SI ANALYTICS

빽 투 더 Representation Learning: Visual Self-supervision을 중심으로

2019 2nd 함께하는 딥러닝 컨퍼런스 @ ETRI 융합기술연구생산센터

발표자: 서정훈

(Research Scientist, Satrec Initiative)

Date: 4st July, 2019

연사 자기 소개

- 서정훈. GIST 전기전자컴퓨터공학부 4학년 휴학 중입니다.
- 위성항공전문 기업 쎄트렉아이 신기술 연구팀에서 ML/CV 연구원으로 2년째 재직 중입니다.
- 최근 2년간 물체 탐지, 모델 해석 가능성, Domain Transfer 등을 주제로 ICML 워크샵, ICLR 워크샵, ACML 워크샵, ACM SIGSPATIAL 등에 논문을 내왔습니다.

들어가며…

- 본 발표에서 저의 목표는 ‘Self-supervised Learning라는 연구 주제를 큰 맥락에서 환기하기’입니다. 해당 주제에 익숙하지 않은 분들을 대상으로 자료와 발표를 준비했기 때문에, 이에 익숙하신 분들에게는 신선한 내용은 없을 듯 합니다.
- 본 자료 혹은 발표에서 논쟁의 여지가 있는 개인적인 의견이 들어가 있을 수 있습니다. 이러한 의견은 저 개인의 것이며, 소속 집단과는 일절 관련이 없습니다.
- 소개하는 내용들에 디테일이 숭숭 빠져 있습니다.
  - 구현 상의 주의사항이나 논문에 관한 비판 등을 소개하기는 시간상 힘들 것 같습니다
- 소개하는 도메인이 컴퓨터 비전으로 꽤 편향되어 있습니다.
  - 제가 컴퓨터 비전을 하는 사람이라서 그럽니다
- 수식 하나도 없습니다.

01 사전 학습, 다시 돌아보기

사전 학습 (Pre-training)

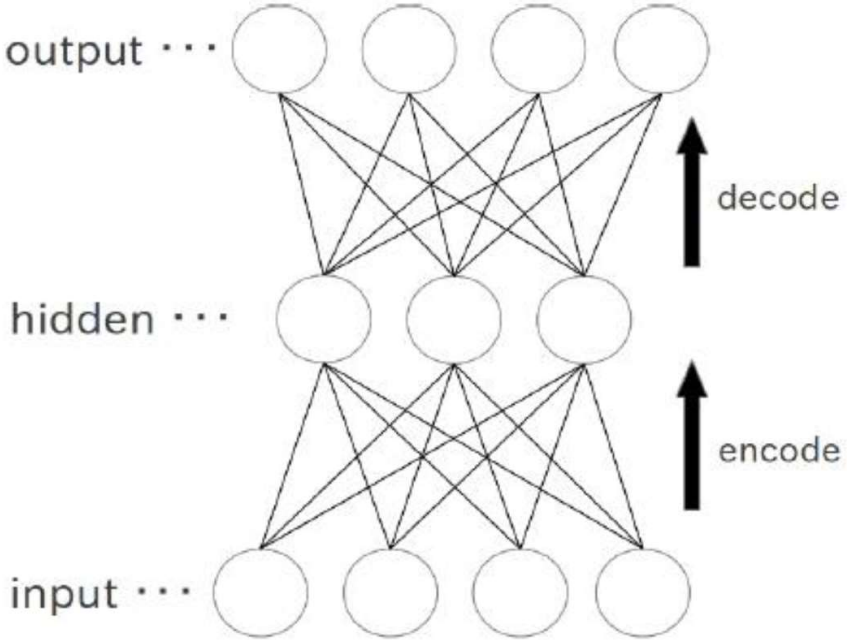
- Q. “사전 학습이 뭘까요?”
  - A1. “RBM이나 DBN 같은 것처럼 먼저 학습시키는 거잖아요. 아니면 Layer-wise Training이라든가.”
  - A2. “이미지넷에서 학습시킨 ResNet이나 VGG 같은 모델을 다른 데이터에 학습시키는 거요.“

- 물론, 둘 다 맞습니다.
- 근데 사전 학습에 해당되는 내용은 이것 뿐만이 아닙니다.

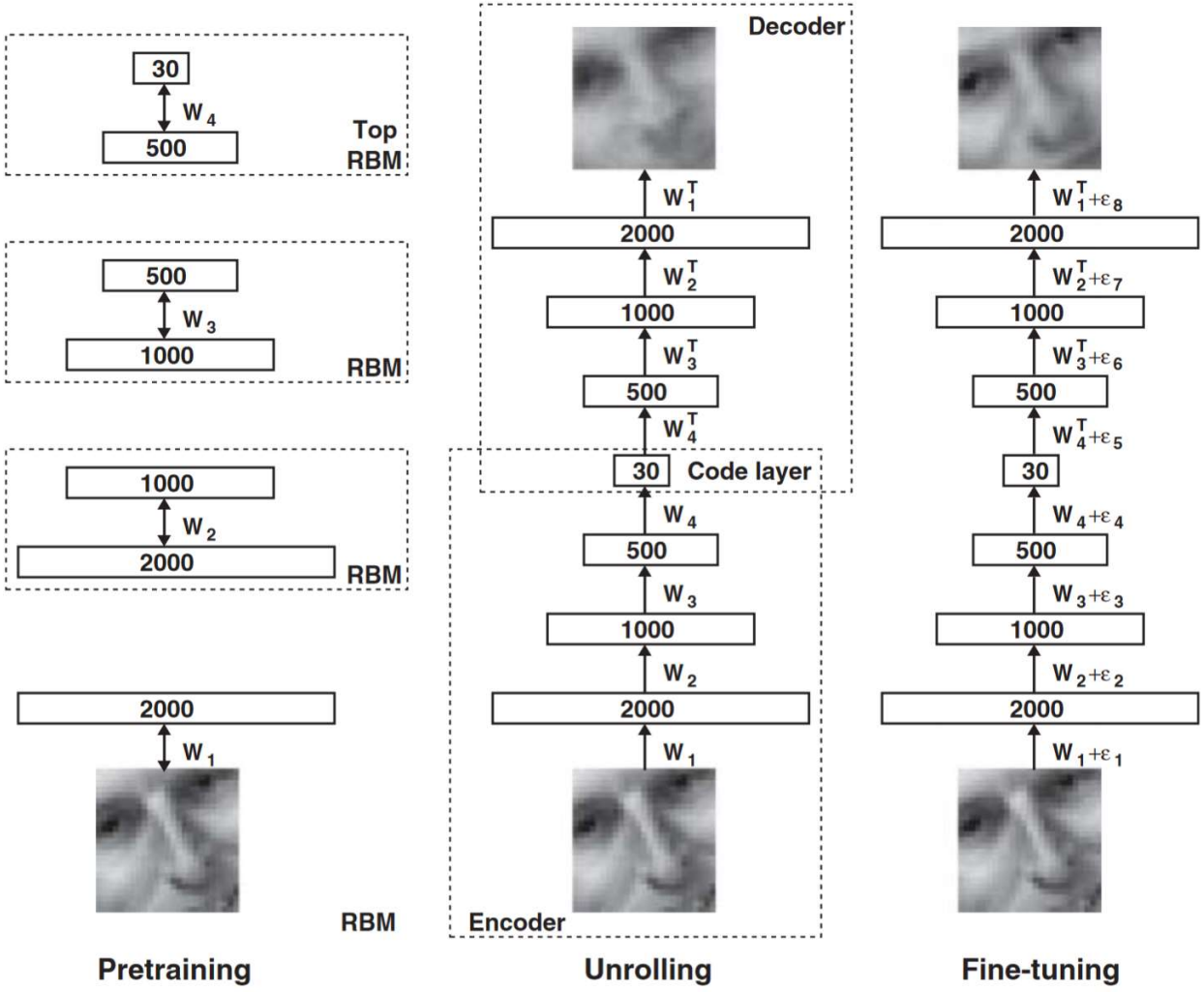
우리의 Pre-training을 찾아서…

Pre-training with RBM [Hinton, 2006b]

- 목표: 차원 축소 (Dimension Reducion)
- 어떻게: Auto-encoder를 학습시켜서
- 문제점: Random Initial Weight로는 Auto-encoder가 학습이 잘 안 됨



Pre-training with RBM [Hinton, 2006b]



Greedy Layer-wise Training ([Bengio, 2007])

Greedy Layer-Wise Training of Deep Networks

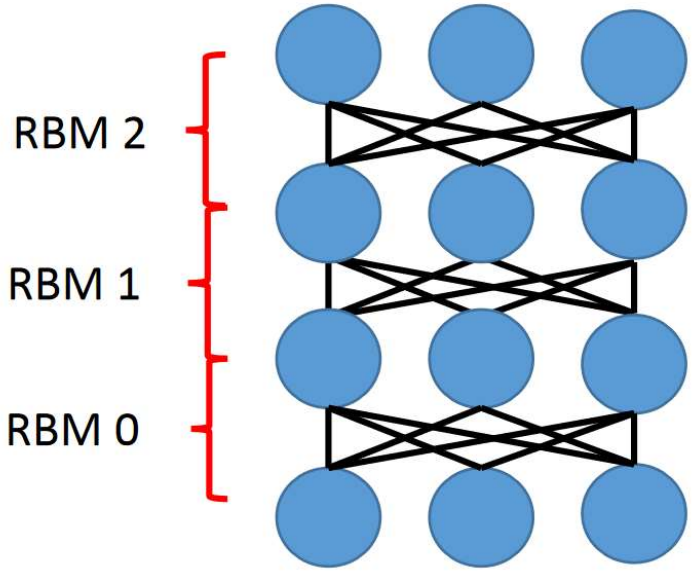
Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle

NIPS 2007

Presented by

Ahmed Hefny

Greedy Layer-wise Training ([Bengio, 2007])



- Extends the concept to:
  - Continuous variables
  - *Uncooperative* input distributions
  - Simultaneous Layer Training
- Explores variations to better understand the training method:
  - What if we use greedy **supervised** layer-wise training ?
  - What if we replace RBMs with auto-encoders ?

Greedy Layer-wise Training ([Bengio, 2007])

	Abalone			Cotton		
	train.	valid.	test.	train.	valid.	test.
1. Deep Network with no pre-training	4.23	4.43	4.2	45.2%	42.9%	43.0%
2. Logistic regression	·	·	·	44.0%	42.6%	45.0%
3. DBN, binomial inputs, unsupervised	4.59	4.60	4.47	44.0%	42.6%	45.0%
4. DBN, binomial inputs, partially supervised	4.39	4.45	4.28	43.3%	41.1%	43.7%
5. DBN, Gaussian inputs, unsupervised	4.25	4.42	4.19	35.7%	34.9%	35.8%
6. DBN, Gaussian inputs, partially supervised	4.23	4.43	4.18	27.5%	28.4%	31.4%

Greedy Layer-wise Training ([Bengio, 2007])

"Backprop would be dead by 2010…"

Greedy Lay[illegible]io, 2007])

[illegible] would be
[illegible]010…"

Q: Layer-wise Pre-Training은 왜 죽었는가?

- A: 없어도 됐기 때문에
- Sigmoid -> ReLU [Glorot, 2011]
- Optimizers
  - AdaGrad [Duchi, 2011]
  - RMSProp [Hinton, 2012]
  - Adam [Kingma, 2015]
- Regularization (e.g. Dropout [Srivastava, 2014])
- 진보된 Weight Initialization 방법론들 (e.g. Xavier Init. [Glorot, 2011])
- Computational Resource -> Over-parameterization (2010~)
  - 그때 당시에는 왜인지는 몰랐지만…
  - [Zhang, 2017], [Shen, 2018], [Arora, 2018], [Simon, 2019a], [Simon, 2019b]


R I P
Backpropagation
1963-200?

Long Live the Back-prop!

그러나 사전 학습이 사라지는 일은 없었다..

- Q: 왜?
- A: 딥러닝 모델을 학습시키기 위해선 데이터가 되게 많아야 했기 때문.
- 따라서 2010년부터 사전 학습은 모델 자체를 학습시키기 위해서가 아니라, 데이터 효율성(Data-efficiency)을 올리기 위한 관점에서 연구되었음.

- (개인적으로 생각하기에) 이 범주에 들어가는 연구 주제들:
  - 1) 전이 학습 (Transfer Learning)
  - 2) 메타 학습 (Meta-Learning)
  - 3) Domain Adaptation
- 중요한 것은, 셋 다 일종의 '사전 학습'이라는 것.

전이 학습 (Transfer Learning)

- Classic Works: [Pratt, 91], [Pratt, 93], [Caruana, 95], [Thrun, 96]
  - "Multi-task Learning"이라 불림.
- http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html

NIPS*95 Post-Conference Workshop

"Learning to Learn: Knowledge Consolidation and Transfer in Inductive Systems"

혼란스러운 용어 사용 (~2000)

- Transfer Learning
- Learning to Learn
- Multi-task Learning
- Life-long Learning
- Continual Learning

- 이 용어들이 전부 다 서로 논문마다 혼재되어 사용되었음

혼란스러운 용어 사용 (~2010)

- Inductive Transfer Learning
  - Target Domain/Task Label 있음
  - Source Domain/Task Label
    - ✓ 없음: Self-taught Learning
    - ✓ 있음: Multi-task Learning
- Transductive Transfer Learning
  - Source Domain/Task Label 있음
  - Target Domain/Task Label 없음
  - Domain Transfer, Sample Selection Bias
- Unsupervised Transfer Learning
  - Source Domain/Task Label 없음
  - Target Domain/Task Label 없음

- Taxonomy Reference: [Pan, 09]

드디어 우리가 아는, “Transfer Learning”가 정의됨

- , [Bengio, 12] ~~감사합니다, 선생님~~
- “Deep Learning relies heavily on unsupervised or semi-supervised learning, and assumes that representations of X that are useful to capture P(X) are also in part useful to capture P(Y|X).”

지금까지의 정리,

- 초기 사전 학습 연구의 목표는 학습이 잘 안 되는 뉴럴넷을 어떻게든 잘 학습시키기 위한 것이었음.
- 10년도 이후 이러한 Motivation은 사실상 별 의미가 없게 됐음.
  - 여러 이유로 그런 사전 학습 방식 없이도 일단 학습해서 수렴은 시킬 수 있게 됐기 때문.
  - 대신, Data-efficiency라는 관점이 더 중요해짐.

- 90년 초중반 초기 전이 학습은 Multi-task Learning과 연관되어 제시되었음.
- 지금 사용하는 (좁은 의미의) 전이 학습이라는 용어는 12년에서야 정립됨.
- 사전 학습의 범주와 목표를 조금 더 확장해서 생각할 필요가 있음.
  - 사전 학습은, 단순히 학습 테크닉이 아니라 표현 학습에 있어 핵심적인 연구 주제.
  - 데이터 효율성, 전이 학습 / 메타 러닝 / Domain Adaptation

Transfer Learning은 정말 효과적인가?

- **(+) "ImageNet-trained Model은 효과적이다!"**
  - ImageNet-pretrained Model을 사용하는 수많은 연구들
  - [Hendrycks, 2019] "Robustness나 다른 유용한 side-effect를 남기더라"
  - [Li, 2019] "Object Detection 작업에서 Data-efficient하다"
- **(△) "되는 것 같지만, 사람들이 말하는 만큼 다 잘되는 것 같지는 않은데?"**
  - [Kornblith, 2019] "좀 되긴 하는데 Fine-grained Task에선 잘 안 먹히는데…"
  - [He, 2018] "Data-efficiency에서는 꽤 효과적인데, 데이터가 적당히 많아지면 최종적인 성능에는 별로 상관 없다"
  - [Zamir, 2018] "같은 도메인에서도 Task마다 배우는 표현의 형태가 되게 많이 다를 수 있다."
- **(-) "별로 안 효과적인 것 같은데…"**
  - [Paghu, 2019] "의료 영상에서 벤치마크 해보니까 효과가 되게 부정적인데?"

[Kornblith, 2019]

ImageNet-pretrained Model

Dataset	Classes	Size (train/test)	Accuracy metric
Food-101 [5]	101	75,750/25,250	top-1
CIFAR-10 [43]	10	50,000/10,000	top-1
CIFAR-100 [43]	100	50,000/10,000	top-1
Birdsnap [4]	500	47,386/2,443	top-1
SUN397 [84]	397	19,850/19,850	top-1
Stanford Cars [41]	196	8,144/8,041	top-1
FGVC Aircraft [55]	100	6,667/3,333	mean per-class
PASCAL VOC 2007 Cls. [22]	20	5,011/4,952	11-point mAP
Describable Textures (DTD) [10]	47	3,760/1,880	top-1
Oxford-IIIT Pets [61]	37	3,680/3,369	mean per-class
Caltech-101 [24]	102	3,060/6,084	mean per-class
Oxford 102 Flowers [59]	102	2,040/6,149	mean per-class

Table 1. Datasets examined in transfer learning

[Kornblith, 2019]

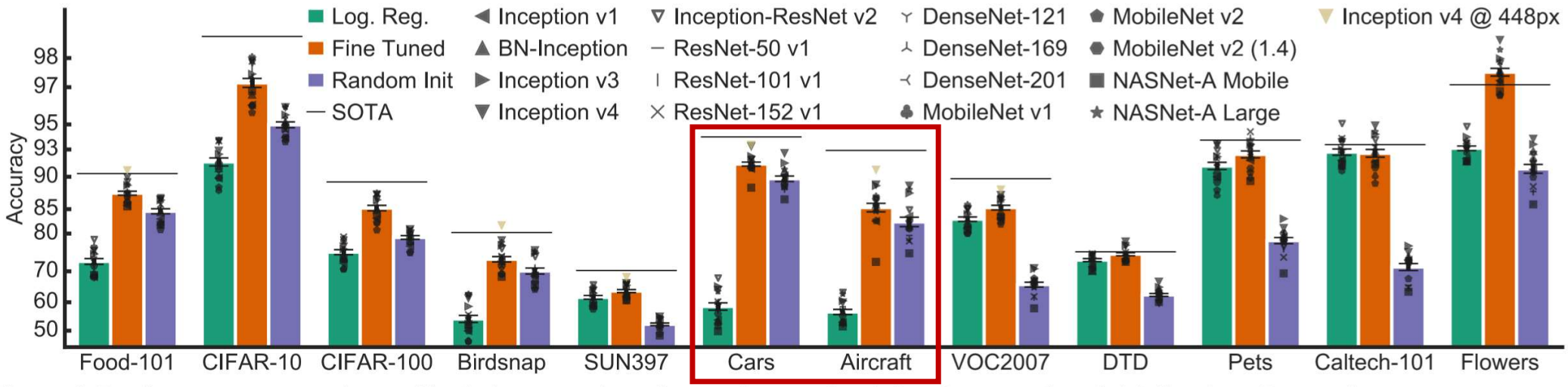
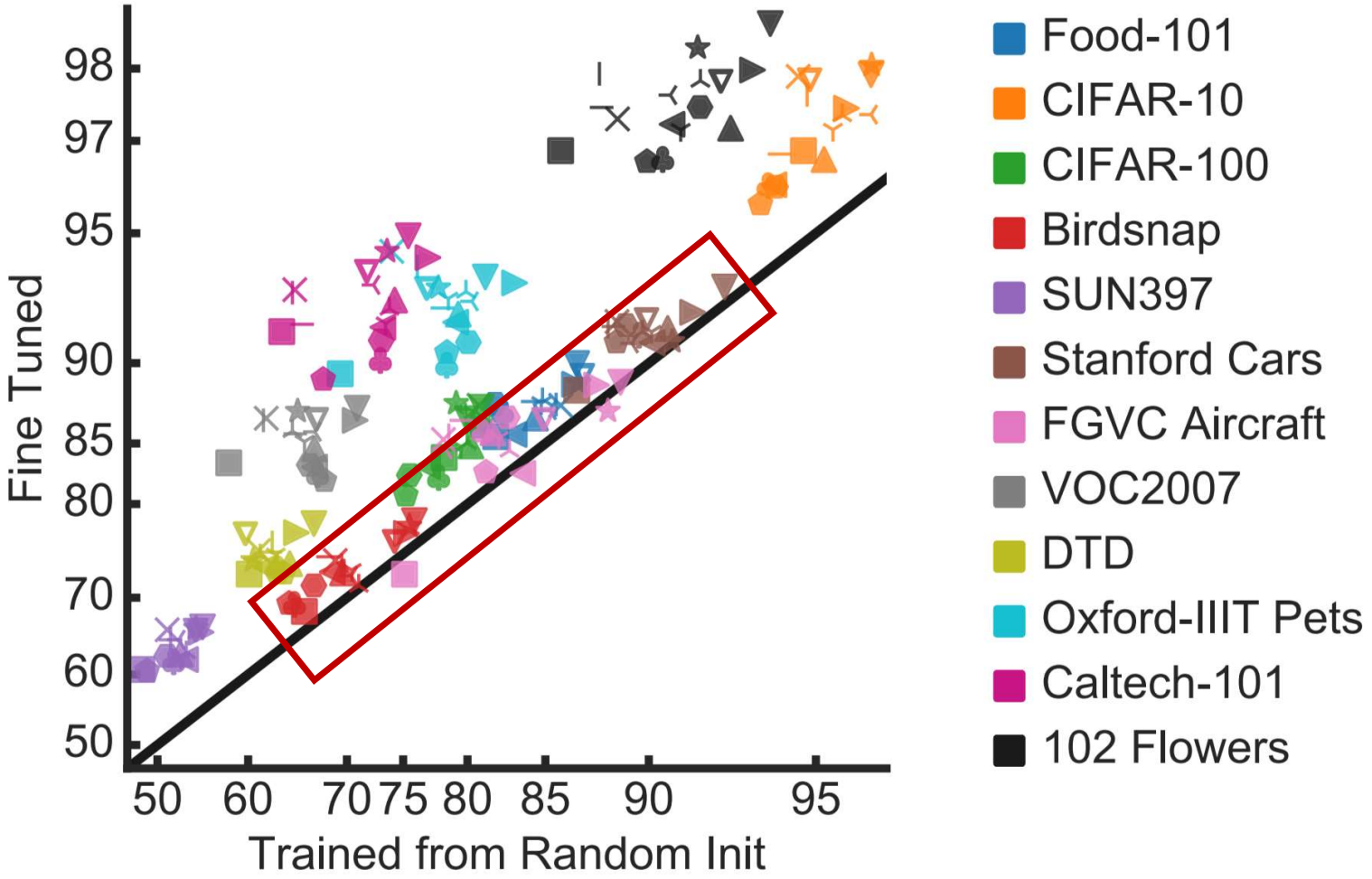


Figure 6. Performance comparison of logistic regression, fine-tuning, and training from random initialization. Bars reflect accuracy across models (excluding VGG) for logistic regression, fine-tuning, and training from random initialization. Error bars are standard error. Points represent individual models. Lines represent previous state-of-the-art. Best viewed in color.

