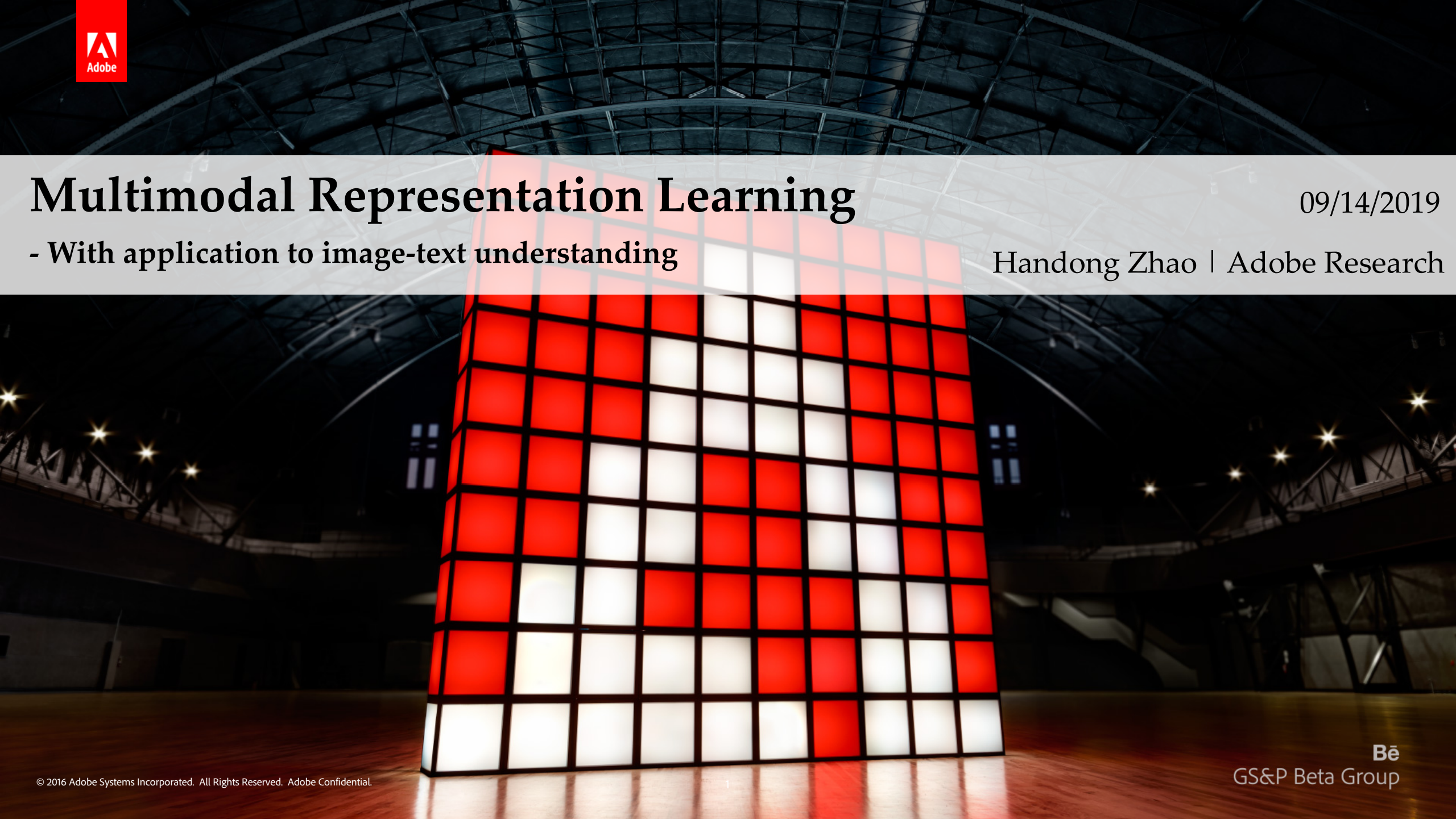


Multimodal Representation Learning

09/14/2019

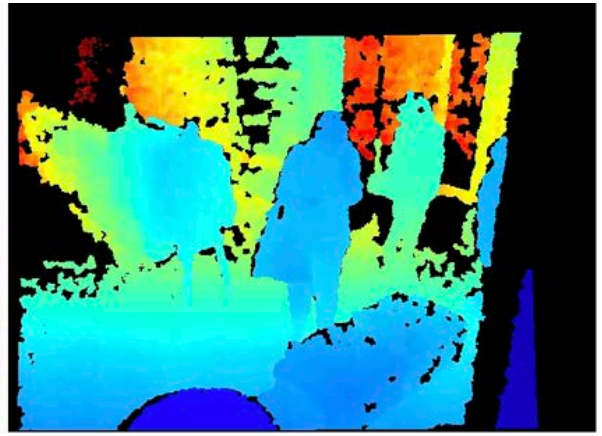
- With application to image-text understanding

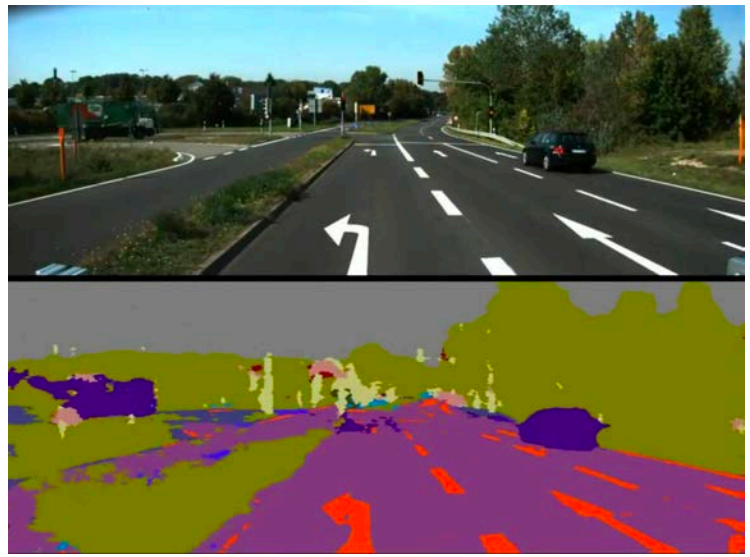
Handong Zhao | Adobe Research

Bē
GS&P Beta Group

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.



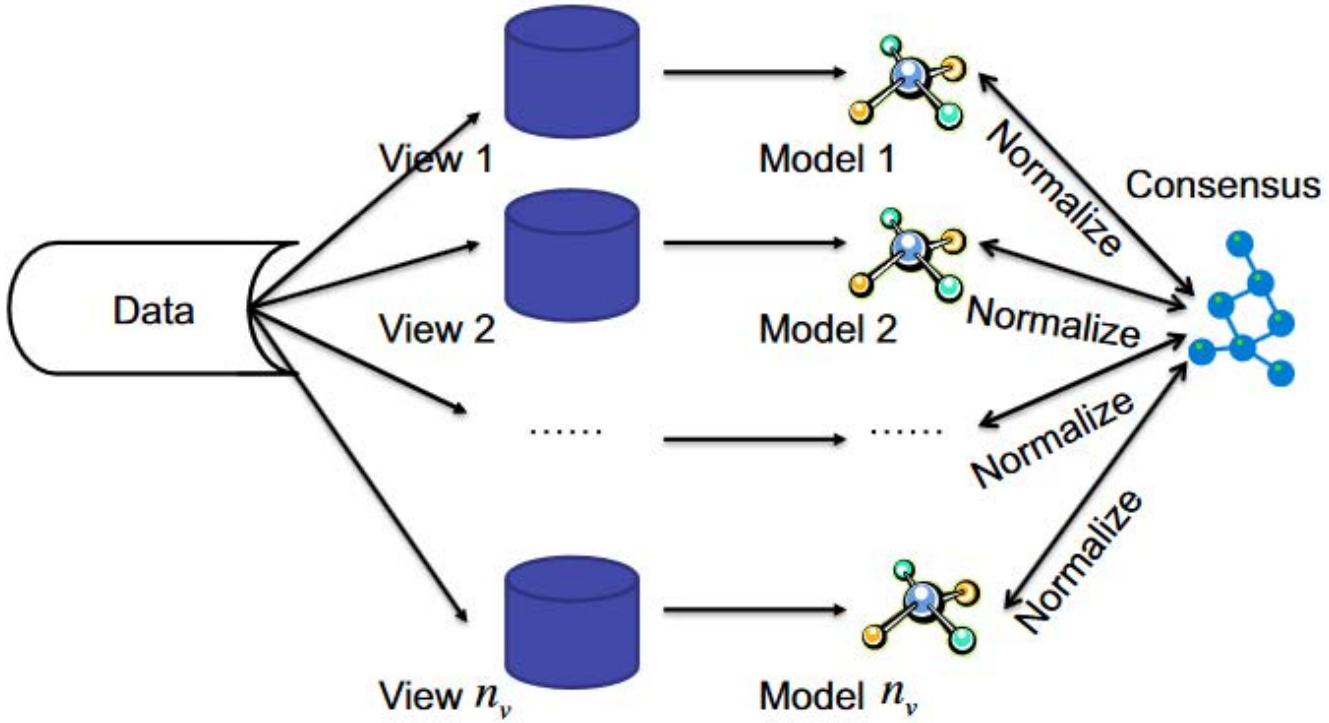


Outline

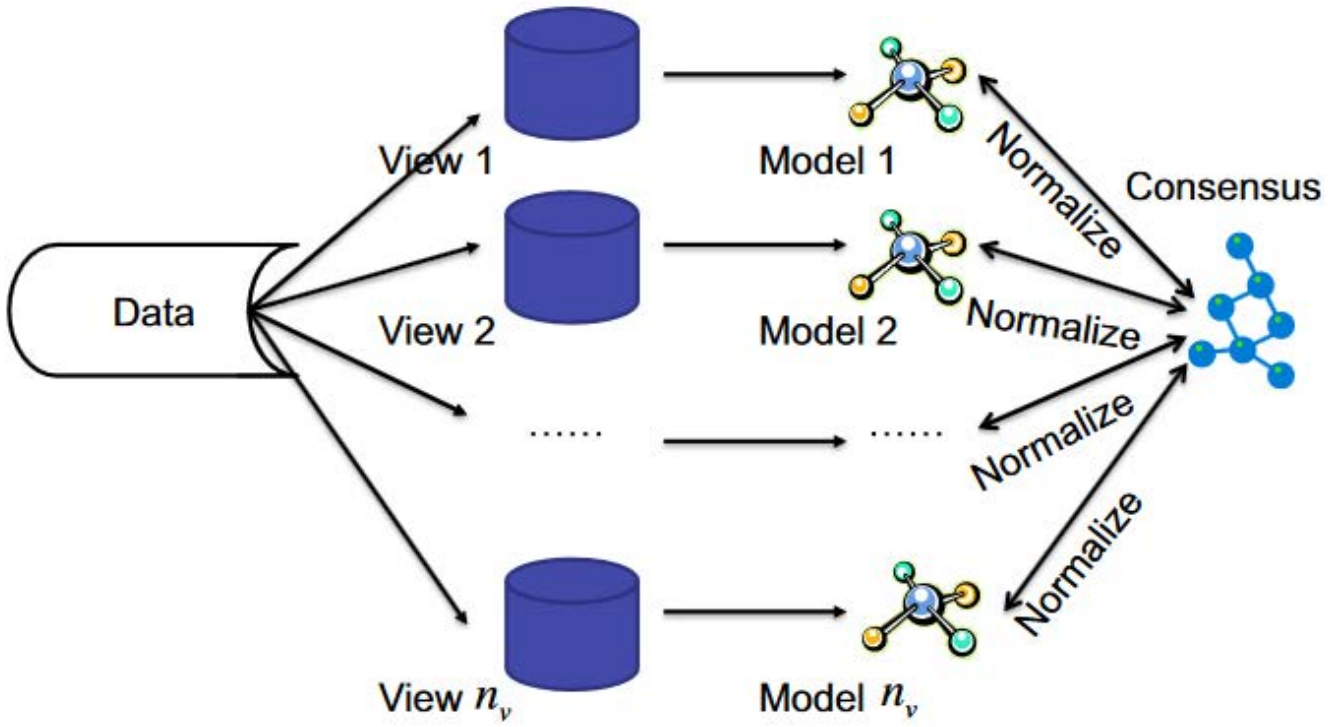
1. Motivation
2. Methodology:
   - *Complete*-modality Case
   - *Incomplete*-modality Case
3. Applications on Image-text Understanding:
   - Scene Graph Generation
   - Image Captioning
4. Conclusion

[IJCAI'16] Handong Zhao, Hongfu Liu and Yun Fu, "Incomplete Multimodal Visual Data Grouping", in IJCAI, 2016.
[AAAI'17] Handong Zhao, Zhengming Ding and Yun Fu, "Multi-view Clustering via Deep Matrix Factorization", in AAAI, 2017.
[TIP'17] Handong Zhao, Hongfu Liu, Zhengming Ding, and Yun Fu, "Consensus Regularized Multi-view Outlier Detection", IEEE TIP, 2017.
[CVPR'19] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai and Mingyang Ling, "Scene Graph Generation with External Knowledge and Image Reconstruction", CVPR, 2019.
[ICCV'19] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang, "Unpaired ImageCaptioning via SceneGraph Alignments", ICCV, 2019.

Problem Statement



Problem Statement



X	Original input data
U	Projection for each view data
$V^{(*)}$	Consensus coefficient in latent space
λ	Weight for each view

weight different projections

$$\sum_{v=1}^{n_v} \lambda_v \left\| X^{(v)} - U^{(v)} \left(V^{(*)}\right)^T \right\|_F^2$$

input multi-modal data

consensus coefficient in latent space

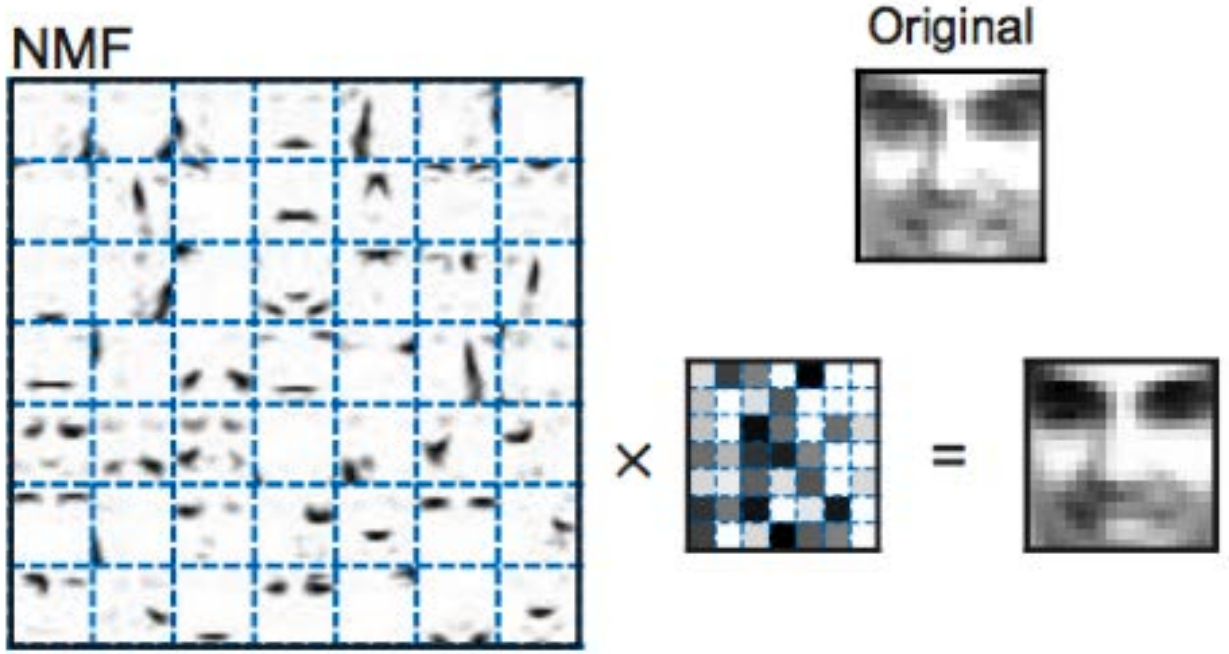
[Liu et al., 2013] Jialu Liu, Chi Wang, Jing Gao, Jiawei Han: Multi-view Clustering via Joint Nonnegative Matrix Factorization. SDM 2013: 252-260

Methodology:

Multimodal Representation Learning in *Complete*-modality Case

Preliminary Knowledge

NMF: Non-negative Matrix Factorization [Lee and Seung, 1999]

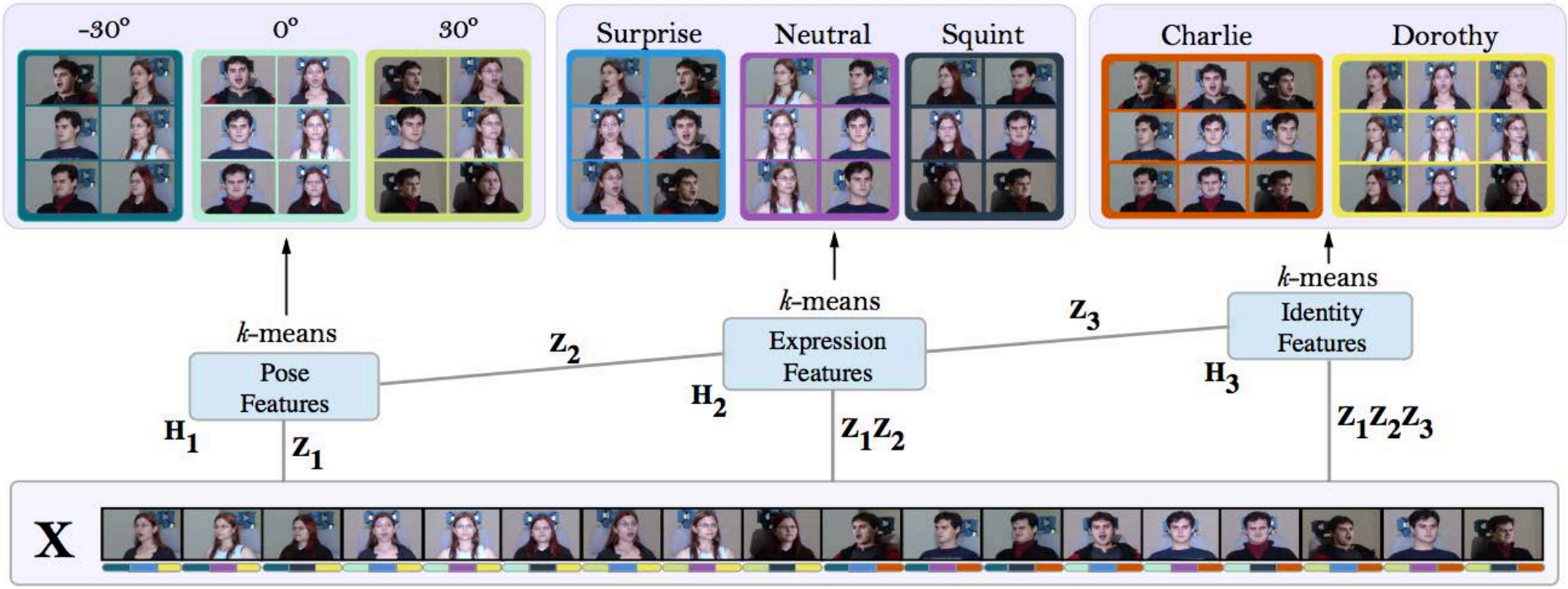


$$\min_{Z,H} \|X - ZH\|_{\mathrm{F}}^2$$
$$\text{s.t. } Z \geq 0,\ H \geq 0$$

[Lee and Seung, 1999] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, 1999

Methodology - *Complete* Modality Case

(Deep) Multi-layer NMF [Trigeorgis et al. 2014, Zhao et al., 2015]



