

論文

[http://openaccess.thecvf.com/  
content\\_cvpr\\_2017/papers/  
Gorji\\_Attentional\\_Push\\_A\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Gorji_Attentional_Push_A_CVPR_2017_paper.pdf)



(Left) Attentional Push  
(center) Attentional Pull  
(Right) Attentional Saliency



# 【86】David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, Léon Bottou, "Discovering Causal Signals in Images", in CVPR, 2017

Keywords: Causal dispositions, Object features and Context features

## 概要

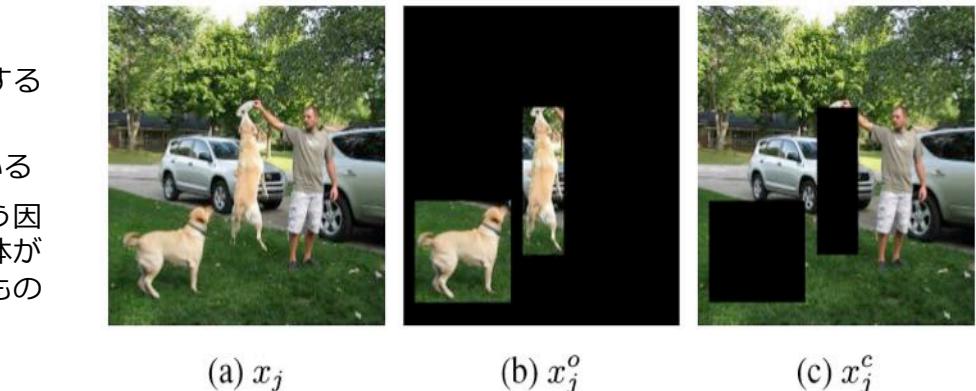
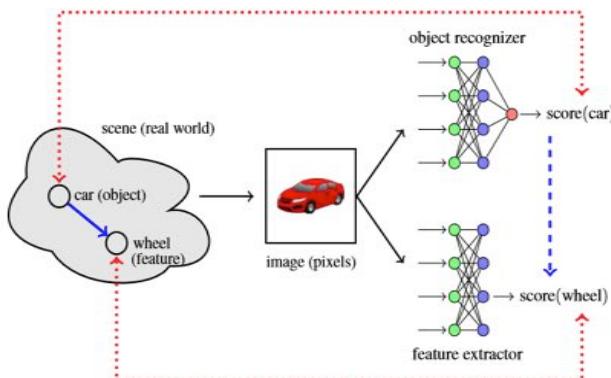
- ・シーンの中に物体特徴とコンテンツ特徴の因果関係を反映するもの (Signal) を探す研究
- ・実験結果としてはその因果関係を反映するものは存在している
- ・シーンのコンテンツであるこそ、ある物体が存在するという因果関係を反映する明らかなものが存在している。逆にある物体が存在すること、シーンがそであるという因果関係を反映するものがあまり存在しない (nonexistent or much weaker)

## 新規性・差分

- ・新しい因果関係を探せるNCCを提案し、関連されるproxy variablesのペアの共同分布に適用する。
- ・proxy variables → 画像ピクセルにCNNを適用することによって計算されるもの

## Links

論文: <https://arxiv.org/abs/1605.08179>



- ・物体特徴(Object features): 目標の物体を反映できる特徴
- ・コンテンツ特徴(Context features): 目標の物体以外のシーンの内容を反映できる特徴

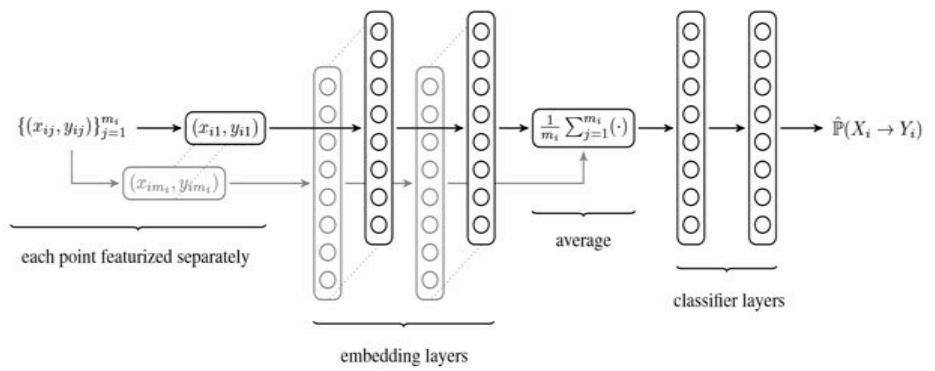


Figure 4: Scheme of the Neural Causation Coefficient (NCC) architecture.

# [87] Weihua Chen et al., "Beyond triplet loss: a deep quadruplet network for person re-identification", in CVPR, 2017. (spotlight)

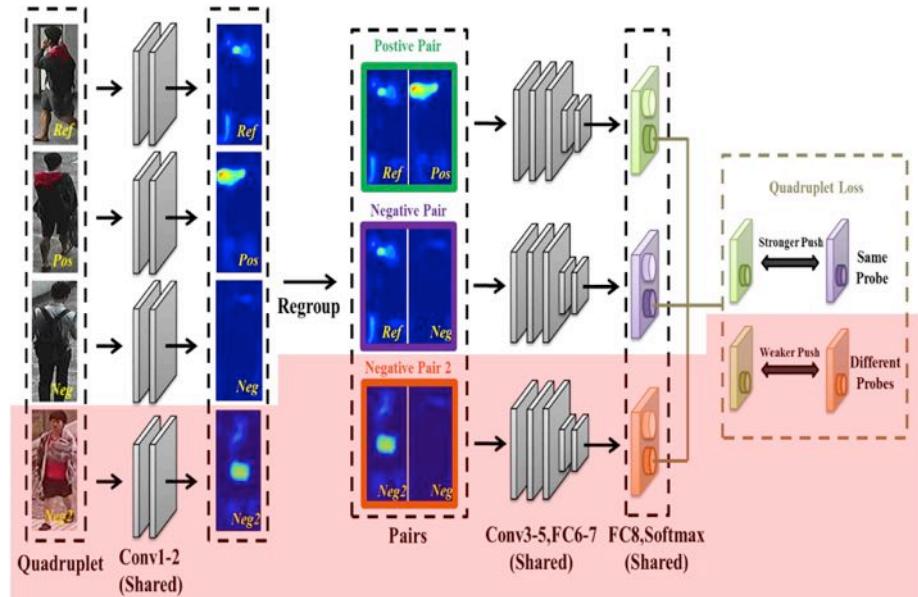
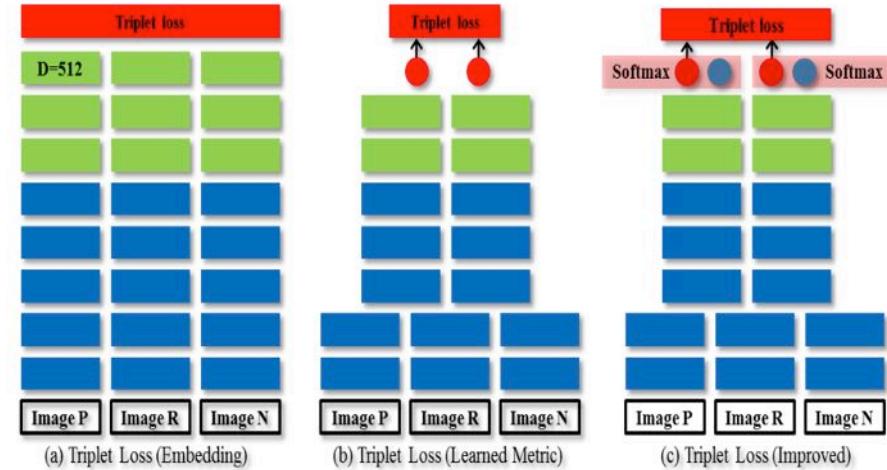
Keywords: Triplet, Quadruplet

## 概要

- TripletならぬQuadruplet-lossを提案した。通常、Triplet学習ではReferenceとなる画像とPositive（類似する画像）、Negative（非類似の画像）を用いてPと距離が近くなるよう、Nと距離が離れるように学習する。それに対してQuadruplet学習ではさらに非類似画像Negative2を増加させて距離の精度を高くする。

## 新規性・差分

- Triplet学習にさらにネガティブ画像を増やすことで距離学習の性能を高めるQuadruplet-lossを提案。本研究ではPerson Re-identificationに応用した。



## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Chen\\_Beyond\\_Triplet\\_Loss\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Chen_Beyond_Triplet_Loss_CVPR_2017_paper.pdf)

# 【88】Jonthan Krause, Justin Johnson, Ranjay Krishna, Li Fei-Fei, "A Hierarchical Approach for Generating Descriptive Image Paragraphs", in CVPR, 2017. (spotlight)

Keywords: Paragraph Generation

## 概要

- ・画像説明文だと画像を説明することに関して表現力の面で不足してしまうため、文章(Paragraph)を生成する研究を提案した(右図)。提案のアーキテクチャは右下図に示すとおりであり、画像を入力として候補領域(RPN)抽出、候補領域ごとの特徴を記述、Pooled Vectorを取り出す。さらに、文章(や単語レベル)を生成するためのRNNを再帰的に処理する。

## 新規性・差分

- ・画像説明文ではなく、パラグラフにすることで表現力を向上した。
- ・物体検出に習い候補領域やその特徴記述を行い、センテンスレベル、単語レベルのRNNを実行してパラグラフを生成した。

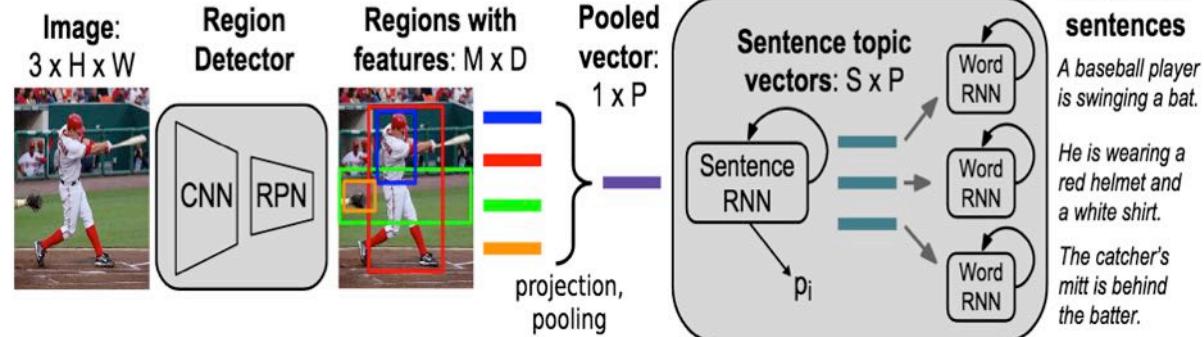
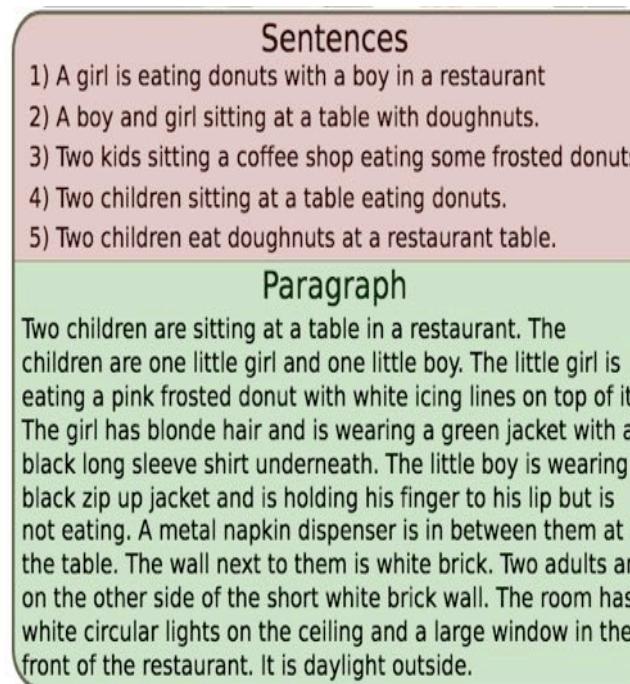
## Links

論文

[http://openaccess.thecvf.com/  
content\\_cvpr\\_2017/papers/  
Krause\\_A\\_Hierarchical\\_Approach\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Krause_A_Hierarchical_Approach_CVPR_2017_paper.pdf)

プロジェクト

<https://github.com/chenxinpeng/im2p>

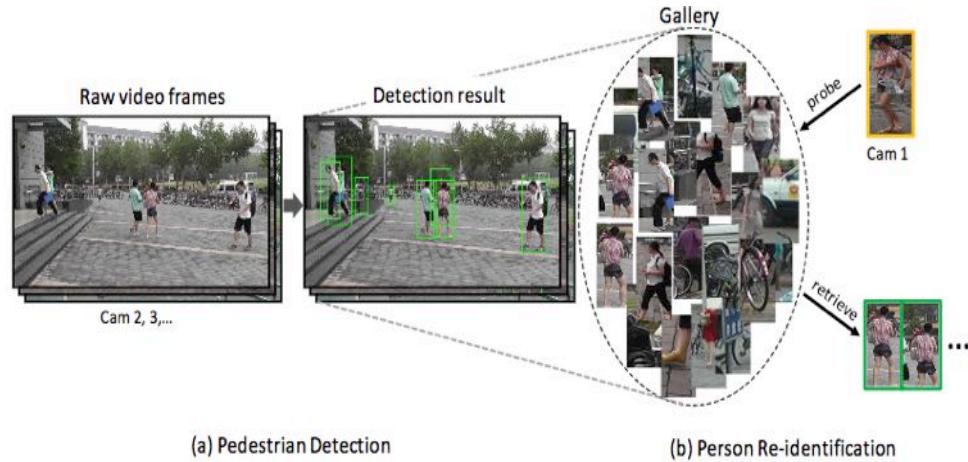


# [89] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, Qi Tian, "Person Re-identification in the Wild", in CVPR, 2017. (spotlight)

Keywords: Joint Detection and Re-identification

## 概要

- 監視カメラの環境にて人物再同定 (Person Re-identification) を行うために、人物検出と人物再同定を同時に行った。932の人物の対応が含まれる11,816枚の画像を含んだデータPRW (Person Re-identification in the Wild)を用いた。



## 新規性・差分

- 認識結果は右の表に示すとおりである。Detectorと組み合わせてRe-IDの認識を行った方が良好な結果を得ることができる。

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Zheng\\_Person\\_Re-Identification\\_in\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zheng_Person_Re-Identification_in_CVPR_2017_paper.pdf)



(a) Pedestrian Detection

(b) Person Re-identification

Detector	Recognizer	#detection=3			#detection=5			#detection=10		
		mAP	r1	r20	mAP	r1	r20	mAP	r1	r20
DPM	BOW	8.9	30.4	58.3	9.7	31.1	58.6	9.6	30.5	57.7
DPM	IDE	12.7	37.2	72.2	13.7	36.9	72.1	13.7	36.6	70.8
DPM	IDE <sub>det</sub>	17.2	45.9	77.9	18.8	45.9	77.4	19.2	45.7	76.0
DPM-Alex	SDALF+Kiss.	12.0	32.6	63.8	13.0	32.5	63.4	12.4	31.8	
DPM-Alex	LOMO+XQ.	13.4	34.9	66.5	13.0	34.1	64.0	12.4	33.6	62.5
DPM-Alex	HistLBP+DNS	14.1	36.8	70.0	13.6	35.9	67.8	12.7	35.0	65.7
DPM-Alex	IDE	15.1	38.8	74.1	14.8	37.6	71.4	14.1	36.9	69.8
DPM-Alex	IDE <sub>det</sub>	<b>20.2</b>	<b>48.2</b>	<b>78.1</b>	20.3	47.4	77.1	19.9	47.2	76.4
DPM-Alex	IDE <sub>det</sub> +CWS	20.0	<b>48.2</b>	<b>78.8</b>	<b>20.5</b>	<b>48.3</b>	<b>78.8</b>	<b>20.5</b>	<b>48.3</b>	<b>78.8</b>
ACF	LOMO+XQ.	10.5	31.5	61.6	10.5	30.9	59.5	9.7	29.7	57.4
ACF	gBiCov+Kiss.	9.8	31.1	60.1	9.9	30.3	58.3	9.0	29.0	55.9
ACF	IDE <sub>det</sub>	16.6	44.8	75.9	17.5	43.8	76.0	17.0	42.9	74.5
ACF-Res	IDE	12.4	35.0	70.4	12.5	33.8	68.6	11.5	33.0	66.7
ACF-Alex	LOMO+XQ.	10.5	31.8	60.7	10.3	30.6	59.4	9.5	29.6	57.1
ACF-Alex	IDE <sub>det</sub>	17.0	45.2	76.6	17.5	43.6	75.1	16.6	42.7	73.7
ACF-Alex	IDE <sub>det</sub> +CWS	17.0	45.2	76.8	17.8	45.2	76.8	17.8	45.2	76.8
LDCF	BoW	8.2	30.1	56.9	9.1	29.8	57.0	8.3	28.3	55.3
LDCF	LOMO+XQ.	11.2	31.6	62.9	11.0	31.1	62.2	10.1	29.6	58.6
LDCF	gBiCov+Kiss.	9.5	30.7	58.8	9.6	30.1	58.4	8.8	28.7	56.7
LDCF	IDE	12.7	35.3	70.1	13.4	34.4	71.1	12.2	33.1	68.0
LDCF	IDE <sub>det</sub>	17.5	45.3	76.2	18.3	44.6	75.6	17.7	43.8	74.3
LDCF	IDE <sub>det</sub> +CWS	17.5	45.5	76.3	18.3	45.5	76.4	18.3	45.5	76.4

# 【90】Hao Jiang, Kristen Grauman, "Seeing Invisible Poses: Estimating 3D Body Pose From Egocentric Video", in CVPR, 2017.

Keywords: pose estimation, egocentric view

## 概要

- ・胸部装着カメラから、カメラには直接映らない、装着者の全身姿勢（3D関節位置）の推定をする。
- ・ポーズインスタンスは学習データのk-meansクラスタリングにより定める。カメラの動きによるオプティカルフローからランダムフォレストによって各姿勢に対する確率を求める。また、静止中の特徴も考慮するため、画像からCNNで立ち座りを判別する。これらを組み合わせてコスト関数Uを構成する。次に、ポーズ遷移を推定する。遷移時間T、ポーズ遷移の速さV、静止ペナルティSの各コストを結合し、 $U+T+V+S$ を最小化するようなポーズ遷移を求める。  
装着者と環境が異なる実験ケースでもstate-of-the-artの性能を達成。



## 新規性・差分

- ・胸部装着カメラ
- ・state-of-the-artな装着者のポーズ推定、ポーズ遷移推定

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Jiang\\_Seeing\\_Invisible\\_Poses\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Jiang_Seeing_Invisible_Poses_CVPR_2017_paper.pdf)

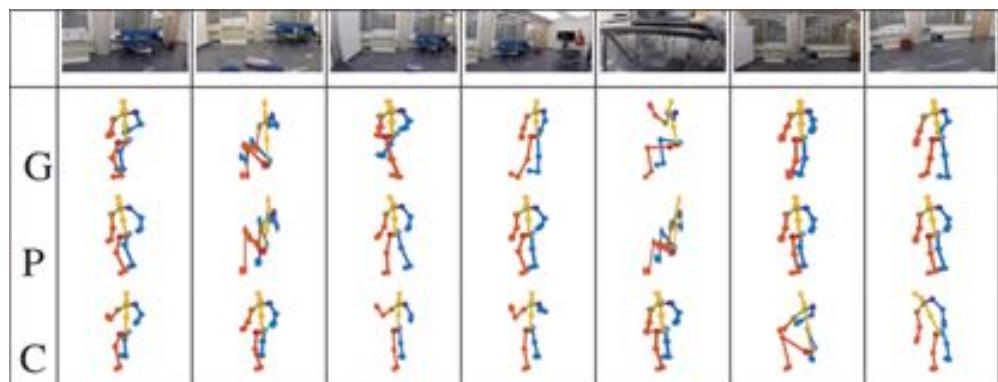


Figure 4. Comparison with the DeepPose [5] method retrained for our task. G: ground truth. P: proposed method. C: CNN-Regression baseline.

# 【91】Baoguang Shi, Xiang Bai, Serge Belongie, "Detecting Oriented Text in Natural Images by Linking Segments", in CVPR Spotlight, 2017.

Keywords: Oriented text detection

## 概要

・(2D的)回転しているテキストを検出するSegLinkを提案する。テキストを「セグメント」と「リンク」に分ける。回転している画像中のテキストを、語やテキストラインのようなセグメントに分け、回転矩形で表現する。セグメントのうち同じ行に属しているものをリンクで連結する。セグメント、リンクはFCNNによりend-to-endで学習され、マルチスケールで密に検出される。従来手法と比較し、精度、速度、学習の容易さにおいて性能が向上した。また、非ラテンの文字列も同じフレームワークで検出可能である。スペースが空きすぎず、直線の行なら検出可能。



## 新規性・差分

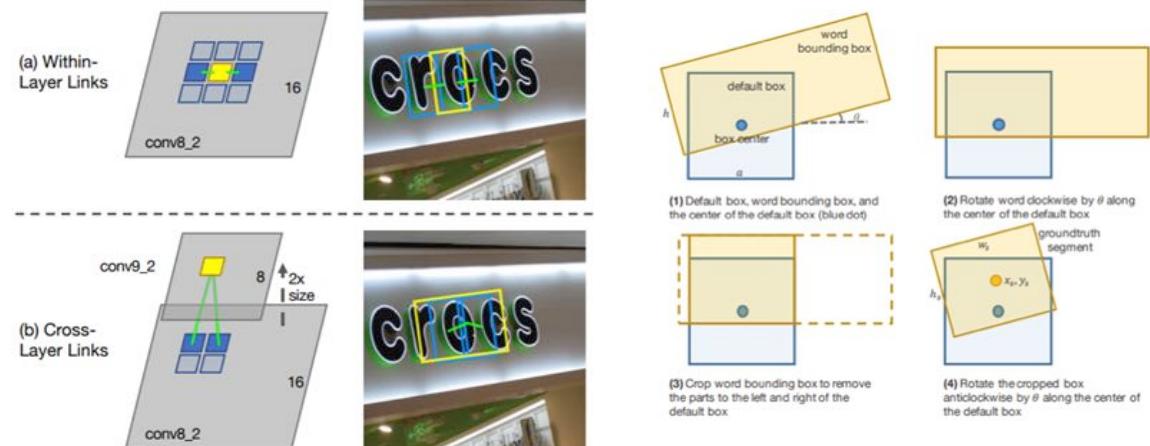
- ・非ラテン文字OK
- ・20fpsで処理可能



## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Shi\\_Detecting\\_Oriented\\_Text\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Shi_Detecting_Oriented_Text_CVPR_2017_paper.pdf)



[92] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, Alexander Sorkine-Hornung, "Learning Video Object Segmentation from Static Images", in CVPR Spotlight, 2017.

