EMBEDDED SPARSE CODING FOR SUMMARIZING MULTI-VIEW VIDEOS

Rameswar Panda[⋆] *Abir Das*[†] *Amit K. Roy-Chowdhury*[⋆]

[⋆] Electrical and Computer Engineering Department, University of California, Riverside
[†] Computer Science Department, University of Massachusetts, Lowell

ABSTRACT

Most traditional video summarization methods are designed to generate effective summaries for single-view videos, and thus they cannot fully exploit the complicated intra- and inter-view correlations in summarizing multi-view videos. In this paper, we introduce a novel framework for summarizing multi-view videos in a way that takes into consideration both intra- and inter-view correlations in a joint embedding space. We learn the embedding by minimizing an objective function that has two terms: one due to intra-view correlations and another due to inter-view correlations across the multiple views. The solution is obtained by using a Majorization-Minimization algorithm that monotonically decreases the cost function in each iteration. We then employ a sparse representative selection approach over the learned embedding space to summarize the multi-view videos. Experiments on several multi-view datasets demonstrate that the proposed approach clearly outperforms the state-of-the-art methods.

Index Terms— Video Summarization, Sparse Coding, Frame Embedding

1. INTRODUCTION

Summarizing a video sequence is of considerable practical importance as it helps the user in several video analysis applications like content-based search, interactive browsing, retrieval and semantic storage, among others. There is a rich literature in computer vision and multimedia developing a variety of ways to summarize a single-view video in form of a key-frame sequence or a video skim (see below for details). However, another important problem but rarely addressed in this context is to find an informative summary from *multi-view* videos [1, 2, 3, 4]. Similar to the single-view summarization problem, the multi-view summarization approach seeks to take a set of input videos captured from different cameras and produce a reduced set of output videos or key-frame sequence that presents the most important portions of the inputs within a short duration.

Summarizing multi-view videos is different from single-view videos in two important ways. First, these videos have large amount of inter-view content correlations along with intra-view correlations. Second, different environmental factors like difference in illumination, pose and synchronization issues among the cameras also pose a great challenge in multi-view settings. So, methods that attempt to extract summary from single-view videos usually do not produce an optimal set of representatives while summarizing multi-view videos.

Prior Work. Various strategies have been studied for summarizing single-view videos, including clustering [5, 6], attention modeling [7], super frame segmentation [8], temporal segmentation [9], crowd-sourcing [10], storyline graphs [11], submodular maximization [12, 13], point process [14], and maximal biclique finding [15]. Generating personalized summaries is another recent trend for video summarization [16, 17]. Interested readers can check [18, 19] for a more comprehensive summary.

To address the challenges encountered in a multi-view camera network, some state-of-the-art approaches use random walk over spatio-temporal shot graphs [1] and rough sets [2] to summarize multi-view videos. A very recent work in [4] uses optimum path forest clustering to solve the problem of summarizing multi-view videos. An online method for summarization can also be found in [3]. The work in [20, 21] also addresses a similar problem of summarization in multi-camera settings with non-overlapping field of views.

More recently, there has been an growing interest in using sparse coding (SC) to solve the problem of video summarization [22, 23, 24, 25] since the sparsity and reconstruction error term in SC naturally fit into the problem of summarization. Specifically, the summary length should be as small as possible and at the same time, the original video should be reconstructed with high accuracy using the extracted summary. These approaches can be used to summarize multi-view videos in two straightforward ways. First, by applying SC to each view of the multi-view videos, and then combining the results to produce a single summary and second, by simply concatenating all the multi-view videos into a long video along the time line and then generating a single video summary. However, both of the strategies fail to exploit any statistical interdependencies between the views, and hence produces a lot of redundancies in the output summary.

Following the importance of multi-view correlations, we split the problem into two sub-problems, namely capturing the content correlations via an embedded representation and then applying sparse representative selection over the learned embedding space to generate the summaries. Specifically, our work builds upon the idea of subspace learning, which typically aim to obtain a latent subspace shared by multiple views by assuming that the input views are generated from this latent subspace [26, 27].