[Kornblith, 2019]

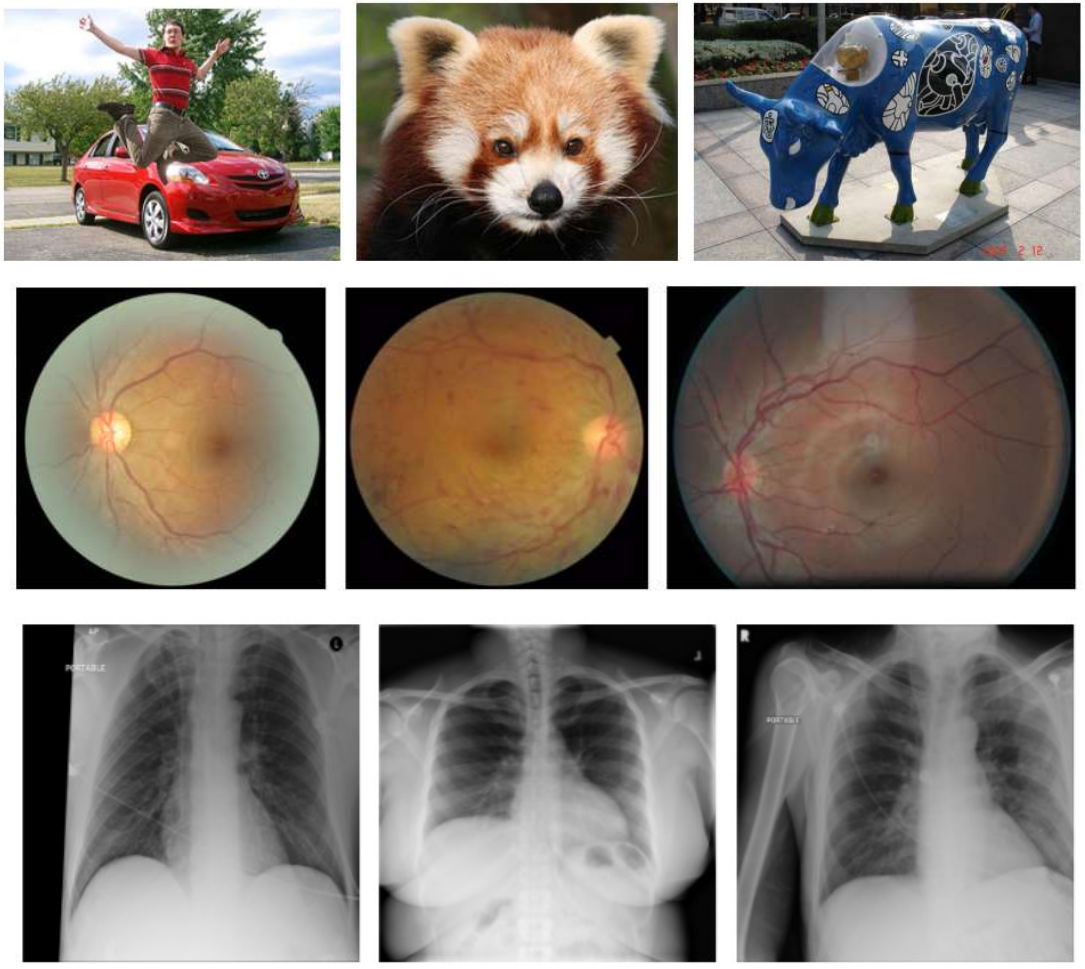


[Kornblith, 2019]

Dataset	Classes	Size (train/test)	Accuracy metric
Food-101 [5]	101	75,750/25,250	top-1
CIFAR-10 [43]	10	50,000/10,000	top-1
CIFAR-100 [43]	100	50,000/10,000	top-1
Birdsnap [4]	500	47,386/2,443	top-1
SUN397 [84]	397	19,850/19,850	top-1
Stanford Cars [41]	196	8,144/8,041	top-1
FGVC Aircraft [55]	100	6,667/3,333	mean per-class
PASCAL VOC 2007 Cls. [22]	20	5,011/4,952	11-point mAP
Describable Textures (DTD) [10]	47	3,760/1,880	top-1
Oxford-IIIT Pets [61]	37	3,680/3,369	mean per-class
Caltech-101 [24]	102	3,060/6,084	mean per-class
Oxford 102 Flowers [59]	102	2,040/6,149	mean per-class

Table 1. Datasets examined in transfer learning

[Paghu, 2019]



ImageNet

retinal fundus photographs

CheXpert datasets

[Paghu, 2019]

Dataset	Model Architecture	Random Init	Transfer	Parameters	ImageNet Top5
Retina	Resnet-50	96.4% ± 0.05	96.7% ± 0.04	23570408	92.% ± 0.06
Retina	Inception-v3	96.6% ± 0.13	96.7% ± 0.05	22881424	93.9%
Retina	CBR-LargeT	96.2% ± 0.04	96.2% ± 0.04	8532480	77.5% ± 0.03
Retina	CBR-LargeW	95.8% ± 0.04	95.8% ± 0.05	8432128	75.1% ± 0.3
Retina	CBR-Small	95.7% ± 0.04	95.8% ± 0.01	2108672	67.6% ± 0.3
Retina	CBR-Tiny	95.8% ± 0.03	95.8% ± 0.01	1076480	73.5% ± 0.05

Table 1: **Transfer learning and random initialization of two standard ImageNet architectures and a family of simple convolutional neural networks have very similar AUCs for diagnosing moderate DR.** Model performance on DR diagnosis is also not closely correlated with ImageNet performance, with the small models performing poorly on ImageNet but very comparably on the medical task.

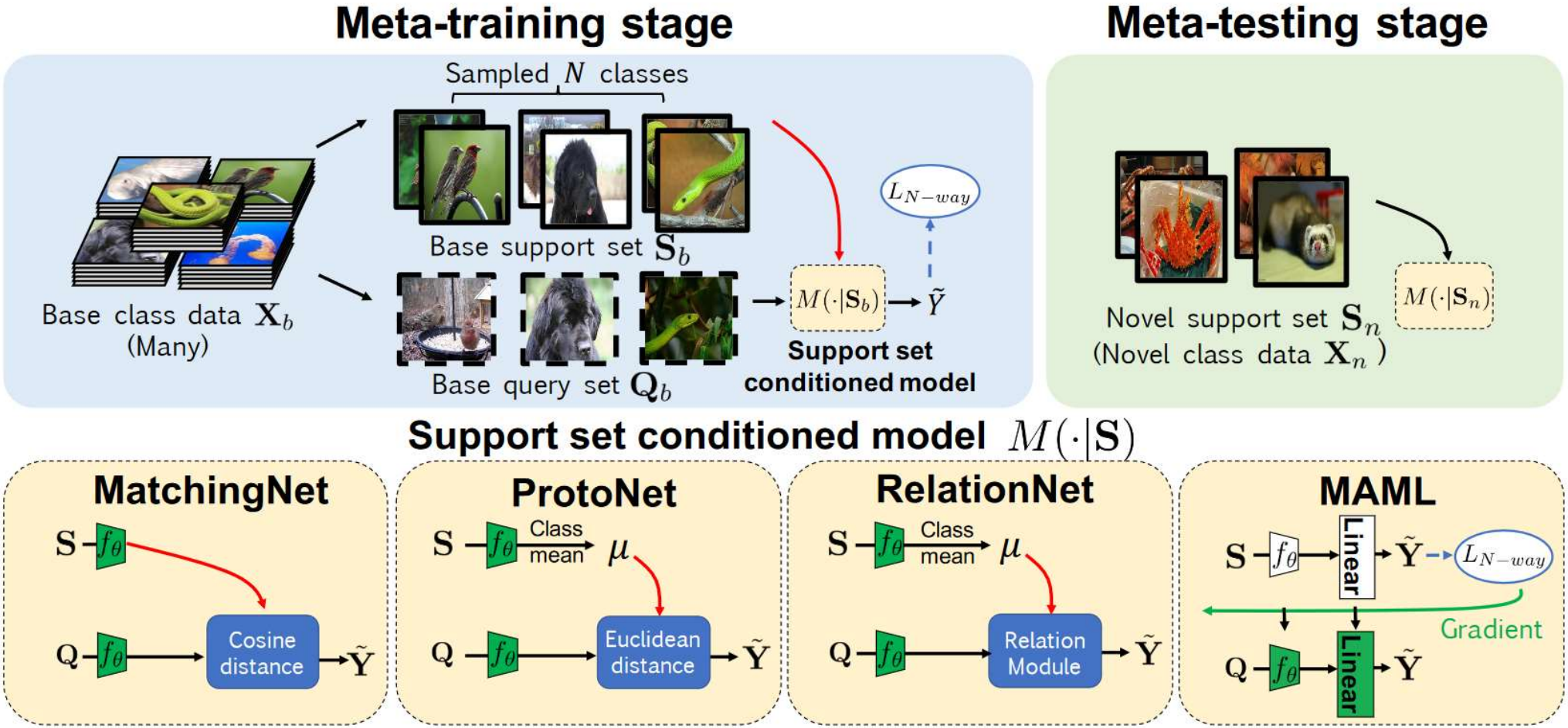
[Paghu, 2019]

Model Architecture	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
Resnet-50	79.52±0.31	75.23±0.35	85.49±1.32	88.34±1.17	88.70±0.13
Resnet-50 (trans)	79.76±0.47	74.93±1.41	84.42±0.65	88.89±1.66	88.07±1.23
CBR-LargeT	81.52±0.25	74.83±1.66	88.12±0.25	87.97±1.40	88.37±0.01
CBR-LargeT (trans)	80.89±1.68	76.84±0.87	86.15±0.71	89.03±0.74	88.44±0.84
CBR-LargeW	79.79±0.79	74.63±0.69	86.71±1.45	84.80±0.77	86.53±0.54
CBR-LargeW (trans)	80.70±0.31	77.23±0.84	86.87±0.33	89.57±0.34	87.29±0.69
CBR-Small	80.43±0.72	74.36±1.06	88.07±0.60	86.20±1.35	86.14±1.78
CBR-Small (trans)	80.18±0.85	75.24±1.43	86.48±1.13	89.09±1.04	87.88±1.01
CBR-Tiny	80.81±0.55	75.17±0.73	85.31±0.82	84.87±1.13	85.56±0.89
CBR-Tiny (trans)	80.02±1.06	75.74±0.71	84.28±0.82	89.81±1.08	87.69±0.75

모델 크기

Table 2: **Transfer learning provides mixed performance gains on chest x-rays.** Performances (AUC%) of diagnosing different pathologies on the CheXpert dataset. Again we see that transfer learning does not help significantly, and much smaller models performing comparably.

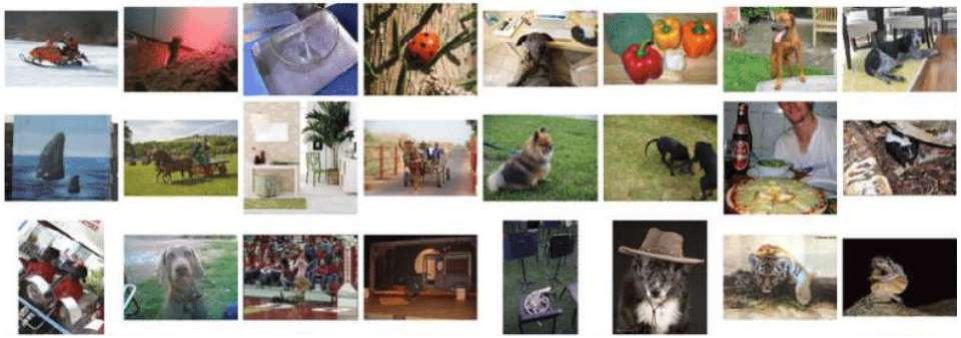
메타-러닝 또한 모든 문제에서의 해결법이 아니다



[Chen, 2019]

메타-러닝 또한 모든 문제에서의 해결법이 아니다

Meta-Training on mini-ImageNet



Meta-Testing on CUB

(a) Calilfornia gull

(b) Glaucous gull

Fine-tuning ↓	***mini*-ImageNet →CUB**
Baseline	**65.57±0.70**
Baseline++	62.04±0.76
MatchingNet	53.07±0.74
ProtoNet	62.02±0.70
MAML	51.34±0.72
RelationNet	57.71±0.73

Table 3: **5-shot accuracy under the cross-domain scenario with a ResNet-18 backbone.** Baseline outperforms all other methods under this scenario.

[Chen, 2019]

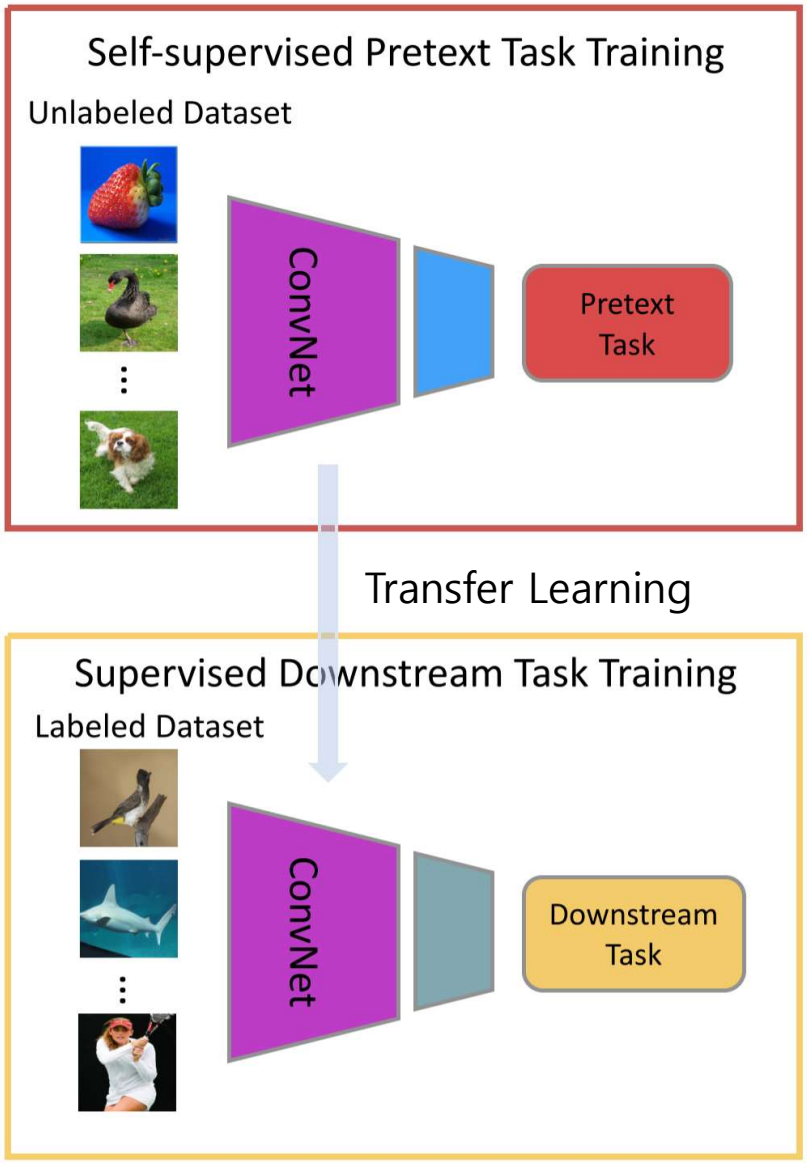
지금까지의 정리

- 전이 학습이든, 메타 러닝이든 사전 학습에 의한 모델 성능은
- Domain에 의존적이고 ([Paghu, 2019])
  - Source Domain과 Target Domain이 비슷할 수록 좋고, 다를 수록 안 좋다
- Task에 의존적이고 ([Zamir, 2018], [Kornblith, 2019])
  - Source Task와 Target Task가 비슷할 수록 좋고, 다를 수록 안 좋다
- Architecture에 의존적이고 ([Paghu, 2019])
- Regularization에 의존적이고 ([Kornblith, 2019])
  - Source에서의 성능이 잘 나올 수록 Target에서 잘 나온다는 보장은 없고,
- ~~아무튼 이게 어떻게 상호적으로 작용해서 결과가 나오는 건진 모르겠다…~~
- 아직까지도 연구가 더 필요하다.

지금까지의 정리

- 전이 학습이든, 메타 러닝이든 사전 학습에 의한 모델 성능은
- Domain에 의존적이고 ([Paghu, 2019])
  - Source Domain과 Target Domain이 비슷할 수록 좋고, 다를 수록 안 좋다
- Task에 의존적이고 ([Zamir, 2018], [Kornblith, 2019])
  - Source Task와 Target Task가 비슷할 수록 좋고, 다를 수록 안 좋다
- Architecture에 의존적이고 ([Paghu, 2019])
- Regularization에 의존적이고 ([Kornblith, 2019])
  - Source에서의 성능이 잘 나올 수록 Target에서 잘 나온다는 보장은 없고,
- ~~아무튼 이게 어떻게 상호적으로 작용해서 결과가 나오는 건진 모르겠다…~~
- 아직까지도 연구가 더 필요하다.

02 Self-supervised Learning: 개괄



Self-supervised Learning

1) Self-supervised Learning의 목적 자체는 기존의 Pre-training과 동일하다.

- "사전학습한 모델을 통해 Downstream Task에서 이득을 본다."

2) "Self-supervised"의 뜻?

- "라벨 없이 (Unsupervised) 지도 학습 (Supervised Learning)"
- == 라벨 없는 데이터에서 Task와 라벨을 만들어야 한다

3) 기존 Pre-training 방식들보다 훨씬 Domain-dependent하다

- 이미지는 이미지 만의 방식이 있고,
- 음성은 음성 만의 방식이 있고,
- 비디오는 비디오 만의 방식이 있고,
- 자연어는 자연어 만의 방식이 있다
- 결국, 도메인에 내제된 구조를 Task화하는 경우가 많음

4) 결국 중요한 것은 Downstream Task에서의 성능

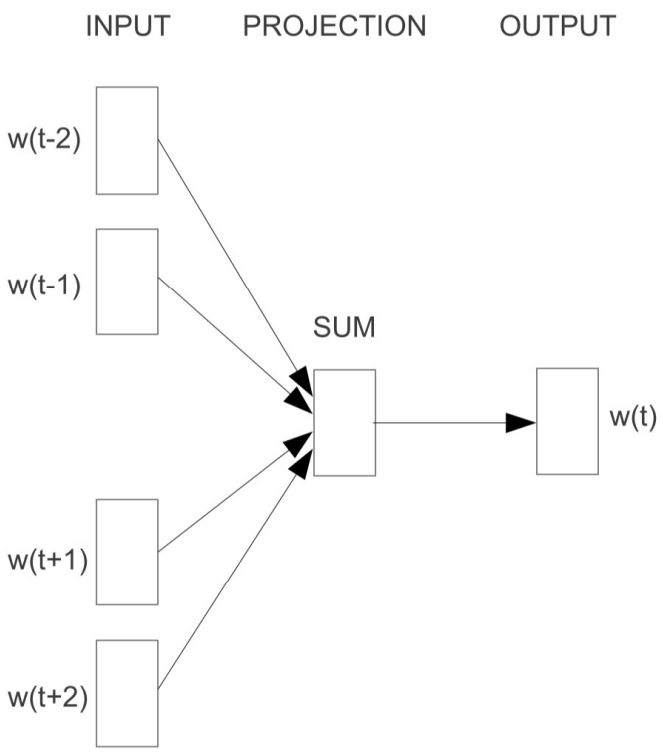
- Downstream Task에서 얼마나 적은 라벨로 성능을 잘 내느냐?
- 그래서 Pretext Task Training에서의 성능 자체는 중요하지 않게 보는 경우가 많음

Self-supervised Learning

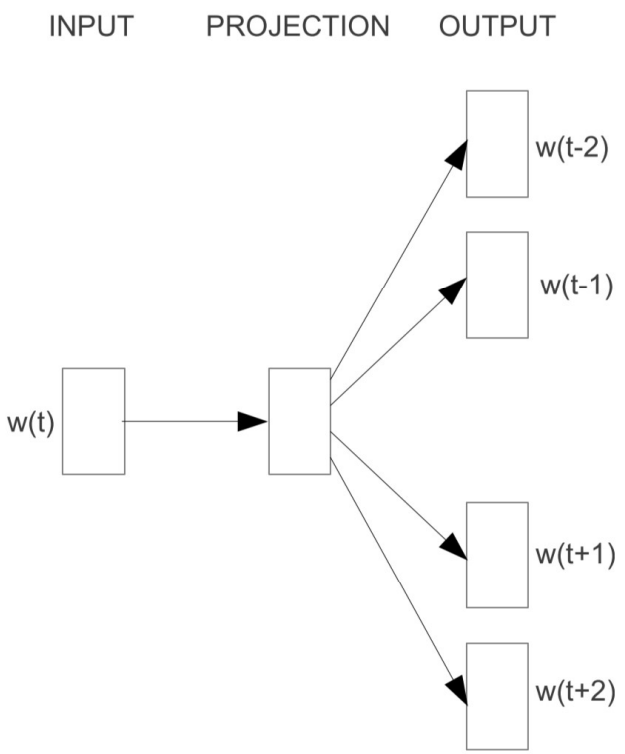
- 일반적으로 말하는 가장 최초의 self-supervised learning 연구는 Context Prediction ([Doersch, 2015]).

- 개인적으로 생각했을 때, 최초의 (유행을 탄) self-supervised learning은…
  - [Mikolov, 2013a], [Mikolov, 2013b]
  - Word2Vec

