Supplementary Material for Siamese Networks: The Tale of Two Manifolds

Soumava Kumar Roy[1,4], Mehrtash Harandi[2,4], Richard Nock[1,3,4], Richard Hartley[1]
[1]The Australian National University; [2]Monash University;
[3]The University of Sydney; [4]DATA61-CSIRO, Australia
Soumava.KumarRoy@anu.edu.au, Mehrtash.Harandi@monash.edu,
Richard.Nock@data61.csiro.au, Richard.Hartley@anu.edu.au

1. Riemannian SGD with Momentum (rSGD-M)

We devised the Riemannian extension of the NAG algorithm or rNAG in § 4 of our paper. In doing so, we rely heavily on differential geometry. To do justice, we review some basics below.

Riemannian Geometry

Definition 1 (Manifold)**.** *A manifold is a locally Euclidean Hausdorff space whose topology has a countable base.*

Locally Euclidean just means that each point has a neighborhood that is homeomorphic to an open ball in $\mathbb{R}^n$ for some n. Moreover, being a Hausdorff space means that distinct points have disjoint neighborhoods. This property is useful for establishing the notion of a differential manifold, as it guarantees that convergent sequences have a single limit point. The exponential map $\mathrm{Exp}_{\boldsymbol{x}}(\cdot) : T_{\boldsymbol{x}}\mathcal{M} \rightarrow \mathcal{M}$ and its inverse, the logarithm maps, $\mathrm{Log}_{\boldsymbol{x}}(\cdot) : \mathcal{M} \rightarrow T_{\boldsymbol{x}}\mathcal{M}$ are defined over Riemannian manifolds to switch between the manifold and its tangent space at $\boldsymbol{x}$. The exponential operator maps a tangent vector Δ to a point $\boldsymbol{y}$ on the manifold. The property of the exponential map ensures that the length of Δ becomes equal to the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. The logarithm map is the inverse of the exponential map, and maps a point on the manifold to the tangent space $T_{\boldsymbol{x}}$. The exponential and logarithm maps vary as point $\boldsymbol{x}$ moves along the manifold.

A natural way to measure nearness on a manifold is by considering the geodesic distance between two points on the manifold. Such distance is defined as the length of the shortest curve connecting the two points. The shortest curves are known as geodesics and are analogous to straight lines in $\mathbb{R}^n$. The tangent space at a point $\boldsymbol{x}$ on the manifold, $T_{\boldsymbol{x}}\mathcal{M}$, is a vector space that consists of the tangent vectors of all possible curves passing through $\boldsymbol{x}$. Note that we will assume that the key conditions needed for these maps to exist are satisfied.

Remark 1. *In devising the rNAG, we opt for the most general and mathematically rigorous solution that benefits from the exponential map and its inverse. For the sake of discussion, We repeat this general form of rNAG below.*

$$\theta^{(t)} = \mathrm{Exp}_{\boldsymbol{m}}\Big(-\eta \mathrm{grad}_{\boldsymbol{m}}\big(J\big)\Big) , \tag{1}$$

$$\boldsymbol{m} = \mathrm{Exp}_{\theta^{(t)}}\Big(-\mu \mathrm{Log}_{\theta^{(t)}}\big(\theta^{(t-1)}\big)\Big) . \tag{2}$$

In many cases of interest, such mappings are known. However, for some manifolds this is not the cases, meaning that even such mappings are not known (or a closed-form might not yet available). Furthermore, even if those mappings are at our disposal, computationally they might not be the most attractive solutions. In Riemannian optimization, it is very common to replace the exponential map with a local approximation of it known as a retraction $r_{\boldsymbol{x}}(\cdot) : T_{\boldsymbol{x}}\mathcal{M} \rightarrow \mathcal{M}$. Moreover, if projection onto the tangent space of a manifold is known, the logarithm map can be approximated easily. This is indeed the case for the quotient geometry developed in our paper. As such, a computationally more attractive version of rNAG can be written as;

$$\theta^{(t)} = r_{\boldsymbol{m}}\Big(-\eta \mathrm{grad}_{\boldsymbol{m}}\big(J\big)\Big) , \tag{3}$$

$$\boldsymbol{m} = r_{\theta^{(t)}}\Big(-\mu \pi_{\theta^{(t)}}\big(\theta^{(t)} - \theta^{(t-1)}\big)\Big) . \tag{4}$$