Contributions. To summarize, the contributions of the paper are the followings. (1) We propose a multi-view frame embedding which is able to preserve both intra- and inter-view correlations without assuming any prior correspondences/alignment between the multi-view videos. (2) We propose a sparse representative selection method over the learned embedding to summarize the multi-view videos, which provides scalability in generating summaries *(analyze once, generate many)*. (3) We demonstrate the effectiveness of our summarization approach on several multi-view datasets including both indoor and outdoor environments.

2. MULTI-VIEW FRAME EMBEDDING

Problem Statement: Consider a set of K different videos captured from different cameras, in a D-dimensional space where $X^k = \{x_i^k \in \mathbb{R}^D, i = 1, \cdots, N_k\}, k = 1, \cdots, K$. Each x_i represents the feature descriptor (e.g., color, texture) of a video frame in D-dimensional feature space. As the videos are captured non-synchronously, the number of frames in each video might be different and hence there is no optimal one-to-one correspondence that can be assumed. We use N_k to denote the number of frames in k-th video and N to denote the total number of frames in all videos.

Given the multi-view videos, our goal is to find an embedding for all the frames into a common space while satisfying some constraints. Specifically, we are seeking a set of embedded coordinates $Y^k = \{y_i^k \in \mathbb{R}^d, i = 1, \cdots, N_k\}, k = 1, \cdots, K$, where, d ($<< D$) is the dimensionality of the embedding space, with the following two constraints: (1) *Intra-view correlations:* content correlations between frames of a video should be preserved in the embedding space, (2) *Inter-view correlations:* frames from different videos with high feature similarity should be close to each other in the embedding space as long as they do not violate the intra-view correlations present in an individual view.

Modeling Multi-view Correlations: To achieve an embedding that preserves the above two constraints, we introduce two proximity matrices based on intra- and inter-view frame feature distances respectively. The intra-view proximity matrix is represented by P^k where P_{ij}^k measures the pairwise proximity between two frames i and j in the k-th view. Similarly, the inter-view proximity matrices are represented by P^{mn} where P_{ij}^{mn} denote the pairwise proximity between the i-th frame in view m and the j-th frame in view n.

Intra-view proximity should reflect spatial arrangement of feature descriptors in each view. Hence, we use a Gaussian kernel on the Euclidean distance between two frames to calculate the intra-view proximities, *i.e.*,

$$P_{ij}^k = e^{-\|x_i^k - x_j^k\|^2}/2\sigma^2, \tag{1}$$

where σ is a scale parameter that determines the extent of similarity between any two frames. As suggested in [28], we set $\sigma = \beta \cdot max(E_d)$, where $\beta \leq 0.2$ and E_d is the set of all pairwise Euclidean distances between the frames.

One seemingly obvious choice for measuring the inter-view proximities is to use the same Gaussian kernel (Eq. 1) on the Euclidean distance between frames of two different videos. However, such a choice is not suitable for multi-view frame embedding as the proximities do not satisfy the exclusion principle [29]. The exclusion principle tries to maintain the local structure of a view while mapping frames from different views into the embedding space. Hence we use Scott and Longuet-Higgins (SLH) algorithm [29] over proximities generated with Gaussian kernel over two different views to enforce the exclusion principle. Notice that the proximity matrix P_{ij}^{mn} is not symmetric and there exists a hyper-symmetry structure, *i.e.*, $P^{mn} = P^{nm^T}$.

Objective Function: The aim of the embedding is to correctly match the proximity score between two frames x_i and x_j to the score between the corresponding embedded points y_i and y_j respectively. Motivated by this observation, we reach the following objective function on the embedded points Y, which needs to be minimized. The objective function is

$$\mathcal{F}(Y) = \sum_k \sum_{i,j} P_{ij}^k \ln \frac{P_{ij}^k}{Q_{ij}^k} + \sum_{\substack{m,n \\ m \neq n}} \sum_{i,j} P_{ij}^{mn} \ln \frac{P_{ij}^{mn}}{Q_{ij}^{mn}}, \tag{2}$$