Word2Vec



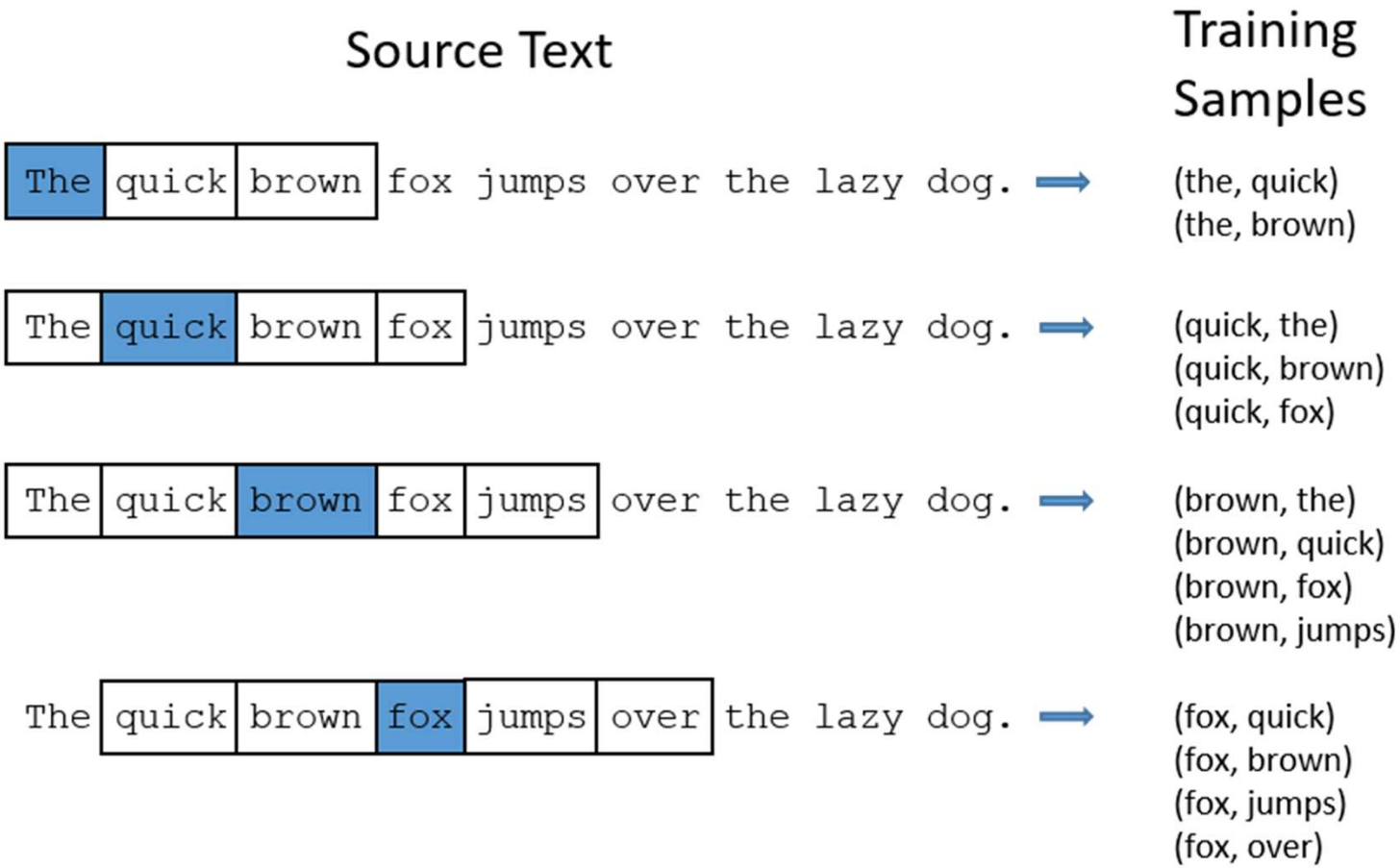
CBOW



Skip-gram

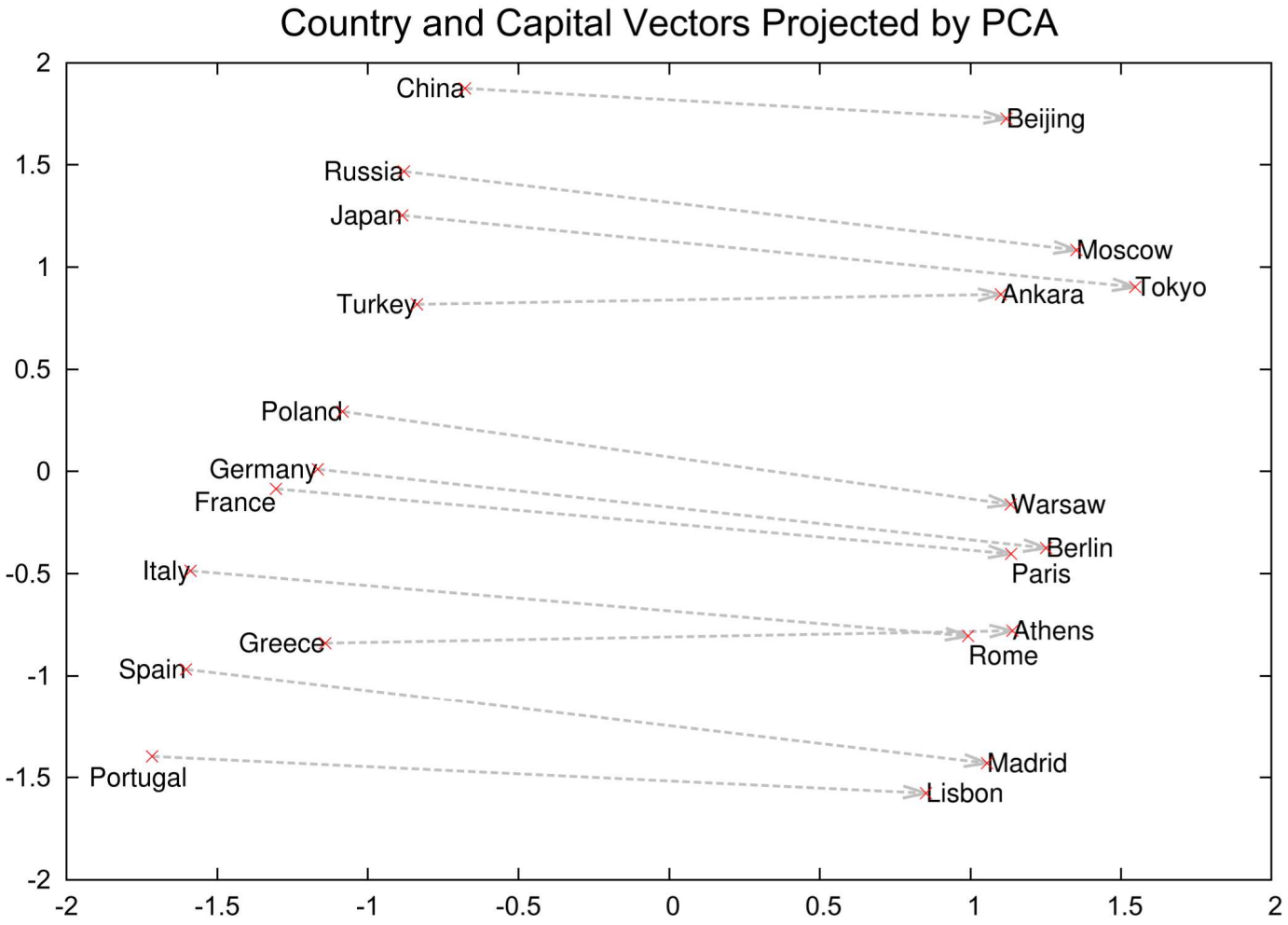
NOTE] Full Softmax는 지나치게 연산량이나 메모리 요구량이 많음 -> Hierarchical Softmax or NCE 사용

Word2Vec - Skip-Gram



이미지 출처: ratsgo's blog (https://ratsgo.github.io)

2-D PCA of the 1000-D Skip-gram vectors



Word2Vec ∈ Self-supervised Learning

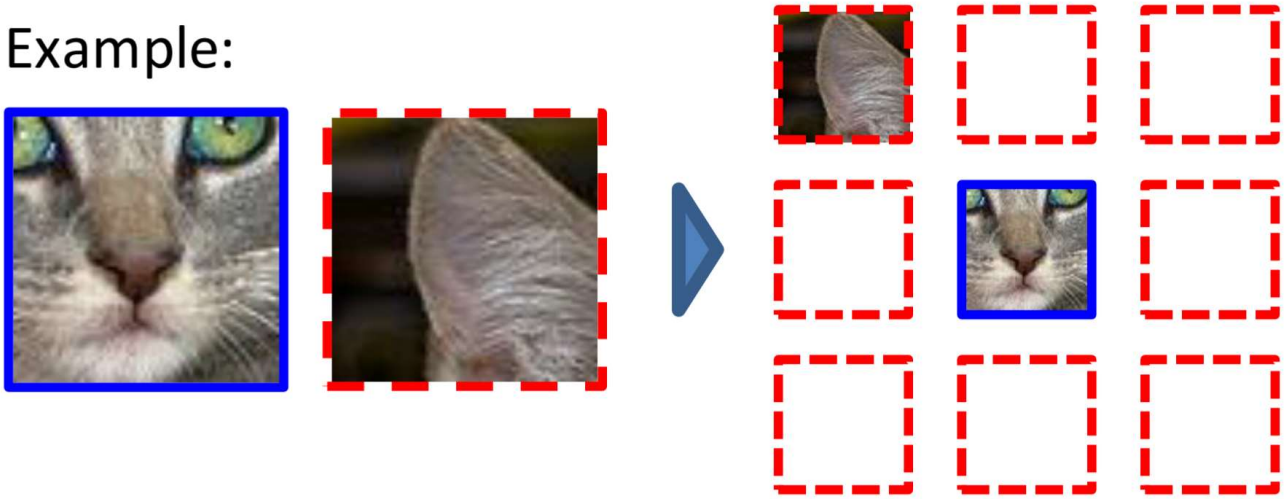
1. 사람의 라벨링이 필요한가? ☞ X

2. 각 단어의 임베딩 자체가 목적이라기보다는 이러한 임베딩을 통해 다른 Task를 푸는 것에 관심이 있음.

3. 모델이 서로 인접한 단어를 예측하는 문제를 품으로서 단어의 구조에 관한 정보를 간접적으로 배울 것이라는 도메인 의존적인 지식을 이용함.
   - Word2Vec의 학습 과정을 Pretext Task라고 봤을 때, 이러한 학습 과정은 이후에 이를 사용해서 할 Downstream Task와는 별로 관계가 없음.

Context Prediction ([Doersch, 2015])

Example:



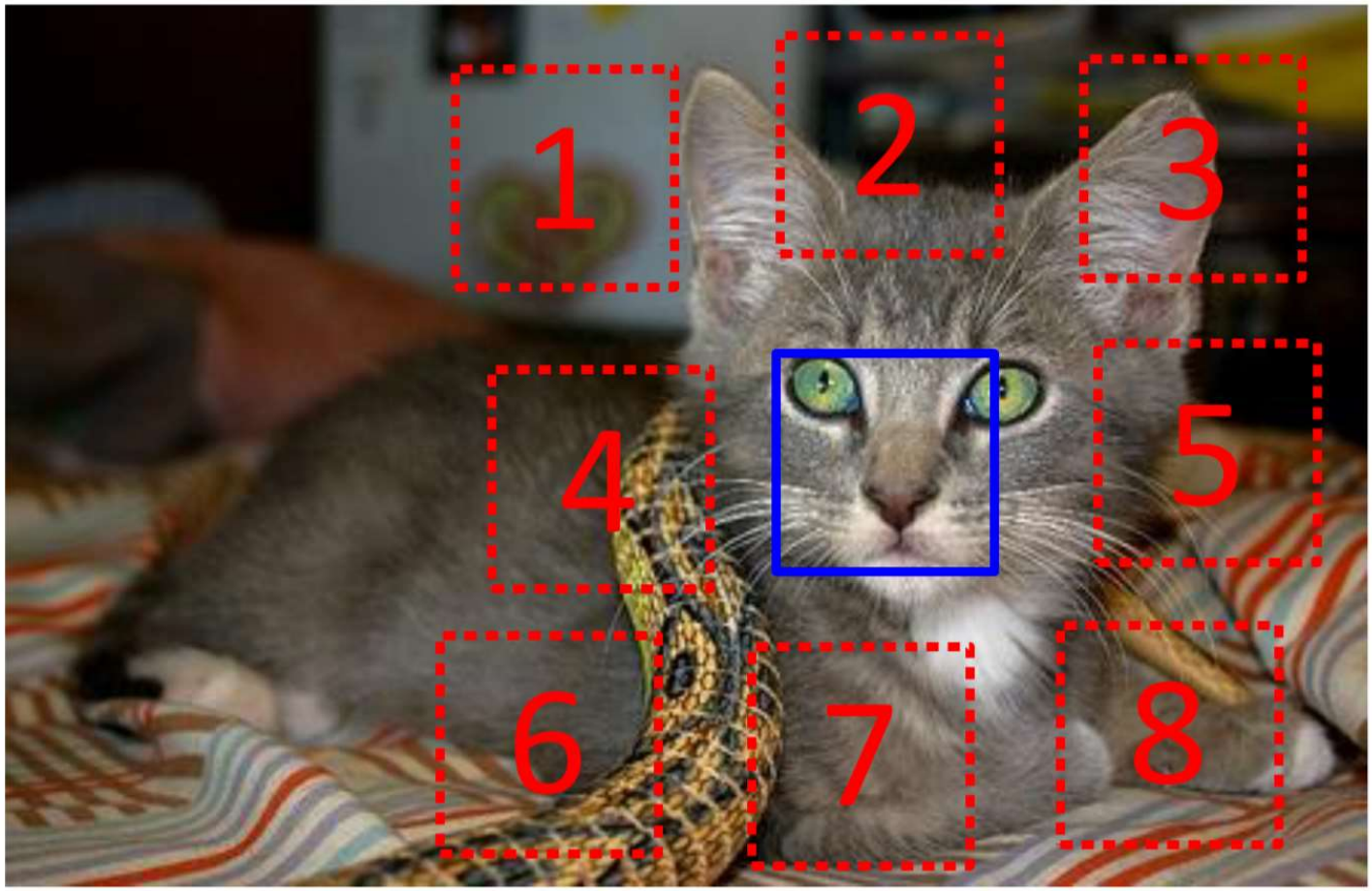
중앙 패치를 기준으로 다른 패치의 상대적 위치를 예측하는 문제를 만들어 학습

Question 1:

Question 2:

"Note that the task is much easier once you have recognized the object!"

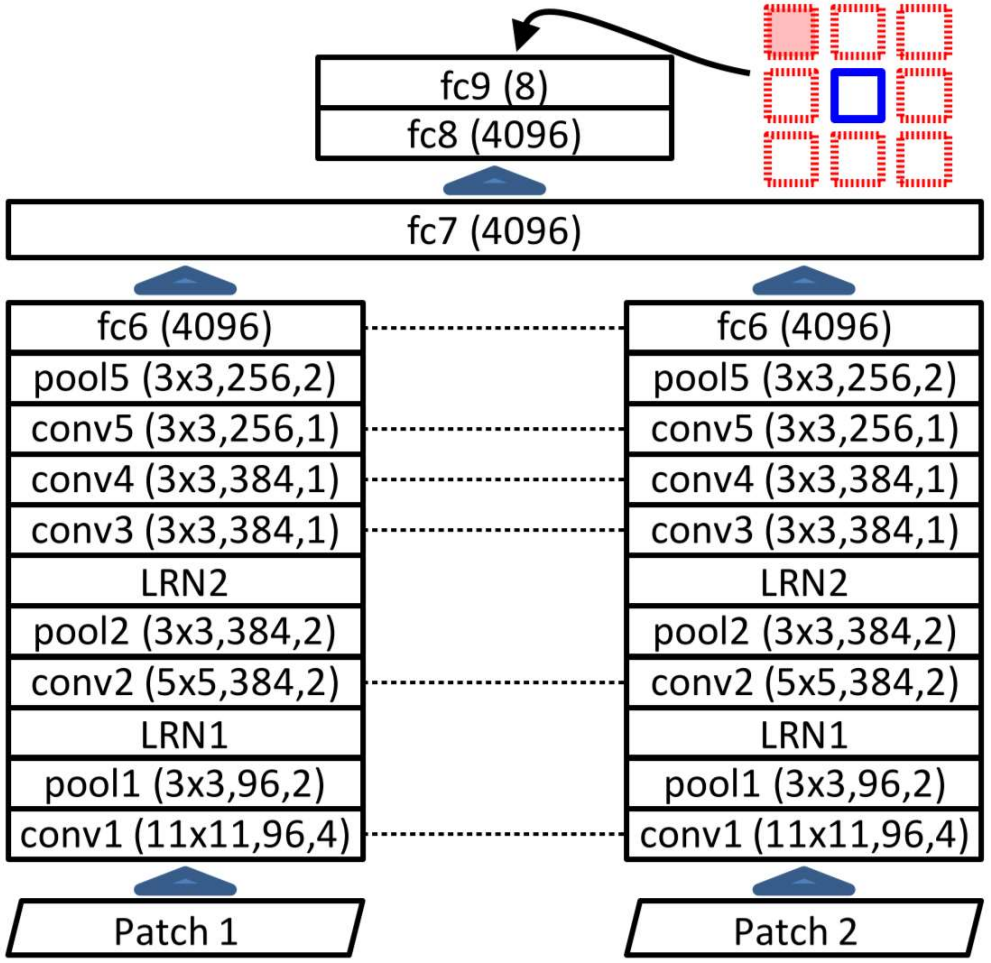
Context Prediction



X = ( , ); Y = 3

Context Prediction - Architecture

X = ( , ); Y = 3



Context Prediction - VOC07 Benchmark

VOC-2007 Test	mAP
Scratch-Ours	39.8
Ours-projection	45.7
Ours-color-dropping	46.3

Random 초기화

VGG-K-means-rescale	42.4
VGG-Ours-rescale	61.7
VGG-ImageNet-rescale	68.6

[Krähenbühl, 2016]

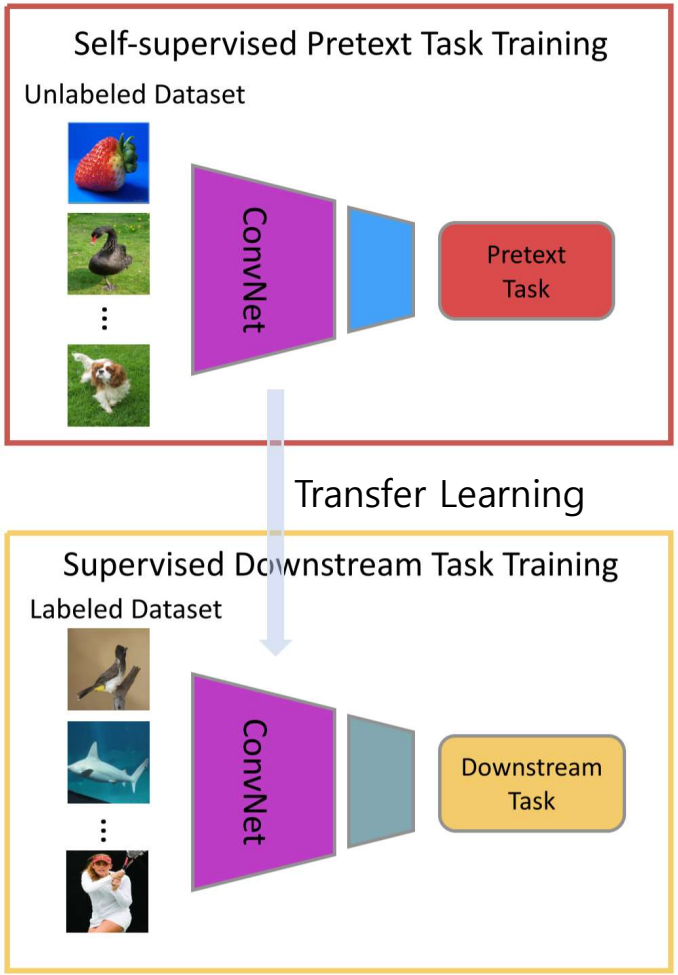
ImageNet-pretrained

정리,

- Self-supervised Learning이란
  - 도메인의 구조나 데이터 획득 과정의 특성을 이용하여,
  - 특정 테스크를 정의해 라벨을 어거지로 만들고,
  - 이를 통해 도메인에 관한 표현 학습을 시키는 것을 말한다

- Pretext/Downstream Task는 서로 연관이 없을 수도 있다

- Pretext Task 성능에는 관심이 없고, Downstream Task의 성능에만 관심이 있다
  - Linear Separability를 벤치마크로 다루는 경우가 있음



03 주요 Self-supervised Learning 연구, 초스피드로 훑어보기

여기서 언급할 Self-supervision 연구들

- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015] & Flow Prediction [Luo, 2017]
  - Ordering [Misra, 2018] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Relative Depth [Jiang, 2018]

여기서 언급할 Self-supervision 연구들

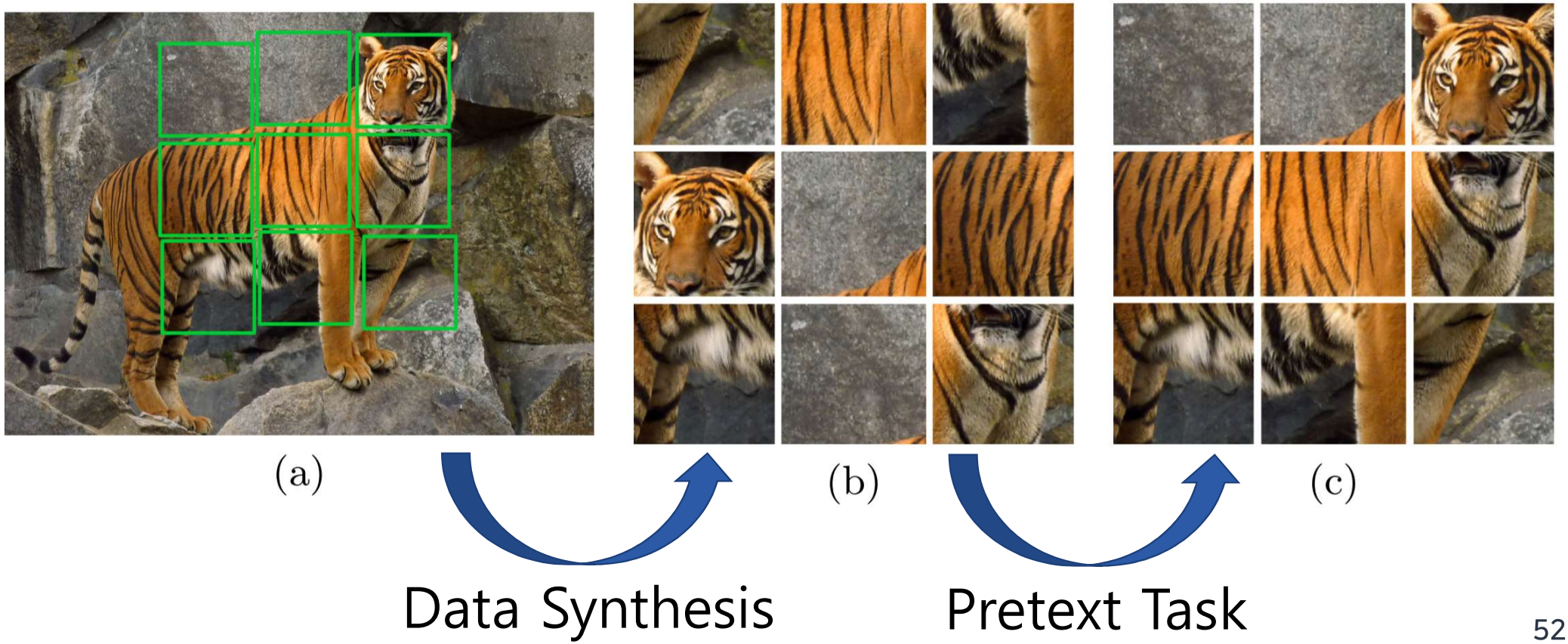
- 오디오 (w/ Cross Modality)
  - Self-supervised synchronization ([Owens, 2018], [Korbar, 2018])
- ~~텍스트~~
  - ~~BERT [Owens, 2018] & GPT [Radford, 2018] & XLNet [Yang, 2019]~~

- 주로 메소드에 관해서만 다루고, 벤치마크에 대해선 일부만 다룹니다
- 다른 도메인들의 연구에 대해선 다룰 여유가 없을 것 같습니다 흑흑

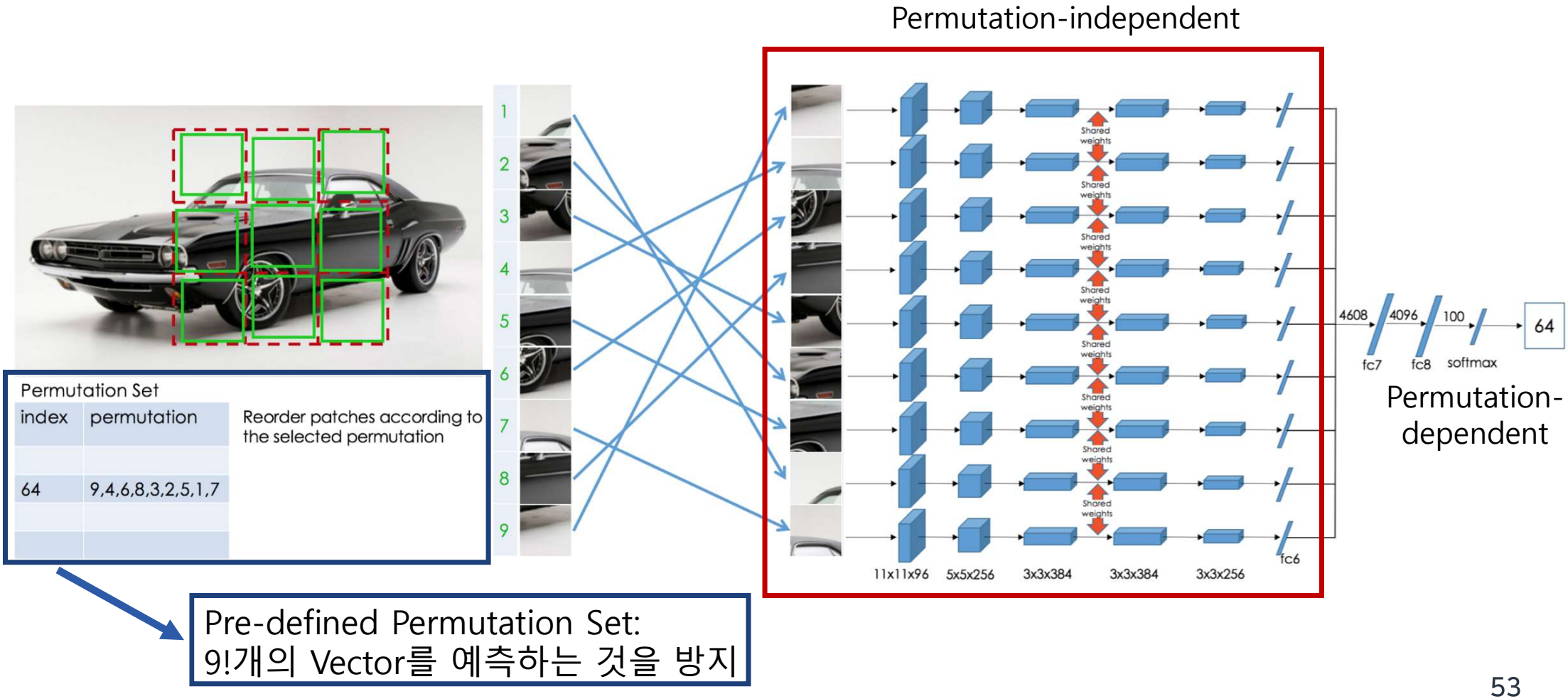
여기서 언급할 Self-supervision 연구들

- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015] & Flow Prediction [Luo, 2017]
  - Ordering [Misra, 2018] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Relative Depth [Jiang, 2018]

