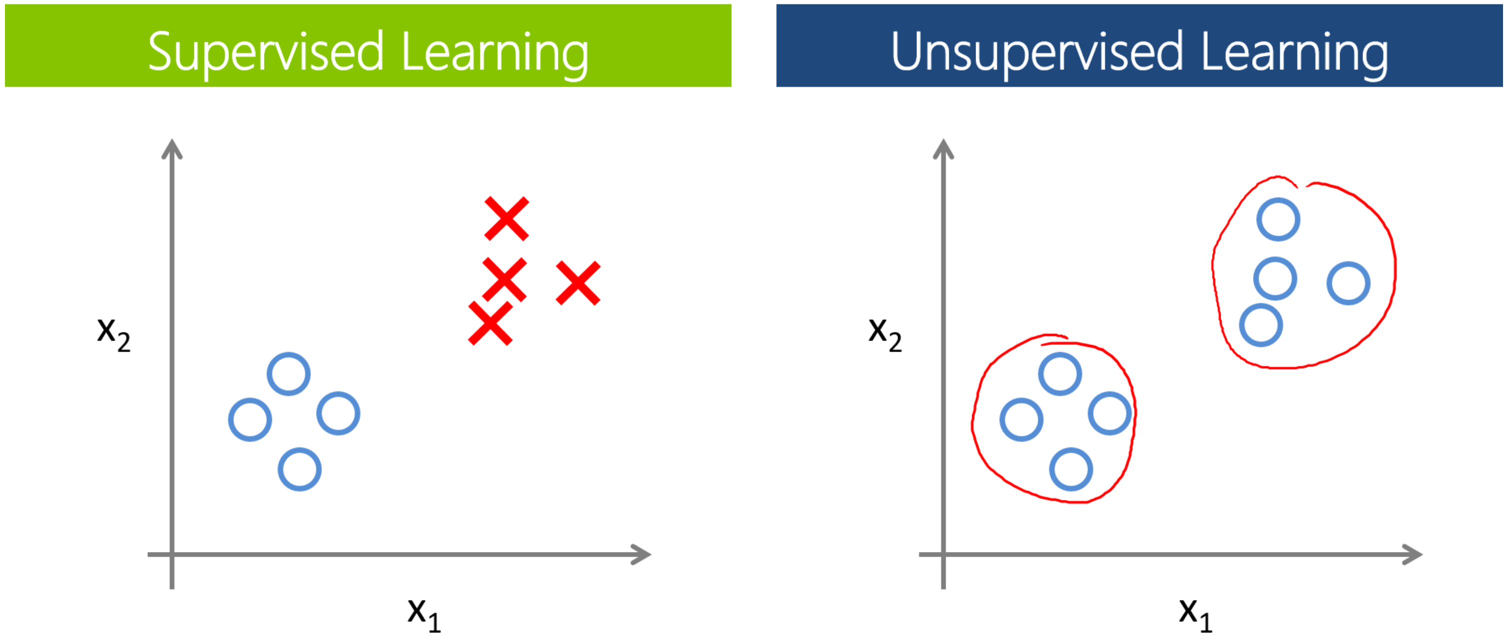
A Survey to Self-Supervised Learning

Naiyan Wang

Paradigm of Learning

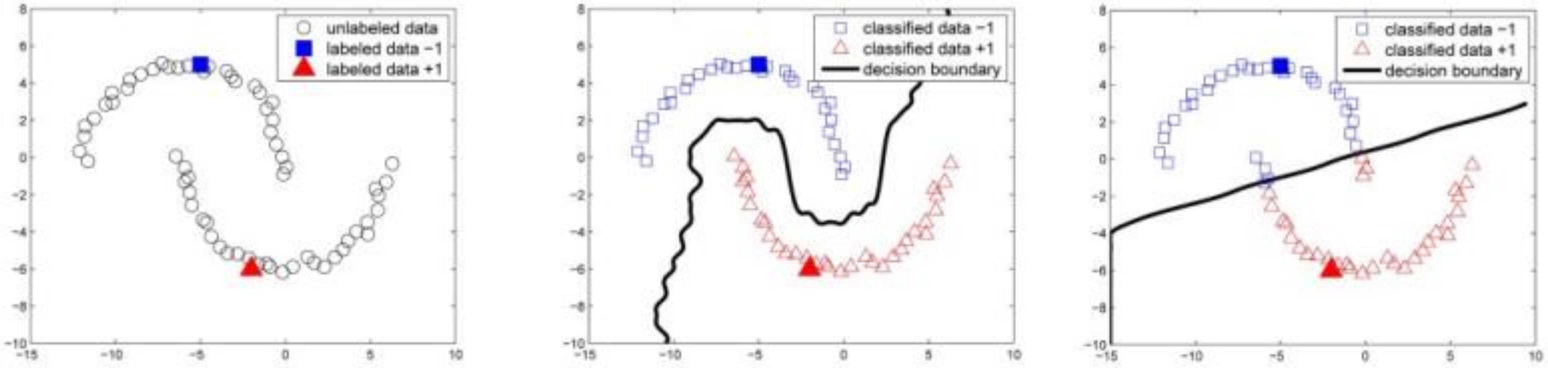
- Supervised Learning & Unsupervised Learning
  - Given desired output vs. No guidance at all



http://oliviaklose.azurewebsites.net/content/images/2015/02/2-supervised-vs-unsupervised-1.png

Paradigm of Learning

- In Between…
  - Semi-Supervised Learning
    - Mix labeled and unlabeled data

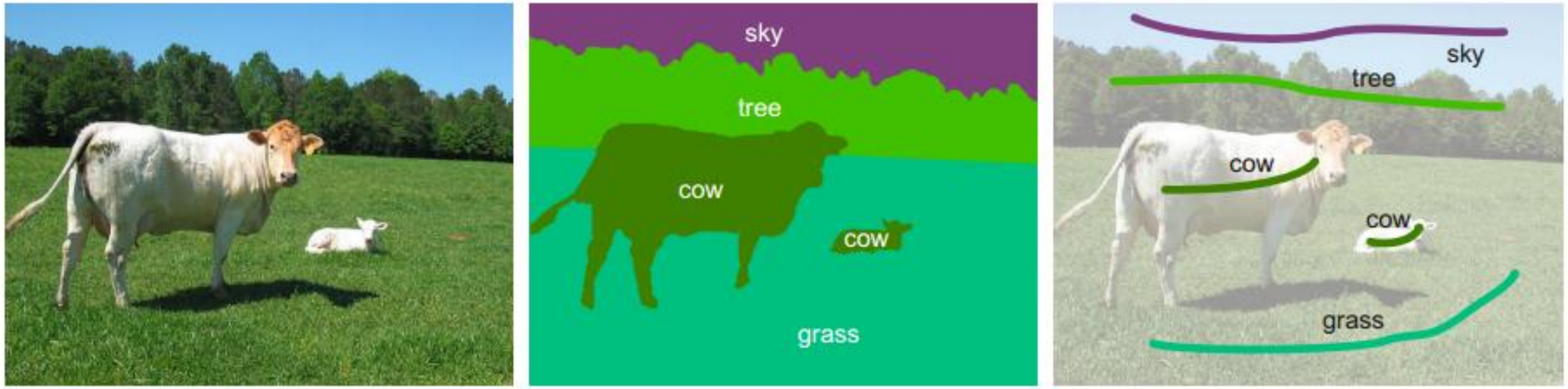


(a) (b) (c)

https://openi.nlm.nih.gov/imgs/512/371/4299091/PMC4299091_sensors-14-23871f4.png

Paradigm of Learning

- In Between…
  - Weakly-Supervised Learning
    - Use somewhat coarse or inaccurate supervision, e.g.
      - Given image level label, infer object level bounding box/ pixel level segmentation
      - Given video level label, infer image level label
      - Given scribble, infer the full pixel level segmentation
      - Given bounding box, infer the boundary of object



(a) image (b) mask annotation (c) scribble annotation

Lin, D., Dai, J., Jia, J., He, K., & Sun, J. (2016). Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR2016*.

Paradigm of Learning

- In Between…
  - Transfer Learning
    - Train on one problem, but test on a different but related problem, e.g.
      - Multi-Task learning
      - Train on one domain, test on another domain (possibly unlabeled)

A source image. Possible target images.

Yoo, D., Kim, N., Park, S., Paek, A. S., & Kweon, I. S. (2016, October). Pixel-level domain transfer. In *ECCV2016*

Paradigm of Learning

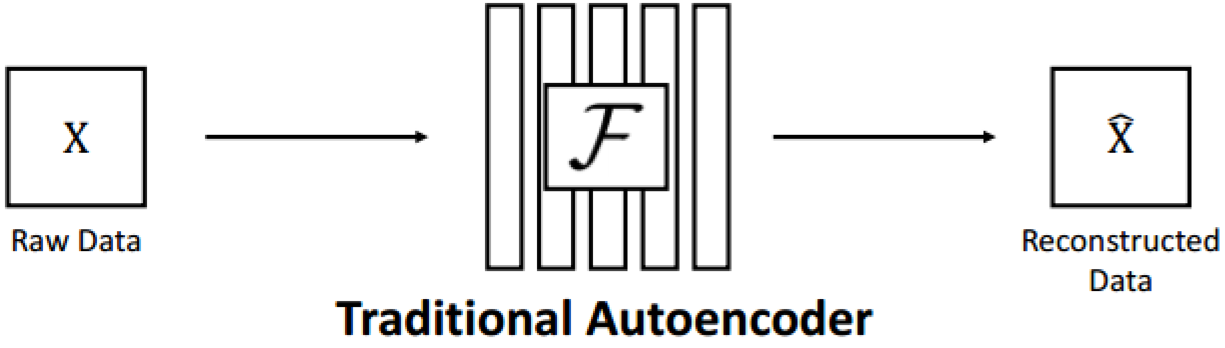
- More to mention…
  - Reinforcement Learning
  - Active Learning
  - Zero/One/Few-Shot Learning

Self-Supervised (Feature) Learning

- What is it?
  - Use naturally existed supervision signals for training.
  - (Almost) no human intervention
- Why do we need it?
  - The age of "representation learning"! (Pre-training – Fine-tune pipeline)
  - Self-Supervised learning can leverage self-labels for representation learning.
- How can we realize it?
  - That is in this talk!