Layer-wise formulation:

$$\boldsymbol{X}^{\pm} \approx \boldsymbol{Z}_1^{\pm}\boldsymbol{H}_1^{+}$$

$$\boldsymbol{X}^{\pm} \approx \boldsymbol{Z}_1^{\pm}\boldsymbol{Z}_2^{\pm}\boldsymbol{H}_2^{+}$$

$$\boldsymbol{X}^{\pm} \approx \boldsymbol{Z}_1^{\pm}\boldsymbol{Z}_2^{\pm}\boldsymbol{Z}_3^{\pm}\boldsymbol{H}_3^{+}$$

$$\Rightarrow \quad \boldsymbol{X}^{\pm} \approx \boldsymbol{Z}_1^{\pm}\boldsymbol{Z}_2^{\pm}\cdots\boldsymbol{Z}_m^{\pm}\boldsymbol{H}_m^{+}$$

[Trigeorgis et al. 2014] Trigeorgis, G.; Bousmalis, K.; Zafeiriou, S.; and Schuller, B. W. A deep semi-nmf model for learning hidden representations. In ICML, 2014

[Zhao et al. 2015] Handong Zhao, Zhengming Ding, Ming Shao and Yun Fu, "Robust Semi-Nonnegative Coding for Semi-supervised Learning", in ICDM, 2015.

Methodology - *Complete* Modality Case

Proposed objective function for complete multimodal setting:

Weighting parameters

Deep Semi-NMF on each view

Consensus graph regularizer

$$\min_{\substack{Z_i^{(v)}, H_i^{(v)} \\ H_m, \alpha^{(v)}}} \sum_{v=1}^{V} (\alpha^{(v)})^{\gamma} \left(\| X^{(v)} - Z_1^{(v)} Z_2^{(v)} \dots Z_m^{(v)} H_m \|_{\mathrm{F}}^2 + \beta \mathrm{tr}(H_m L^{(v)} H_m^{\mathrm{T}}) \right)$$

$$\text{s.t. } H_i^{(v)} \geq 0,\ H_m \geq 0,\ \sum_{v=1}^{V} \alpha^{(v)} = 1,\ \alpha^{(v)} \geq 0$$

Methodology - *Complete* Modality Case

Proposed objective function for complete multimodal setting:

Deep Semi-NMF
on each view

$$\min_{\substack{Z_i^{(v)}, H_i^{(v)} \\ H_m, \alpha^{(v)}}} \sum_{v=1}^{V} (\alpha^{(v)})^{\gamma} \left(\| X^{(v)} - Z_1^{(v)} Z_2^{(v)} \dots Z_m^{(v)} H_m \|_{\mathrm{F}}^2 + \beta \mathrm{tr}(H_m L^{(v)} H_m^{\mathrm{T}}) \right)$$

$$\text{s.t. } H_i^{(v)} \geq 0,\ H_m \geq 0,\ \sum_{v=1}^{V} \alpha^{(v)} = 1,\ \alpha^{(v)} \geq 0$$

<u>Remark 1:</u> Due to the homology of multi-view data, the final layer representation $H_m^{(v)}$ for v-th view data should be close to each other.

Methodology - *Complete* Modality Case

Proposed objective function for complete multimodal setting:

Consensus graph regularizer

$$\min_{\substack{Z_i^{(v)}, H_i^{(v)} \\ H_m, \alpha^{(v)}}} \sum_{v=1}^{V} (\alpha^{(v)})^{\gamma} \left(\| X^{(v)} - Z_1^{(v)} Z_2^{(v)} \dots Z_m^{(v)} H_m \|_{\mathrm{F}}^2 + \beta \mathrm{tr}(H_m L^{(v)} H_m^{\mathrm{T}}) \right)$$

$$\text{s.t. } H_i^{(v)} \geq 0,\ H_m \geq 0,\ \sum_{v=1}^{V} \alpha^{(v)} = 1,\ \alpha^{(v)} \geq 0$$

<u>Remark 2:</u> Multiple graphs are constructed to constrain the common representation learning so that the geometric structure in each view could be well preserved for the final clustering.

Methodology - *Complete* Modality Case

Proposed objective function for complete multimodal setting:

Weighting parameters

$$\min_{\substack{Z_i^{(v)}, H_i^{(v)} \\ H_m, \alpha^{(v)}}} \sum_{v=1}^{V} (\alpha^{(v)})^{\gamma} \left(\|X^{(v)} - Z_1^{(v)} Z_2^{(v)} \dots Z_m^{(v)} H_m\|_{\mathrm{F}}^2 + \beta \mathrm{tr}(H_m L^{(v)} H_m^{\mathrm{T}}) \right)$$

$$\text{s.t. } H_i^{(v)} \geq 0,\ H_m \geq 0,\ \sum_{v=1}^{V} \alpha^{(v)} = 1,\ \alpha^{(v)} \geq 0$$

Remark 3: With **only one** parameter γ, we could control the different weights for different views.

Methodology - *Complete* Modality Case

Experiments

Dataset:

a) Yale
b) Extended Yale B
c) Notting-Hill [Cao et al. 2015]

Baseline:

a) BestSV:
b) ConcatFea:
c) ConcatPCA:
d) Co-Reg (SPC) [Kumar et al. 2011]
e) Co-Training [Kumar et al. 2011]
f) Min-D(isagreement) [de Sa 2005]
g) MultiNMF [Liu et al. 2013]
h) NaMSC [Cao et al. 2015]
i) DiMSC [Cao et al. 2015]

Metric:

a) Normalized Mutual Information (NMI)
b) Accuracy (ACC)
c) Adjusted Rand Index (AR)
d) F-score
e) Precision
f) Recall.

Yale

Notting-Hill

Adobe

Methodology - *Complete* Modality Case

Table 1: Results (Mean ± standard deviation) on dataset Yale.

Method	NMI	ACC	AR	F-score	Precision	Recall
BestSV	0.654 ± 0.009	0.616 ± 0.030	0.440 ± 0.011	0.475 ± 0.011	0.457 ± 0.011	0.495 ± 0.010
ConcatFea	0.641 ± 0.006	0.544 ± 0.038	0.392 ± 0.009	0.431 ± 0.008	0.415 ± 0.007	0.448 ± 0.008
ConcatPCA	0.665 ± 0.037	0.578 ± 0.038	0.396 ± 0.011	0.434 ± 0.011	0.419 ± 0.012	0.450 ± 0.009
Co-Reg	0.648 ± 0.002	0.564 ± 0.000	0.436 ± 0.002	0.466 ± 0.000	0.455 ± 0.004	0.491 ± 0.003
Co-Train	0.672 ± 0.006	0.630 ± 0.001	0.452 ± 0.010	0.487 ± 0.009	0.470 ± 0.010	0.505 ± 0.007
Min-D	0.645 ± 0.005	0.615 ± 0.043	0.433 ± 0.006	0.470 ± 0.006	0.446 ± 0.005	0.496 ± 0.006
MultiNMF	0.690 ± 0.001	0.673 ± 0.001	0.495 ± 0.001	0.527 ± 0.000	0.512 ± 0.000	0.543 ± 0.000
NaMSC	0.671 ± 0.011	0.636 ± 0.000	0.475 ± 0.004	0.508 ± 0.007	0.492 ± 0.003	0.524 ± 0.004
DiMSC	0.727 ± 0.010	0.709 ± 0.003	0.535 ± 0.001	0.564 ± 0.002	0.543 ± 0.001	0.586 ± 0.003
Ours	**0.782** ± 0.010	**0.745** ± 0.011	**0.579** ± 0.002	**0.601** ± 0.002	**0.598** ± 0.001	**0.613** ± 0.002

Table 3: Results (Mean ± standard deviation) on dataset Notting-Hill.

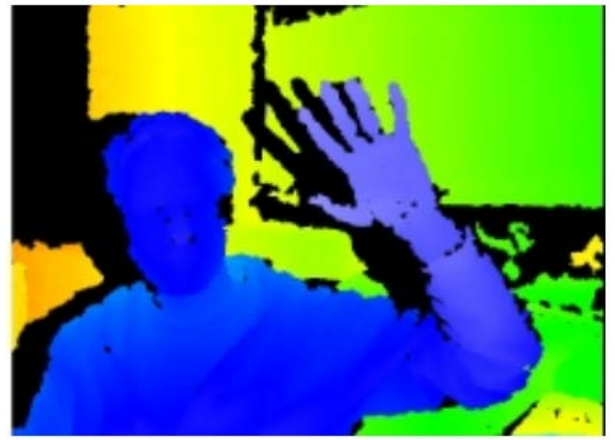
Method	NMI	ACC	AR	F-score	Precision	Recall
BestSV	0.723 ± 0.008	0.813 ± 0.000	0.712 ± 0.020	0.775 ± 0.015	0.774 ± 0.018	0.776 ± 0.013
ConcatFea	0.628 ± 0.028	0.673 ± 0.033	0.612 ± 0.041	0.696 ± 0.032	0.699 ± 0.032	0.693 ± 0.031
ConcatPCA	0.632 ± 0.009	0.733 ± 0.008	0.598 ± 0.015	0.685 ± 0.012	0.691 ± 0.010	0.680 ± 0.014
Co-Reg	0.660 ± 0.003	0.758 ± 0.000	0.616 ± 0.004	0.699 ± 0.000	0.705 ± 0.003	0.694 ± 0.003
Co-Train	0.766 ± 0.005	0.689 ± 0.027	0.589 ± 0.035	0.677 ± 0.026	0.688 ± 0.030	0.667 ± 0.023
Min-D	0.707 ± 0.003	0.791 ± 0.000	0.689 ± 0.002	0.758 ± 0.002	0.750 ± 0.002	0.765 ± 0.003
MultiNMF	0.752 ± 0.001	0.831 ± 0.001	0.762 ± 0.000	0.815 ± 0.000	0.804 ± 0.001	0.824 ± 0.001
NaMSC	0.730 ± 0.002	0.752 ± 0.013	0.666 ± 0.004	0.738 ± 0.005	0.746 ± 0.002	0.730 ± 0.011
DiMSC	**0.799** ± 0.001	0.843 ± 0.021	0.787 ± 0.001	0.834 ± 0.001	0.822 ± 0.005	0.836 ± 0.009
Ours	0.797 ± 0.005	**0.871** ± 0.009	**0.803** ± 0.002	**0.847** ± 0.002	**0.826** ± 0.007	**0.870** ± 0.001

Methodology:

Multimodal Representation Learning in *Incomplete*-modality Case

Methodology - *Incomplete* Modality Case

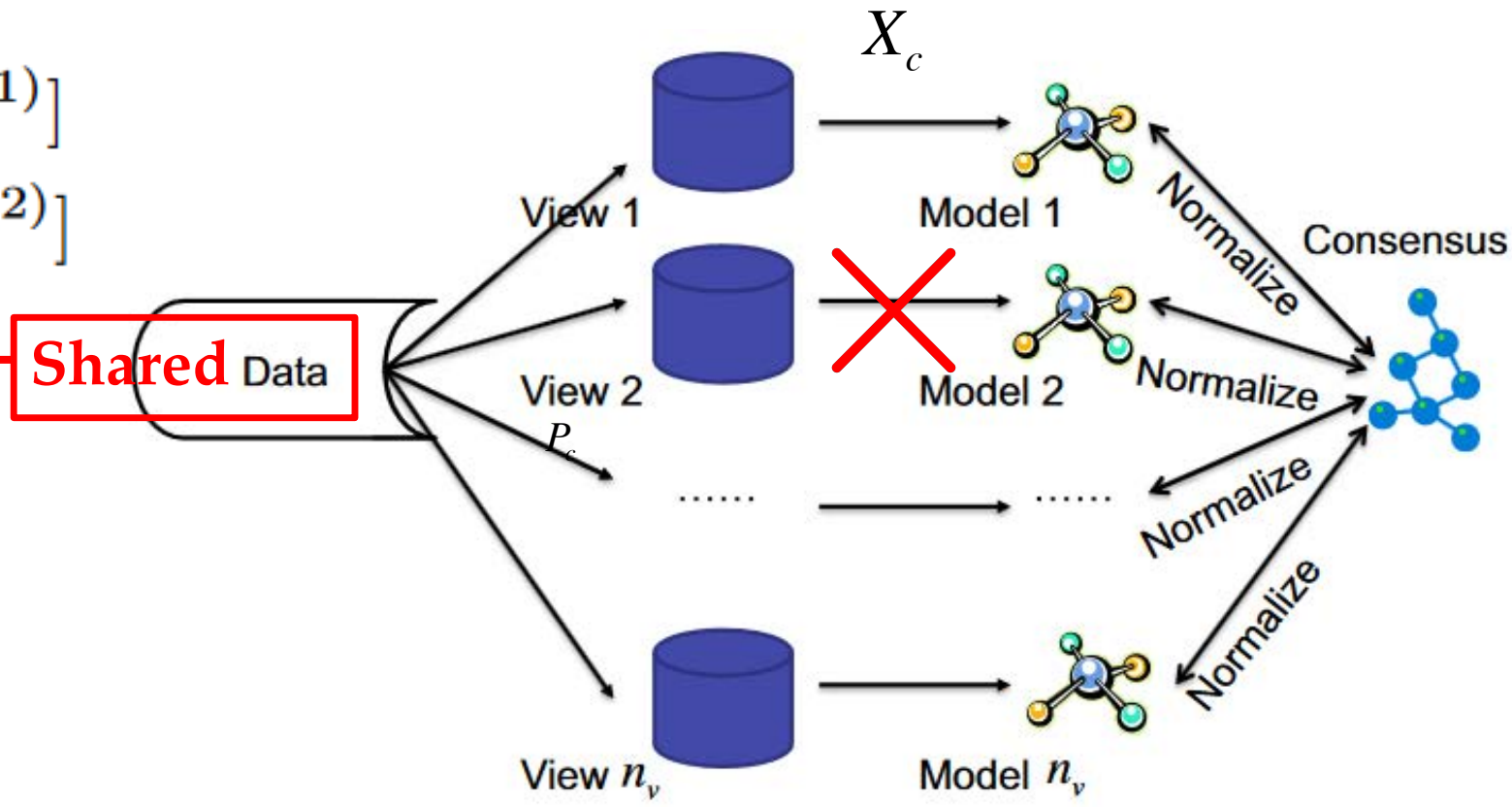
When the data from one modality/more modalities are inaccessible because of sensor failure or other reasons, most MVC methods would inevitably degenerate or even fail.



Methodology - *Incomplete* Modality Case

Unlike the traditional MVC, the numbers of data samples in different view are not the same. However, we assume partial of the data has all the view info.

$$\bar{\mathbf{X}}^{(1)} = [\mathbf{X}_c^{(1)}; \hat{\mathbf{X}}^{(1)}]$$
$$\bar{\mathbf{X}}^{(2)} = [\mathbf{X}_c^{(2)}; \hat{\mathbf{X}}^{(2)}]$$



Methodology - *Incomplete* Modality Case

Unlike the traditional MVC, the numbers of data samples in different view are not the same. However, we assume partial of the data has all the view info.

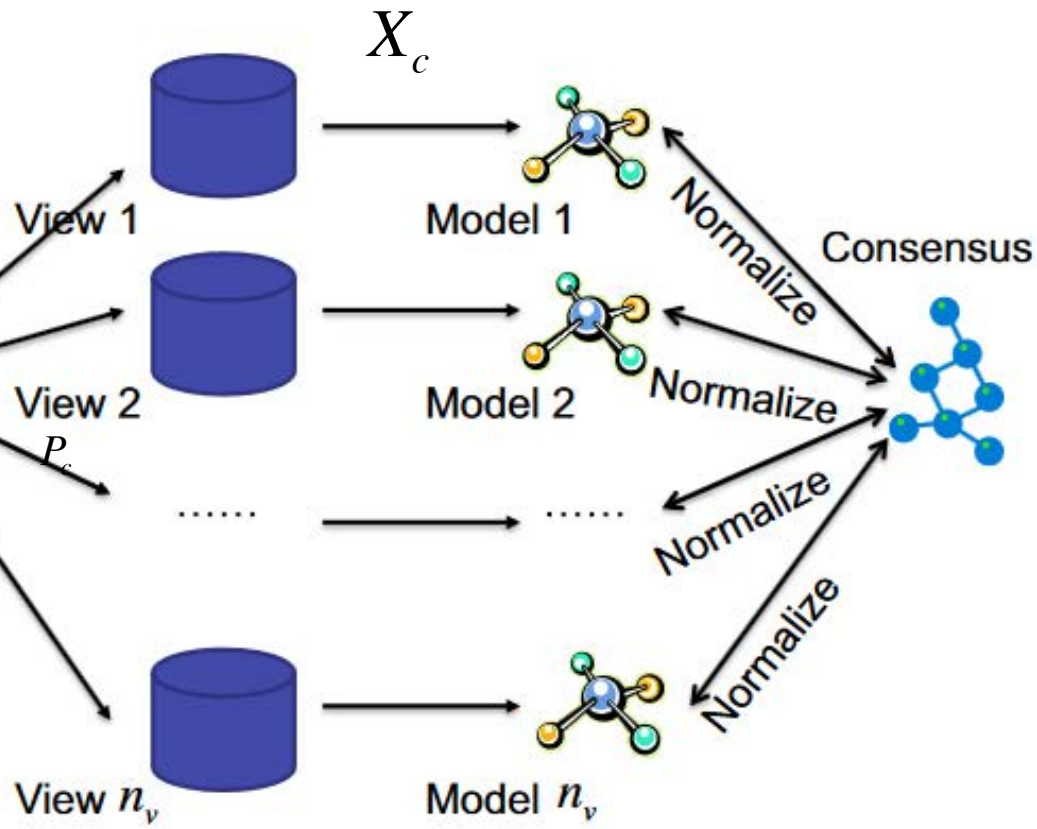
$$\bar{\mathbf{X}}^{(1)} = [\mathbf{X}_c^{(1)}; \hat{\mathbf{X}}^{(1)}]$$

$$\bar{\mathbf{X}}^{(2)} = [\mathbf{X}_c^{(2)}; \hat{\mathbf{X}}^{(2)}]$$

Shared Data

$$O \equiv \left\| \begin{bmatrix} \mathbf{X}_c^{(1)} \\ \hat{\mathbf{X}}^{(1)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(1)} \end{bmatrix} U^{(1)} \right\|_F^2$$

$$+ \left\| \begin{bmatrix} \mathbf{X}_c^{(2)} \\ \hat{\mathbf{X}}^{(2)} \end{bmatrix} - \begin{bmatrix} P_c \\ \hat{P}^{(2)} \end{bmatrix} U^{(2)} \right\|_F^2$$



Methodology - *Incomplete* Modality Case

- Challenge:

Only partial shared data is used to learn the projections for each view. The learned representation in the latent space is hard to preserve the good grouping structure when the shared data is insufficient.

