Supplementary Material for Diffeomorphism Matching for Fast Unsupervised Learning on Radiographs

Thanh M. Huynh*[1]
v.thanhhuynh@vinbrain.net
Chanh D. T. Nguyen*[1, 2]
v.chanhndt@vinbrain.net
Ta Duc Huy[1]
v.huyta@vinbrain.net
Hoang Cao Huyen[1]
v.huyenhc@vinbrain.net
Trung H. Bui[3]
bhtrung@yahoo.com
Steven QH Truong[1]
brain01@vinbrain.net

[1] VinBrain JSC
7 Bang Lang, Viet Hung District
Ha Noi, Vietnam

[2] VinUniversity
Vinhomes Ocean Park, Gia Lam District
Ha Noi, Vietnam

[3] Independant Researcher
USA

1 Proof of theorem

Theorem 2. *if T and S are C^1 mapping from data manifold M to some feature manifolds $T(M)$ and $S(M)$ such that there exists a diffeomorphism H with the property $H \circ T = S$ and $H^{-1} \circ S = T$, then T and S are equivalent.*

Proof. Consider 2 data points $x, y \in \mathcal{M}$, such that $T(x) = T(y)$, one has

$$H \circ T(x) = H \circ T(y) \tag{1}$$

with the property $H \circ T = S$, Eq. 1 reduces to

$$S(x) = S(y) \tag{2}$$

Same argument can be applied in the direction of H^{-1}, which implies S and T are equivalent. □

2 Medical Imaging Architecture Design Space

We studied many neural architecture designs using the approach of RegNet [2]. Our design space consists of squeeze ratio s for SE module [1], width multiplier wm, number of downsampling stages s, and bottleneck ratio b for residual skip connection. We sample a total of

*Equal Contribution.

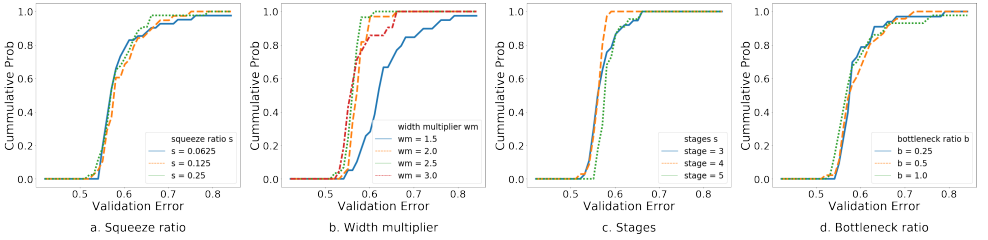

Figure 1: EDF of model's error on the CheXpert validation set for various network design choices of squeeze ratio s, width multiplier wm, stages s, and bottleneck ratio b. The further the chart is to the left, the more models with specific design choice concentrate in low validation error region.

2K models with various initial witdth and $s \in \{0.0625, 0.125, 0.25\}$, $wm \in \{1.5, 2, 2.5, 3\}$, $s \in \{3, 4, 5\}$, and $b \in \{0.25, 0.5, 1.0\}$. Fig. 1 shows the error distribution of models with various design choices. Surprisingly, many configurations of SE module's squeeze ratio and bottleneck ratio have the same performance in medical imaging. For width multiplier *wm*, the findings are the same as in the original paper, i.e., higher width multiplication leads to better performance. For the number of stages s, it's intriguing that models with $s = 5$ have lower performance than models with $s = 4$. We hypothesize that due to the locality of the abnormal findings, going deeper resulted in adding more noisy information to local features.

3 Matching Error Distribution

Diffeomorphism Matching loss can be expressed as

$$\begin{aligned}\mathcal{L}_{\text{DM}} = \sum_{x \in D} (&\|HS(x) - T(x)\|_2^2 + \|H^{-1}HS(x) - S(x)\|_2^2 \\ &+ \|H^{-1}T(x) - S(x)\|_2^2 + \|HH^{-1}T(x) - T(x)\|_2^2),\end{aligned} \tag{3}$$

where we call the first term as Student Projection loss, the second term as Student Cyclic loss, the third term as Teacher Projection loss, and the fourth term as Teacher Cyclic loss respectively.