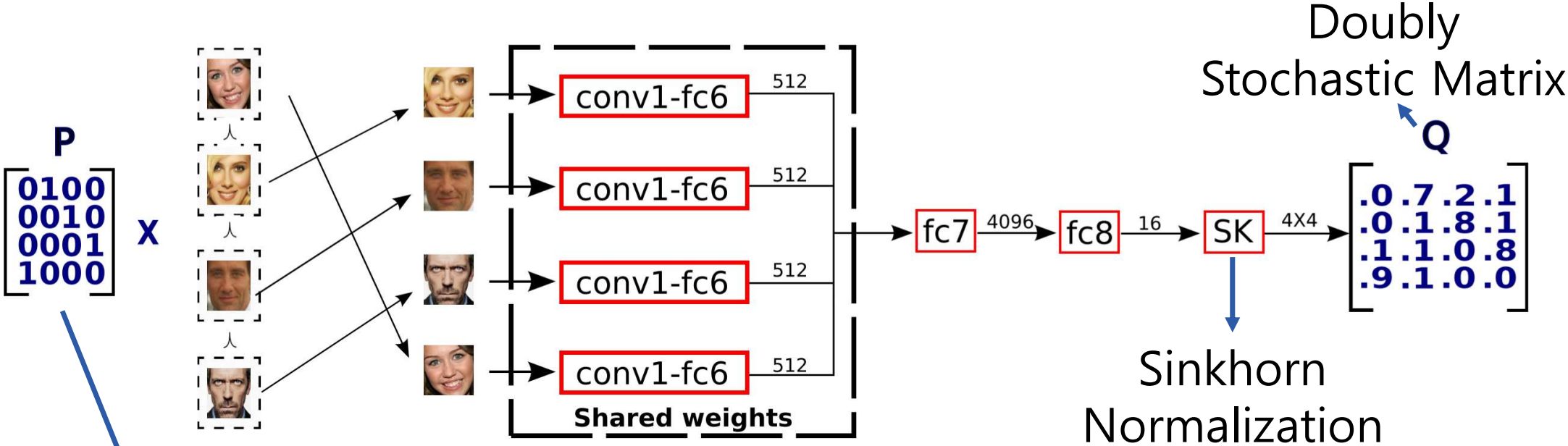
이미지 - Jigsaw [Noroozi, 2016]



이미지 - Jigsaw [Noroozi, 2016]

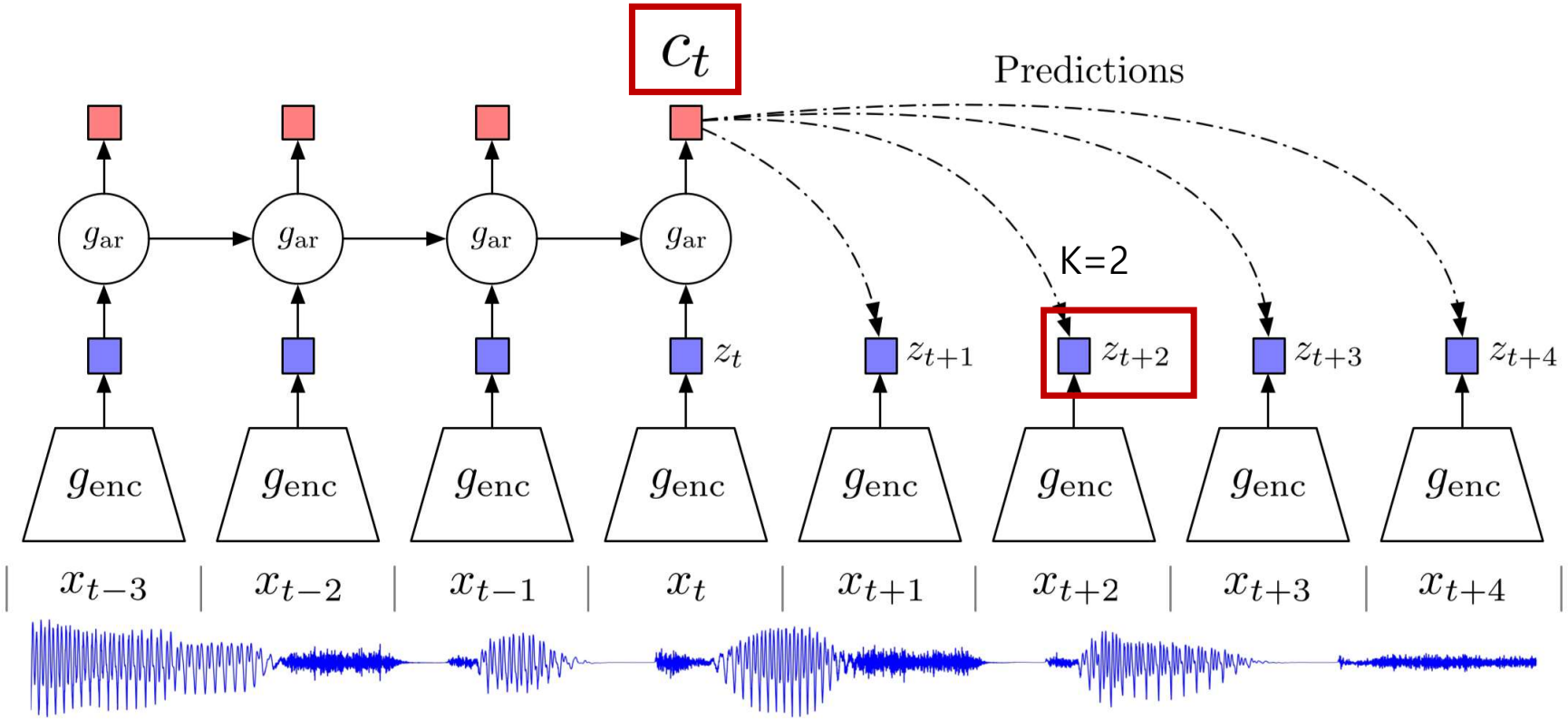


이미지 - DeepPermNet [Santa, 2017]



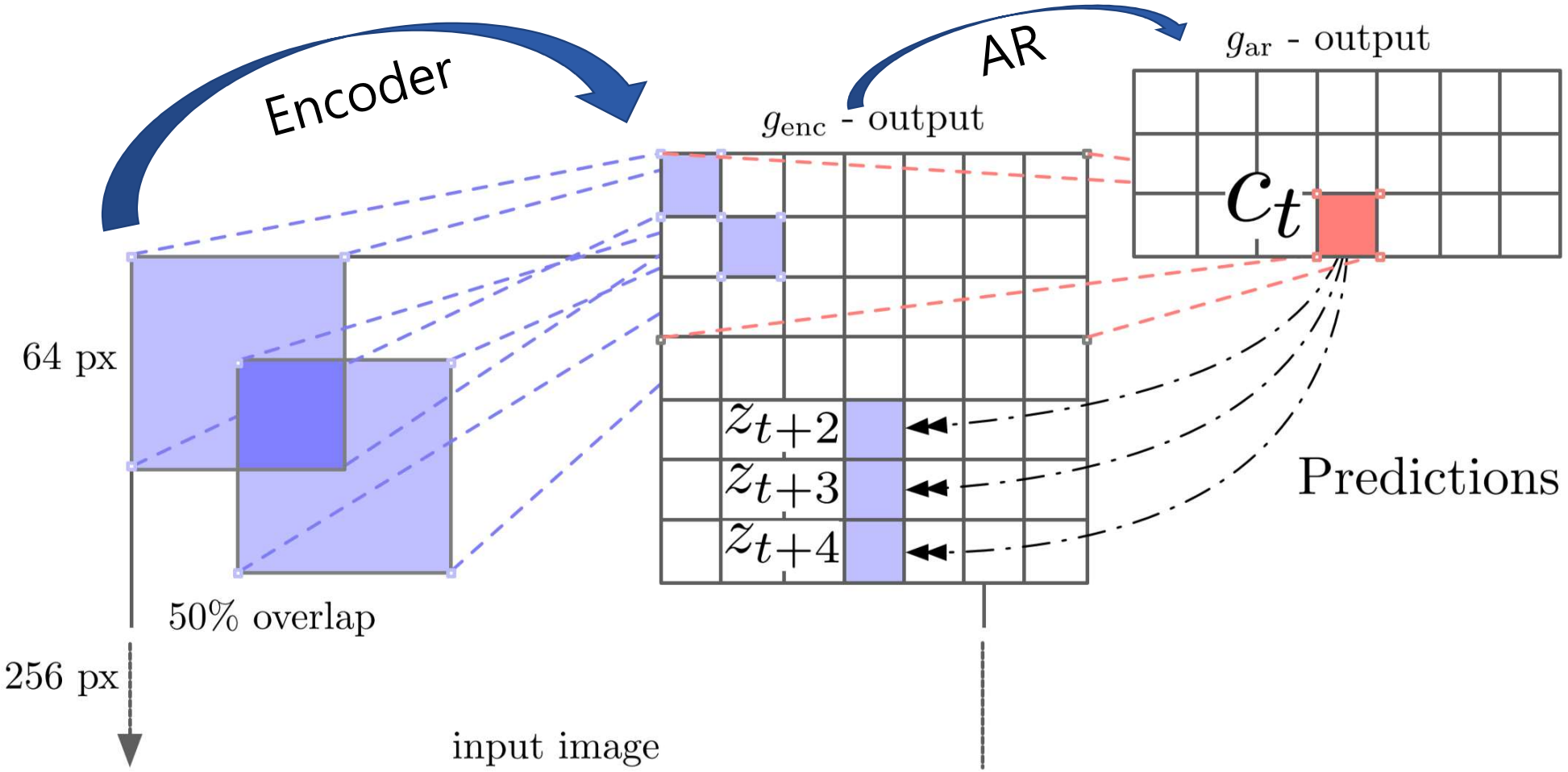
Jigsaw [Noroozi, 2016]와 비교하여, 모든 Permutation을 학습 라벨로 사용할 수 있다는 강점이 있음.

이미지 - Contrast Predictive Coding [Oord, 2018]



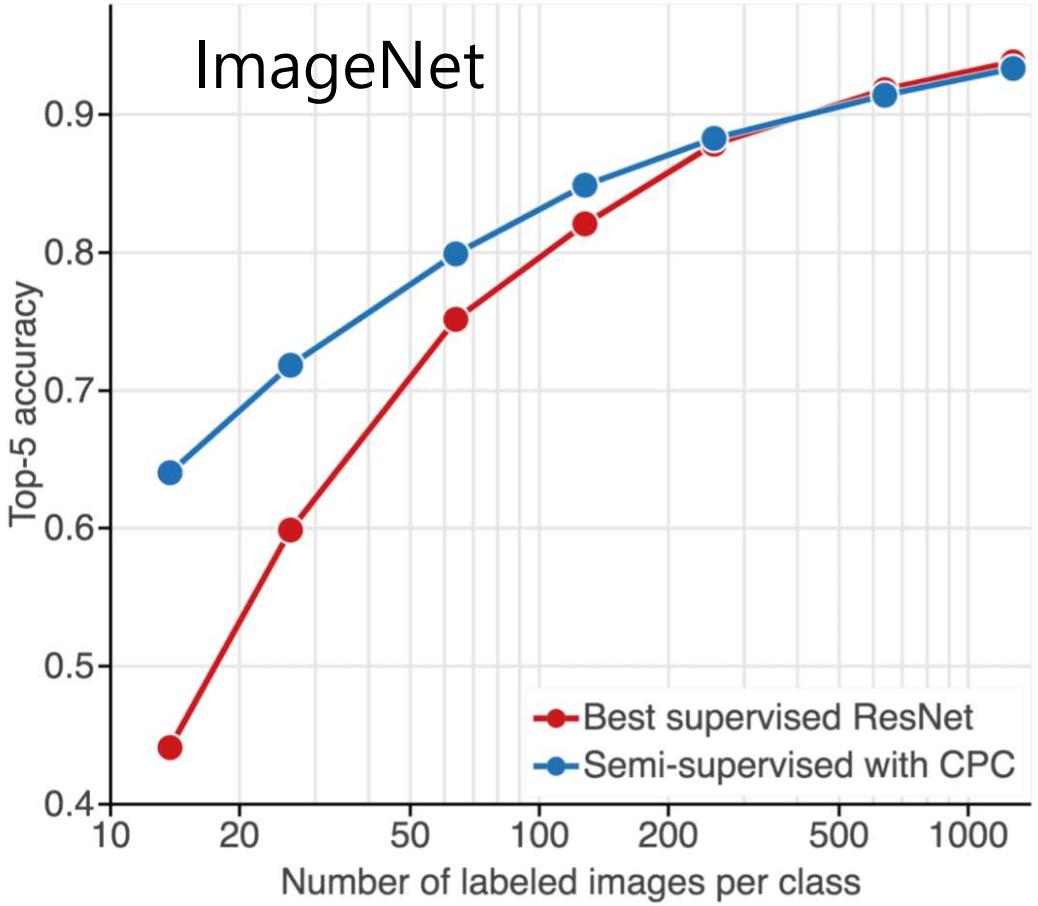
1) Auto-regressive Prediction을 Noise Contrastive Estimation 변형 문제로 변환
2) 이러한 InfoNCE Formulation이 (c_t, x_(t+k)) 간 Mutual Information의 Lower Bound를 최대화하는 것과 동치임을 밝힘

이미지 - Contrast Predictive Coding [Oord, 2018]



이미지 - Contrast Predictive Coding [Hénaff, 2019]

- CPC [Oord, 2018]의 개량 모델을 다룬 후속 연구

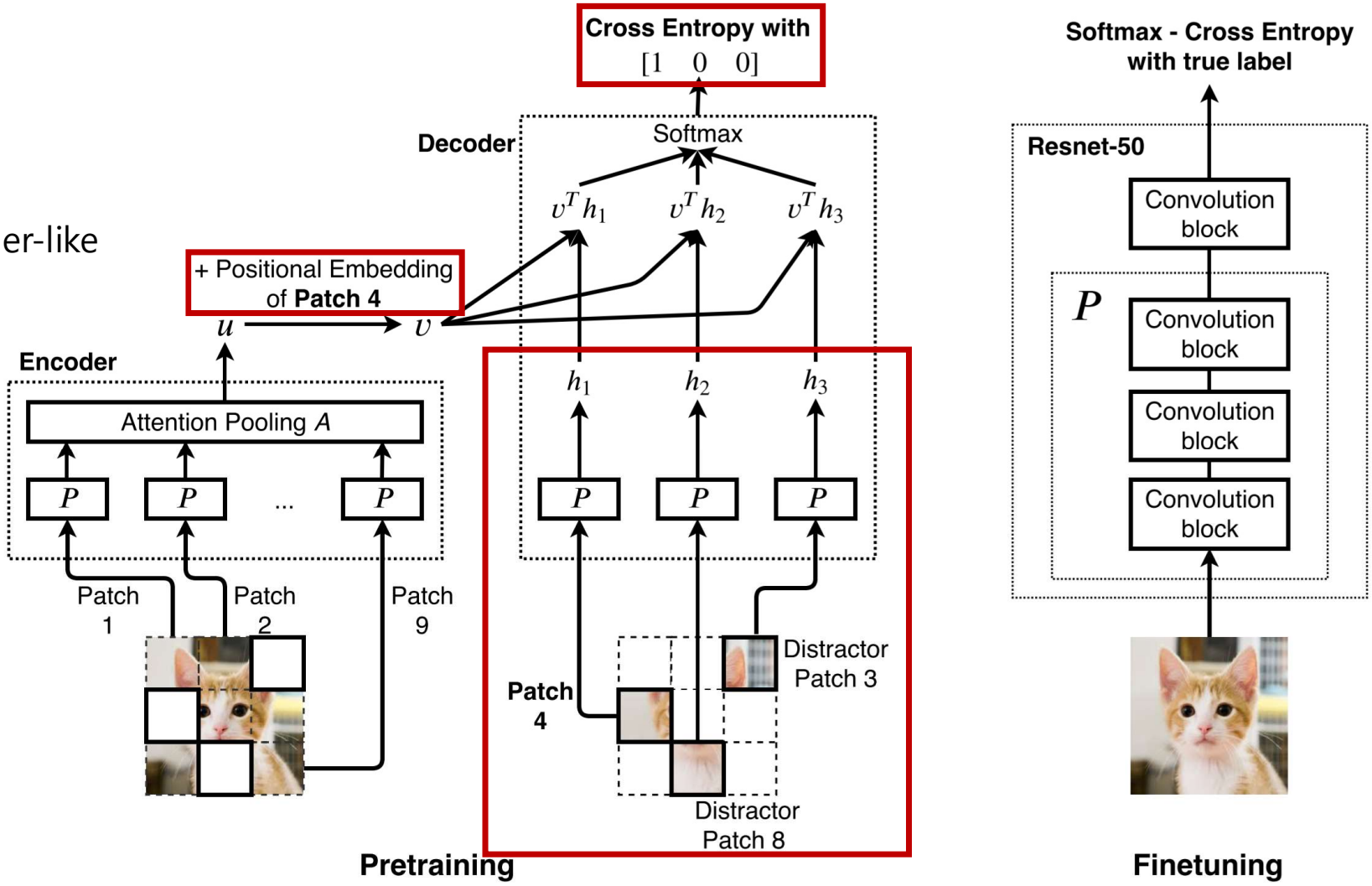


Pascal VOC07

Method	mAP
Transfer from labeled ImageNet:	
Supervised - ResNet-152	74.7
Transfer from unlabeled ImageNet:	
Exemplar (Ex) [17]	60.9
Motion Segmentation (MS) [50]	61.1
Colorization (Col) [69]	65.5
Relative Position (RP) [14]	66.8
Combination of Ex + MS + Col + RP [15]	70.5
Deep Cluster [8]	65.9
Deeper Cluster [9]	67.8
CPC - ResNet-101	70.6
CPC - ResNet-170	**72.1**

이미지 - Selfie [Trinh, 2019]

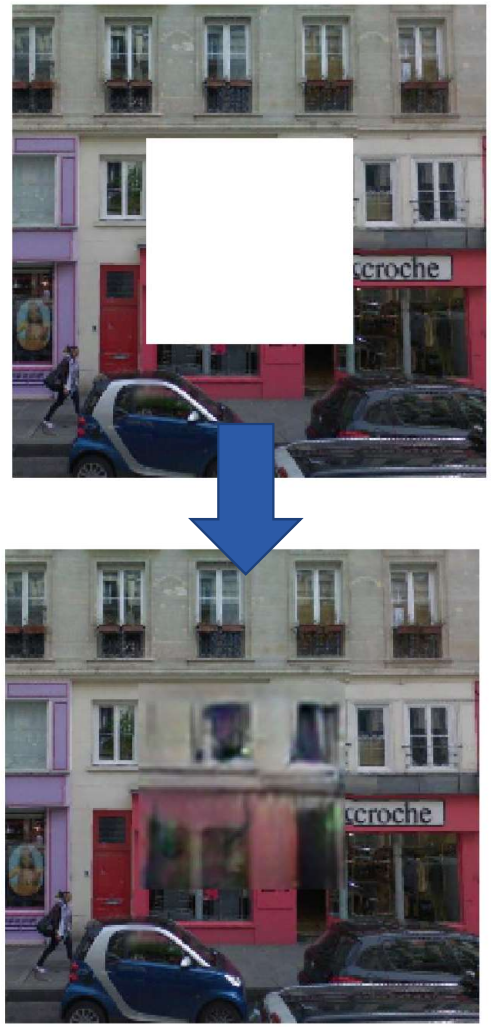
A: Transformer-like

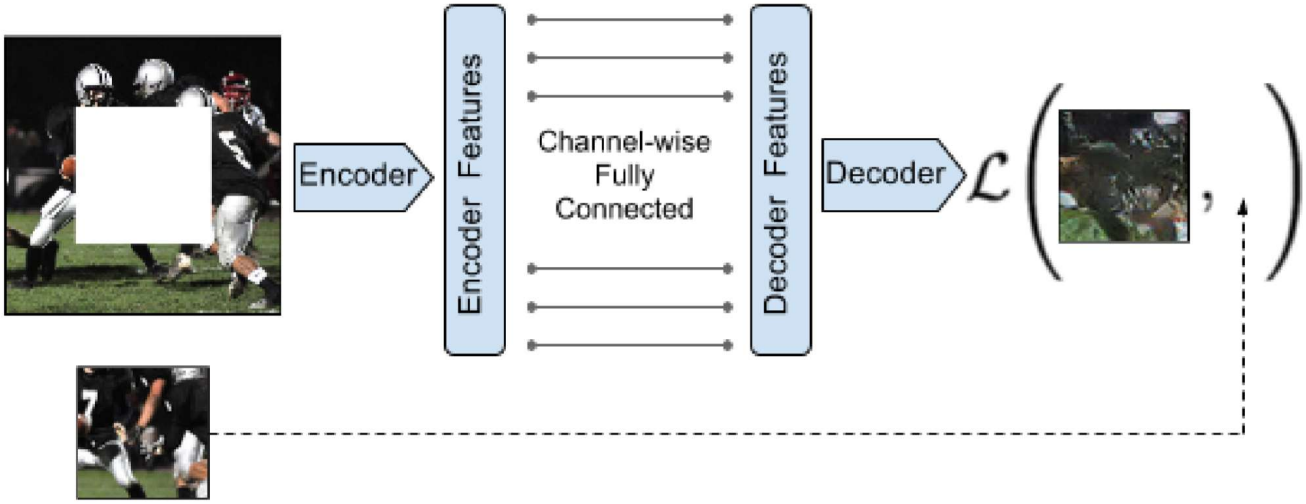


여기서 언급할 Self-supervision 연구들

- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015] & Flow Prediction [Luo, 2017]
  - Ordering [Misra, 2018] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Relative Depth [Jiang, 2018]

이미지 - Image Inpainting [Pathak, 2016]





Pretraining Method	Supervision	Pretraining time	Classification	Detection	Segmentation
ImageNet [26]	1000 class labels	3 days	78.2%	56.8%	48.0%
Random Gaussian	initialization	< 1 minute	53.3%	43.4%	19.8%
Autoencoder	-	14 hours	53.8%	41.9%	25.2%
Agrawal *et al.* [1]	egomotion	10 hours	52.9%	41.8%	-
Wang *et al.* [39]	motion	1 week	58.7%	47.4%	-
Doersch *et al.* [7]	relative context	4 weeks	55.3%	46.6%	-
Ours	context	14 hours	56.5%	44.5%	30.0%

Table 2: Quantitative comparison for classification, detection and semantic segmentation. Classification and Fast-RCNN Detection results are on the PASCAL VOC 2007 test set. Semantic segmentation results are on the PASCAL VOC 2012 validation set from the FCN evaluation described in Section 5.2.3, using the additional training data from [18], and removing overlapping images from the validation set [28].

