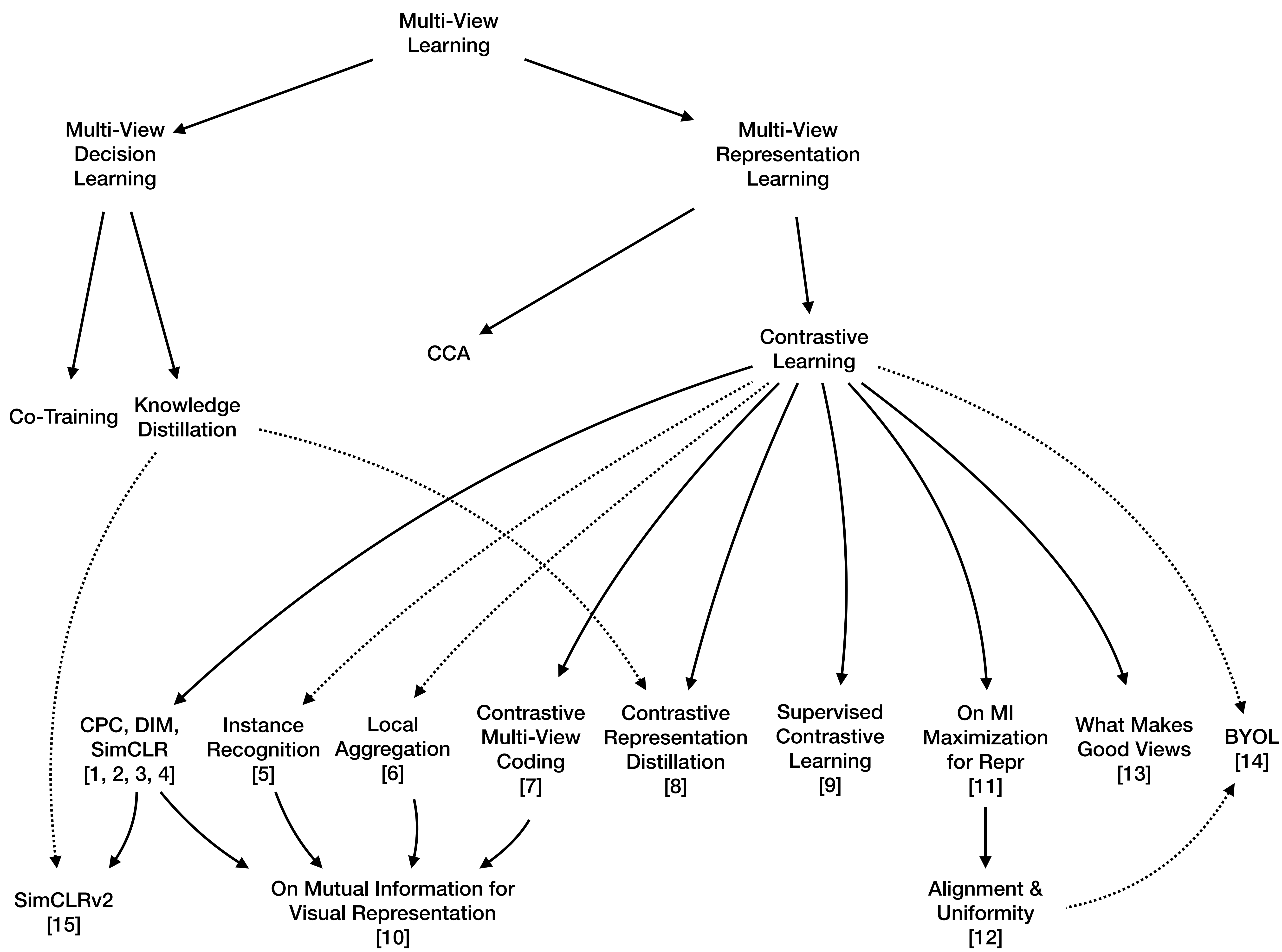
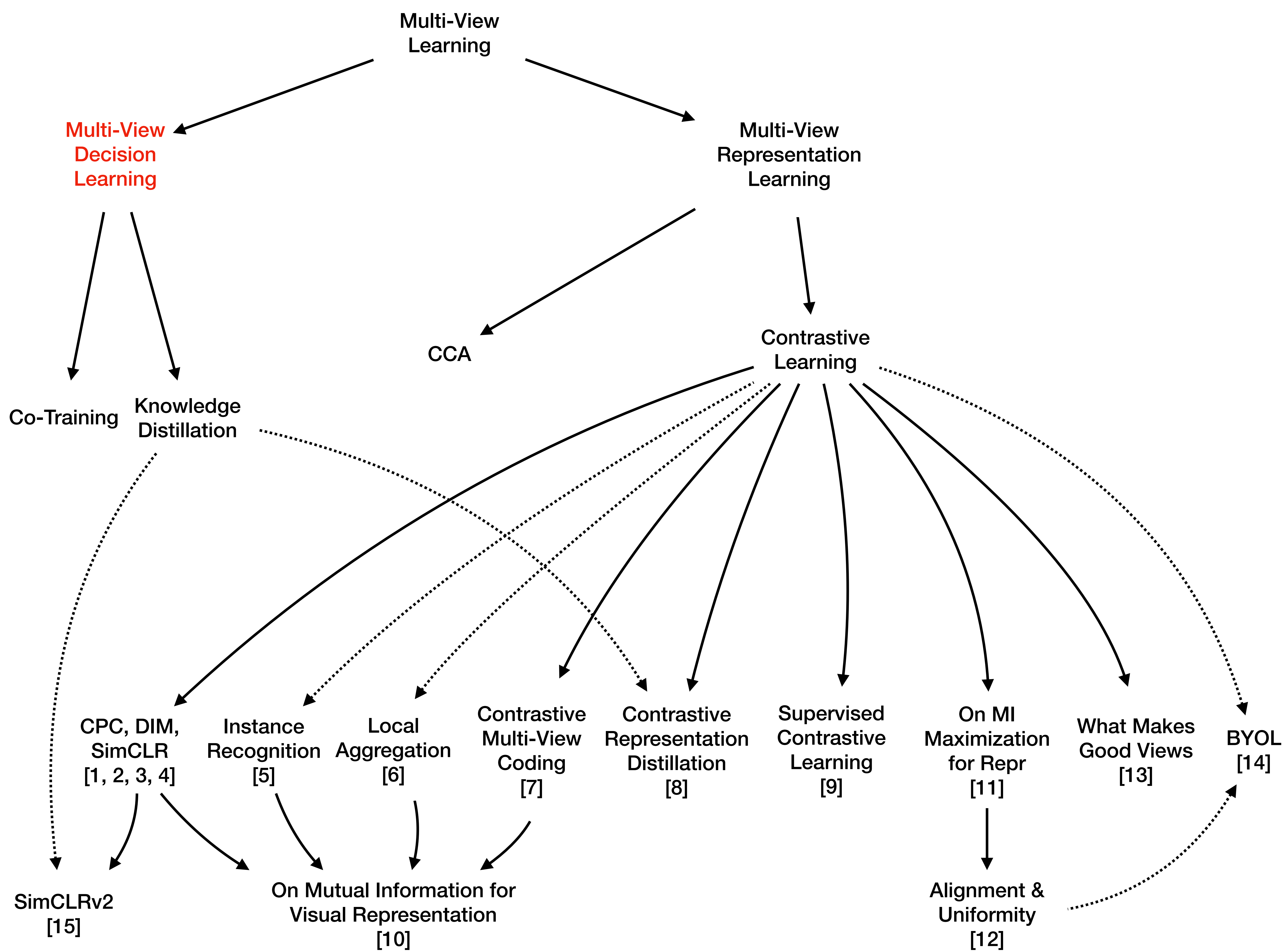
Multi-View Learning