Keywords: Segmentation, Fine-tuning

## 概要

- ビデオの物体の注目領域マスクを追跡する。コンセプトのカギは、オフラインとオンラインの2つの学習の戦略の組み合わせによる。オフライン学習では以前のフレームでの推定によりリファインドなマスクを得る(MaskRefinementNetwork)。オンライン学習では追跡領域の画像を学習データとしたファインチューニングを逐次行う。ラベリングのネットワークDeepLabv2-VGGを用いた。

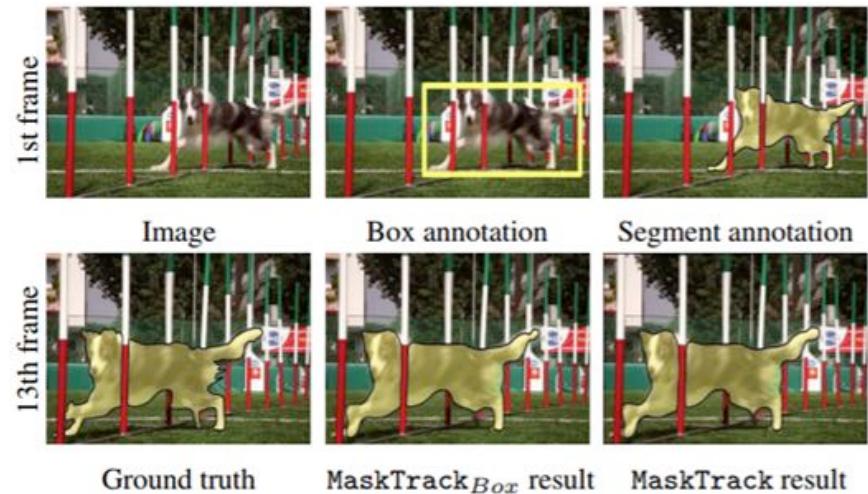
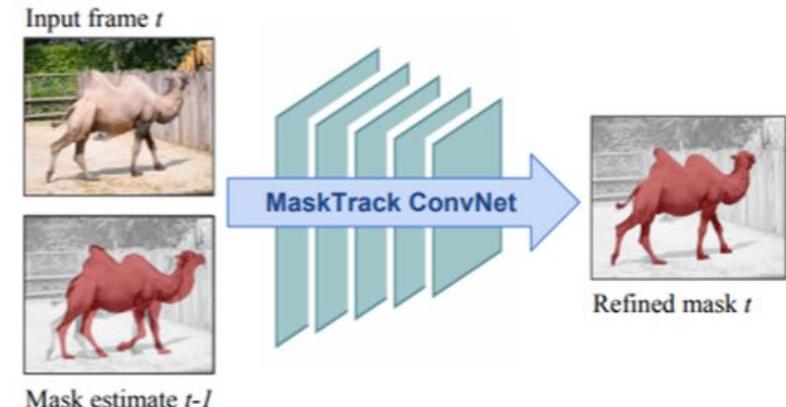
## 新規性・差分

- 入力マスクはラフでOK
- アノテーションはセグメンテッドラベルだけでなくBoundingBoxでもオプティカルフローでもOK
- 実験的論文。問題設定に沿ったデータオーグメンテーションの仕方などが丁寧に書かれている。どこまでデータを精製すればよいか、アノテーションの種類による性能変化、ファインチューニングの是非を評価実験により調査している。
- すごい論文力。

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Perazzi\\_Learning\\_Video\\_Object\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Perazzi_Learning_Video_Object_CVPR_2017_paper.pdf)



- [93] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J. Crandall, Michael S. Ryoo, "Identifying First-Person Camera Wearers in Third-Person Videos", in CVPR Poster, 2017.

Keywords: sensors data identification, wearable camera vs. static camera , triplet loss function

## 概要

- ・ウェアラブルカメラ装着者と、第三者視点の固定カメラ映像中の人物を同定する。semi-Siamseネットワークにより実現する。同じ対応付けを近づけ、違う対応付けを遠くするような距離を定義するTriplet loss関数を定義する。それぞれのネットワークはTwo-stream構造を持ち、空間・動きの両特徴を考慮する。

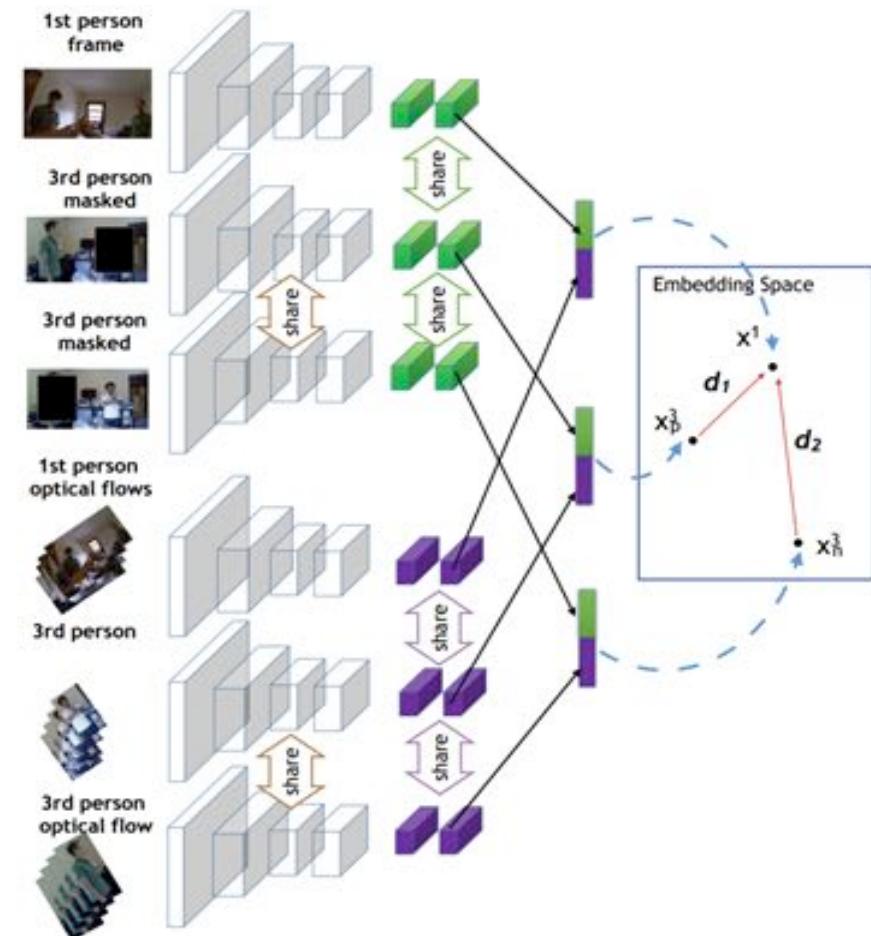
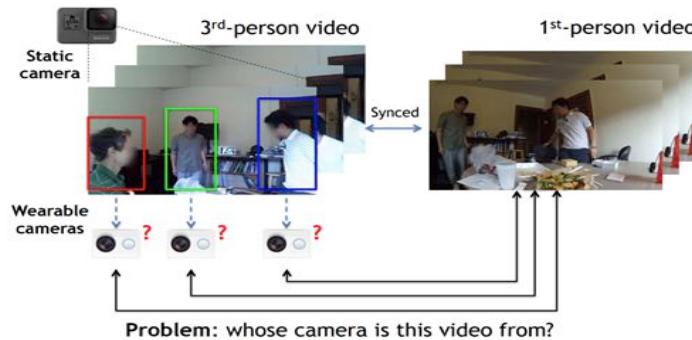
## 新規性・差分

- ・新規的な問題に対する、新しい構造によるCNNの適用。
- ・ベースラインとの比較で、三人称映像中の人物をクエリとした一人称視点画像の同定タスクにおいて性能が3割程度向上
- ・一人称映像をクエリとした三人称視点画像中の人物同定タスクにおいて性能が5%程度向上

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Fan\\_Identifying\\_First-Person\\_Camera\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Fan_Identifying_First-Person_Camera_CVPR_2017_paper.pdf)



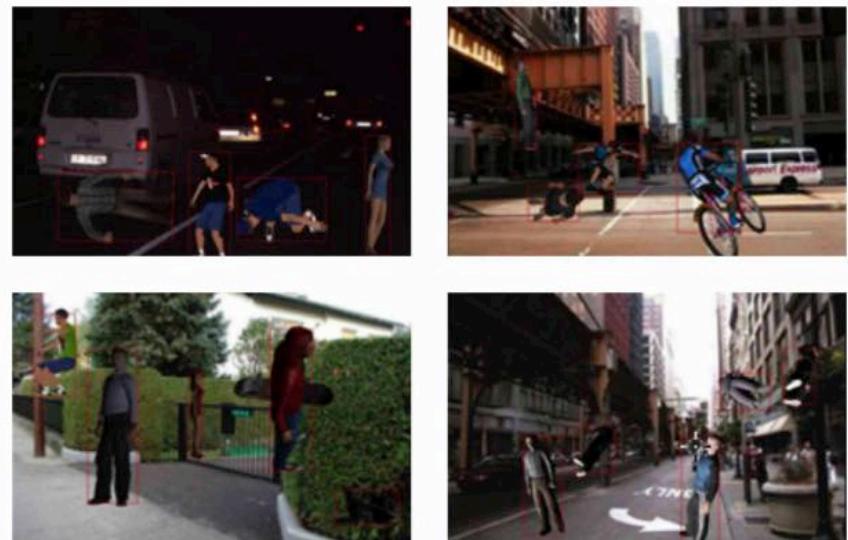
(d) Two-stream semi-triplet network

# 【94】Shiyu Huang, Deva Ramanan, "Expecting the Unexpected: Training Detectors for Unusual Pedestrians with Adversarial Imposters", in CVPR, 2017.

Keywords: Pedestrian Detection, GAN, Synthetic Data

## 概要

- ・従来の歩行者検出データセットでは、ほとんどが普通に歩いている歩行者のデータで、危険につながりやすい遊んでいる子供やスケボーに乗っている人などのデータは少なかった。そこで、そのようなデータを集めたデータセットを新たに作成。しかし、このようなデータを大量に集めるのは困難なので、人工的に合成データを生成して学習データ量を増やす。そこで、より実データに近い合成データを生成するために、GANベースの手法を提案。ノイズベクトルではなく合成データをRenderingするためのパラメータを入力とできるようにネットワークを修正。実データに近い合成データの利用により高い精度の検出を実現。



## 新規性・差分

- ・危険につながりやすい歩行者データを含む新たなデータセットを構築
- ・実データに近い合成データを生成するためのGANベースの手法を提案

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Huang\\_Expecting\\_the\\_Unexpected\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Expecting_the_Unexpected_CVPR_2017_paper.pdf)

Fine-tuning method	50% overlap	70% overlap
$S$	83.49%	95.18%
$T$	72.39%	93.70%
$S \Rightarrow T$	48.45%	77.14%
$S \Rightarrow (T \cup I)$	45.97%	74.94%
$S \Rightarrow (T \cup I) \Rightarrow T$	<b>42.47%</b>	<b>73.70%</b>

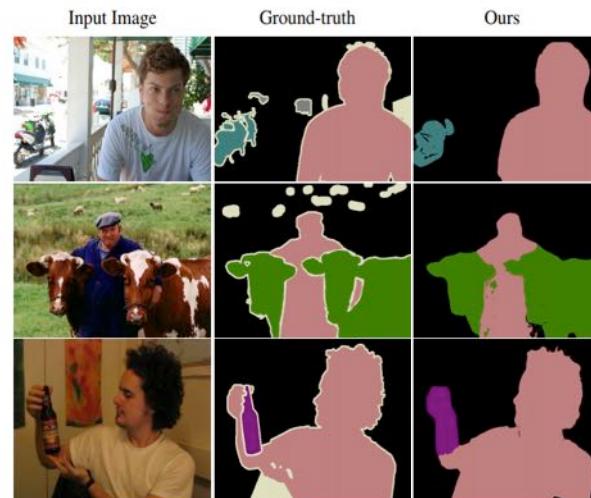
・ $I$ : 合成データ

# [95] Seunghoon Hong, Donghun Yeo, Suha Kwak, Honglak Lee, Bohyung Han, “Weakly Supervised Semantic Segmentation using Web-Crawled Videos”, in CVPR, 2017.

Keywords: Semantic Segmentation, Weakly Supervised, Web Video

## 概要

- 画像単位でラベルが与えられているWeakly Supervised Semantic Segmentationのための新しい手法を提案。画像単位のラベルのみの利用では学習が難しいので、YouTube動画を利用して精度を上げるための手法。動画は画像よりも前景・背景のセグメンテーションがしやすいという利点があるが、ノイズが多いという欠点もある。そこで、画像中のObject Localization手法に基づいて動画中の対象が含まれている領域を探し、動画を利用してセグメンテーションマスクを獲得。画像と動画の両者の利点を活かすことにより効果的なWeakly Supervised Learningを実現。



Method	mean
<b>Image labels:</b>	
EM-Adapt [26]	33.8
CCNN [28]	35.3
MIL+seg [30]	42.0
SEC [17]	50.7
<b>+Extra annotations:</b>	
Point supervision [2]	46.0
Bounding box [26]	58.5
Bounding box [6]	62.0
Scribble [20]	63.1
Transfer learning [13]	52.1
<b>+Videos (unannotate</b>	
MCNN [36]	38.1
Ours	<b>58.1</b>

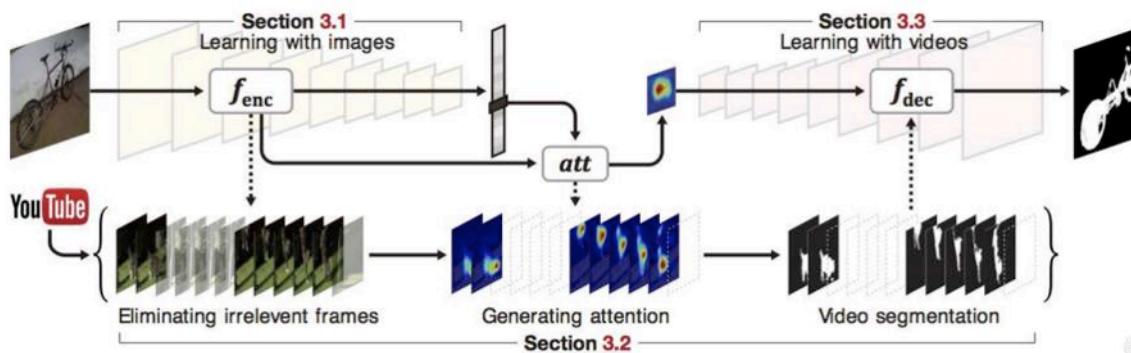
## 新規性・差分

- Web動画を用いたWeakly Supervised Learningを提案
- 画像単位のラベル以外に人手を用いず学習可能な手法であり高精度

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Hong\\_Weakly\\_Supervised\\_Semantic\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Hong_Weakly_Supervised_Semantic_CVPR_2017_paper.pdf)

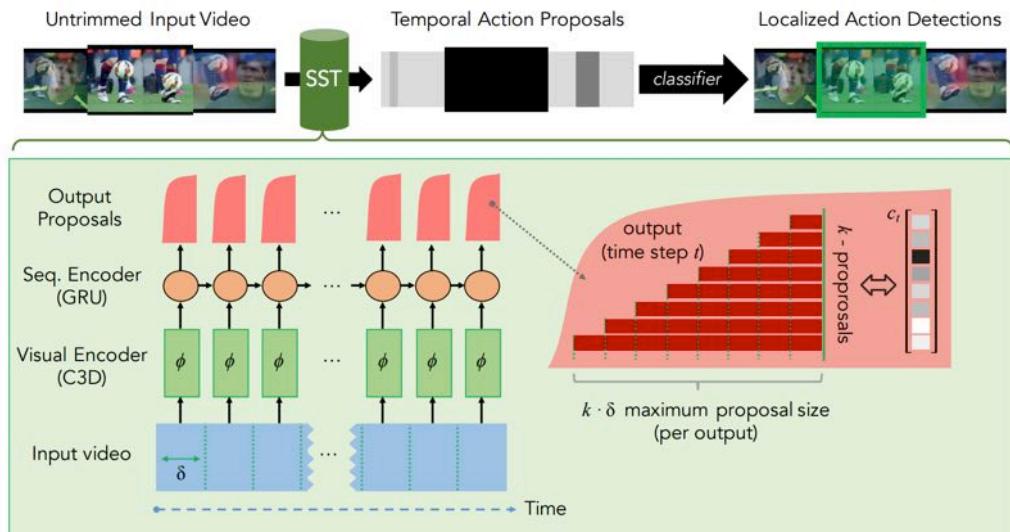


【96】Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, Juan Carlos Niebles, "SST: Single-Stream Temporal Action Proposals", in CVPR, 2017.

Keywords: Action Proposals, Action Localization

## 概要

- 動画像中から行動の候補領域（開始，終了）を検出する Temporal Action Proposalsにおける新たな手法を提案。従来手法はマルチスケールで重なりを含むSliding Windowをベースとした手法が多かった。この研究では、重なりなし、単一スケールのSliding Windowを一度動画全体に適用するだけで様々なスケールのAction Proposalsを生成する手法を提案。CNNにより各Windowをエンコードした後、RNNに入力して、各Windowに基づくProposalsを出力。評価実験から、従来手法よりも高い精度で高速に動作することを確認した。



## 新規性・差分

- 効率的にマルチスケールのAction Proposalsを生成する手法を提案
- RNNにより非常に長い動画に対しても同様に処理可能な手法を実現

## Links

論文  
[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Buch\\_SST\\_Single-Stream\\_Temporal\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Buch_SST_Single-Stream_Temporal_CVPR_2017_paper.pdf)  
 Github <https://github.com/ranjaykrishna/SST>

Method	Recall		
	tIoU= 0.6	tIoU= 0.8	FPS
DAPs	0.916	0.573	134
S-CNN-prop	<b>0.938</b>	0.524	60
<b>SST (Ours)</b>	<b>0.920</b>	<b>0.672</b>	<b>308</b>

- [97] Licheng Yu, Hao Tan, Mohit Bansal, Tamara L. Berg, "A Joint Speaker-Listener-Reinforcer Model for Referring Expressions", in CVPR Spotlight, 2017.

Keywords: referring expression

## 概要

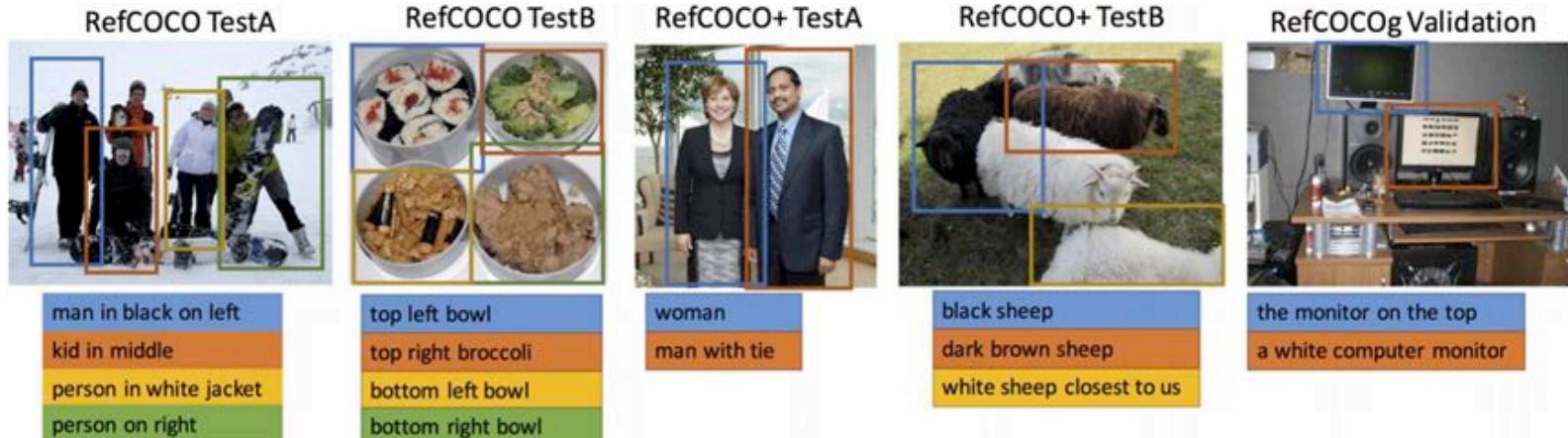
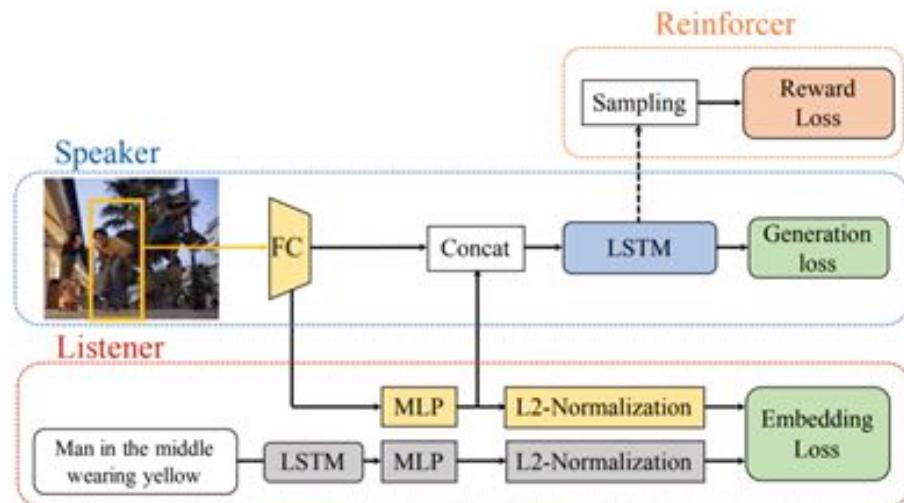
- 参照表現という自然言語上の枠組みがある。参照表現の理解および生成のフレームワークを提案する。(1)Speaker (2)Listener (3)Reinforcer の3つのモジュールから構成される。Speakerは参照表現を生成し、Listenerは参照表現を理解し、Reinforcerはより判読可能な表現に対し褒賞する。Listener-Speakerのモジュールは、end-to-endで学習される。Reinforcerによるフィードバックは学習中にSpeaker, Listenerに共有される。

## 新規性・差分

- 参照表現の理解・生成のタスクにおいてstate-of-the-art
- 参照表現研究については（特に言語学においては）旧来から行われてきたが、深層学習化は数点

## Links

論文[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Yu\\_A\\_Joint\\_Speaker-Listener-Reinforcer\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Yu_A_Joint_Speaker-Listener-Reinforcer_CVPR_2017_paper.pdf)



- [98] Vamsi Kiran Adhikarla, Marek Vinkler, Denis Sumin, Rafał K. Mantiuk, Karol Myszkowski, Hans-Peter Seidel, Piotr Didyk, "Towards a quality metric for dense light fields", in CVPR Poster, 2017.

Keywords: light field, distortion, evaluation

## 概要

- ・3次元シーンの表現に使われ、コモディティ化が進むライトフィールドは質の低下の原因がいくつかあるが、その悪化の程度の定量化が望まれる。そこで、合成・実写の両方を含むライトフィールドのデータセットを作成し、いくつかのパターンによる歪みをシミュレートし、主観評価実験を行った。
- ・歪みのパターンは次の通り。

伝達時：動画圧縮（HEVC条件），6つの量子化パラメータ

復元時：補間方法4つ（NN, LINEAR, OPT, DQ条件），6つの角度分解能パラメータ

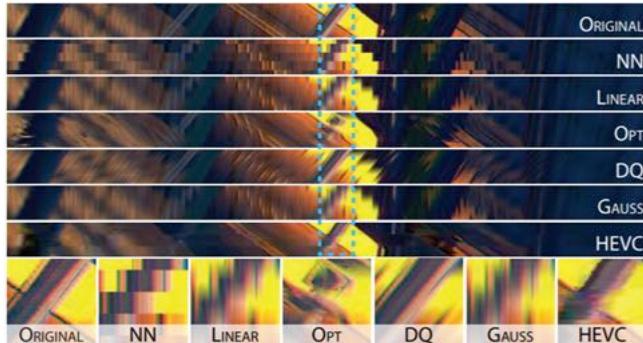
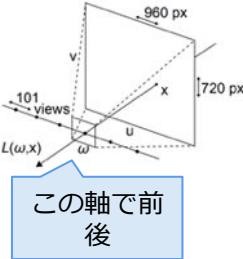
表示時：左右ステレオ画像の混信を角度方向のガウシアンブラーとしてモデル化（GAUSS条件），6つの角度分解能パラメータ

## 新規性・差分 ・新技術へのフォーカスシフトにあたっての先進的，示唆的研究

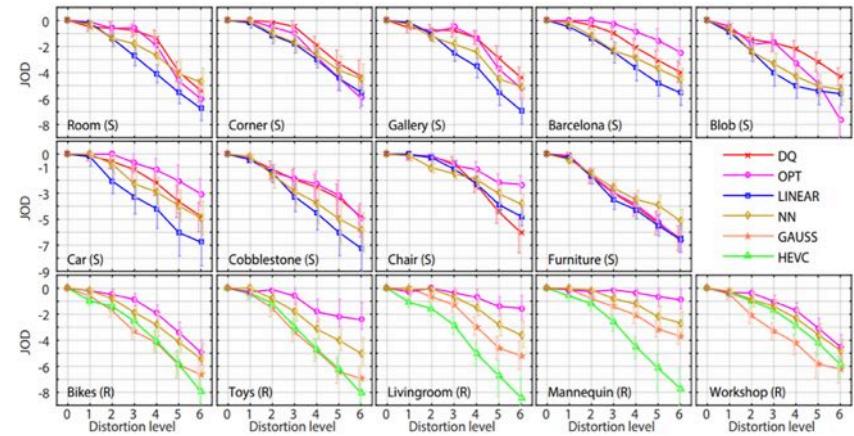
## Links

論文[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Adhikarla\\_Towards\\_a\\_Quality\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Adhikarla_Towards_a_Quality_CVPR_2017_paper.pdf)

## 結果



縦軸：上ほど高評価  
横軸：パラメータ，右ほど低質



【99】 Wadim Kehl, Federico Tombari, Slobodan Ilic, Nassir Navab, "Real-Time 3D Model Tracking in Color and Depth on a Single CPU Core", in CVPR Poster, 2017.

