RESEARCH ARTICLE

LAMBO: Landmarks Augmentation With Manifold-Barycentric Oversampling

IAROSLAV BESPALOV¹, NAZAR BUZUN¹, ², OLEG KACHAN¹,
AND DMITRY V. DYLOV¹ (Member, IEEE)

¹Skolkovo Institute of Science and Technology, 121205 Moscow, Russia
²Artificial Intelligence Research Institute, 121170 Moscow, Russia

Corresponding author: Dmitry V. Dylov (d.dylov@skoltech.ru)

The work of Dmitry V. Dylov was supported by RFBR under Grant 21-51-12012.

ABSTRACT We propose the first data augmentation method based on optimal transport theory, with the generated data being guaranteed to belong to the original data manifold. The proposed algorithm randomly samples a clique in the nearest-neighbors graph representing the data knowledge and computes the Wasserstein barycenter between the neighbours with random uniform weights. Being extremely natural-looking, many such barycenters are then produced iteratively to overpopulate the original dataset. We apply this approach to the problem of landmarks detection in unsupervised and semi-supervised scenarios in the popular tasks of face keypoints extraction, pose detection, and the segmentation of anatomical contours in medical imaging. The barycentric oversampling approach is shown to outperform state-of-the-art data augmentation methods. The code is available at https://github.com/cviaai/LAMBO/.

INDEX TERMS Augmentation, barycenter, cyclic generative adversarial networks, landmarks, manifold, optimal transport theory, oversampling, wasserstein distance.

I. INTRODUCTION

Excessive labeling is required to obtain stable neural network models of high image-to-image mapping quality. The ability to prepare a training dataset of sufficient size strongly depends on the research area, with domains like medicine, space imagery, or geology demanding immense expertise, time, and financial resources to systematize, clean, and label the data. Yet, despite the unquestionable demand, the unlabelled data usually prevail the labeled ones regardless of the domain, creating a welcoming setting for the development of unsupervised/semi-supervised approaches.

Most of the existing unsupervised methods do not achieve the quality of models optimized in a supervised training mode. However, some Cyclic Generative Adversarial Networks (Cyclic GANs), e.g., already a classic CycleGAN [1], DualGAN [2], or others [3], [4], [5], oftentimes yield results comparable to the supervised learning, without requiring the aligned (paired) data for training. This class of architectures is efficient, e.g., in a semi-supervised setup in the presence of a large number of unlabelled images.

Data augmentation is a way to fictitiously inflate the available annotation, conventionally entailing either basic geometric or deep learning-based approaches [6]. While the latter is deservedly the subject of active research today [7], the former is prone to expanding the manifold of the annotations beyond the original distribution. Namely, augmentations that include non-trivial transformations (beyond basic image rotation and flips, e.g., a linear stretching) will inevitably lead to expansion of the original manifold, disturbing the physical nature of the data. This becomes a major problem if one wants to understand why the augmented data improves the performance of a certain model, obscuring the overall interpretability. Moreover, disciplines like medicine are known for their resilience to the classical augmentation methods such as affine transformations [6], [8], demanding the development of more advanced methods.

Existing data augmentation approaches dismiss advances of modern optimal transport (OT) theory [9], [10], which – allegedly – could be attributed to the lack of user-friendly solutions to handle the complex topology of large datasets.
FIGURE 1. Landmarks augmentation with manifold-barycentric oversampling. The barycentric manifold is obtained by calculating weighted Wasserstein barycenters between the simplices of the nearest neighbours of the landmarks graph. First, we sample new landmarks from random uniform barycentric coordinates. The landmarks sampled this way are close by distribution to the empirical ones, preserving the physical nature of the data and acquiring a natural look. In the second step, the sampled landmarks are used to train a given mapping model. The augmented dataset increases the accuracy of the target task performed by the mapping model.

In OT, the manifold hypothesis [11], [12] states that data do not occupy the whole ambient space, but (due to redundancy and co-relation of features) the data concentrate near a submanifold of intrinsically lower dimension than that of the entire ambient space. Yet, even if the ambient space is flat and lacks curvature, the data submanifold can be curved and topologically nontrivial, having clusters and topological holes of higher dimensions.