Shengchao Liu
MILA & UdeM


Multi-View Learning
Multi-View Decision Learning
Multi-View Representation Learning
Co-Training
Knowledge Distillation
CCA
Contrastive Learning
CPC, DIM, SimCLR [1, 2, 3, 4]
Instance Recognition [5]
Local Aggregation [6]
Contrastive Multi-View Coding [7]
Contrastive Representation Distillation [8]
Supervised Contrastive Learning [9]
On MI Maximization for Repr [11]
What Makes Good Views [13]
BYOL [14]
SimCLRv2 [15]
On Mutual Information for Visual Representation [10]
Alignment & Uniformity [12]



Multi-View Learning
Multi-View Decision Learning
Multi-View Representation Learning
Co-Training
Knowledge Distillation
CCA
Contrastive Learning
CPC, DIM, SimCLR [1, 2, 3, 4]
Instance Recognition [5]
Local Aggregation [6]
Contrastive Multi-View Coding [7]
Contrastive Representation Distillation [8]
Supervised Contrastive Learning [9]
On MI Maximization for Repr [11]
What Makes Good Views [13]
BYOL [14]
SimCLRv2 [15]
On Mutual Information for Visual Representation [10]
Alignment & Uniformity [12]

Co-Training

<u>Combining Labeled and Unlabeled Data with Co-Training</u>, COLT 1998

- Co-training assumption $f(x) = f_1(v_1) = f_2(v_2), \forall x = (v_1, v_2) \sim X$

  1. Learn a separate classifier for each view on S (labeled data)
  2. Predictions of two classifiers on U (unlabeled data) are gradually added to S

  - Two views are different and provide complementary info

Co-Training

Deep Co-Training for Semi-Supervised Image Recognition, ECCV 2018

- View Difference Constraint assumption (encourages the networks to be different)
  $\exists X', f_1(v_1) \neq f_2(v_2), \forall x = (v_1, v_2) \sim X'$
- Deep Co-Training
  - Co-training assumption: different views agree on predictions
    $L(x) = H\big(\frac{1}{2}(p_1(x) + p_2(x))\big) - \frac{1}{2}\big(H(p_1(x)) + H(p_2(x))\big)$
  - View Difference Constraint:
    - Adversarial images D' where $p_1(x) \neq p_2(x), \forall x \in D'$, i.e.,$D \cap D' = \varnothing$
    - Adversarial images $D' = \{g(x) \,|\, x \in D\}$
    - $g(x)$ is an adversarial example that fools the network p_2 but not network p_1
    - Thus we propose to train the network p_1 to be resistant to adversarial examples $g_2(x)$ of p_2 by minimizing the CE between $p_2(x)$ and $p_1(g_2(x))$,
      $L(x) = H\big(\frac{1}{2}(p_1(x) + p_2(g_1(x)))\big) + H\big(\frac{1}{2}(p_2(x) + p_1(g_2(x)))\big)$