Methodology - *Incomplete* Modality Case

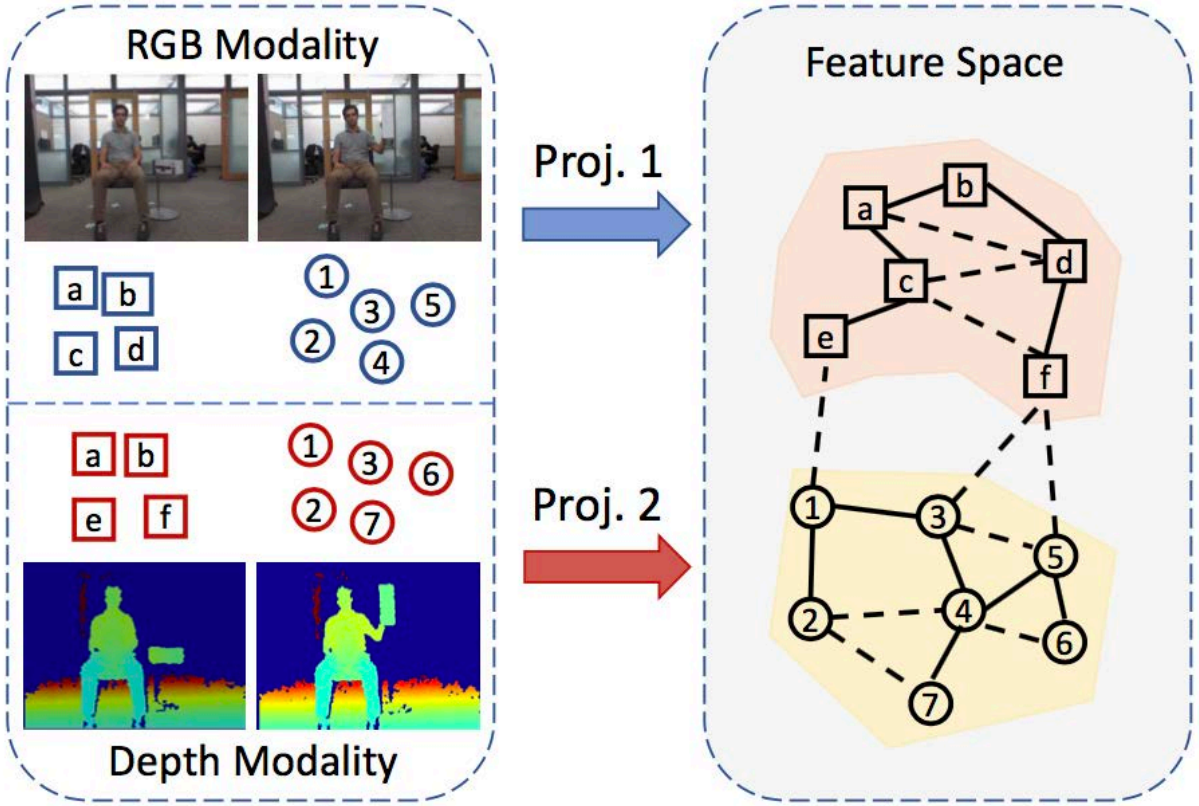
- Challenge:

Only partial shared data is used to learn the projections for each view. The learned representation in the latent space is hard to preserve the good grouping structure when the shared data is insufficient.

- Solution:

Graph Laplacian:

- Note that the similarity among data points is impossible measured in original feature space.
- Measure the similarity in the latent space as a graph, and jointly optimized this similarity matrix among data points with other variables.



Methodology - *Incomplete* Modality Case

The proposed objective function:

$$\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\U^{(1)},U^{(2)}}} \mathcal{P}roj(X,U)+\mathcal{R}(U,A)+\beta\mathrm{tr}(P^{\mathrm{T}}L_AP).$$

$$\text{s.t. } \forall i\ A_i^T\mathbf{1}=1,\ A_i\succeq 0,$$

Graph Laplacian term, where L_A is the Laplacian matrix of similarity matrix A, which is learned on the latent representation P.

Where, $P=[P_c;\hat{P}^{(1)};\hat{P}^{(2)}]$

$$\mathcal{P}roj(X,U)=\sum_{v\in\{1,2\}}\left\|\begin{bmatrix}X_c^{(v)}\\ \hat{X}^{(v)}\end{bmatrix}-\begin{bmatrix}P_c\\ \hat{P}^{(v)}\end{bmatrix}U^{(v)}\right\|_{\mathrm{F}}^2$$

$$\mathcal{R}(U,A)=\lambda\sum_{v\in\{1,2\}}\|U^{(v)}\|_{\mathrm{F}}^2+\gamma\|A\|_{\mathrm{F}}^2$$ ⇨ Regularizers to prevent the trivial solution.

Methodology - *Incomplete* Modality Case

The proposed objective function:

$$\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\U^{(1)},U^{(2)}}} \mathcal{P}roj(X,U)+\mathcal{R}(U,A)+\beta\mathrm{tr}(P^{\mathrm{T}}L_AP).$$

$$\text{s.t. } \forall i\ A_i^T\mathbf{1}=1,\ A_i\succeq 0,$$

Graph Laplacian term, where L_A is the Laplacian matrix of similarity matrix A, which is learned on the latent representation P.

Any problem with L_A??

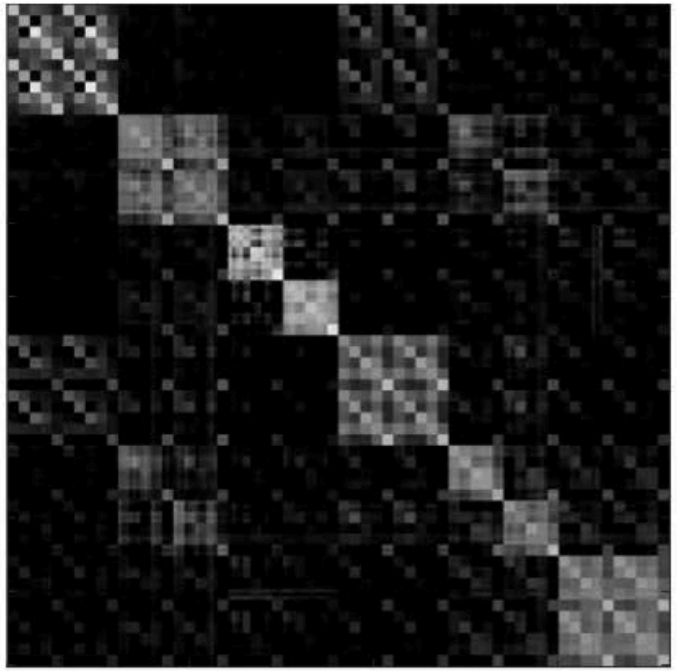
Where, $P=[P_c;\hat{P}^{(1)};\hat{P}^{(2)}]$

$$\mathcal{P}roj(X,U)=\sum_{v\in\{1,2\}}\left\|\begin{bmatrix}X_c^{(v)}\\ \hat{X}^{(v)}\end{bmatrix}-\begin{bmatrix}P_c\\ \hat{P}^{(v)}\end{bmatrix}U^{(v)}\right\|_{\mathrm{F}}^2$$

$$\mathcal{R}(U,A)=\lambda\sum_{v\in\{1,2\}}\|U^{(v)}\|_{\mathrm{F}}^2+\gamma\|A\|_{\mathrm{F}}^2$$

Methodology - *Incomplete* Modality Case

Visualization of the constructed graph A.



Methodology - *Incomplete* Modality Case

The proposed objective function:

$$\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\U^{(1)},U^{(2)}}} \mathcal{P}roj(X,U)+\mathcal{R}(U,A)+\beta\mathrm{tr}(P^{\mathrm{T}}L_AP).$$

$$\text{s.t. } \forall i\ A_i^T\mathbf{1}=1,\ A_i\succeq 0,$$

Graph Laplacian term, where L_A is the Laplacian matrix of similarity matrix A, which is learned on the latent representation P.

Where, $P=[P_c;\hat{P}^{(1)};\hat{P}^{(2)}]$

$$\mathcal{P}roj(X,U)=\sum_{v\in\{1,2\}}\left\|\begin{bmatrix}X_c^{(v)}\\\hat{X}^{(v)}\end{bmatrix}-\begin{bmatrix}P_c\\\hat{P}^{(v)}\end{bmatrix}U^{(v)}\right\|_{\mathrm{F}}^2$$

$$\mathcal{R}(U,A)=\lambda\sum_{v\in\{1,2\}}\|U^{(v)}\|_{\mathrm{F}}^2+\gamma\|A\|_{\mathrm{F}}^2$$

Problem: The constructed graph L_A is computed based on the Euclidean distance between two points in latent space, which is vulnerable to noises and random corruptions especially in the incomplete multi-view setting.

Methodology - *Incomplete* Modality Case

The proposed objective function:

$$\min_{\substack{P_c,\hat{P}^{(1)},\hat{P}^{(2)}\\U^{(1)},U^{(2)}}} \mathcal{P}roj(X,U)+\mathcal{R}(U,A)+\beta\mathrm{tr}(P^{\mathrm{T}}L_AP).$$

$$\text{s.t. } \forall i\ A_i^T\mathbf{1}=1,\ A_i\succeq 0,$$

Graph Laplacian term, where L_A is the Laplacian matrix of similarity matrix A, which is learned on the latent representation P.

Where, $P=[P_c;\hat{P}^{(1)};\hat{P}^{(2)}]$

$$\mathcal{P}roj(X,U)=\sum_{v\in\{1,2\}}\left\|\begin{bmatrix}X_c^{(v)}\\\hat{X}^{(v)}\end{bmatrix}-\begin{bmatrix}P_c\\\hat{P}^{(v)}\end{bmatrix}U^{(v)}\right\|_{\mathrm{F}}^2$$

$$\mathcal{R}(U,A)=\lambda\sum_{v\in\{1,2\}}\|U^{(v)}\|_{\mathrm{F}}^2+\gamma\|A\|_{\mathrm{F}}^2$$

Problem: The constructed graph L_A is computed based on the Euclidean distance between two points in latent space, which is vulnerable to noises and random corruptions especially in the incomplete multi-view setting.

Solution: Robust graph regularization.

Methodology - *Incomplete* Modality Case

Let's take a closer look at the graph Laplacian.

Applying eigen decomposition on $L_A = V_A^T S_A V_A$ and denoting $W_A = V_A^T S_A^{\frac{1}{2}}$, we have

$$\mathrm{tr}(P^{\mathrm{T}} L_A P) = \mathrm{tr}(P^{\mathrm{T}} V_A^{\mathrm{T}} S_A V_A P) = \|P^{\mathrm{T}} W_A\|_{\mathrm{F}}^2$$

If we replace the vulnerable Frobenius norm with **sparse** regularization:

$$\min_{\substack{P_c, \hat{P}^{(1)}, \hat{P}^{(2)} \\ U^{(1)}, U^{(2)}}} \mathcal{P}roj(X, U) + \mathcal{R}(U, A) + \beta \|P^{\mathrm{T}} W_A\|_1$$

$$\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1, \; A_i \succeq 0.$$

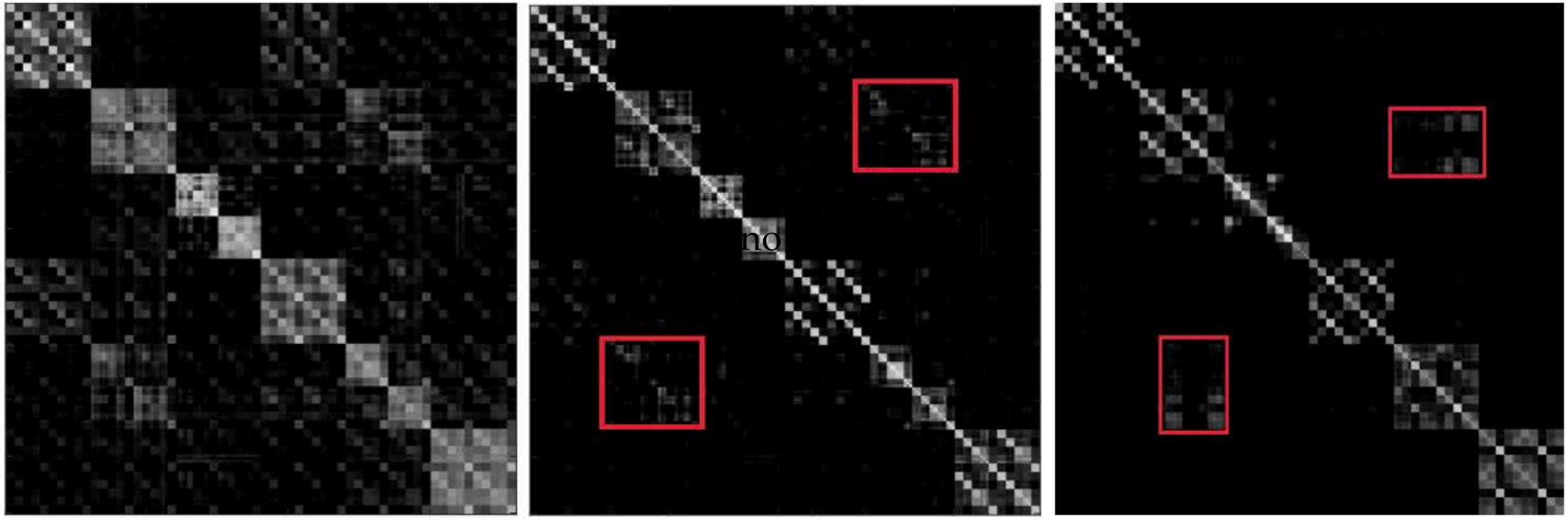
If we replace the vulnerable Frobenius norm with **low-rank** regularization:

$$\min_{\substack{P_c, \hat{P}^{(1)}, \hat{P}^{(2)} \\ U^{(1)}, U^{(2)}}} \mathcal{P}roj(X, U) + \mathcal{R}(U, A) + \beta \|P^{\mathrm{T}} W_A\|_*$$

$$\text{s.t. } \forall i \; A_i^T \mathbf{1} = 1, \; A_i \succeq 0,$$

Methodology - *Incomplete* Modality Case

Visualization of the constructed graph A.



(a) Ours-IMG (b) Ours-Sparse (c) Ours-LR

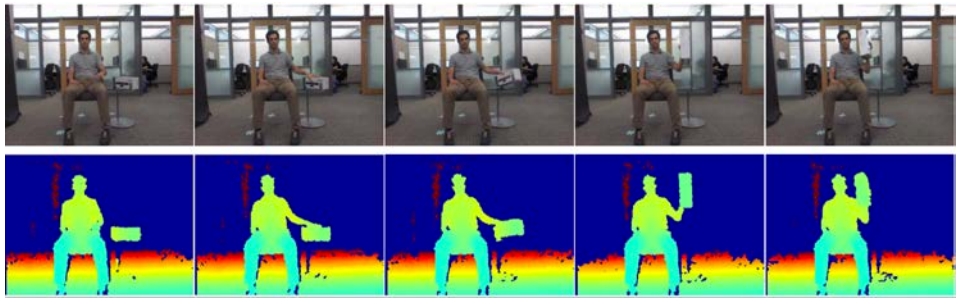
Visualization of the constructed graph A by (a) the probabilistic graph model, (b) the proposed sparse model and (c) the low-rank model. The first 6 actions in MSR Daily Activity are used.

Methodology - *Incomplete* Modality Case

Databases and settings:

Video:
- MSR Action Pairs dataset [Oreifej and Liu, 2013]
- MSR Daily Activity dataset [Wang et al., 2012]

Image:
- UCI handwritten digit [1]
- BUAA NirVis [Huang et al., 2012]



MSR Action Pairs Dataset

BUAA NirVis Dataset

For all the RGB-Depth video sequences, we temporally normalize each video clip to 10 frames with spatial resolution of 120 × 160. Histograms of gradient oriented feature is extracted from both depth and RGB videos with patch size 8 × 8. Thus, a total of 3000 patches are extracted from each video, with the feature dimensionality of 31.