Motivated by these challenging attributes of the submanifold, in this work, we propose to construct a local non-parametric approximation of it by means of Wasserstein distance to devise an efficient augmentation arrangement. Each local region in such a representation will be a convex set of neighboring points, corresponding to some wanted landmarks or segmentation contours. Hence, to sample new points inside the region, we propose to compute random weighted Wasserstein barycenters [12] which inherit the geometric properties of the basis of landmark measures and such local regions will automatically prevent us from leaving the submanifold. We introduce the notion of barycentric manifold that we define as a set of weighted barycenters generated in the local regions of the manifold. Fig. 2 summarizes the proposed concept for the case of a mapping model that performs some image-to-image translation task.

The contribution of this paper is in the following:

1) A new type of sample-efficient and non-parametric data augmentation approach based on weighted Wasserstein barycenters. This is the first adaptation of optimal transport theory to data augmentation problem, which guarantees to sample the new data that asymptotically converges to the original data submanifold.

2) Efficient (compatible to supervised outcome) mapping model for unpaired and semi-supervised detection of landmarks and segmentation contours.

3) Extensive line of experiments that compares our method to modern data augmentation approaches. Our augmentation outperforms both classical (geometric, Gaussian mixture model (GMM)) and deep-learning based models (generative adversary networks (GAN), variational autoders (VAE), and normalizing flows (NF)).

II. RELATED WORK

Rapid progress of deep learning instigated a series of supervised landmark detection approaches: cascade of convolutional neural networks (CNNs) [13], multi-task learning (pose, smile, the presence of glasses, sex) [14], and recurrent attentive-refinements via Long Short-Term Memory networks (LSTMs) [15]. Special loss functions (e.g., wing loss [16]) were shown to further improve the accuracy of CNN-based facial landmark localisation. Ultimately, Hourglass [17] has proven to be a universal model for detecting landmarks in the supervised learning mode, having a stack of U-Net [18] type architectures at its core.

Unsupervised pretraining has attracted major interest in the community with the advent of data-hungry deep networks [19]. A classical approach for such a task is to use the geometric transformations for feature descriptors encoding, comprising different variations of embeddings [19], [20], [21], [22] where pretrained descriptors had to reflect the image changes synchronously. Recent work [23] proposes an extension to this approach by matching descriptors from different images via deep clustering. However, none of these
works report extraction of the landmarks in a truly unsupervised way. Their unsupervised nature is only as good as a generic pretraining could be, being prone to encoding some redundant information within the bottleneck. Only the authors of [5] managed to distill just the landmarks data in their encoder, establishing the first fully unsupervised extraction method.

Another class of unsupervised pretraining methods [24], [25], [26] uses conditional image generation. If images $I_1$ and $I_2$ have different landmarks but have the same style (e.g., sequential frames from the video), the reconstruction network $G$ can generate $I_2$ from $I_1$ and the landmarks $L(I_2)$ of the second image $I_2$. The corresponding loss function then minimizes the difference between $I_2$ and $G(I_1, L(I_2))$ and the additional condition of sparsity on the landmarks heatmap corresponding to $L(I_2)$. Methods [27], [28] are similar, but they have an additional discriminator network to compare predicted landmarks to the landmarks from some unaligned dataset. Thus, the use of unaligned dataset is an intermediate step between the unsupervised and the supervised training. The method proposed herein also belongs to this class, significantly improving the precision of the GAN model and the quality of training with limited annotation.

An important part of the Cyclic GAN implementation proposed herein is the content-style decomposition of an image [29]. We use the architecture of stylegan2 [4] that allows to mix style and landmarks through weights modulation in convolutional layers. It makes landmarks extraction invariant to the style changes and vice-versa: the style is invariant to local geometric perturbations of the image [26]. Lastly, paper [30] explores style translation between images, augmenting the dataset for supervised training.

We conversely focus on datasets with large number of unlabelled images that require landmark augmentation. The idea of augmentation via barycentric sampling was motivated by papers [12] and [31]. The authors of [12] introduce barycentric coordinates (defined in Section III) for low-dimensional embedding. To augment available data, we propose to sample such coordinates w.r.t. arbitrary collections of data points being in the neighborhood and then to map them into the landmarks domain.