Knowledge Distillation

Distilling the Knowledge in a Neural Network, NIPS'15 Workshop

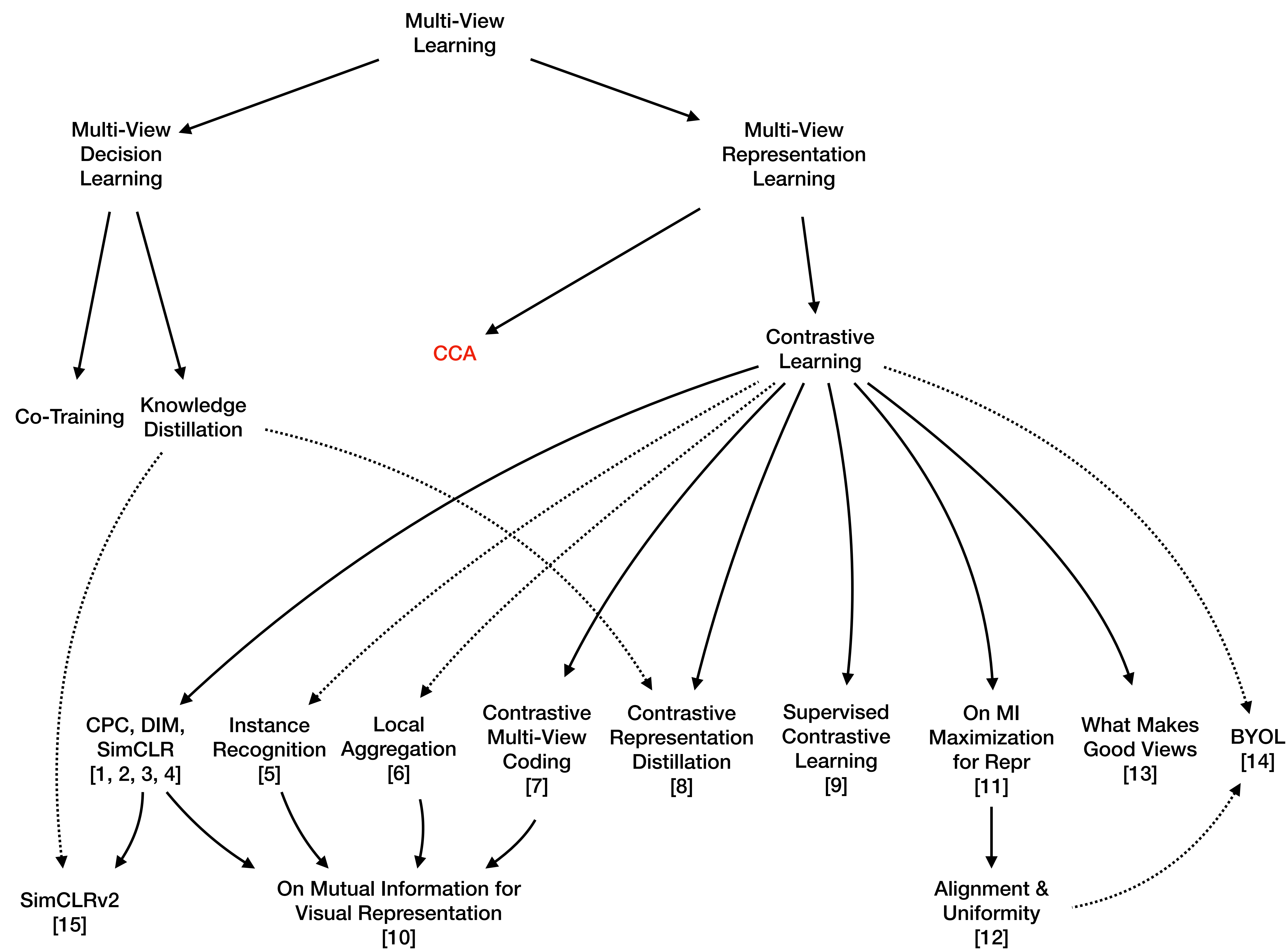
Geoffrey Hinton, etc.