[Oreifej and Liu, 2013] O. Oreifej and Z. Liu, "HON4D: histogram of oriented 4d normals for activity recognition from depth sequences," in CVPR, 2013, pp. 716–723.
[Wang et al., 2012] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in CVPR, 2012, pp. 1290–1297.
[Huang et al., 2012] D. Huang, J. Sun, , and Y. Wang, "The buaa-visnir face database instructions," IRIP-TR-12- FR-001, 2012.
1. http://archive.ics.uci.edu/ml/datasets.html

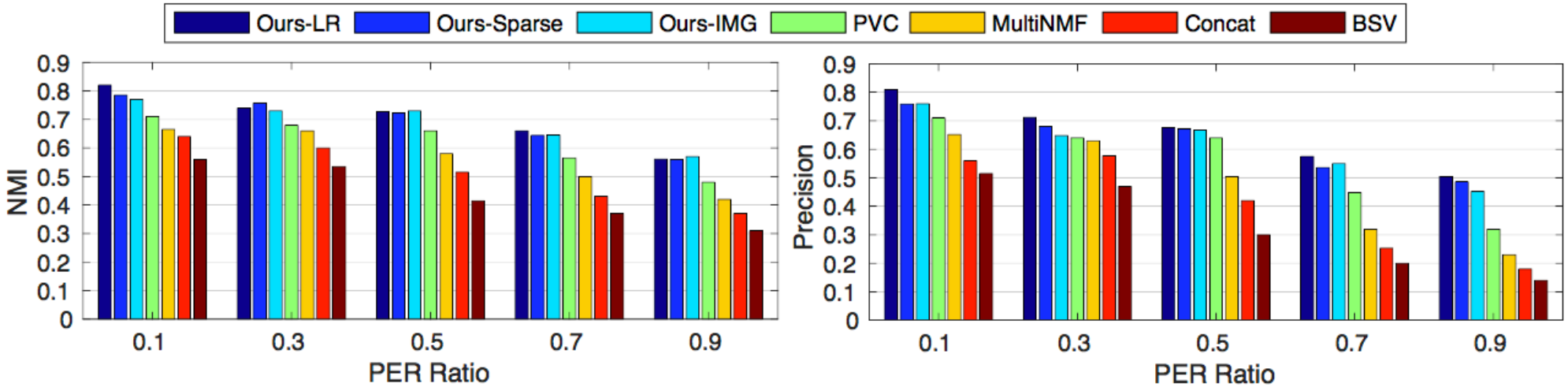
Methodology - *Incomplete* Modality Case

NMI/Precision results on *MSR Action Pairs* dataset (top) and *MSR Daily Activity* dataset (bottom) under different PER settings. The best two are highlighted in **bold**.

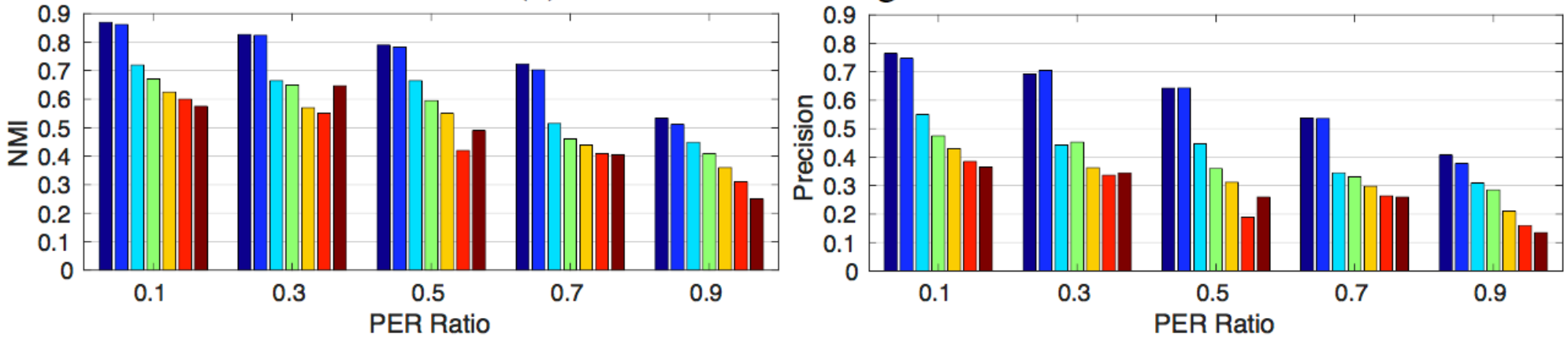
Method\PER	0.1	0.3	0.5	0.7	0.9
BSV	0.4807 / 0.2687	0.4807 / 0.2687	0.3691 / 0.1660	0.2874 / 0.1190	0.2779 / 0.1085
Concat	0.6270 / 0.3538	0.5803 / 0.3306	0.5512 / 0.3030	0.5123 / 0.2750	0.4685 / 0.2268
MultiNMF	0.6033 / 0.4038	0.5149 / 0.2984	0.5008 / 0.2828	0.4816 / 0.2539	0.4463 / 0.2267
PVC	0.6917 / 0.4490	0.6501 / 0.3998	0.6356 / 0.3734	0.6012 / 0.3662	0.5882 / 0.3629
Ours-IMG	0.6859 / 0.4504	0.6763 / 0.4431	0.6504 / 0.3836	0.6468 / 0.3774	0.6396 / 0.3734
Ours-Sparse	**0.7160 / 0.4382**	**0.6999 / 0.4338**	**0.6953 / 0.4250**	**0.6895 / 0.4095**	**0.6581 / 0.4109**
Ours-LR	**0.7017 / 0.4560**	**0.6926 / 0.4574**	**0.6841 / 0.4427**	**0.6841 / 0.4400**	**0.6639 / 0.4166**

Method\PER	0.1	0.3	0.5	0.7	0.9
BSV	0.2012 / 0.0826	0.1851 / 0.0765	0.1683 / 0.0680	0.1487 / 0.0641	0.1328 / 0.0626
Concat	0.2499 / 0.1137	0.2354 / 0.0997	0.2261 / 0.0843	0.2031 / 0.0755	0.1878 / 0.0758
MultiNMF	0.2077 / 0.0841	0.2057 / 0.0911	0.1924 / 0.0806	0.1823 / 0.0713	0.1655 / 0.0674
PVC	0.2605 / 0.1385	0.2487 / 0.1275	0.2236 / 0.1086	0.2175 / 0.1049	0.2062 / 0.0902
Ours-IMG	0.2807 / 0.1489	0.2554 / 0.1263	0.2512 / 0.1241	0.2421 / 0.1108	0.2201 / 0.0907
Ours-Sparse	**0.2957 / 0.1669**	**0.2723 / 0.1477**	**0.2639 / 0.1383**	**0.2566 / 0.1318**	**0.2379 / 0.1178**
Ours-LR	**0.2814 / 0.1593**	**0.2720 / 0.1465**	**0.2646 / 0.1305**	**0.2464 / 0.1182**	**0.2444 / 0.1186**

