噪声关联学习：
一种新的噪声标注学习范式

四川大学 计算机学院（软件学院）
彭玺

2022年1月5号

Partially View-aligned Representation Learning with Noise-robust Contrastive Loss

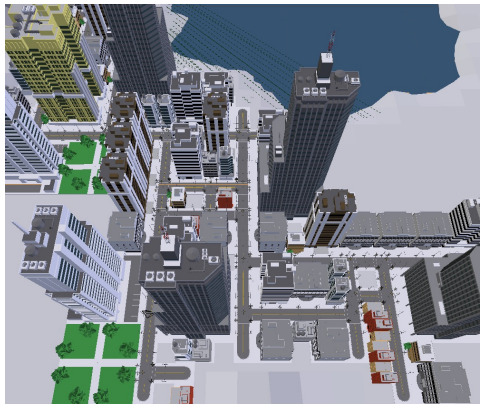
Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu, **Xi Peng***

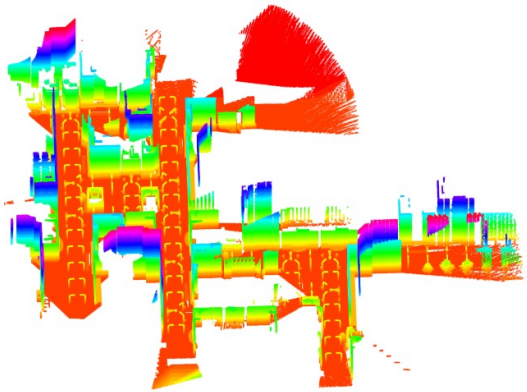
CVPR 2021

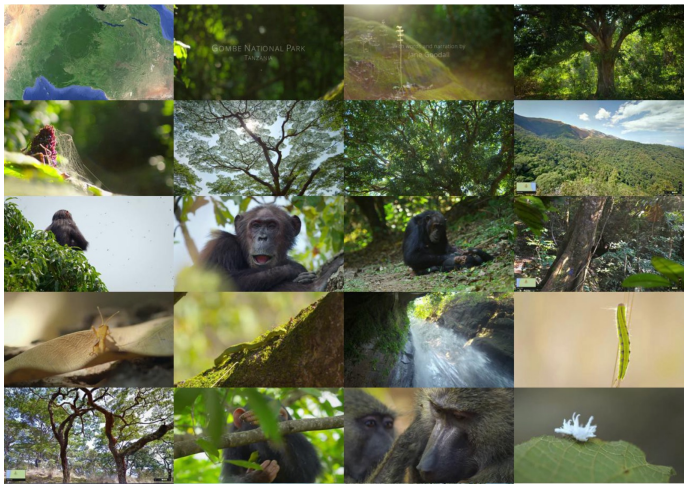
Background

- In real world, almost all data is presented in multiple modalities/views.

- Multi-modal learning has a wide range of such as visual navigation, cross-view retrieval, and so on.

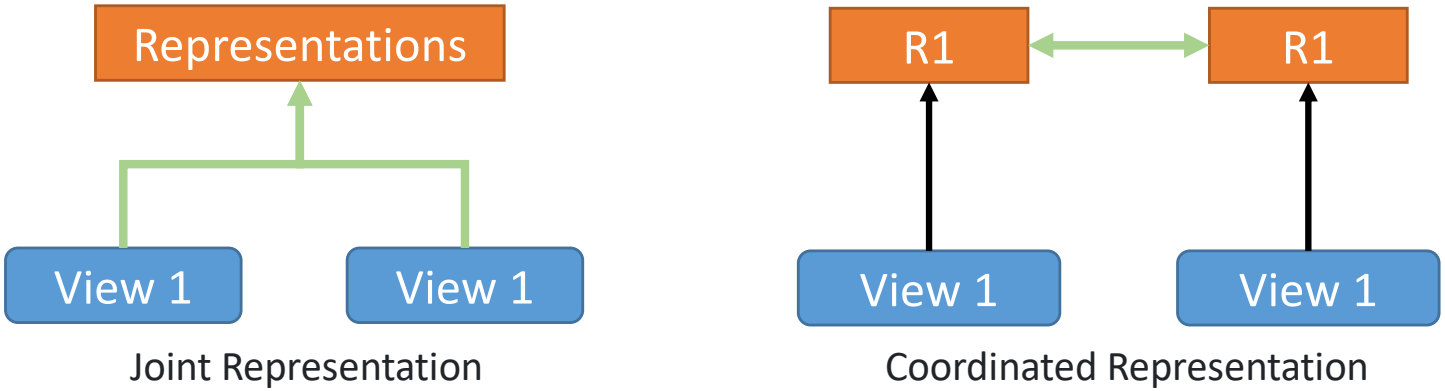






Observation

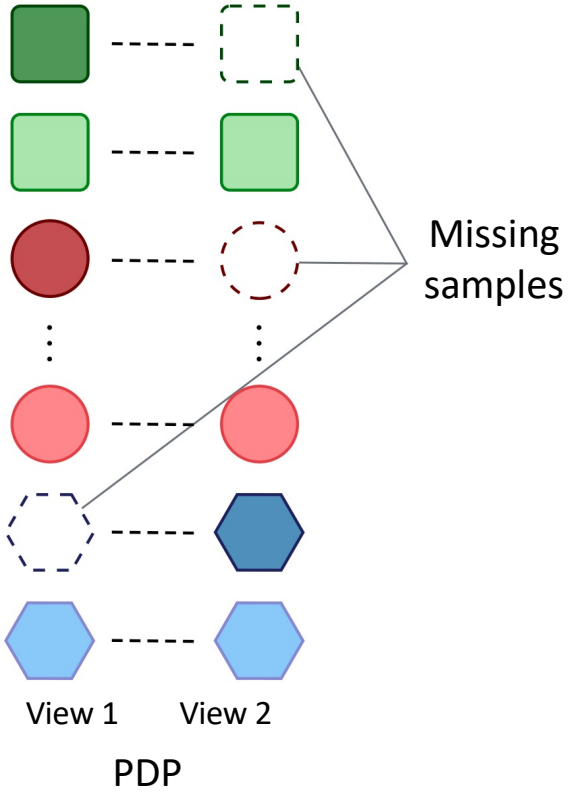
- Multi-view Learning

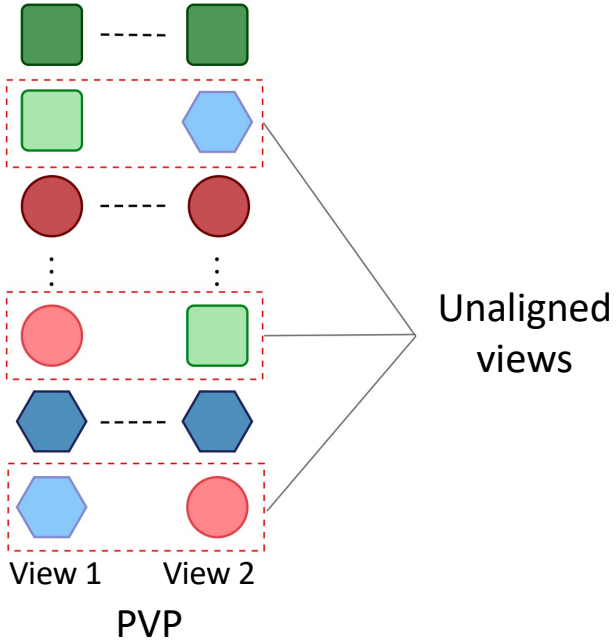


- Either of JR and CR explicitly use cross-view consistency which implicitly satisfies:
  - Completeness of data
  - Correspondence between views

Observation

- Completeness of data
  - Assumption: all samples will be present in all views.
  - Partially Data-missing Problem (PDP): some samples are missed in some views.

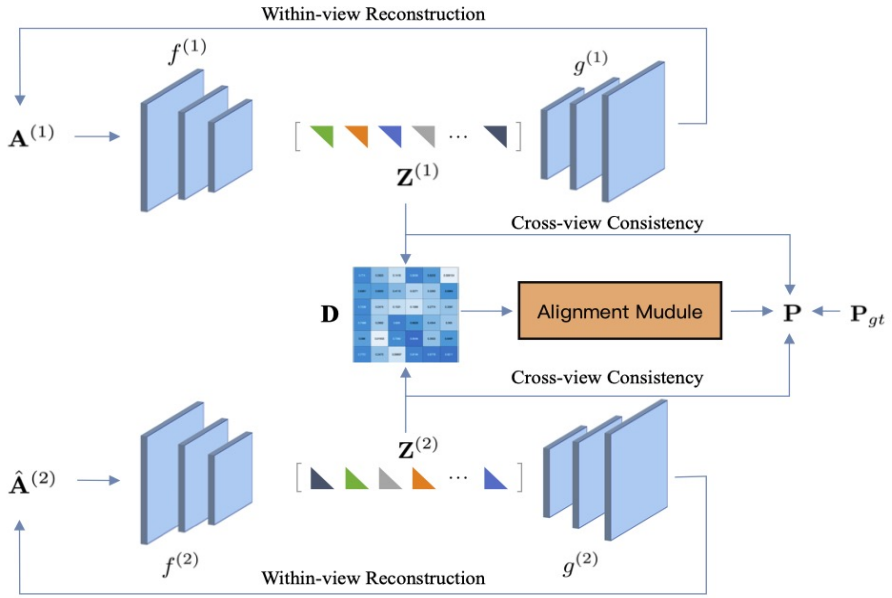




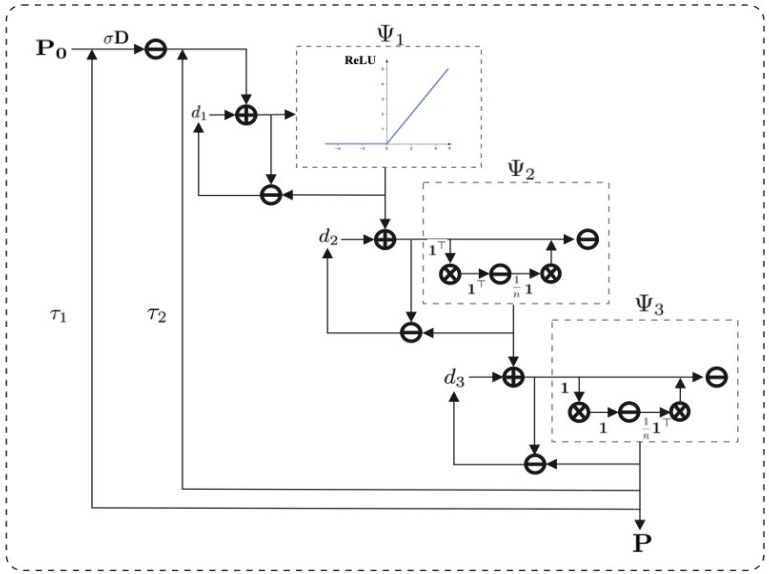
- Correspondence between views
  - Assumption: data from different views must be strictly aligned.
  - Partially View-aligned Problem (PVP): only a portion of the correspondences are known.

Observation

- Existing works
  - PDP: Some works have made remarkable progress on this problem.
  - PVP: Only PVC[1] explicitly considers this challenging problem and its goal is to achieve the *instance-level alignment* (IA), which is daunting and over-sufficient to the discriminative tasks such as classification and clustering.