Why not use construction?

- What is wrong with autoencoder?
  - Use pixel-wise loss, no structural loss incorporated
  - Reconstruction can hardly represent semantic information
- GAN may alleviate the first issue (e.g. BiGAN)

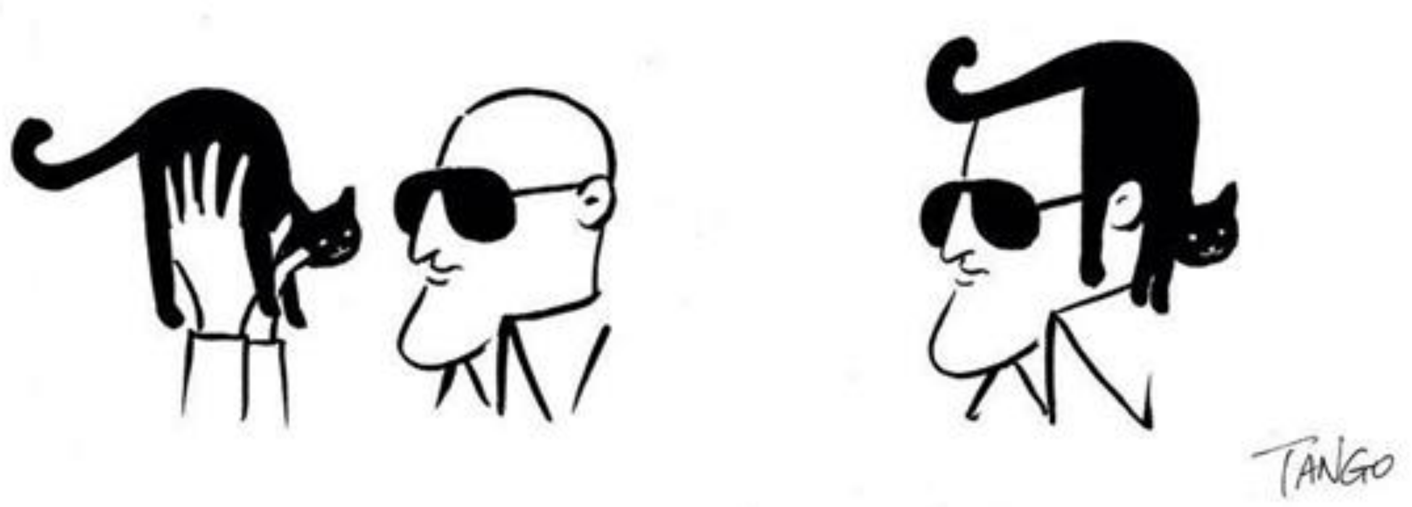


Outline

- Context
- Video
- Cross-Modality
- Exemplar Learning

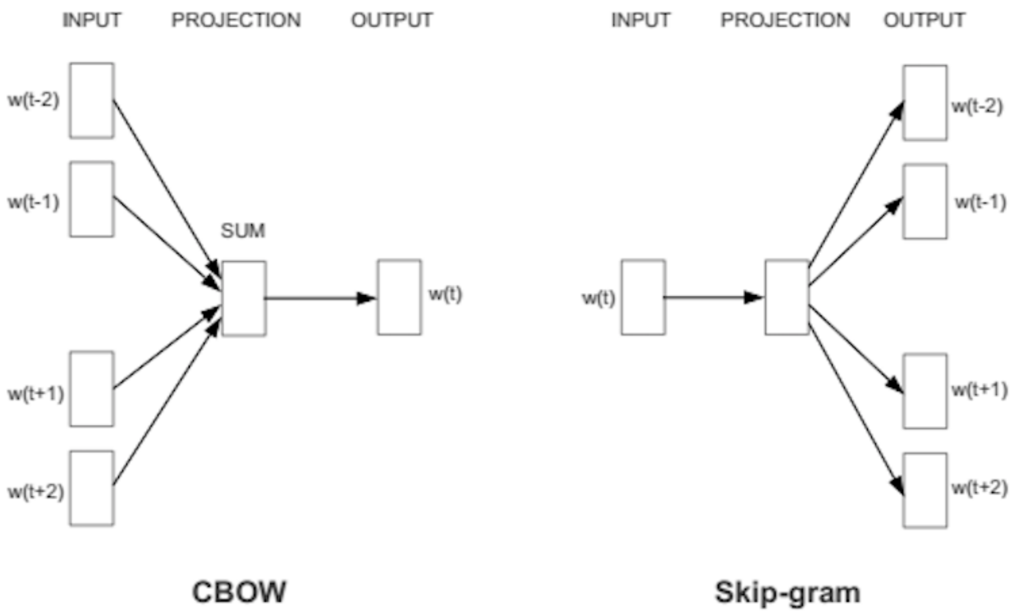
Context

- Context is ubiquitous in CV/NLP
  - 管中窥豹 & 断章取义
  - Cat or hair?
  - Beyond using it to improve performance, can you use it as supervision directly?



Context

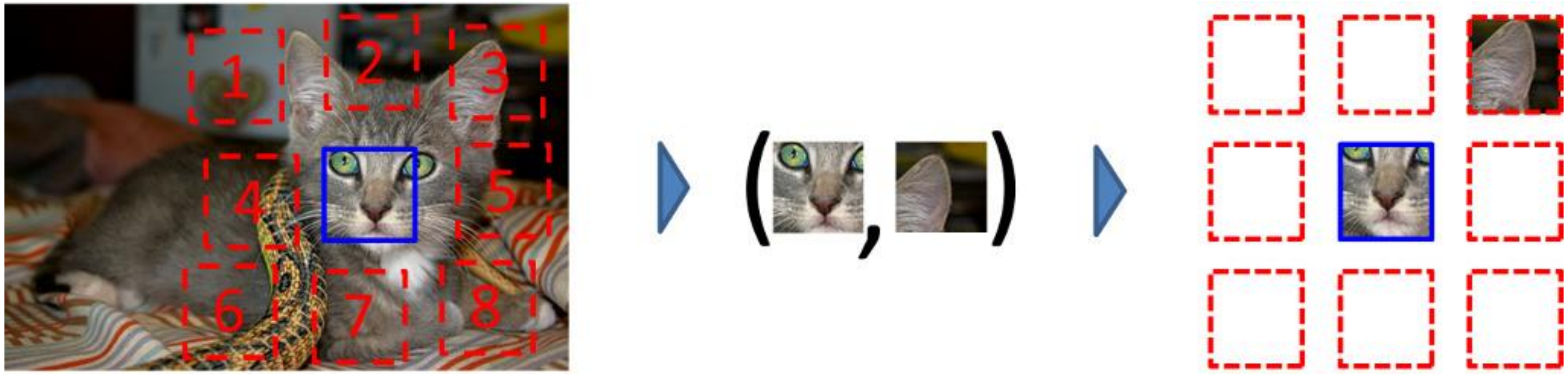
- Word2Vec: 1-dim context in NLP



https://deeplearning4j.org/img/word2vec_diagrams.png

Context

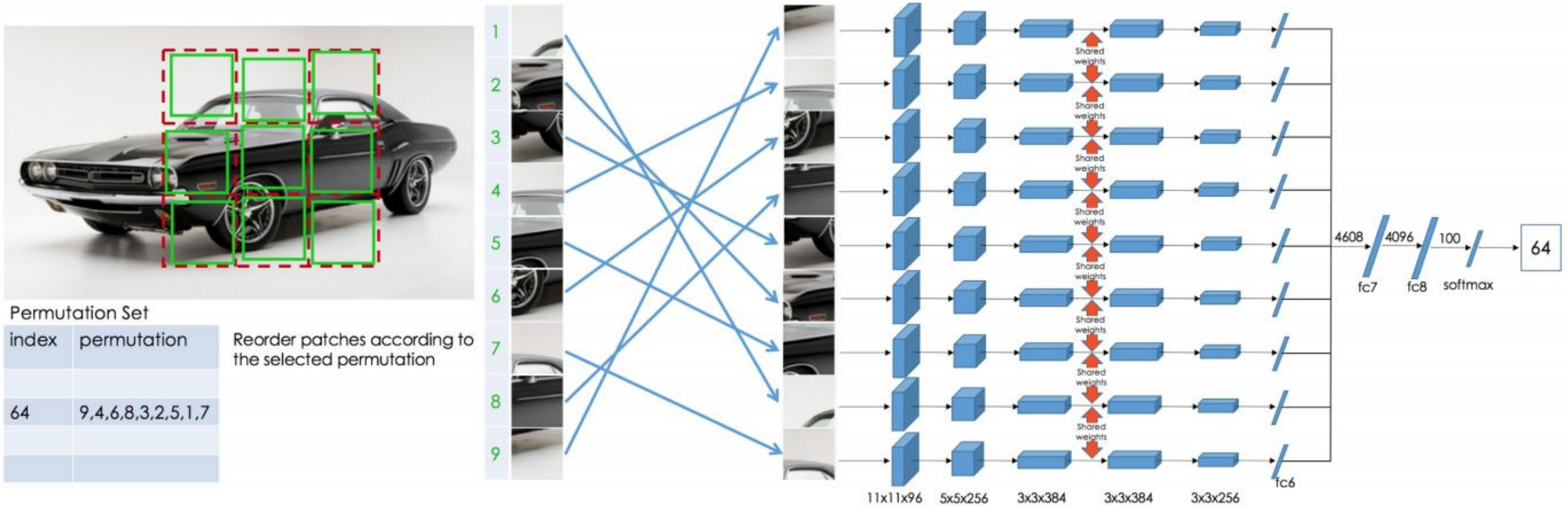
- Solving the Jigsaw
  - Predict relative positions of patches
  - You have to understand the object to solve this problem!
  - Be aware of trivial solution! CNN is especially good at it



Carl Doersch, Abhinav Gupta, and Alexei A. Efros. **Unsupervised Visual Representation Learning by Context Prediction.** In *ICCV 2015*

Context

- Solving the Jigsaw
  - Use stronger supervision, solve the real jigsaw problem
  - Harder problem, better performance



Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV 2016*.

