Supplementary Material for “Ensemble Diffusion for Retrieval”

Song Bai[1]*, Zhichao Zhou[1]*, Jingdong Wang[2], Xiang Bai[1]†, Longin Jan Latecki[3], Qi Tian[4]
[1]Huazhong University of Science and Technology, [2]Microsoft Research Asia
[3]Temple University, [4]University of Texas at San Antonio
{songbai,zzc,xbai}@hust.edu.cn, jingdw@microsoft.com, latecki@temple.edu, qi.tian@utsa.edu

The document contains the supplementary materials for “Ensemble Diffusion for Retrieval”. The primary goal of this document is to provide the proof of used propositions in Sec. 1, which are omitted in the main paper due to the space limitation.

As a secondary goal, we present additional experiments for a more comprehensive evaluation, including the quantitative evaluations in Sec. 2, the qualitative evaluations in Sec. 3, and the parameter discussions in Sec. 4, respectively.

1. Propositions

Proposition 1. *Eq. (5) converges to Eq. (6).*

Proof. By applying $vec(\cdot)$ to both sides of Eq. (6), we obtain

$$\tilde{A}^{(t+1)} = \alpha \mathbb{S} \tilde{A}^{(t)} + (1-\alpha)\tilde{I},$$

where $\mathbb{S} = S^{(1)} \otimes S^{(2)}$. Therefore, $\tilde{A}^{(t+1)}$ can be expanded as

$$\tilde{A}^{(t+1)} = (\alpha\mathbb{S})^t \tilde{A}^{(1)} + (1-\alpha)\sum_{i=0}^{t-1}(\alpha\mathbb{S})^i \tilde{I}.$$

It is known that the spectral radius of both $S^{(1)}$ and $S^{(2)}$ are no larger than 1. Hence, all the eigenvalues of $\mathbb{S}$ are in $[-1, 1]$. Considering that $0 < \alpha < 1$, we have

$$\lim_{t\to\infty} (\alpha\mathbb{S})^t \tilde{A}^{(1)} = 0,$$
$$\lim_{t\to\infty} (1-\alpha)\sum_{i=0}^{t-1}(\alpha\mathbb{S})^i \tilde{I} = (1-\alpha)(I-\alpha\mathbb{S})^{-1}\tilde{I}.$$

Therefore, one can easily induce that

$$\lim_{t\to\infty} \tilde{A}^{(t+1)} = (1-\alpha)(I-\alpha\mathbb{S})^{-1}\tilde{I},$$

which is identical to Eq. (6) after applying vec^{-1} to both sides. □

Proposition 2. *The closed-form solution of Eq. (7) is Eq. (6).*

Proof. Define $Y \equiv N(i-1)+k$, $Z \equiv N(j-1)+l$, $\mathbb{D} = D^{(1)} \otimes D^{(2)} \in \mathbb{R}^{N^2\times N^2}$ and $\mathbb{W} = W^{(1)} \otimes W^{(2)} \in \mathbb{R}^{N^2\times N^2}$. Then the left term of Eq. (7) can be transformed to

$$\begin{aligned}
&\frac{1}{2}\sum_{Y,Z=1}^{N^2} \mathbb{W}_{YZ}\left(\frac{\tilde{A}_Y}{\sqrt{\mathbb{D}_{YY}}} - \frac{\tilde{A}_Z}{\sqrt{\mathbb{D}_{ZZ}}}\right)^2 \\
=&\sum_{Y,Z=1}^{N^2} \mathbb{W}_{YZ}\frac{\tilde{A}_Y^2}{\mathbb{D}_{YY}} - \sum_{Y,Z=1}^{N^2} \tilde{A}_Y \frac{\mathbb{W}_{YZ}}{\sqrt{\mathbb{D}_{YY}\mathbb{D}_{ZZ}}}\tilde{A}_Z \\
=&\sum_{Y=1}^{N^2} \tilde{A}_Y^2 - \tilde{A}^{\mathrm{T}}\mathbb{D}^{-0.5}\mathbb{W}\mathbb{D}^{-0.5}\tilde{A} \\
=&\tilde{A}^{\mathrm{T}}\left(I - \mathbb{D}^{-0.5}\mathbb{W}\mathbb{D}^{-0.5}\right)\tilde{A}, \\
=&\tilde{A}^{\mathrm{T}}(I-\mathbb{S})\tilde{A}.
\end{aligned}$$