Data augmentation techniques can vary from simple geometric transformations [32] to more sophisticated ones, requiring interpolation of inputs in the feature space [33] or a targeted use of deep learning [34], [35], [36]. Where the oversampling is entailed, these methods oftentimes rely on Synthetic Minority Oversampling Technique (SMOTE) [33] – an established geometric approach to random oversampling and data augmentation in Euclidean space, followed by many extensions [37], [38]. For faces, the survey [39] covers a large set of augmentation methods, ranging from attribute transformations [40] to generative methods (e.g., GAN-based [1], [41]). Other modern augmentation methods include VAEs [42], [43] and NFs [44]. Wasserstein barycenters were used to enhance color representation in special stains [45]; however, their consideration was limited to classical color transfer [46] for a non-generic application.

Inexplicably, there is still no data augmentation method based on optimal transport theory, motivating this effort.

III. METHOD

We begin with an unpaired dataset of 2D images and landmarks (fixed number of keypoints in $\mathbb{R}^2$ that may be either ordered or randomly permuted). The images and the landmarks are assumed to be mapped using a class of generative image-to-image translation models, generically referred to as a Mapping model. We emphasize that any mapping model can be used; however, cyclic generative models, such as CycleGAN [1] or DualGAN [2], are studied below as the most representative and popular I2I architectures.

A. CYCLIC MODEL TRAINING

The Cyclic GAN architecture considered herein includes two generators, two discriminators, and a style encoder (refer to [4]). The first generator maps images to landmarks and uses hourglass model [17]. The second generator does the inverse mapping, taking landmarks and style as input and producing an image. The style may be either encoded from the image (to reconstruct the output image) or generated from a Gaussian noise (to approximate image distribution). We, thus, train bi-directional mapping $image \rightarrow (landmarks, style) \rightarrow image$ and $(noise, landmarks) \rightarrow image \rightarrow landmarks$. Corresponding loss functions ($l_{\text{rec}}, l_{\text{rec}}$) compare the reconstructed images/landmarks with the original ones (Section IV). The training also includes mono-directional mappings, where we compare the generated image/landmarks with the real ones by distribution. Algorithm 1 concisely covers universal optimization routine of Cyclic GAN training, regardless of the image-to-image translation task.

For given landmarks coordinates $X^l$, we engage intermediate mapping to Gaussian heatmaps $L = L(X^l)$ which are computed as follows. For the $k$-th point of the coordinates $X^l(k)$ and integer coordinates $(i, j)$ in the heatmap, denote the distance from $(i, j)$ to $X^l(k)$ by $d(k, i, j)$. Then, up to the normalisation term, if the landmarks are ordered,

$$L[k][i,j] \sim \exp\left\{-\frac{d^2(k, i, j)}{2\sigma^2}\right\}.$$  

In the un-ordered case (e.g., segmentation contours), one should use the landmarks summed by channels, such that $L[i,j] = \sum_k L[k][i,j]$.

B. MANIFOLD-BARYCENTRIC OVERSAMPLING

We consider point clouds as empirical probability distributions supported over finite sets of points in the Euclidean plane $\mathcal{P}(\mathbb{R}^2)$, endowed with the 2-Wasserstein distance $W_2$, a.k.a., the Wasserstein space [47] $W_2(\mathbb{R}^2) = (\mathcal{P}(\mathbb{R}^2), W_2)$. So, the distance between two landmarks can be measured using $W_2$ as the distance between two empirical measures.\textsuperscript{1}

\textsuperscript{1}Wasserstein distance is identical to $L_2$ for the ordered landmarks but is indispensable in the unordered case.
Algorithm 1: Cyclic GAN Training Iteration

Input: image $I$, unpaired landmarks coordinates $X^l$, hyperparameters of reconstruction losses $c_l, c_{GL}$, generators $G_L, G_I$, discriminators $D_L, D_I$, heatmap variance hyperparameter $\sigma$;
compute Gaussian heatmap: $L = L(X^l, \sigma)$;