Representation Learning Module

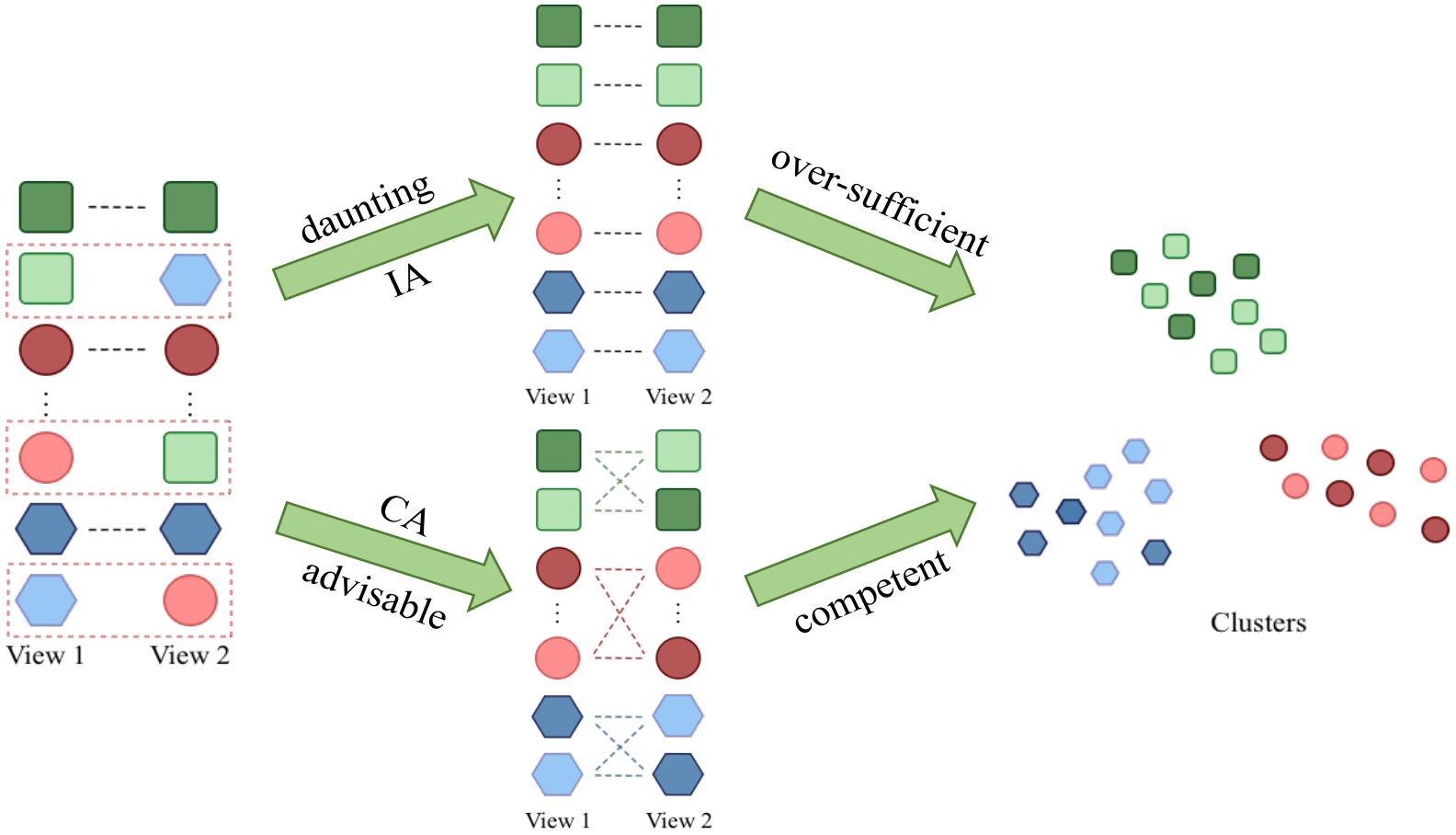


Differentiable Alignment Module

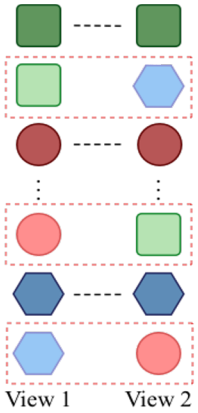
1. Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng*, Partially View-aligned Clustering, Neural Information Processing Systems (NeurIPS), 2020. (Oral)

Motivation

- Our basic idea
  - Category-level Alignment (CA), which embraces higher accessibility. Intuitively, an instance have a probability of $1/K$ *vs.* $1/N$ to be aligned in CA than IA.

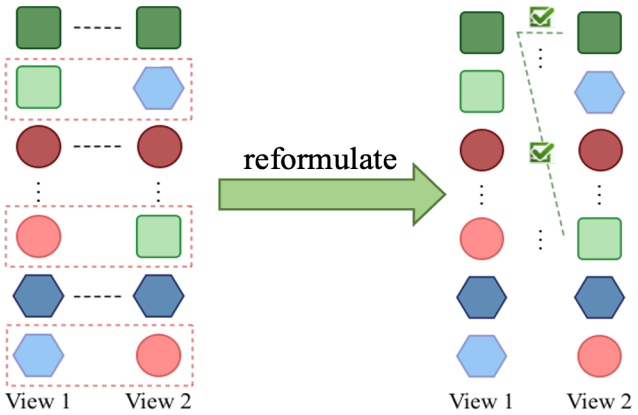


Key Idea



(a) Partially view-aligned

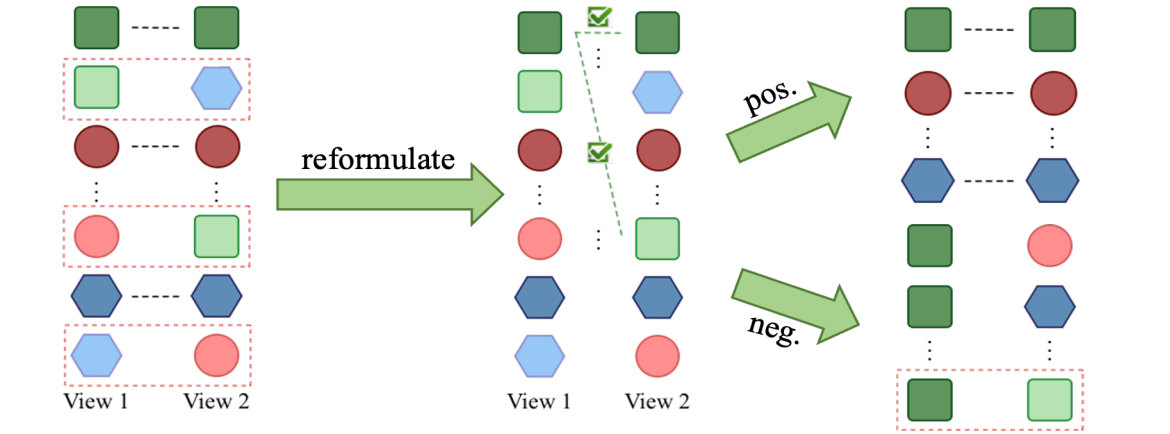
Key Idea



(a) Partially view-aligned (b) Categorial identification

- We reformulate CA as as an Categorial Identification (CI) task, which could be achieved by contrastive learning.

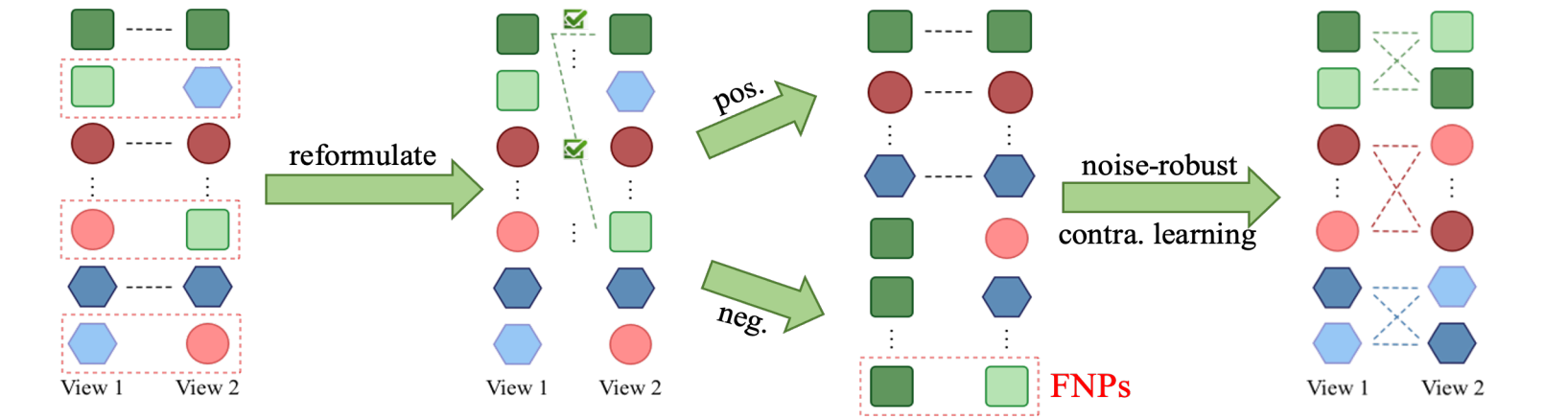
Key Idea



(a) Partially view-aligned (b) Categorial identification (c) Noisy correspondence

- We reformulate CA as as an Categorial Identification (CI) task, which could be achieved by contrastive learning.
- Under the unlabeled setting, we construct the positive pairs using the available aligned data and the Negative Pairs (NPs) using random sampling.

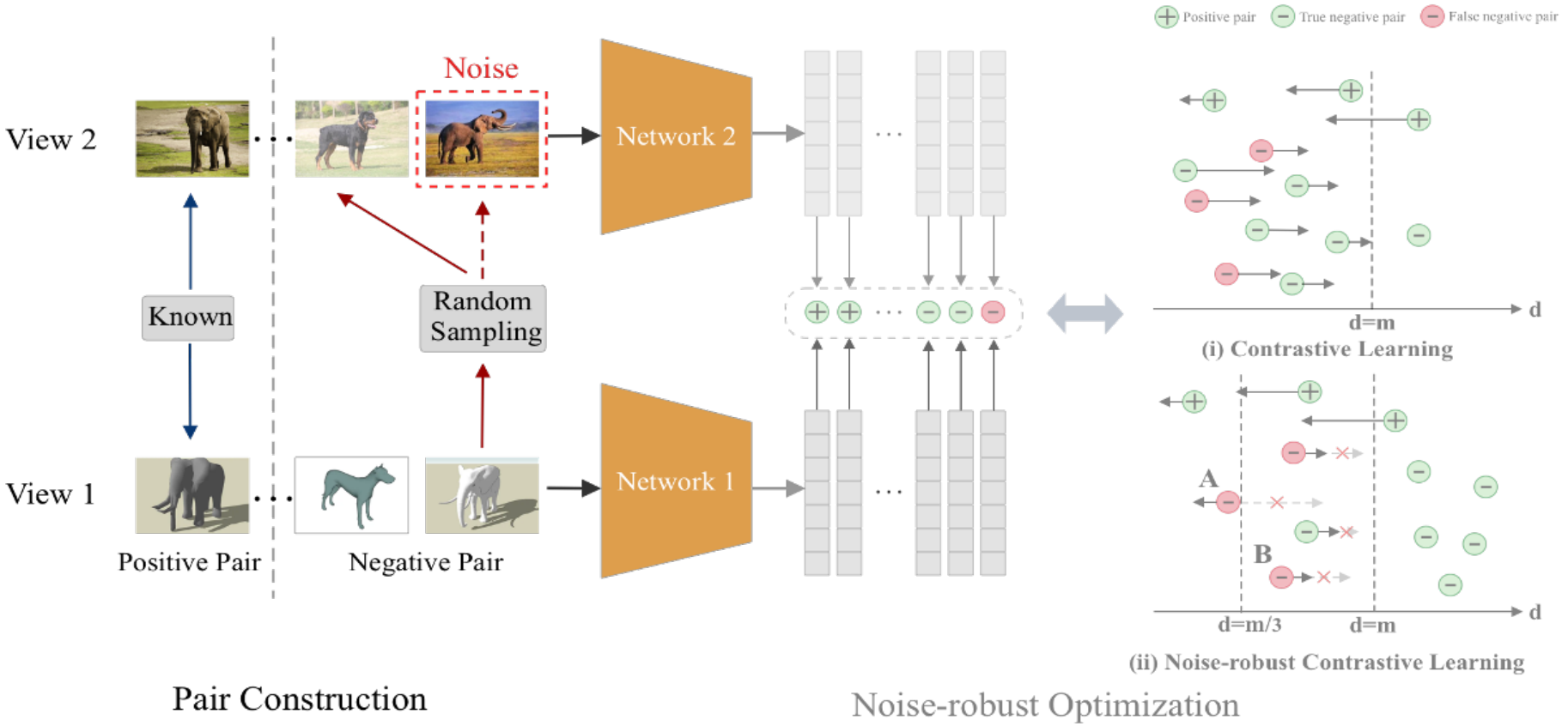
Key Idea



(a) Partially view-aligned (b) Categorial identification (c) Noisy correspondence (d) Category-level aligned