The right term can be described as $\mu\|\tilde{A} - \tilde{I}\|_2^2$. Therefore, the objective function in Eq. (7) becomes

$$\min_{\tilde{A}} \tilde{A}^{\mathrm{T}}(I-\mathbb{S})\tilde{A} + \mu\|\tilde{A}-\tilde{I}\|_2^2,$$

whose partial derivative is

$$2(I-\mathbb{S})\tilde{A} + 2\mu(\tilde{A}-\tilde{I}).$$

In order to get the optimal solution of the above problem, set the derivative to zero and substitute $\mu = \frac{1}{\alpha} - 1$. We have

$$\tilde{A} = (1-\alpha)(I-\alpha\mathbb{S})^{-1}\tilde{I},$$

which is equivalent to Eq. (6) after applying vec^{-1} to both sides. □

Proposition 3. *Eq. (14) converges to the solution in Eq. (12).*

Proof. Eq. (14) can be vectorized to

$$\begin{aligned}
\tilde{A}^{(t+1)} &= \mathbb{S}\tilde{A}^{(t)} + (1 - \sum_{v=1}^{M}\alpha_v)\tilde{I} \\
&= \mathbb{S}^t\tilde{A}^{(1)} + (1 - \sum_{v=1}^{M}\alpha_v)\sum_{i=0}^{t-1}\mathbb{S}^i\tilde{I}.
\end{aligned}$$

*indicates equal contributions.

†corresponding author.

where $\mathbb{S} = \sum_{v=1}^{M} \alpha_v \mathbb{S}^v$.

Similar to Proposition 1, we only need to prove that all the eigenvalues of $\mathbb{S}$ are in $(-1, 1)$. Since all the eigenvalues of $\mathbb{S}^v (1 \leq v \leq M)$ are in $[-1, 1]$, the spectral radius of $\mathbb{S}$ is bounded by $\sum_{v=1}^{M} \alpha_v$. Considering $\mu > 0$, $\beta_v > 0$ and Eq. (13), we have

$$\sum_{v=1}^{M} \alpha_v = \frac{\sum_{v'=1}^{M} \beta_{v'}}{\mu + \sum_{v'=1}^{M} \beta_{v'}} < 1.$$

Therefore, we have

$$\lim_{t \to \infty} \mathbb{S}^t \tilde{A}^{(1)} = 0,$$

Then,

$$\lim_{t \to \infty} \tilde{A}^{(t+1)} = (1 - \sum_{v=1}^{M} \alpha_v)(I - \sum_{v=1}^{M} \alpha_v \mathbb{S}^v)^{-1} \tilde{I}.$$

The proof is complete. □

2. Quantitative Evaluation

2.1. Face Retrieval

We follow [3] to evaluate the performances on ORL face dataset. ORL dataset is comprised of 400 grayscale face images, divided into 40 categories. The evaluation metric is called bull's eye score, which counts the average recall before the top-15 candidates in the ranking list. For each face image, we extract 128 dimensional SIFT [6], 124 dimensional HoG [2], 232 dimensional LBP [8] and 512 dimensional GIST [9]. Their baseline performances are 84.95%, 73.70%, 73.15% and 83.67% respectively.

The detailed performances of tensor product fusion is given in Table 1, and the performance comparison of different fusion methods is given in Table 2. Again, RED yields the best performances. Compared with [3] which enumerates 72 kinds of diffusion processes (by varying 4 different affinity initializations, 6 different transition matrices and 3 different update schemes), RED gives nearly 20 percent gain in the performance. The reason for the performance gain are two folds. First, instead of using only one similarity, we leverage more than two similarities, which can be well handled by the proposed framework. Second, the proposed RED is a more robust fusion with diffusion method with automatic weight learning.

2.2. The performances of Tensor Product Fusion

Table 3 and Table 4 present the detailed retrieval performances of tensor product fusion on Holidays dataset [5] and Ukbench dataset [7], respectively.

The best results (marked in red) and the worse results (marked in blue) are the upper and the lower bounds of the performances of tensor product fusion, which are presented in the main paper.

	SIFT	HOG	LBP	GIST
SIFT	-	90.00	89.75	**97.17**
HOG	90.60	-	87.25	92.45
LBP	90.50	**87.13**	-	94.38
GIST	96.83	91.60	93.05	-

Table 1. The bull's eye scores of tensor product fusion on ORL dataset.

Methods	Bull's eye score
Generic Diffusion Process [3]	77.42
Naive Fusion	93.53
Tensor Product Fusion	87.13∼97.17
RED	**97.75**

Table 2. The performance (%) comparison on ORL face dataset.