Context

- Solving the Jigsaw
  - Visualization of filters

Noroozi, M., & Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV 2016*.

Context

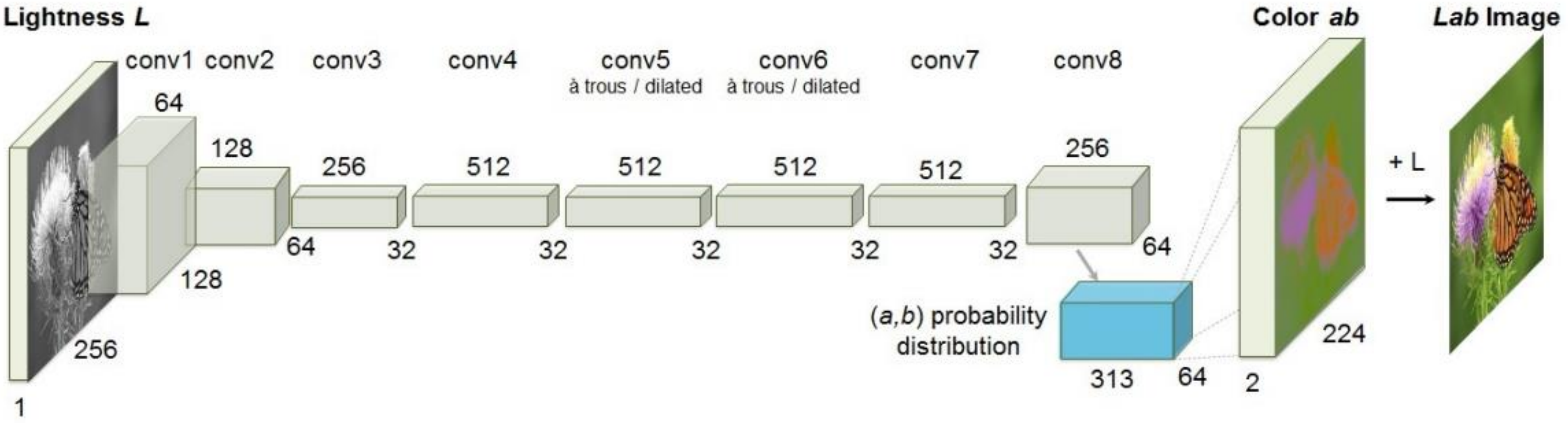
- Why not directly predict the missing parts?
  - With the advancement of adversarial loss

(a) Input context | (b) Human artist | (c) Context Encoder ($L2$ loss) | (d) Context Encoder ($L2$ + Adversarial loss)

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *CVPR 2016*.

Context

- Colorization
  - You have to know what the object is before you predict its color
  - E.g. Apple is red/green, sky is blue, etc.



Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. In *ECCV 2016*

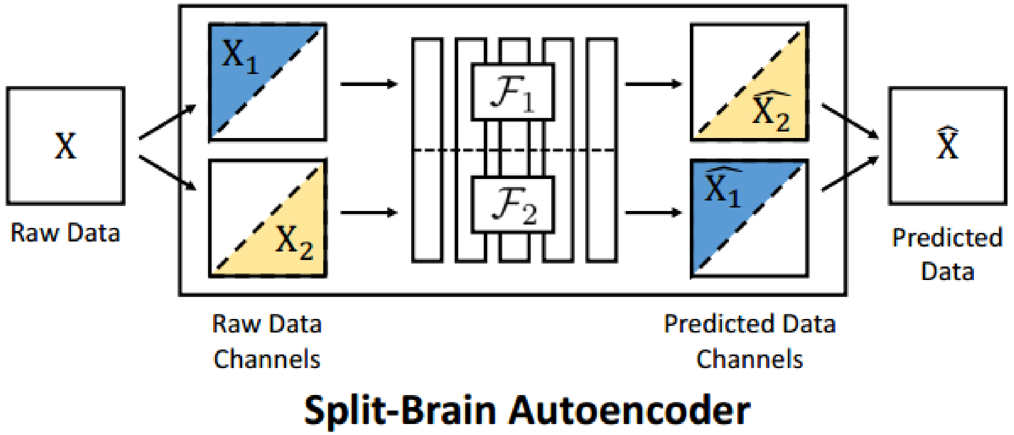
Context

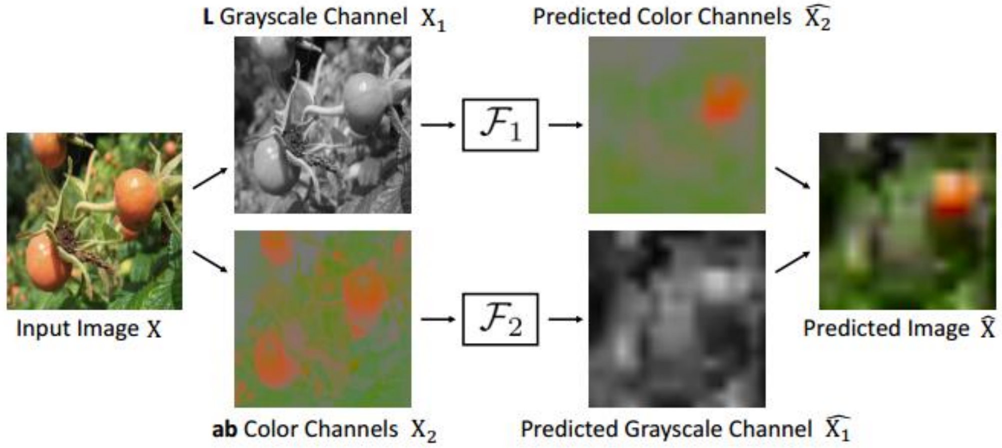
- Colorization
  - It is important how to interpret your work!
  - Example colorization of Ansel Adams's B&W photos

Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. In *ECCV 2016*

Context

- Colorization
  - Stronger supervision, cross-supervision of different parts of data





(a) ***Lab* Images**

Zhang, R., Isola, P., & Efros, A. A. Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction. *In CVPR 2017*

Video

- Video can provide rich information
  - Temporal continuity
  - Motion consistency
  - Action order

Video

- Slow feature
  - Neighborhood frames should have similar features

$$\mathcal{U}_2 = \{\langle (j,k), p_{jk} \rangle : \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathcal{U} \text{ and } p_{jk} = \mathbb{1}(0 \leq j - k \leq T)\},$$

$$R_2(\boldsymbol{\theta}, \mathcal{U}) = \sum_{(j,k) \in \mathcal{U}_2} D_\delta(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_k), p_{jk})$$

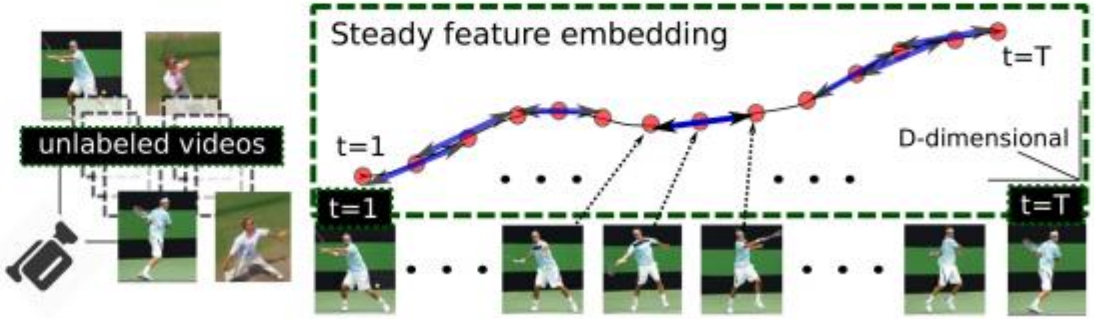
$$= \sum_{(j,k) \in \mathcal{U}_2} p_{jk}\, d(\mathbf{z}_{\boldsymbol{\theta}j}, \mathbf{z}_{\boldsymbol{\theta}k}) + \overline{p_{jk}}\, \max(\delta - d(\mathbf{z}_{\boldsymbol{\theta}j}, \mathbf{z}_{\boldsymbol{\theta}k}), 0),$$

Mobahi, H., Collobert, R., & Weston, J. Deep learning from temporal coherence in video. In *ICML 2009*.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, *14*(4), 715-770.