In Eqn. (4) $\pi_{\boldsymbol{x}}(\cdot) : \mathbb{R}^n \rightarrow T_{\boldsymbol{x}}\mathcal{M}$ denotes the mapping from the embedded space $\mathbb{R}^n$ onto the tangent space at $\boldsymbol{x}$. In the developed quotient geometry, this is indeed the horizontal part of a tangent vector.

Empirical Evaluations

In this part, we empirically compare rNAG against two baselines, **1.** Riemannian Stochastic Gradient Descent (rSGD) [2] and **2.** Riemannian SVRG (rSVRG) [14]. Briefly,

- **rSGD** is the extension of SGD to the Riemannian manifolds and under some mild conditions enjoys convergence properties.

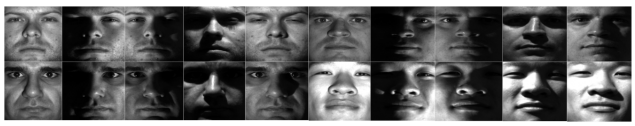

Figure 1. Examples of the YaleB Dataset [9].

- **rSVRG** is the extension of SVRG [6] to the Riemannian framework to solve constrained problems. rSVRG cyclically stores an optimal estimate of the parameters and ensures the subsequent updates do not deviate too far from this store optimal estimate.

Towards our goal, we will consider two classical learning tasks, namely PCA and and compare rNAG to rSGD and rSVRG. Before delving deeper, our tests (which simply go beyond the two experiments here) show that rNAG converges faster and usually to a lower loss as compared to rSGD. Comparing to rSVRG, we have observed that rNAG usually converges faster but in some cases (*e.g.*, second experiment here) the solutions obtained by rNAG and rSVRG are close (which is indeed a positive result). Having said this, a proper and detailed study of the rNAG algorithm deserves a separate and dedicated work. We believe that our current work is the first step to employ rNAG algorithm in training deep structures.

1.1. Experiment#1. PCA

The PCA objective function is to minimize a form of reconstruction error as

$$J(\boldsymbol{U}) = \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X} \right\|_F^2, \quad \text{s.t.} \quad \boldsymbol{U}^\top \boldsymbol{U} = \mathbf{I}_p \,. \tag{5}$$

Here, $\boldsymbol{X} \in \mathbb{R}^{n \times N}$ is a matrix whose columns are N data points and $\boldsymbol{U} \in \mathbb{R}^{n \times p}$ is the PCA projection. We stress that while (5) can be solved by eigenvalue decomposition, it is a classical problem to study when it comes to Riemannian optimization [3, 14] and hence our choice here. The orthogonality constraint may imply that this problem shall be addressed using the geometry of the Stiefel manifold. However, we note that the objective function satisfies $J(\boldsymbol{U}) = J(\boldsymbol{U}\boldsymbol{R})$, for $\boldsymbol{R} \in \mathcal{O}(p)$. This indeed leads to a quotient space of the Stiefel manifold which is known as the Grassmannian. Formally, a Grassmann manifold $\mathcal{G}(p,n)$ is the space of p-dimensional linear subspaces of $\mathbb{R}^n$ for $0 < p < n$ [1]. As such minimizing J can be view as an optimization problem over the Grassmann manifold $\mathcal{G}(p,n)$

$$\min_{\boldsymbol{U} \in \mathcal{G}(p,n)} \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X} \right\|_F^2 \tag{6}$$

In Fig. 2, we compare the results of **rNAG** against rSGD and rSVRG for solving (6) on $\mathcal{G}(256, 2016)$ ($2016 = 48 \times 42$). In this experiment, we use the YaleB [9] dataset which consists of images of 38 subjects captured at 64 different illumination conditions. We have used a preprocessed version of the database, where each face is cropped and downsampled to a 48×42 image. We use the first 52 images of every subject for training (*YALEB-Train*) and the remaining 12 images for testing (*YALEB-Test*) (see Fig. 1 for examples). Empirically, we observed very similar behavior for the studied algorithms when the dimensionality of the subspaces varies (*i.e.*, p). Studying Fig. 2 clearly shows that rNAG is superior to both rSGD and rSVRG when speed and accuracy of the solution are considered.