The standard deviations of Student Cyclic and Teacher Cyclic Error are 0.004 and 0.006 respectively (Fig. 2). Those standard deviations are of order 10^{-3}, which is compatible with our definition of approximately equal. Therefore, we can treat $H^{-1}H$ and HH^{-1} as approximately the same as the identity mapping. Fig 2 shows that each term in Eq. 3 can be modeled using a Normal distribution with zero mean and small standard deviation. Interestingly, Student Projection Error distribution has a bigger standard deviation than Teacher Projection Error. This difference in standard deviation indicates a gap in the domain of features extracted using S and T, i.e., features extracted by S are projected versions of features extracted by T. Hence, it is easier for feature extracted by T to match feature extracted by S than the other way around. Therefore, we hypothesize that the feature manifold of S is diffeomorphic to a feature submanifold of T containing features relevant to radiographs like edges or shapes.

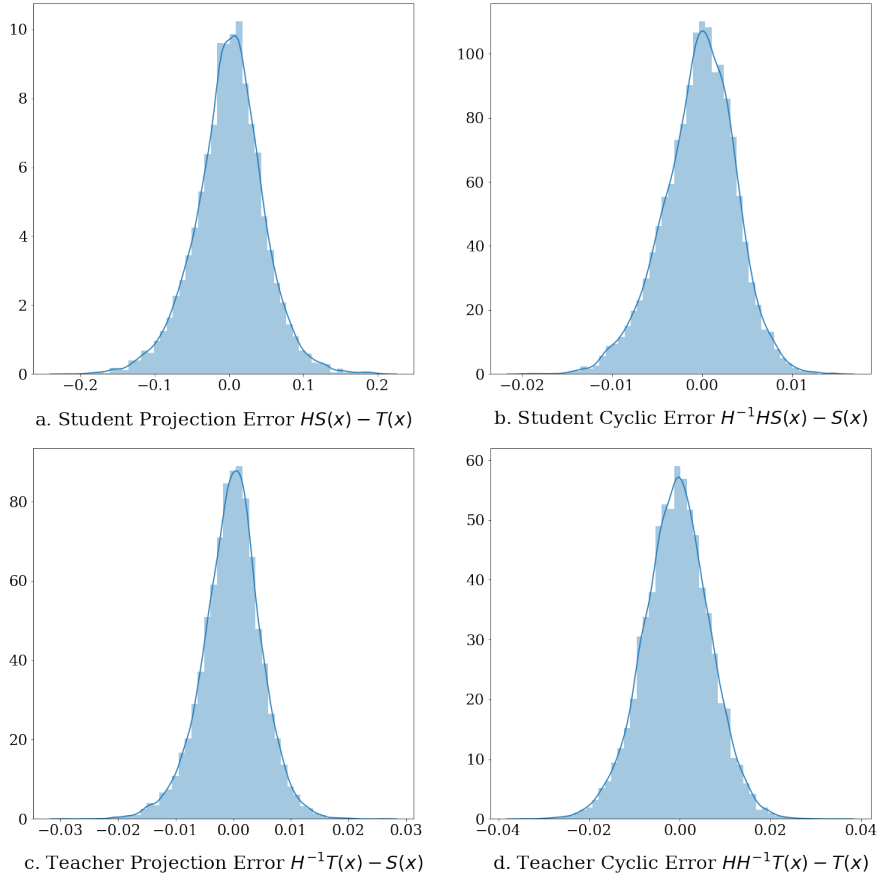

Figure 2: Probability distribution estimations for each term in $\mathcal{L}_{DM}$ on training data. The distributions can be modeled using Normal distribution with zero mean ($\mu = 0$) and small standard deviation. **a.** Student Projection Error distribution $\sigma = 0.047$. **b.** Student Cyclic Error distribution $\sigma = 0.004$. **c.** Teacher Projection Error distribution $\sigma = 0.005$. **d.** Teacher Cyclic Error distribution $\sigma = 0.006$.

References

[1] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. doi: 10.1109/TPAMI.2019.2913372. URL https://doi.org/10.1109/TPAMI.2019.2913372.

[2] Ilija Radosavovic, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10425–10433. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01044. URL https://doi.org/10.1109/CVPR42600.2020.01044.