# Image GAN model
sample noise: $z \sim \mathcal{N}(0, 1)^{\dim(z)}$;
generate random style: $S_o = S_o(z)$;
generate fake image: $I_f = G_I(G_L(I), S_o)$;
max $D_I$ images disc. loss: $L_d(D_I, I_f, I)$;
calculate fake style $S_f = S(I_f)$;
min $G_I$ generator loss: $L^g_r(L_f)$;
# Landmarks GAN model
generate fake landmarks: $L_f = G_L(I)$;
max $D_L$ landmarks disc. loss: $L_d(D_L, L_f, L)$;
min $G_L$ generator loss: $L^g_r(L_f)$;

# Discriminator penalty
if iteration mod 4 == 0 then
    $\min G_L, D_L \lambda E \|\nabla D_L(I)\|^2$;
    $\min D_I, G_I \lambda E \|\nabla D_I(I)\|^2$;
end

# Cycle mapping
initialize $g$: random geometric transformation;
reconstruct image: $I_R = G_I(G_L(I), S(g(I)))$;
min $G_L, G_I, c_L E_{\text{rec}}(I_R, I)$;
reconstruct landmarks: $L_R = G_L(G_I(L, S_o))$;
min $G_L, G_I, c_L E_{\text{rec}}(L_R, L)$;

Under the manifold hypothesis, the data comprise a submanifold in the ambient space $W_2(\mathbb{R}^2)$. Due to the data submanifold curvature, we can only trust the ambient metric locally. Thus, our idea is to non-linearly interpolate the densely populated local regions of the space of point clouds to generate probable landmark point clouds, without leaving the data manifold. Otherwise, exceeding the data manifold could generate improbable, unnaturally-looking landmarks.

We formalize the Wasserstein-barycentric manifold sampling for the point cloud data augmentation in Algorithm 2. Given a hyper-parameter $k$ we construct a symmetricated $k$-NN neighborhood graph $G_k$ over the landmark point clouds, endowed with the Wasserstein distance. Then, we construct the clique complex $K(G_k)$ by finding the maximal cliques (i.e., not contained in any other clique) of the neighborhood graph, associating a $k$-simplex with a maximal $(k-1)$-clique. Then, $N_{\text{aug}}$ simplices $\sigma_i$ are randomly selected into the sampling set $\{\sigma_i\}_{i=1}^{N_{\text{aug}}}$, with replacement according to the multinomial distribution (to provide uniform selection of the graph’s vertices). From each $k_i$-simplex $\sigma_i$, we sample a new point $\bar{\mu}_i$ according to the uniform distribution over the simplex. This allows to sample the barycentric coordinates $\lambda_i \sim \text{Dir}(\alpha)$, with $\alpha = (1, \ldots, 1) \in \mathbb{R}_{\geq 0}^{k+1}$ and to compute the weighted Wasserstein barycenter w.r.t. the vertices of $k_i$-simplex corresponding to those coordinates.

Notably, Algorithm 2 entails time-consuming pairwise computation of edges with OT weights for the $k$-nearest neighbors (kNN) graph. Its complexity is $O(N^2 s^2 / \varepsilon^2)$, where $s$ is the domain size (the number of key-points in a landmark) and $\varepsilon$ is the accuracy of the Sinkhorn method [48]. For large datasets ($N \gg 1000$), one may further improve Algorithm 2 by using faster kNN graph computation, taking $O(N \log s)$ steps instead of $N^2$ [49]. The algorithm’s complexity also includes maximal-cliques search, $O(\min\{k^4, (Nk)^{3/2}\})$ [50], and the computation of barycenters, $O(N_{\text{aug}} s^2 k / \varepsilon^2)$ [51].

1) WASSERSTEIN BARYCENTER
Barycentric coordinate system specifies the location of each point on a simplex with a reference to the points spanning it. For Euclidean space, the transition between barycentric and the ambient coordinates is, by definition, a convex combination equivalent to the weighted sum. However, for generic metric spaces, it is a solution to an optimization problem. In particular, the Wasserstein space is not Euclidean, with a weighted mean generalized as the minimizer of the sum of squared distances, generally known as the weighted Fréchet mean [52] or, in $W_2(X)$, the weighted Wasserstein barycenter [10]:

$$\bar{\mu}(\lambda) = \arg \min_{\mu} \sum_{i=0}^{k} \lambda_i W_2^2(\mu_i, \mu).$$ (2)
For measures with free support, the solution of the Wasserstein barycenter problem is a locally optimal measure with a discrete support [53].