이미지 - Colorization [Larsson, 2017]

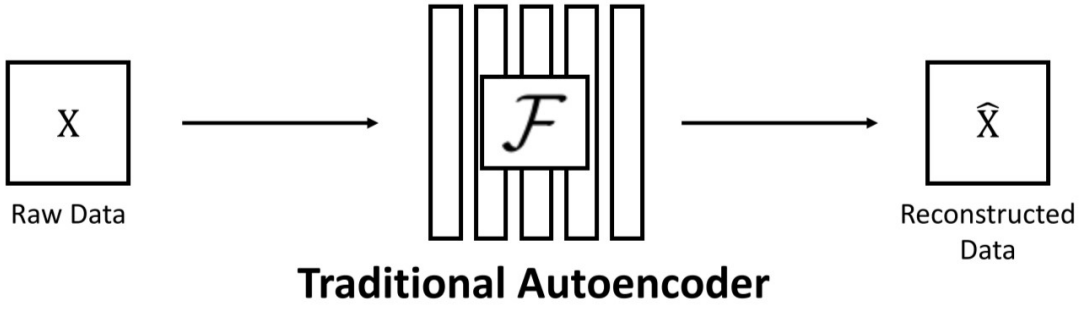
Ex. 3: **Colorization** (predict color given intensity)

Initialization	Architecture	Class. %mAP	Seg. %mIU
ImageNet (+FoV)	VGG-16	86.9	69.5
Pathak *et al.* [31]	AlexNet	56.5	29.7
Wang & Gupta [38]	AlexNet	58.7	-
Donahue *et al.* [5]	AlexNet	60.1	35.2
Doersch *et al.* [4, 5]	AlexNet	65.3	-
Zhang *et al.* (col) [42]	AlexNet	65.6	35.6
Zhang *et al.* (s-b) [43]	AlexNet	67.1	36.0
Noroozi & Favaro [28]	Mod. AlexNet	68.6	-
Larsson *et al.* [20]	VGG-16	-	50.2
Our method	AlexNet	65.9	38.4
(+FoV)	VGG-16	**77.2**	56.0
(+FoV)	ResNet-152	**77.3**	**60.0**

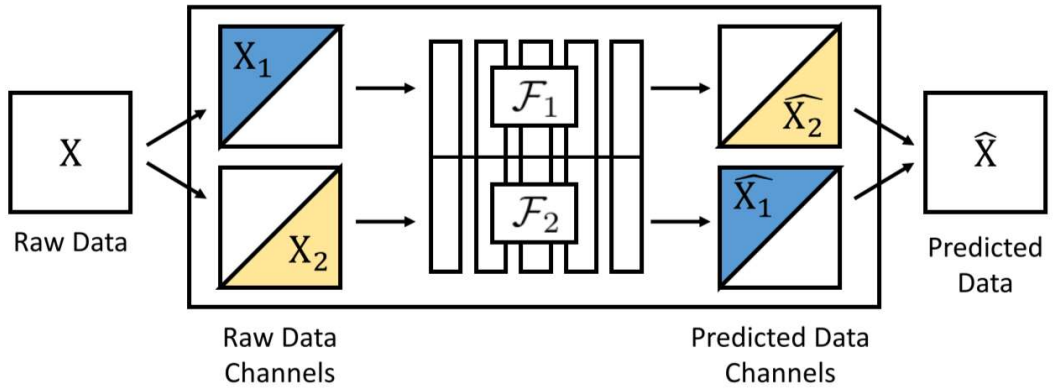
Domain Mismatch

Initialization	Grayscale input		Color input
Classification	66.5	<	69.5
Colorization	56.0		55.9

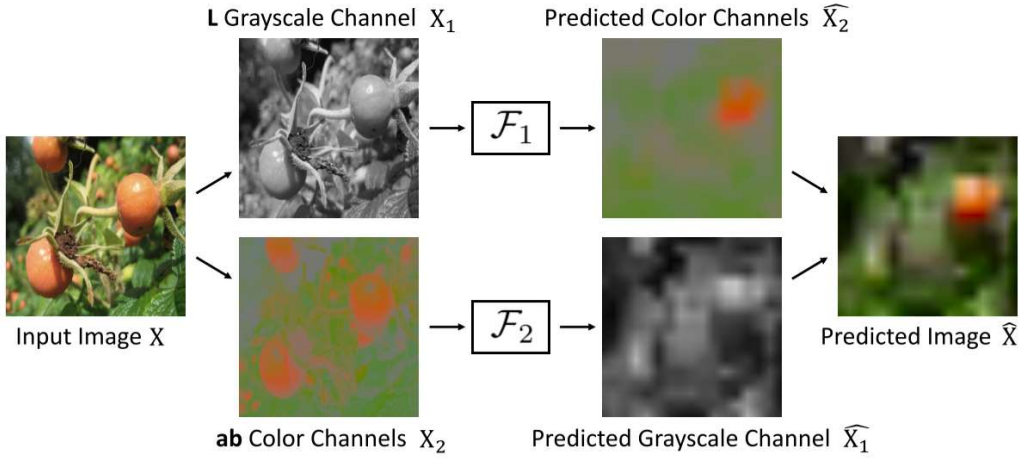
이미지 - Split-brain [Zhang, 2017]



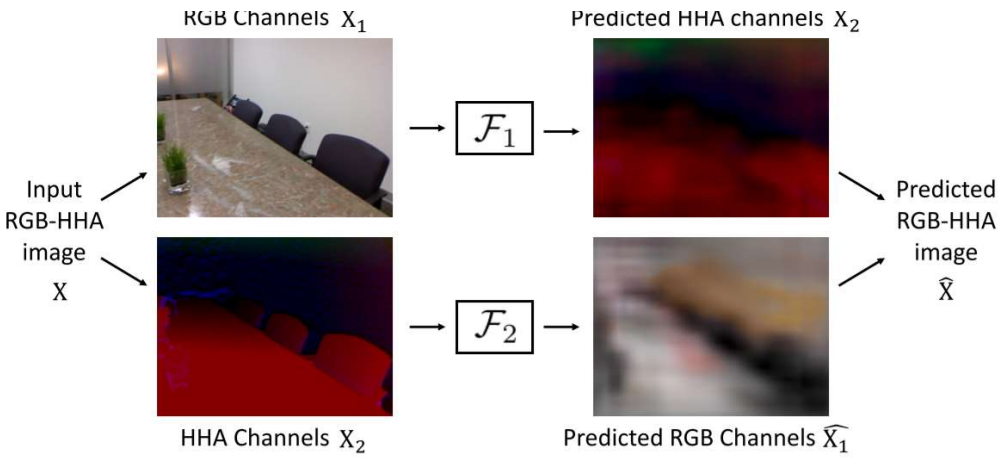
Traditional Autoencoder



Split-Brain Autoencoder



(a) ***Lab*** **Images**



(b) **RGB-D Images**

여기서 언급할 Self-supervision 연구들

- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015] & Flow Prediction [Luo, 2017]
  - Ordering [Misra, 2018] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Relative Depth [Jiang, 2018]

이미지 - Counting [Noroozi, 2017]

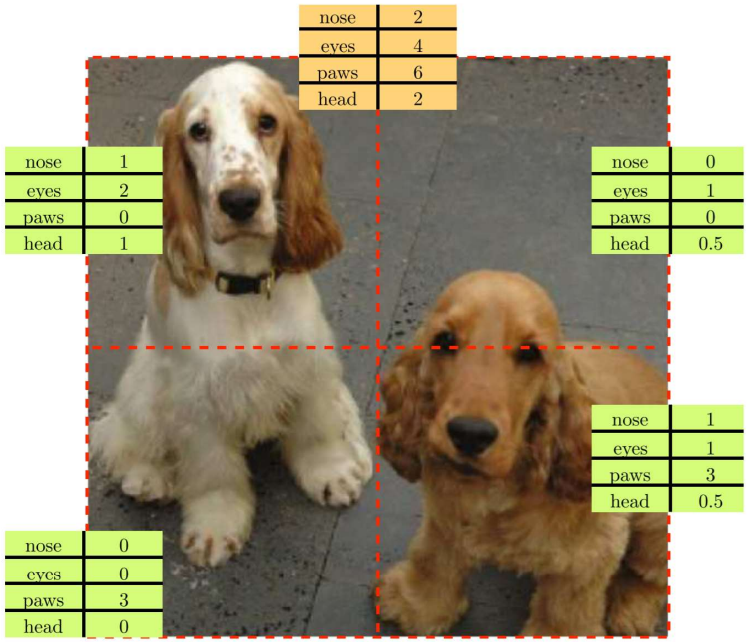
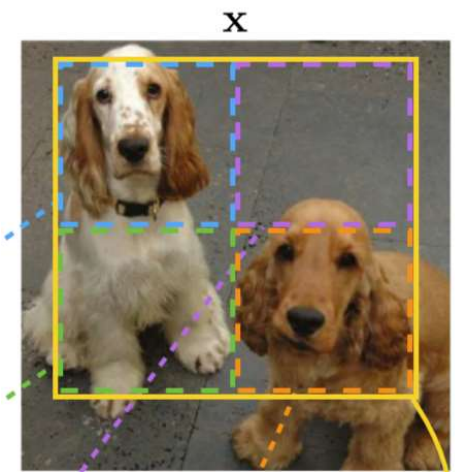
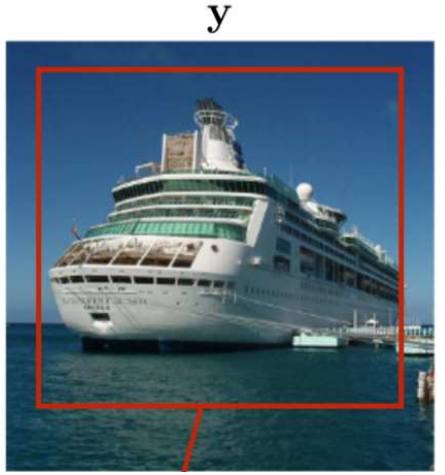


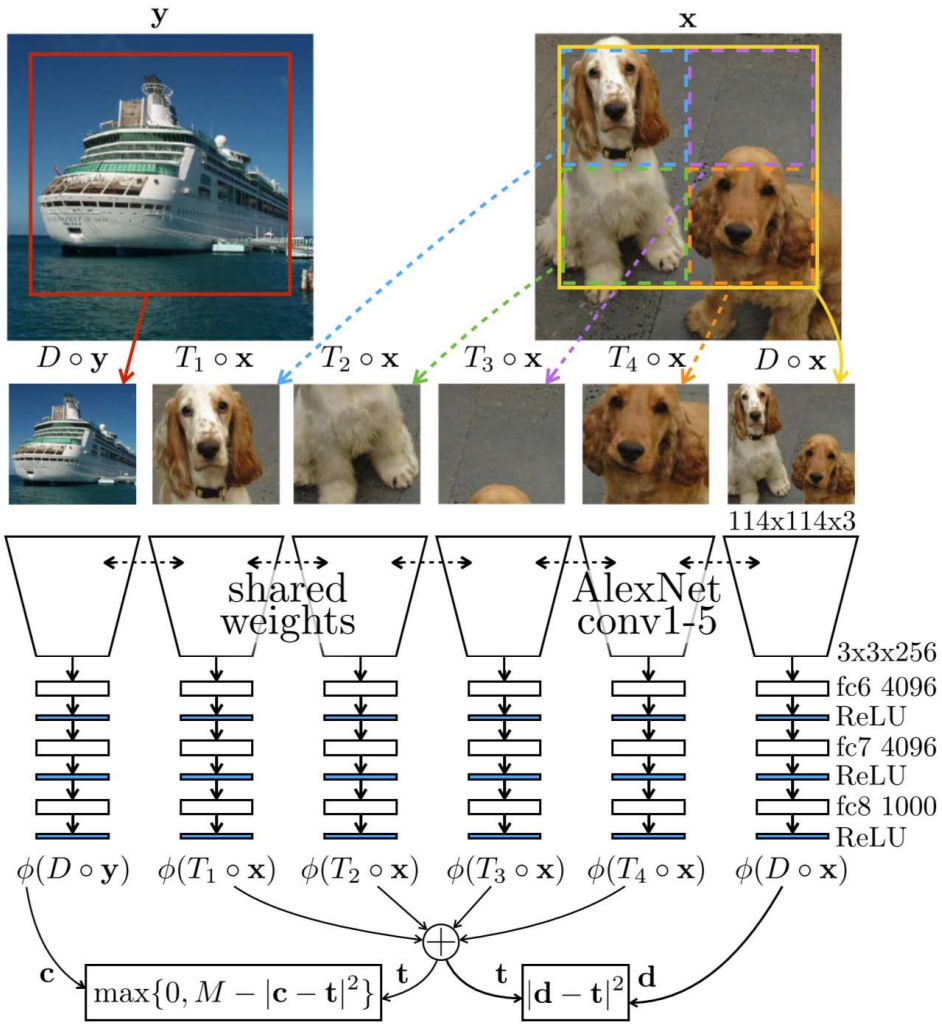
Figure 1: The number of visual primitives in the whole image should match the sum of the number of visual primitives in each tile (dashed red boxes).





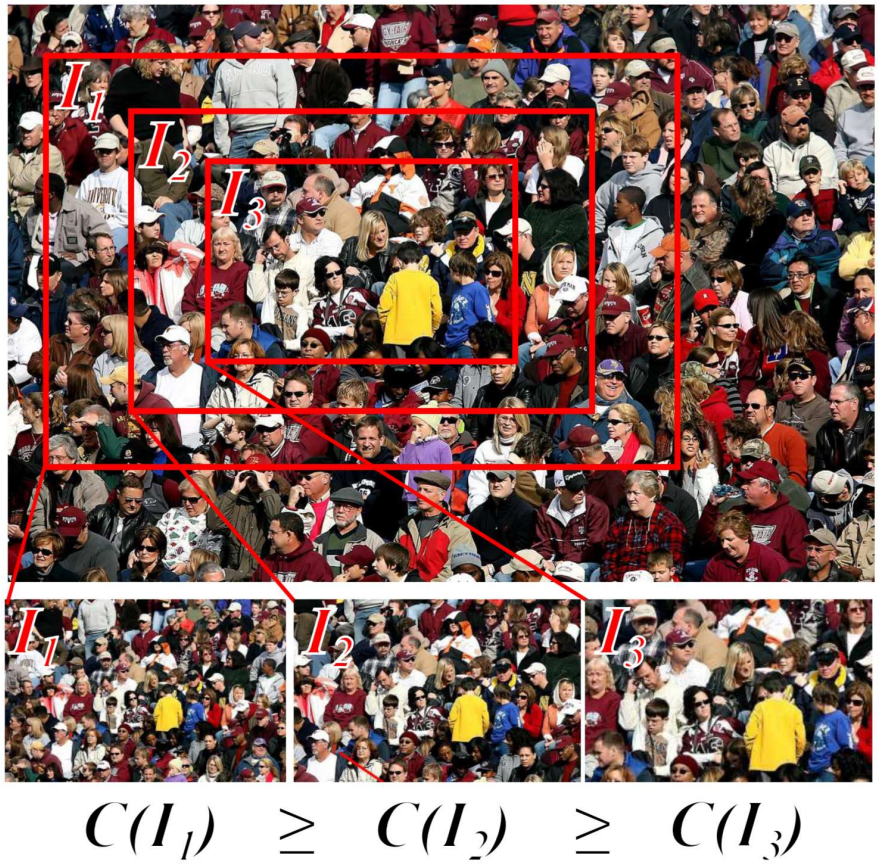
1) X에서 나타나는 Semantic한 Object는 Y에서 나타나지 않는다.
2) 전체 X의 이미지에서 나타나는 Object의 개수는 X의 (겹치지 않는) 패치 이미지에서 나타나는 모든 Object의 개수를 합친 것과 같다

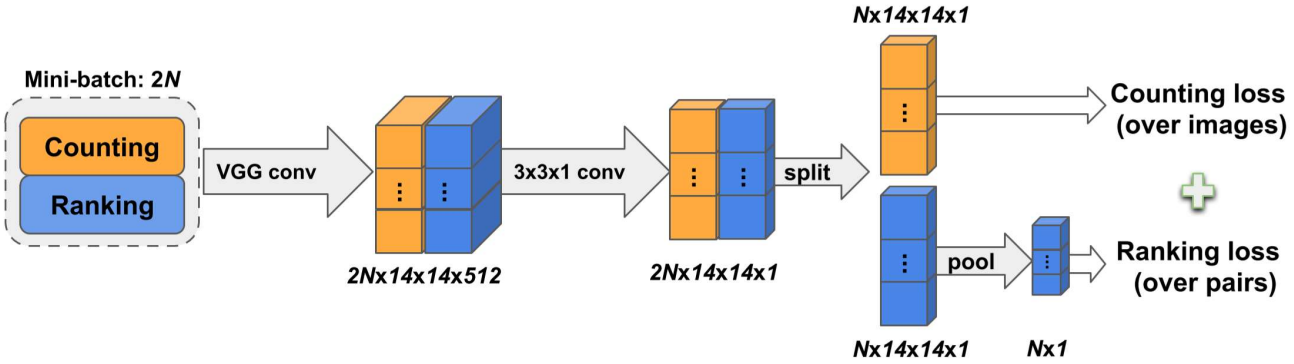
이미지 - Counting [Noroozi, 2017]



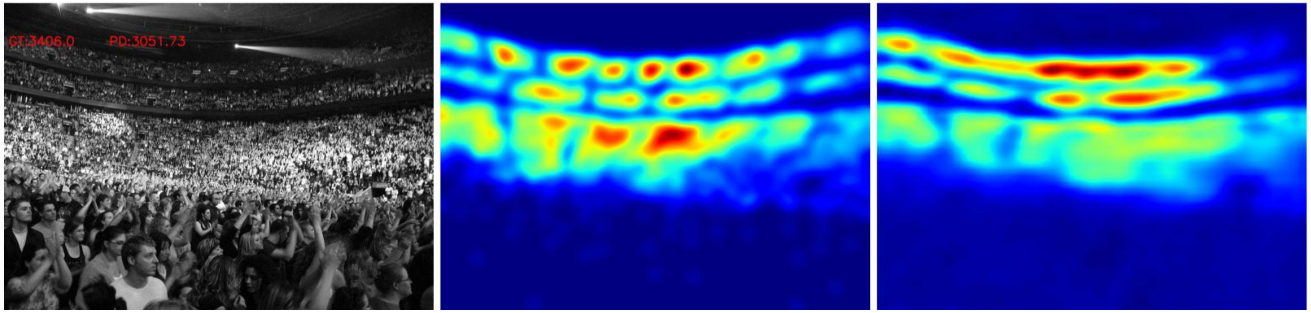
Method	Ref	Class.	Det.	Segm.
Supervised [20]	[43]	79.9	56.8	48.0
Random	[33]	53.3	43.4	19.8
Context [9]	[19]	55.3	46.6	-
Context [9]*	[19]	65.3	51.1	-
Jigsaw [30]	[30]	67.6	**53.2**	37.6
ego-motion [1]	[1]	52.9	41.8	-
ego-motion [1]*	[1]	54.2	43.9	-
Adversarial [10]*	[10]	58.6	46.2	34.9
ContextEncoder [33]	[33]	56.5	44.5	29.7
Sound [31]	[44]	54.4	44.0	-
Sound [31]*	[44]	61.3	-	-
Video [41]	[19]	62.8	47.4	-
Video [41]*	[19]	63.1	47.2	-
Colorization [43]*	[43]	65.9	46.9	35.6
Split-Brain [44]*	[44]	67.1	46.7	36.0
ColorProxy [22]	[22]	65.9	-	**38.0**
WatchingObjectsMove [32]	[32]	61.0	52.2	-
Counting		**67.7**	51.4	36.6

이미지 - Ranking [Liu, 2019]

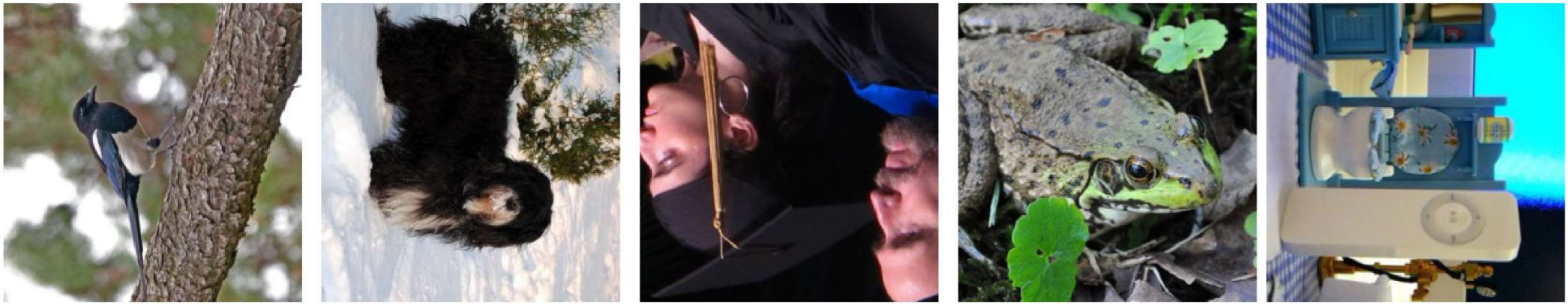




Input GT (3406) Prediction (3052)

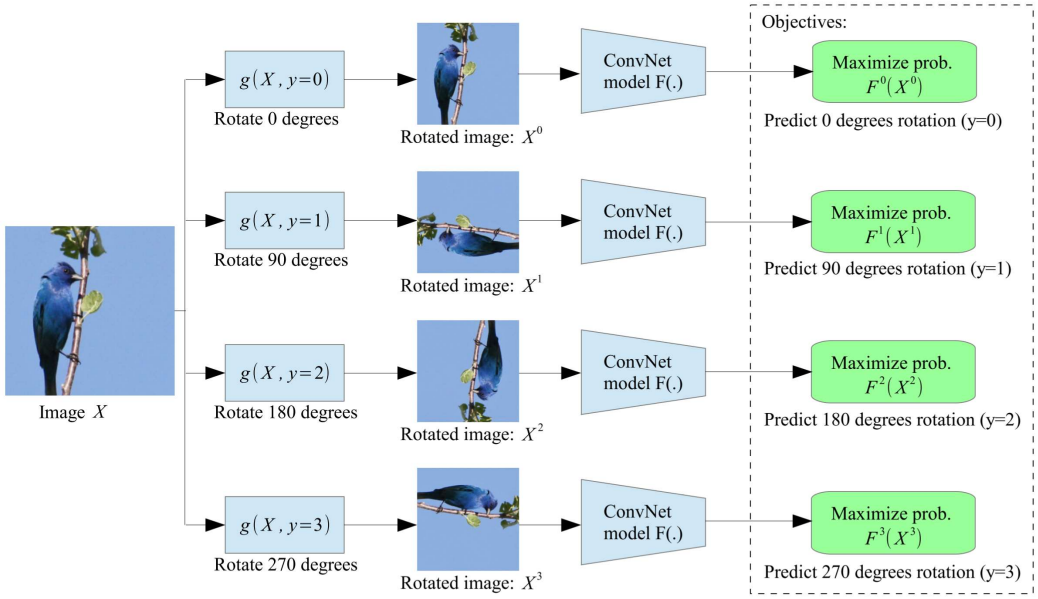


이미지 - RotNet [Gidaris, 2018]



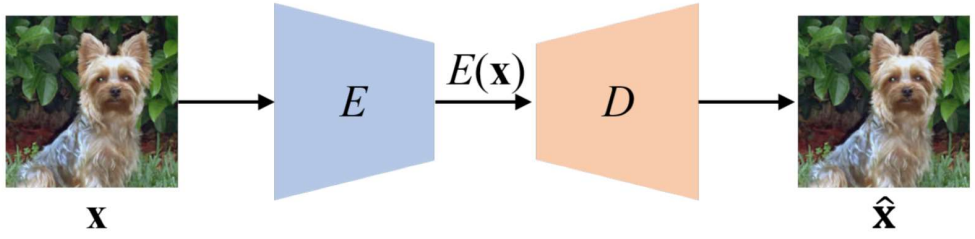
90° rotation　270° rotation　180° rotation　0° rotation　270° rotation

이미지 - RotNet [Gidaris, 2018]