Keywords: 3D Tracking, low computational cost

## 概要

- RGB-D映像による3次元追跡を、シングルコアCPUで、2msの処理速度を実現する。 (1)色で前景抽出し、輪郭をとる(2)輪郭を8度刻みでプリレンダする(3)疎にサンプリング(4)2D比較の輪郭一致エネルギーと3D比較のICPエネルギーを定義し、エネルギー最小化で位置・姿勢を逐次追跡する。高速に動作するものの、平均誤差がサブミリ、サブ度オーダーで、従来手法に匹敵する。また、処理速度が6msに増加するものの、輪郭抽出について距離情報も併用すると、さらに精度が向上する。

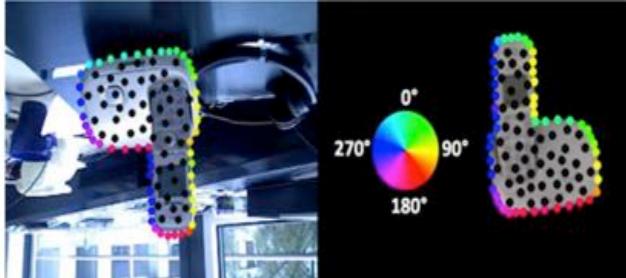
## 新規性・差分

- No-CNN
- シングルコアで高速で高精度と圧倒的スペック

## Links

論文  
[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Kehl\\_Real-Time\\_3D\\_Model\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Kehl_Real-Time_3D_Model_CVPR_2017_paper.pdf)

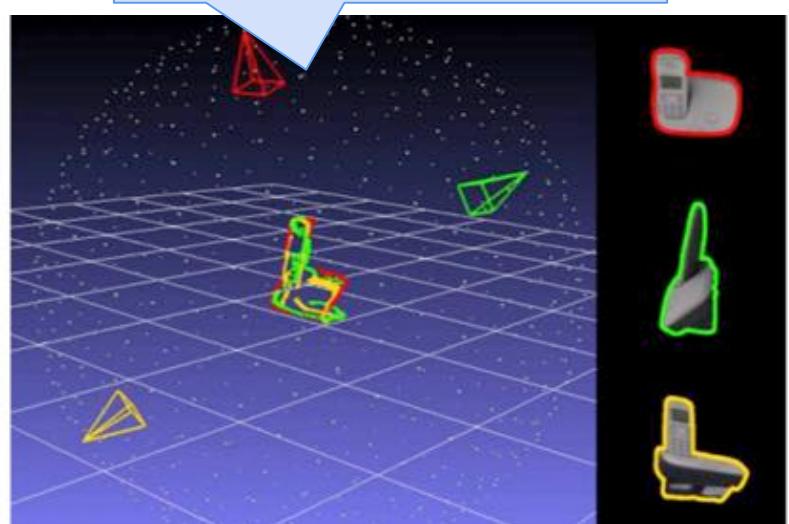
疎にサンプリング



前景マスクがあるので  
オクルージョン対応可能



あらかじめ様々な視点  
から撮つておく



[100] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, Eli Shechtman, "Controlling Perceptual Factors in Neural Style Transfer", in CVPR Poster, 2017.

Keywords: style transfer, multiple sources, flexibility

概要 • スタイルトランスターファーの研究。既存手法を拡張し、複数の入力に対して、(1)空間的配置、(2)色情報、(3)空間スケールのコントロールができるようになる。地面のテクスチャが空領域に適用されるような失敗ケースが緩和される。また、画像Aの照明情報を優先するなど、非常にフレキシブルなスタイルトランスターファーが行えるようになる。

## 新規性・差分

- 非常に高い柔軟性

## Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Gatys\\_Controlling\\_Perceptual\\_Factors\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Gatys_Controlling_Perceptual_Factors_CVPR_2017_paper.pdf)

(1)空間配置自在



(2)色情報自在



(3)スケーリング自在

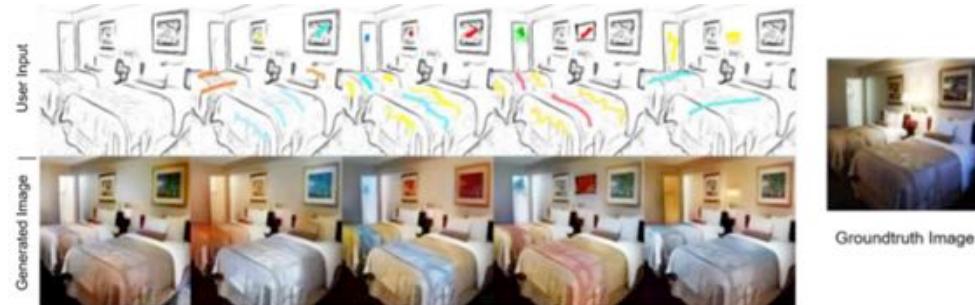


【101】 Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, James Hays,  
 “Scribbler: Controlling Deep Image Synthesis With Sketch and Color”,  
 in, CVPR, 2017.

Keywords: image generation, sketch, user interactive, GAN

## 概要

- ・スケッチ画像からリアルな画像を生成する方法があるが、グレースケールのスケッチに対して、ユーザが疎に色付けすることで、好きな配色で、しかもよりリアルな画像を生成できる手法を提案。ユーザの入力に対してフィードフォワードで高速に動作するので、インタラクティブに入力スケッチ画像を編集できる。  
背景に色がリークするなどの、現状うまくいっていない点についても考察。



## 新規性・差分

- ・スケッチ画像に線でちょっと色を塗るだけで配色変更可能
- ・インタラクティブ編集できる

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Sangkloy\\_Scribbler\\_Controlling\\_Deep\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Sangkloy_Scribbler_Controlling_Deep_CVPR_2017_paper.pdf)

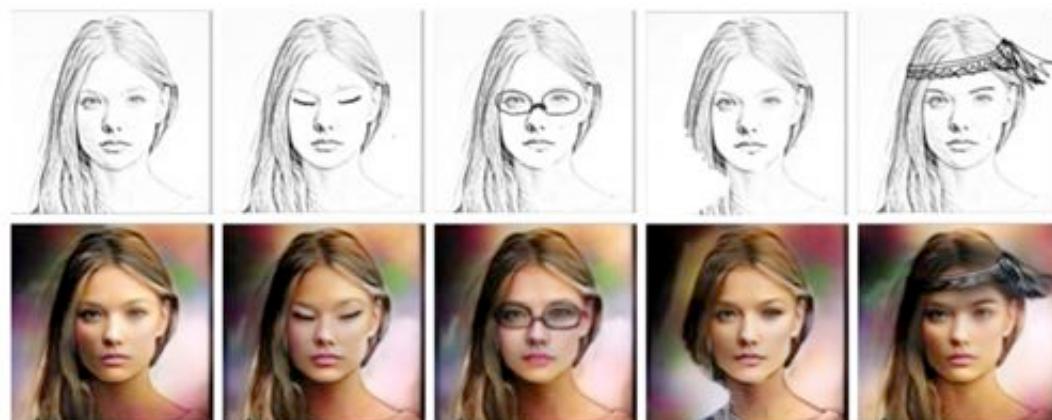


Figure 8. Interactive image editing. The user can incrementally modify the sketch to change the eyes, hair, and head decorations.

# 【102】Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, “Temporal Residual Networks for Dynamic Scene Recognition”, in CVPR, 2017.

Keywords: Residual Network, Dynamic Scene Recognition

## 概要

動画像を入力としてシーンを認識するDynamic Scene Recognitionに関する研究。時空間の情報を畳み込んで記述するための新たなCNNを提案(T-ResNet)。ResNetをベースとしたCNNであり、従来のResidual Unitに隣接したフレームの情報をつなげるための新たなfilterを追加。従来のデータセットやこの研究で新たに提案したデータセットで評価した結果、従来手法を上回る精度を示した。

## 新規性・差分

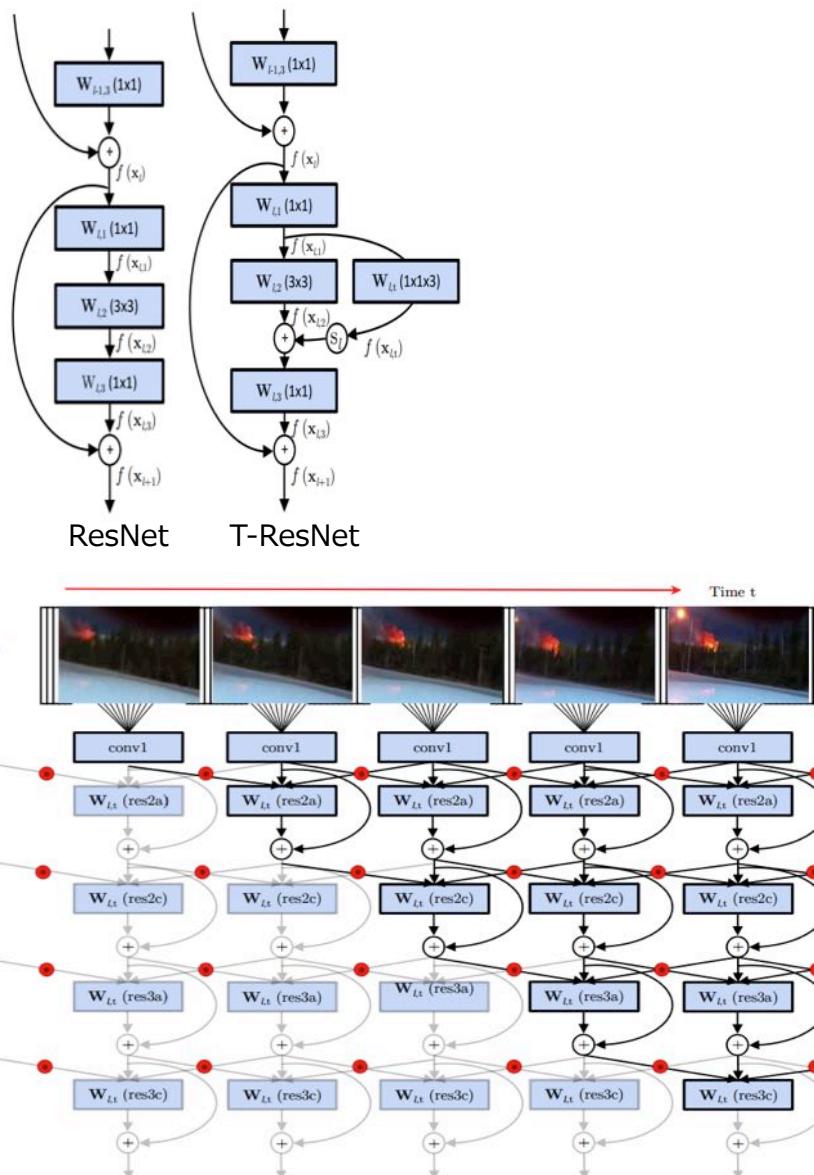
- ・時空間情報を記述するための新たなネットワークを提案
- ・新しいDynamic Scene Recognitionのデータセットを提案

Class	SFA	BoSE	T-CNN	S-CNN	IDT	C3D	ResNet	T-ResNet
Average	56.9	77.0	50.6	82.0	85.6	84.0	85.9	89.0

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Feichtenhofer\\_Temporal\\_Residual\\_Networks\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Feichtenhofer_Temporal_Residual_Networks_CVPR_2017_paper.pdf)

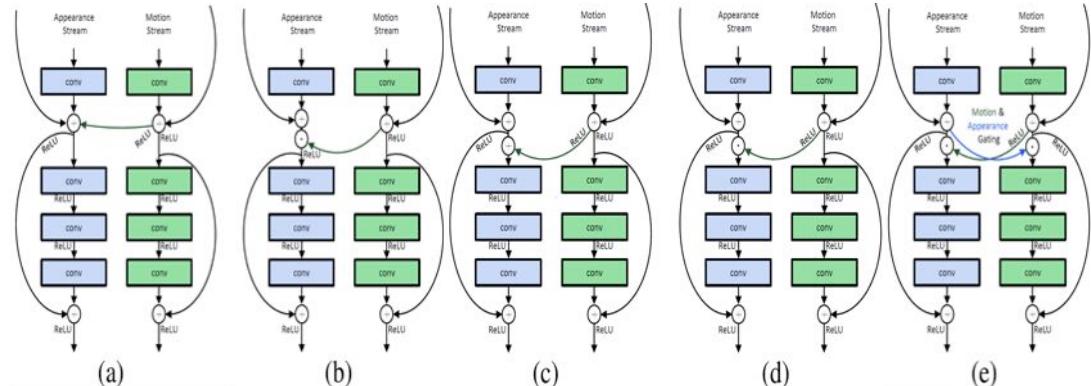


# 【103】Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, “Spatiotemporal Multiplier Networks for Video Action Recognition”, in CVPR, 2017.

Keywords: Action Recognition, Two-stream CNN

## 概要

Two-stream CNNをベースとした新たなAction Recognitionの手法を提案。従来にもTwo-streamもAppearanceとMotionを統合するための手法は提案されていたが十分な検討はされていなかった。この研究では様々な統合方法を検討し、従来用いられていなかった乗算による統合がより精度を向上させることを示した。



## 新規性・差分

- Two-Stream CNNにおける有効な統合方法を検討
- 乗算による統合が有効な理由を考察

case	into	Fig.	UCF101	HMDB51
direct $\oplus$	$\leftarrow$	Fig. 2(a)	24.78	54.85
direct $\odot$	$\leftarrow$	Fig. 2(b)	81.98	77.89
residual $\oplus$ Sect. 3.2.1	$\leftarrow$	Fig. 2(c)	9.38	41.89
residual $\odot$ Sect. 3.2.2	$\leftarrow$	Fig. 2(d)	<u>8.72</u>	<u>37.23</u>
residual $\oplus$	$\rightarrow$	Fig. 2(c)	16.76	49.54
residual $\odot$	$\rightarrow$	Fig. 2(d)	16.68	48.43
residual $\odot$	$\leftrightarrow$	Fig. 2(e)	15.15	48.56

Two-streamの  
様々な統合方法  
を検討

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Feichtenhofer\\_Spatiotemporal\\_Multiplier\\_Networks\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Feichtenhofer_Spatiotemporal_Multiplier_Networks_CVPR_2017_paper.pdf)

プロジェクト

[104]

Jan Hosang, Rodrigo Benenson, Bernt Schiele, "Learning non-maximum suppression", in CVPR, 2017.

Keywords: Object Detection, Non-Maximum Suppression

## 概要

物体検出などの検出課題において、一つの対象に対する複数の検出結果を統合するために、後処理としてNon-Maximum Suppression (NMS) という処理が行われていた。近年主流となっているのはEnd-to-End Frameworkであり、後処理としてNMSをかけるというのは、真のEnd-to-End Detectionには邪魔となる。そこで、この研究では、Bounding Boxとそのスコアを入力としてNMSを行うためのCNNを提案。従来のGreedyなClusteringをベースとするNMSよりも高い精度を示した。検出処理とNMS処理の完全な統合につながっていくことを期待。

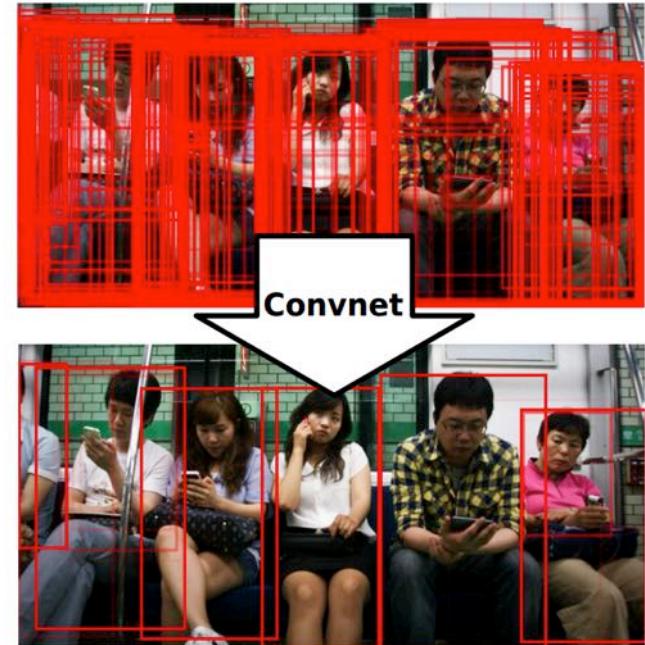
## 新規性・差分

- ・NMSを行うためのCNNを提案
- ・従来のNMSよりも高い精度を実現

## Links

論文

[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Hosang\\_Learning\\_Non-Maximum\\_Suppression\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Hosang_Learning_Non-Maximum_Suppression_CVPR_2017_paper.pdf)



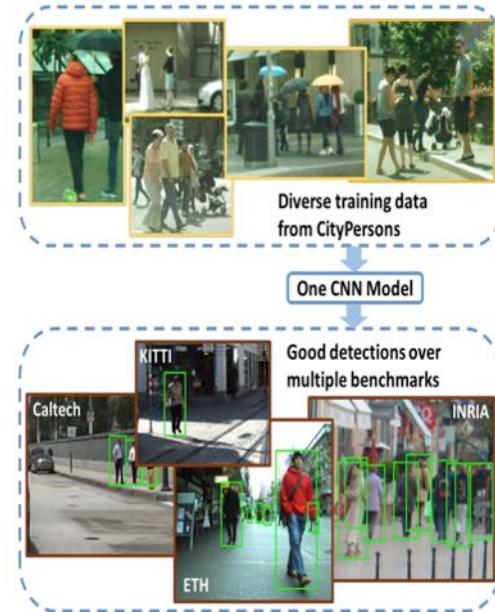
Method	All		Occlusion [0, 0.5)		Occlusion [0.5, 1]		
	AP <sub>0.5</sub>	AP <sub>0.5</sub> <sup>0.95</sup>	AP <sub>0.5</sub>	AP <sub>0.5</sub> <sup>0.95</sup>	AP <sub>0.5</sub>	AP <sub>0.5</sub> <sup>0.95</sup>	
val	GreedyNMS>0.5	65.6	35.6	65.2	35.2	35.3	12.1
	Gnet, 8 blocks	<b>67.3</b>	<b>36.9</b>	<b>66.9</b>	<b>36.7</b>	<b>36.7</b>	<b>13.1</b>
test	GreedyNMS>0.5	65.0	35.5	61.8	33.8	30.3	11.0
	Gnet, 8 blocks	<b>66.6</b>	<b>36.7</b>	<b>66.8</b>	<b>36.1</b>	<b>33.9</b>	<b>12.4</b>

# 【105】 Shanshan Zhang, Rodrigo Benenson , Bernt Schiele, “CityPersons: A Diverse Dataset for Pedestrian Detection”, in CVPR, 2017.

Keywords: Pedestrian Detection, New Dataset

## 概要

新しい歩行者検出のためのCityPersonsデータセットを提案. 元々 Semantic Segmentation用に提案されていたCityScapesデータセットをベースに, 歩行者検出用のBounding Boxをアノテーション. Semantic Segmentationでは見えていている部分にアノテーションされている（オクルージョン領域は含まない）ので, 従来の歩行者検出用データセットに合わせて, 全身を含むBounding Boxをアノテーションした. 加えて, pedestrian, rider, sitting person, other personという4種類のラベルも付与. CityPersonsデータセットでFaster R-CNNを学習することでstate-of-the-artを達成.



## 新規性・差分

- ・新しい歩行者検出用のデータセットを提案
- ・提案データセットでのPre-trainingが有効であることを確認

認

## Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Zhang\\_CityPersons\\_A\\_Diverse\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhang_CityPersons_A_Diverse_CVPR_2017_paper.pdf)  
プロジェクト

	Caltech	KITTI	CityPersons
# country	1	1	3
# city	1	1	18
# season	1	1	3
# person/image	1.4	0.8	7.0
# unique person	1 273	6 336	19 654

# 【106】 Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrel, Bharath Hariharan “Learning Features by Watching Objects Move”, in CVPR, 2017.

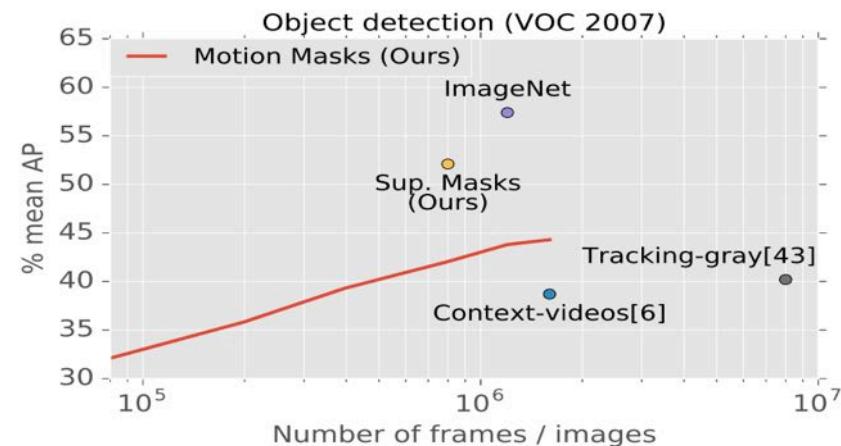
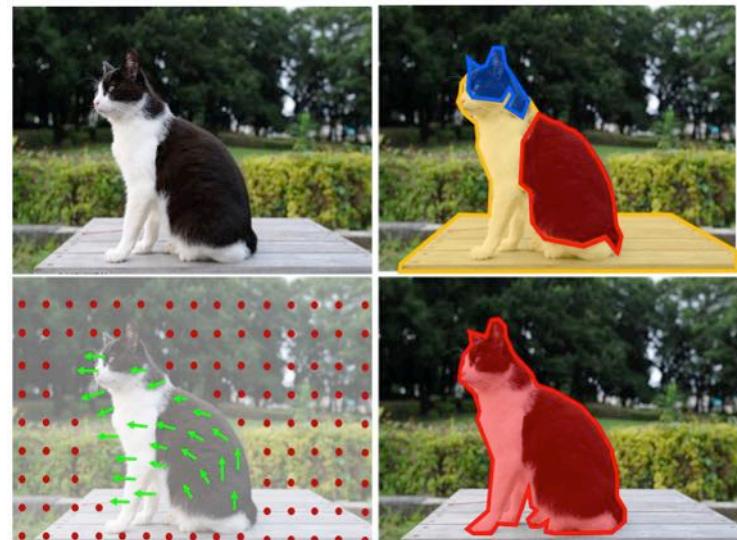
Keywords: Feature Learning, Unsupervised, Segmentation

## 概要

ラベル付きデータなしで特徴表現を学習するための新しい手法を提案。提案手法は動画像を利用。動画像の各フレームをモーションベースのセグメンテーションにより前景と背景に分割。各フレームを入力、これによって得られる前景領域のマスクを教師信号としてCNNを学習すると、画像認識や検出に有効な高次の特徴表現が学習可能であることを示した。

## 新規性・差分

- ・従来手法よりも高精度なラベル付きデータなしでの特徴表現学習手法を提案
- ・モーションベースのセグメンテーションマスクの利用が特徴表現学習に有効なことを示した



## Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Pathak\\_Learning\\_Features\\_by\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Pathak_Learning_Features_by_CVPR_2017_paper.pdf)  
プロジェクト

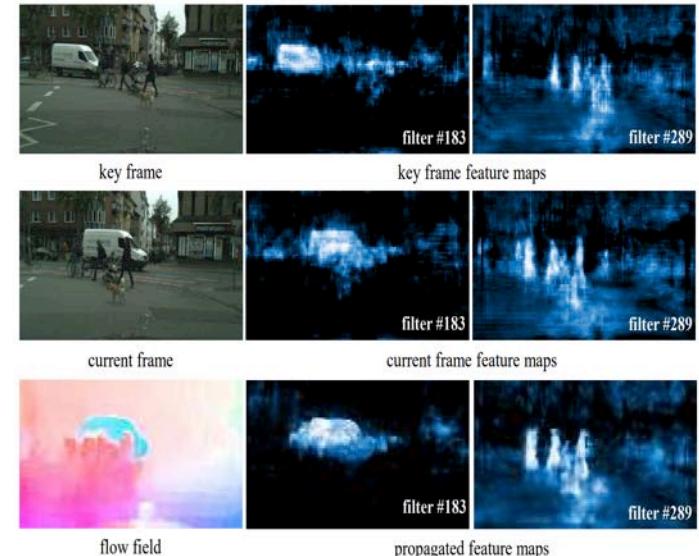
【107】

Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, Yichen Wei,  
“Deep Feature Flow for Video Recognition”, in CVPR, 2017.