where k, m and $n = 1, \cdots, K$. Q denotes the matrix of proximities between the embedded points Y. The first term of the objective function preserves the intra-view correlations whereas, the second term tries to preserve the inter-view correlations by bringing embedded points y_i^m and y_j^n close to each other if their pairwise proximity score P_{ij}^{mn} is high. The above function in Eq. 2 can be rewritten using one proximity matrix defined over the whole set of frames as:

$$\mathcal{F}(Y) = \sum_{m,n} \sum_{i,j} P_{ij}^{mn} \ln \frac{P_{ij}^{mn}}{Q_{ij}^{mn}} \tag{3}$$

where the total proximity matrix is defined as

$$P_{ij}^{mn} = \begin{cases} P_{ij}^k & \text{if } m = n = k \\ P_{ij}^{mn} & \text{otherwise} \end{cases} \tag{4}$$

This construction defines a $N \times N$ similarity matrix where the diagonal blocks represent the intra-view correlations and off-diagonal blocks represent inter-view correlations.

Given this construction, the objective function in Eq. 3 reduces to the problem of stochastic neighbor embedding [30, 31] of the frames defined by the proximity matrix P. The normalized pairwise proximity matrix P can be considered as a joint probability distribution over the frames and the objective function minimizes the KL divergence between two probability distribution P and Q. Similar to the *t*-distributed SNE (*t*-SNE) [31], we define the matrix of proximities Q_{ij} between the embedded points y_i and y_j as

$$Q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{a \neq b} \left(1 + \|y_a - y_b\|^2\right)^{-1}} \tag{5}$$

Optimization: The objective function in Eq. 3 is not convex and gradient descent algorithm can be used to find the local solution. However, constant step sizes in gradient algorithm do not guarantee decrease of the objective; expensive line

searches are often needed to decrease the objective in successive steps. In contrast, algorithms such as Majorization-Minimization (MM) [32] would be guaranteed to monotonically decrease the cost in each update. MM algorithms are based on finding a tight auxiliary upper bound of a cost function and then minimizing the cost by analytically updating the parameters at each step. We therefore resort to a two phase Quadratification-Lipschitzation (QL) procedure [32] based MM algorithm to optimize Eq. 3.

3. SPARSE REPRESENTATIVE SELECTION

Once the frame embedding is over, our next goal is to find an optimal subset of all the embedded frames, such that each frame can be described as weighted linear combination of a few of the frames from the subset. The subset is then referred as the informative summary of the multi-view videos. Given the above stated goals, we formulate the following objective function on the embedded frames Y, which needs to be minimized. The objective function is

$$\Phi(C) = \frac{1}{2}\|Y - YC\|_F^2 + \lambda\|C\|_{2,1} \tag{6}$$

where $C \in \mathbb{R}^{N\times N}$ is the sparse coefficient matrix and λ is the regularization parameter that balances the weight of the two terms. $\|C\|_{2,1} \triangleq \sum_{i=1}^{N}\|C^i\|_2$ is the sum of ℓ_2 norms of the rows of C. The first term represents the error using the selected subset to reconstruct the whole set of frames and the second term follows that the minimization of Eq. 6 leads to a sparse solution for C in terms of rows, *i.e.*, the sparse coefficient matrix C contains few nonzero rows which constitute the video summary. Notice that, unlike traditional sparse coding algorithms, the formulation in Eq. 6 is constrained to have a fixed basis selection range. In other words, we set the dictionary to be the matrix of same data points Y. In video summarization, the fixed dictionary Y is logical as the representatives for summary should come from the original frame set. Notice that, our approach is also computationally efficient as the sparse coding is done in a lower-dimensional space and at the same time, it preserves the locality and correlations among the original frames which has a great impact on the summarization output.

Optimization: Eq. 6 involves convex but non-smooth terms due to the presence of ℓ_{21} norm that require special attention. Proximal methods are specifically tailored towards it because of their fast convergence rate. It finds the minimum of a cost function of the form $g(C) + h(C)$ where g is convex, differentiable but h is closed, convex and non-smooth. We use a fast proximal algorithm, FISTA [33] to solve Eq. 6 which maintains two variables in each iteration and then combines them to find the solution.