Video

- Slow and steady feature
  - Not only similar, but also smooth
  - Extend to triplet setting (Not triplet loss!)



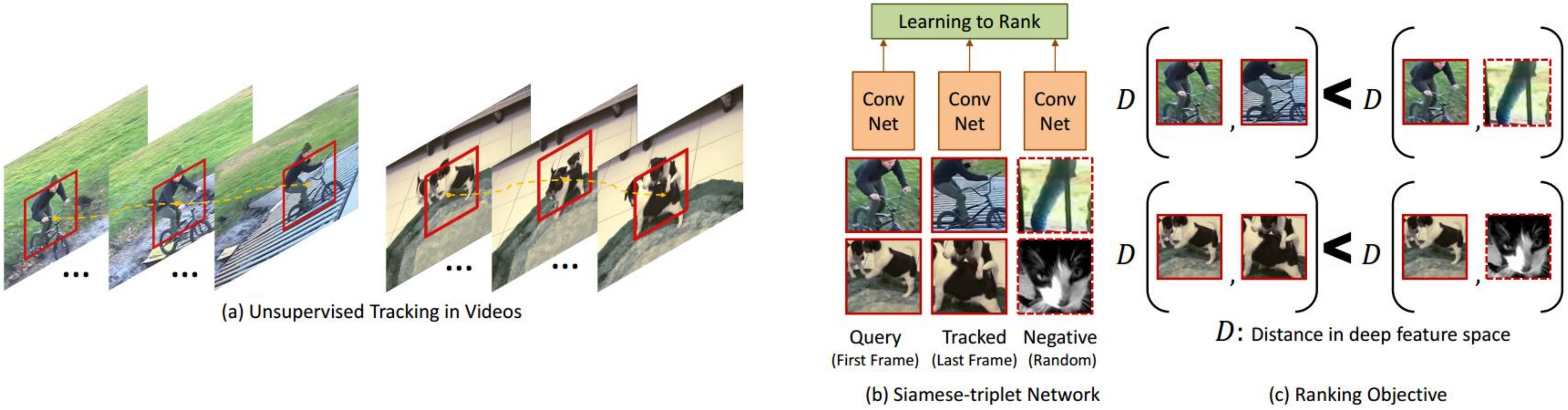
$$\mathcal{U}_3 = \{\langle (l,m,n), p_{lmn}\rangle : \boldsymbol{x}_l, \boldsymbol{x}_m, \boldsymbol{x}_n \in \mathcal{U} \text{ and } p_{lmn} = \mathbb{1}(0 \le m - l = n - m \le T)\}.$$

$$R_3(\boldsymbol{\theta},\mathcal{U}) = \sum_{(l,m,n)\in\mathcal{U}_3} D_\delta(\mathbf{z}_{\boldsymbol{\theta}l} - \mathbf{z}_{\boldsymbol{\theta}m},\ \mathbf{z}_{\boldsymbol{\theta}m} - \mathbf{z}_{\boldsymbol{\theta}n},\ p_{lmn}),$$

Jayaraman, D., & Grauman, K. Slow and steady feature analysis: higher order temporal coherence in video. In *CVPR 2016*.

Video

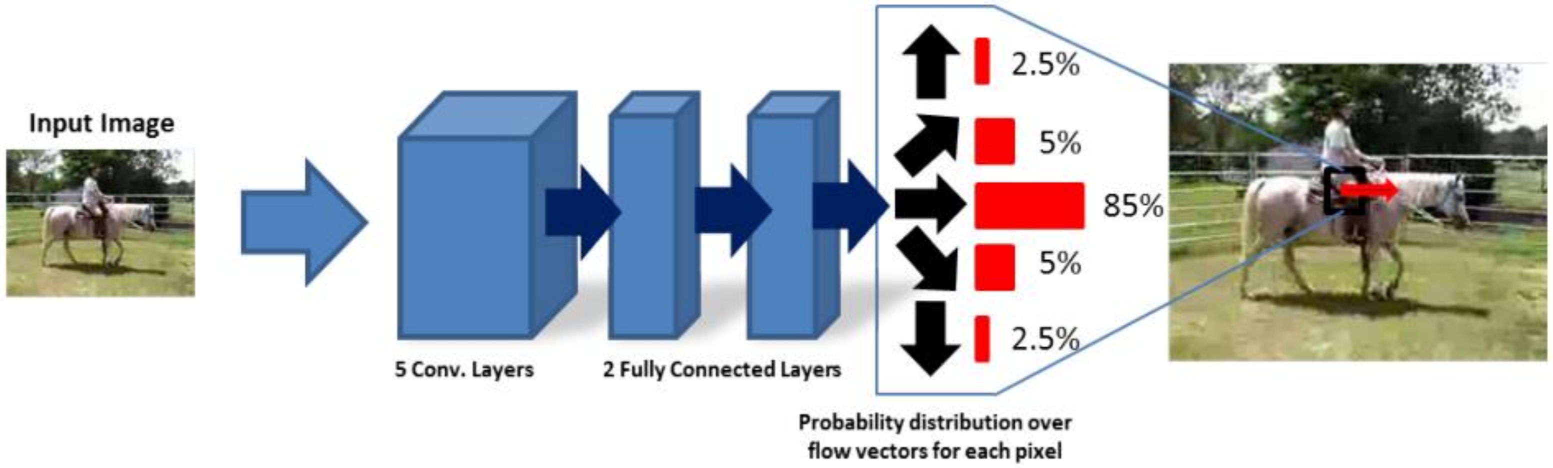
- Find corresponding pairs using visual tracking



Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *ICCV2015*

Video

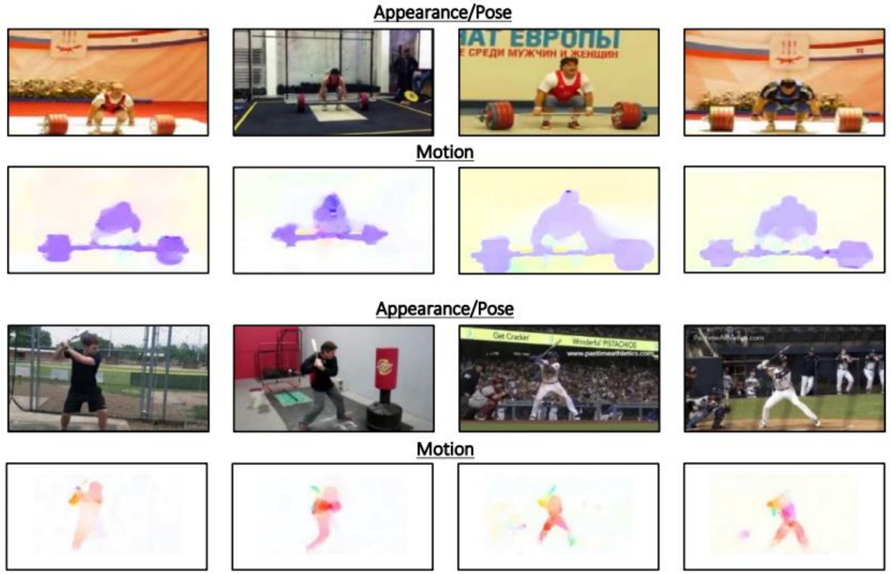
- Directly predict motion
  - Motion is not predictable by its nature
  - The ultimate goal is not to predict instance motion, but to learn common motion of visually similar objects

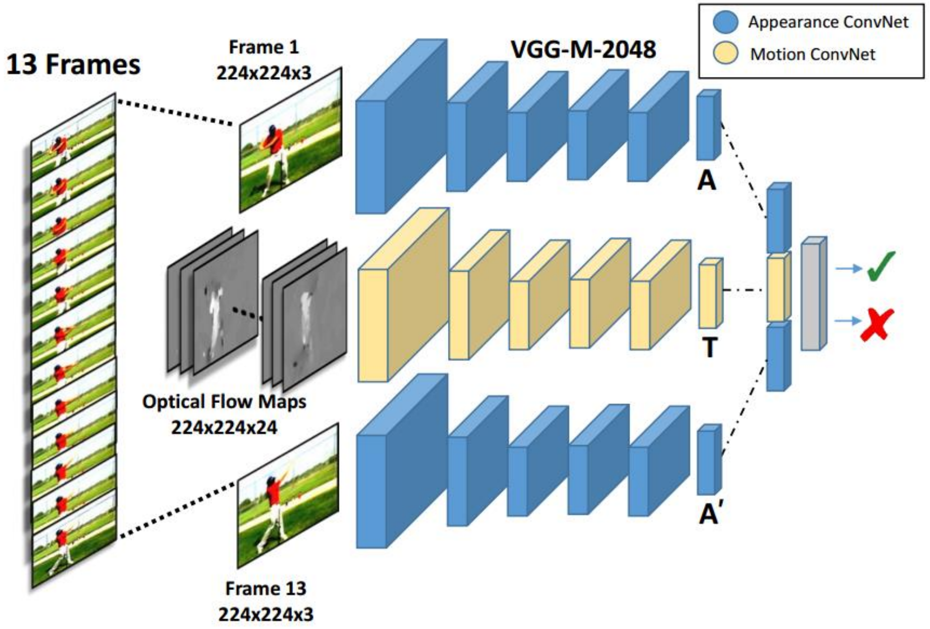


Walker, J., Gupta, A., & Hebert, M. Dense optical flow prediction from a static image. In *ICCV 2015*