Methodology - *Incomplete* Modality Case



(b) UCI handwritten digits dataset

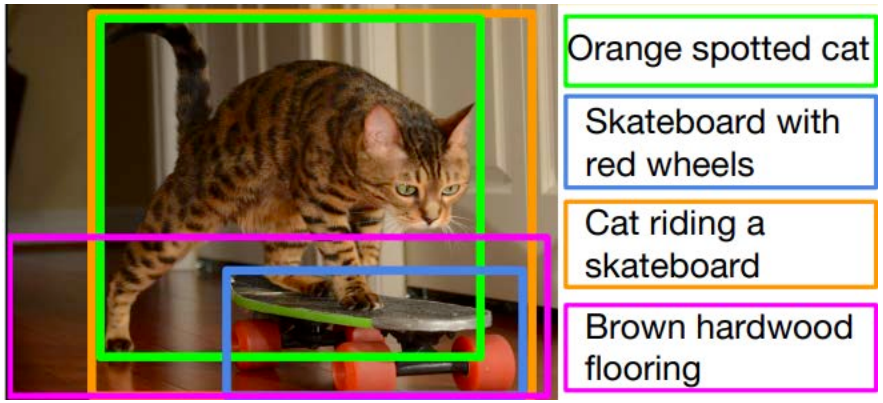


(a) BUAA-NirVis dataset

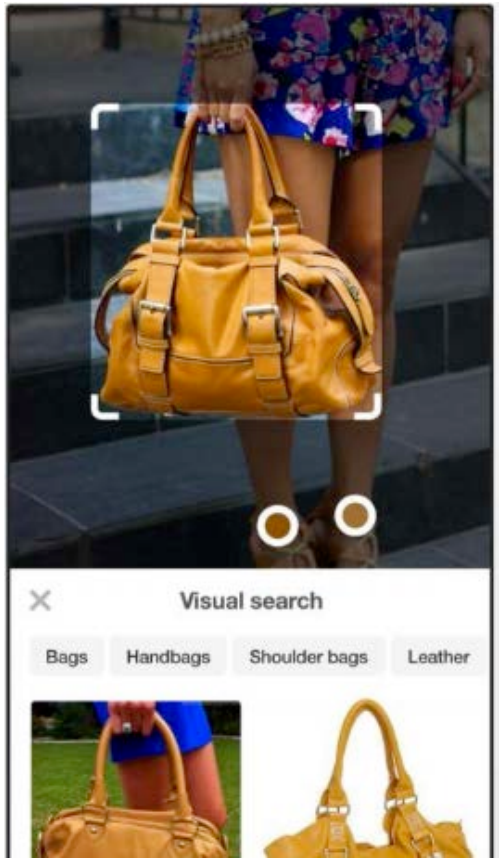
Image-Text Understanding Applications

Image editing

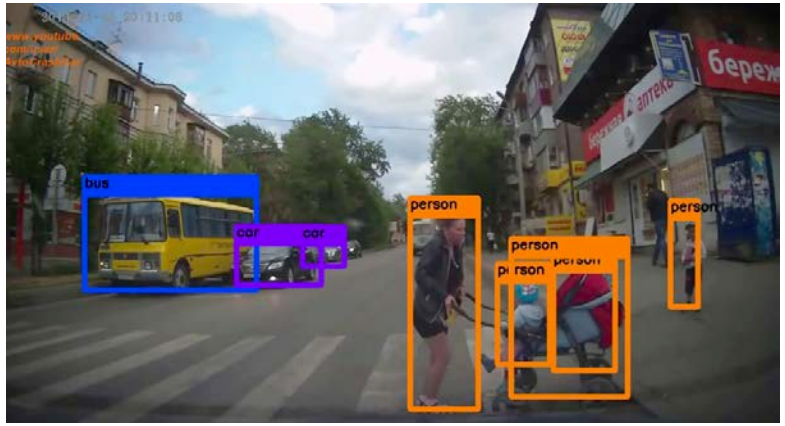
Vision-language tasks



Visual search



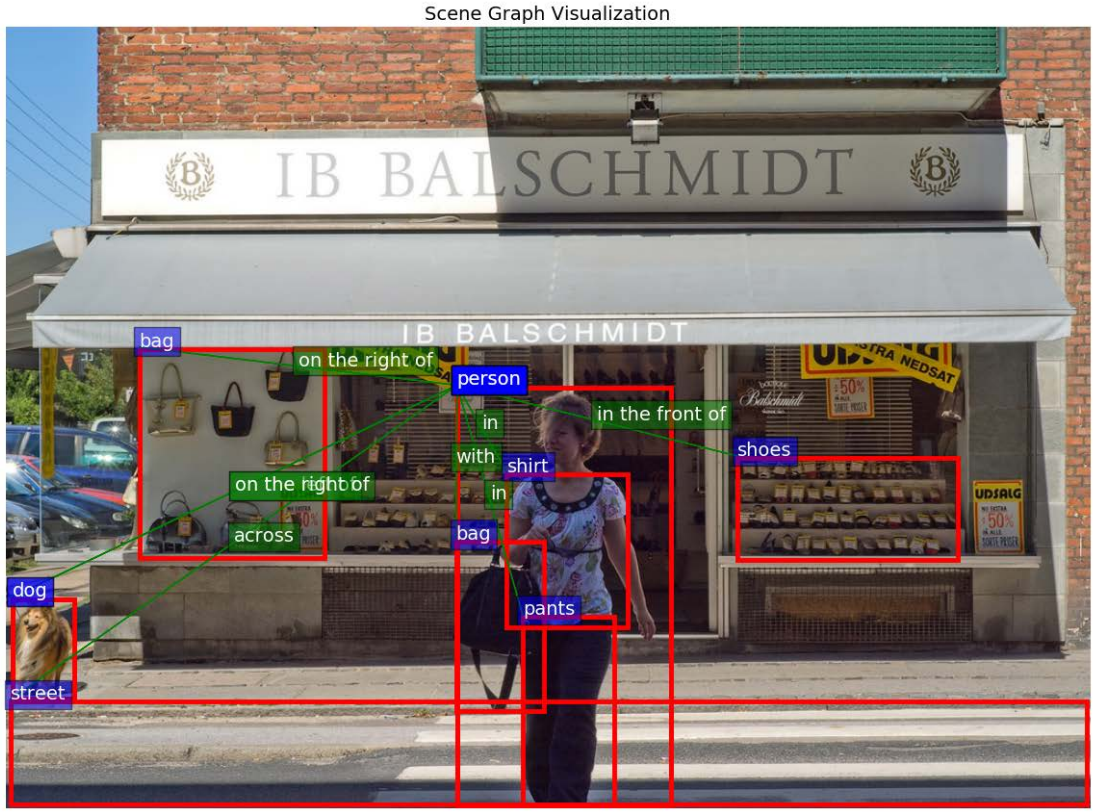
Self-driving cars

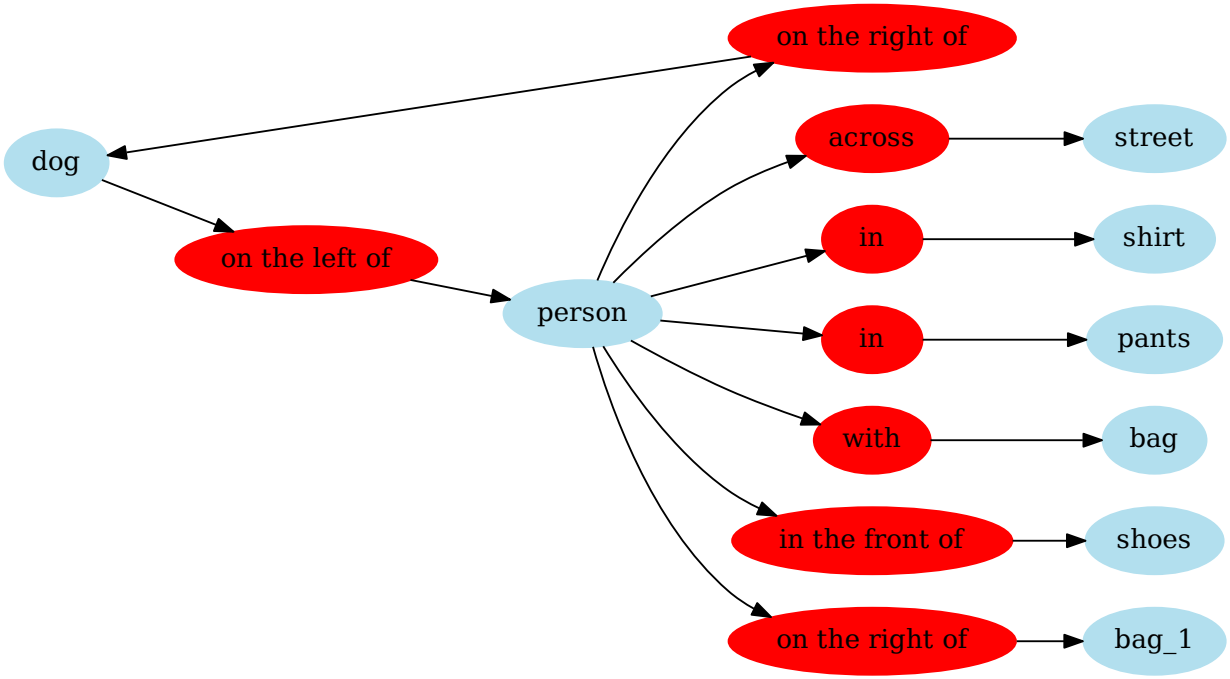


Robotics

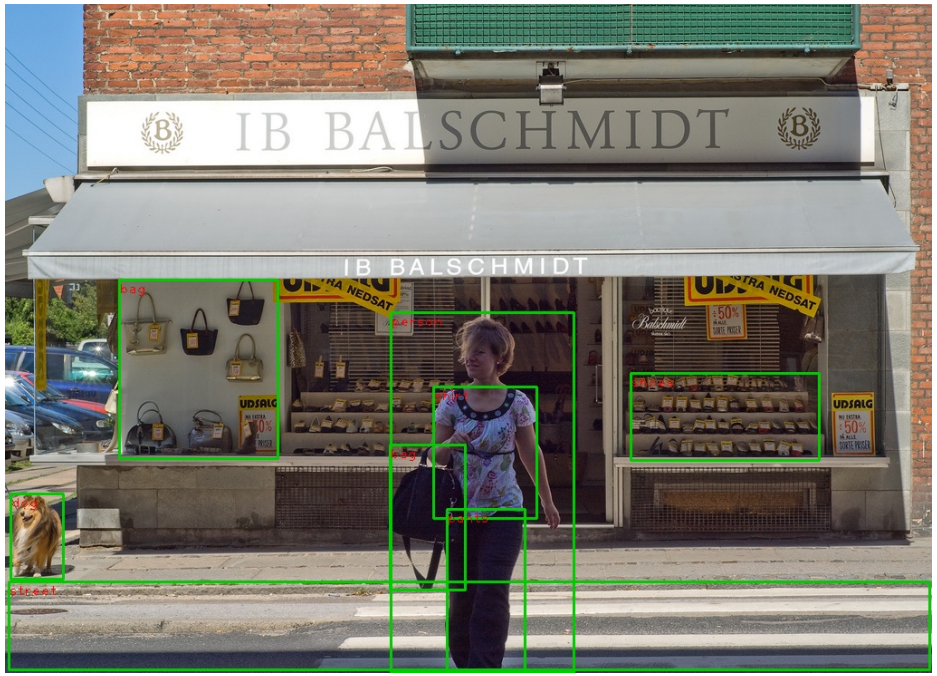
Application 1: Scene Graph Generation

Scene Graph Generation



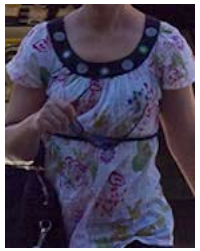


Current Approaches



Current Approaches

shirt



bag

bag

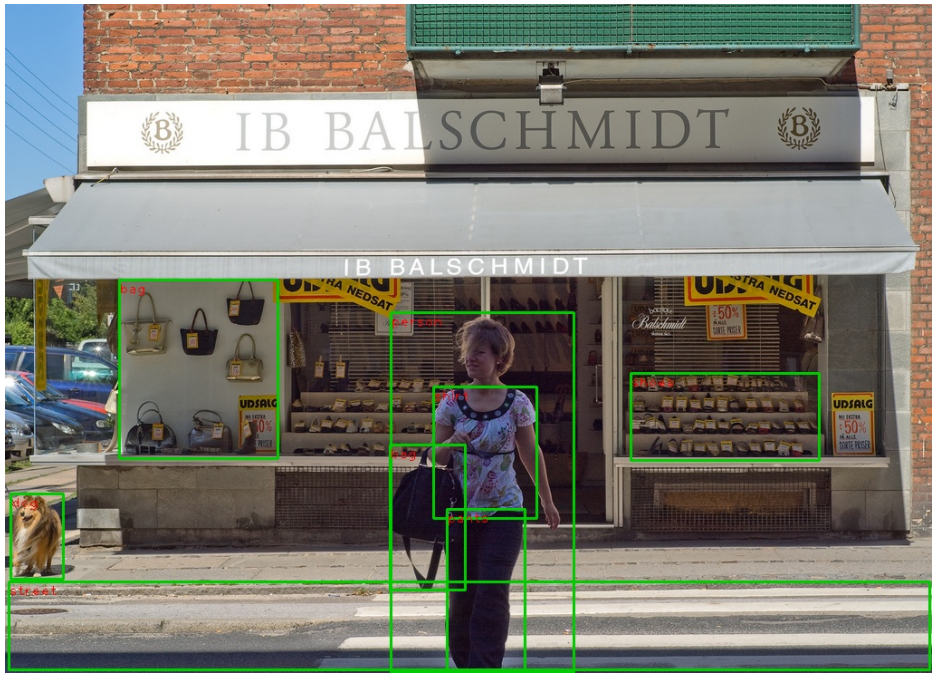
pants

person

dog

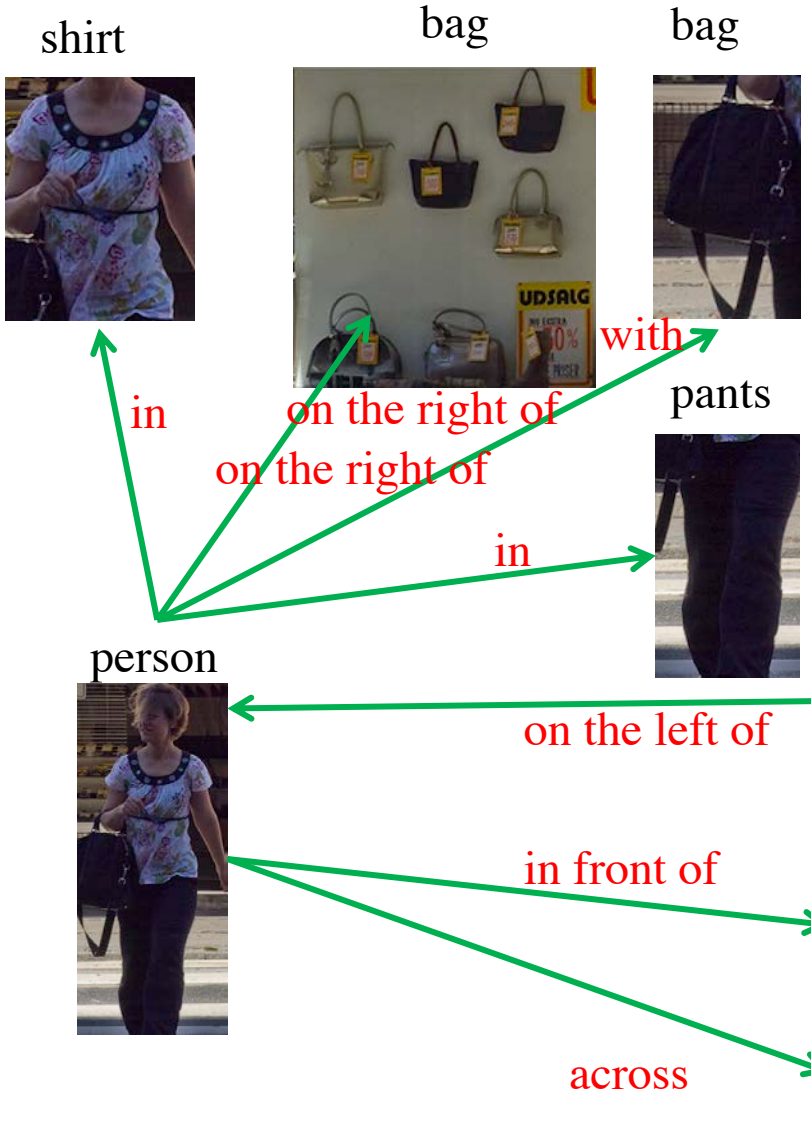
shoes

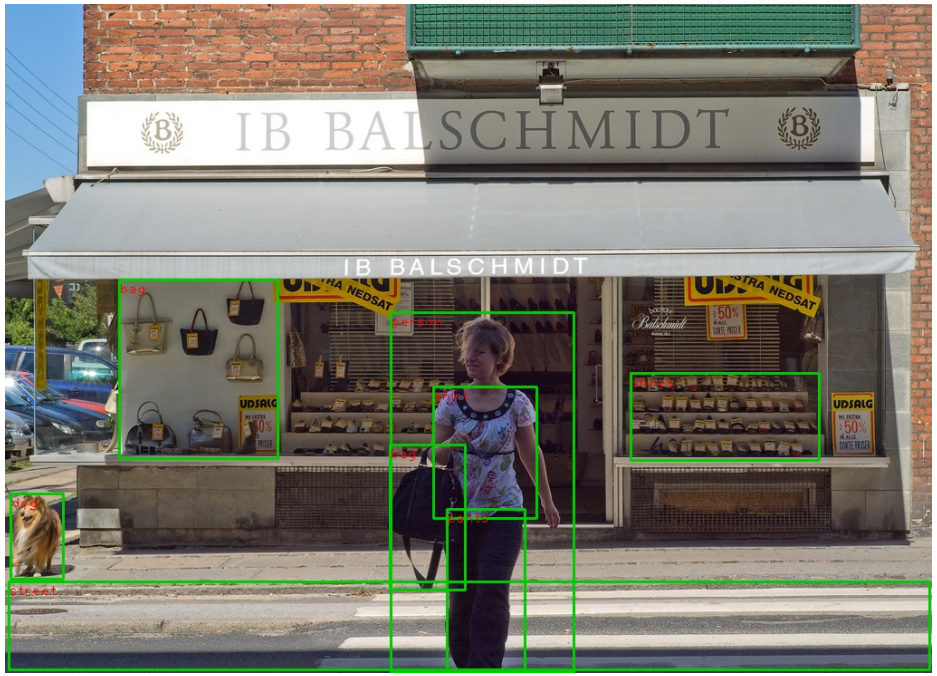
street



Adobe

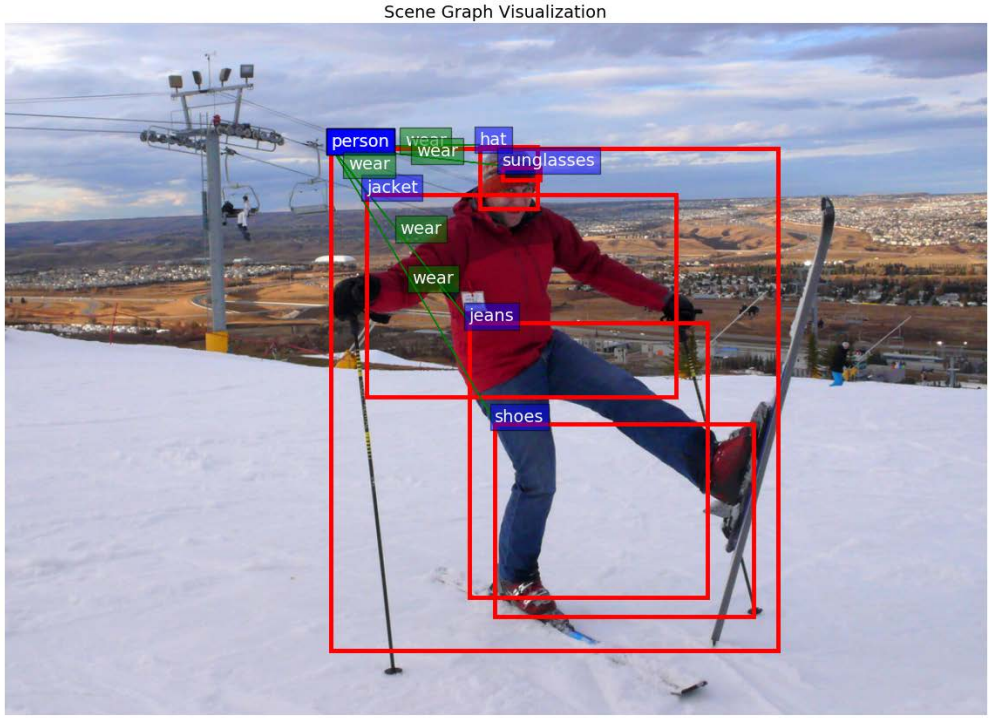
Current Approaches

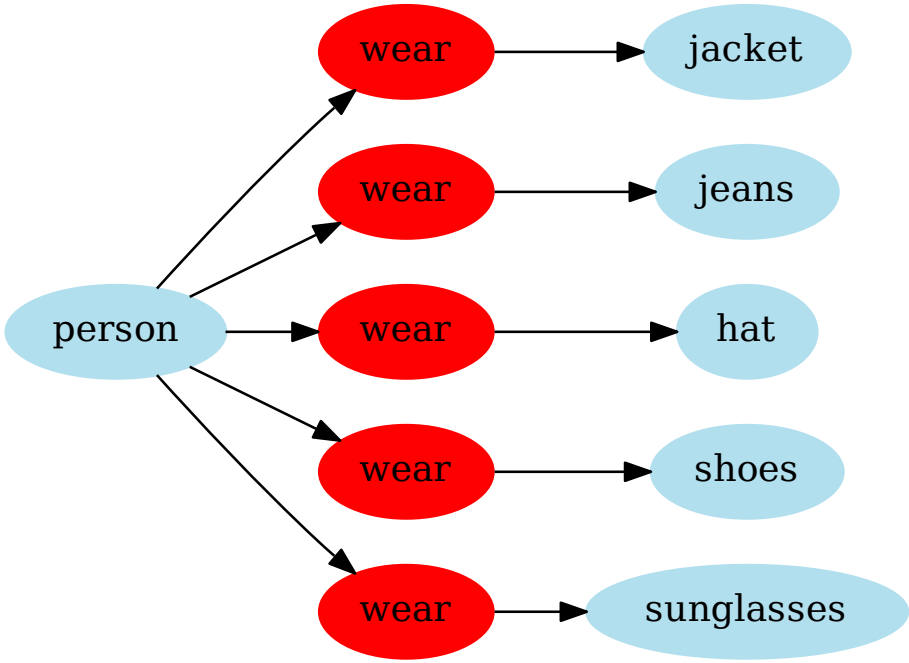




Challenges

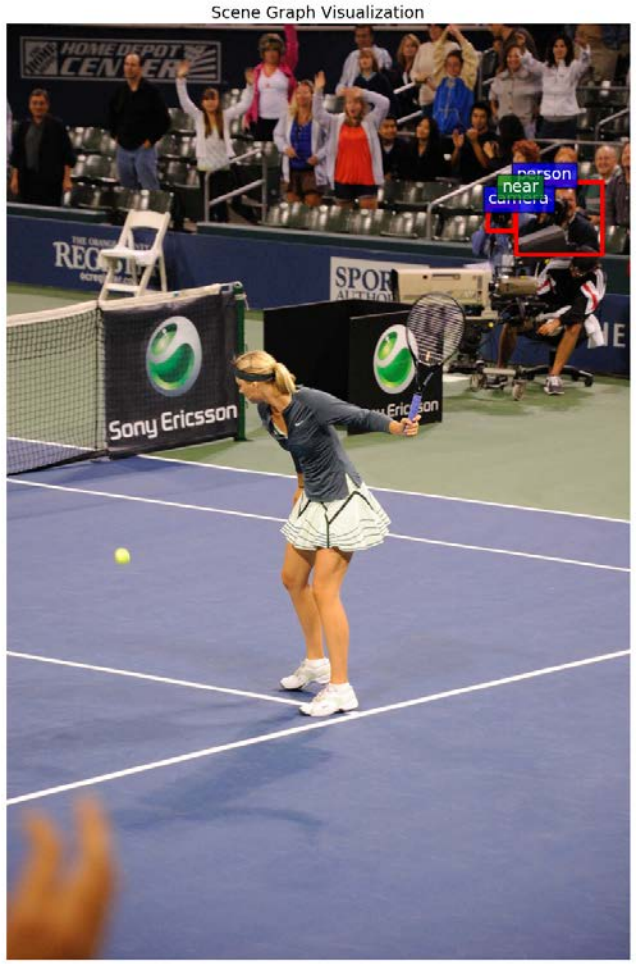
Biased relationship distribution

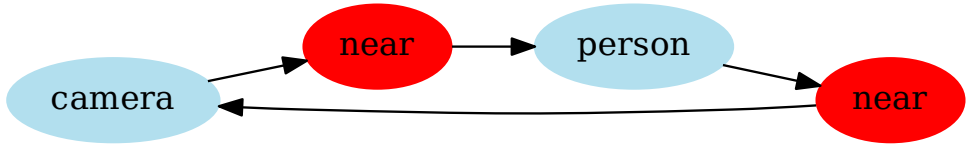




Challenges

Missing annotations





Adobe

Our Solution

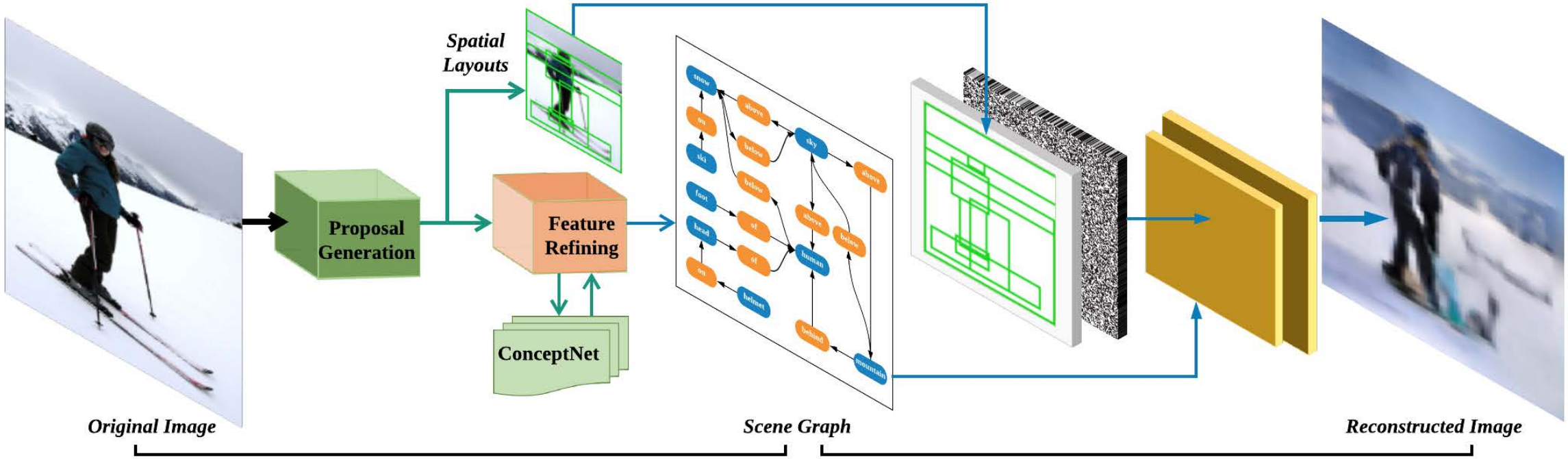
Biased relationship distribution

Enhance the visual relationship detection network with **External-knowledge**

Missing annotations

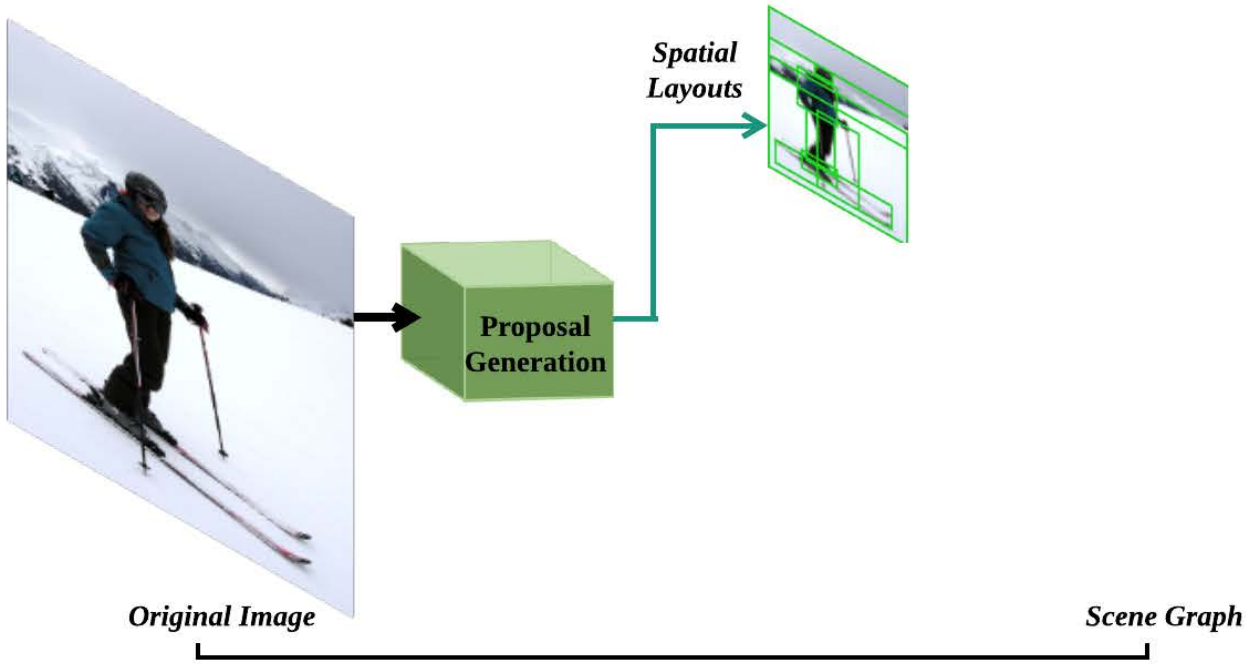
Using image-level supervision to regularize the object detection process