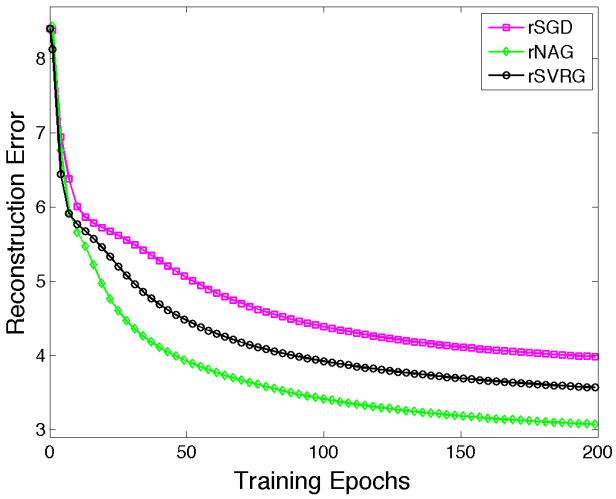

Figure 2. Convergence behavior of the rNAG, rSGD and rSVRG to address the PCA problem using $\mathcal{G}(256, 2016)$.

1.2. Experiment#2. Fréchet Mean

Symmetric Positive Definite (SPD) matrices are imperative in computer vision [5]. As our second experiment, we consider the problem of computing the Freéchet mean of a set of SPD matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \cdots \boldsymbol{A}_k$ such that $\boldsymbol{A}_i \in \mathcal{S}_{++}^n$, $i = 1, \ldots k$. The objective function can be expressed as the follows

$$\min \sum_i^k \left\| \log(\boldsymbol{A}_i^{-1/2} \boldsymbol{M} \boldsymbol{A}_i^{-1/2}) \right\|_F^2 \quad \text{s.t.} \quad \boldsymbol{M} \in \mathcal{S}_{++}^n \,. \tag{7}$$

In (7), log denotes the principal matrix logarithm [4] and should not be confused with the logarithm map on a Riemannian manifold. For this experiment, we use the Kylberg texture dataset [8]. The Kylberg dataset consists of patches from 28 different texture classes with 160 unique samples per class. We scaled the images to 128×128 pixels and computed 5×5 covariance descriptors [11] using following

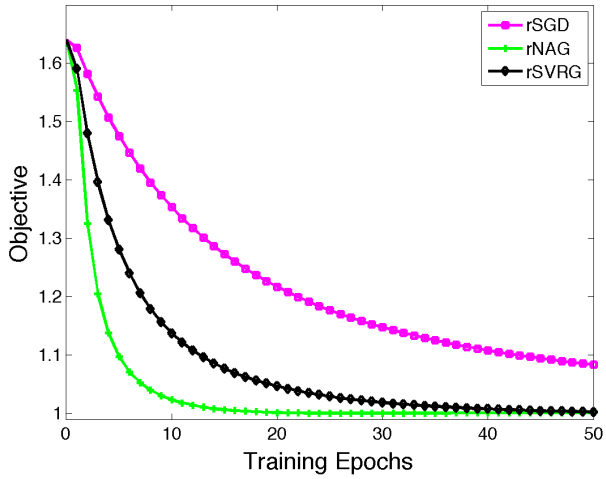

Figure 3. Convergence behavior of the rNAG, rSGD and rSVRG for computing the Freéchet Mean on $\mathcal{S}^5_{++}$.

features

$$x_{u,v} = \left[I_{u,v}, \left|\frac{\partial I}{\partial u}\right|, \left|\frac{\partial I}{\partial v}\right|, \left|\frac{\partial^2 I}{\partial u^2}\right|, \left|\frac{\partial^2 I}{\partial v^2}\right| \right],$$

where $I_{u,v}$ represents the intensity at location (u, v). Fig. 3 plots the behavior of rNAG, rSGD and rSVRG for this experiment. As before, rNAG is superior to both rSGD and rSVRG when the convergence speed is considered. In terms of the performance and the value of the objective function, rNAG is clearly superior to rSGD while providing very similar results to rSVRG.