IV. LOSS FUNCTIONS

A. RECONSTRUCTION

Let us denote the generator (decoder) model by \( G \). It maps a pair \((\text{landmarks}, \text{style})\) into an image, where the style may be either encoded from the image \( S(I) \) or generated from the noise \( S_0(z) \). We apply random geometric transformation \( g \) before extracting style from image for better separation of style and content. Let \( G_L \) be the landmarks encoder. Then, the following loss term compares the original image to the reconstructed one, for a given heatmap of landmarks and the encoded style [54]:

\[
L_{\text{rec}}^I(I_R, I) = c_1 \| \text{Alex}(I_R) - \text{Alex}(I) \|^2 + c_2 (1 - \text{Id}(I_R)^T \text{Id}(I)) + c_3 \| I_R - I \|, \tag{3}
\]

where \( I_R = G_L(G_L(I), S(g(I))) \). This loss has three terms: Euclidean image similarity, features from Alex network (\( \text{Alex}(I) \)) [55], and features from ResNet ArcFace (\( \text{Id}(I) \)) [56].

Similarly, for the landmarks reconstruction, obtained from a generated image (fake) \( L_R = G_L(I_F) \):

\[
L_{\text{rec}}^L(L_R, L) = c_4 H[I_R | L] + c_5 W_1(X^I(L_R), X^L(L)), \tag{4}
\]

where \( H \) is cross-entropy function, \( W_1 \) is Wasserstein distance with Euclidean ground metric,

\[
I_F = G_L(L, S_0(z)), z \sim \mathcal{N}(0, 1)^{\dim(z)}. \tag{5}
\]

Component \( H \) enables more accurate comparison of ‘close’ pairs of landmarks and \( W_1 \) speeds up the convergence of distant elements.

B. ADVERSARIAL

We further increase the overlap of the real and the generated image distributions by employing the GAN losses for discriminator and generator:

\[
L_d(D_I, I_F, I) = -E \log \left( 1 + e^{-D_I(I)} \right) - E \log \left( 1 + e^{D_I(I_F)} \right), \tag{6}
\]

\[
L_g^I(I_F, S_F, S_0) = E \log \left( 1 + e^{-D_I(I_F)} \right) + c_6 \| S_F - S_0 \|_1. \tag{7}
\]

They are classical losses from binary \text{LogitBoost} classifier [57] that have been shown to provide a stable training of GANs with non-vanishing gradients [4]. The discriminator optimizes separation of real and fake images and the generator does the classification with opposite label signs. Similarly, for the landmarks distribution, we use the same discriminator loss and \( L_g^L(I_F) = E \log \left( 1 + e^{-D_I(I_F)} \right) \).

V. ARCHITECTURE DETAILS

We enhanced the \text{stylegan2} architecture [4] by introducing several important modifications.

A. IMAGE GENERATOR

We consecutively downsample the heatmap of the landmarks by convolutional layers from the size \( 256 \times 256 \) to the sizes \( 4 \times 4, 8 \times 8, \ldots, 256 \times 256 \), and then, we concatenate them with the outputs from the progression of the modulated convolution blocks (\text{ModulatedConvBlock}) in network \( G_I \). The term “modulated convolution” means that its weights are obtained from the style. The noise \( z \) passes through a series of linear layers to place the style \( S_0(z) \) on the style manifold. These styles are further used in \text{ModulatedConvBlock} to obtain the corresponding modulated weights. So, each \text{ModulatedConvBlock} receives a pre-processed heatmap, multiplies it by the modulated weights, computes some non-linear transform, and passes
I. Bespalov et al.: LAMBO: Landmarks Augmentation With Manifold-Barycentric Oversampling

FIGURE 4. Wasserstein distance between augmented and separated validation data as a function of parameter $k$ in $k$NN-graph for different dataset sizes $N \in \{100, 300, 600\}$. Variance is computed on 20 runs with random re-sampling. Bottom plot: The dependence of the KL distance between the distributions with and without the barycenter augmentation on the dataset size (in # of images).

it to the next block. In other words, it is a core block for combining the style and the content (landmarks) information. Finally, the very last block returns the fake image.