Framework



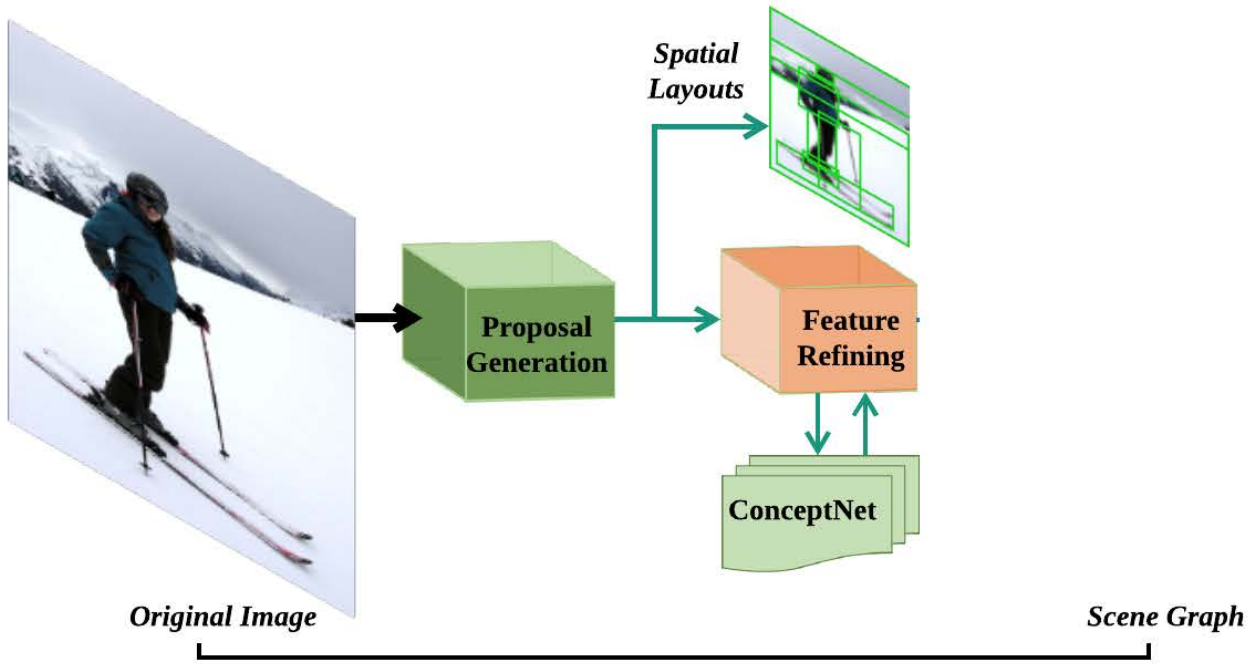
Framework

The entire process can be summarized as the following steps:
(1) **generates object and subgraph proposals for given image;**



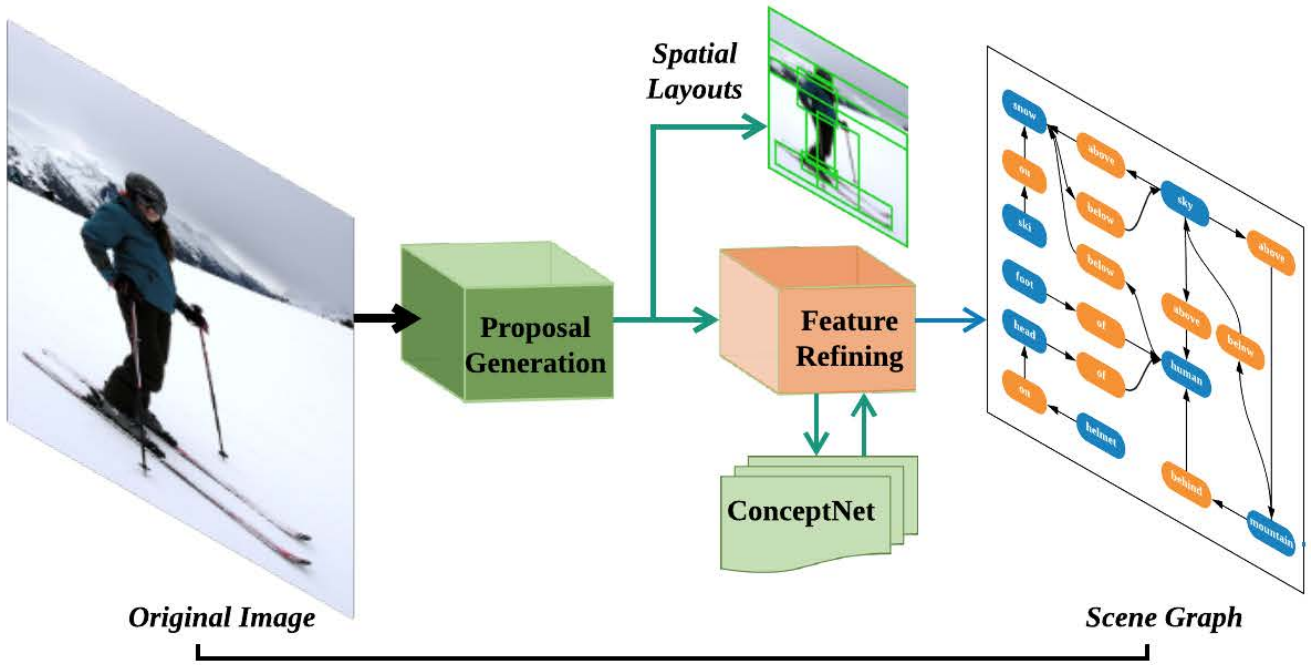
Framework

The entire process can be summarized as the following steps:
(1) generates object and subgraph proposals for given image;
(2) **refine the object and subgraph features with external knowledge;**



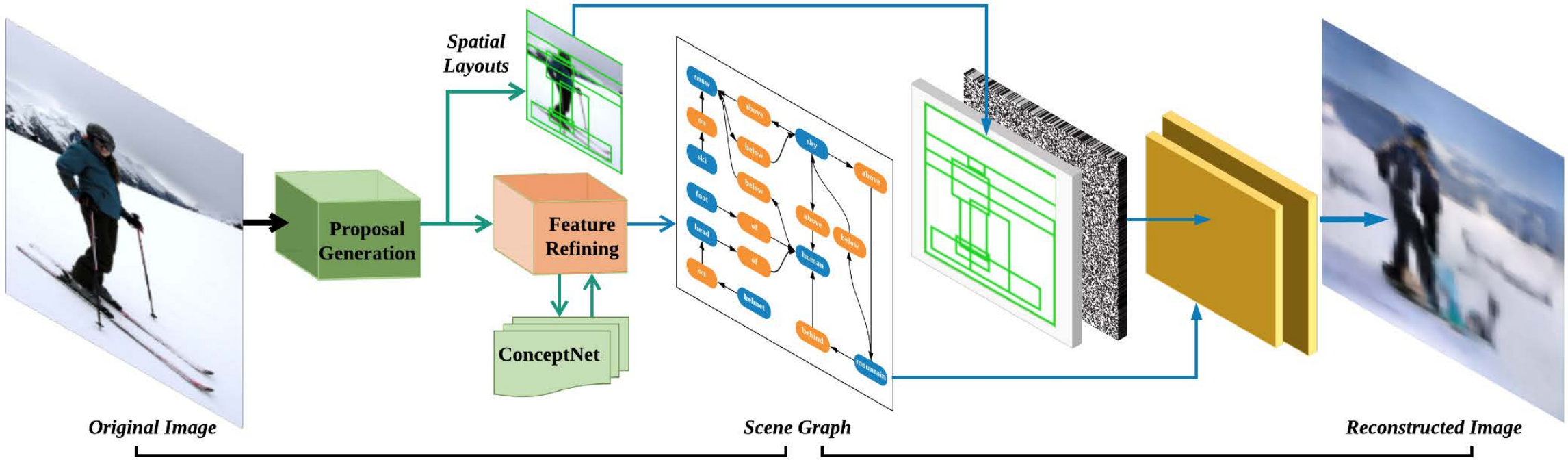
Framework

The entire process can be summarized as the following steps:
(1) generates object and subgraph proposals for given image;
(2) refine the object and subgraph features with external knowledge;
(3) **recognizes the object categories and their relations by fusing the subgraph features and object feature pairs;**

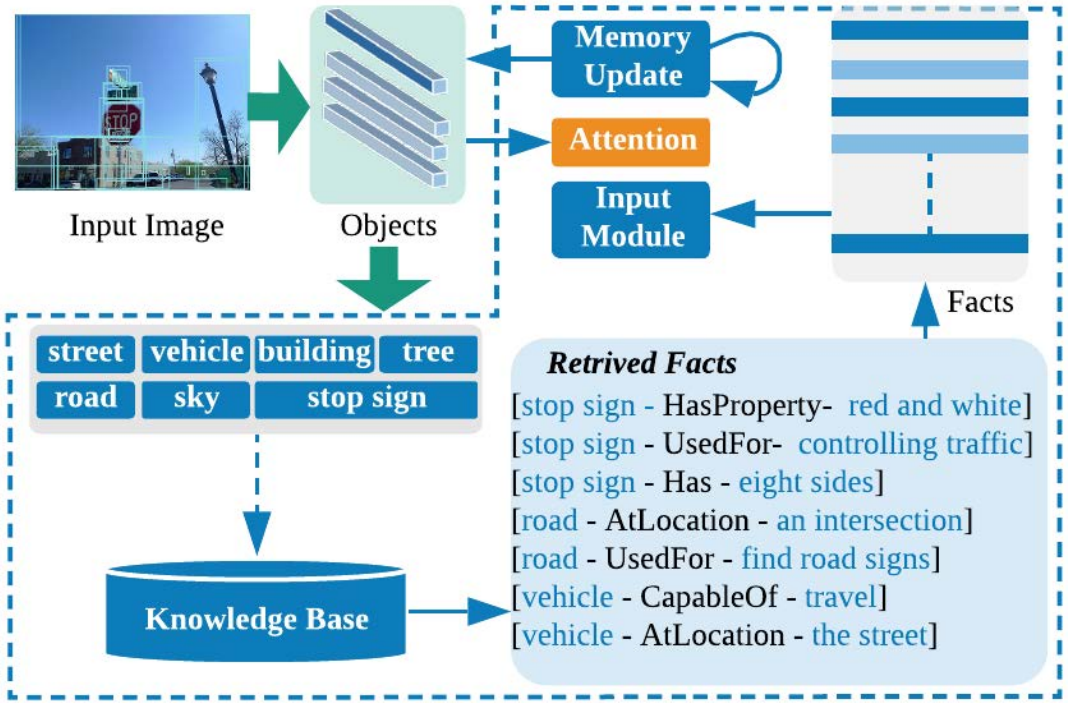


Framework

The entire process can be summarized as the following steps:
(1) generates object and subgraph proposals for given image;
(2) refine the object and subgraph features with external knowledge;
(3) recognizes the object categories with object features and their relations by fusing the subgraph features and object feature pairs;
(4) **reconstruct the image based on the predicted results.**



Contribution 1: Scene Graph Generation with External Knowledge



1. We first use the Region Proposal Network (RPN) to extract a set of object proposals
2. We extract the object label from the refined object vector and match those labels with the corresponding semantic entities in KB.

$$\mathrm{a}_i \xrightarrow{\text{retrive}} \langle \mathrm{a}_i, \mathrm{a}^r_{i,j}, \mathrm{a}^o_{i,j}, w_{i,j} \rangle, j \in [0, K-1]$$

3. We transform the j-th triplet into T words, and map those words into a continuous vector space using word embedding: $\mathbf{x}_j = \mathbf{W}_e X_j$
4. Then, those word vectors are fed into a RNN-based encoder:

$$\mathbf{h}^t_j = \mathrm{RNN}_{\text{fact}}(\mathbf{x}^t_j, \mathbf{h}^{t-1}_j),\ t \in [0, T-1]$$

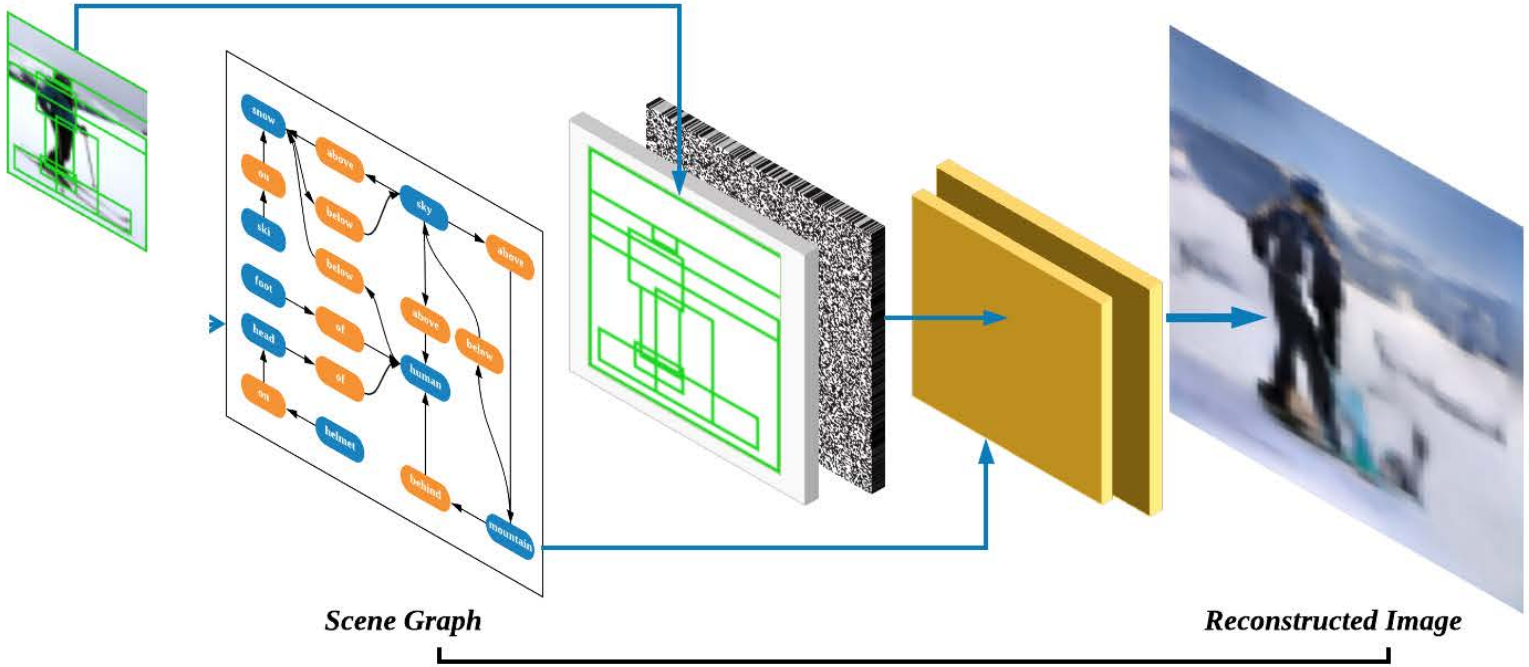
5. The Dynamic Memory Network is adopted to attend the facts:

$$\mathbf{q}_i = \tanh(\mathbf{W}_q[\bar{\mathbf{o}}_\mathbf{i} + \mathbf{b}_q])$$
$$\mathbf{m}^t = \mathrm{ReLU}(\mathbf{W}_m[\mathbf{m}^{t-1}; \sum_k \alpha^t_k \mathbf{f}^k_i; \mathbf{q}_i] + \mathbf{b}_m)$$
$$\alpha^t = \mathrm{softmax}(f_{\text{att}}([\mathbf{F}; \mathbf{m}^{t-1}; \mathbf{q}_i]))$$

6. The refined object vector is:

$$\tilde{\mathbf{o}}_i = \mathrm{ReLU}(\mathbf{W}_c[\bar{\mathbf{o}}_\mathbf{i}; \mathbf{m}^{T-1}] + \mathbf{b}_c)$$

Contribution 2: Scene Graph To Image Reconstruction with GAN



Adobe

Experiments

- We evaluate our approach on two datasets: Visual Relationship Detection (VRD) and Visual Genome (VG).
  - VRD is a small benchmark dataset where most of the existing methods are evaluated.
  - Compared to VRD, raw Visual Genome contains too many noisy labels, so dataset cleansing should be done to make it available for model training and evaluation. Another dataset is the cleansed-version Visual Genome.

Table 1: Dataset statistics. #Img denotes the number of images. *#Rel* denotes the number of subject-predicate-object relation pairs. #Obj and #Pred denotes the number of object and predicate categories respectively.

Dataset	**Training Set**		**Testing Set**		#Obj	#Pred
	#Img	#Rel	#Img	#Rel		
VRD [13]	4,000	30,355	1,000	7,638	100	70
VG [10]	46,164	507,296	10,000	111,396	150	50

Component Analysis

1. **Models:**
   - **Baseline**. This baseline model is reimplementation of the Factorizable Net in our codebase.
   - **KB**. This model is a knowledge-enhanced version of the baseline model. Features are refined jointly with external knowledge.
   - **GAN**. This model is a GAN-enhanced version of the baseline model, but without the external knowledge.
   - **KB-GAN**. This is our full model containing all the KB and GAN.
2. We evaluate our models on two tasks: **Visual Phrase Detection (PhrCls)** and **Scene Graph Generation (SGGen).**
   - **PhrCls** is to detect the <*subject-predicate-object*> phrases.
   - **SGGen** is to detect the *objects* within the image and recognize their *pair-wise relationships*.

Table 2: Ablation studies of each components.

KB	GAN	PhrDet		SGGen	
		Rec@50	Rec@100	Rec@50	Rec@100
-	-	25.57	31.09	18.16	22.30
✓	-	27.02	34.04	19.85	24.58
-	✓	26.65	34.06	19.56	24.64
✓	✓	**27.65**	**34.60**	**20.33**	**25.01**

Table 3: Ablation study using mean average procession.