Keywords: Video Recognition,

## 概要

動画に対してフレームごとに検出やセマンティックセグメンテーションをする場合、フレームごとの特徴マップ計算が高コストであり、高速な動作が困難という問題があった。この研究では、スパースなキーフレームに対してのみ特徴マップを計算し、残りのフレームに対してはキーフレームの特徴マップをベースに変換することで高速な動作を実現。キーフレームからのFlow Fieldを計算し、Flowに基づいて特徴マップを変化することで、そのフレームから直接計算した特徴マップを非常に類似したものが計算可能。Flow Fieldの計算もCNNで行われ、それも含めてEnd-to-Endに全体を最適化可能。実験結果から、僅かに精度を低下させるものの、フレームごとに処理するよりも高速な処理を実現。



## 新規性・差分

・スパースなキーフレーム以外の処理を簡略化することで動画に対する処理の高速化を実現

Methods	Cityscapes ( $l = 5$ )		ImageNet VID ( $l = 10$ )	
	mIoU(%)	runtime (fps)	mAP(%)	runtime (fps)
Frame	71.1	1.52	73.9	4.05
DFF	69.2	5.60	73.1	20.25

## Links

論文 [http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Zhu\\_Deep\\_Feature\\_Flow\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Zhu_Deep_Feature_Flow_CVPR_2017_paper.pdf)  
プロジェクト

【108】

## Title: Zero-Shot Learning - The Good, the Bad and the Ugly

- Zero-Shotが非常に注目されている(good).しかし、一致な評価プロトコルがない(bad).テストクラスがイメージネットをオーバー(ugly).
- Standard Split(SS) と比べ、 Proposal Split(PS) を提案した。

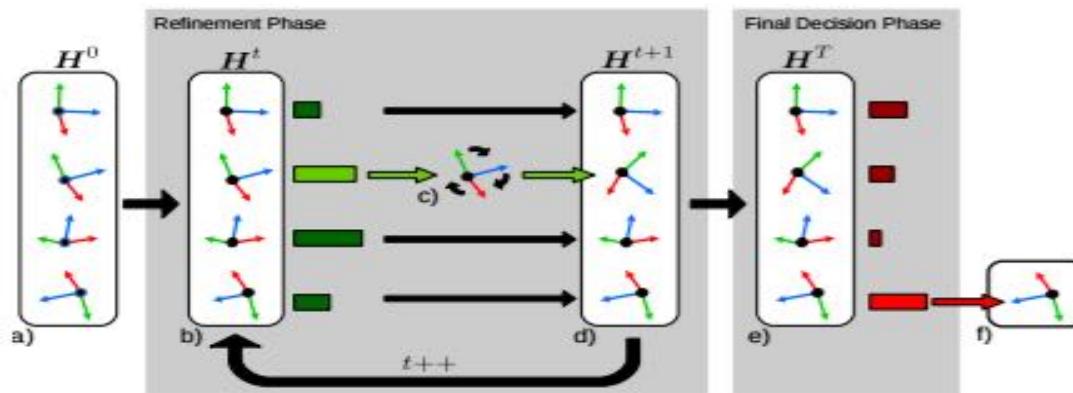
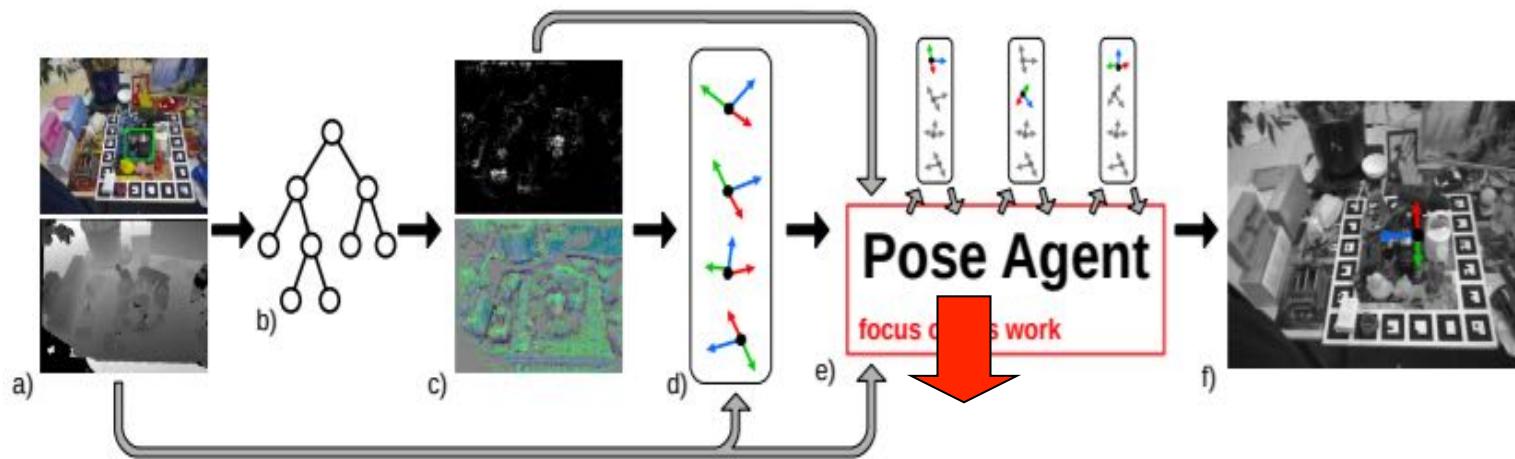
Method	SUN			CUB			AWA			aPY		
	ts	tr	H									
DAP [22]	4.2	25.1	7.2	1.7	67.9	3.3	0.0	<b>88.7</b>	0.0	4.8	78.3	9.0
CONSE [26]	6.8	<b>39.9</b>	11.6	1.6	<b>72.2</b>	3.1	0.4	88.6	0.8	0.0	<b>91.2</b>	0.0
CMT [34]	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	1.4	85.2	2.8
CMT* [34]	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	<b>10.9</b>	74.2	<b>19.0</b>
SSE [42]	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	0.2	78.9	0.4
LATEM [39]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	0.1	73.0	0.2
ALE [3]	<b>21.8</b>	33.1	<b>26.3</b>	23.7	62.8	<b>34.4</b>	<b>16.8</b>	76.1	<b>27.5</b>	4.6	73.7	8.7
DEVISE [11]	16.9	27.4	20.9	<b>23.8</b>	53.0	32.8	13.4	68.7	22.4	4.9	76.9	9.2
SJE [4]	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	3.7	55.7	6.9
ESZSL [32]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	2.4	70.1	4.6
SYNC [7]	7.9	43.3	13.4	11.5	70.9	19.8	8.9	87.3	16.2	7.4	66.3	13.3

Table 5: Generalized Zero-Shot Learning on Proposed Split (PS) measuring ts = Top-1 accuracy on  $\mathcal{Y}^{ts}$ , tr=Top-1 accuracy on  $\mathcal{Y}^{tr+ts}$ , H = harmonic mean (CMT\*: CMT with novelty detection). We measure top-1 accuracy in %.

【109】

## PoseAgent: Budget-Constrained 6D Object Pose Estimation

- 一枚のRGBD画像から既知物体の6D poseを予測。
- 6D poseを予測するタスクに初めて強化学習を導入。



## Detecting Oriented Text in Natural Images by Linking Seg

- textをローカル検出できる (segments, links) に解析する。
- segments: ボックス links:二つのsegmentsをリンクする
- end-to-end, full-convolutional neural networkを用いてデンスでマルチスケールで (segments, links) を検出する

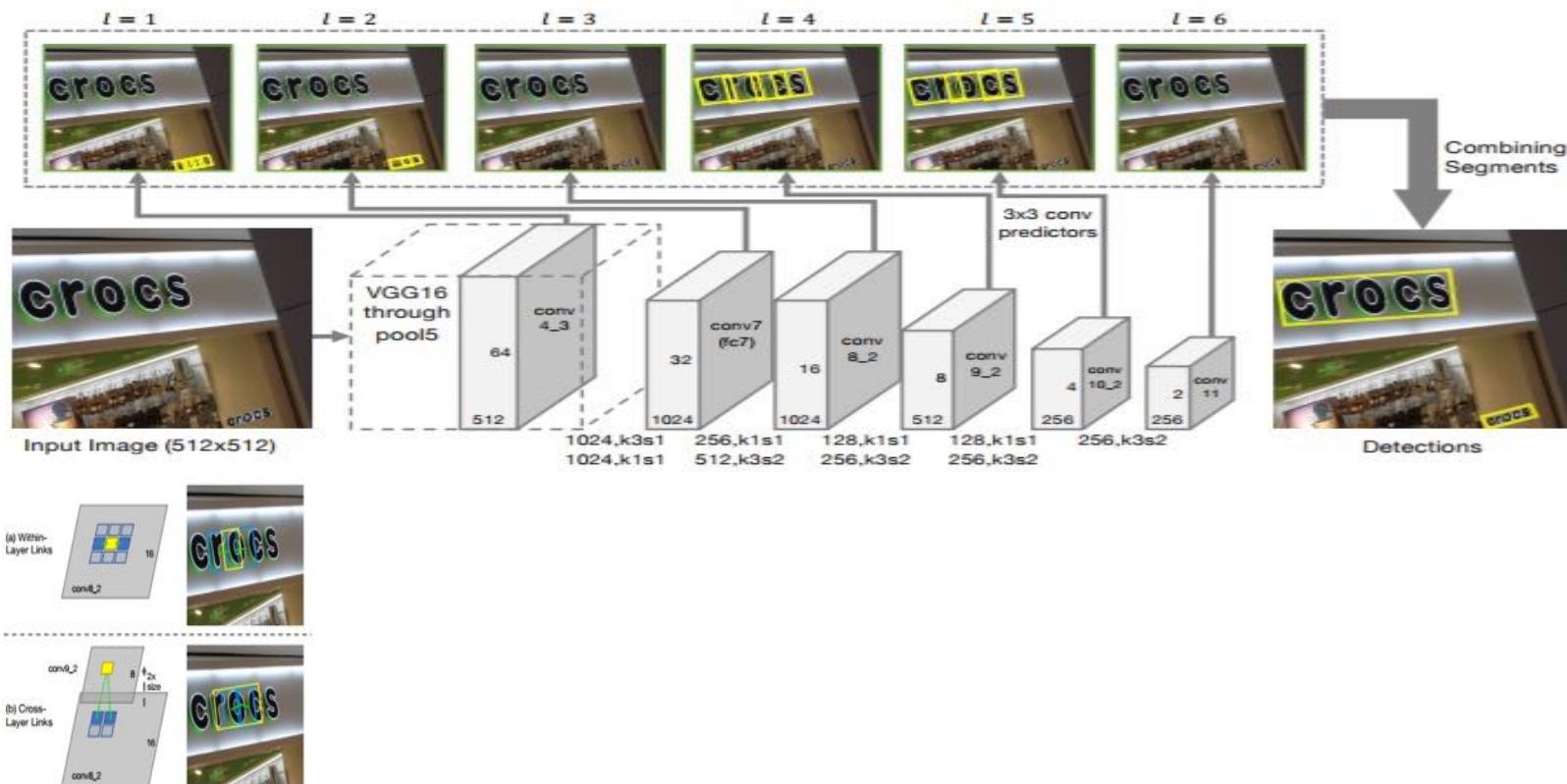


Figure 3. Within-Layer and Cross-Layer Links. (a) A location on conv8\_2 (yellow block) and its 8-connected neighbors (blue blocks with and without fill). The detected within-layer links (green lines) connect a segment (yellow box) and its two neighboring segments (blue boxes) on the same layer. (b) The cross-layer links connect a segment on conv9\_2 (yellow box) and two segments on conv8\_2 (blue boxes).

【111】

## Awesome Typography: Statistics-Based Text Effects Transfer

- 自動的に文字の様々な模様、textureを別の文字にコピーする。
- テキストの効果の空間分部の高レギュラー性を利用し、アピアランスと分部と心理効果の三つの面から、構成している。



目標関数 :

$$\min_q \sum_p E_{\text{app}}(p, q) + \lambda_1 E_{\text{dist}}(p, q) + \lambda_2 E_{\text{psy}}(p, q), \quad (10)$$

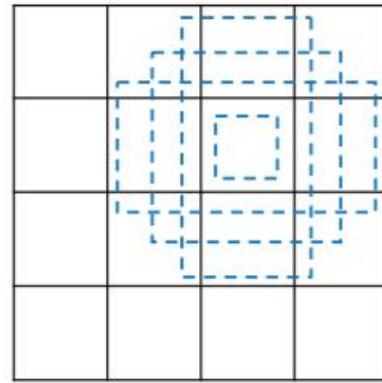
[112]

## Deep matching prior network:toward tighter multi-oriented text detection

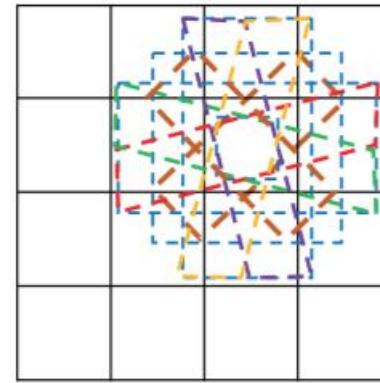
- 従来のバウンディングボックス検出を任意な四角い形にする
- マルチスケールsliding windowからアンカー、ポリゴン間の重なり計算、候補検出、さらに相関オフセット



a) Comparison of recalling scene text.



(b) Horizontal sliding windows.



(c) Proposed quadrilateral sliding windows.



### Our shared Monte-Carlo method

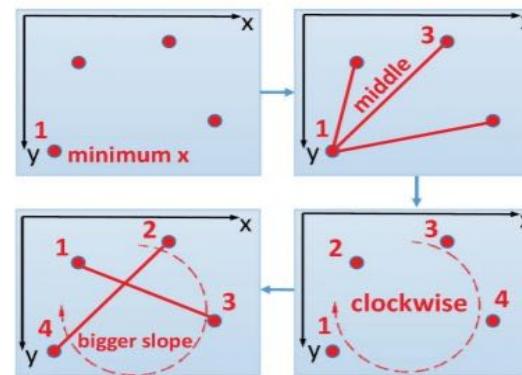
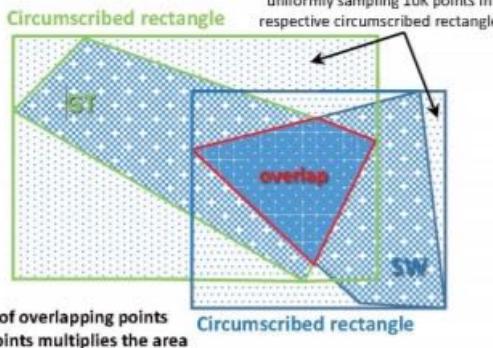


Figure 4. Procedure of uniquely determining the sequence of four points from a plane convex quadrangle.



# A Multi-View Stereo Benchmark with high-resolution images and multi-camera videos

- 高解像度DSLR画像使用
- 手持ちマルチカメラビデオからリコンストラクション
- 様々なシーン
- オンラインでできる
- 生成できる3次元ポイントクラウド色んな場面で用いられる

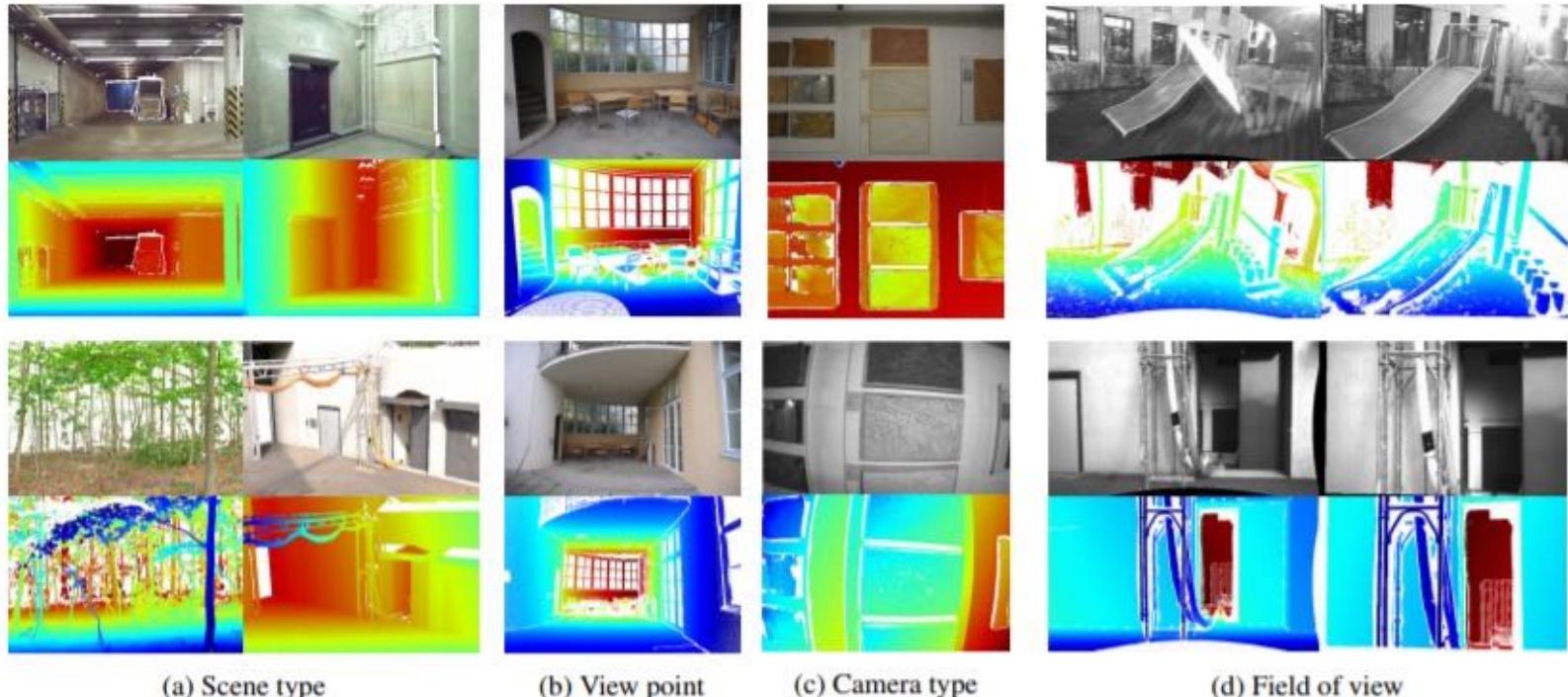
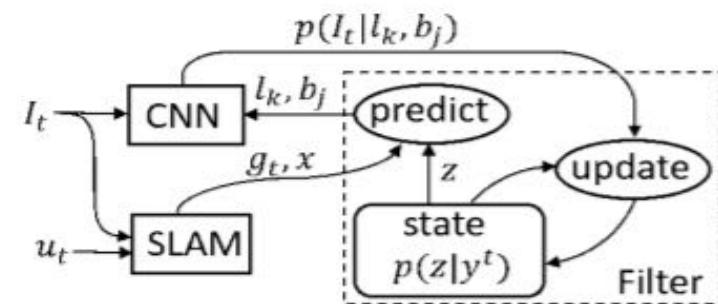
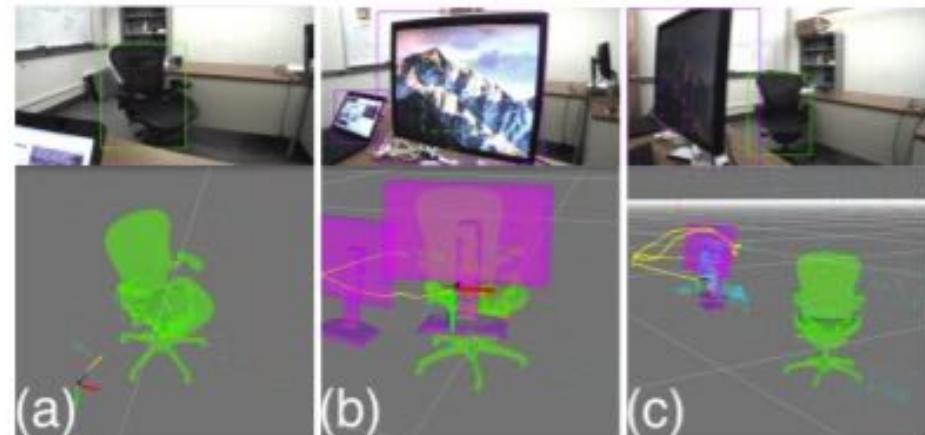
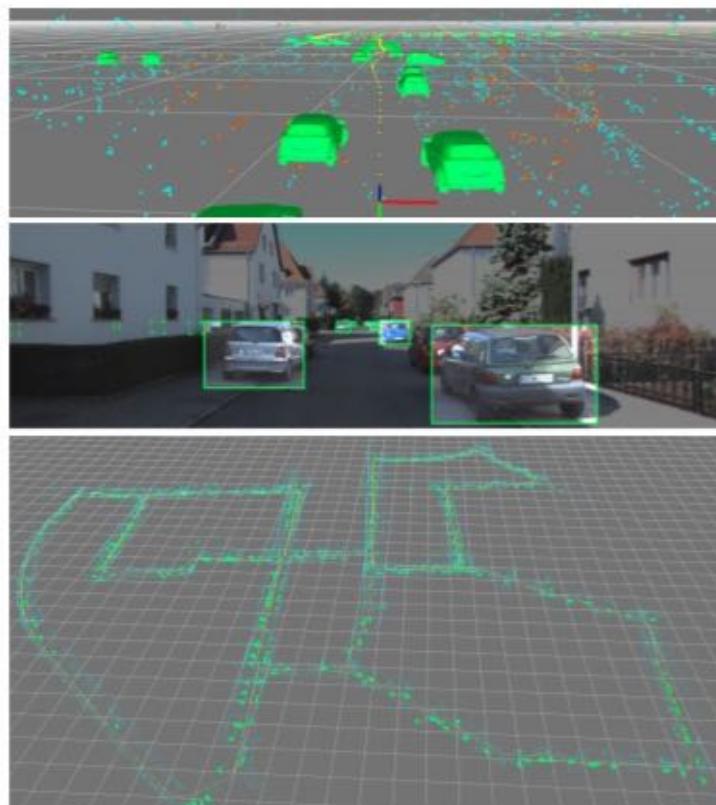


Figure 1. Examples demonstrating the variety of our dataset in terms of appearance and depth. (a) Colored 3D point cloud renderings of different natural and man-made scenes. (b) DSLR images taken from different view points. (c) DSLR image (top) and image from our multi-camera rig (bottom) of the same scene. (d) Camera rig images with different fields-of-view.