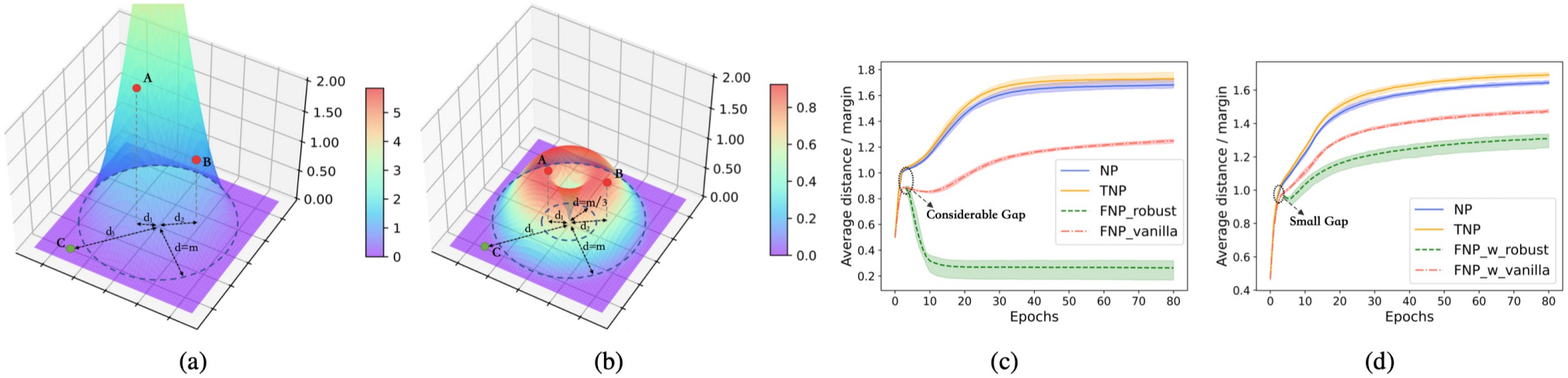
- We reformulate CA as as an Categorial Identification (CI) task, which could be achieved by contrastive learning.
- Under the unlabeled setting, we construct the positive pairs using the available aligned data and the Negative Pairs (NPs) using random sampling.
- To alleviate or even eliminate the influence of the False-Negative Pairs (FNPs) caused by random sampling, our model is with a novel noise-robust contrastive loss.

Pair Construction



- Positive and negative pairs: use the known aligned portion $\{\mathbf{A}\}_{i=1}^{2}$.
- Negative pair: randomly choose two samples $\mathbf{a}_i^1, \mathbf{a}_j^2$ $(i \neq j)$ from $\{\mathbf{A}\}_{i=1}^{2}$. Intuitively, these pairs have a probability of $1/K$ to be False Negative Pairs (FNPs).
- Pass the pairs into two encoders f_1 and f_2 to get view-specific representations.

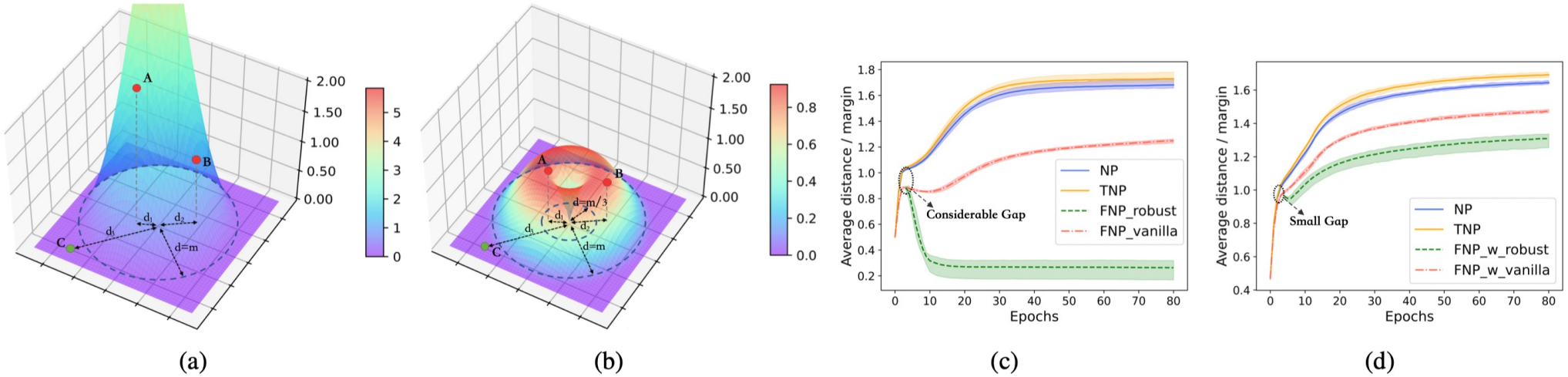
Overview



- Model optimized with vanilla loss would wrongly fit FNPs (**red** line in Fig. c and d).

$$\mathcal{L}_{van} = \frac{1}{2N}\sum_{i=1}^{N}\left(Pd(\mathbf{a}_i^1, \mathbf{a}_i^2) + (1-P)\max(m - d(\mathbf{a}_i^1, \mathbf{a}_j^2), 0)^2\right)$$

Overview



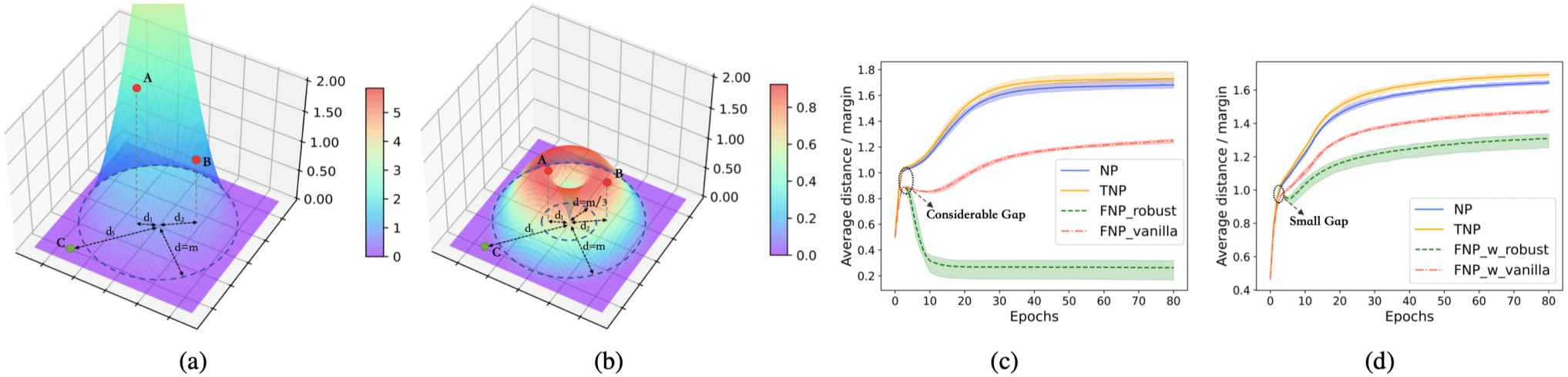
- Model optimized with vanilla loss would wrongly fit FNPs (red line in Fig. c and d).

$$\mathcal{L}_{van} = \frac{1}{2N}\sum_{i=1}^{N}\left(Pd(\mathbf{a}_i^1, \mathbf{a}_i^2) + (1-P)\max(m - d(\mathbf{a}_i^1, \mathbf{a}_j^2), 0)^2\right)$$

- The proposed noise-robust contrastive loss could alleviate (**green** line in Fig. d) or even eliminate (**green** line in Fig. c) the influence of FNPs.

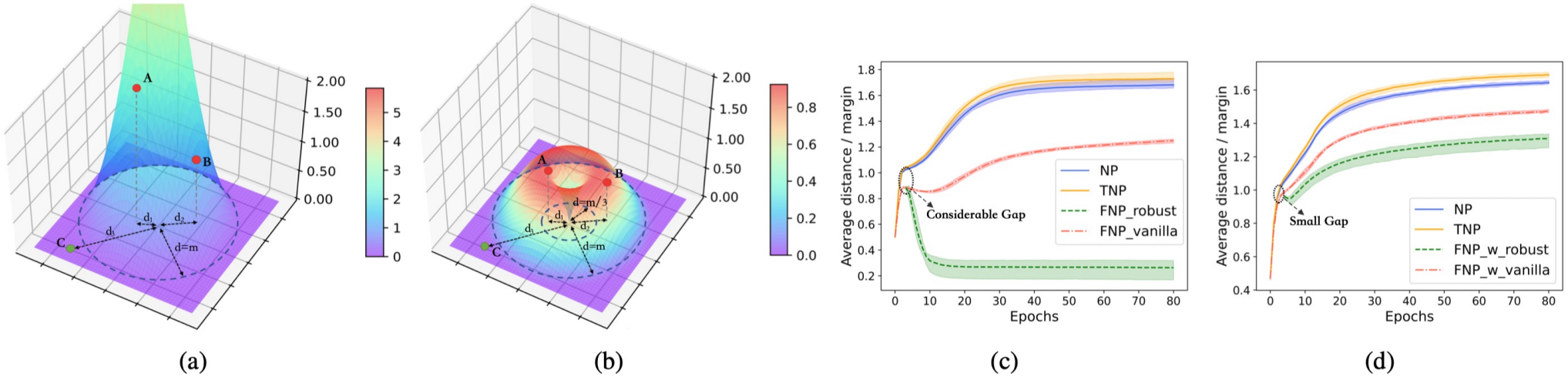
$$\mathcal{L}_{ncl} = \frac{1}{2N}\sum_{i=1}^{N}\left(Pd(\mathbf{a}_i^1, \mathbf{a}_i^2) + (1-P)\frac{1}{m}\max(md^{\frac{1}{2}}(\mathbf{a}_i^1, \mathbf{a}_j^2) - d^{\frac{3}{2}}(\mathbf{a}_i^1, \mathbf{a}_j^2), 0)^2\right)$$

Overview



- Reverse optimization ($0 < d < m/3$): For the negative pairs locating into the hole area (see A for example), the gradient of our loss will be reversed, and thus the distance of negative pairs will decrease.
- Slow optimization ($m/3 < d < m$): For the pairs locating into the slope area (see B for example), the optimization speed of our loss will be slower than that of the vanilla loss.
- The challenge remained is that $\mathcal{L}_{ncl}$ may inevitably impede the optimization of TNPs.

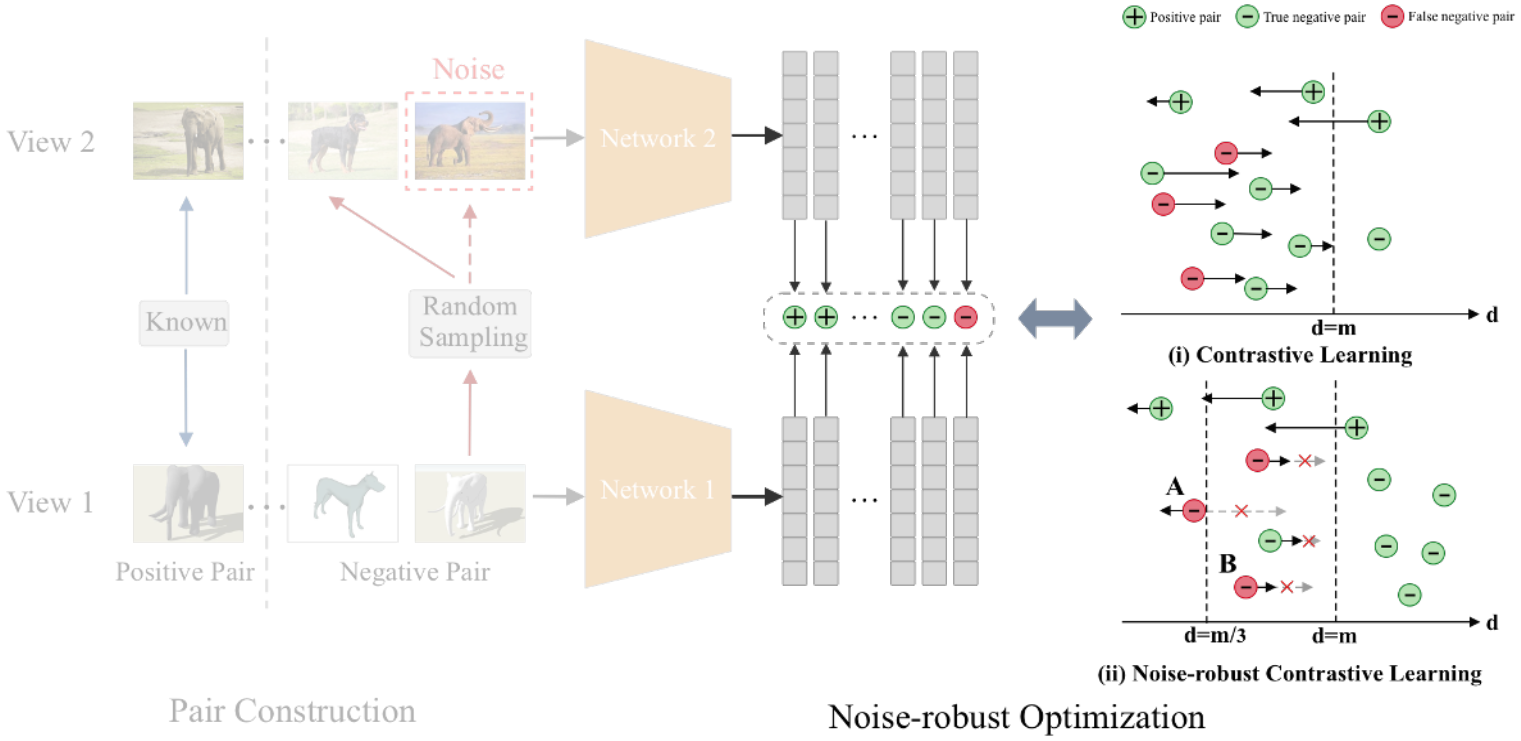
Overview



- Bengio *et al.*[1] have empirically found that the neural networks apt to fit the simple patterns first. Motived by this, we propose that TNPs could be regarded as simple patterns and FNPs could be treated as the complex ones.
- Thanks to the above observation, we propose adopting a two-stage optimization strategy to prevent FNPs from dominating the network optimization as follows:
  - Employ the vanilla lo$\mathcal{L}_{van}$ until the average distance of all NPs is larger than m.
  - Switch into the second stage with the proposed noise-robust contrastive lo$\mathcal{L}_{ncl}$.