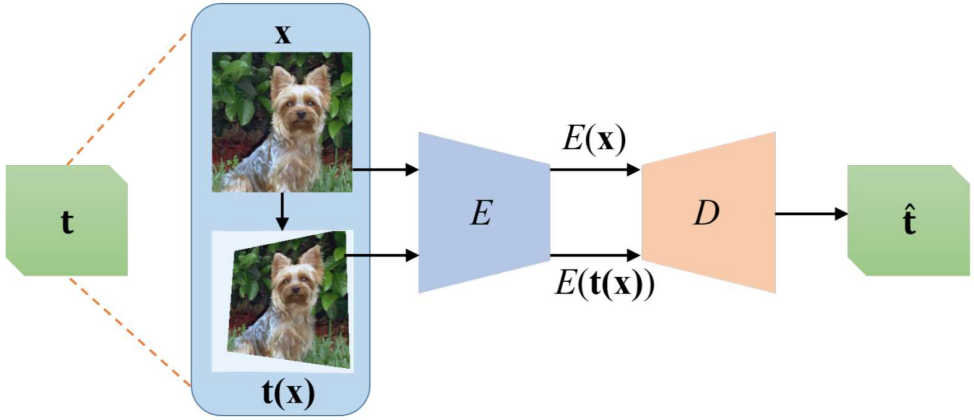


Method	Conv4	Conv5
ImageNet labels from (Bojanowski & Joulin, 2017)	59.7	59.7
Random from (Noroozi & Favaro, 2016)	27.1	12.0
Tracking Wang & Gupta (2015)	38.8	29.8
Context (Doersch et al., 2015)	45.6	30.4
Colorization (Zhang et al., 2016a)	40.7	35.2
Jigsaw Puzzles (Noroozi & Favaro, 2016)	45.3	34.6
BIGAN (Donahue et al., 2016)	41.9	32.2
NAT (Bojanowski & Joulin, 2017)	-	36.0
(Ours) RotNet	50.0	43.8

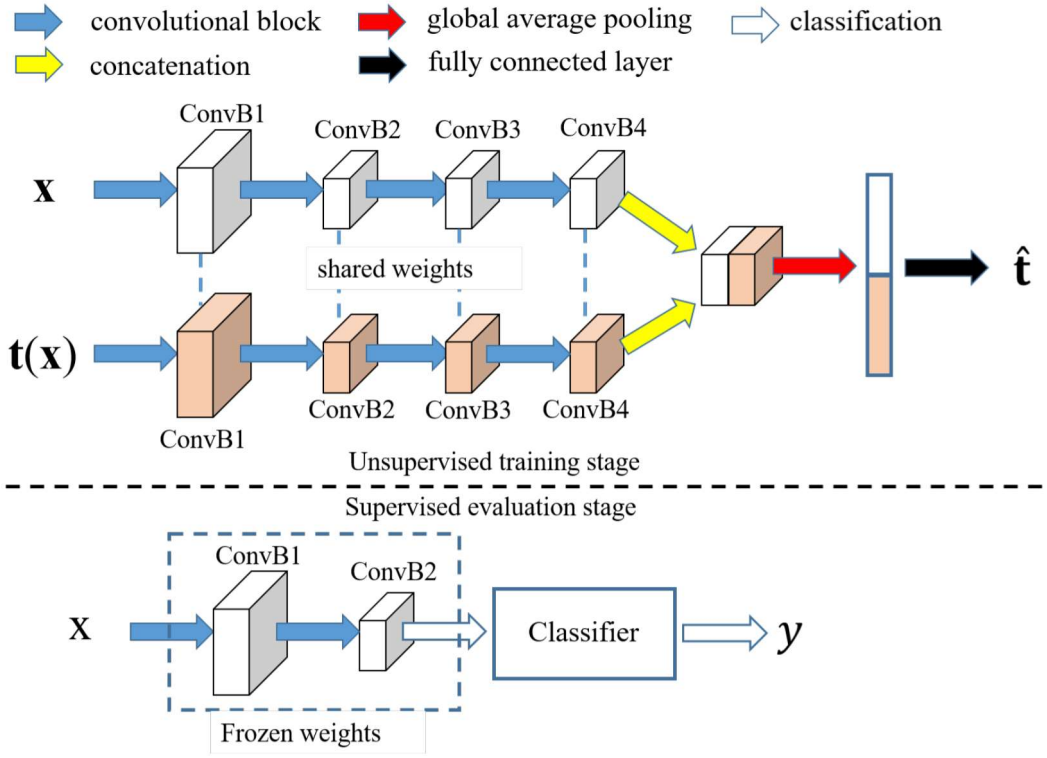
이미지 - AET [Zhang, 2019]



(a) Auto-Encoding Data (AED)



(b) Auto-Encoding Transformation (AET)



여기서 언급할 Self-supervision 연구들

- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015]
  - Ordering [Misra, 2016] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Flow Prediction [Luo, 2017], Relative Depth [Jiang, 2018]

비디오 - Frame Prediction [Srivastava, 2015]

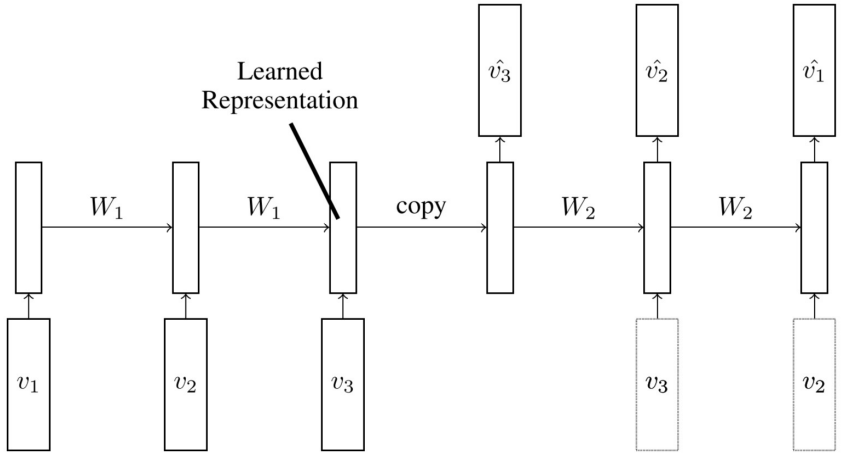


Figure 2. LSTM Autoencoder Model

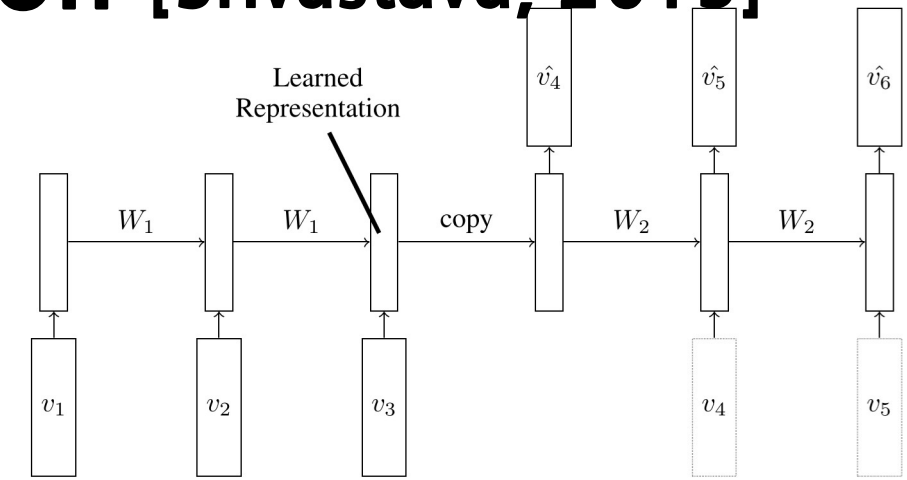
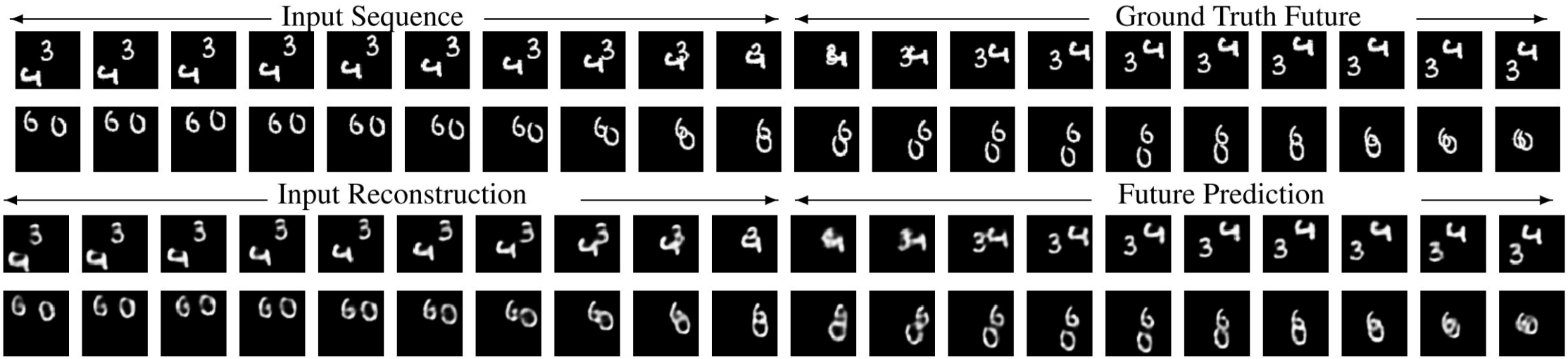
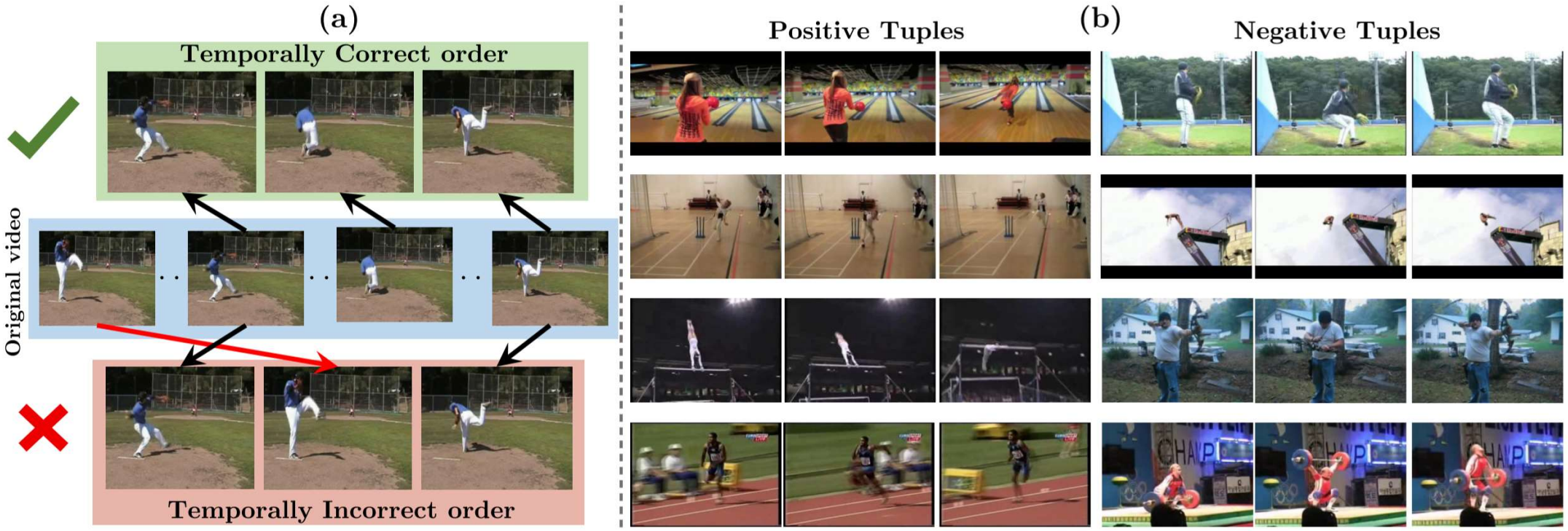


Figure 3. LSTM Future Predictor Model

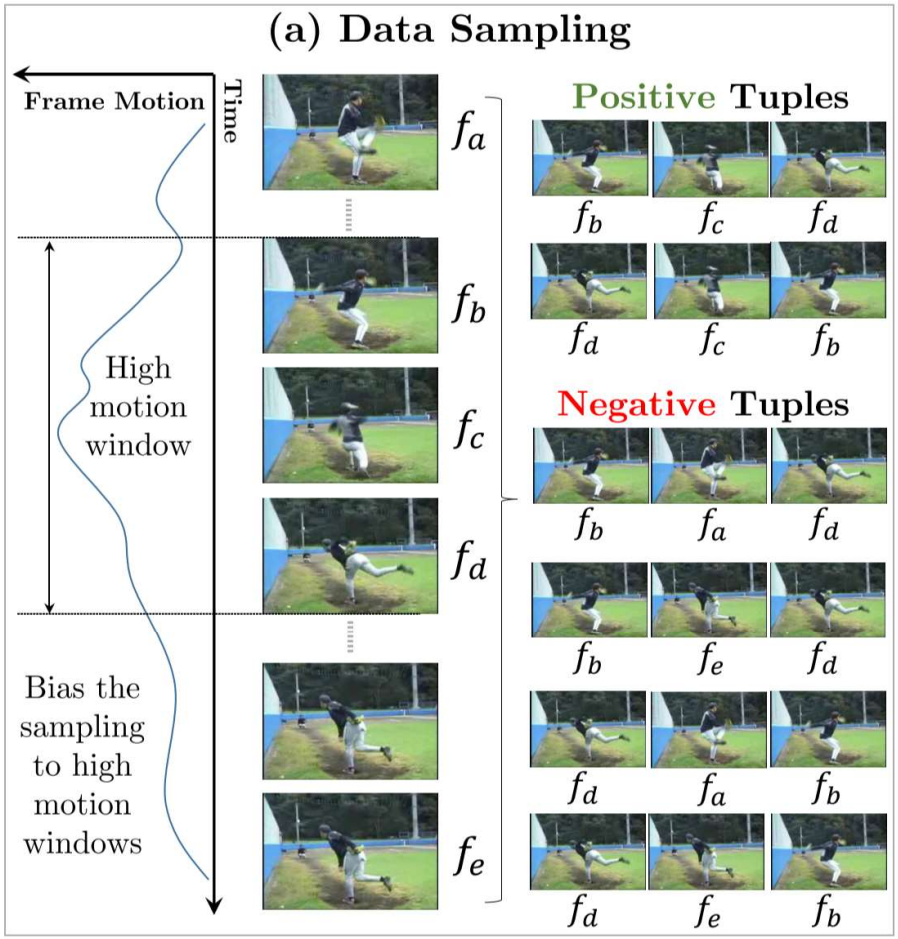


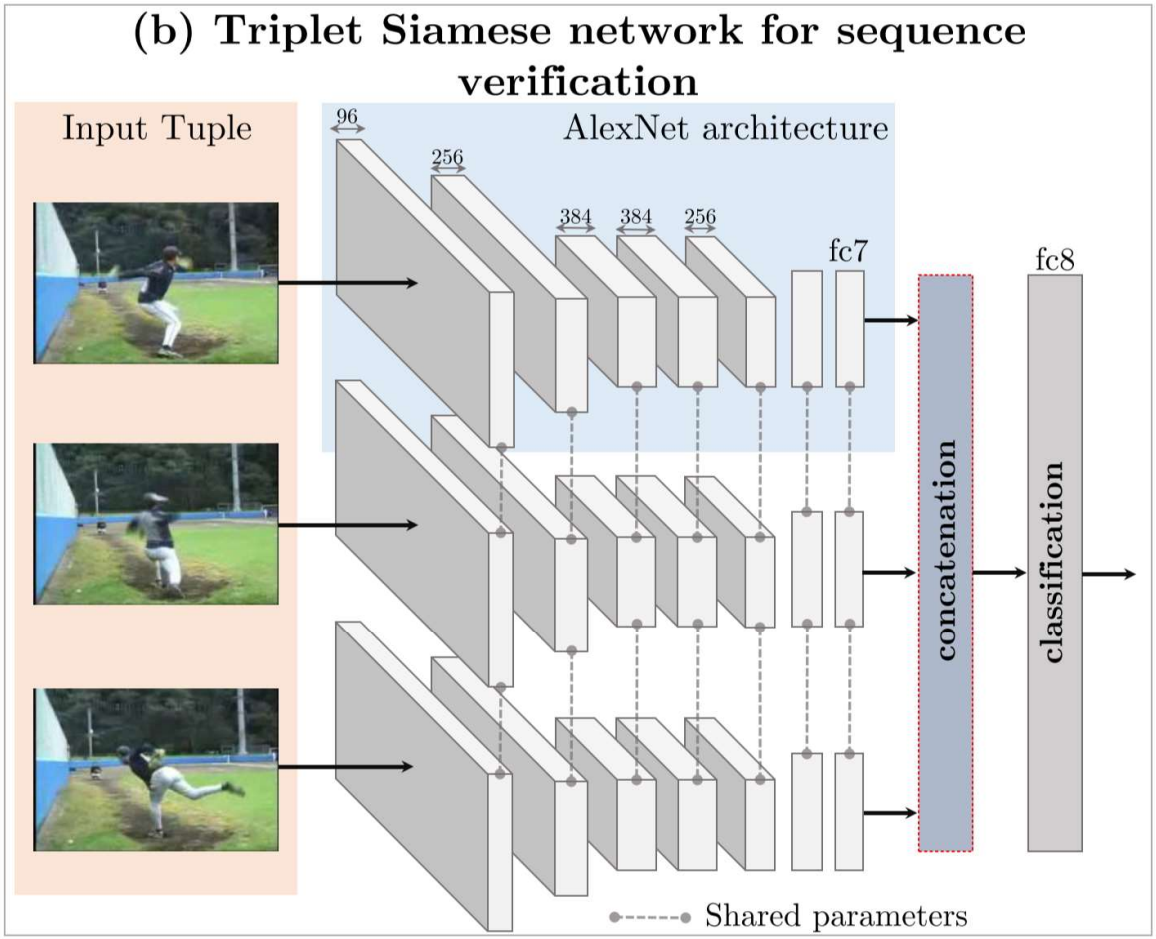
Two Layer Composite Model with a Conditional Future Predictor

비디오 - Frame Ordering [Misra, 2016]

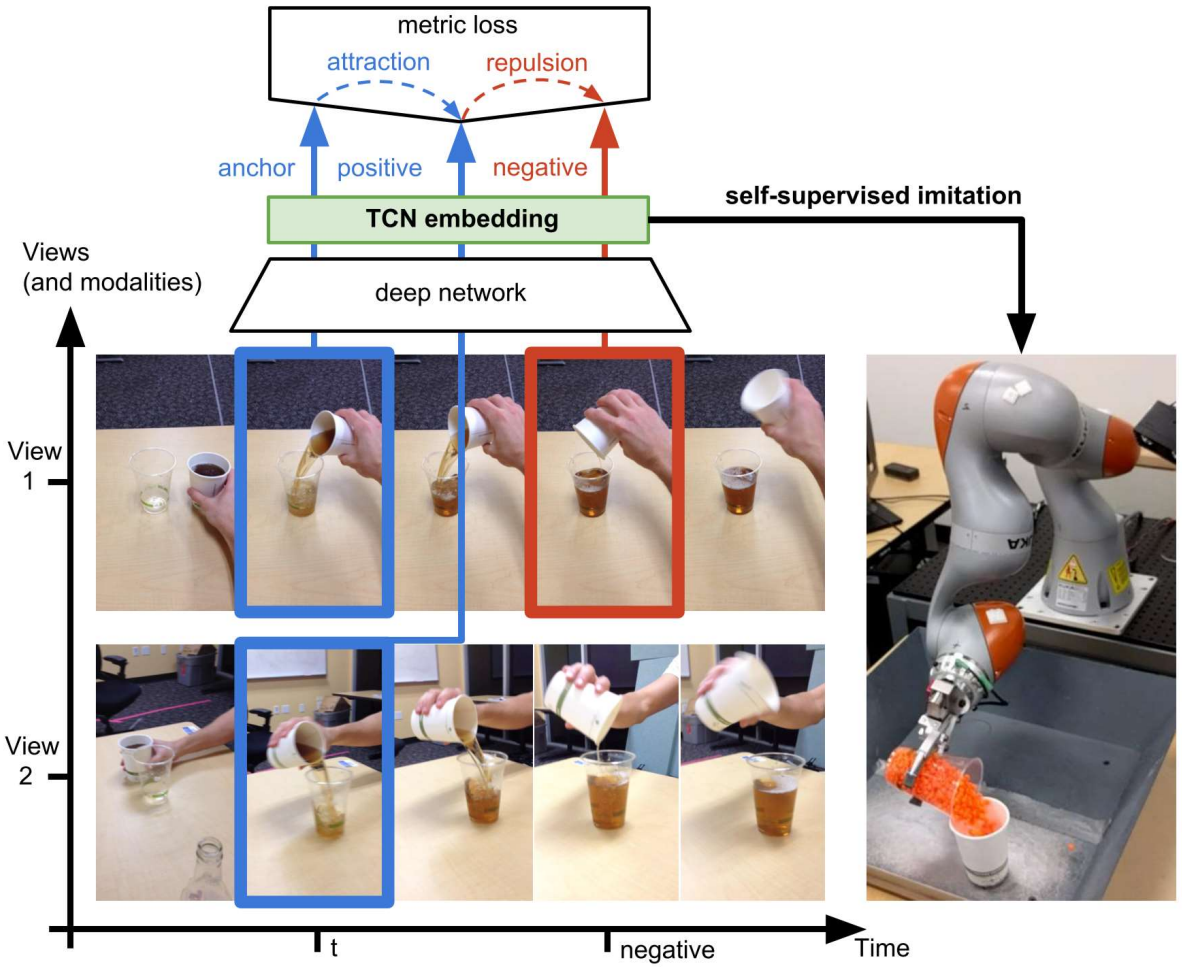


비디오 - Frame Ordering [Misra, 2016]





비디오 - Time-contrastive Learning [Sermanet, 2018]



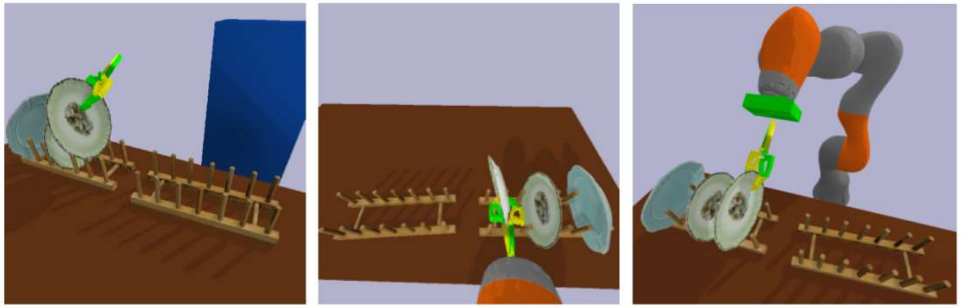


Fig. 5: **Simulated dish rack task.** *Left:* Third-person VR demonstration of the dish rack task. *Middle:* View from the robot camera during training. *Right:* Robot executing the dish rack task.

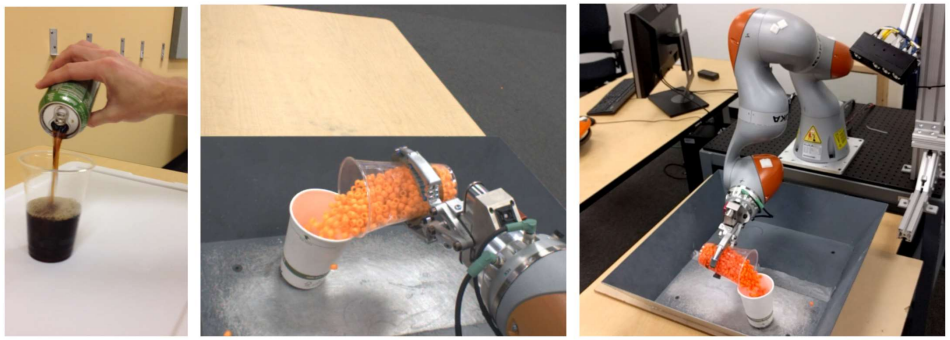


Fig. 6: **Real robot pouring task.** *Left:* Third-person human demonstration of the pouring task. *Middle:* View from the robot camera during training. *Right:* Robot executing the pouring task.