Scalability in Generating Summaries: Apart from indicating the representatives for the summary, the non-zero rows of C also provide information about the relative importance of the representatives for describing the whole videos. A higher ranking representative frame takes part in the reconstruction of many frames in the multi-view videos as compared to a lower ranked frame. This provides scalability to our summarization approach as the ranked list can be used as a scalable representation to provide summaries of different lengths as per user request (*analyze once, generate many*).

4. EXPERIMENTS

Datasets and Performance Measures: We conduct experiments using three publicly available multi-view datasets: (i) Office dataset captured with 4 stably-held web cameras in an indoor environment [1], (ii) Campus dataset taken with 4 handheld ordinary video cameras in an outdoor scene [1] and (iii) Lobby dataset captured with 3 cameras in a large lobby area [1]. We represent each video frame by a 256-dimensional feature vector obtained from a color histogram using HSV color space (16 ranges of H, 4 ranges of S, and 4 ranges of V) [5].

To provide an objective comparison, we use three quantitative measures on all experiments, including Precision, Recall and F-measure [1]. For all these metrics, the higher value indicates better summarization quality.

Implementation Details: For QL based MM algorithm, we set the parameters, as in [32] and kept constant throughout all experiments. We set the regularization parameter $\lambda = \lambda_0/\mu$, where $\mu > 1$ and λ_0 is analytically computed from the input data Y [22]. Our approach can produce both static video summary in form of key frames or dynamic summary in form of video skims. For static summary, we extract the key frames based on the nonzero rows of C and the corresponding ℓ_2 norm gives the relative importance of that frame. The generated key frames are then used to produce a skim based on the desired skim length. Moreover, one can also produce a video skim by segmenting the videos into shots and then finding the representative shots based on the nonzero rows to constitute the multi-view summary.

Compared Methods: We compare our approach with total of seven existing approaches including four baselines (ConcateAttention [7], ConcateSparse [22], AttentionConcate [7], SparseConcate [22]) that use single-view summarization approach over multi-view videos to extract summary and three state-of-the-art methods (RandomWalk [1], RoughSets [2], BipartiteOPF [4]) which are specifically designed for multi-view video summarization. The first two baselines (ConcateAttention, ConcateSparse) concatenate all the views into a single video and then apply attention model [7] and sparse coding [22] (i.e., applying Eq. 6 to the concatenated video without any embedding) respectively, whereas in the other two baselines (AttentionConcate, SparseConcate), the corresponding approach is first applied to each view and then the resulting summaries are combined to form a single summary. The purpose of comparing with single-view methods is to show that methods that attempt to extract summary from single-view videos usually do not produce an optimal set of representatives while summarizing multi-view videos. We employ the ground truth of important events reported in [1] for a fair comparison. In our approach, an event is taken to be cor-

Table 1. Performance comparison with several baselines including both single and multi-view methods applied on the three multi-view datasets. **P**: Precision in percentage, **R**: Recall in percentage and **F**: F-measure. Ours perform the best.

Methods	Office			Campus			Lobby		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
ConcateAttention [7]	100	38	55.07	56	48	51.86	31	95	81.98
ConcateSparse [22]	100	50	66.67	64	31	41.81	90	62	73.91
AttentionConcate [7]	100	46	63.01	40	28	32.66	100	70	82.21
SparseConcate [22]	94	58	71.34	66	41	50.98	87	67	77.30
RandomWalk [1]	100	61	76.19	70	55	61.56	100	77	86.81
RoughSets [2]	100	61	76.19	69	57	62.14	97	74	84.17
BipartiteOPF [4]	100	69	81.79	75	69	71.82	100	79	88.26
Ours	**100**	**70**	**81.79**	**80**	**69**	**74.09**	**100**	**79**	**88.26**