1. A Closer Look at Memorization in Deep Networks, ICML 2017.

Overview



- TNPs (green circles) are fit faster than FNPs (red circles), which results in that most TNPs and FNPs will locate into the areas of $d > m$ and $d < m$, respectively.
- The distance of FNPs will either increase slowly (see red circle B) or decrease (see red circle A), thus alleviating or even eliminating the influence of noisy labels.

Experiment

- Baselines
  - Canonically Correlated Analysis (CCA) (A. Vinokourov et al., 2002)
  - Kernel canonically Correlated Analysis (KCCA) (Bach et al., 2002)
  - Deep Canonical Correlation Analysis (DCCA) (Andrew et al., 2013)
  - Deep Canonically Correlated AutoEncoders (DCCAE) (Wang et al., 2015)
  - Multi-View Clustering via Deep Matrix Factorization (DMF) (Zhao et al., 2017)
  - Latent multi-view subspace clustering (LMSC) (Zhang et al., 2017)
  - Self-weighted Multi-view Clustering (SwMC) (Nie et al., 2017)
  - Binary Multi-View Clustering (BMVC) (Zhang et al., 2018).
  - Autoencoder in Autoencoder Networks (AE2-Nets) (Zhang et al., 2019)
  - Partially View-aligned Clustering (PVC) (Huang et al., 2020)
- Datasets
  - Scene-15: 4,485 images associated with 15 indoor and outdoor scene categories.
  - Caltech101: 9,144 images distributed over 102 object and background categories.
  - Reuters: a subset, 18,758 samples from six classes.
  - NoisyMNIST: 70,000 samples of 10 classes.

Experiment

❖ Clustering performance

Aligned	Methods	Scene-15			Caltech-101			Reuters			NoisyMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Fully	CCA (NeurIPS'03)	36.37	36.91	19.82	20.25	45.41	16.34	44.31	20.34	14.52	71.31	52.60	48.46
	KCCA (JMLR'02)	37.93	37.42	21.38	21.45	45.58	17.62	**50.87**	22.34	**20.61**	**96.85**	**92.10**	**93.23**
	DCCA (ICML'13)	36.61	39.20	21.03	27.60	47.84	**30.86**	47.95	**26.57**	12.71	89.64	88.33	83.95
	DCCAE (ICML'15)	34.58	39.01	19.65	19.84	45.05	14.57	41.98	20.30	8.51	78.00	81.24	68.15
	LMSC (CVPR'17)	38.46	35.50	20.54	26.87	48.80	18.06	38.56	20.12	15.48	–	–	–
	MvC-DMF (AAAI'17)	30.99	31.35	15.68	24.35	44.98	14.82	33.83	14.89	12.59	74.39	63.22	49.79
	SwMC (IJCAI'17)	33.89	32.98	11.78	**30.74**	36.07	7.75	33.65	16.02	5.90	–	–	–
	BMVC (TPAMI'18)	**40.74**	**41.67**	**24.19**	27.59	46.43	21.28	42.39	21.86	15.14	88.31	77.01	76.58
	AE^2-Nets (CVPR'19)	37.17	40.47	22.24	20.79	45.01	15.89	42.39	19.76	14.87	42.11	43.38	30.42
Partially	CCA (NeurIPS'03)	32.73	34.24	18.80	20.06	41.56	**16.62**	40.87	**15.82**	12.68	34.46	29.83	17.89
	KCCA (JMLR'02)	33.09	31.43	16.35	12.57	31.36	7.65	40.08	11.80	11.27	26.57	18.19	10.55
	DCCA (ICML'13)	34.27	36.55	18.83	12.52	32.13	7.63	39.71	13.83	**14.38**	29.22	20.24	11.08
	DCCAE (ICML'15)	33.62	36.56	18.54	11.75	30.54	6.60	**41.42**	12.82	13.61	27.61	19.45	10.00
	LMSC (CVPR'17)	26.27	20.45	10.93	**21.54**	**40.26**	15.51	32.17	11.34	7.19	–	–	–
	MvC-DMF (AAAI'17)	28.49	24.31	11.22	9.54	23.41	3.84	32.58	12.36	11.08	27.34	22.96	6.85
	SwMC (IJCAI'17)	31.03	30.39	12.94	19.03	22.75	3.73	31.92	11.03	5.40	–	–	–
	BMVC (TPAMI'18)	**36.81**	**36.55**	**20.20**	12.13	31.33	7.11	38.15	11.57	12.07	28.47	24.69	14.19
	AE^2-Nets (CVPR'19)	28.56	26.58	12.96	10.45	29.51	7.90	35.49	10.61	8.07	**38.25**	**34.32**	**22.02**
Partially	PVC (NeurIPS'20)	37.88	39.12	20.63	22.11	**47.82**	17.98	42.07	20.43	16.95	81.84	82.29	82.03
	MvCLN (Mean)	**38.53**	**39.90**	**24.26**	**30.09**	43.07	**38.34**	**50.16**	**30.65**	**24.90**	**91.05**	**84.15**	**83.56**
	MvCLN (Best)	39.87	40.47	24.83	35.72	45.25	51.44	56.62	33.62	27.37	94.51	86.77	88.42

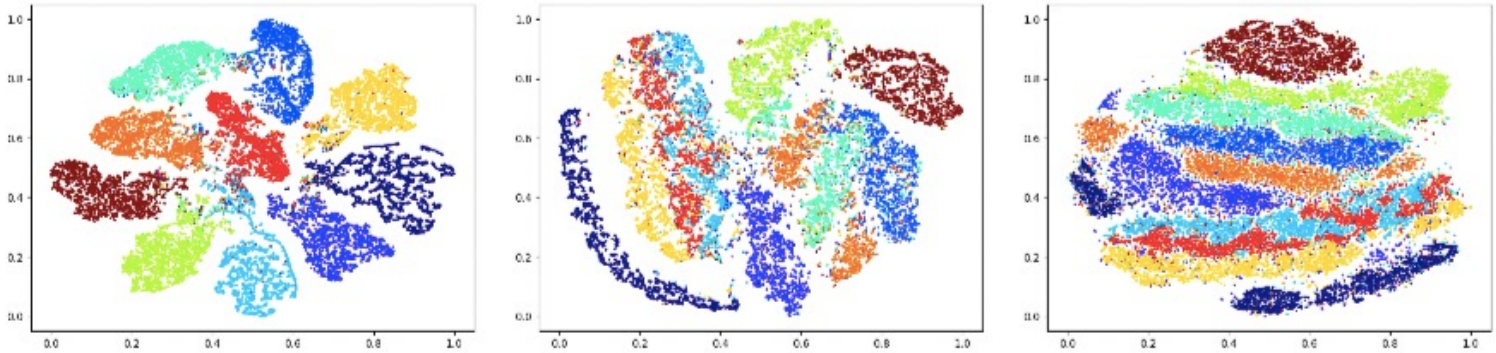
Experiment

❖ Classification performance

Aligned	Methods	Scene-15			Caltech-101			Reuters			NoisyMNIST		
		8 / 2	5 / 5	2 / 8	8 / 2	5 / 5	2 / 8	8 / 2	5 / 5	2 / 8	8 / 2	5 / 5	2 / 8
Fully	CCA (NeurIPS'03)	57.44	56.21	51.07	37.70	36.14	32.79	69.13	68.67	67.07	87.85	86.09	82.06
	KCCA (JMLR'02)	50.19	50.18	47.26	38.50	36.95	33.72	64.75	64.63	64.63	**97.20**	**97.18**	**97.08**
	DCCA (ICML'13)	63.61	61.72	57.3	38.89	37.23	33.75	71.92	72.33	71.54	96.22	96.34	96.08
	DCCAE (ICML'15)	50.42	48.84	46.48	38.61	37.53	34.03	72.00	71.65	70.63	96.45	96.37	96.08
	LMSC (CVPR'17)	51.28	51.08	48.99	53.92	51.25	42.80	56.09	55.53	54.99	–	–	–
	MvC-DMF (AAAI'17)	43.07	42.45	40.48	48.27	46.71	40.53	42.97	43.08	76.45	75.83	74.05	49.79
	BMVC (TPAMI'18)	66.32	65.16	61.73	**58.57**	**55.69**	**49.92**	**78.65**	**78.20**	**77.73**	92.45	92.47	92.05
	AE^2-Nets (CVPR'19)	**72.03**	**69.76**	**64.66**	35.24	34.38	31.72	65.47	64.82	63.28	89.74	89.33	87.90
Partially	CCA (NeurIPS'03)	**52.49**	**51.52**	**47.95**	35.72	34.56	31.47	64.73	64.70	63.91	65.53	65.05	64.07
	KCCA (JMLR'02)	50.49	48.82	45.75	32.87	31.29	28.90	64.06	63.95	62.83	57.08	56.63	56.06
	DCCA (ICML'13)	51.68	50.64	46.85	35.72	33.97	31.20	**65.92**	**65.80**	**65.15**	60.95	60.90	60.16
	DCCAE (ICML'15)	46.24	45.37	43.75	31.95	30.75	28.14	61.88	61.58	60.67	47.42	47.17	46.26
	LMSC (CVPR'17)	39.15	38.10	36.88	**45.21**	**43.51**	**38.02**	45.03	44.76	44.49	–	–	–
	MvC-DMF (AAAI'17)	36.74	36.42	34.71	20.78	20.08	18.93	41.59	41.34	41.27	33.04	32.64	32.03
	BMVC (TPAMI'18)	50.35	49.83	46.39	33.56	32.83	30.09	64.69	64.20	63.27	72.49	72.03	70.92
	AE^2-Nets (CVPR'19)	48.19	47.64	42.61	23.30	22.65	20.61	62.74	62.40	60.65	**76.58**	**75.87**	**73.75**
Partially	PVC (NeurIPS'20)	48.77	45.97	40.46	36.78	36.50	35.54	72.63	72.08	71.11	93.09	93.12	93.06
	MvCLN (Mean)	**57.93**	**57.15**	**55.52**	**46.69**	**45.89**	**43.87**	**81.77**	**81.63**	**81.11**	**96.19**	**96.18**	**96.15**