비디오 - Flow Prediction [Luo, 2017]

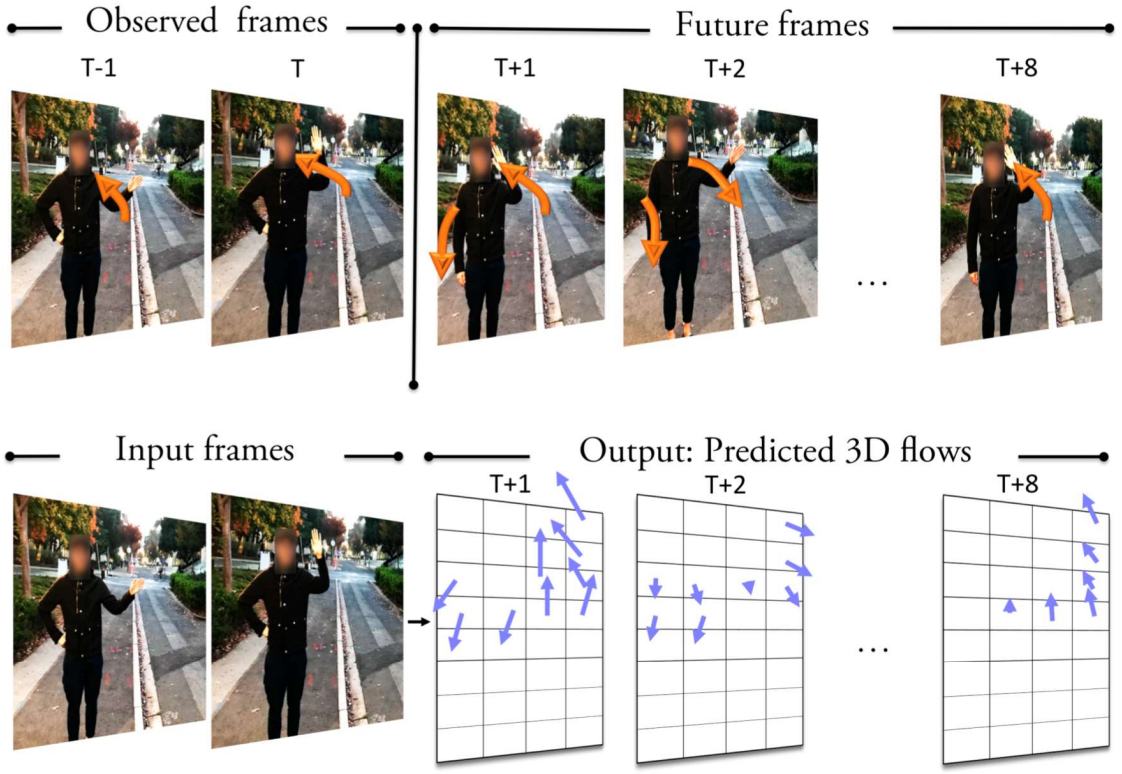
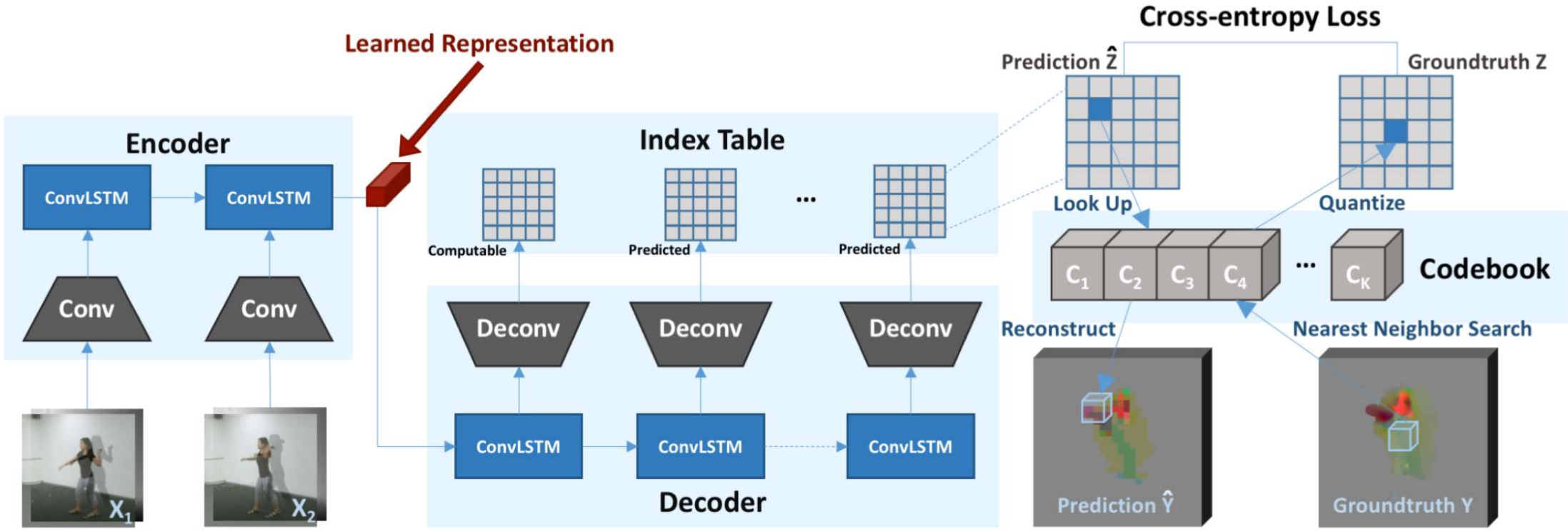


Figure 1. We propose a method that learns a video representation by predicting a sequence of basic motions described as atomic 3D flows. The learned representation is then extracted from this model to recognize activities.

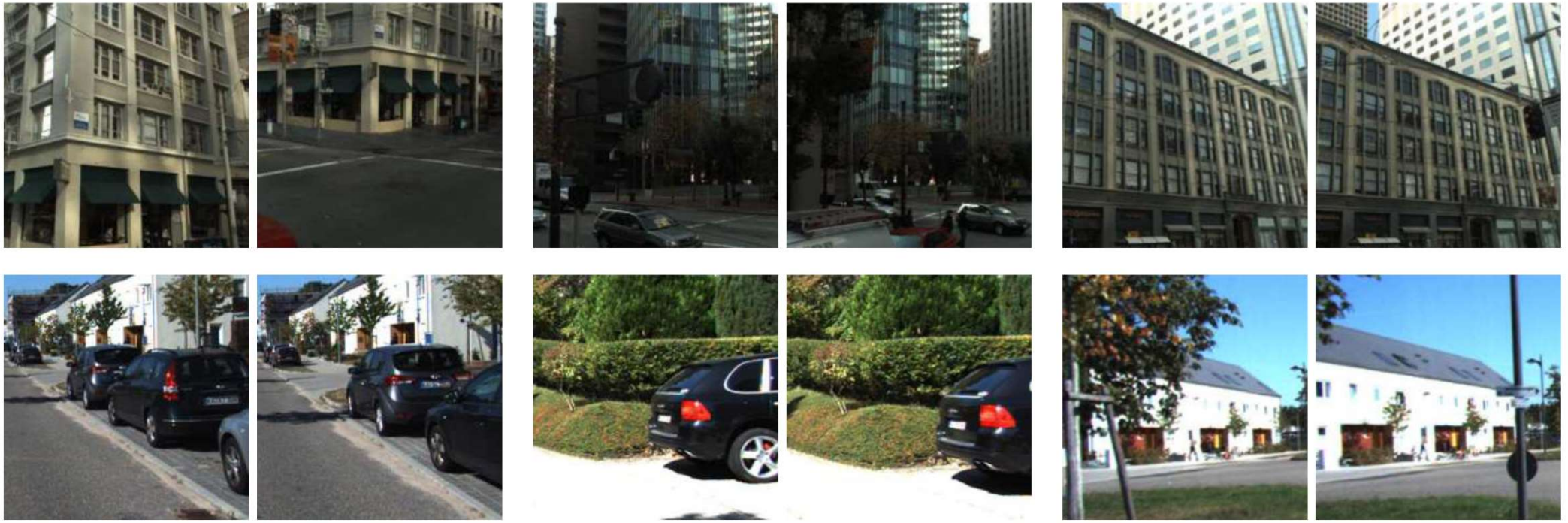
비디오 - Flow Prediction [Luo, 2017]



여기서 언급할 Self-supervision 연구들

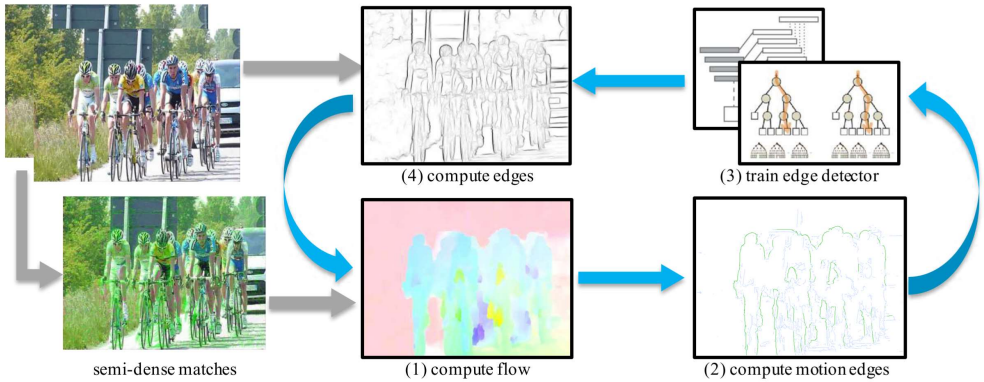
- 이미지
  - Context Prediction 계열
    - ✓ Jigsaw [Noroozi, 2016] & DeepPermNet [Santa, 2017]
    - ✓ Contrastive Predictive Coding ([Oord, 2018], [Hénaff, 2019]) & Selfie [Trinh, 2019]
  - Generation 계열
    - ✓ Image Inpainting [Pathak, 2016] & Colorization [Larsson, 2017] & Split-Brain [Zhang, 2017]
  - Geometric Prediction 계열
    - ✓ Counting [Noroozi, 2017] & Ranking [Liu, 2019] & RotNet [Gidaris, 2018] & AET [Zhang, 2019]
- 비디오
  - Frame Prediction [Srivastava, 2015]
  - Ordering [Misra, 2016] & Time Contrastive Learning [Sermanet, 2018]
  - Ego-motion: [Agrawal, 2015]
  - Label Generation from Hard Program
    - ✓ Edge [Li, 2016], Moving Object [Pathak, 2017], Flow Prediction [Luo, 2017], Relative Depth [Jiang, 2018]

비디오 - Ego-Motion [Agrawal, 2015]



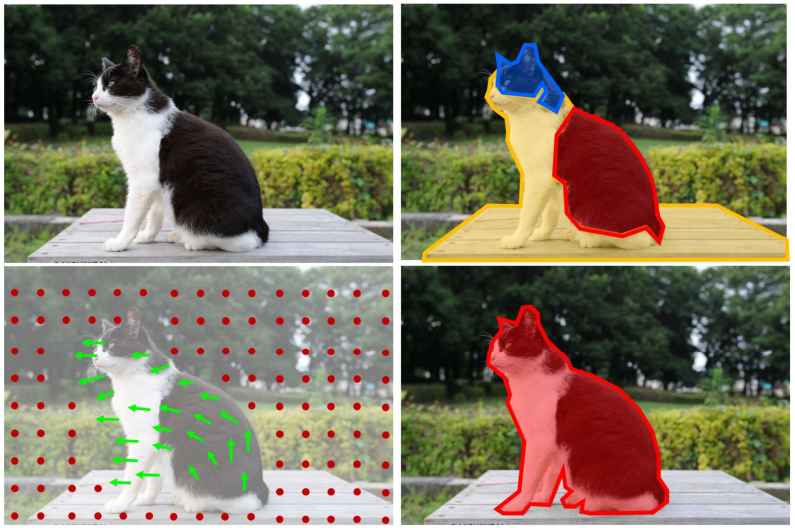
비디오 - Label Generation

Edge [Li, 2016]

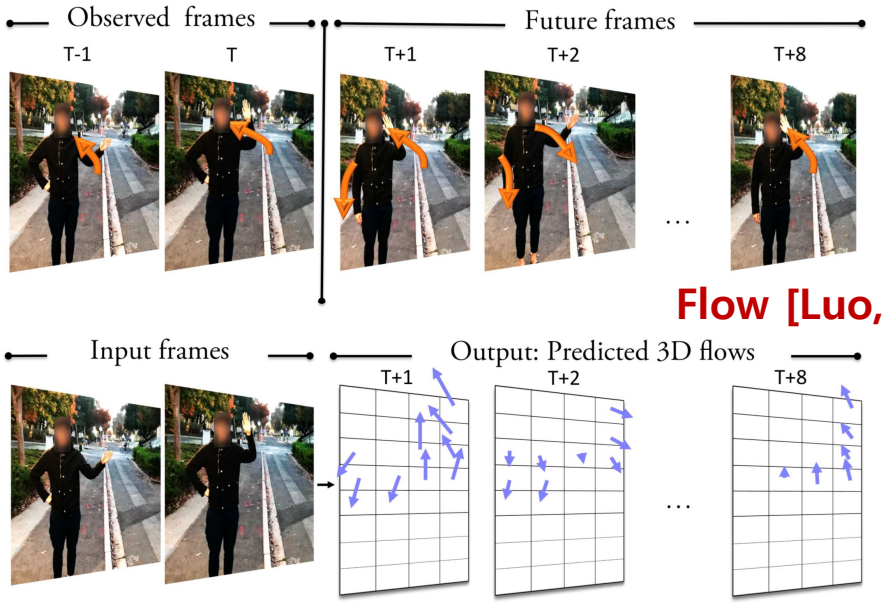


Relative Depth [Jiang, 2018]

Moving Object [Pathak, 2017]



Flow [Luo, 2017],

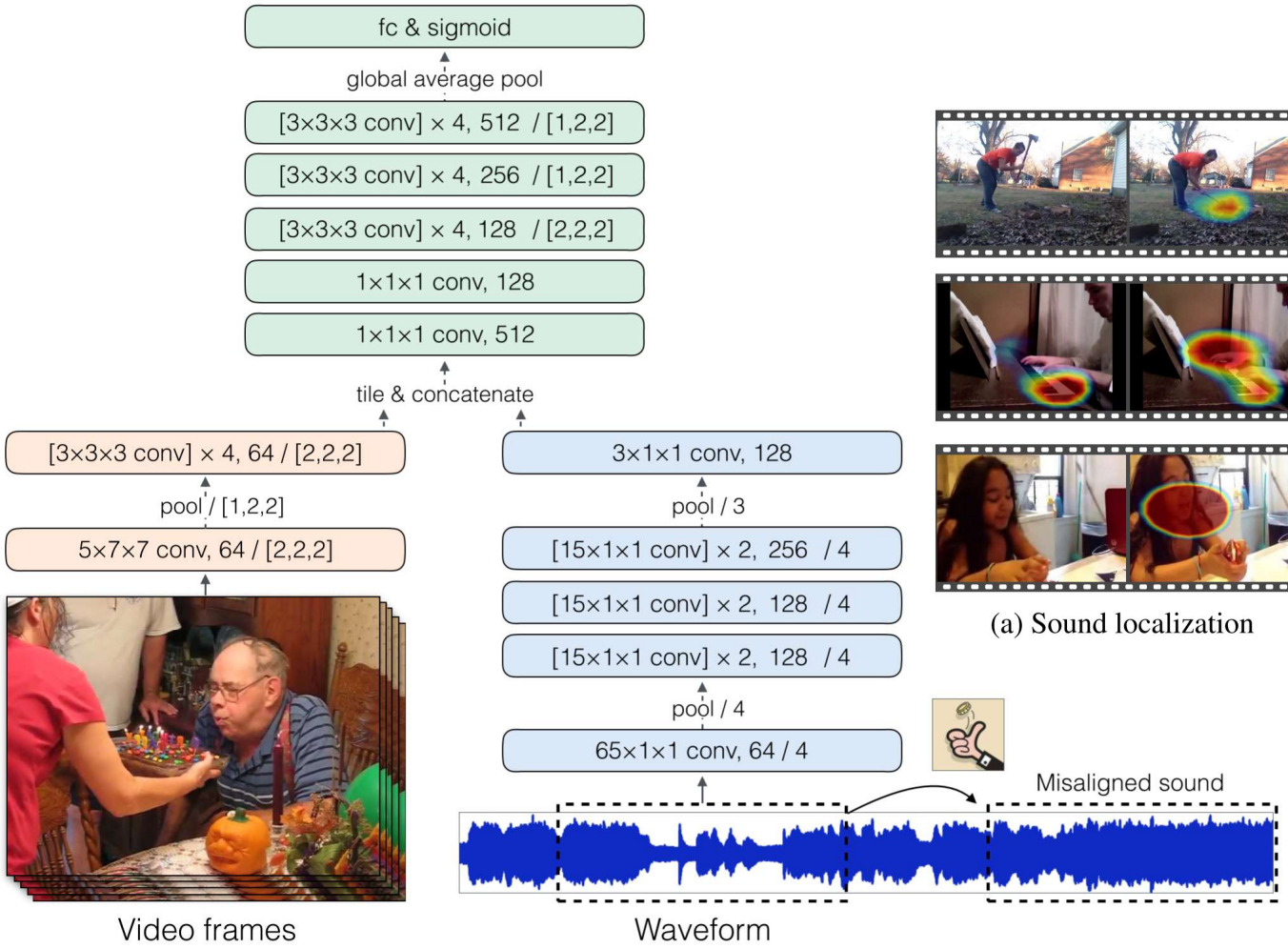


여기서 언급할 Self-supervision 연구들

- 오디오 (w/ Cross Modality)
  - Self-supervised synchronization ([Owens, 2018], [Korbar, 2018])
- ~~텍스트~~
  - ~~BERT [Owens, 2018] & GPT [Radford, 2018] & XLNet [Yang, 2019]~~

- 주로 메소드에 관해서만 다루고, 벤치마크에 대해선 일부만 다룹니다
- 다른 도메인들의 연구에 대해선 다룰 여유가 없을 것 같습니다 흑흑

오디오 w/ Cross-modality - [Owens, 2018]



04 Self-supervised Learning, 어디로 가야 하나

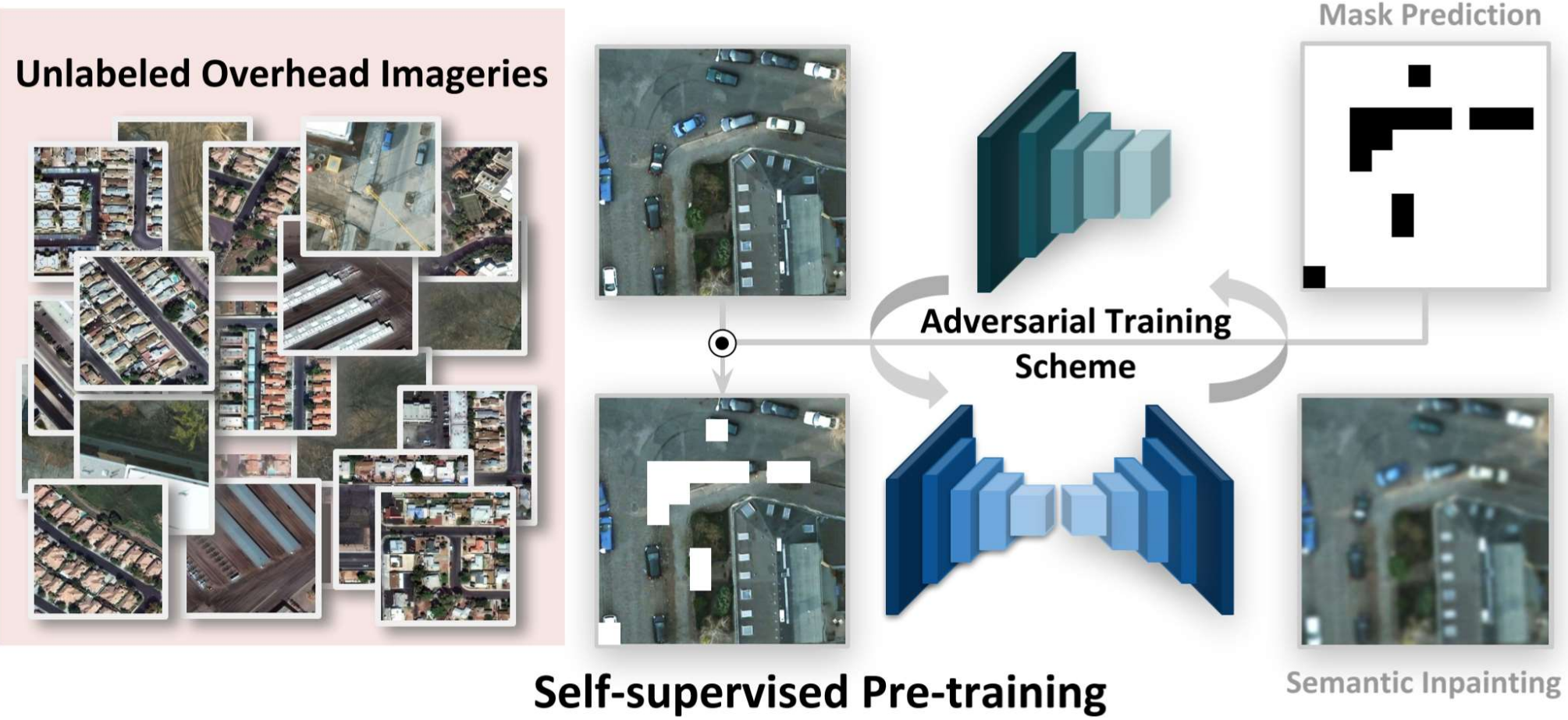
SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

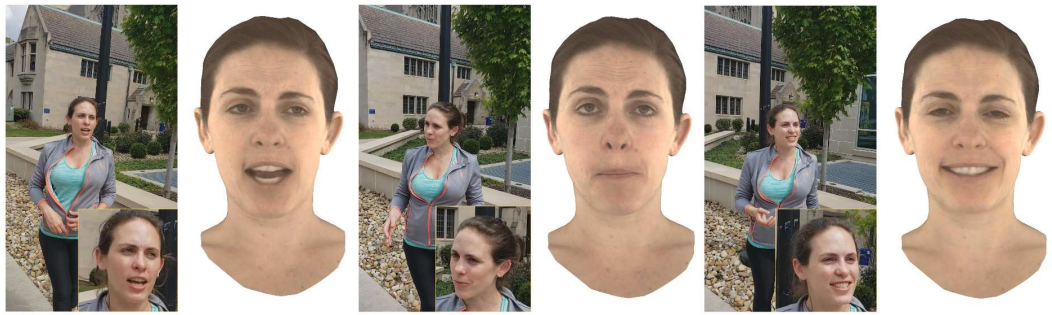
[Suriya, 2018]


Unlabeled Overhead Imageries
Mask Prediction
Adversarial Training Scheme
Semantic Inpainting
Self-supervised Pre-training

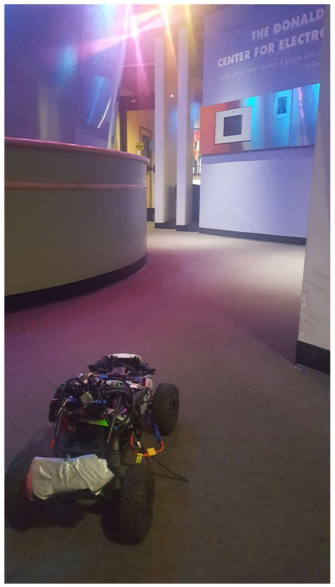
[Suriya, 2018]