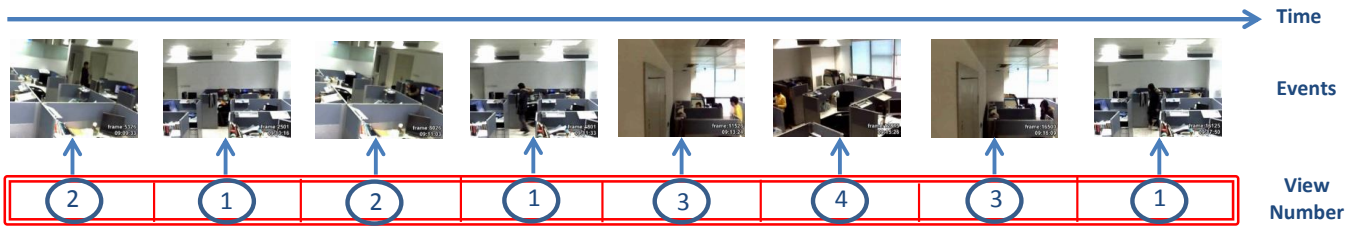

Fig. 1. Some summarized events for the Office dataset. Each event is represented by a key frame and is associated with a number that indicates the view from which the event is detected. As an illustration, we have shown only eight events arranged in temporal order. As per the ground truth [1]: A_0 represents a girl with a black coat, A_1 represents the same girl with a yellow sweater and B_0 indicates another girl with a black coat. The sequence of events are: 1st: A_0 enters the room, 2nd:A_0 stands in Cubicle 1, 3rd: A_0 is looking for a thick book to read, 4th: A_0 leaves the room, 5th: A_1 enters the room and stands in Cubicle 1, 6th: A_1 goes out of the Cubicle, 7th: B_0 enters the room and goes to Cubicle 1, and 8th: B_0 goes out of the Cubicle.

rectly detected if we get a representative frame from the set of ground truth frames between the start and end of the event.

Comparison with State-of-the-art Multi-view Summarization: Tab. 1 shows that the precision of our method as well as that of RandomWalk and BipartiteOPF are 100% for the Office and Lobby datasets and somewhat low for the Campus dataset. This is obvious since the Campus dataset contains many trivial events as it was captured in an outdoor environment, thus making the summarization more difficult. Still, for this challenging dataset, F-measure of our method is about 15% better than that of RandomWalk and 5% better than that of BipartiteOPF. Tab. 1 also reveals that while comparing to the very recent work BipartiteOPF, our method produces similar results for both Office and Lobby datasets but outperforms in the challenging Campus dataset both in precision and F-measure. Moreover, with the same precision as RandomWalk, our method produces summaries with better recall value which indicates the ability of our method in keeping more important information in the summary compared to RandomWalk. Overall, on all datasets, our approach outperforms all the baselines in terms of F-measure. This corroborates the fact that sparse representative selection coupled with multi-view frame embedding produces better summaries in contrast to the state-of-the-art methods.

Comparison with Single-view Summarization: Despite our focus on multi-view summarization, we also compare our method with several mono-view summarization approaches (ConcateAttention, ConcateSparse, AttentionConcate, SparseConcate) to show their performance on multi-view videos. Table. 1 reveals that our method significantly outperforms all the single-view baselines to generate high quality summaries. We observe that directly applying single-view summarization approaches over multi-view videos produce a lot of redundancies (simultaneous presence of most of the events) since they fail to exploit the complicated inter-view frame correlations present in multi-view videos. However, our proposed framework efficiently explores these correlations via an embedding to generate a more informative summary from multi-view videos.

Limited to the space, we only present a part of the summarized events for the Office dataset as illustrated in Fig. 1. The detected events are assembled along the time line across multiple views. Each event is represented by a key frame and is associated with a number, given inside a box below it, that illustrates the view from which the event is detected.

5. CONCLUSIONS

In this paper, we presented a novel framework for summarizing multi-view videos by exploiting the content correlations via a frame embedding. We then employed a sparse coding method over the embedding that provides scalability in generating the summaries. Our empirical study suggests that the proposed approach can effectively explore the underlying data correlations in multi-view videos and outperform all other state-of-the-art methods used in the experiments.

Acknowledgments: This work was partially supported by NSF grants IIS-1316934 and CPS-1544969.

6. REFERENCES

[1] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z. H. Zhou, "Multi View Video Summmarization," *TMM*, vol. 12, no. 7, pp. 717–729, 2004.