Video

- Similar pose should have similar motion
  - Learning appearance transformation

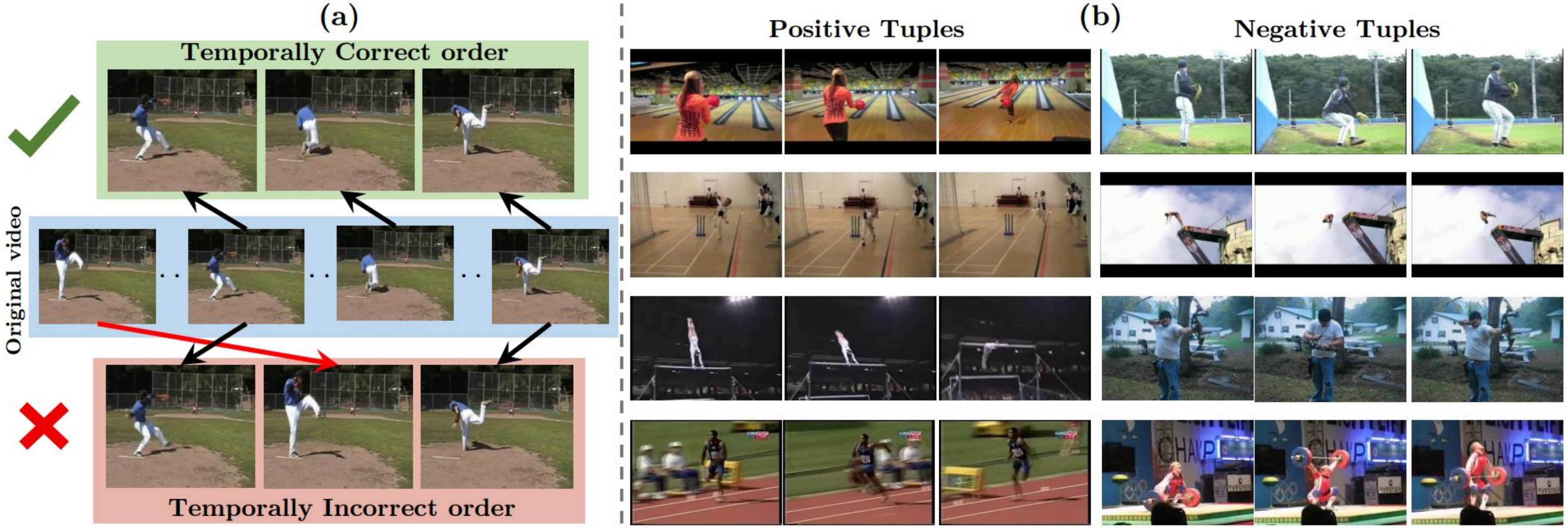




Purushwalkam, S., & Gupta, A. Pose from Action: Unsupervised Learning of Pose Features based on Motion. In *ECCVW 2016*.

Video

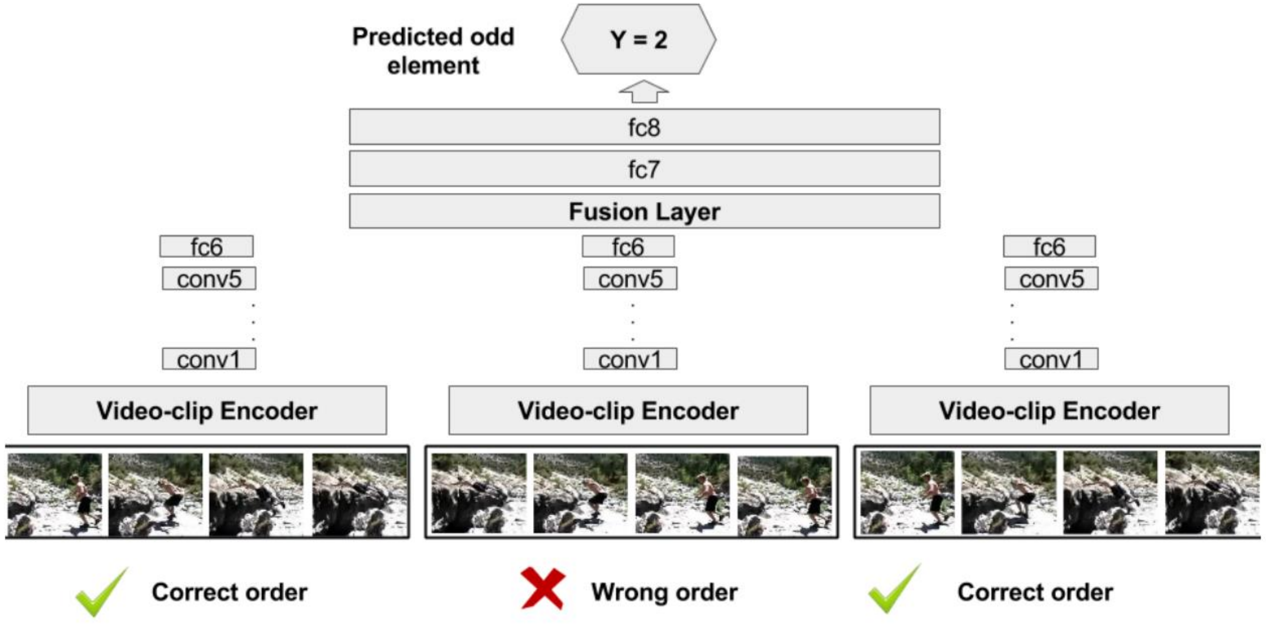
- Is the temporal order of a video correct?
  - Encode the cause and effect of action



Misra, I., Zitnick, C. L., & Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV 2016*.

Video

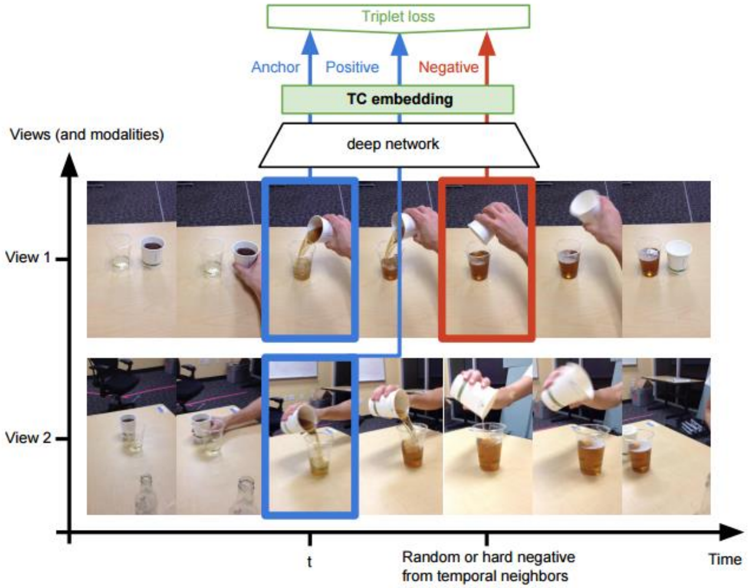
- Is the temporal order of a video correct?
  - Find the odd sequence

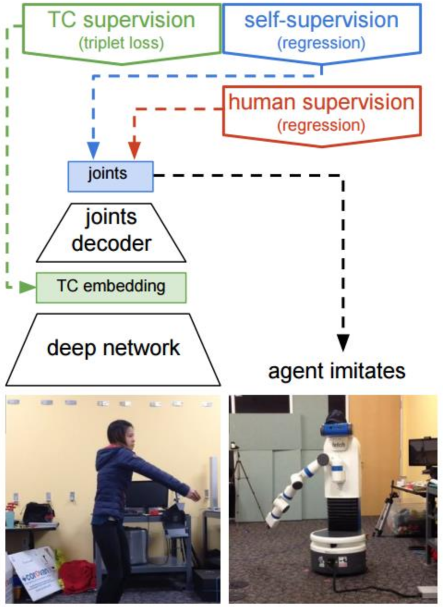


Fernando, B., Bilen, H., Gavves, E., & Gould, S. Self-Supervised Video Representation Learning With Odd-One-Out Networks. *In CVPR2017*.

Video

- Multi-view
  - Same action, but different view
  - View and pose invariant fetures

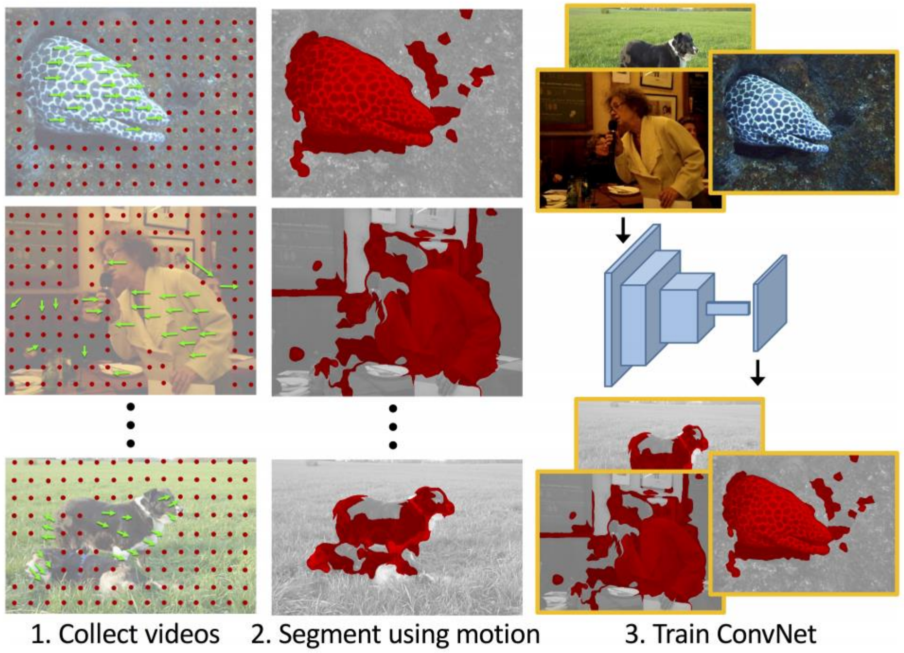




Sermanet, P., Lynch, C., Hsu, J., & Levine, S. Time-Contrastive Networks: Self-Supervised Learning from Multi-View Observation. *arXiv preprint arXiv:1704.06888*.