- Transfer knowledge from teacher (cumbersome model) to student (distilled model)
- Knowledge Distillation: $\mathscr{L}_{KD} = (1-\alpha)H(y, y^S) + \alpha\rho^2 H(\sigma(\frac{z^T}{\rho}), \sigma(\frac{z^S}{\rho}))$,

  where $H(\sigma(\frac{z^T}{\rho}), \sigma(\frac{z^S}{\rho})) = KL(\sigma(\frac{z^T}{\rho}), \sigma(\frac{z^S}{\rho})) + H(\sigma(\frac{z^T}{\rho}))$

Notice: Matching logits is a special case of distillation

https://nervanasystems.github.io/distiller/knowledge_distillation.html#hinton-et-al-2015



Multi-View Learning
Multi-View Decision Learning
Multi-View Representation Learning
Co-Training
Knowledge Distillation
CCA
Contrastive Learning
CPC, DIM, SimCLR [1, 2, 3, 4]
Instance Recognition [5]
Local Aggregation [6]
Contrastive Multi-View Coding [7]
Contrastive Representation Distillation [8]
Supervised Contrastive Learning [9]
On MI Maximization for Repr [11]
What Makes Good Views [13]
BYOL [14]
SimCLRv2 [15]
On Mutual Information for Visual Representation [10]
Alignment & Uniformity [12]

Canonical Correlation Analysis (CCA)

Relations between two sets of variates, Biometrika 1936

Deep Canonical Correlation Analysis, ICML'13

On deep multi-view representation learning, ICML'15

- CCA

$$(w_1^*, w_2^*) = \arg\max_{w_1, w_2} corr(w1^T X_1, w_2^T X_2) = \arg\max_{w_1, w_2} \frac{w_1^T \Sigma_{12} w_2}{\sqrt{w_1^T \Sigma_{11} w1 w_2^T \Sigma_{22} w_2}}$$

- Solution:
  - $T = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$
  - U_k, V_k are top k left- and right- singular values of T
  - $(A_1^*, A_2^*) = (\Sigma_{11}^{-1/2} U_k, \Sigma_{22}^{-1/2} V_k)$

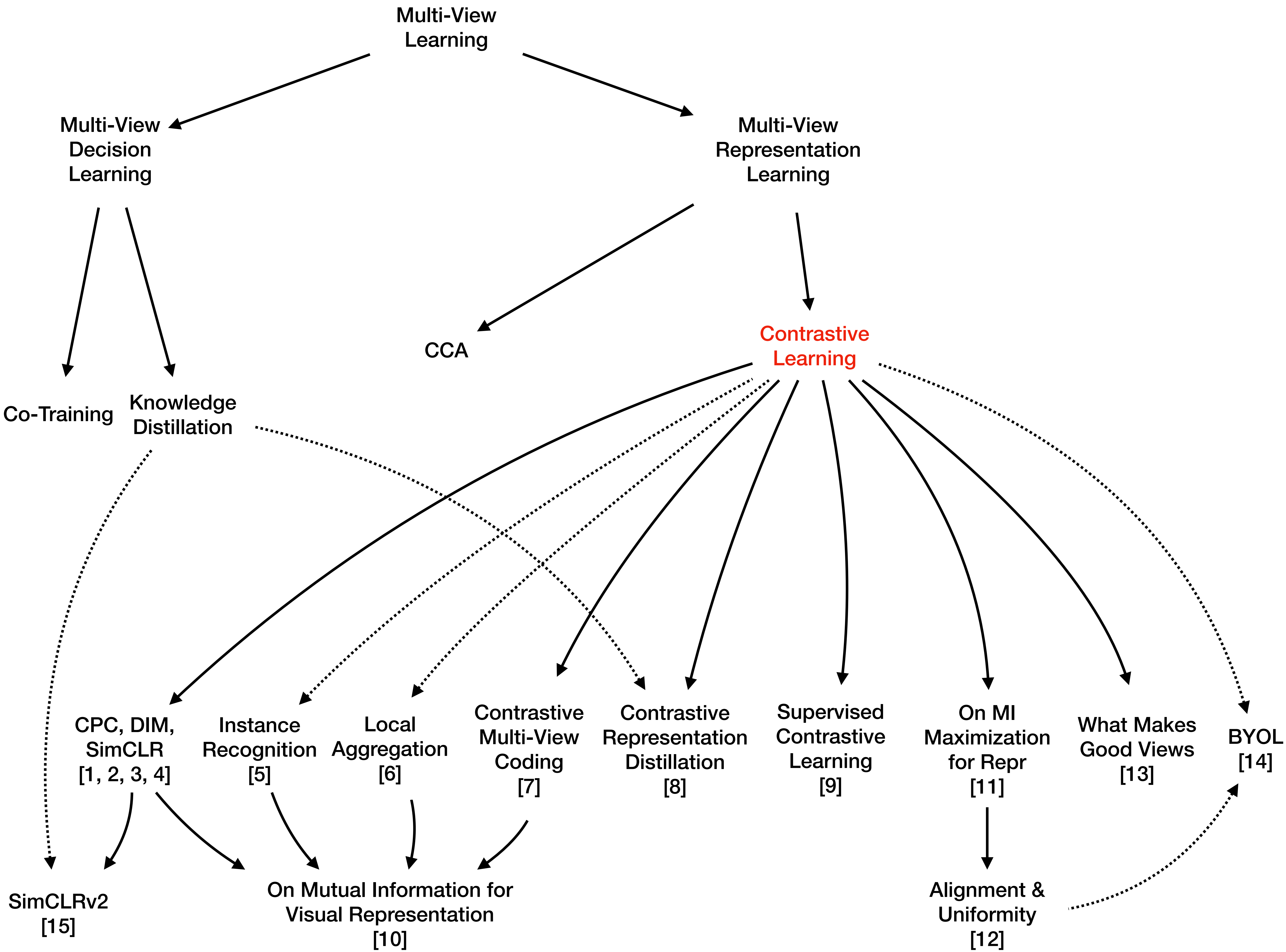
Canonical Correlation Analysis (CCA)