	NetVLAD	SPoC	ResNet	HSV
NetVLAD	-	92.36	91.85	88.24
SPoC	**92.46**	-	90.02	87.55
ResNet	91.85	90.09	-	85.44
HSV	87.77	87.33	**85.12**	-

Table 3. The mAPs of tensor product fusion on Holidays dataset.

	NetVLAD	SPoC	ResNet	HSV
NetVLAD	-	3.871	3.874	**3.626**
SPoC	3.876	-	3.854	3.629
ResNet	**3.884**	3.861	-	3.629
HSV	3.680	3.685	3.682	-

Table 4. The N-S scores of tensor product fusion on Ukbench dataset.

3. Qualitative Evaluation

In this section, we present three sample retrieval results for qualitative evaluations on ModelNet40 dataset [12]. The retrieval results of the 4 baseline similarities (Volumetric CNN [11], GIFT [1], ResNet [4], PANORAMA [10]) are presented in the first 4 rows. The retrieval results of the 3 fusion methods (naive fusion, tensor product fusion, RED) are presented in the next 3 rows. The performances of tensor product fusion are obtained by fusing ResNet [4] and GIFT [1].

As can be seen from Fig. 2 and Fig. 3, fusion methods clearly outperform baseline similarities. Moreover, RED yield 100% retrieval precision in the top-10 retrieved list, which is superior to other fusion methods.

In Fig. 4, one can observe that all the 4 baseline similarities fail with this query. However, by exploiting the complementary nature among them, RED still improves the retrieval performance remarkably.

4. Parameter Discussion

In Fig. 1(a), we plot the influence of μ used by RED on the retrieval performance with Holidays dataset. As can be seen, RED is not sensitive to the change of μ as long as it is in a reasonable range.

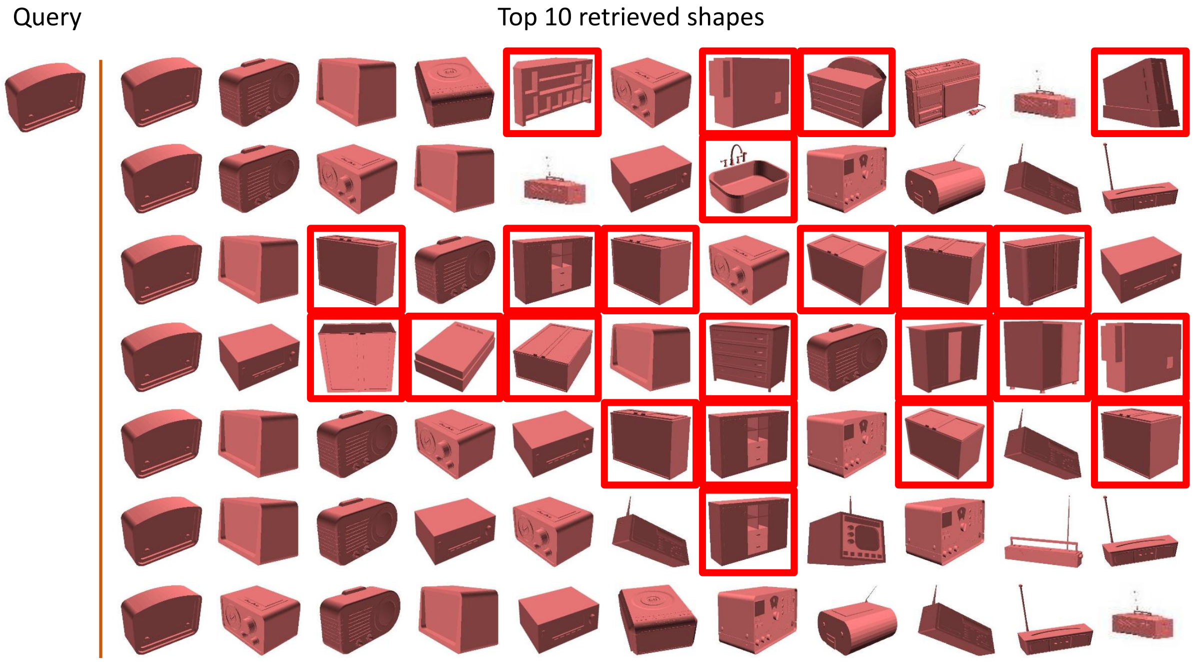


Figure 2. The first sample retrieval results on ModelNet40 dataset. False positives are in red boxes