2. Overview of Datasets used for Fine-Grained Image Classification

- The **CUB-200-2011** [13] consists of 11,788 images of birds from 200 different varieties. The first 100 categories are considered for training (5,864 images), while the rest 100 categories are considered for testing (5,924 images).

- The **CARS196** dataset [7] consists of 16,185 images of cars from 196 different categories. The first 98 categories (8,054 images) are considered in the training of the network, while the next 98 categories (8,131 images) are used in the testing phase.

- The **SOP** dataset [10] consist of 120,053 images of 22,634 different products sold in eBay.com . The first 11,318 categories is used for training and the remaining 11,316 categories is used for testing.

Fig. 4 shows examples of the three datasets.

(a) CUB-200-2011 [13]

(b) CARS196 [7]

(c) SOP [10]

Figure 4. Exemplar samples of the three fine-grained image classification datasets used for evaluation of qConv and Stiefel layers.

Qualitative Measures

Apart from the quantitative measures reported in the paper, here we provide the Barnes-Hut t-SNE visualization [12] of the proposed geometrical embedding spaces, *i.e.* qConv and Stiefel, for all the three datasets used for in the fine-grained image classification experiments. Fig. 5, 6 and 7 and Fig. 8, 9 and 10 show the t-SNE plots for the Stiefel and qConv embedding configurations respectively for CUB-200-2011, CARS196 and SOP datasets. These plots are best viewed when zoomed in.