Relations between two sets of variates, Biometrika 1936

Deep Canonical Correlation Analysis, ICML'13

On deep multi-view representation learning, ICML'15

- Deep CCA $(w_1^*, w_2^*) = \arg\max_{w_1, w_2} corr(f_1(X_1; \theta_1), f_2(X_2; \theta_2))$
- Solution:
  - H_1, H_2 are feature matrices
  - $\bar{H}_1 = H_1 - \frac{1}{m} H_1 1, \bar{H}_2 = H_2 - \frac{1}{m} H_2 1$
  - $\hat{\Sigma}_{12} = \frac{1}{m-1} \bar{H}_1 \bar{H}_2^T, \hat{\Sigma}_{11} = \frac{1}{m-1} \bar{H}_1 \bar{H}_1^T + r_1 I$
  - $T = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$
  - $corr(H_1, H_2) = \|T\|_{tr} = tr(T^T T)^{1/2}$
  - $\frac{\partial corr(H_1, H_2)}{\partial H_1} = \frac{1}{m-1}(2 \nabla_{11} \bar{H}_1 + \nabla_{12} \bar{H}_2), \frac{\partial corr(H_1, H_2)}{\partial H_2} = \frac{1}{m-1}(2 \nabla_{22} \bar{H}_2 + \nabla_{12} \bar{H}_1)$



Multi-View Learning
Multi-View Decision Learning
Multi-View Representation Learning
Co-Training
Knowledge Distillation
CCA
Contrastive Learning
CPC, DIM, SimCLR [1, 2, 3, 4]
Instance Recognition [5]
Local Aggregation [6]
Contrastive Multi-View Coding [7]
Contrastive Representation Distillation [8]
Supervised Contrastive Learning [9]
On MI Maximization for Repr [11]
What Makes Good Views [13]
BYOL [14]
SimCLRv2 [15]
On Mutual Information for Visual Representation [10]
Alignment & Uniformity [12]

InfoNCE

[1] Representation learning with contrastive predictive coding (CPC), ArXiv'19

[2] Learning Deep Representations By Mutual Information Estimation and Maximization (DIM), ICLR'19

[3] On variational bounds of mutual information, ICML'19

[4] A Simple Framework for Contrastive Learning of Visual Representations (SimCLR), ICML'20

[*] Noise-contrastive estimation: A new estimation principle for unnormalized statistical models (NCE), AISTAT'10

- $\mathscr{L}_{\text{contrast}} = -\mathbb{E}\left[\frac{h_\theta(v_1^1, v_2^1)}{\sum_{j=1}^{k+1} h_\theta(v_1^1, v_2^j)}\right]$
- $I(z_i; z_j) \geq \log(k) - \mathscr{L}_{\text{contrast}}$

[5] <u>Unsupervised feature learning via non-parametric instance discrimination</u>, CVPR'18
Zhirong Wu, etc.

- Observation: class-level classification can implicitly learn class-wise similarity
    - For a leopard image, the confidence is leopard > jaguar > bookcase
- Extend this to the instance-level:
    - instance-level classification can implicitly learn the instance-wise similarity
        - Memory bank: θ, f_i are updated with SGD first, then $f_i \rightarrow v_i$
        - $$P(i \mid f_i) = \frac{\exp(v_i^T f_i / \tau)}{\sum_{j=1}^{n} \exp(v_j^T f_i / \tau)}, J(\theta) = -\sum_{i=1}^{n} \log P(i \mid f_\theta(x_i))$$
        - Too many classes / n is too large => NCE
        - $$h(i; v) = P(D = 1 \mid i, v) = \frac{P(i \mid v)}{P(i \mid v) + mP_n(i)},$$ $$J_{NCE}(\theta) = -\mathbb{E}_{P_d}[\log h(i, v)] - m\mathbb{E}_{P_n}[\log(1 - h(i, v))]$$

Not between views, but between instances

[6] Local Aggregation for Unsupervised Learning of Visual Embeddings, ICCV'19

Chengxu Zhang, etc. Stanford

- Local Aggregation: contrastive learning on class
- B_i : k nearest neighbors to x_i
- C_i : the set of nodes belong to the same cluster as x_i

  (usually C_i is a subset of B_i)
- $P(A \mid v) = \sum_i p(i \mid v)$, where $p(i \mid v) = \dfrac{\exp(v_i^T v/\tau)}{\sum_j \exp(v_j^T v/\tau)}$
- $L(C_i, B_i \mid \theta, x_i) = -\log \dfrac{P(C_i \cap B_i \mid v_i)}{P(B_i \mid v_i)}$
  - B_i is background neighbors/sampled pairs
  - C_i is close neighbors/positive pairs.