# Visual-Inertial-Semantic Scene representation for 3D object detection

- 正しい識別が画像の枚数が増やすに従って上昇する。また、物体が三次元形状とスケールがあるため、グローバル形状と方向などが得られる、そういう情報によってシステムが平衡できる。
- シーンとオブジェクトパートを仮定し、観察が増加することによりシーンを拡大しオブジェクトの認識を更新する



# Learning category-specific 3D shape models from weakly labeled 2D images

- 従来の2次元画像から3次元モデルを予測する手法の入力は：精密な物体モデル、2次元画像、キーポイントとカメラ姿勢の対応関係
- 弱labeled 2 D画像から 3 次元形状モデルを学習する手法を提案
- (2次元→キーポイント→弱三次元→繰り返し最適化)

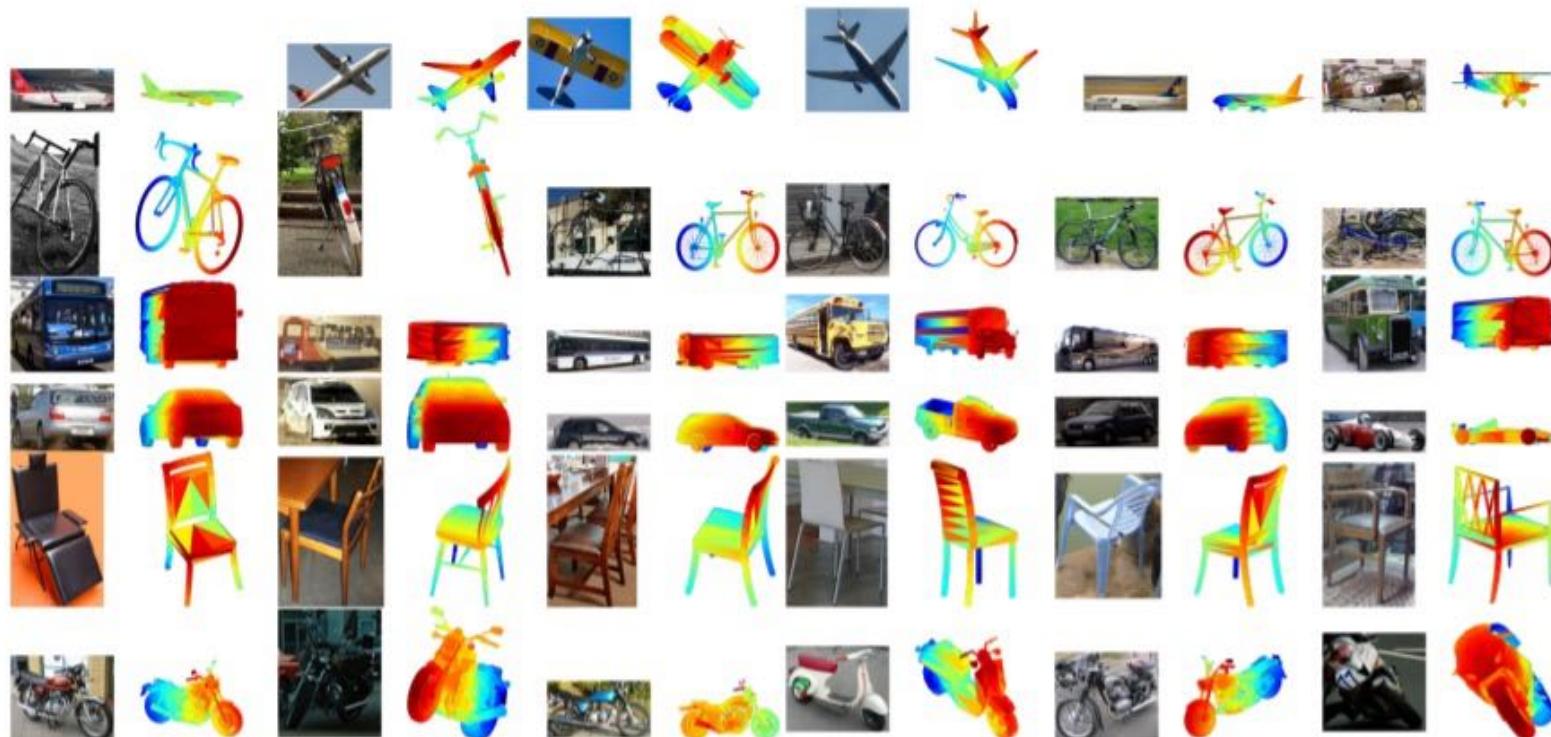
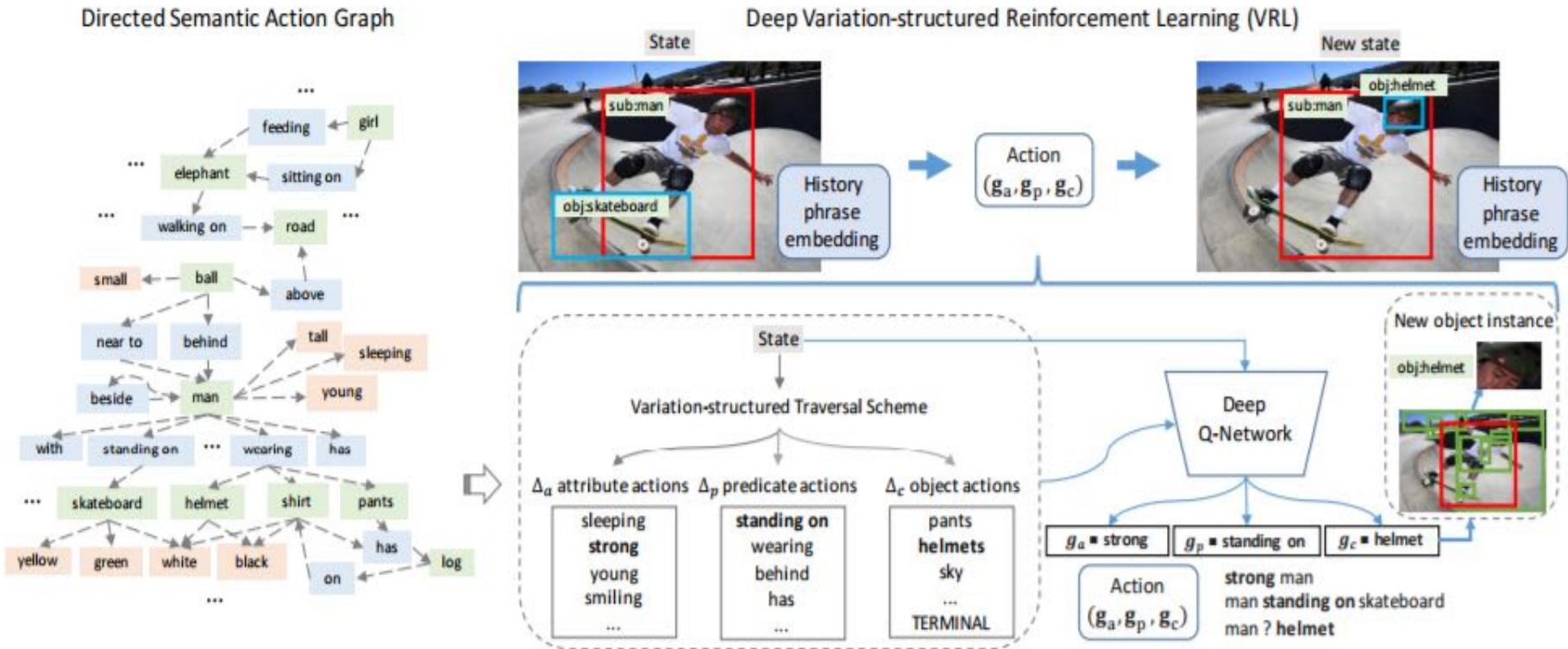


Fig. 6: Viewpoint predictions for unoccluded groundtruth instances using our algorithm. The columns show 15th, 30th, 45th, 60th, 75th and 90th percentile instances respectively in terms of the error. We visualize the predictions by rendering a 3D model using our predicted viewpoint.

[116]

# Deep Variation-structured reinforcement learning for visual relationship and attribute detection

- 画像の中のオブジェクトの関係の研究、従来手法はほとんどグローバルなコンテキストcueをキャプションしない。
- 画像中の物体のグローバルな相互依存性をキャプチャーするため、オブジェクトの関係を順次に探索していく手法VRLを提案した。

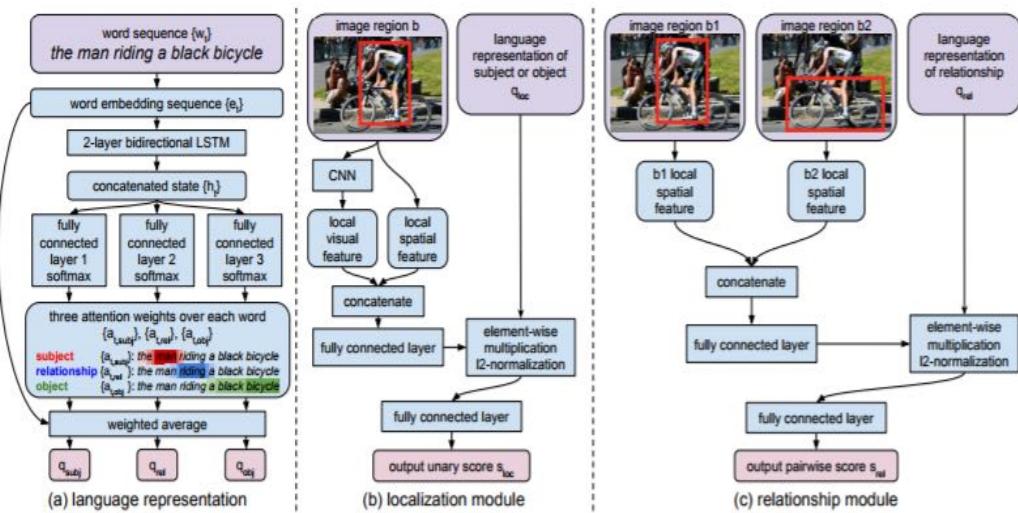
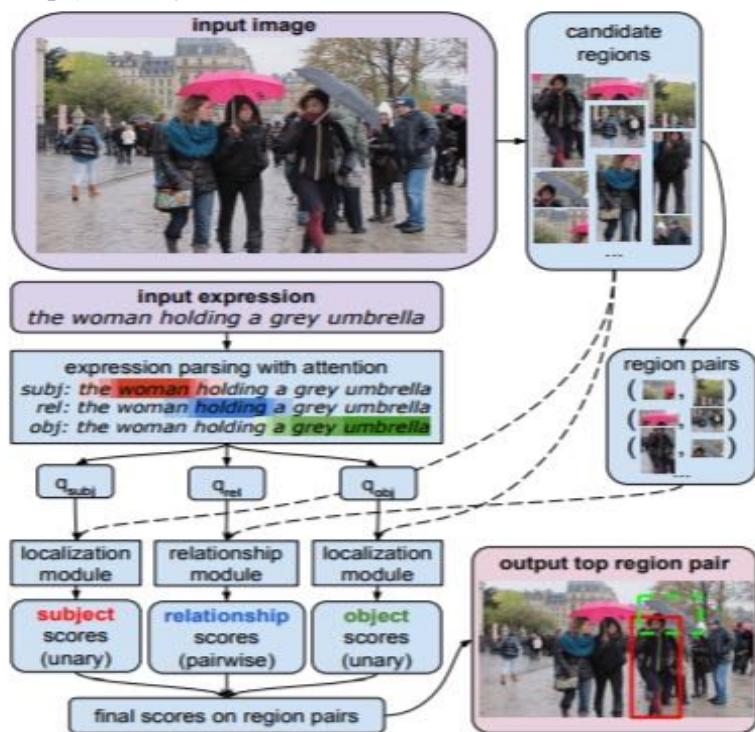


# Modeling relationships in referential expressions with compositional modular networks

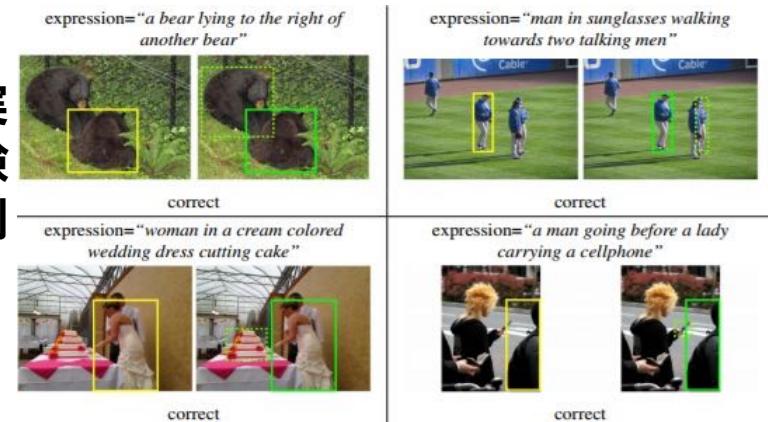
- end-to-end で言語分析と視覚推定を同時に学習する手法CMNを提案した。
- 画像と言語表現ペアを学習、言語からオブジェクトと関係を推定、画像の中からオブジェクトの領域の推定を行う。

具体的構造→

イメージ：



実  
験  
例



[118]

## Guesswhat?! Visual object discovery through multi-modal dialogue

- Guesswhat: 2プレイヤーゲーム、次々と質問することにより画像中にunknown物体を位置を導く。
- 150kデータセット : human-played games (800K 質問、66k画像) を提案

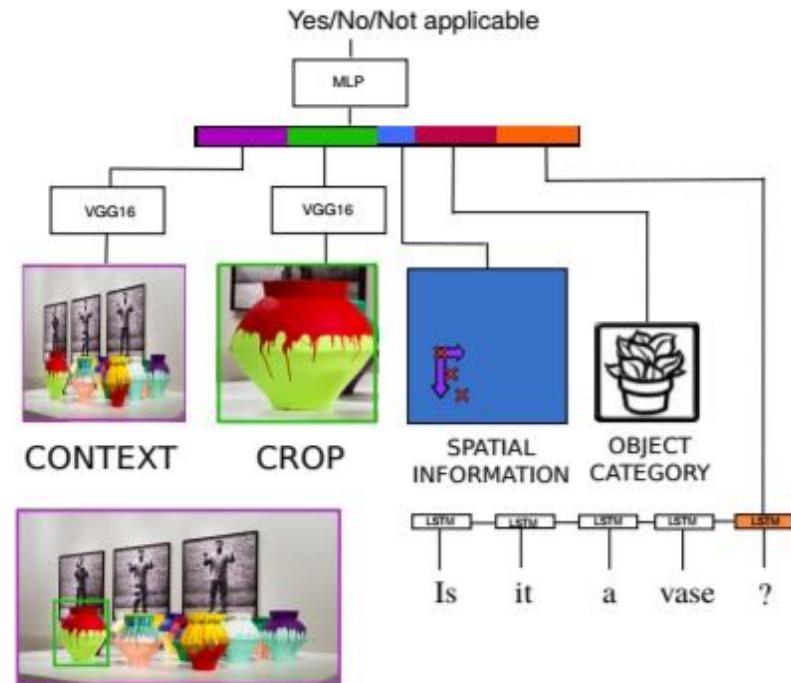


### Questioner

Is it a vase?  
Is it partially visible?  
Is it in the left corner?  
Is it the turquoise and purple one?

### Oracle

Yes  
No  
No  
Yes



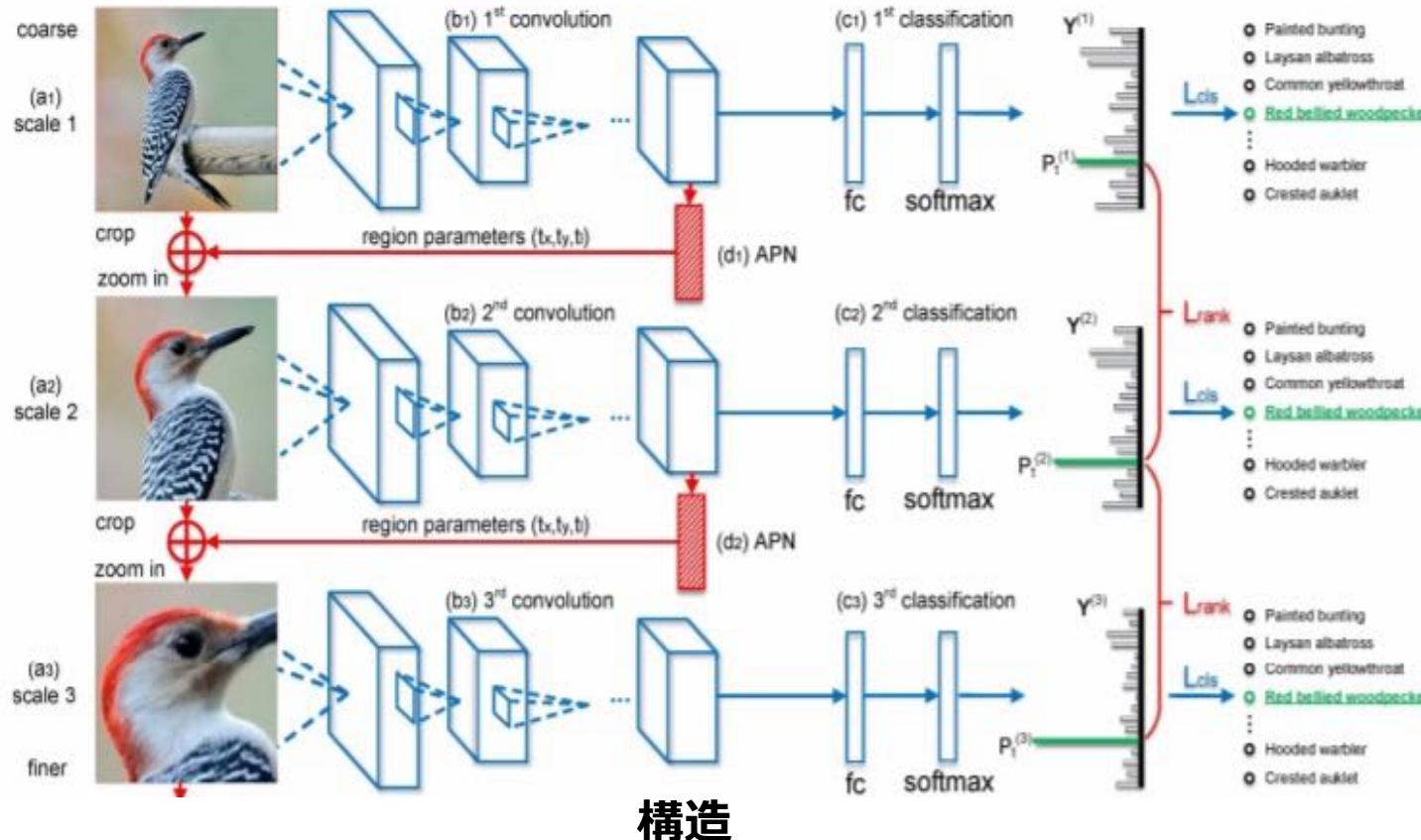
ゲーム例

画像+質問  
+クロップ+空間+カテゴリ  
構造

[119]

# Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition

- 従来のfine-grained(きめ細かい)カテゴリ認識は、識別領域ローカリゼーションとfine-grained(特徴の学習の2つのタスクを分けて解決する
- recurrent attention CNN(RA-CNN)を提案し、マルチスケールで (coarse-to-fine構造) 相互に強化された方法で上記の2つのタスクを同時に再帰的に学習



**stanford  
dog:dataset**

Approach	Accuracy
NAC (AlexNet) [26]	68.6
PDFR (AlexNet) [34]	71.9
VGG-16 [27]	76.7
DVAN [35]	81.5
FCAN [20]	84.2
RA-CNN (scale 2)	85.9
RA-CNN (scale 3)	85.0
RA-CNN (scale 1+2)	86.7
RA-CNN (scale 1+2+3)	87.3

**stanford  
car:dataset**

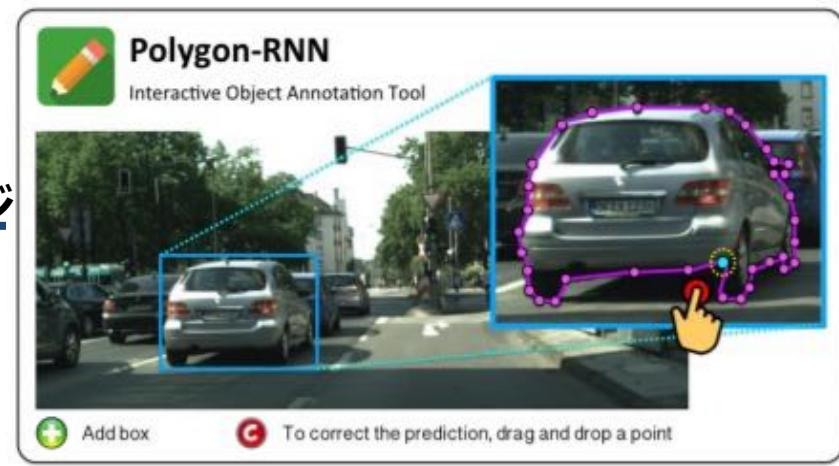
Approach	Train Anno.	Accuracy
R-CNN [7]	✓	88.4
FCAN [20]	✓	91.3
PA-CNN [14]	✓	92.8
VGG-19 [27]		84.9
DVAN [35]		87.1
FCAN [20]		89.1
B-CNN (250k-dims) [19]		91.3
RA-CNN (scale 2)		90.0
RA-CNN (scale 3)		89.2
RA-CNN (scale 1+2)		91.8
RA-CNN (scale 1+2+3)		92.5

[120]

## Annotating object instances with a polygon-RNN

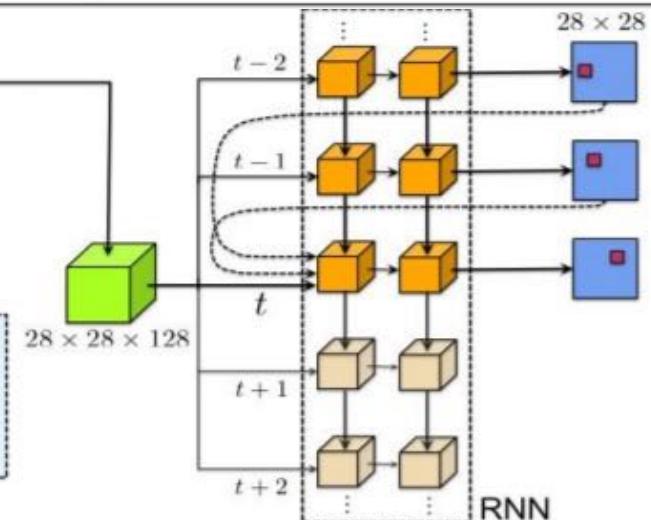
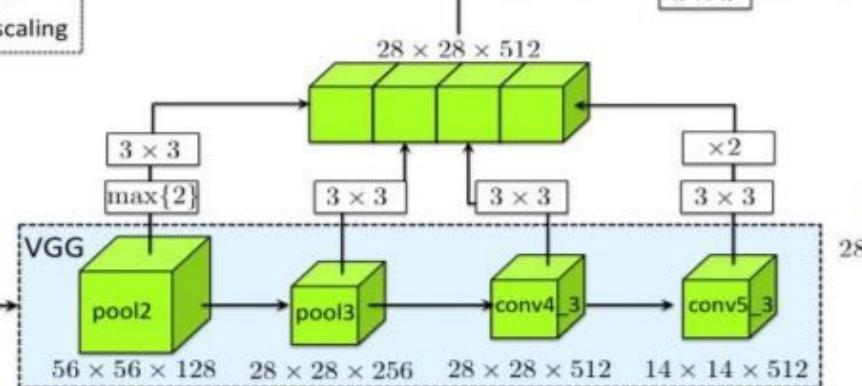
- ポリゴンを生成してポリゴンで囲まれる部分もピクセルレベルのラベルつき、しかし semantic segmentation より作業量が大幅に軽減 → 半自動物体インスタンス annotation とポリゴン生成システム Polygon-RNN tool を提案
- ユーザ入力： 物体のバウンディングボックス（システムでは2回クリックして生成できる）  
→ 出力： ポリゴン物体領域

### ツールのイメージ



### 半自動ポリゴン領域生成構造

$3 \times 3$	... convolution ( $3 \times 3$ )
$\max\{2\}$	... max-pooling ( $2 \times 2$ )
$\times 2$	... bilinear upscaling