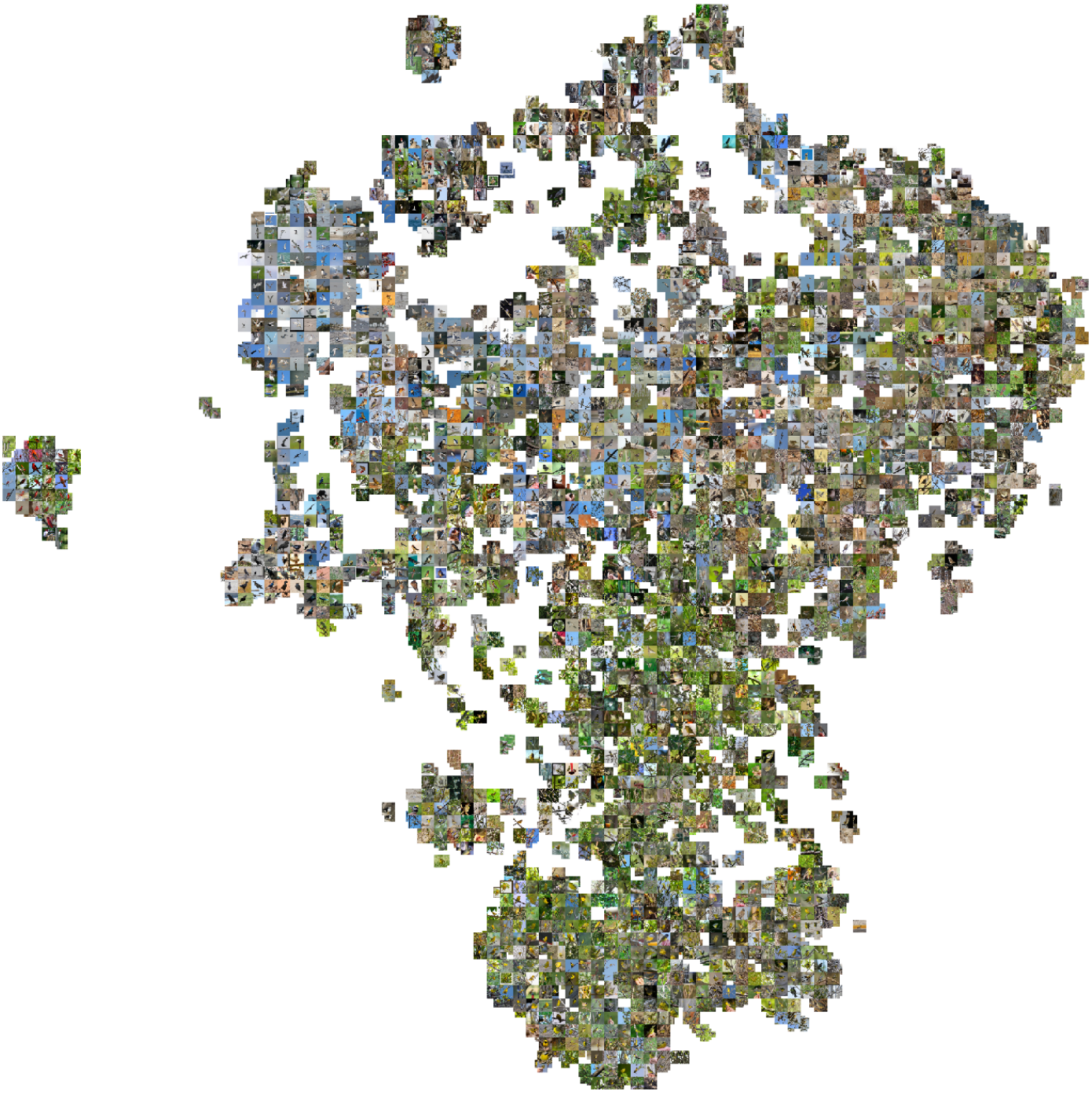

Figure 5. Barnes-Hut t-SNE [12] visualization of our Stiefel embedding on the test set of the CUB-200-2011 [13]. Best viewed when zoomed in.

Figure 6. Barnes-Hut t-SNE [12] visualization of our Stiefel embedding on the test set of the CARS196 [7]. Best viewed when zoomed in.

Figure 7. Barnes-Hut t-SNE [12] visualization of our Stiefel embedding on the 20000 random samples chosen from the test set of the SOP [10]. Best viewed when zoomed in.