Not between views, but between instances

[7] Contrastive Multi-View Coding, ArXiv'19
Yonglong Tian, Dilip Krishnan, Phillip Isola

- $\mathscr{L}^{V_1,V_2}_{\text{contrast}} = -\mathbb{E}_{\{v_1^1, v_2^1, v_2^2, \ldots, v_2^{k+1}\}}\left[\frac{h_\theta(v_1^1, v_2^1)}{\sum_{j=1}^{k+1} h_\theta(v_1^1, v_2^j)}\right]$
- $\mathscr{L}_{contrast} = \mathscr{L}^{V_1,V_2}_{\text{contrast}} + \mathscr{L}^{V_2,V_1}_{\text{contrast}}$
- $I(z_i; z_j) \geq \log(k) - \mathscr{L}_{\text{contrast}}$

[8] Contrastive Representation Distillation, ICLR'20
Yonglong Tian, Dilip Krishnan, Phillip Isola

- Knowledge Distillation: $\mathscr{L}_{KD} = (1-\alpha)H(y, y^S) + \alpha\rho^2 H(\sigma(\frac{z^T}{\rho}), \sigma(\frac{z^S}{\rho}))$

- $$\begin{aligned} f^{S*} &= \arg\max_{f^S} \max_h \mathscr{L}_{critic}(h) \\ &= \arg\max_{f^S} \max_h \mathbb{E}_{q(T,S|C=1)}[\log h(T,S)] + N\mathbb{E}_{q(T,S|C=0)}[\log(1-h(T,S))] \end{aligned}$$

- $h(T,S) = \dfrac{\exp\big((g^T(T)' g(S)^S)/\tau\big)}{\exp\big((g^T(T)' g(S)^S)/\tau\big) + N/M}$

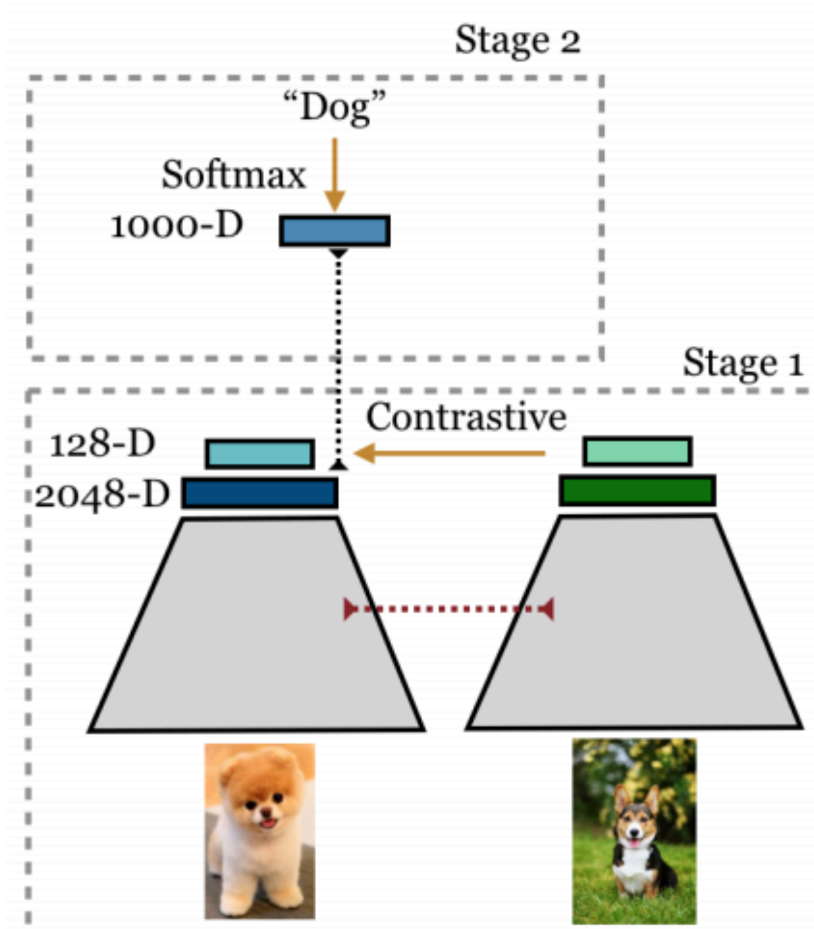
[9] Supervised Contrastive Learning, ArXiv'20
Google Research, Yonglong Tian, Phillip Isola, etc.

- $$L^{sup} = \sum_{i=1}^{2N} L_i^{sup}$$
- $$L_i^{sup} = -\frac{1}{2N_{\tilde{y}_i} - 1} \sum_{j=1}^{2N} 1_{i \neq j} 1_{\tilde{y}_i = \tilde{y}_j} \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k=1}^{2N} 1_{i \neq k} \exp(z_i \cdot z_k / \tau)}$$
- InfoNCE is motivated by NCE and N-pair losses:

One important property:

The ability to discriminate between signal and noise (negatives)

is obtained by adding more negative examples.



(c) Supervised Contrastive

[10] On Mutual Information in Contrastive Learning for Visual Representations, NIPS'20 In Submission
Mike Wu, Chengxu Zhang, etc., Stanford

- Three types of contrastive learning (IR, LA, CMC) are equivalent with InfoNCE
- Choices of views and negative sample distribution influence the performance

[11] On Mutual Information Maximization for Representation Learning, ICLR'20
Michael Tschannen, etc.