Video

- The world is rigid, or at least piecewise rigid
  - Motion provide evidence of how pixels move together
  - The pixels move together are likely to form an object



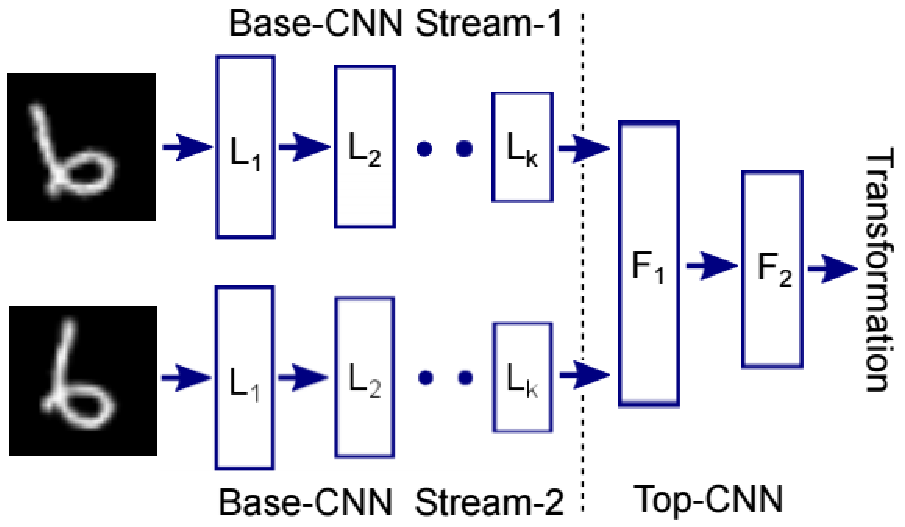
Pathak, D., Girshick, R., Dollár, P., Darrell, T., & Hariharan, B. Learning Features by Watching Objects Move. *In CVPR 2017*.

Cross-Modality

- In some applications, it is easy to collect and align the data from several modalities
  - Lidar & GPS/IMU & Camera
  - RGB & D
  - Image & Text
- How to utilize them for cross-supervision?

Cross-Modality

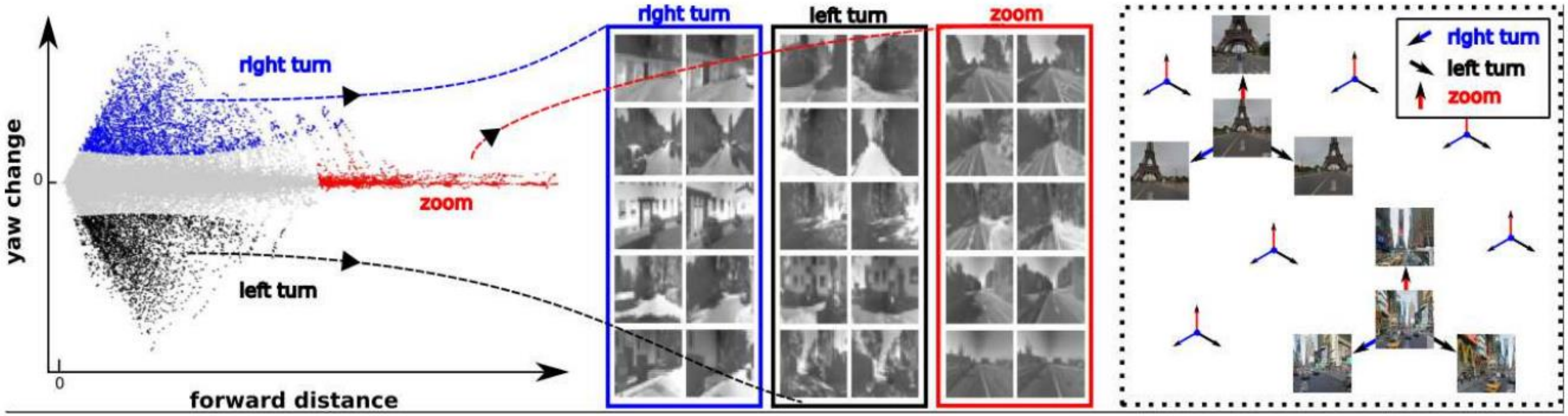
- Ego-motion
  - “We move in order to see and we see in order to move” - J.J Gibson
  - Ego-motion data is easy to collect
  - Siamese CNN to predict camera translation & Rotation along 3-axises. (Visual Odometry)



Agrawal, P., Carreira, J., & Malik, J. Learning to see by moving. In *ICCV 2015*

Cross-Modality

- Ego-motion
  - Learning features that are equivariant to ego-motion



Jayaraman, D., & Grauman, K. Learning image representations tied to ego-motion. In *ICCV 2015*

Cross-Modality

- Ego-motion
  - Siamese networks with contrastive loss
  - M_g is the transformation matrix specified by the external sensors

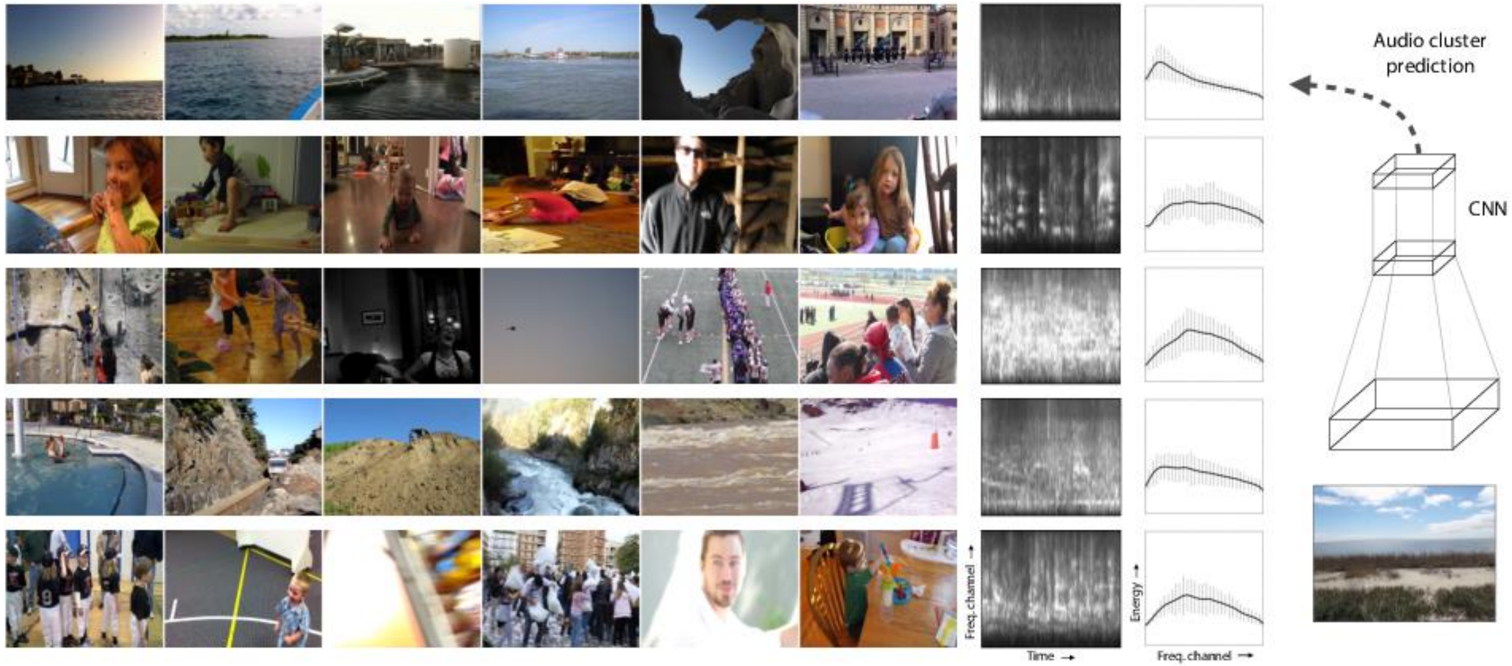
$$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \underset{\boldsymbol{\theta}, \mathcal{M}}{\arg\min} \sum_{g,i,j} d_g \left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), p_{ij} \right),$$

$$d_g(\boldsymbol{a}, \boldsymbol{b}, c) = \mathbb{1}(c = g) d(\boldsymbol{a}, \boldsymbol{b}) + \mathbb{1}(c \neq g) \max(\delta - d(\boldsymbol{a}, \boldsymbol{b}), 0),$$

Jayaraman, D., & Grauman, K. Learning image representations tied to ego-motion. In *ICCV 2015*