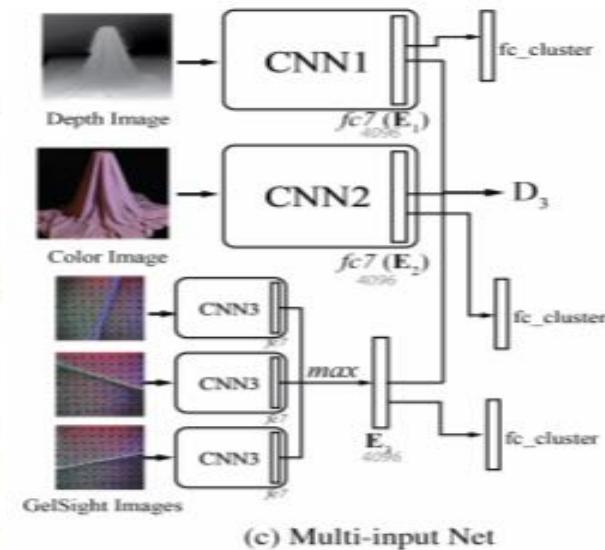
[121]

## Connecting look and feel: associating the visual and tactile properties of physical materials

- 布の生地（118種収集）を実験対象とし、touch（触感）とビージョンの対応付けを学習
- GelSight sensor(物体をタッチし、触感のデータを画像にするセンサー)を用い、布のtouch(触感)画像を生成
- 布の触感画像、色画像、デプス画像の対応付けを学習するために、joint NNを用い、同じ布に対し触感画像、模様画像、デプス画像から特徴をできるだけ近くすることを学習



touch(触感)センサー、  
touch(触感)画像



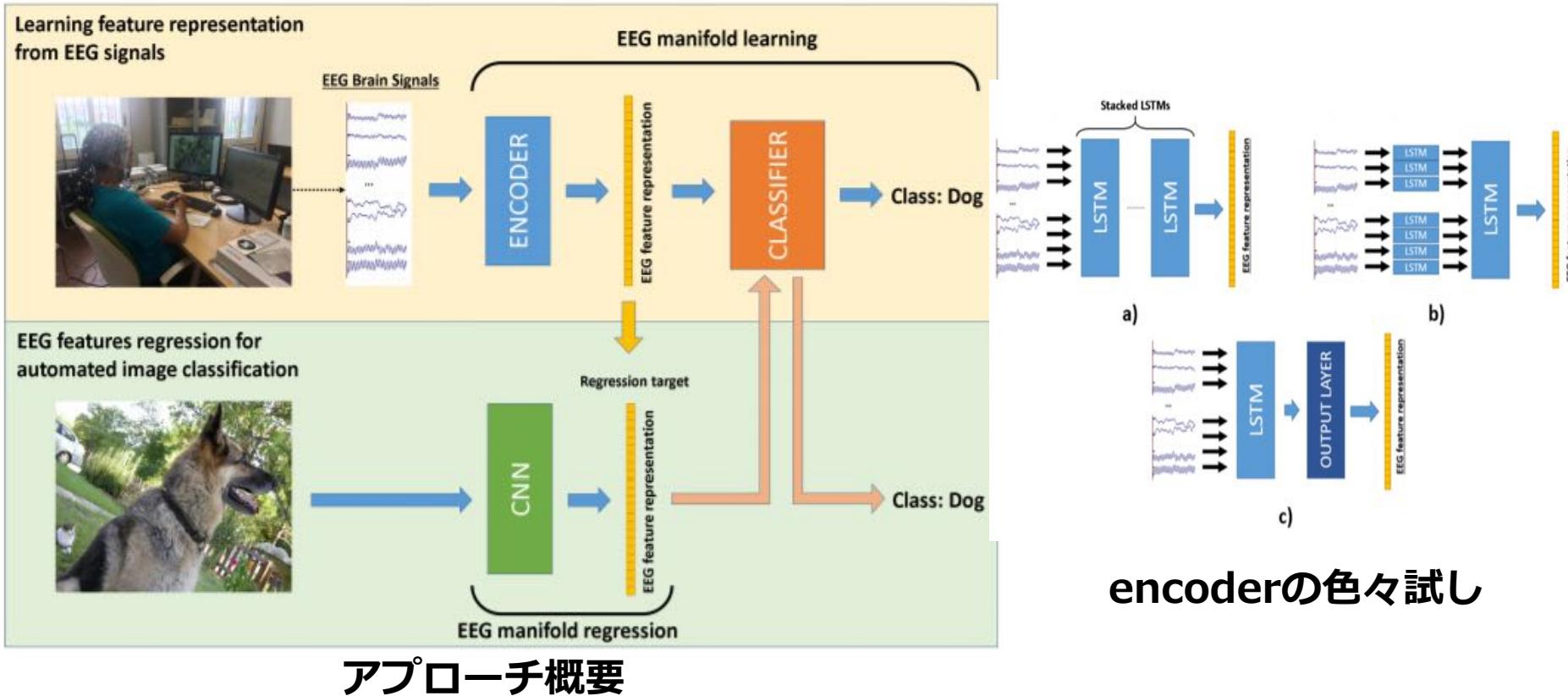
joint Neural Network :  
提案手法のネットワーク構造

$$D_3 = \|E_1 - E_2\| + \|E_2 - E_3\| + \|E_3 - E_1\| \quad (1)$$

[122]

# Deep learning human mind for automated visual classification

- 初めての：脳信号からのコンピュータビージョンアプローチ
- 初めての：直接人の神経から視覚descriptorsを抽出し分類に用いる研究
- 性能がstate-of-the-art
- 現状最大な視覚物体分析用のEEG（脳波）データセット公開



【123】

## Learning to align semantic segmentation and 2.5D maps for geolocalization

- 都市の画像からポーズ情報を推定する
- "動かない" () と"rotateしない"方向によって、ポーズが得られる
- 2つのネットワークをトレーン： (1) センサーポーズからのポズ予測； (2) セマンティックマスクからのポズ予測。2つの推定結果が一致になるようにポズを更新。

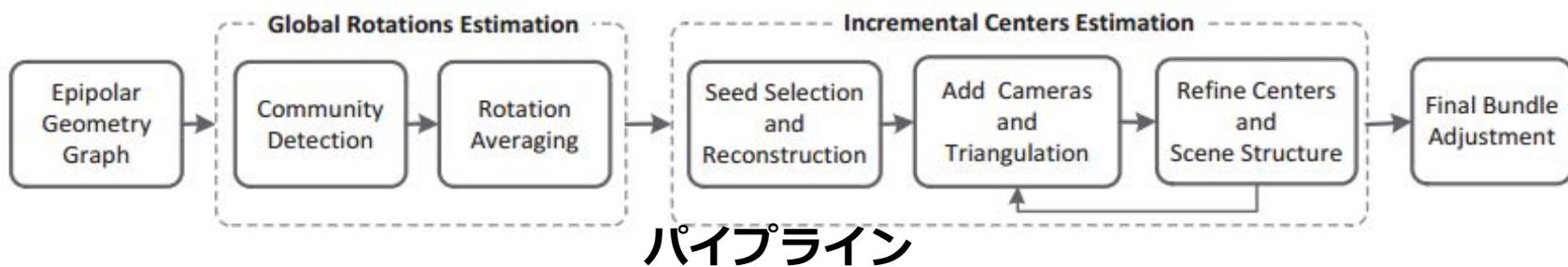


(ポズ更新のiteration, 右列: 最終結果)

【124】

## HSfm: Hybrid structure-from-motion

- 従来のSfM手法が初期カメラ姿勢推定の違いから : incrementalSfM (口バスト、正確、遅い) とグローバルSfM (速い、口バストと正確さがよくない) の2種類に分けられる。
- epipolar geometry graphによってカメラ位置を推定し、そのうえグローバルリコンストラクションというハイブリッドなSfMを提案した (口バストで、正確、incrementalSfMより速い)
- グローバルリコンストラクションを行前、カメラ姿勢をカメラ間距離の差などによりグループ化し、グループ内で先にalignすることにより口バスト化と正確さを高める



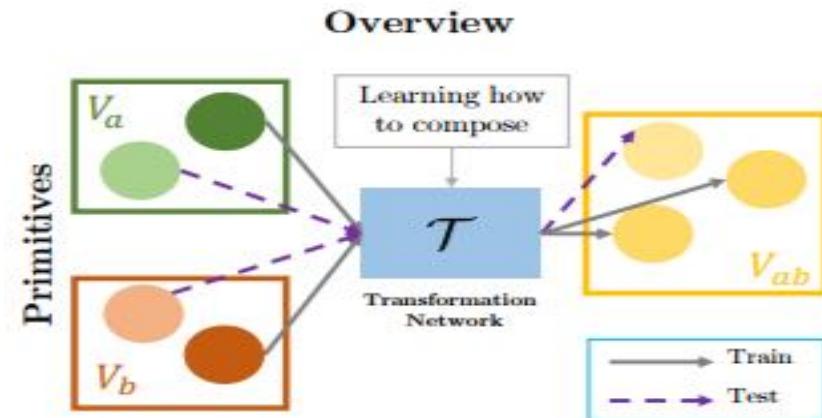
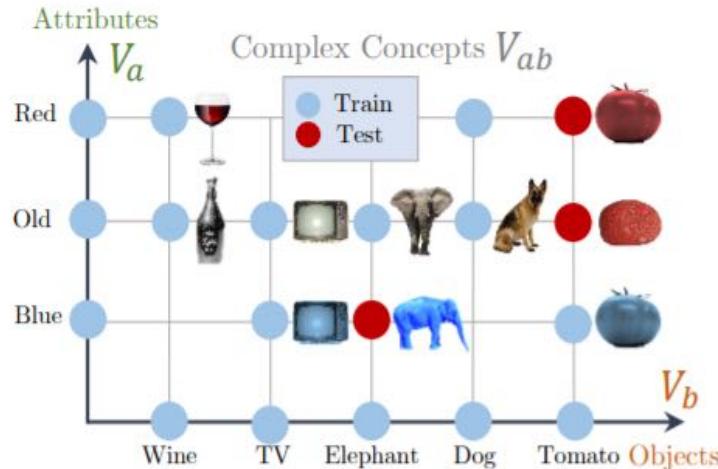
	COLMAP	LUD	Theia	Our HSfm	Bundler
Gendar-menmarkt					
Temple					

効果

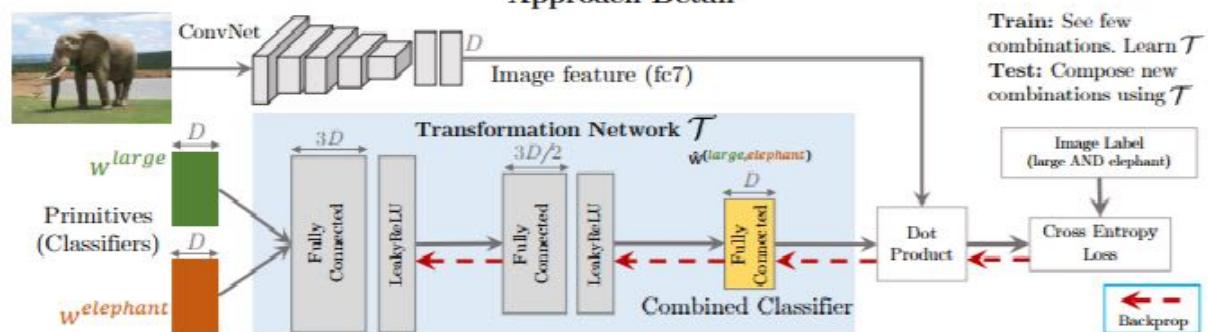
【125】

## From red wine to red tomato: Composition with context

- 既知の視覚コンセプトの弁別器を生成する手法を提案した
- 弁別器が構成的変換をモデル化することができるスムースな空間上に分部することが提案手法のベース
- 提案手法が特徴合成、オブジェクト合成、subjectの合成と予測に用いられる



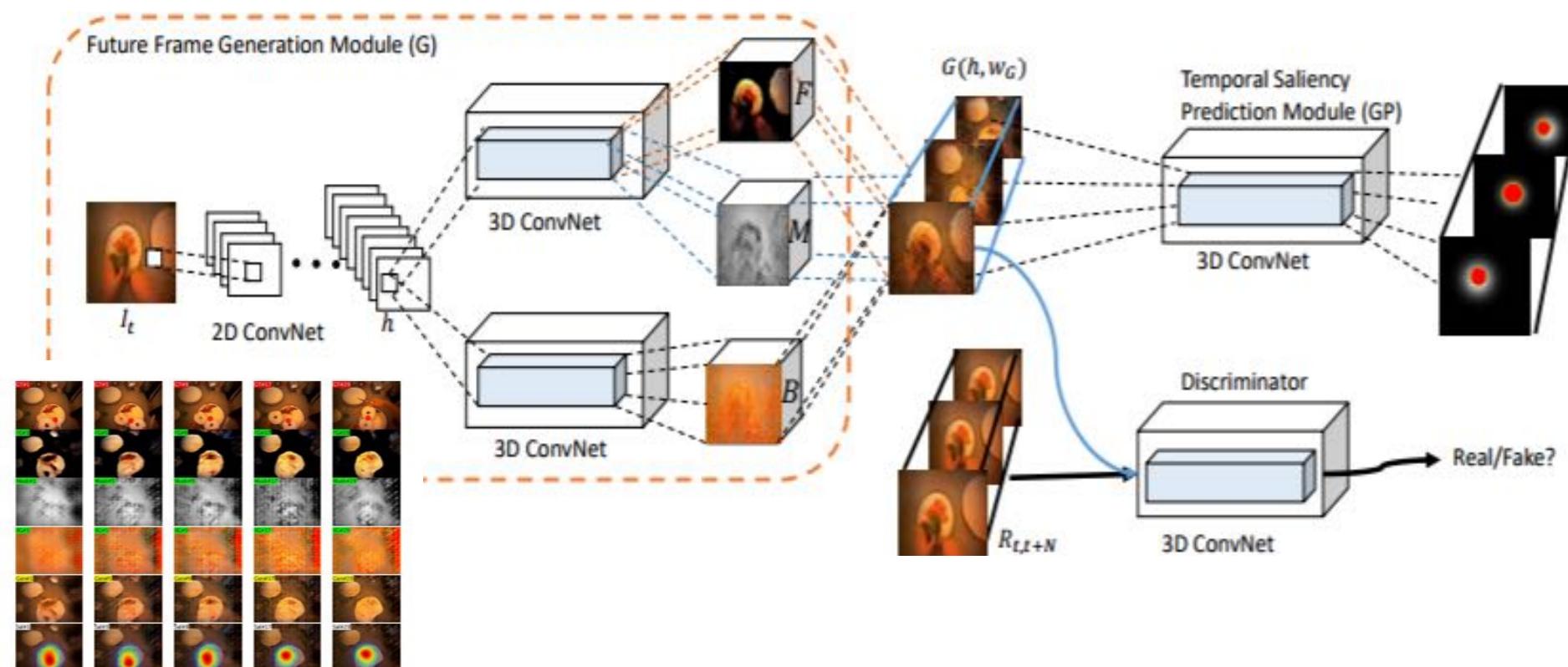
複雑なビージョンコンセプトが簡単な  
ビージョンコンセプトから合成できること  
を仮定する



[126]

## Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks

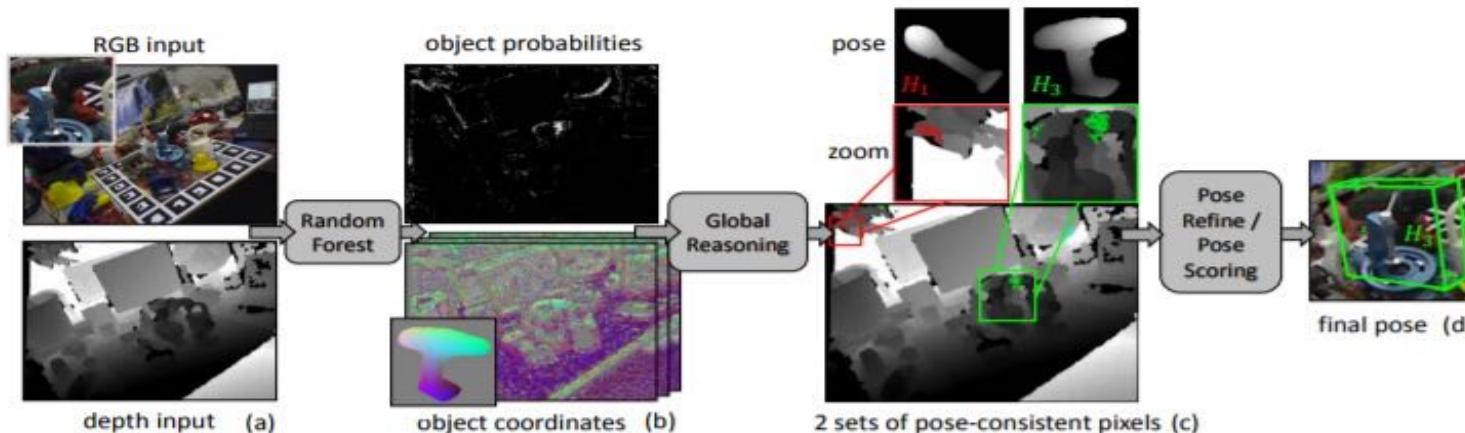
- 手持ちカメラから撮ったビデオから視線を予測する手法を提案した
- GANをベースにDeep Future Gaze (DFG) を提案した。 DFGが一つのフレームから数フレームを生成し、同時に未来の数秒の視線を予測する。
- DFGモデル構造 : G : 2 stream : 3D CNN + CNN (背景前景分ける) ; D : 3 DCNN  
(vganとほぼ同じような気がする)



【127】

## Global hypothesis generation for 6D object-pose estimation

- 1枚のRGB-D画像から既知の物体の6D姿勢を求める
- 従来手法の一般的な流れ：（1）局所的特徴量を求める；（2）ポーズ－仮定のプールを生成する；（3）プールから姿勢を選択と最適化。従来手法の（2）のところ、現状は全部ローカルな推定を生成している。
- 提案手法が前述のところに対し、局所的ではなく、fully connected CRFにより全局な推定を生成する。スピードが従来手法より大幅に上昇した



パイプライン

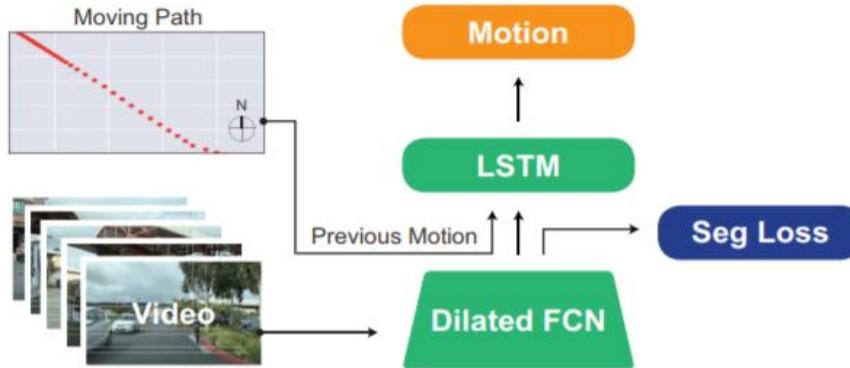
精度結果

Object	Method	Scores			
		Our method	Hinterstoisser et al.[10]	Krull et al.[18]	Brachmann et al.[3]
Ape		80.7%	<b>81.4%</b>	68.0%	53.1%
Can		88.5%	<b>94.7%</b>	87.9%	79.9%
Cat		<b>57.8%</b>	55.2%	50.6%	28.2%
Driller		<b>94.7%</b>	86.0%	91.2%	82.2%
Duck		74.4%	<b>79.7%</b>	64.7%	64.3%
Eggbox		47.6%	<b>65.5%*</b>	41.5%	9.0%
Glue		<b>73.8%</b>	52.1%	65.3%	44.5%
Hole Puncher		<b>96.3%</b>	95.5%	92.9%	91.6%
Average		<b>76.7%</b>	76.2%	70.3%	56.6%

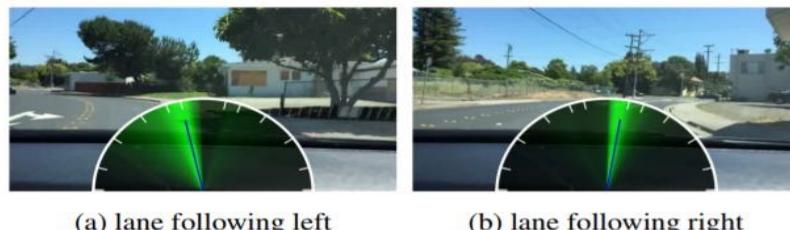
【128】

## End-to-end learning of driving models from large-scale video datasets

- 大規模な学習データから車載RGBカメラのcurrent写真と過去の運行状態から未来の運行方向を予測するend-to-endなカメラネットワークを提案した。
- 新しいFCN-LSTMネットワークを提案し、privileged学習を用いてそのネットワークは学習可能です。
- 大規模なcrowd-sourced driving behaviorデータセットBDDVを提案した。

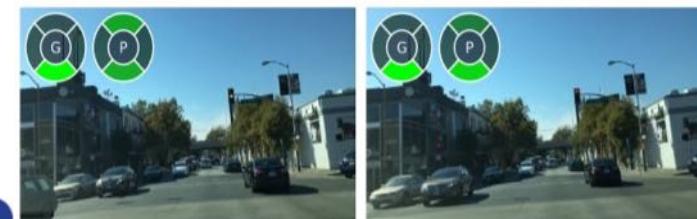


提案したdriving behavior予測モデル



(a) lane following left

(b) lane following right



(a) go at yellow light

(b) stop at red light



(c) stop & go equal weight at medium distance  
(d) stop when too close to vehicle ahead

果図

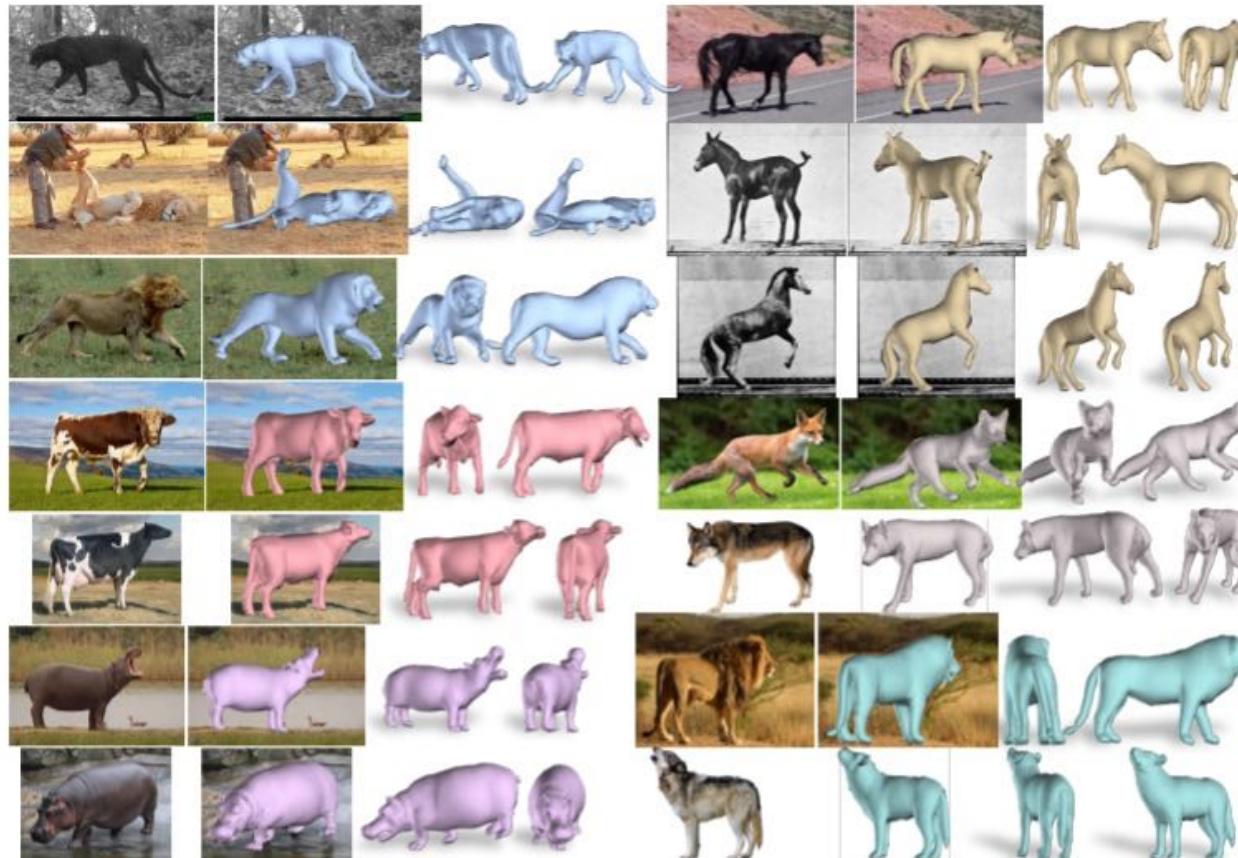
離散driving behavior予測効

連続driving behavior  
予測効果図

【129】

## 3D Menagerie: Modeling the 3D shape and Pose of animals

- 動物の3次元shapeとposeを推定するモデルを提案した
- 5種類の動物（猫科動物、犬、馬、牛、カバ）のおもちゃ（毛がなし）をスキヤンしモデルを集めた（**データセット公開**）
- Align のGlobal-Local Shape Space model(GLOSS)手法を提案、PCAにより形状をclusteringし、それにより**unseen**動物のshapeとposeを推定可能