- Maximizing MI is not directly connected to the improved downstream performance
- Looser bounds with simpler critics can lead to better representations
- Connection between InfoNCE and deep metric learning
  - The deep metric learning
    $$L = \frac{1}{K}\sum_{i=1} K \log(1 + \sum_{j \neq i} \exp(\phi(x_i)^T\phi(y_j) - \phi(x_i)^T\phi(y_i)))$$
  - critic is $f(x, y) = \phi(x)^T\phi(y)$
  - Then I_{NCE} is equivalent to metric learning
  - Add more negative samples may not help

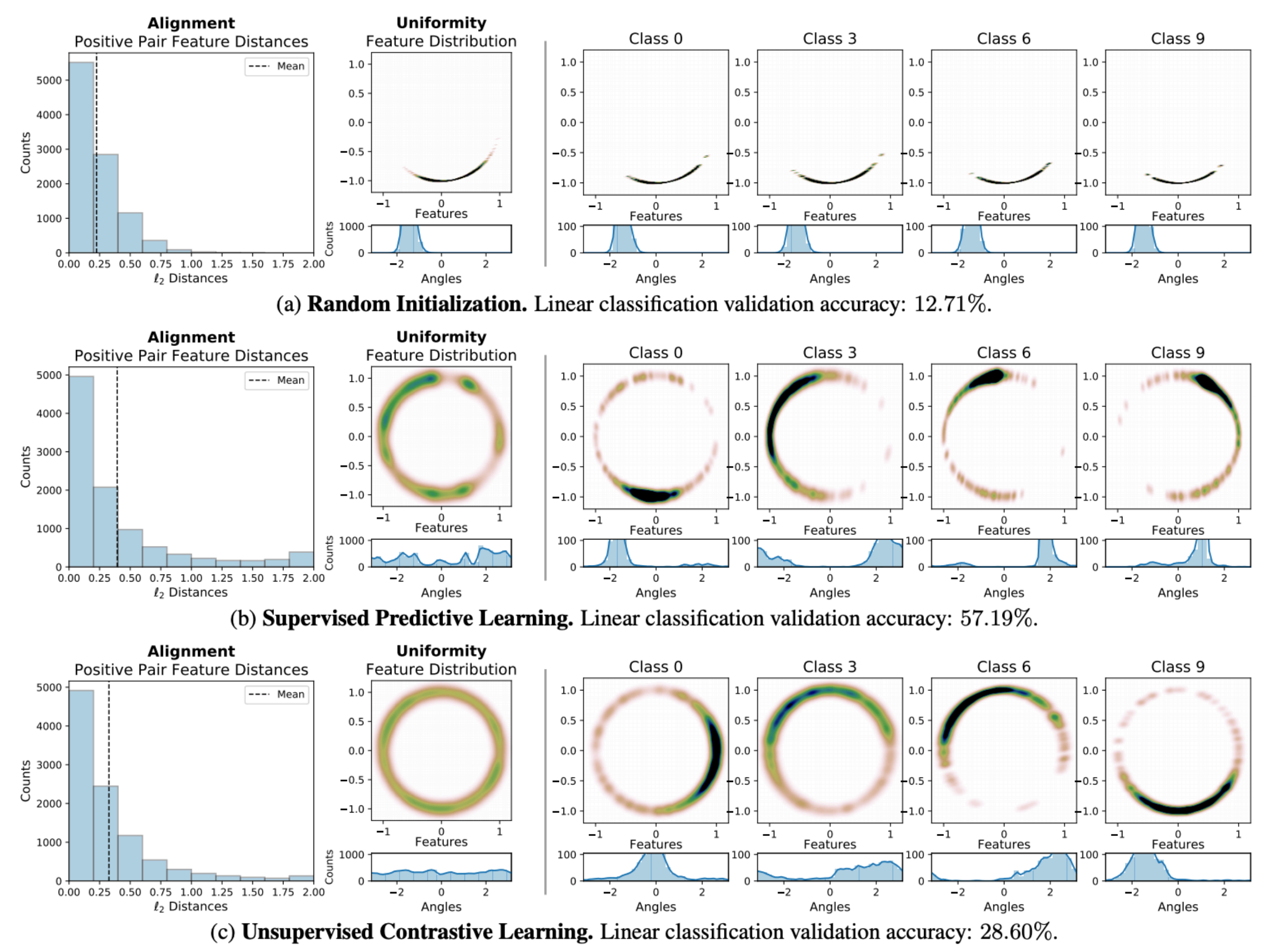
[12] Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, ICML'20
Tongzhou Wang, Phillip Isola

- $L_{contrastive} = \mathbb{E}_{(x,y)\sim p_{pos}}[-f(x)^T f(y)/\tau] + \mathbb{E}_{(x,y)\sim p_{pos}, x\sim p_{data}}[\log(\exp(f(x)^T f(y)/\tau) + \sum_i \exp(f(x)^T f(x_i)/\tau))]$

- Two key properties of contrastive loss, with metric to quantify each property
    - Alignment/closeness: Learned pos pairs should be similar, thus invariant to noise factors. $L_{align}(f) = -\mathbb{E}_{(x,y)\sim p_{pos}}[\|f(x) - f(y)\|_2^\alpha], \alpha > 0$
    - Uniformity: features should be roughly uniformly distributed on the unit hypersphere. $L_{uniform} = \log \mathbb{E}_{(x,y)\sim p_{data}}[\exp(-t\|f(x) - f(y)\|_2^2)], t > 0$

Instead of interpreted with InfoMAX, what contrastive loss doing is to learn an aligned and information-preserving encoder.