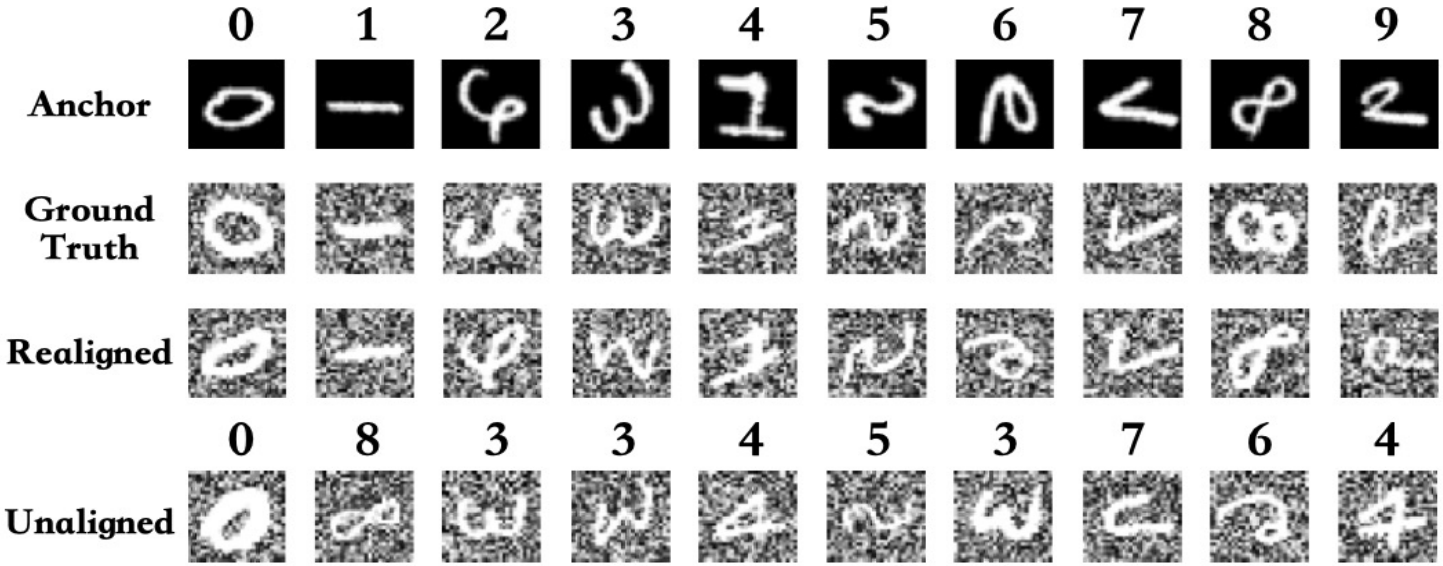
Experiment

- t-SNE visualization on the Noisy MNIST dataset



(a) MvCLN (b) PVC (c) AE^2-Nets

- Visualization of the reestablished correspondences on NoisyMNIST

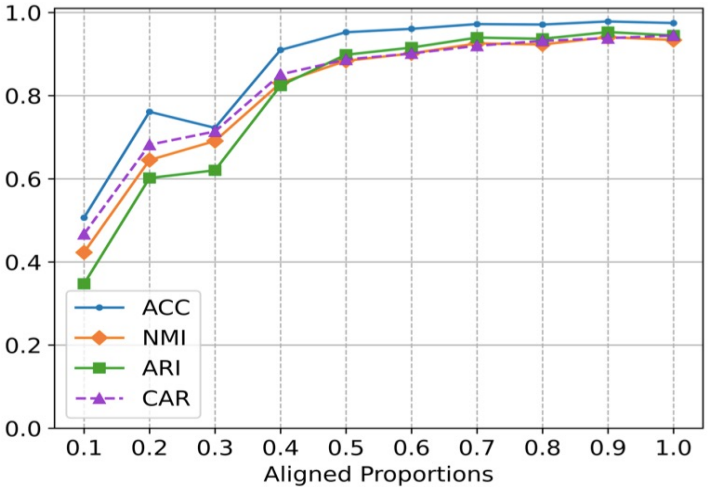


Experiment

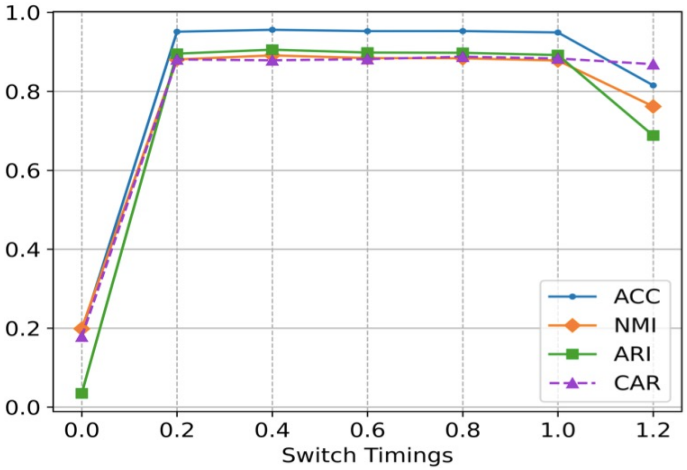
❖ Time consumption

Dataset	Method	training time (s)	inferring time (s)
Scene-15	Hungarian	–	2.69
	PVC	10,907.27	2.01
	MvCLN	**155.53**	**0.72**
Caltech-101	Hungarian	–	48.87
	PVC	11,839.74	7.2
	MvCLN	**388.58**	**1.75**
Reuters	Hungarian	–	289.82
	PVC	18,715.34	30.36
	MvCLN	**790.30**	**3.48**
NoisyMNIST	Hungarian	–	3,778.39
	PVC	53,070.87	34.36
	MvCLN	**1,202.77**	**5.76**

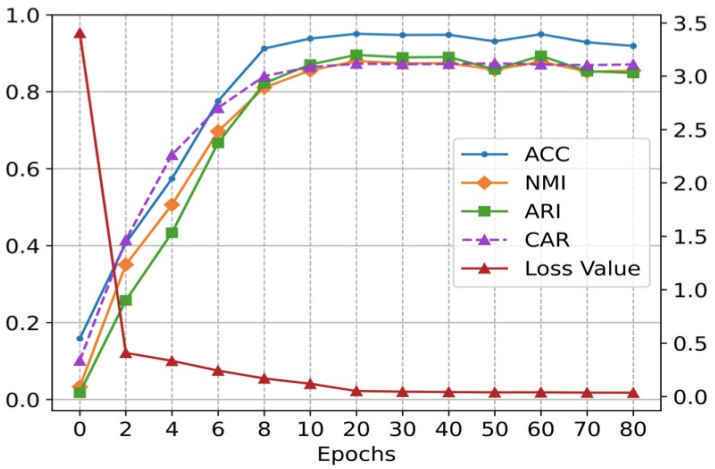
❖ Performance w.r.t. aligned proportions



❖ Influence of switching times



❖ Convergence analysis



Learning with Noisy Correspondence for Cross-modal Matching

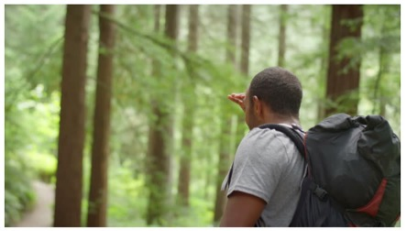
Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, **Xi Peng***

NeurIPS 2021 (Oral， acc rate=0.6%)

Background

Many applications highly rely on the data correspondence

Cross-modal Retrieval



A man hiking through the woods looks around the forest and then smiles at the camera , in slow motion

Visual Grounding

- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

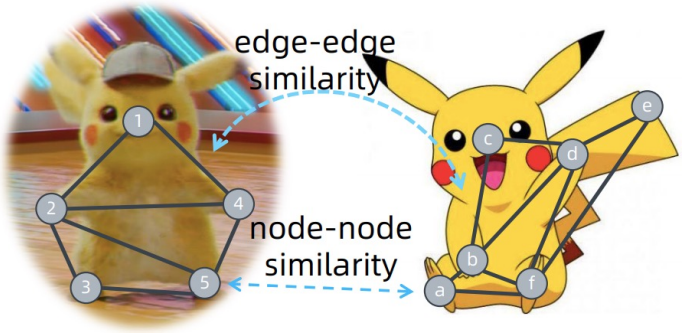
Data correspondence

Machine Reading Comprehension

Q: When did Reginald Eppes wake up?
A: Five in the morning

Q: What was the first thing he checked?
A: the weather forecast

Graph Matching



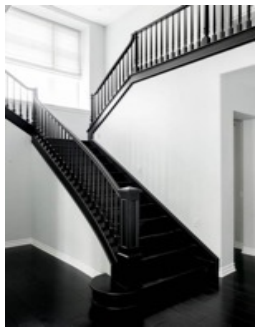
Motivation

Noisy pairs are common in the real world

Read between the lines , and your dream about person will be clear

A large crowd turned out for show .

There is no need to be sad.



See pictures of first home .

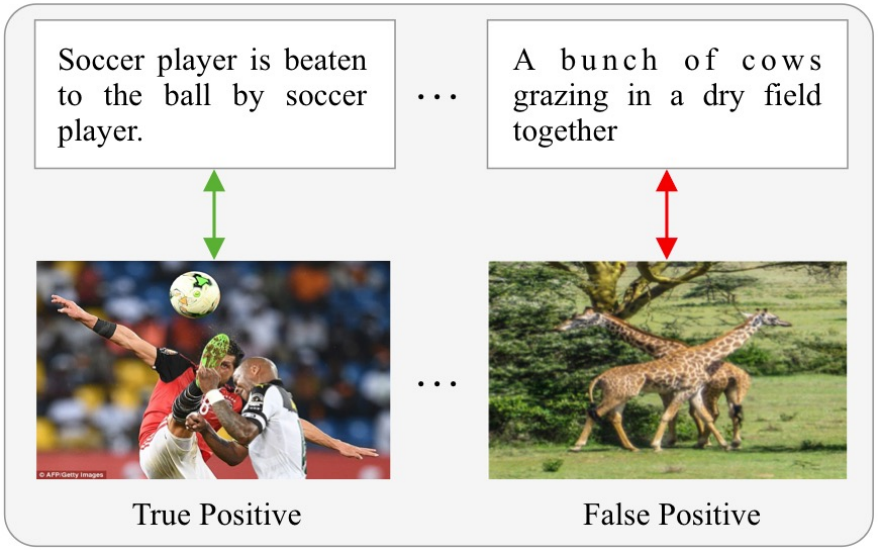
WRONGLY matched image-text pairs from Conceptual Captions dataset*

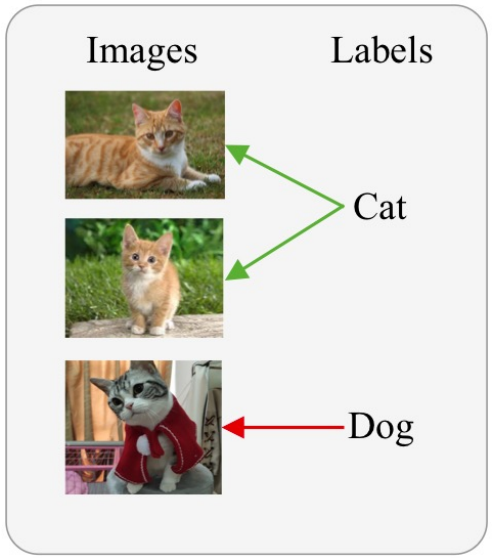
*Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning, ACL

Motivation

We reveal and define the mismatched pairs as Noisy Correspondence (NC)