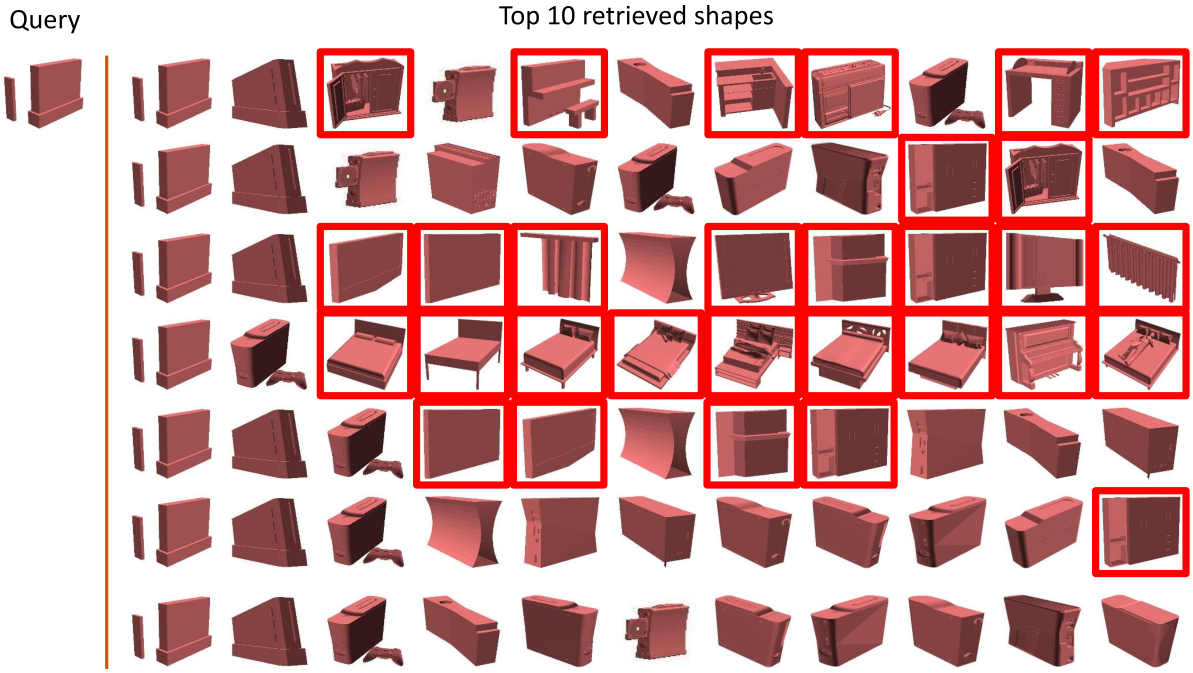


Figure 3. The second sample retrieval results on ModelNet40 dataset. False positives are in red boxes

As suggested in [3, 13], it is crucial to determine the number of nearest neighbors k on the affinity graph. Fig. 1(b) analyzes the influence of k on the retrieval performance with Holidays dataset. It can be observed that when $k \geq 5$, the performance is significantly improved (around 93 in mAP). When k keeps increasing, the performance drops slightly due to the inclusion of noisy edges on the affinity graph. We can infer that it is still an open issue to automatically select a proper k on the affinity graph.

References

[1] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki. Gift: A real-time and scalable 3d shape search engine. In *CVPR*, 2016. 2