[2] P. Li, Y. Guo, and H. Sun, "Multi key-frame abstraction from videos," in *ICIP*, 2011.

[3] S. H. Ou, C. H. Lee, V.S. Somayazulu, Y. K. Chen, and S. Y. Chien, "On-Line Multi-View Video Summarization for Wireless Video Sensor Network," *JSTSP*, vol. 9, no. 1, pp. 165–179, 2015.

[4] S. Kuanar, K. Ranga, and A. Chowdhury, "Multi-view video summarization using bipartite matching constrained optimum-path forest clustering," *TMM*, vol. 17, no. 8, pp. 1166–1173, 2015.

[5] J. Almeida, N. J. Leite, and R. S. Torres, "VISON: VIdeo Summarization for ONline applications," *PRL*, vol. 33, no. 4, pp. 397–409, 2012.

[6] G. Guan, Z. Wang, S. Mei, M. Ott, M. He, and D. D. Feng, "A Top-Down Approach for Video Summarization," *TOMCCAP*, vol. 11, no. 4, pp. 56–68, 2014.

[7] Y. F. Ma, X.S Hua, and H.J. Zhang, "A Generic Framework of User Attention Model and Its Application in Video Summarization," *TMM*, vol. 7, no. 5, pp. 907–919, 2005.

[8] M. Gygli, H. Riemenschneider H. Grabner, and L. V. Gool, "Creating summaries from user videos," in *ECCV*, 2014.

[9] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *ECCV*, 2014.

[10] A. Khosla, R. Hamid, C. J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *CVPR*, 2013.

[11] G. Kim, L. Sigal, and E. P. Xing, "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *CVPR*, 2014.

[12] M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *CVPR*, 2015.

[13] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *CVPR*, 2015.

[14] B. Gong, W. L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," in *NIPS*, 2014.

[15] W. S. Chu, Y. Song, and A. Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *CVPR*, 2015.

[16] H Boukadida, S.A Berrani, and P Gros, "Automatically creating adaptive video summaries using constraint satisfaction programming: Application to sport content," *TCSVT*, 2015.

[17] F Chen, C De Vleeschouwer, and A Cavallaro, "Resource allocation for personalized video summarization," *TMM*, vol. 16, no. 2, pp. 455–469, 2014.

[18] B. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *TOMM*, vol. 3, no. 1, 2007.

[19] A. G. Money and H. Agius, "Video summarisation: A conceptual framework and survey of the state of the art," *JVCIR*, vol. 19, no. 2, pp. 121–143, 2008.

[20] C. De Leo and B. S. Manjunath, "Multicamera video summarization from optimal reconstruction," in *ACCV Workshop*, 2011.

[21] C. De Leo and B. S. Manjunath, "Multicamera Video Summarization and Anomaly Detection from Activity Motifs," *ACM Transaction on Sensor Networks*, vol. 10, no. 2, pp. 1–30, 2014.

[22] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *CVPR*, 2012.

[23] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *CVPR*, 2014.

[24] Y. Cong, J. Yuan, and J. Luo, "Towards Scalable Summarization of Consumer Videos Via Sparse Dictionary Selection," *TMM*, vol. 14, no. 1, pp. 66–75, 2012.

[25] S Mei, G Guan, Z Wang, S Wan, M He, and D. D Feng, "Video summarization via minimum sparse reconstruction," *PR*, vol. 48, no. 2, pp. 522–533, 2015.

[26] Y Pan, T Yao, T Mei, H Li, C-W Ngo, and Y Rui, "Click-through-based cross-view learning for image search," in *SIGIR*, 2014.

[27] C Xu, D Tao, and C Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *PAMI*, vol. 22, no. 8, pp. 885–905, 2000.

[29] G. Scott and H. Longuett-Higgins, "An algorithm for associating the features of two images," *The Royal Society of London*, 1991.

[30] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *NIPS*, 2002.

[31] L. V. Maaten and G. Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, pp. 2579–2605, 2008.

[32] Z. Yang, J. Peltonen, and S. Kaski, "Majorization-minimization for manifold embedding," in *AISTATS*, 2015.

[33] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.