Noisy Correspondence **VS** Noisy Label

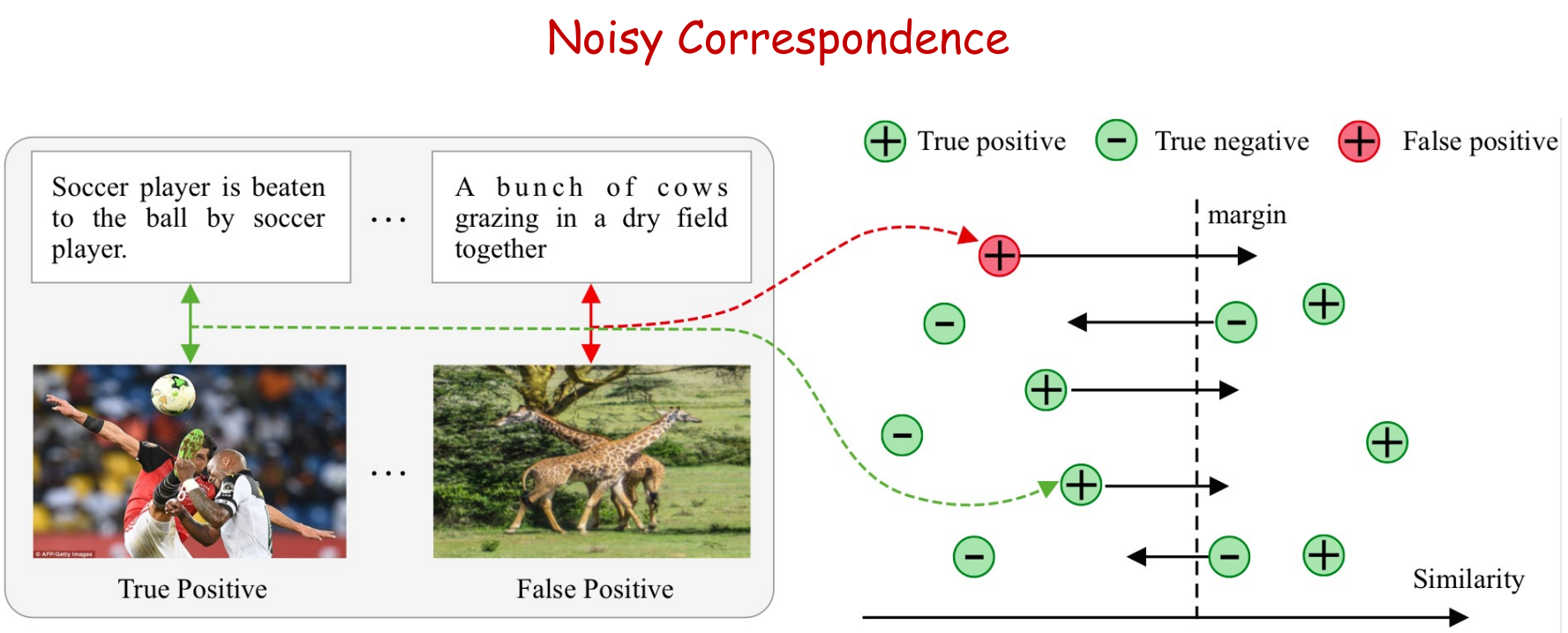




NC refers to the alignment errors in paired data rather than the errors in category annotations

Motivation

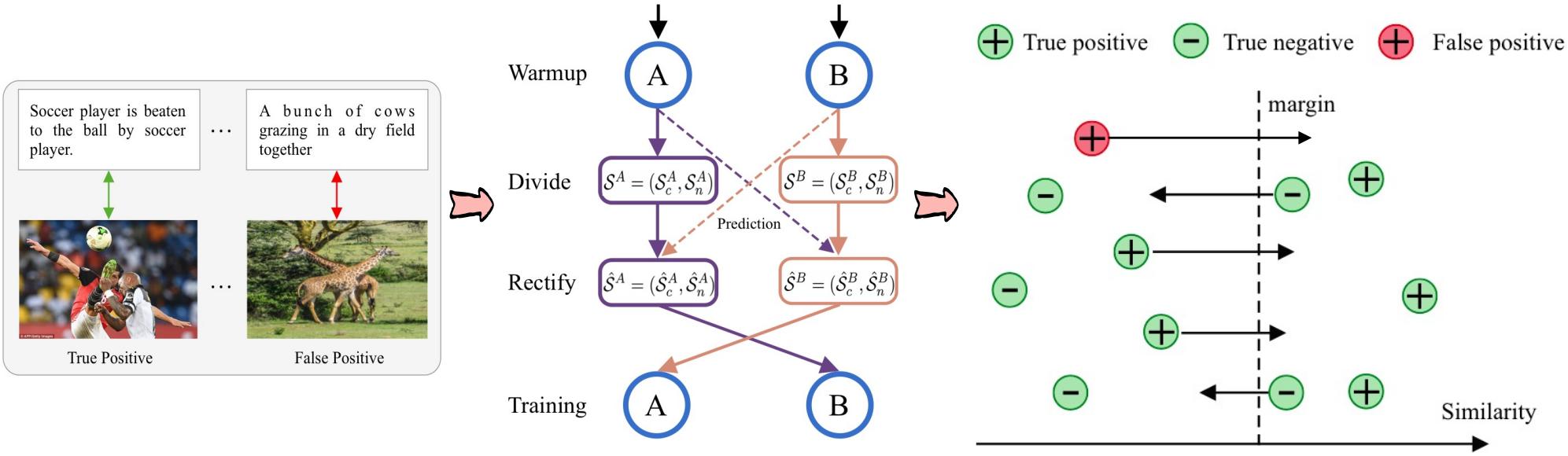
Noisy Correspondence will degrade the performance of various tasks including cross-modal matching



The false positives from the noisy correspondence vastly degrade the matching performance

The Proposed Solution to NC

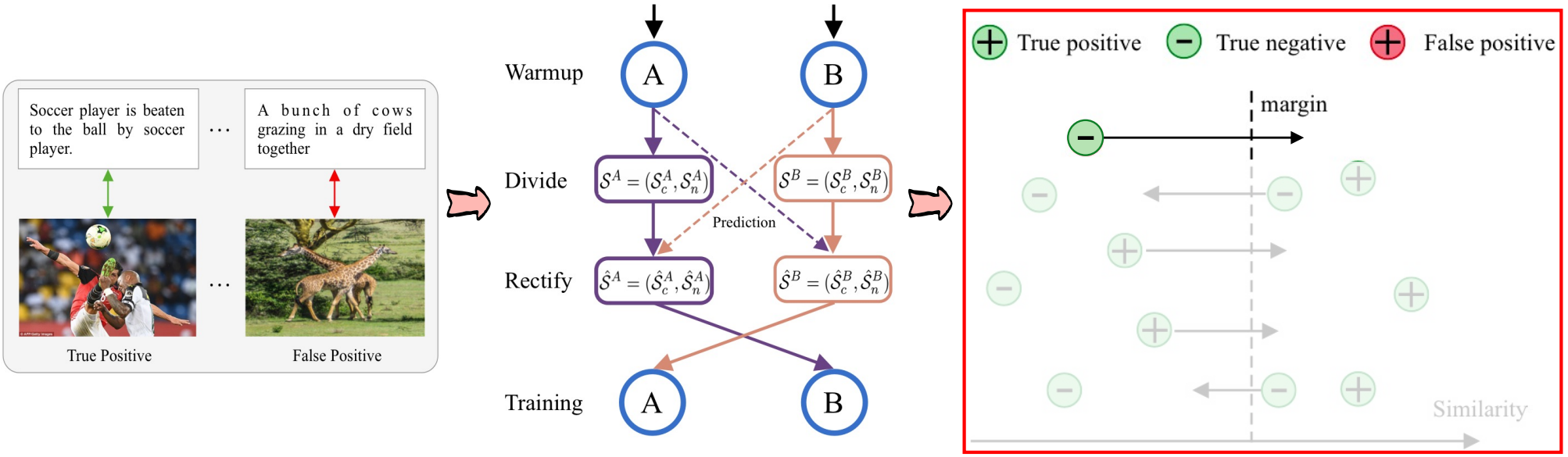
Noisy Correspondence Rectifier (NCR)



NCR divides the data into clean and noisy partitions based on the memorization effect of neural networks and then rectifies the correspondence via an adaptive prediction model in a co-teaching manner. Finally, NCR achieves robust matching with soft-margin based loss.

The Proposed Solution to NC

Noisy Correspondence Rectifier (NCR)

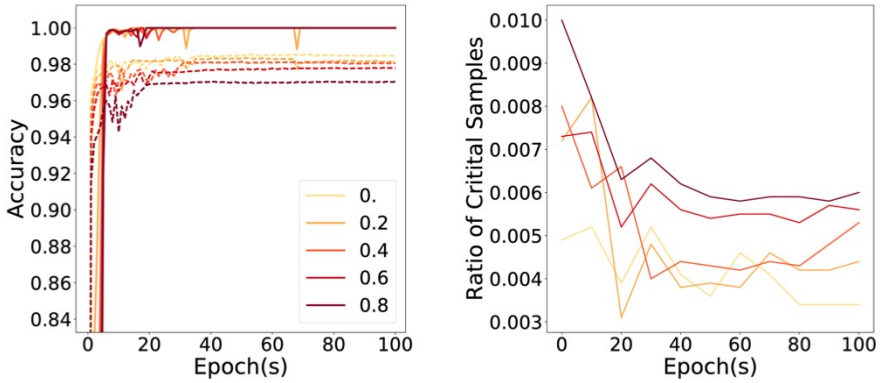


NCR divides the data into clean and noisy partitions based on the memorization effect of neural networks and then rectifies the correspondence via an adaptive prediction model in a co-teaching manner. Finally, NCR achieves robust matching with soft-margin based loss.

Co-divide

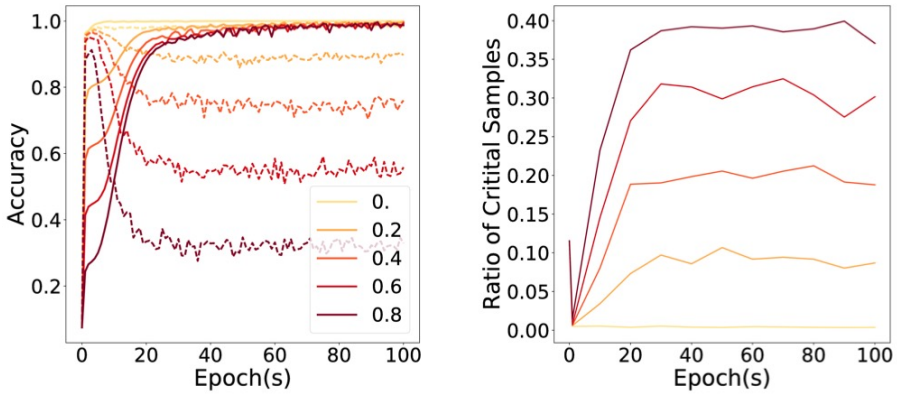
DNN Memorization Effect*

Solid is train, dotted is validation



(a) Noise added on classification inputs.

Solid is train, dotted is validation



(b) Noise added on classification labels.

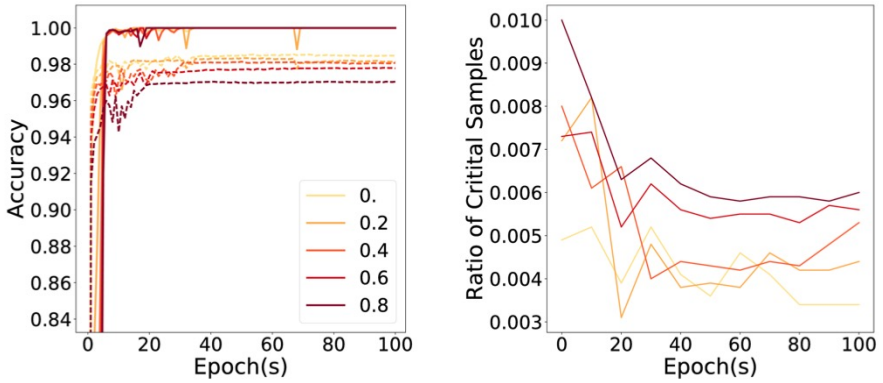
The DNN models first learn the simple and general patterns of the real data before fitting the noise.

*Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... & Lacoste-Julien, S. (2017, July). A closer look at memorization in deep networks. In *International Conference on Machine Learning* (pp. 233-242). PMLR.

Co-divide

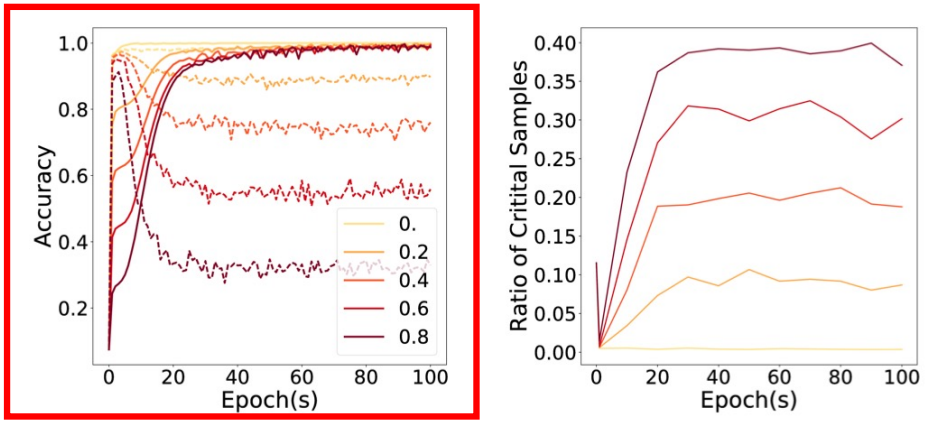
DNN Memorization Effect*

Solid is train, dotted is validation



(a) Noise added on classification inputs.