(perfectly uniform is the most entropic)



(a) **Random Initialization.** Linear classification validation accuracy: 12.71%.

(b) **Supervised Predictive Learning.** Linear classification validation accuracy: 57.19%.

(c) **Unsupervised Contrastive Learning.** Linear classification validation accuracy: 28.60%.

[13] What Makes for Good Views for Contrastive Learning?, ArXiv'20

Yonglong Tian, Phillip Isola, etc.

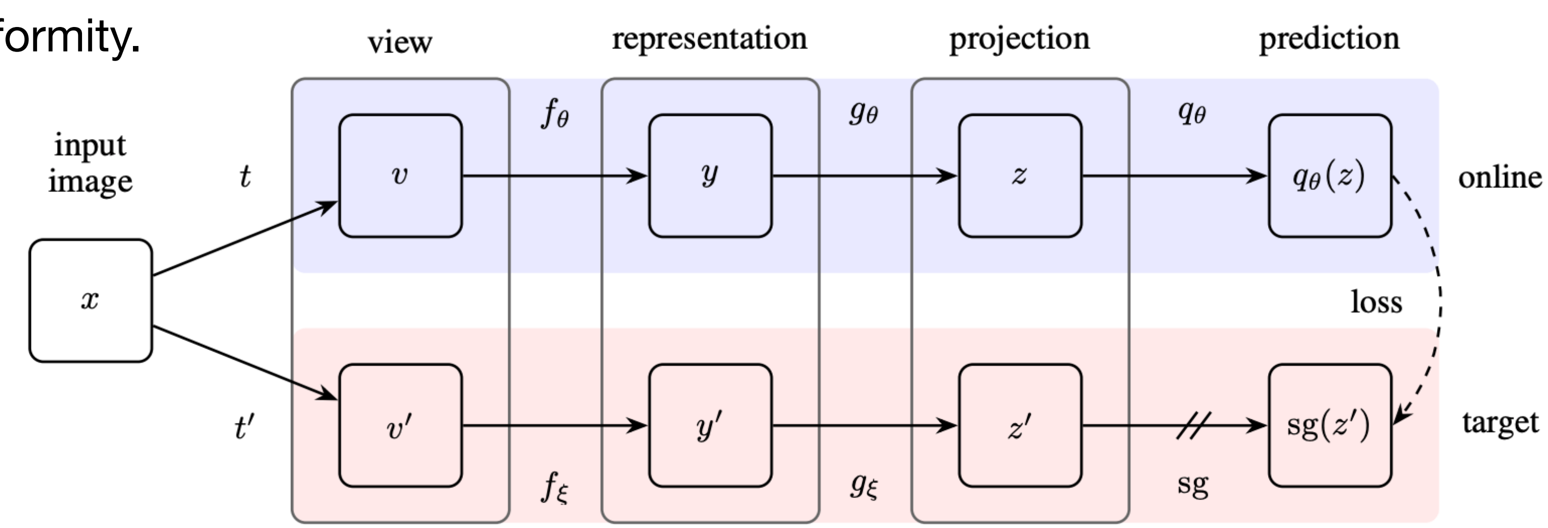
- InfoMin Principle:
  - Keep task-relevant semantics
  - Reduce the mutual information between views
  - => minimal sufficient encoders will ignore task-irrelevant information
  - => minimal sufficient encoders are still able to predict y

[14] Bootstrap Your Own Latent A New Approach to Self-Supervised Learning, In Submission NeurIPS'20

- Comparison between BYOL and contrastive learning
  - No Negative Sampling
  - More robust to the choice of image augmentation
  - Iteratively refine its representation
- Two networks and two views.
  1. Online network: $v_1 \rightarrow f_\theta, g_\theta \rightarrow z_1$
  2. Target network: $v_2 \rightarrow f_\xi, g_\xi \rightarrow z_2$
  3. Use online network (representation) to predict target network (representation)

$$\|\bar{q}_\theta(z_1) - \bar{z}_2\|^2 = 2 - 2 \cdot \frac{q_\theta(z_1)^T, z_2}{\|q_\theta(z_1)\|_2 \|z_2\|_2}$$

- Above is v_1 on online network and v_2 on target network. A symmetric loss is also included.
- BYOL is explicitly doing alignment, no uniformity.
- Moving average is scattering features.



Note: SimCLR suggests adding projection

[15] Big Self-Supervised Models are Strong Semi-Supervised Learners, In Submission NeurIPS'20

- Labeled data for teacher network, unlabeled data for student network.
- 3 steps:
  1. Pre-train $L = \log \frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} \exp(sim(z_i, z_j)/\tau)}$
  2. Fine-tune
  3. Distill $L^{distill} = \sum_{x_i} \left[\sum_{y} P^T(y \mid x_i; \tau) \log P^S(y \mid x_i; \tau)\right]$, where

     $$P(y \mid x_i) = \frac{\exp(f(x_i)[y]/\tau)}{\sum_{y'} \exp(f(x_i)[y']/\tau)}$$