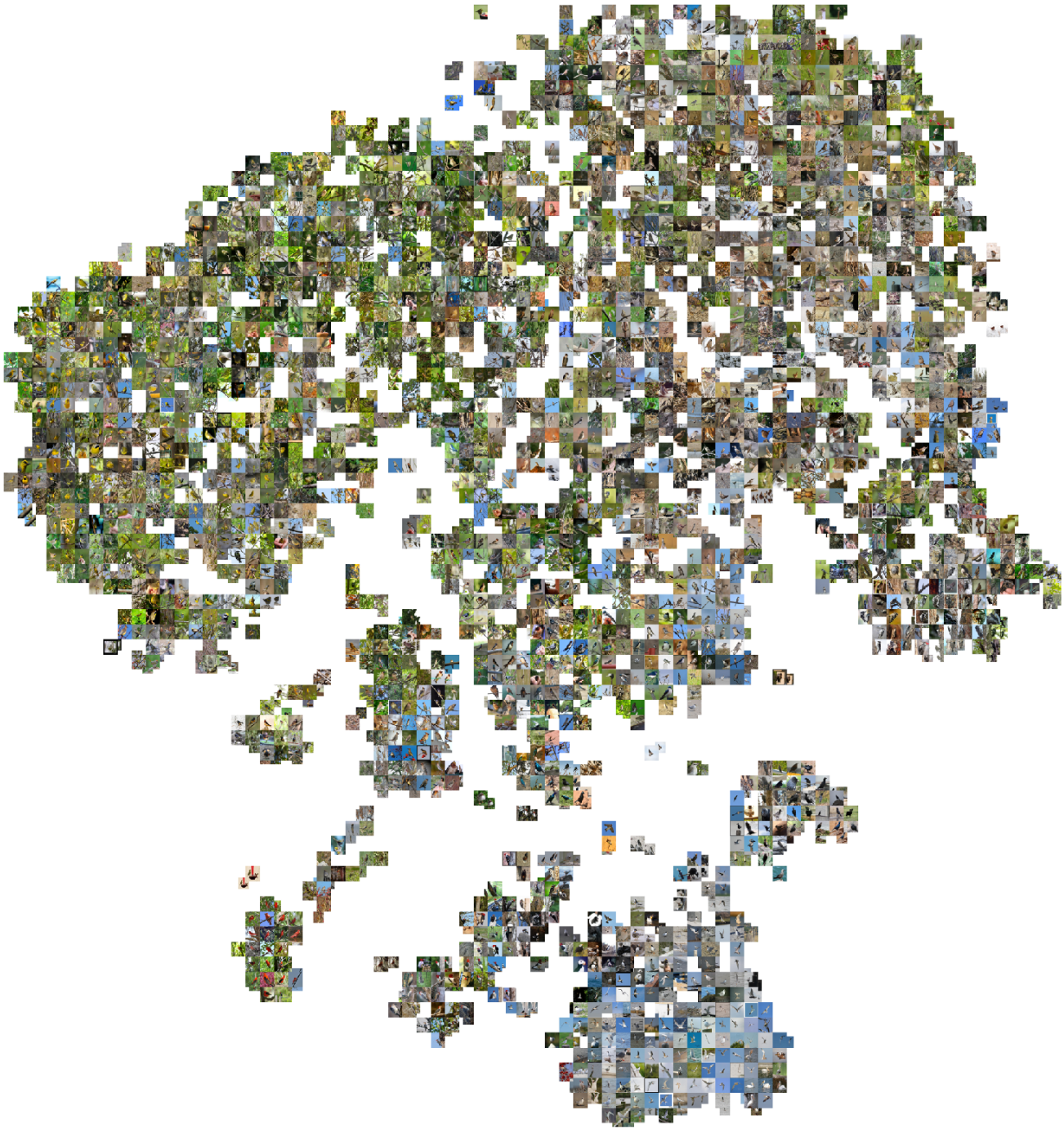

Figure 8. Barnes-Hut t-SNE [12] visualization of our qConv embedding on the test set of the CUB-200-2011 [13]. Best viewed when zoomed in.

Figure 9. Barnes-Hut t-SNE [12] visualization of our qConv embedding on the test set of the CARS196 [7]. Best viewed when zoomed in.

Figure 10. Barnes-Hut t-SNE [12] visualization of our qConv embedding on the 20000 random samples chosen from the test set of the SOP [10]. Best viewed when zoomed in.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. 2

[2] Silvere Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 1

[3] John P Cunningham and Zoubin Ghahramani. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, 2015. 2

[4] Gene H Golub and Charles F Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 4 edition, 2013. 2

[5] M. Harandi, M. Salzmann, and R. Hartley. Dimensionality Reduction on SPD Manifolds: The Emergence of Geometry-Aware Methods. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 2

[6] Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *NIPS*, pages 315–323, 2013. 2

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d Object Representations for Fine-Grained Categorization. In *Proc. of the IEEE Int. Conf. on Computer Vision Workshops*, 2013. 3, 5, 8

[8] Gustaf Kylberg. *Kylberg Texture Dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, 2011. 2

[9] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring Linear Subspaces for Face Recognition under Variable Lighting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005. 2

[10] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. In *CVPR*, pages 4004–4012, 2016. 3, 6, 9

[11] Oncel Tuzel, Fatih Porikli, and Peter Meer. *Region Covariance: A Fast Descriptor for Detection and Classification*, pages 589–600. Springer, 2006. 2

[12] Laurens Van Der Maaten. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014. 3, 4, 5, 6, 7, 8, 9

[13] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 3, 4, 7

[14] Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds. In *NIPS*, pages 4592–4600, 2016. 1, 2