Solid is train, dotted is validation



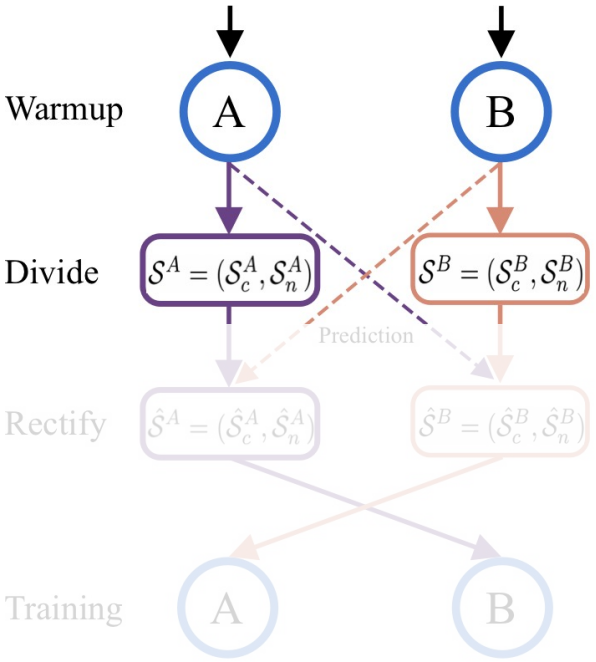
(b) Noise added on classification labels.

The DNN models first learn the simple and general patterns of the real data before fitting the noise.

*Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., ... & Lacoste-Julien, S. (2017, July). A closer look at memorization in deep networks. In *International Conference on Machine Learning* (pp. 233-242). PMLR.

Co-divide

Co-divide the noisy data via the memorization effects

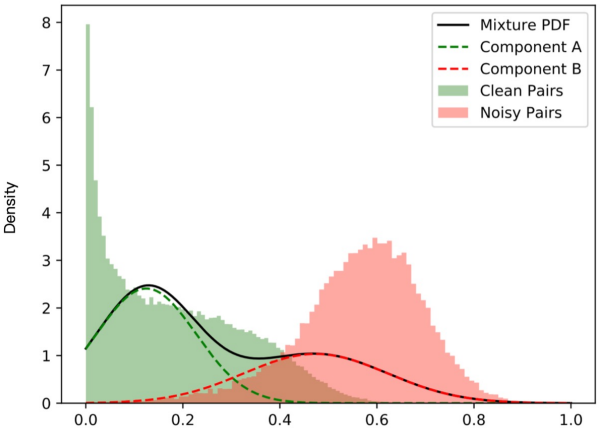


- Model Warmup

$$L_w(I,T) = \sum_{\hat{T}}[\alpha - S(I,T) + S(I,\hat{T})]_+ + \sum_{\hat{I}}[\alpha - S(I,T) + S(\hat{I},T)]_+,$$

- Gaussian Mixture Model

$$p(\ell|\theta) = \sum_{k=1}^{K} \beta_k \phi(\ell|k)$$

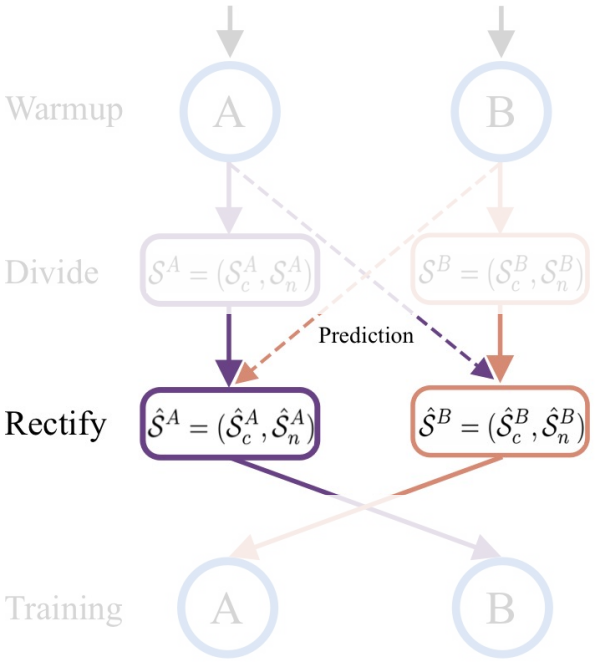


- Confidence

$$w_i = p(k|\ell_i) = p(k)p(\ell_i|k)/p(\ell_i)$$

Co-rectify

Co-rectify the correspondence via model predictions

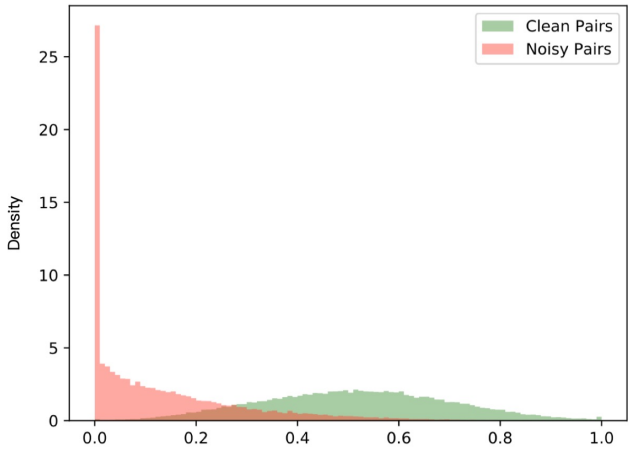


- Model Prediction

$$P(I,T) = \Theta(s)/\tau$$

$$s = S(I,T) - (\frac{1}{b}\sum_{\hat{T}} S(I,\hat{T}) - \frac{1}{b}\sum_{\hat{I}} S(\hat{I},T))/2$$

- Correspondence Refinement

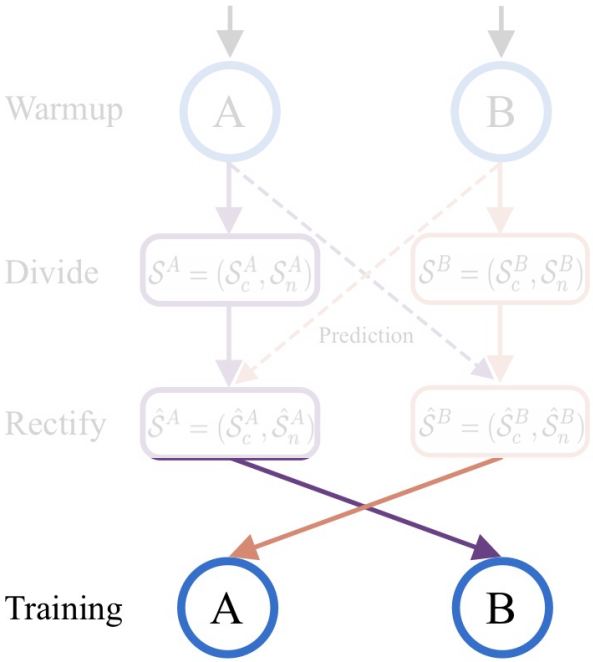


$$\hat{y}_i^c = w_i y_i^c + (1 - w_i) P^k(I_i^c, T_i^c),$$

$$\hat{y}_i^n = (P^A(I_i^n, T_i^n) + P^B(I_i^n, T_i^n))/2,$$

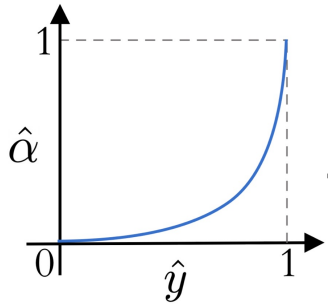
Robust Matching Loss

Soft margin based Triplet loss

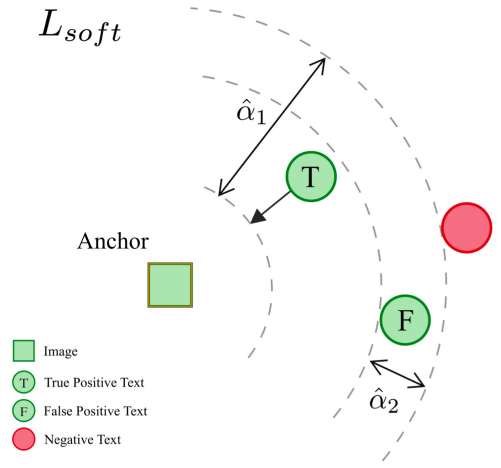


- Refined correspondence > Soft margin

$$\hat{\alpha}_i = \frac{m^{\hat{y}_i}}{m-1}\alpha$$



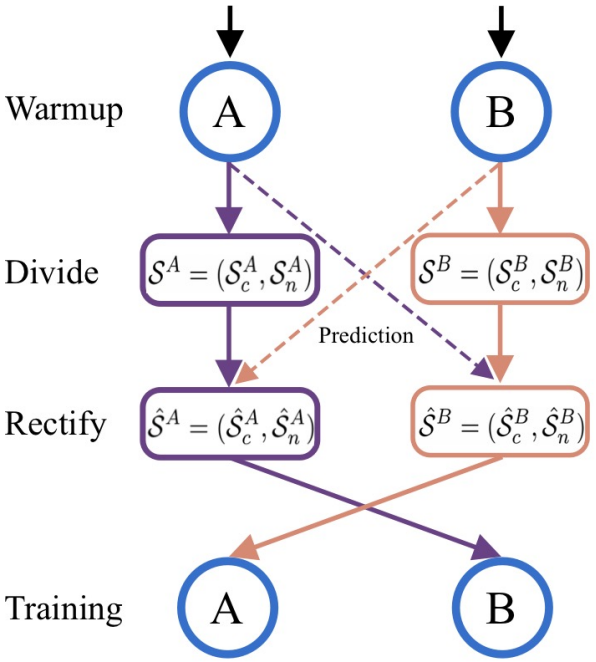
- Soft Triplet loss



$$L_{soft}(I_i, T_i) = [\hat{\alpha}_i - S(I_i, T_i) + S(I_i, \hat{T}_h)]_+ + [\hat{\alpha}_i - S(I_i, T_i) + S(\hat{I}_h, T_i)]_+$$

All Together

Noisy Correspondence Rectifier



- Divide the data into two relatively accurate data partitions based on their loss difference:

$$\mathcal{S}^A = (\mathcal{S}_c^A, \mathcal{S}_n^A) \qquad \mathcal{S}^B = (\mathcal{S}_c^B, \mathcal{S}_n^B)$$

- Rectify the correspondence via an adaptive prediction function:

$$\hat{\mathcal{S}}^A = (\hat{\mathcal{S}}_c^A, \hat{\mathcal{S}}_n^A) \qquad \hat{\mathcal{S}}^B = (\hat{\mathcal{S}}_c^B, \hat{\mathcal{S}}_n^B)$$

- Train the matching model with novel triplet loss by recasting the rectified correspondence as the soft margin:

$$L_{soft}(I_i, T_i) = [\hat{\alpha}_i - S(I_i, T_i) + S(I_i, \hat{T}_h)]_+ \qquad \hat{\alpha}_i = \frac{m^{\hat{y}_i}}{m-1}\alpha$$
$$+ [\hat{\alpha}_i - S(I_i, T_i) + S(\hat{I}_h, T_i)]_+$$

Experiments: MS-COCO and Flickr30K

Baseline: SCAN(ECCV18), VSRN(ICCV18), IMRAM(CVPR20), SGRAF(AAA21)