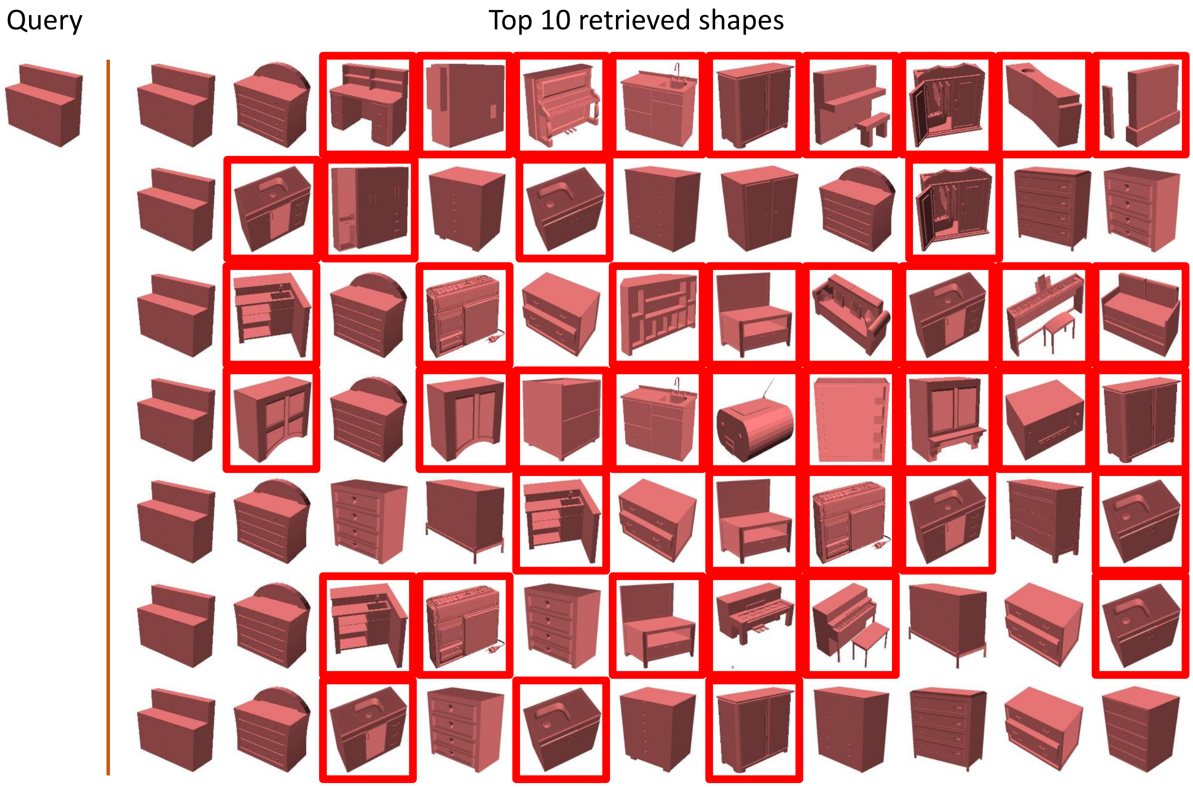
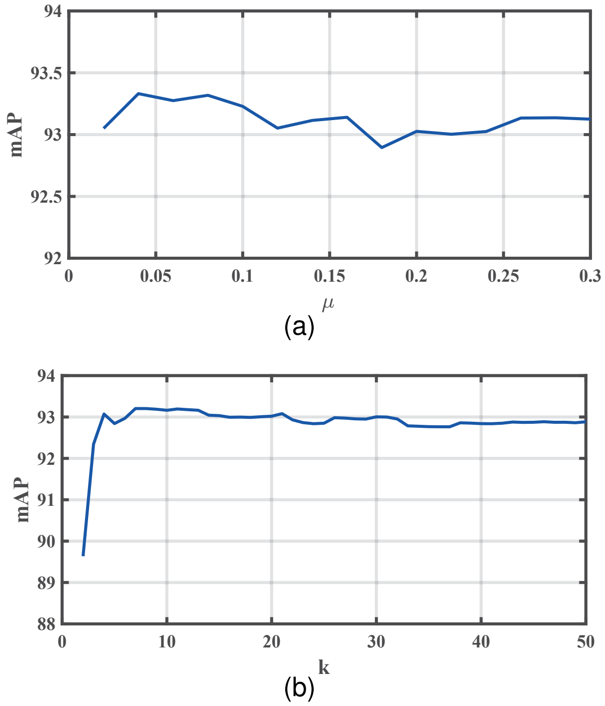


Figure 4. The third sample retrieval results on ModelNet40 dataset. False positives are in red boxes.



(a)

(b)

Figure 1. The influence of μ and the number of nearest neighbors k on Holidays dataset.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 2

[3] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *CVPR*, pages 1320–1327, 2013. 2, 3

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, 2008. 2

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[7] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, 2006. 2

[8] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002. 2

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 2

[10] P. Papadakis, I. Pratikakis, T. Theoharis, and S. J. Perantonis. Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. *IJCV*, 89(2-3):177–192, 2010. 2

[11] C. R. Qi, H. Su, M. Niessner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *CVPR*, 2016. 2

[12] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shape modeling. In *CVPR*, 2015. 2

[13] X. Yang, S. Koknar-Tezel, and L. J. Latecki. Locally constrained diffusion process on locally densified distance s-paces with applications to shape retrieval. In *CVPR*, pages 357–364, 2009. 3