Model	Faster R-CNN [18]	ViP-CNN [11]	Baseline	KB	GAN	KB-GAN
mAP	14.35	20.56	20.70	22.26	22.10	**22.49**

Table 4: Ablation study on spare data.

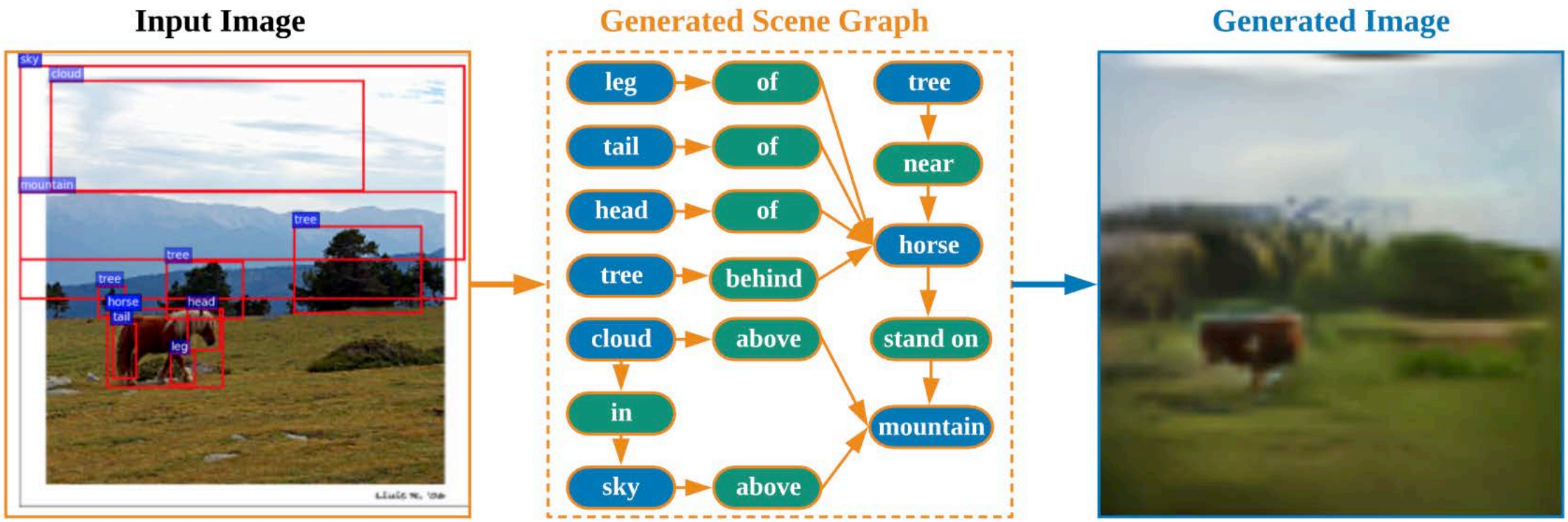
KDMN	GAN	PhrDet		SGGen	
		Rec@50	Rec@100	Rec@50	Rec@100
-	-	15.44	20.96	10.94	14.53
-	✓	24.07	30.89	17.50	22.31
✓	✓	**26.62**	**31.13**	**19.78**	**24.17**

- Top-K Recall is used to evaluate how many labeled relationships are hit in the top K predictions.

Comparison with Existing Methods

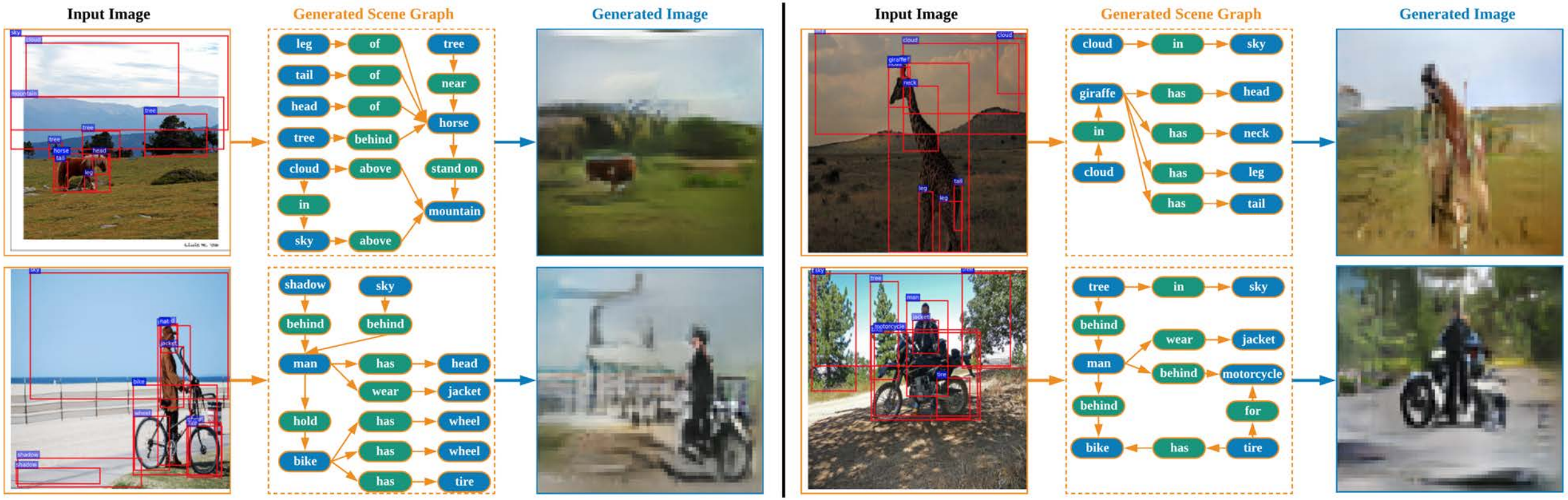
Table 5: Comparison with existing methods on *PhrDet*, and *SGGen*.

Dataset	Model	PhrDet		SGGen	
		Rec@50	Rec@100	Rec@50	Rec@100
VRD [13]	LP [13]	16.17	17.03	13.86	14.70
	ViP-CNN [11]	22.78	27.91	17.32	20.01
	DR-Net [1]	19.93	23.45	17.73	20.88
	ILC [16]	16.89	20.70	15.08	18.37
	U+W+SF+LK: T+S [25]	26.32	29.43	19.17	21.34
	Factorizable Net [9]	26.03	30.77	18.32	21.20
	KB-GAN	**27.65**	**34.60**	**20.33**	**25.01**



Adobe

Qualitative Results



Adobe

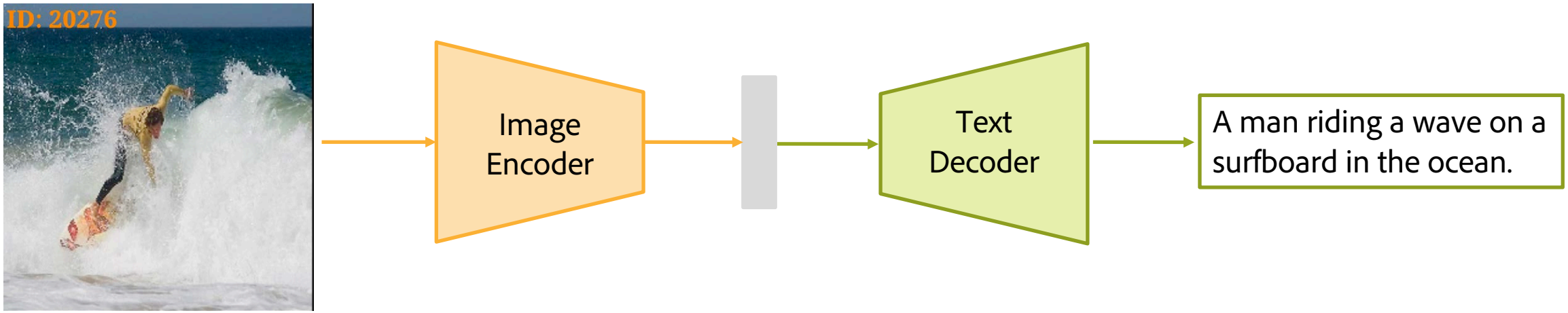
Application 2: Unpaired Image Captioning

Problem Definition

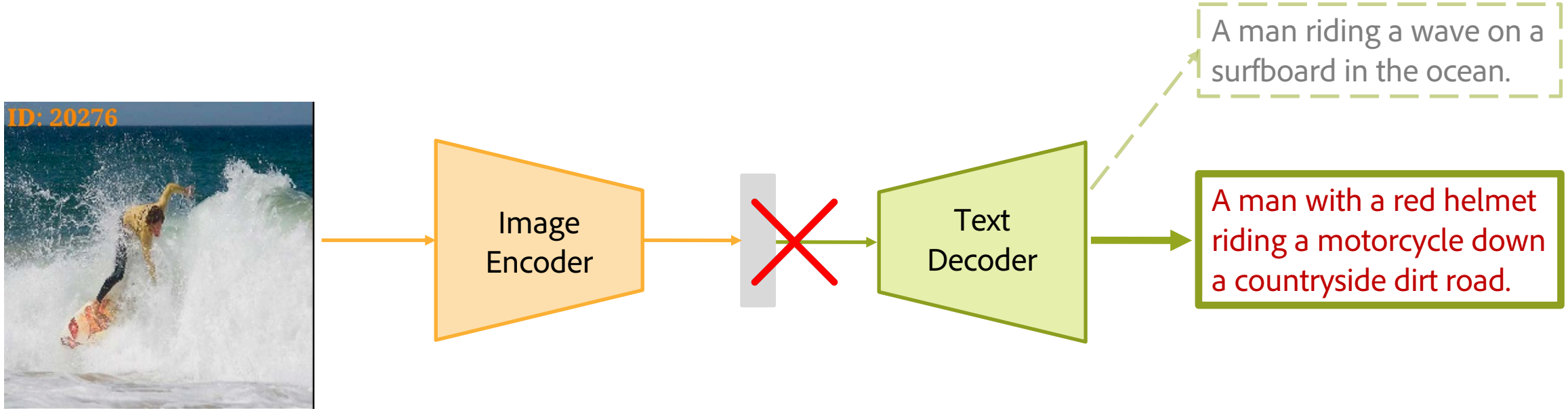
The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Pipeline for Pair Setting

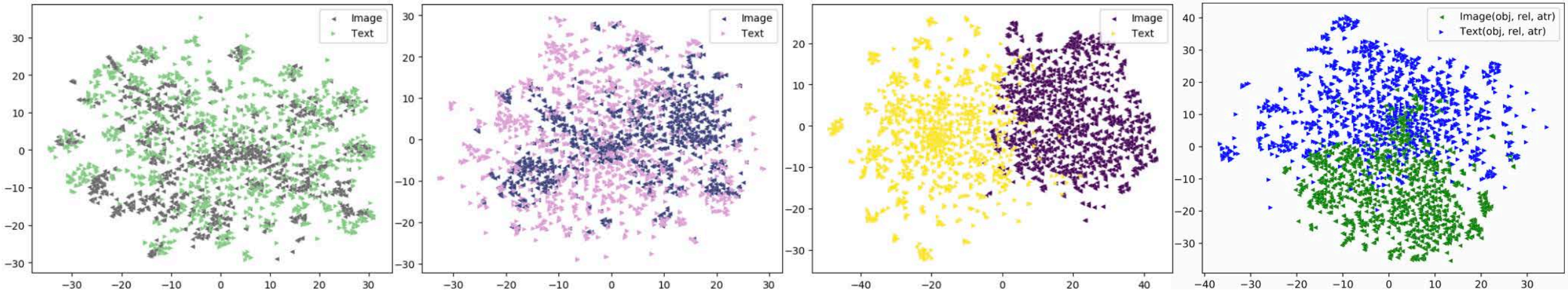


Unpaired Setting?



Challenge: Distribution Divergence

- Randomly sample 1,500 unpaired image/text.

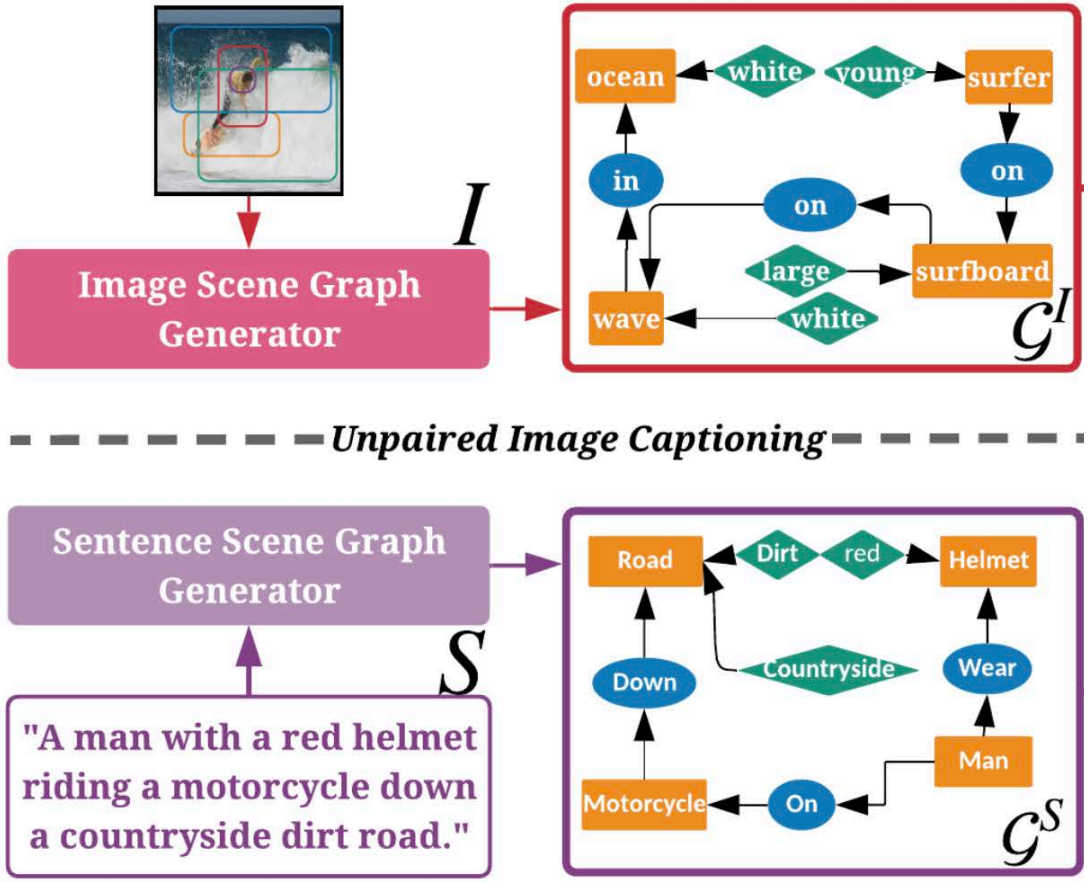


(a) Object Features (Raw) (b) Relation Features (Raw) (c) Attribute Features (Raw) (d) Triplet Features (Raw)

Visualization of features in 2D space by t-SNE.

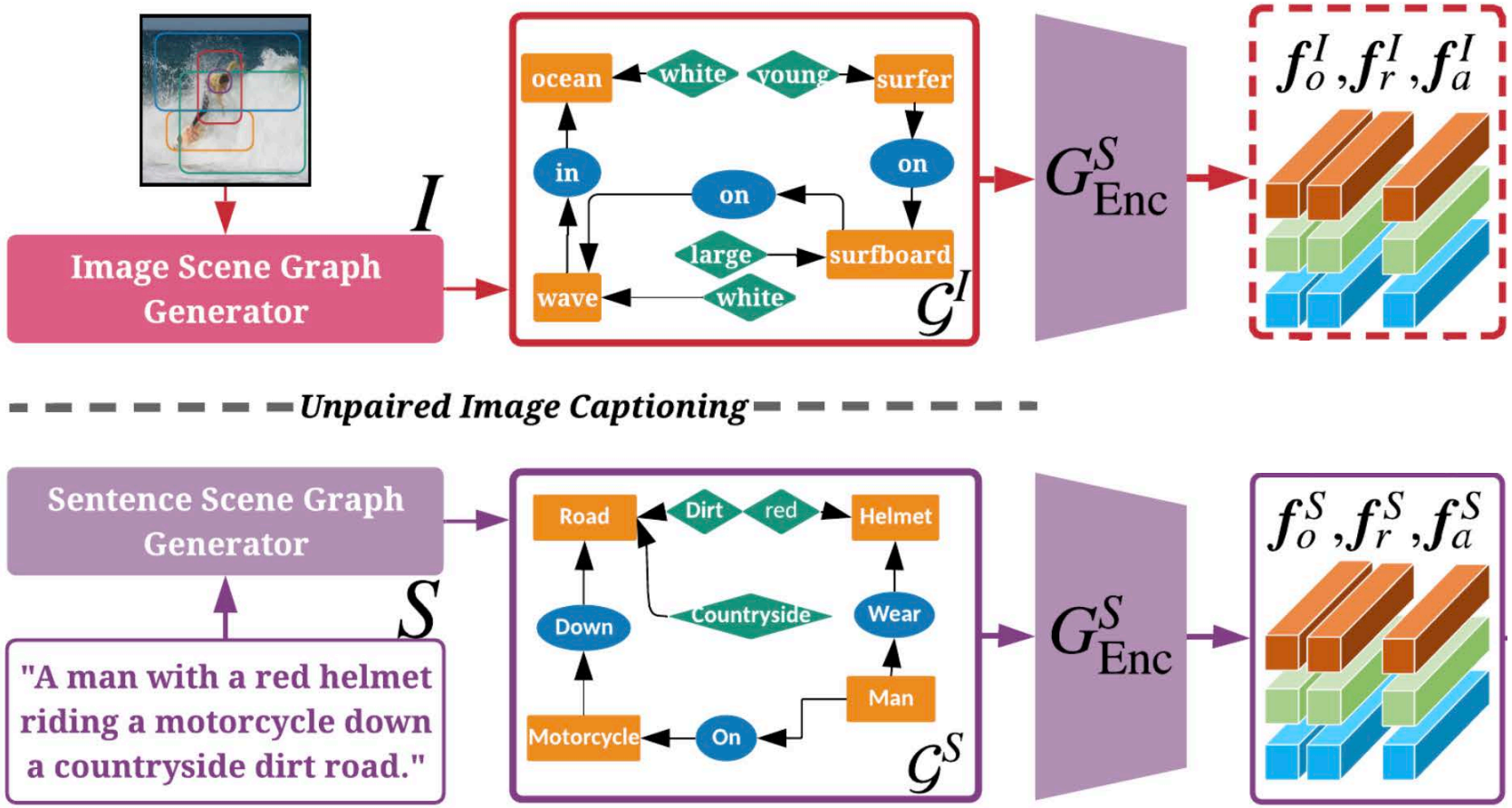
Framework

- Use scene graphs as the explicit representations for image / text.