Method	Encoder	Bottleneck	Decoder	Results			
				Potsdam	SpaceNet	DG Roads	DG Lands
Context Prediction [8]	AlexNet	✗	✗	0.273	0.593	0.478	0.257
Context Encoders [36]		✓	✗	0.298	0.610	0.478	0.339
Splitbrain AE [45]		✓	✗	0.265	0.641	0.482	0.411
ImageNet	ResNet-18	✗	✗	0.493	0.701	0.669	**0.575**
Scratch	ResNet-18	✗	✗	0.414	0.657	0.643	0.495
Scratch		✗	✓	0.418	0.661	0.607	0.507
Autoencoder		✓	✓	0.502	0.748	0.749	0.515
Autoencoder		✗	✓	0.499	0.742	0.742	0.499
Context Encoders (Ours)	ResNet-18	✓	✗	0.540	0.730	0.478	0.501
		✗	✓	0.562	0.762	0.759	0.503
Coach Mask (Ours)	ResNet-18	✗	✓	**0.568**	**0.770**	**0.768**	0.529

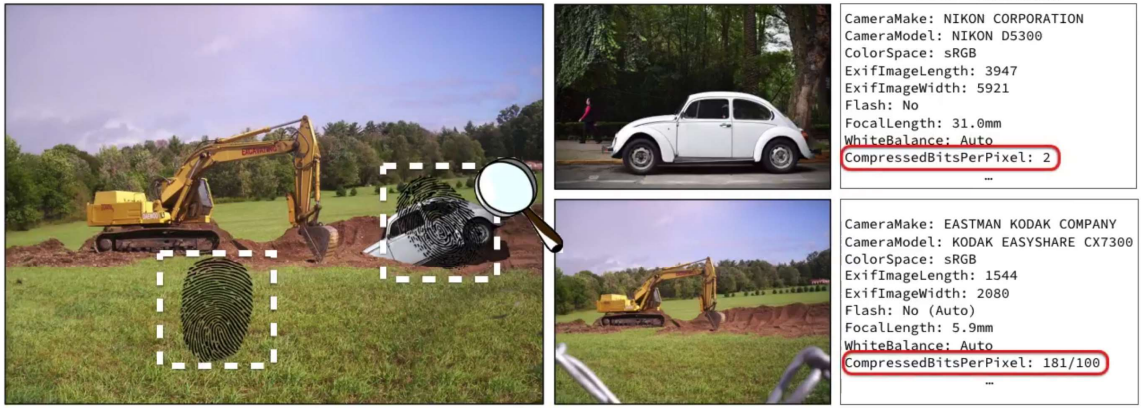
3D Facial Performance Rendering [Yoon, 2019]

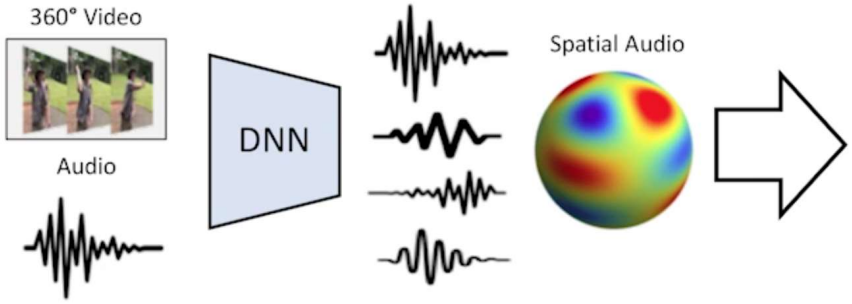


Robot Navigation
[Kahn, 2018]



Manipulation
Detection
[Huh, 2018]



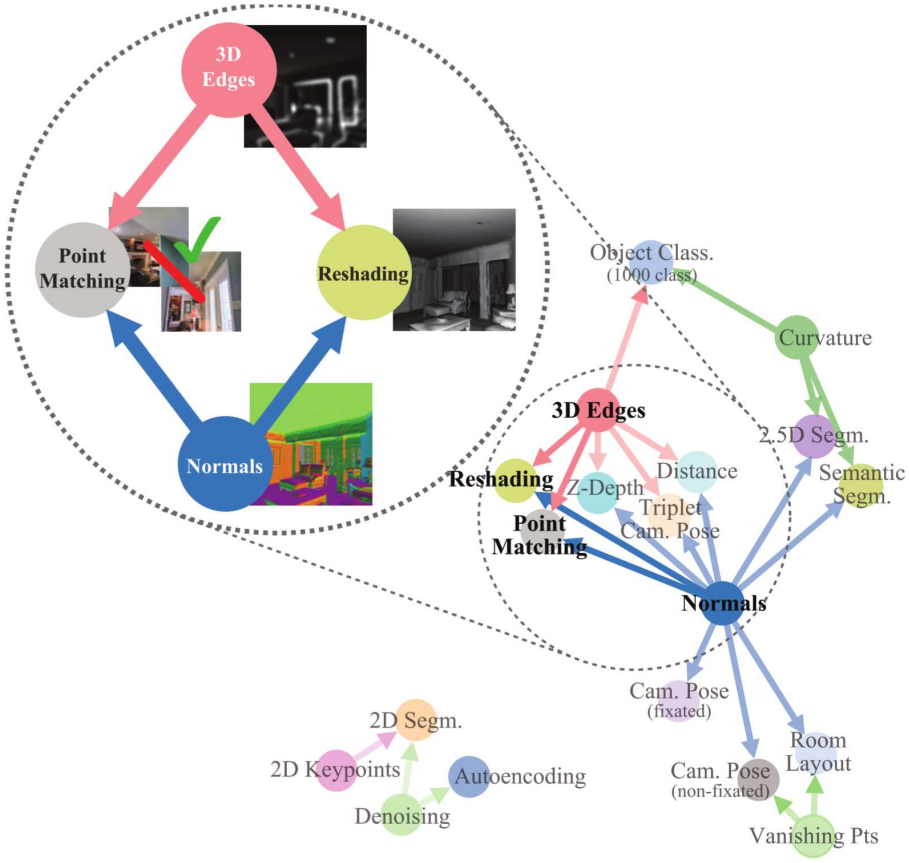


360’ Camera +
Spatial Audio
[Morgado, 2018]

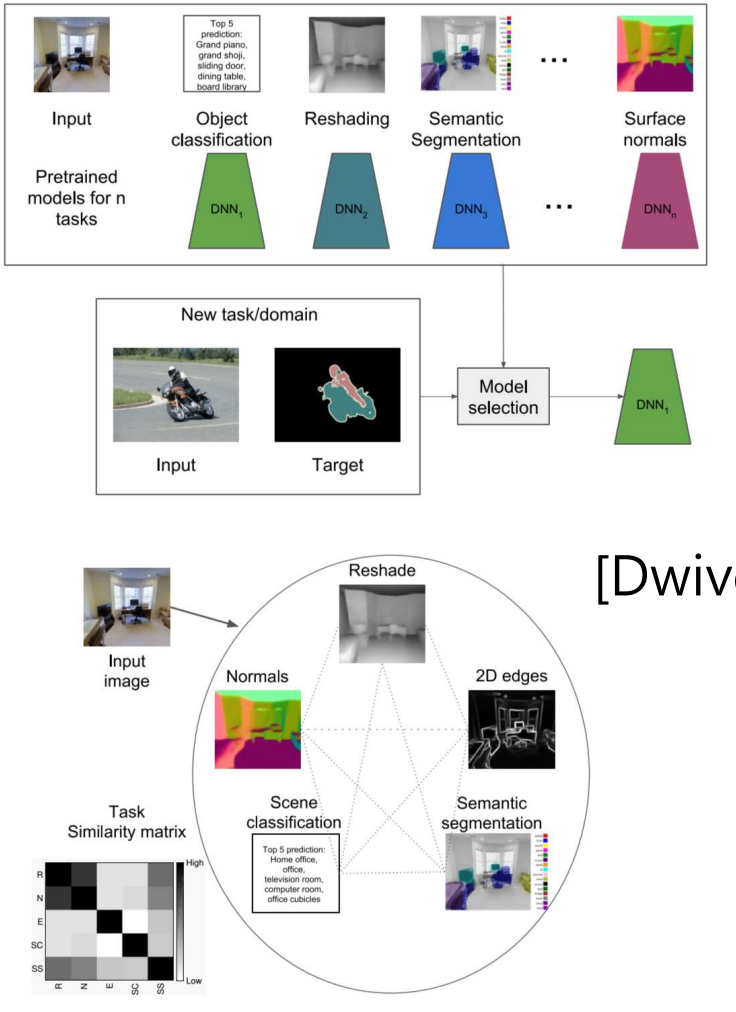
SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

Transfer Learning: Pretext Task <-> Downstream Task



[Zamir, 2018]



[Dwivedi, 2019]

SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

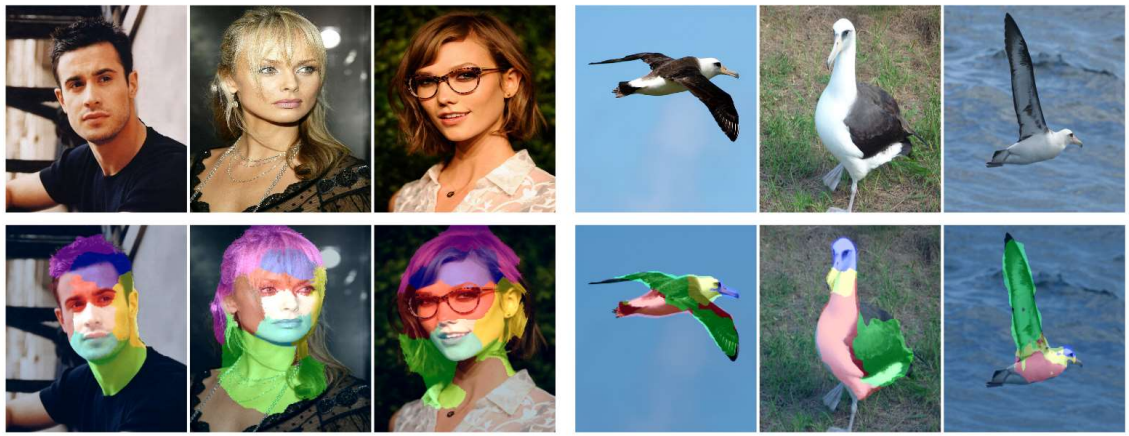
SSL의 적절한 이용 방법은 무엇인가?

- [Doersch, 2017]
  - "다양한 SSL 방법론을 섞어서 적절한 구조로 Multi-task로 학습하면 더 잘 된다"
- [Kolesnikov, 2019]
  - "같은 SSL이라도 모델 아키텍처 영향을 되게 많이 탄다. 일반적으로 잘 통용되는 아키텍처가 SSL에서도 잘 된다는 보장은 없다."
  - "지금의 Linear Separability에 의한 SSL 벤치마크는 실험 세팅에 영향을 많이 받는 면이 있다."
- [Goyal, 2019]
  - "지금 있는 SSL 벤치마크들은 SSL에 사용되는 데이터 크기도 부족하고, 모델 크기도 부족하고, SSL에서 사용되는 Pretext Task 자체도 지나치게 쉬운 것 같다."
  - "만약 데이터 크기, 모델 크기, Pretext Task의 난이도 자체를 키울 수 있다면, SSL이 훨씬 효과적으로 동작할 것이다."

SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

"Tracking Emerges by Colorizing"
[Vondrick, 2018]



Co-Part Segmentation by SSL
[Hung, 2019]

SSL, 어디로 가야 하나?

- 주 도메인에서의 더 좋은 SSL 방법론 고안 (너무나 당연…)
  - 일단은 ImageNet-pretrained 모델부터 가볍게 뛰어넘어야 쓸 만 할 텐데…
- 비교적 마이너한 도메인에서 동작하는 SSL 방법론 고안
  - 주류 도메인보다 더 data-hungry하지만 기존의 방법론들이 잘 안 먹힐 수도.
- Pretext Task와 Downstream Task 간의 관계를 더 정확하게 규정
  - == 'Transfer Learning' 자체를 더 잘 알아야 한다.
- "우리는 제대로 SSL을 하고 있는가? 굳이 해야 하는가?"
- 단순히 특징의 Transferability를 이용하는 게 아니라, 직접 다른 Task를 푸는데 이용할 순 없는가?
- SSL이 ML에서의 다른 근본적인 문제들을 더 잘 풀 수 있도록 도와줄 수 있지 않을까?

다른 근간적인 ML 문제의 해결을 돕는 도구로서의 SSL

- GAN: [Lucic, 2019], [Chen, 2019]
- Semi-supervised Learning: [Zhai, 2019]
- Clustering: [Zhang, 2019]
- Anomaly Detection: [Golan, 2019]
- Robustness & Uncertainty: [Hendrycks, 2019]
- …

05 결론(과 하고 싶은 말)

정리,

- 사전 학습과 전이 학습은 표현 기반 학습에서 감각적으로 느껴지는 것보다 훨씬 오래되고 중요한 주제
  - 이런 맥락에서 SSL은 단순히 유행이 아니라, 나름 핵심적인 연구 주제라고 생각합니다.
- Self-supervised Learning이란
  - 도메인의 구조나 데이터 획득 과정의 특성을 이용하여,
  - 특정 테스크를 정의해 라벨을 어거지로 만들고,
  - 이를 통해 도메인에 관한 표현 학습을 시키는 것
- 최근 5년간 (특히, 3년간) SSL 논문이 정말 사방에서 쏟아지고 있습니다.
- 가야할 길도, 가봐야 할 길도 아직 많다

마지막으로 하고 싶은 말,

- **'아, SSL이라는 게 있구나! 내 도메인도 데이터 부족하고 성능 잘 안 나오는데 바로 적용해볼까.'**
  - SSL을 도입하기 전에, 무조건 우선 관련 도메인에서의 가장 많이 쓰이는 Pre-trained Model을 끌어다 쓰고 벤치마크를 한 번 찍어보는 것을 권장 드립니다.
    - ✓E.g) 이미지 -> ImageNet, NLP -> BERT, GPT, XLNet…
  - 적어도 Natural Image에서는 아직까지 ImageNet-pretrained Model만큼 사용하기 쉬우면서 성능이 나오는 SSL 방법론은 없습니다. 배보다 배꼽?
- **'내 도메인은 좀 특이하다. 그러면서 항상 데이터가 부족하다. SSL 한 번 해 볼 만 할 것 같은데?' or 'SSL 관련 연구를 해보고 싶다'**
  - 환영합니다. 이거 완전 힙하고 재밌어요. ~~내가 연구하고 있는 걸 다른 곳에서 항상 먼저 낸다는 거 빼구요.~~
- **부족한 발표임에도 경청해 주셔서 감사합니다.**

References (Section 1)

- [Hinton, 2006a] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.
- [Hinton, 2006b] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." science 313.5786 (2006): 504-507.
- [Bengio, 2007] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks." NIPS. 2007.
- [Glorot, 2010] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." AISTATS. 2010.
- [Glorot, 2011] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks." AISTATS. 2011.
- [Duchi, 2011] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." JMLR 12.Jul (2011): 2121-2159.
- [Hinton, 2012] Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent." (2012).
- [Srivastava, 2014] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." JMLR 15.1 (2014): 1929-1958.
- [Kingma, 2015] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." ICLR. .2015.
- [Zhang, 2017] Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." ICLR. 2017.
- [Arora, 2018] Arora, Sanjeev, Nadav Cohen, and Elad Hazan. "On the optimization of deep networks: Implicit acceleration by overparameterization." ICML. 2018.
- [Shen, 2018] Shen, Hao. "Towards a mathematical understanding of the difficulty in learning with feedforward neural networks." CVPR. 2018.
- [Simon, 2019a] "Gradient descent provably optimizes over-parameterized neural networks." ICLR. 2019.
- [Simon, 2019b] Du, Simon S., et al. "Gradient descent finds global minima of deep neural networks." ICML 2019.

References (Section 2)

- [Pratt, 91] Pratt, Lorien Y., et al. "Direct Transfer of Learned Information Among Neural Networks." AAAI. Vol. 91. 1991.
- [Pratt, 93] Pratt, Lorien Y. "Discriminability-based transfer between neural networks." NIPS. 1993.
- [Caruana, 95] Caruana, Rich. "Learning many related tasks at the same time with backpropagation." NIPS. 1995.
- [Thrun, 96] Thrun, Sebastian. "Is learning the n-th thing any easier than learning the first?." NIPS. 1996.
- [Pan, 09] Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE TKDE 22.10 (2009): 1345-1359.
- [Bengio, 12] Bengio, Yoshua. "Deep learning of representations for unsupervised and transfer learning." ICML Workshop on Unsupervised and Transfer Learning. 2012.

References (Section 2)

- [Zamir, 2018] Zamir, Amir R., et al. "Taskonomy: Disentangling task transfer learning." CVPR. 2018.
- [He, 2018] He, Kaiming, Ross Girshick, and Piotr Dollár. "Rethinking imagenet pre-training." arXiv preprint arXiv:1811.08883 (2018).
- [Hendrycks, 2019] Hendrycks, Dan, Kimin Lee, and Mantas Mazeika. "Using Pre-Training Can Improve Model Robustness and Uncertainty." ICML (2019).
- [Li, 2019] Li , Hengdu, Bharat Singh, Mahyar Najibi, Zuxuan Wu, Larry S. Davis, “An Analysis of Pre-Training on Object Detection.” arXiv: 1904.05871. 2019.
- [Chen, 2019] Chen, Wei-Yu, et al. "A closer look at few-shot classification." ICLR (2019).
- [Kornblith, 2019] Kornblith, Simon, Jonathon Shlens, and Quoc V. Le. "Do better imagenet models transfer better?." CVPR. 2019.
- [Paghu, 2019] Raghu, Maithra, et al. "Transfusion: Understanding transfer learning with applications to medical imaging." arXiv preprint arXiv:1902.07208 (2019).

References (Section 3)

- [Doersch, 2015] Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." ICCV. 2015.
- [Mikolov, 2013a] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." ICLR Workshop. 2013.
- [Mikolov, 2013b] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." NIPS. 2013.
- [Doersch, 2015] Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." CVPR. 2015.
- [Krähenbühl, 2016] Krähenbühl, Philipp, et al. "Data-dependent initializations of convolutional neural networks." ICLR. 2016.

References (Section 4 - 이미지)

- [Noroozi, 2016] Noroozi, Mehdi, and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles." ECCV. Springer, Cham, 2016.
- [Santa, 2017] Santa Cruz, Rodrigo, et al. "Deeppermnet: Visual permutation learning." CVPR. 2017.
- [Oord, 2018] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).
- [Hénaff, 2019] Hénaff, Olivier J., et al. "Data-efficient image recognition with contrastive predictive coding." ICML Workshop. 2019.
- [Trinh, 2019] Trinh, H. Trieu, Minh-Thang Luong, Quoc V. Le. "Selfie: Self-supervised Pretraining for Image Embedding." arXiv:1906.02940. 2019.
- [Pathak, 2016] Pathak, Deepak, et al. "Context encoders: Feature learning by inpainting." CVPR. 2016.
- [Larsson, 2017] Larsson, Gustav, Michael Maire, and Gregory Shakhnarovich. "Colorization as a proxy task for visual understanding." CVPR. 2017.
- [Zhang, 2017] Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Split-brain autoencoders: Unsupervised learning by cross-channel prediction." CVPR. 2017.
- [Noroozi, 2017] Noroozi, Mehdi, Hamed Pirsiavash, and Paolo Favaro. "Representation learning by learning to count." CVPR. 2017.
- [Liu, 2019] Liu, Xialei, Joost Van De Weijer, and Andrew D. Bagdanov. "Exploiting Unlabeled Data in CNNs by Self-supervised Learning to Rank." TPAMI. 2019.
- [Gidaris, 2018] Gidaris, Spyros, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations." ICLR. 2018.
- [Zhang, 2019] Zhang, Liheng, et al. "Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data." CVPR. 2019.

References (Section 4 - 비디오)

- [Srivastava, 2015] Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhudinov. "Unsupervised learning of video representations using lstms." ICML. 2015.
- [Misra, 2018] Misra, Ishan, C. Lawrence Zitnick, and Martial Hebert. "Shuffle and learn: unsupervised learning using temporal order verification." ECCV. Springer, Cham, 2016.
- [Sermanet, 2018] Sermanet, Pierre, et al. "Time-contrastive networks: Self-supervised learning from video." ICRA. 2018.
- [Agrawal, 2015] Agrawal, Pulkit, Joao Carreira, and Jitendra Malik. "Learning to see by moving." ICCV. 2015.
- [Li, 2016] Li, Yin, et al. "Unsupervised learning of edges." CVPR. 2016.
- [Pathak, 2017] Pathak, Deepak, et al. "Learning features by watching objects move." CVPR. 2017.
- [Luo, 2017] Luo, Zelun, et al. "Unsupervised learning of long-term motion dynamics for videos." CVPR. 2017.
- [Jiang, 2018] Jiang, Huaizu, et al. "Self-supervised relative depth learning for urban scene understanding." ECCV. 2018.

References (Section 4 - 오디오 & 텍스트)

- [Owens, 2018] Owens, Andrew, and Alexei A. Efros. "Audio-visual scene analysis with self-supervised multisensory features."ECCV. 2018.
- [Korbar, 2018] Korbar, Bruno, Du Tran, and Lorenzo Torresani. "Cooperative learning of audio and video models from self-supervised synchronization." NIPS. 2018.

- [Devlin, 2019] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." ACL. 2019.
- [Radford, 2018] Radford, Alec, et al. "Improving language understanding by generative pre-training." URL: https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper.pdf (2018).
- [Yang, 2019] Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." arXiv: 1906.08237. 2019.

References (Section 5 - 1)

- [Suriya, 2018] Suriya, et al. “Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery.” BMVC. 2018.
- [Morgado, 2018] Morgado, Pedro, et al. "Self-Supervised Generation of Spatial Audio for 360 Video." NIPS. 2018.
- [Huh, 2018] Huh, Minyoung, et al. "Fighting fake news: Image splice detection via learned self-consistency." ECCV. 2018.
- [Kahn, 2018] Kahn, Gregory, et al. "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation." ICRA. 2018.
- [Yoon, 2019] Self-Supervised Adaptation of High-Fidelity Face Models for Monocular Performance Tracking. CVPR. 2019.
- [Zamir, 2018] Zamir, Amir R., et al. "Taskonomy: Disentangling task transfer learning." CVPR. 2018.
- [Dwivedi, 2019] Dwivedi, Kshitij, and Gemma Roig. "Representation Similarity Analysis for Efficient Task taxonomy & Transfer Learning." CVPR. 2019.

References (Section 5 - 2)

- [Doersch, 2017] Doersch, Carl, and Andrew Zisserman. "Multi-task self-supervised visual learning." ICCV. 2017.
- [Kolesnikov, 2019] Kolesnikov, Alexander, Xiaohua Zhai, and Lucas Beyer. "Revisiting Self-Supervised Visual Representation Learning." ICML Workshop (2019).
- [Goyal, 2019] Goyal, Priya, et al. "Scaling and benchmarking self-supervised visual representation learning." arXiv preprint arXiv:1905.01235 (2019).
- [Vondrick, 2018] Vondrick, Carl, et al. "Tracking emerges by colorizing videos." ECCV. 2018.
- [Hung, 2019] Hung, Wei-Chih, et al. " SCOPS: Self-Supervised Co-Part Segmentation." CVPR. 2019.

- [Lucic, 2019] Lucic, Mario, et al. "High-Fidelity Image Generation With Fewer Labels." ICML. 2019.
- [Chen, 2019] Chen, Ting, et al. " Self-Supervised GANs via Auxiliary Rotation Loss." CVPR. 2019.
- [Zhai, 2019] Zhai, Xiaohua, et al. "S^4L: Self-Supervised Semi-Supervised Learning." ICML Workshop. 2019.
- [Zhang, 2019] Zhang, Junjian, et al. “Self-Supervised Convolutional Subspace Clustering Network.” CVPR. 2019.
- [Golan, 2019] Golan, Izhak, and Ran El-Yaniv. "Deep Anomaly Detection Using Geometric Transformations." NIPS. 2018.
- [Hendrycks, 2019] “Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty.” ICML Workshop. 2019.

End of Slides

- End of Slides