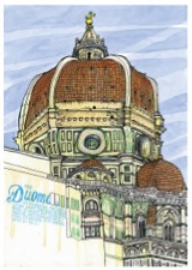
		Flickr30K						MS-COCO					
		Image → Text			Text → Image			Image → Text			Text → Image		
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0%	SCAN	67.4	90.3	95.8	48.6	77.7	85.2	69.2	93.6	97.6	56.0	86.5	93.5
	VSRN	71.3	90.6	96.0	54.7	81.8	88.2	71.3	90.6	96.0	54.7	81.8	88.2
	IMRAM	74.1	93.0	96.6	53.9	79.4	87.2	76.7	95.6	98.5	61.7	89.1	95.0
	SAF	73.7	93.3	96.3	56.1	81.5	88.0	76.1	95.4	98.3	61.8	89.4	95.3
	SGR	75.2	93.3	96.6	56.2	81.0	86.5	78.0	95.8	98.2	61.4	89.3	95.4
	SGRAF	**77.8**	**94.1**	97.4	58.5	83.0	88.8	**79.6**	**96.2**	**98.5**	63.2	**90.7**	**96.1**
	NCR	77.3	94.0	**97.5**	**59.6**	**84.4**	**89.9**	78.7	95.8	**98.5**	**63.3**	90.4	95.8
20%	SCAN	59.1	83.4	90.4	36.6	67.0	77.5	66.2	91.0	96.4	45.0	80.2	89.3
	VSRN	58.1	82.6	89.3	40.7	68.7	78.2	25.1	59.0	74.8	17.6	49.0	64.1
	IMRAM	63.0	86.0	91.3	41.4	71.2	80.5	68.6	92.8	97.6	55.7	85.0	91.0
	SAF	51.0	79.3	88.0	38.3	66.5	76.2	67.3	92.5	96.6	53.4	84.5	92.4
	SGR*	62.8	86.2	92.2	44.4	72.3	80.4	67.8	91.7	96.2	52.9	83.5	90.1
	SGR-C	72.8	90.8	95.4	56.4	82.1	88.6	75.4	95.2	97.9	60.1	88.5	94.8
	NCR	**75.0**	**93.9**	**97.5**	**58.3**	**83.0**	**89.0**	**77.7**	**95.5**	**98.2**	**62.5**	**89.3**	**95.3**
50%	SCAN	27.7	57.6	68.8	16.2	39.3	49.8	40.8	73.5	84.9	5.4	15.1	21.0
	VSRN	14.3	37.6	50.0	12.1	30.0	39.4	23.5	54.7	69.3	16.0	47.8	65.9
	IMRAM	9.1	26.6	38.2	2.7	8.4	12.7	21.3	60.2	75.9	22.3	52.8	64.3
	SAF	30.3	63.6	75.4	27.9	53.7	65.1	30.4	67.8	82.3	33.5	69.0	82.8
	SGR*	36.9	68.1	80.2	29.3	56.2	67.0	60.6	87.4	93.6	46.0	74.2	79.0
	SGR-C	69.8	90.3	94.8	50.1	77.5	85.2	71.7	94.1	97.7	57.0	86.6	93.7
	NCR	**72.9**	**93.0**	**96.3**	**54.3**	**79.8**	**86.5**	**74.6**	**94.6**	**97.8**	**59.1**	**87.8**	**94.5**

Experiments: CC152K from CC dataset

Comparison:

Methods	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN (ECCV'18)	30.5	55.3	65.3	26.9	53.0	64.7
VSRN (ICCV'19)	32.6	61.3	70.5	32.5	59.4	70.4
IMRAM (CVPR'20)	33.1	57.6	68.1	29.0	56.8	67.4
SAF (AAAI'21)	31.7	59.3	68.2	31.9	59.0	67.9
SGR (AAAI'21)	11.3	29.7	39.6	13.1	30.1	41.6
SGR* (AAAI'21)	35.0	63.4	73.3	34.9	63.0	72.8
NCR	**39.5**	**64.5**	**73.5**	**40.3**	**64.6**	**73.2**

Detected Noisy Samples:



digital art selected for the #

look at him, like it 's no work at all

take a look at this !

share some with your friends !

family walking on a beach

Experiments: Comparison with CLIP

Comparison to large pretrained model (CLIP*):

Noise Ratio	Methods	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
20%	CLIP (ViT-B/32)	21.4	49.6	63.3	14.8	37.6	49.6
	NCR	**56.9**	**83.6**	**91.0**	**40.6**	**69.8**	**80.1**
50%	CLIP (ViT-B/32)	10.9	27.8	38.3	7.8	19.5	26.8
	NCR	**53.1**	**80.7**	**88.5**	**37.9**	**66.6**	**77.8**

- Although CLIP utilizes 400 million image-text pairs for pretraining, its performance inevitably degenerates during finetuning.
- NCR achieves the promising matching performance with the presence of noisy correspondence, indicating the necessity of algorithm design.

*Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Experiments: Comparison with CLIP

Comparison to large pretrained model (CLIP*):

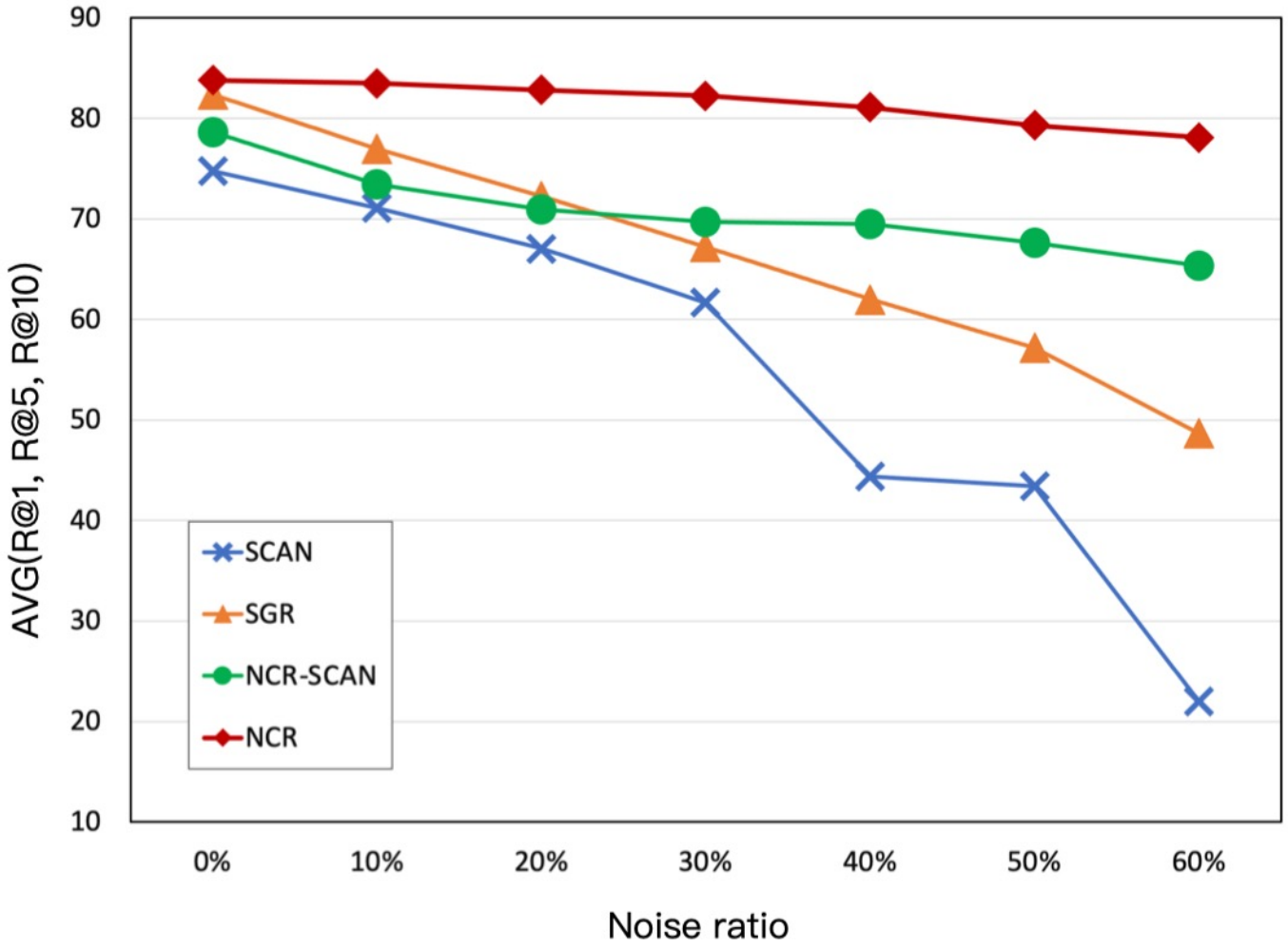
Noise Ratio	Methods	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
20%	CLIP (ViT-B/32)	21.4	49.6	63.3	14.8	37.6	49.6
	NCR	**56.9**	**83.6**	**91.0**	**40.6**	**69.8**	**80.1**
50%	CLIP (ViT-B/32)	10.9	27.8	38.3	7.8	19.5	26.8
	NCR	**53.1**	**80.7**	**88.5**	**37.9**	**66.6**	**77.8**

Having a ton of noisy image-text data isn't enough and could be further boost by handling the possible noise.

*Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

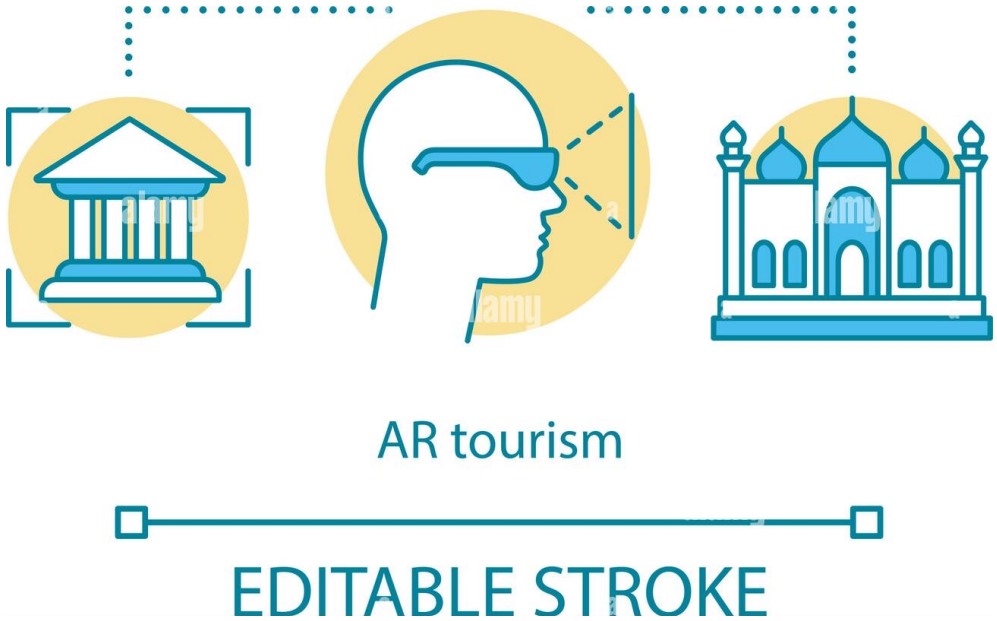
Experiments

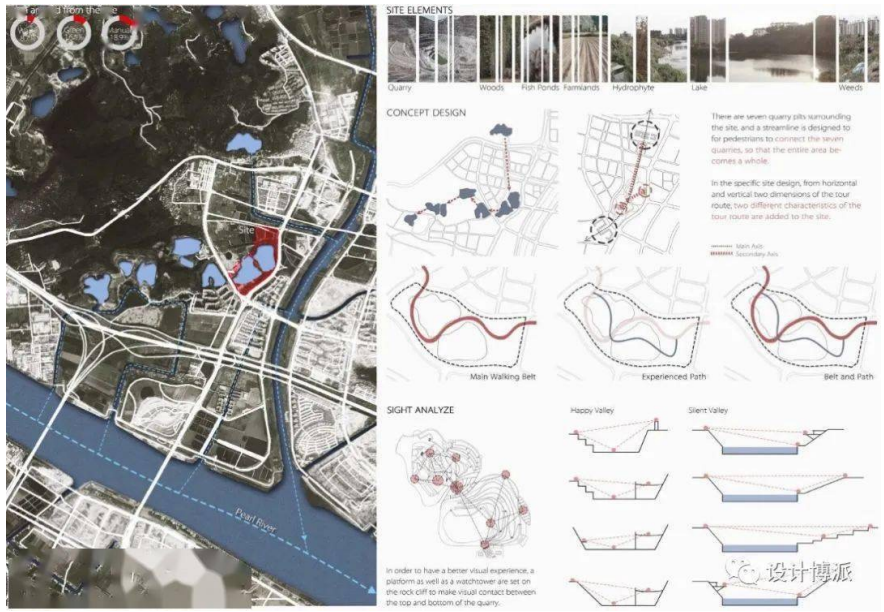
Generalization to different models and noise ratios:



Conclusion

- Reveal a new paradigm for the noisy labels, i.e., noisy correspondence which is totally different from existing noisy label learning;
- Noisy correspondence is general to many techniques of intelligent tourism, including but not limited to cross-modal retrieval, VQA, visual grounding, visual navigation, Re-ID, and so on;





All codes could be downloaded at www.pengxi.me

谢谢，

敬请批评指正！