Cross-Modality

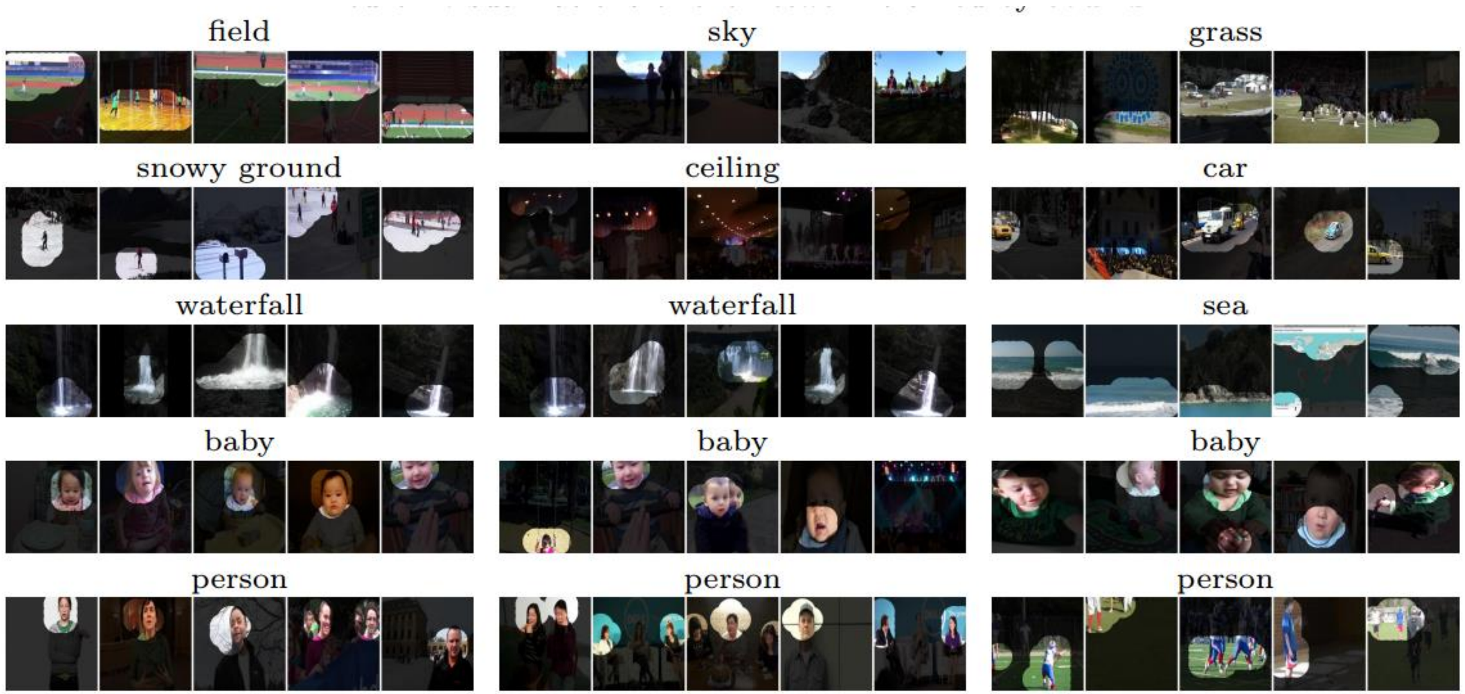
- Acoustics -> RGB
  - Similar events should have similar sound.
  - Naturally cluster the videos.



Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., & Torralba, A. Ambient sound provides supervision for visual learning. In *ECCV 2016*

Cross-Modality

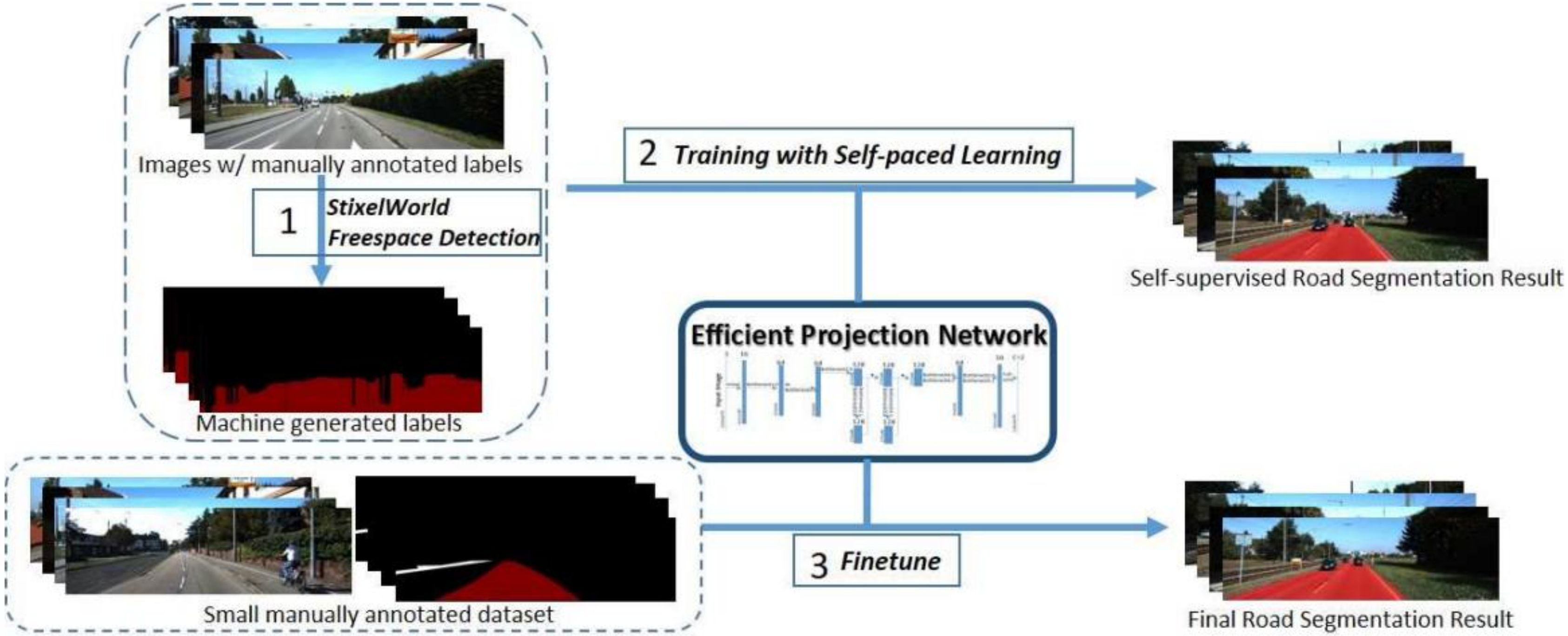
- Acoustics -> RGB
  - What does this CNN learn? Separation of baby and person :-D



Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., & Torralba, A. Ambient sound provides supervision for visual learning. In *ECCV 2016*

Cross-Modality

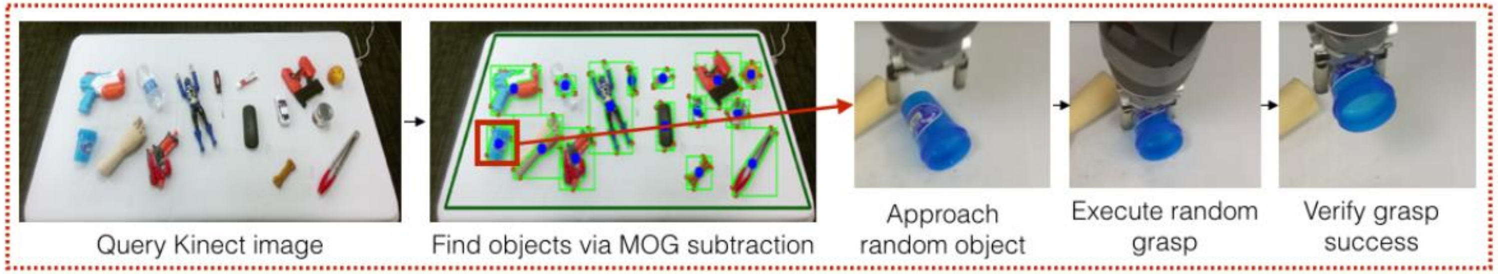
- Features for road segmentation (Depth -> RGB)

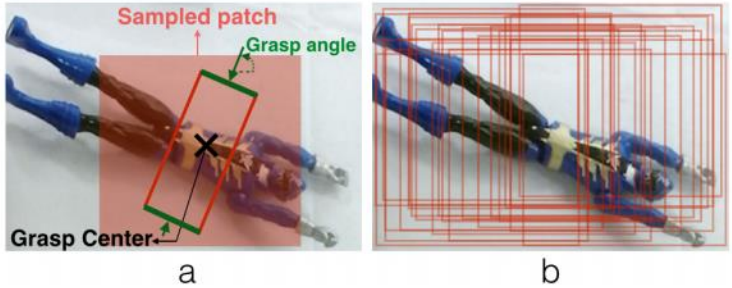


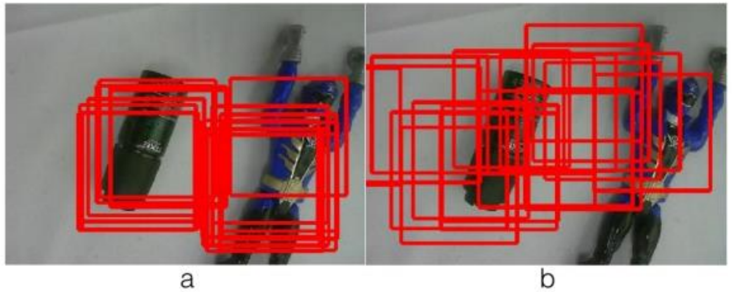
Weiyue W. , Naiyan W. , Xiaomin W. , Suya Y. and Ulrich N. Self-Paced Cross-Modality Transfer Learning for Efficient Road Segmentation. In *ICRA2017*