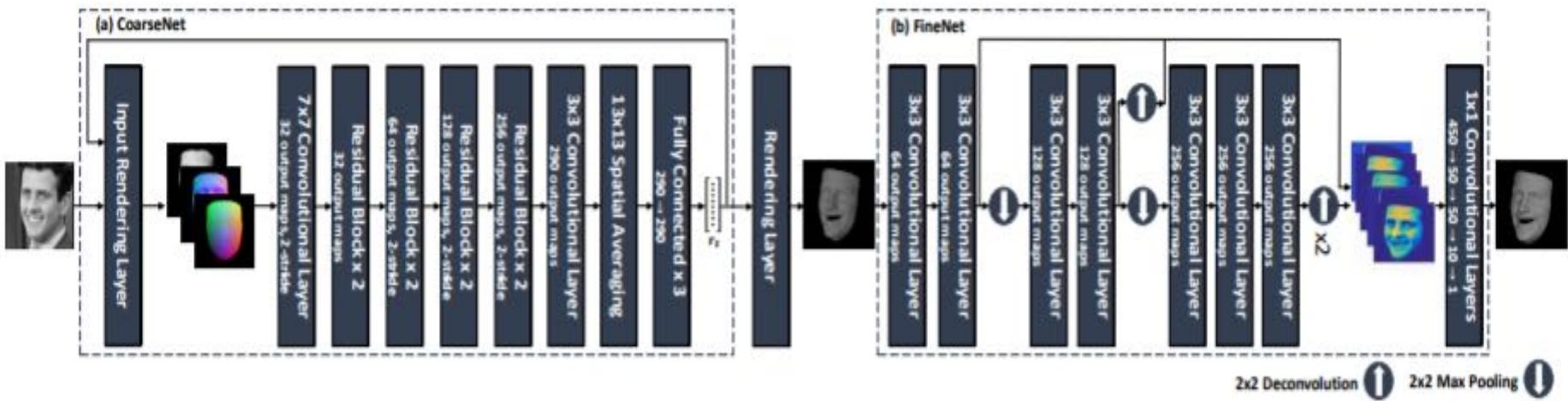


提案モデルの結果

## Learning Detailed face reconstruction from a single image

- end-to-endのDetailed 顔三次元モデルリコンストラクション手法を提案
- 提案手法のメインアイデアは：CoarseNetとFineNetの2段階でリコンストラクションをし、一段階目では形状情報を把握し、2段階目でより細かいところを修正

### ネットワーク構造



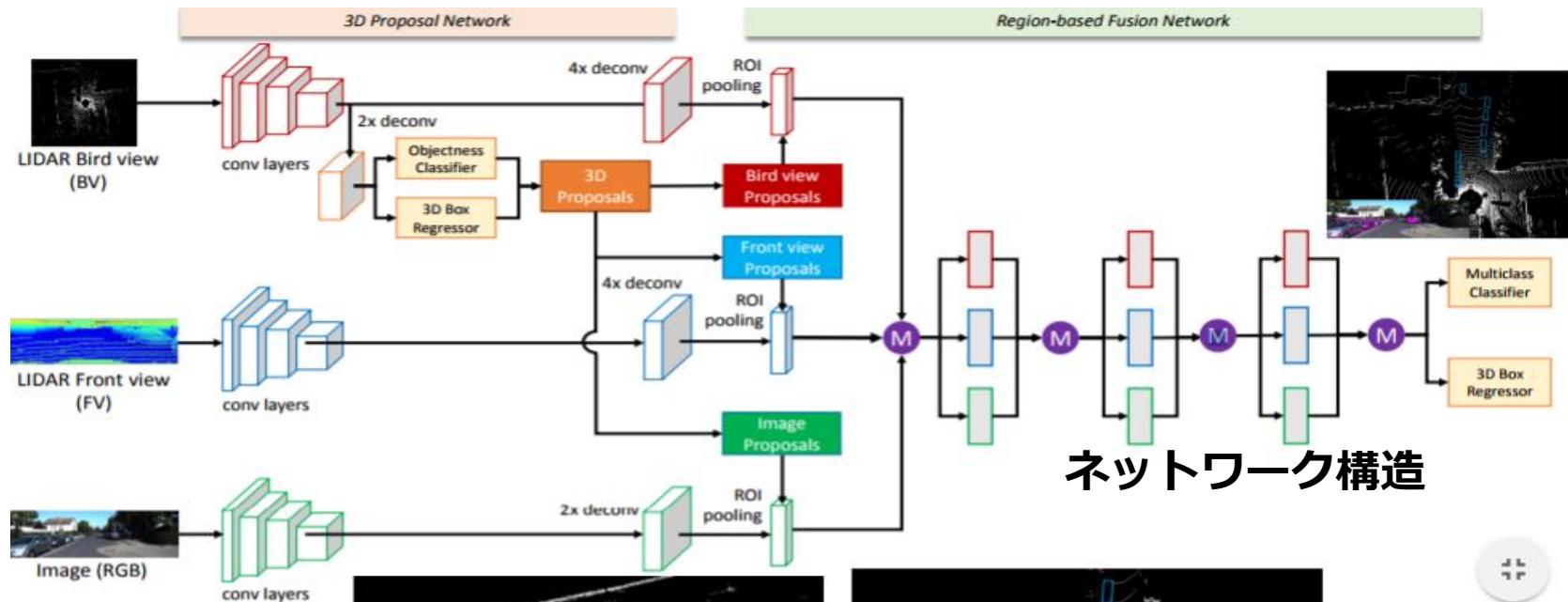
### 提案手法の表現



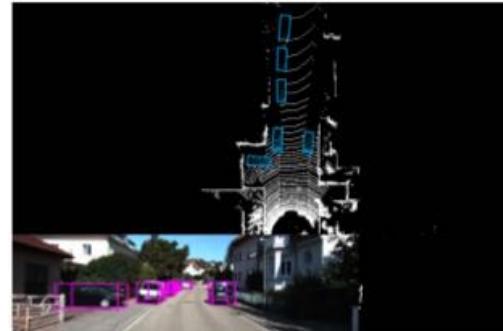
[131]

## Multi-view 3D object detection network for autonomous driving

- 初めての、monocularカメラとLIDARをフージョンするend-to-endの三次元 detection(三次元バウンディングボックス検出)手法を提案した
- 三次元detectionに優れた精度を達した (Recall:99.1% (IoU=0.25) , Recall:91.0% (IoU=0.5) )



提案手法の効果

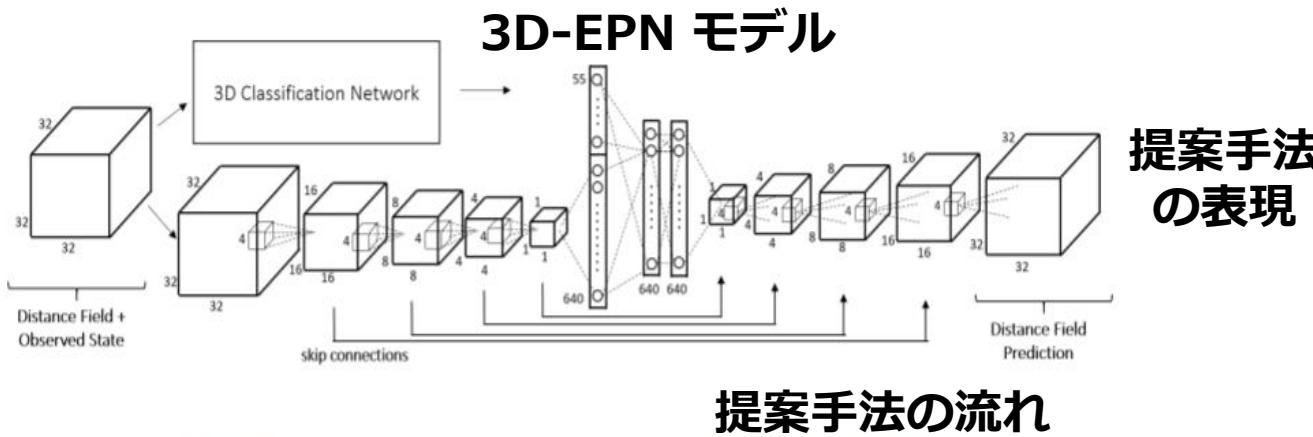


[132]

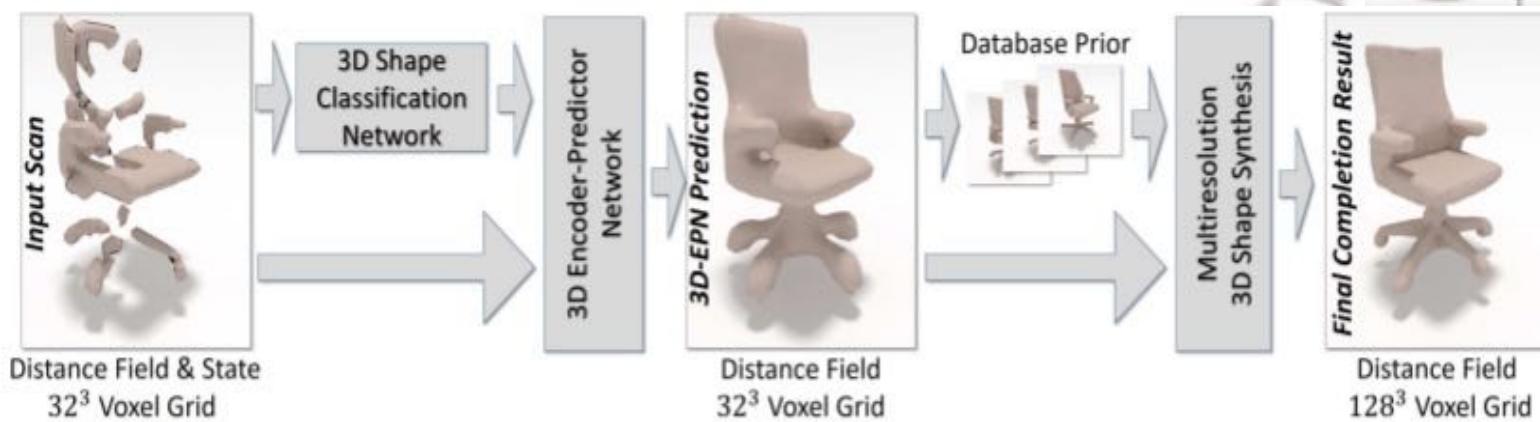
## Shape completion using 3D-encoder-predictor CNNs and shape synthesis

- data-drivenの3次元形状が欠けているから完全な形状を生成する3D Encoder-predictor (3D-EPN) ネットワークを提案
- 提案手法がモデルのグローバル形状復元を高解像度でできる
- 生成できるモデルが細かくないが、形状がcomplete

ground  
truth  
Ours  
Input



### 提案手法の流れ

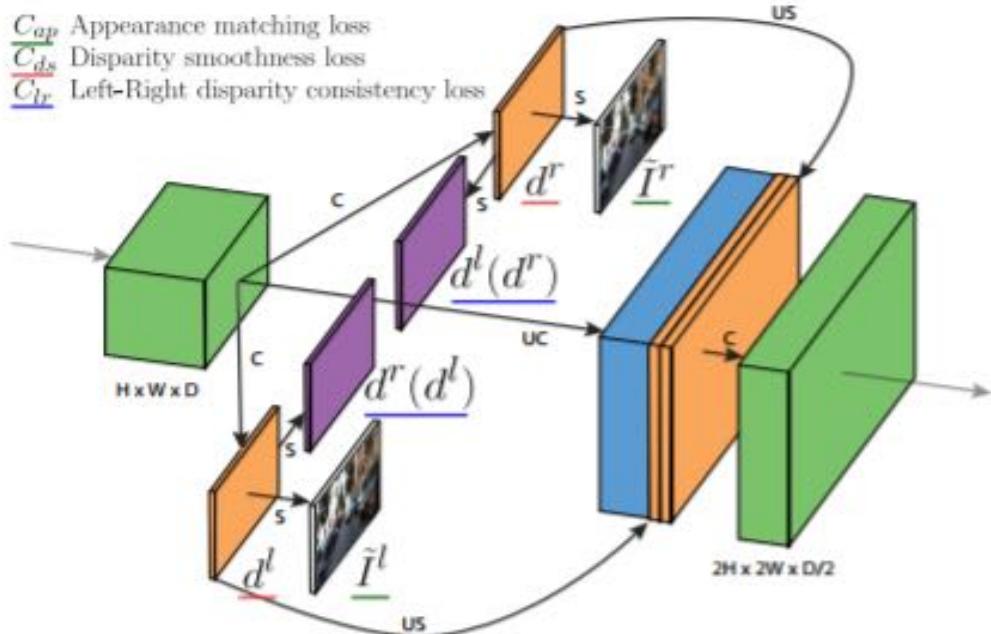


[133]

## Unsupervised Monocular Depth estimation with left-right consistency

- stereo RGB画像からunsupervisedでデプス画像を推定
- supervised手法よりも高い精度でデプス画像推定
- 2枚の画像（左、右）からのstereo視、一枚をCNNに入力し学習してデプス推定（左右 Loss, smoothness Loss）、推定した結果にbilinear samplingを用いりコンストラクション（Reconstruction Loss）。その3つのLossでLoss関数構成。

### ネットワーク構造



### 提案手法の結果



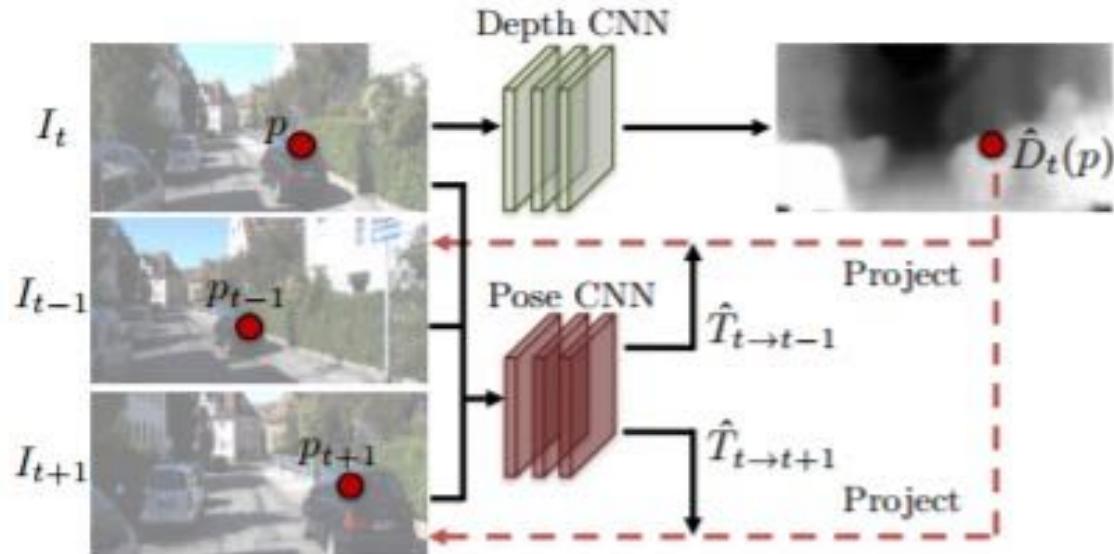
ground-truth  
Ours

supervision!

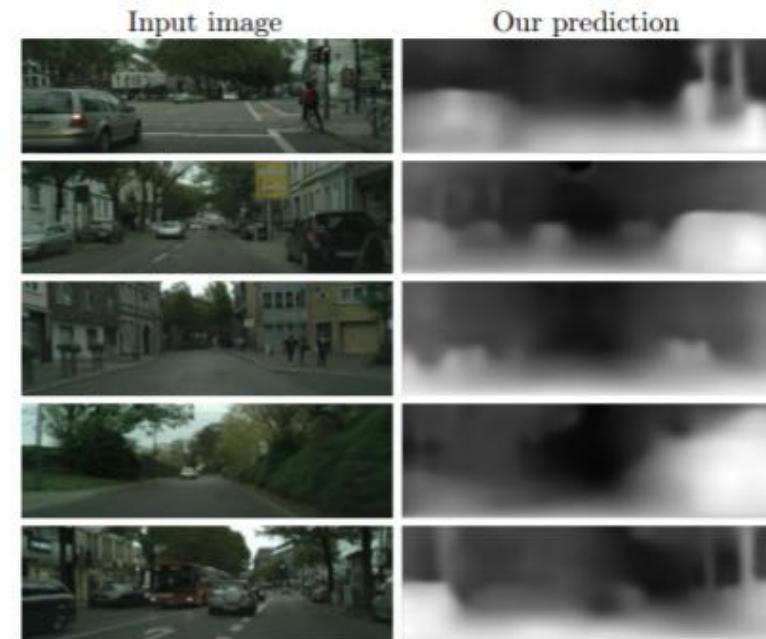
# Unsupervised learning of depth and Ego-motion from video

- 人が2次元投影から現実世界を観察し、3次元を感知できる。これがモチベーションで、3D情報のunsupervisedの学習手法を提案した
- unlabeledビデオからego-motionとデプス画像を同時に学習する
- 同時にデプスとポズを推定し、推定結果を用い投影し元画像と比べ、ロスで推定更新