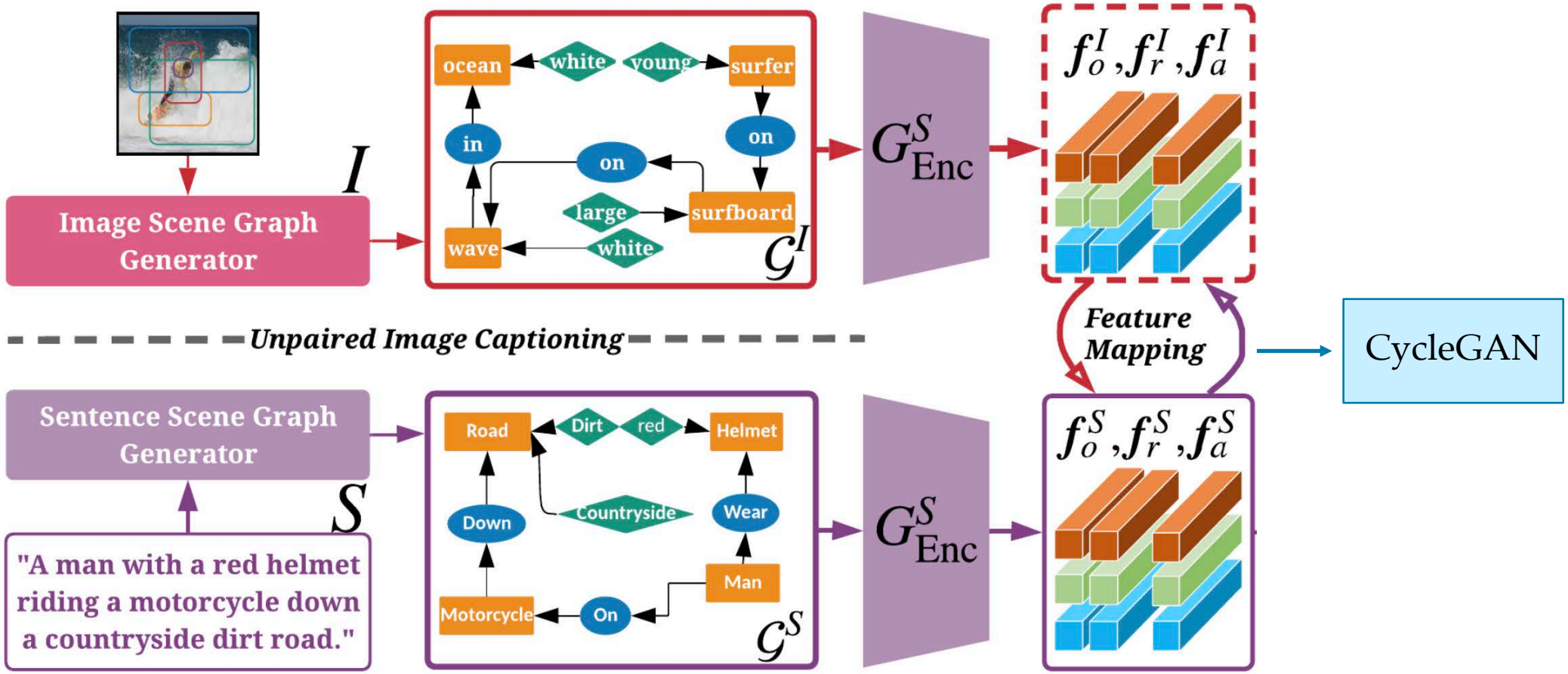
Framework

- Feature extraction based on scene graphs



Framework

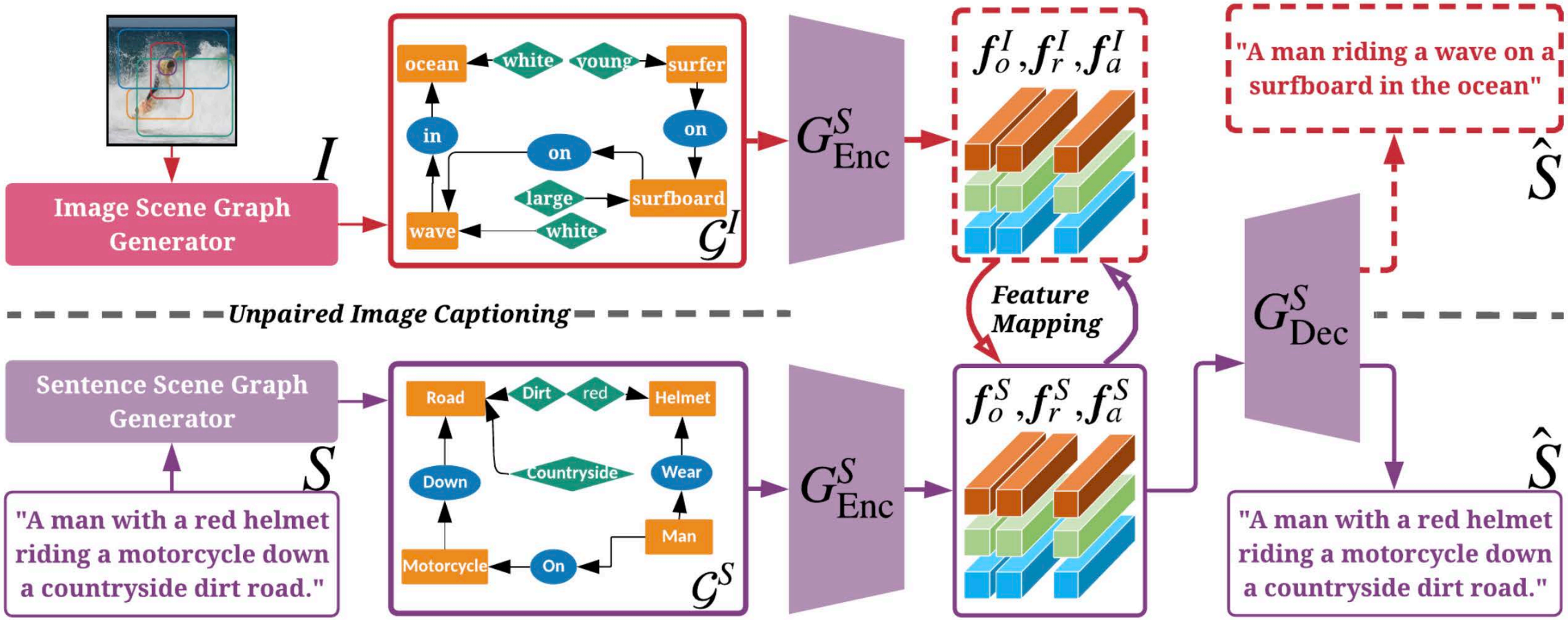
- Unpaired feature mapping by CycleGAN [1].

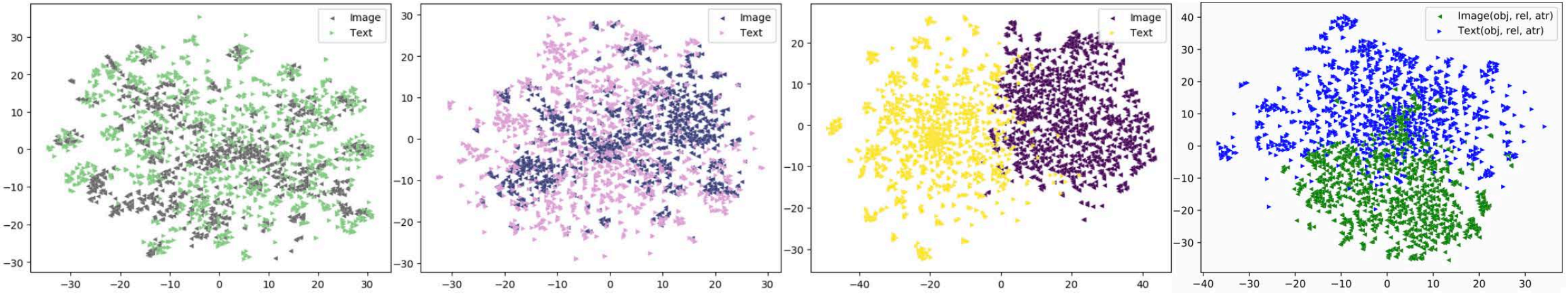


[1] Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, ICCV 2017

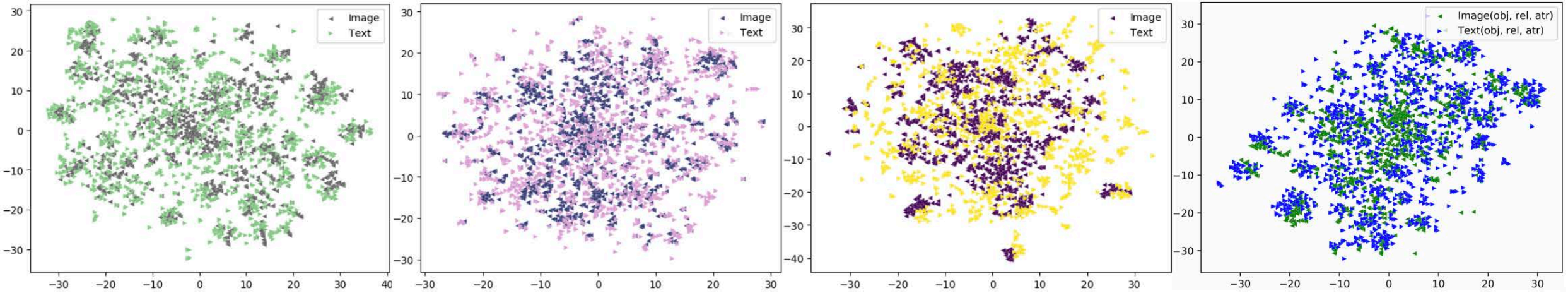
Framework

- Regularization via sentence reconstruction





(a) Object Features (Raw) (b) Relation Features (Raw) (c) Attribute Features (Raw) (d) Triplet Features (Raw)



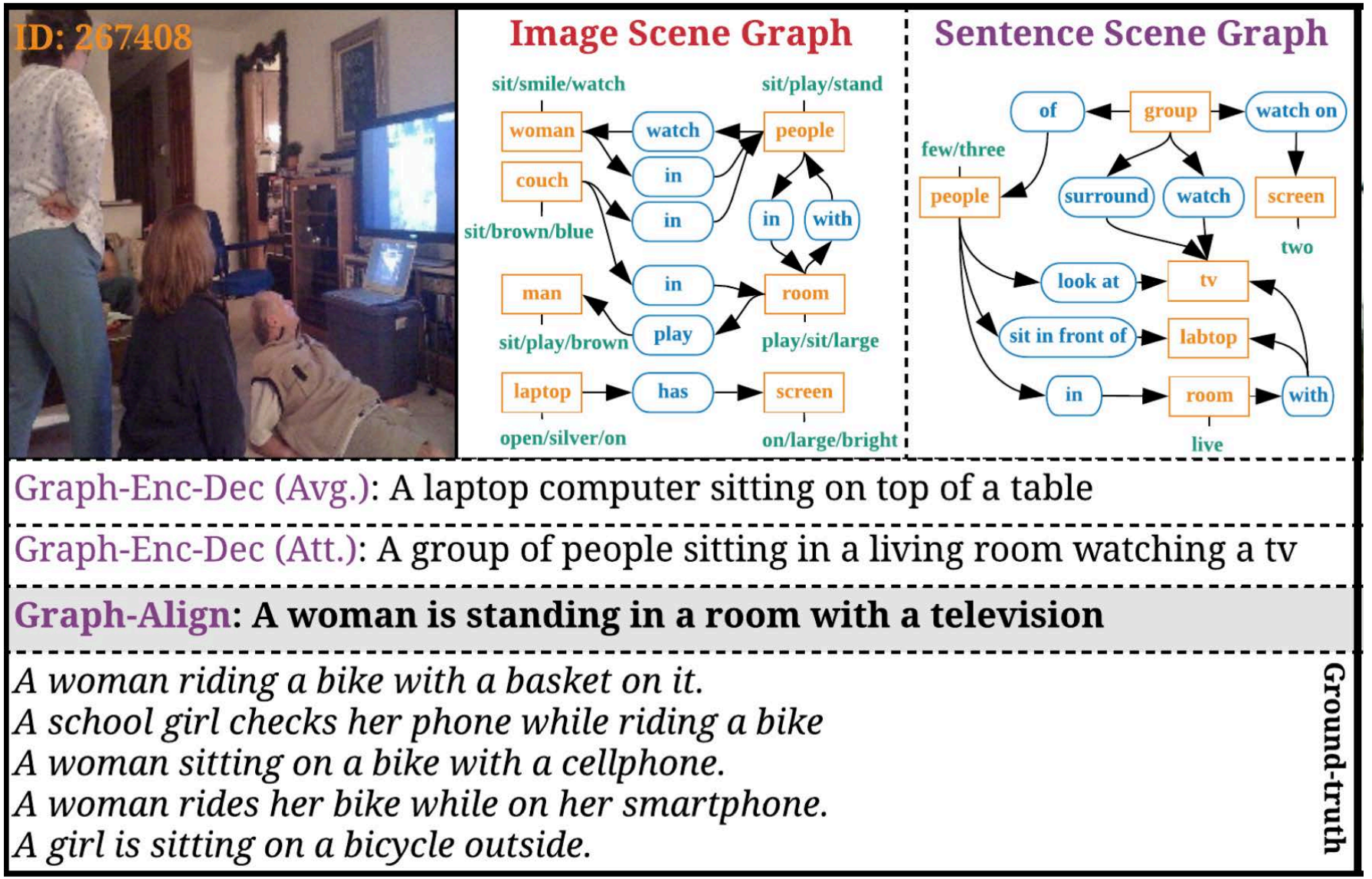
(e) Object Features (Aligned) (f) Relation Features (Aligned) (g) Attribute Features (Aligned) (h) Triplet Features (Aligned)

Visualization of features in 2D space by t-SNE. We plot the scatter diagrams for 1,500 samples

Adobe

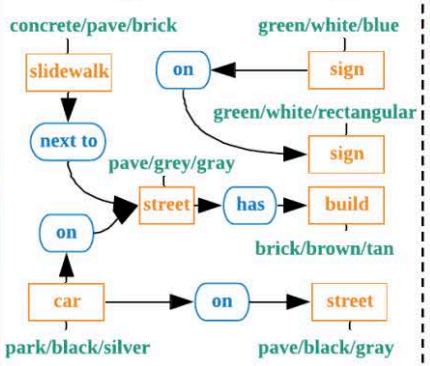
Experiment

Qualitative Example:

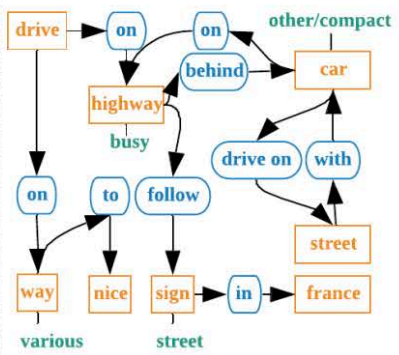


- More examples

Image Scene Graph



Sentence Scene Graph



Graph-Enc-Dec (Avg.): A street sign on the side of a street

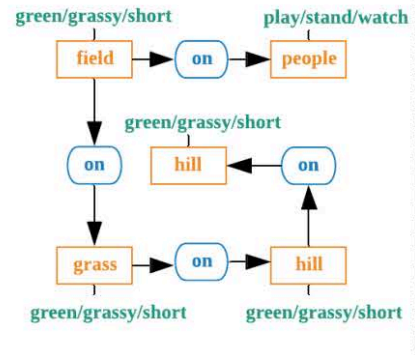
Graph-Enc-Dec (Att.): A busy street with a cars on a highway

Graph-Align: A street sign on a pole with a traffic light

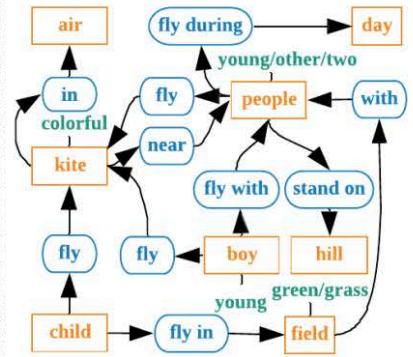
Ground-truth

Driving on a highway behind a compact car, on the way to Nice
A French motorway points travelers in various directions.
A busy road full of street signs and cars
Car on a highway following signs in France
A car is driving on a street with other cars and signs.

Image Scene Graph



Sentence Scene Graph



Graph-Enc-Dec (Avg.): A large kite flying in the sky

Graph-Enc-Dec (Att.): A young boy standing on a hill flying a kite

Graph-Align: A group of people flying a kite on a grassy hill

Ground-truth

Three people standing on a grassy hill flying a kite.
A young boy is flying a colorful kite near two other people.
Three people flying a kite in the air during the day.
A child flying a kite in a green grass field with two other people.
A young boy flying a kite with two other people watching.

Adobe

Experiment

- MSCOCO dataset

Performance comparisons on the test split of the MSCOCO dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Paired Setting								
Soft-Attention [36]	71.8	50.4	35.7	25.0	23.0	–	90.0	–
Stack-Cap [13]	78.6	62.5	47.9	36.1	27.4	56.9	120.4	20.9
SGAE (base) [38]	79.9	–	–	36.8	27.7	57.0	120.6	20.9
Unpaired Setting								
Language Pivoting [14]	46.2	24.0	11.2	5.4	13.2	–	17.7	
Adversarial+Reconstruction [10]	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1
Graph-Align	**67.1**	**47.8**	**32.3**	**21.5**	**20.9**	**47.2**	**69.5**	**15.0**

Conclusion

1. We study the (unsupervised) multimodal representation learning for both *complete*-modal and *incomplete*-modal case.

2. In scene graph generation, external-knowledge as well as image-level supervision are introduced to address the dataset biases.

3. In image captioning, we study the unpaired setting. Scene graphs serves as the bridge to connect the images and texts.

Thanks!