Cross-Modality

- Features for grasping
  - Verify whether we could grasp the center of a patch at a given angle



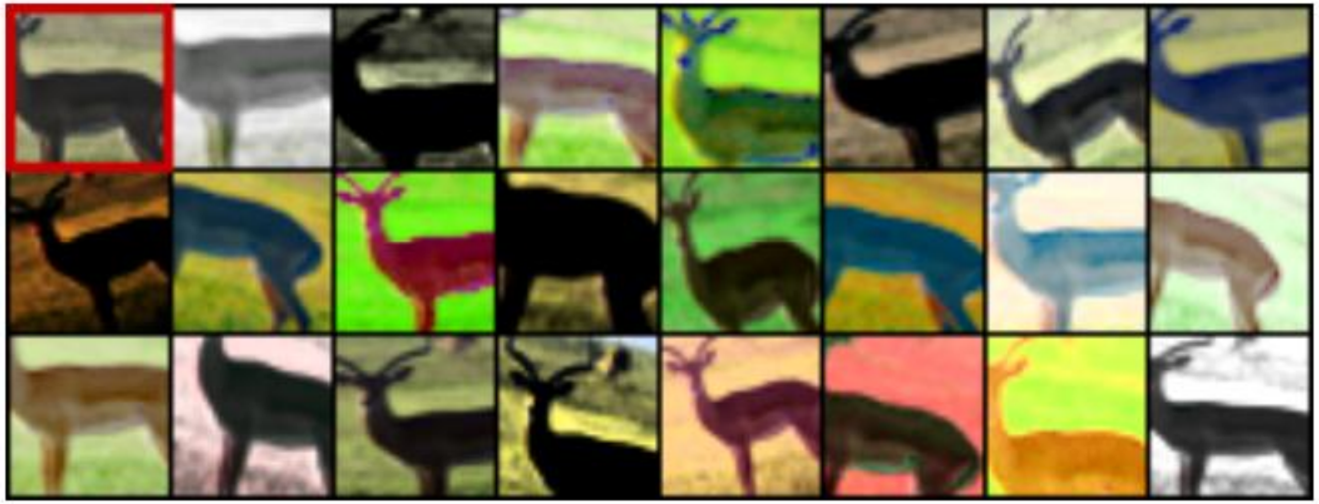




Pinto, L., & Gupta, A. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *ICRA 2016*

Exemplar Learning

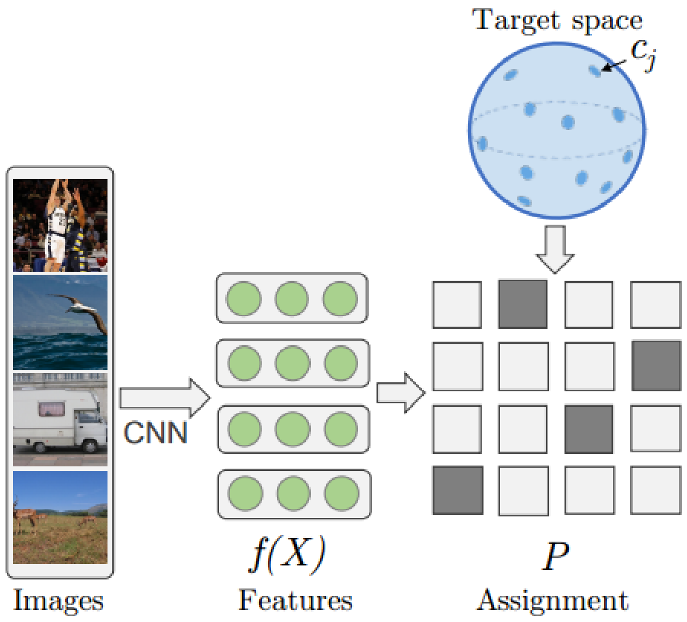
- Learning instance features
  - Each data sample as one class
  - Need strong augmentation



Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. Discriminative unsupervised feature learning with exemplar convolutional neural networks, arXiv preprint. *arXiv preprint arXiv:1506.02753*.

Exemplar Learning

- Learning instance features
  - The key is to avoid trivial solution. (Several tricks in this paper)
  - Project each sample on a random target uniformly samples on a unit ball



Bojanowski, P., & Joulin, A. Unsupervised Learning by Predicting Noise. *arXiv preprint arXiv:1704.05310*.

Evaluation

- Evaluate on general high-level vision tasks (classification, detection)
  - Be caution of different settings!

	Full train set						150 image set						
Method	All	>c1	>c2	>c3	>c4	>c5	All	>c1	>c2	>c3	>c4	>c5	#wins
Supervised													
Imagenet	56.5	57.0	57.1	57.1	55.6	52.5	17.7	19.1	19.7	20.3	20.9	19.6	NA
Sup. Masks (Ours)	51.7	51.8	52.7	52.2	52.0	47.5	13.6	13.8	15.5	17.6	18.1	15.1	NA
Unsupervised													
Jigsaw‡ [30]	49.0	50.0	48.9	47.7	45.8	37.1	5.9	8.7	8.8	10.1	9.9	7.9	NA
Kmeans [23]	42.8	42.2	40.3	37.1	32.4	26.0	4.1	4.9	5.0	4.5	4.2	4.0	0
Egomotion [2]	37.4	36.9	34.4	28.9	24.1	17.1	–	–	–	–	–	–	0
Inpainting [35]	39.1	36.4	34.1	29.4	24.8	13.4	–	–	–	–	–	–	0
Tracking-gray [46]	43.5	44.6	44.6	44.2	41.5	35.7	3.7	5.7	7.4	9.0	9.4	9.0	0
Sounds [33]	42.9	42.3	40.6	37.1	32.0	26.5	5.4	5.1	5.0	4.8	4.0	3.5	0
BiGAN [10]	44.9	44.6	44.7	42.4	38.4	29.4	4.9	6.1	7.3	7.6	7.1	4.6	0
Colorization [51]	44.5	44.9	44.7	44.4	42.6	38.0	6.1	7.9	8.6	10.6	10.7	9.9	0
Split-Brain Auto [52]	43.8	45.6	45.6	46.1	44.1	37.6	3.5	7.9	9.6	10.2	11.0	10.0	0
Context [8]	**49.9**	**48.8**	44.4	44.3	42.1	33.2	6.7	**10.2**	9.2	9.5	9.4	8.7	3
Context-videos† [8]	47.8	47.9	46.6	**47.2**	44.3	33.4	6.6	9.2	10.7	12.2	11.2	9.0	1
Motion Masks (Ours)	48.6	48.2	**48.3**	47.0	**45.8**	**40.3**	**10.2**	**10.2**	**11.7**	**12.5**	**13.3**	**11.0**	**9**

Pathak, D., Girshick, R., Dollár, P., Darrell, T., & Hariharan, B. Learning Features by Watching Objects Move. *In CVPR 2017*.

Evaluation

- Best so far

Initialization		Architecture	Class. %mAP	Seg. %mIU
ImageNet	(+FoV)	VGG-16	86.9	69.5
Random (ours)		AlexNet	46.2	23.5
Random [31]		AlexNet	53.3	19.8
k-means [19, 5]		AlexNet	56.6	32.6
k-means [19]		VGG-16	56.5	-
k-means [19]		GoogLeNet	55.0	-
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**

- Action Recognition

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.	**60.3**	**32.5**
Rand weights - Stack-of-Diff.	51.3	28.3
ImageNet weights - Stack-of-Diff.	70.1	40.8

Discussion

- How to cross the semantic gap between low-level and high-level?
  - Utilize high-level/global context
  - Explore piece-wise rigidity in real-life
  - More to discover…
- What is a useful self-supervised learning?
  - Improve the performance of subsequent task.
  - Task Related Self-Supervised Learning

Active Research Groups

- Alexei Efros (Berkeley)

- Abhinav Gupta (CMU)

- Martial Hebert (CMU)

Uncovered Papers

- **Colorization:**
- Larsson, G., Maire, M., & Shakhnarovich, G. Learning representations for automatic colorization. In *ECCV 2016*.
- Larsson, G., Maire, M., & Shakhnarovich, G. Colorization as a Proxy Task for Visual Understanding. *In CVPR 2017*.
- **Optical Flow**
- J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In ECCVW, 2016.
- Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. G. Guided optical flow learning. *arXiv preprint arXiv:1702.02295*.
- Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., & Zha, H. Unsupervised Deep Learning for Optical Flow Estimation. In *AAAI* 2017
- **Others**
- Cruz, R. S., Fernando, B., Cherian, A., & Gould, S. DeepPermNet: Visual Permutation Learning. *arXiv preprint arXiv:1704.02729*.
- Nair, A., Chen, D., Agrawal, P., Isola, P., Abbeel, P., Malik, J., & Levine, S. Combining Self-Supervised Learning and Imitation for Vision-Based Rope Manipulation. *arXiv preprint arXiv:1703.02018*.
- Pinto, L., Gandhi, D., Han, Y., Park, Y. L., & Gupta, A. The curious robot: Learning visual representations via physical interactions. In *ECCVW 2016*.