提案手法のパイプライン

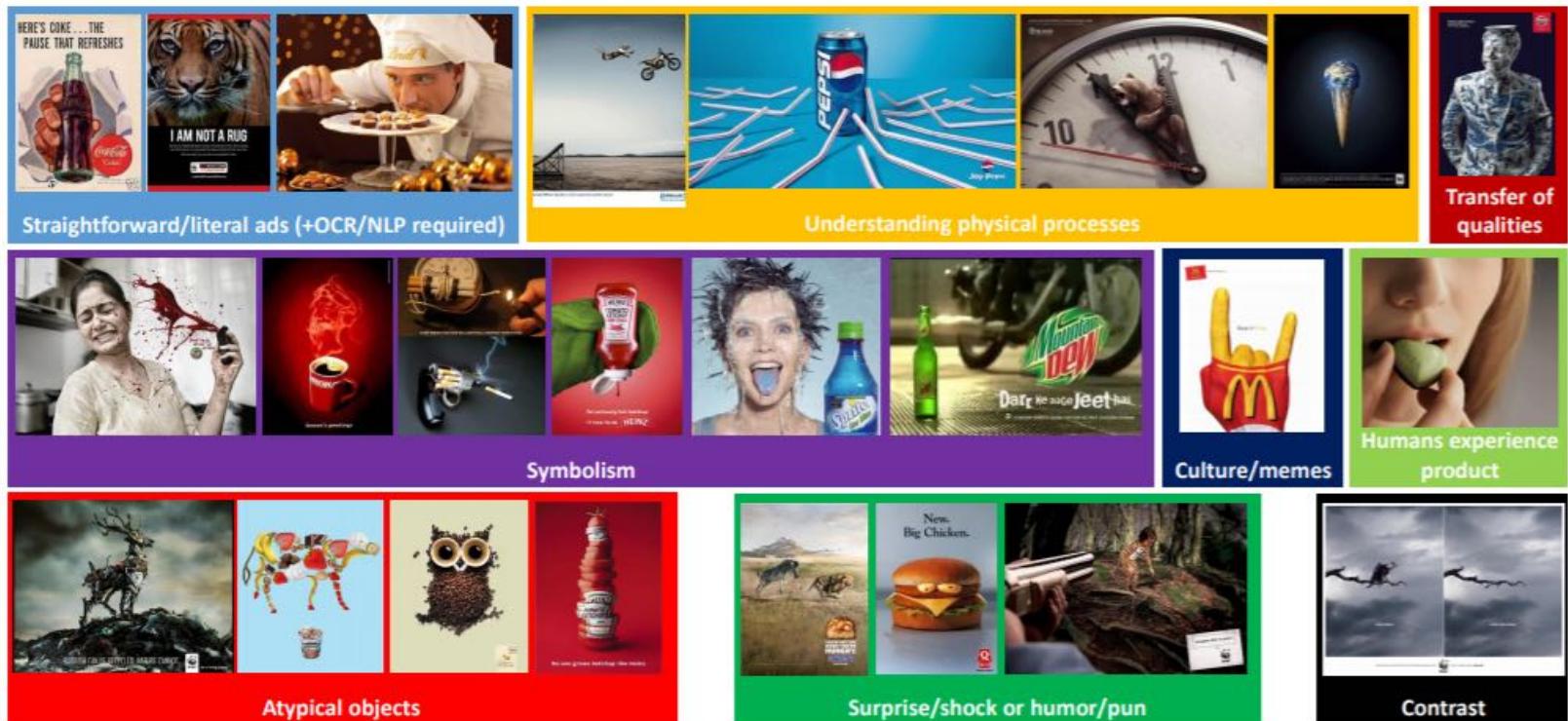


デプス推定結果例



## Automatic Understanding of Image and Video Advertisements

- 約6万枚の広告画像/動画(CM)のデータセット
- 広告の目的(メッセージ性)を予測するためのフレームワークの提案



# One-Shot Metric Learning for Person Re-identification

- 色とテクスチャの距離学習によるRe-ID
- 距離学習には1組分の画像とカラーチャートのみであるため大規模データを必要としない

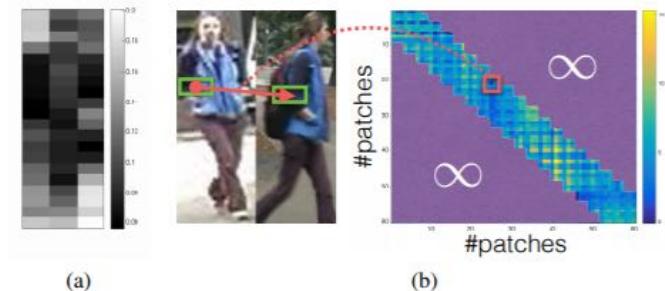
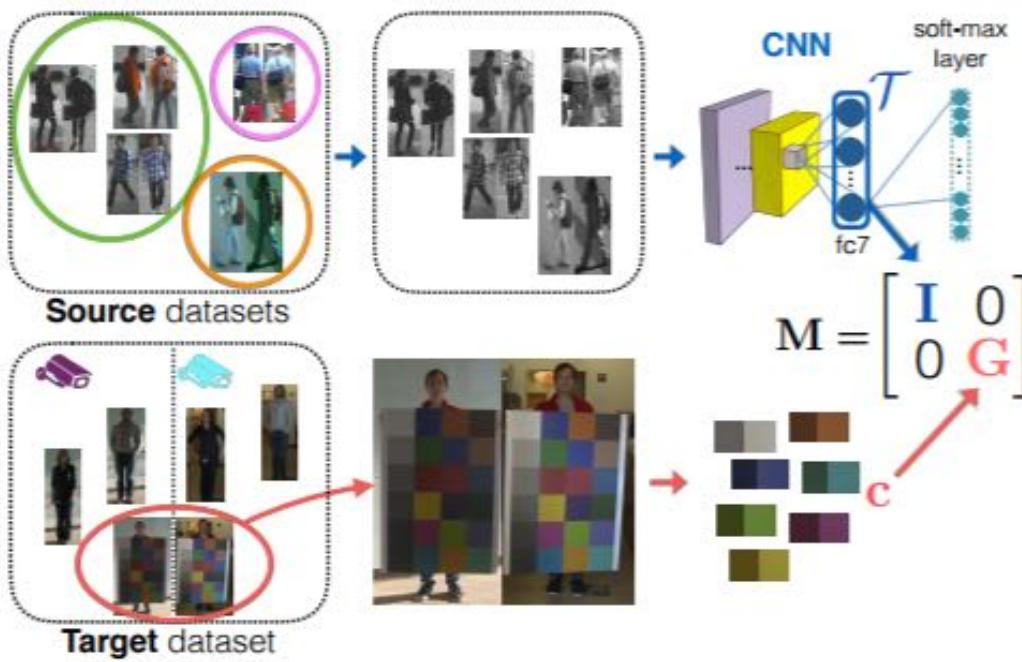
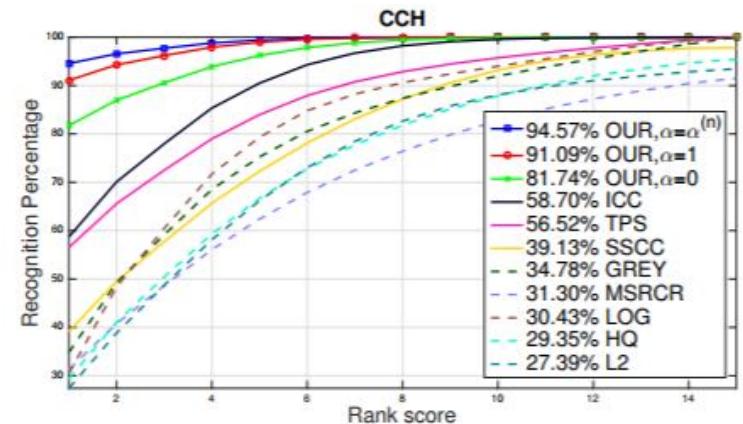
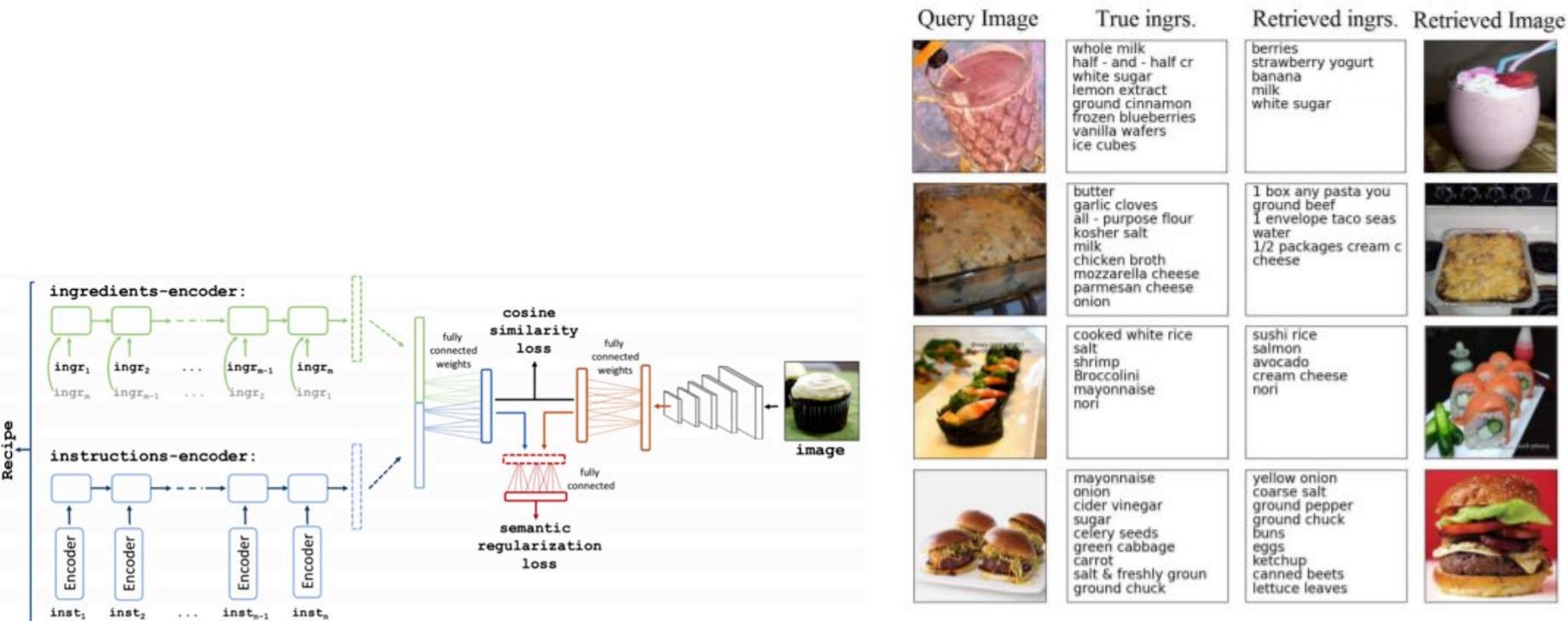


Figure 3: Spatial variations: (a) learned background distortion coefficients  $\alpha^{(n)}$ ; (b)  $N \times N$  cost matrix, which is used as an input to the Hungarian algorithm for finding optimal patch correspondence.



## Learning Cross-modal Embeddings for Cooking Recipes and Food Images

- 画像によるレシピの検索手法の提案
  - レシピはword2vecおよびLSTMによって生成
- 1M以上のレシピと800kの料理画像を含むRecipe1Mを公開



## Growing a Brain: Fine-Tuning by Increasing Model Capacity

- Fine-tuning方法の検討
- Layerの増加、拡張のうち精度向上に貢献する方法を実験的に紹介

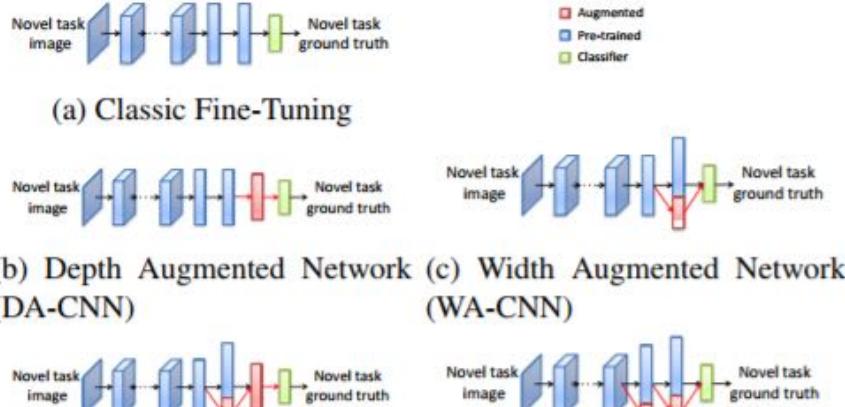
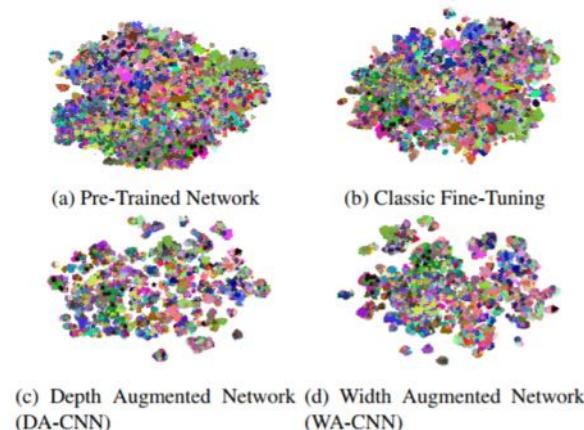


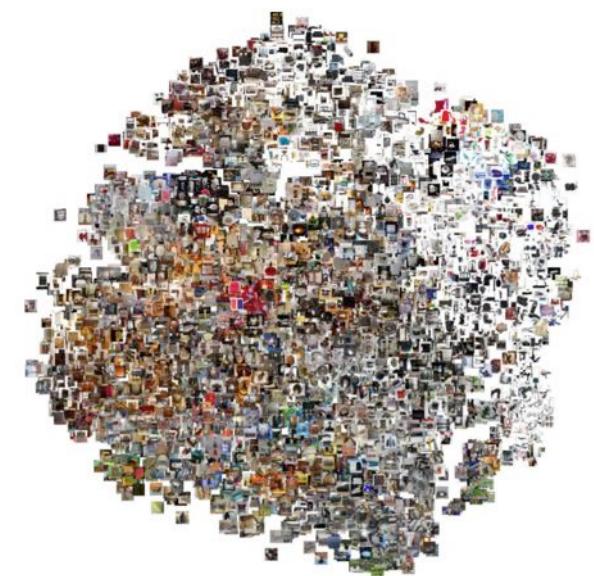
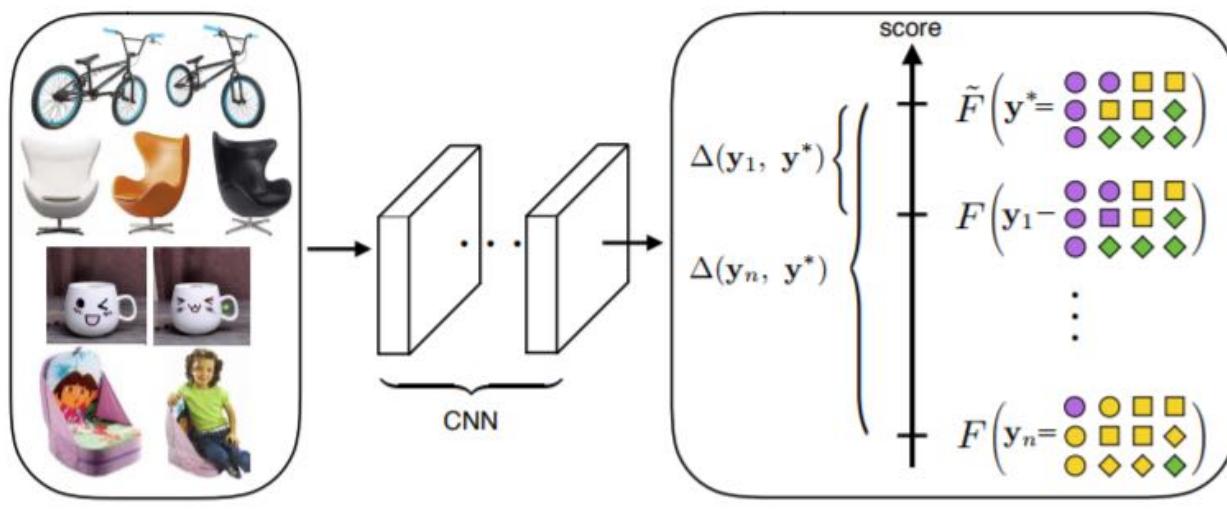
Figure 2: Illustration of classic fine-tuning (a) and variations of our developmental networks with augmented model capacity (b-e).

Network	Type	Method	Acc (%)			
			New	FC <sub>7</sub> -New	FC <sub>6</sub> -New	All
AlexNet	Baselines	Finetuning-CNN [1, 15]	53.63 48.4	54.75 —	54.29 51.6	55.93 52.2
	Single (Ours)	DA-CNN WA-CNN	54.24 <b>56.81</b>	56.48 56.99	57.42 57.84	58.54 58.95
	Combined (Ours)	DWA-CNN WWA-CNN	56.07 56.65	56.41 <b>57.10</b>	56.97 <b>58.16</b>	57.75 <b>59.05</b>
VGG16	Baselines	Finetuning-CNN	60.77	59.09	50.54	62.80
	Single (Ours)	DA-CNN WA-CNN	61.21 <b>63.61</b>	62.85 <b>64.00</b>	63.07 <b>64.15</b>	65.55 <b>66.54</b>



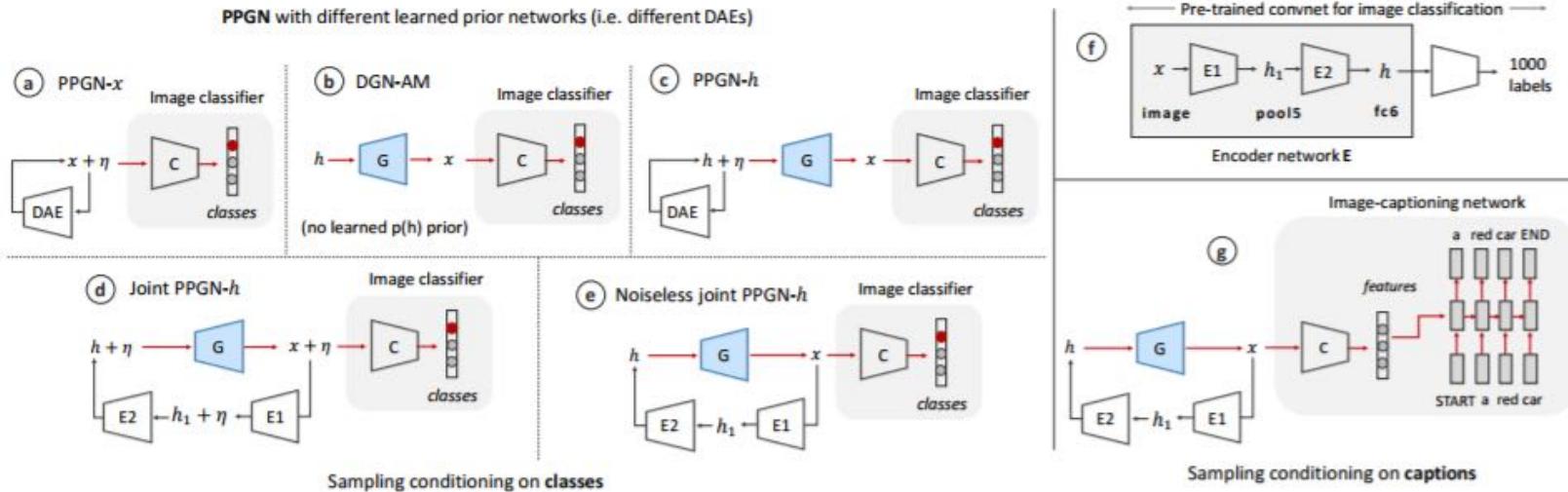
## Deep Metric Learning via Facility Location

- 学習可能なクラスタリングとメトリックを用いてデータセット全体のメトリックをEnd-to-Endで最適化
  - バッチごとのクラスタリングとメトリック
- CUB200-2011, Cars196、Stanford Online Productではそれぞれクラスタリングと検索タスクの両方でベースラインを更新



## Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space

- 高解像度な画像生成手法が可能となるPlug & Play Generative Networksの提案



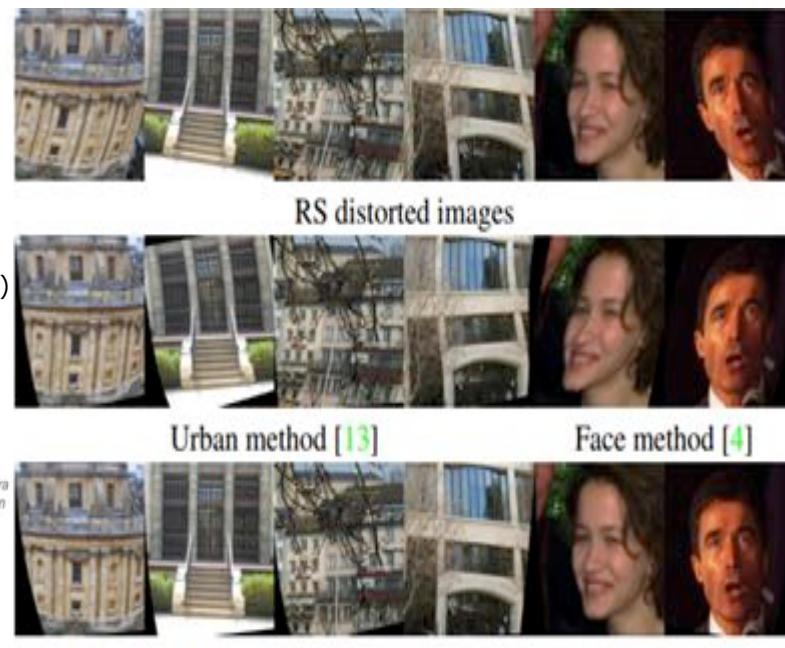
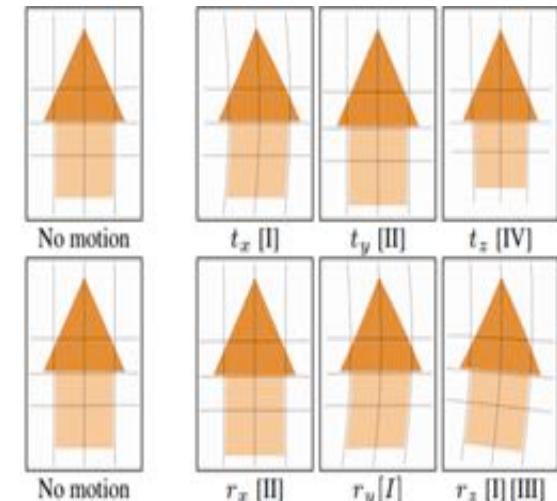
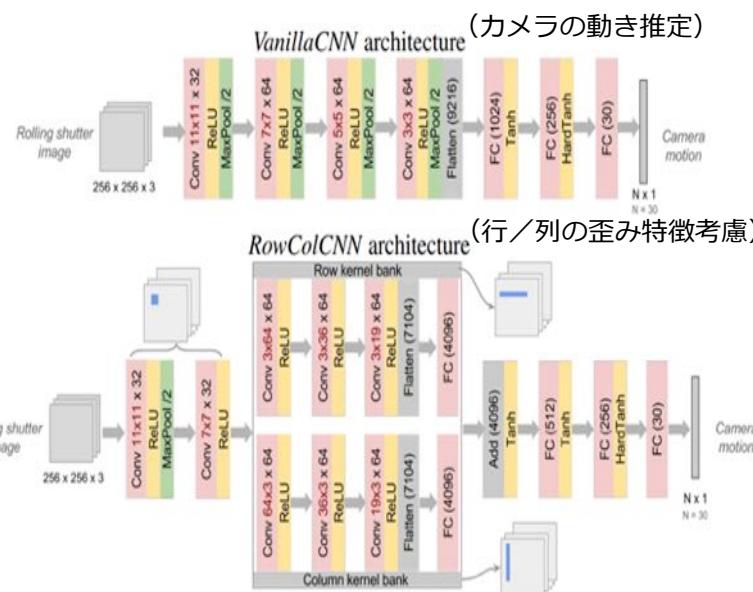
Keywords: rolling shutter correction

## 概要

CMOSカメラのようなローリングシャッターカメラでは、カメラが動いている状況で行ごとの歪みが発生するローリングシャッタ現象がある。これに対し、単画像に対してCNNにより歪み特性を自動で学習し、最初の行に合わせるように補正を行う手法を提案する。長い矩形のカーネルが行単位の歪みを学習する。センス・スペシフィックであった従来手法に対し、より挑戦的な状況においてよい性能を示す。

## 新規性・差分

- 特定のシーン特徴に合わせた抽出でなく、またカメラパラメータの事前知識も無しに実現する、単画像ローリングシャッタ歪みの補正
- 露光のメカニズムにあわせ、動きと、行／列方向の両方のゆがみ特徴を学習する新たな2つのCNNの提案



## Links

### 論文

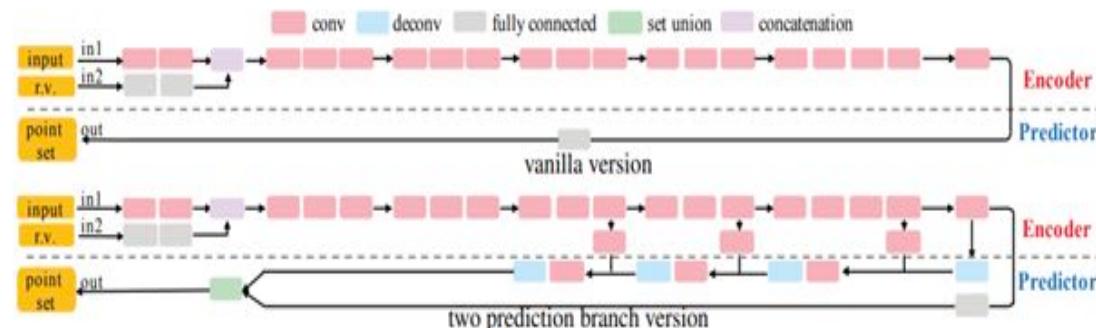
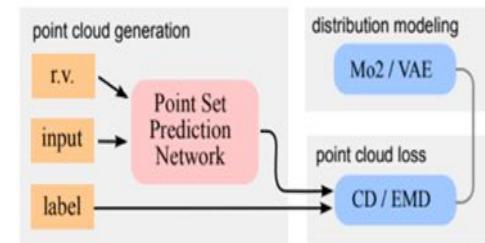
[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Rengarajan\\_Unrolling\\_t he\\_Shutter\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Rengarajan_Unrolling_t he_Shutter_CVPR_2017_paper.pdf)

# [142] Haoqiang Fan, Hao Su, Leonidas J. Guibas, "A Point Set Generation Network for 3D Object Reconstruction From a Single Image", in CVPR Oral, 2017.

Keywords: 3D point cloud generation, 3D reconstruction

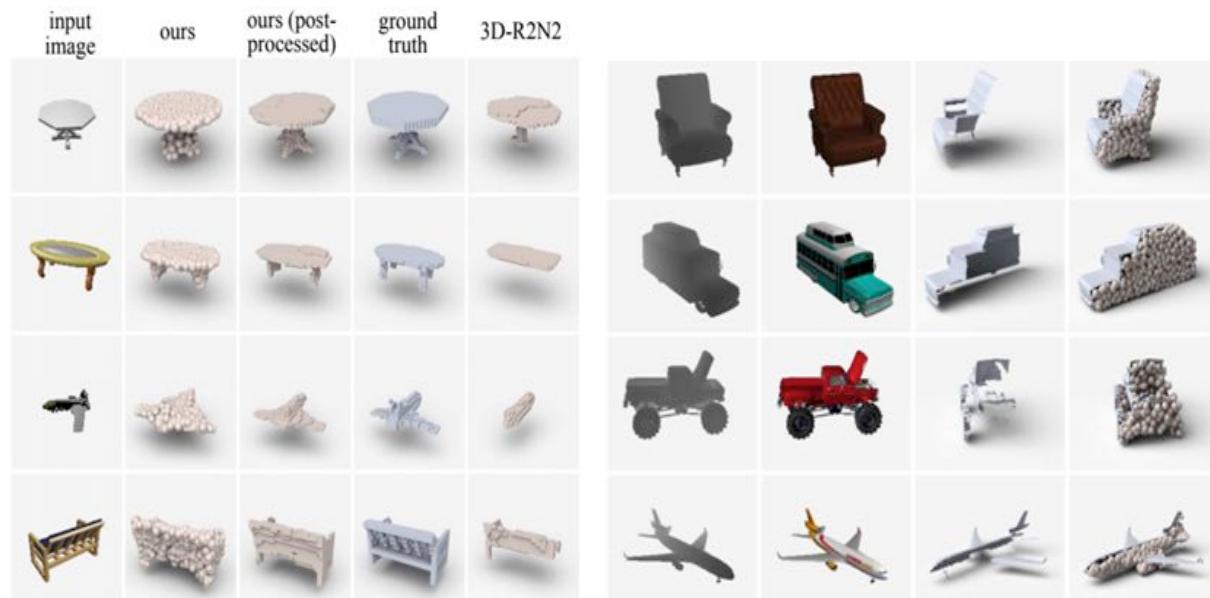
## 概要

単RGB画像から3Dデータの生成・復元を行う。入力画像によれば、曖昧になる領域が存在するという不確かさの問題がある。そこで、新しく設計したCNNにより、もっともらしい複数の3D点群を直に出力する。



## 新規性・差分

- ・単画像からの3Dデータ生成においてstate-of-the-art
- ・真値の不確かさを考慮に入れた、計画的な新CNNの設計



## Links

論文  
[http://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Fan\\_A\\_Point\\_Set\\_CVPR\\_2017\\_paper.pdf](http://openaccess.thecvf.com/content_cvpr_2017/papers/Fan_A_Point_Set_CVPR_2017_paper.pdf)

ご質問・コメント等ありましたら、[cvpaper.challenge@gmail.com](mailto:cvpaper.challenge@gmail.com) / Twitter@CVPaperChallengまでお願いします。