B. LANDMARK GENERATOR (ENCODER)

Our landmarks encoder consists of two principal parts. Due to recent success of the stacked hourglass model [17], we have integrated it in our landmarks encoder, which produces an output with separate channels corresponding to the key-points. Afterwards, the application of the downsampling convolutions yields the coordinates of the landmarks ($X_l$).

C. DISCRIMINATORS

In the discriminators, we convolve the landmarks heatmaps and the images separately and then pass them through a sequence of standard ResNet layers.

D. STYLE ENCODER

Style encoder is a PSP-type network [54] that maps the image to the style matrices of size $14 \times 512$. The rows of such matrices store ModulatedConvBlock layers with different corresponding resolutions (from $4 \times 4$ to $256 \times 256$).

The connection of all components is shown in Fig. 2.

TABLE 1. Results on 300-W/Human3.6M datasets using interocular distance (IOD)/%MSE as a metric. Unpaired denotes training Cyclic GAN with unaligned dataset. In our method, the landmarks from 300-W are randomly permuted and augmented with the barycenters. Note: Jakab et al. [28] train on additional images from MultiPIE dataset. Semi-supervised method refers to complete extraction of landmarks within the bottleneck of encoder, without pretraining. Unsupervised pretraining methods use CelebA dataset to pre-train the encoder. In all tests, the empirical standard deviations are not greater than 0.12 for 300-W and 0.07 for Human, computed on 20 runs.

| Training          | Method            | 300-W | Human |
|-------------------|-------------------|-------|-------|
| Supervised        | TCDCN [14]        | 5.54  | -     |
|                   | RAR [15]          | 4.94  | -     |
|                   | WingLoss [16]     | 4.04  | -     |
|                   | HG2 [17]          | 4.2   | 2.16  |
| Unsupervised      | UDIT [27]         | 5.37  | -     |
| pretraining       | Sparse [21]       | 7.97  | 7.51  |
| + supervised      | Fab-Net [55]      | 5.71  | -     |
| regression        | Dense 3D [20]     | 8.23  | -     |
|                   | DVE HG [19]       | 4.65  | -     |
|                   | Zhang [22]        | -     | 4.14  |
|                   | Lorenz [26]       | -     | 2.79  |
| Semi-supervised   | BRULE [5]         | 3.9   | -     |
| Unpaired          | Jakab et al. [28] | 8.67  | 2.73  |
|                   | LAMBO (Ours)      | 4.96  | 2.97  |

E. IMPLEMENTATION DETAILS

Additional implementation details are provided in the Appendix. The code is available at https://github.com/cviaai/LAMBO/.

VI. EXPERIMENTS

We consider 300-W [58], Human3.6M [59], and Cardio MRI (private) datasets. Dataset 300-W contains 68 annotated key-points per face, treated as unordered, and 3k training pairs. Human3.6M dataset has about 13k training pairs of 32 key-points (ordered), depicting the human pose in the images. Cardio MRI dataset has 701 training pairs of MRI scans and segmentation contours with 200 unordered key-points each.

Typical for Cyclic GANs, the unaligned training needs a labelled dataset but does not need a direct image-mask correspondence. To evaluate the efficiency of augmentation via the barycentric manifold, we sampled 7k barycenters for each dataset. The barycenters themselves are various kinds of key-points distortions (Figs. 3), yet with the physicality of these data (proportions, angles, etc.) preserved.

In agreement with the theory, Fig. 4 also demonstrates asymptotic convergence to empirical data when the dataset size increases. Given such realistic augmentations, the Cyclic GAN performance improves in all three datasets considered (Fig. 5), with the new landmarks always remaining within the manifold of the original data. This could be confirmed both visually and by analyzing the post-augmentation distributions (Fig. 6).

In Table 1, we compare our augmented model with the state-of-the-art (SOTA) approaches, trained both in the supervised and in the unsupervised (pretraining with the supervised regression) scenarios, using typical for these datasets metrics.
of %-MSE, normalised by the image size, or the inter-ocular distance (IOD [60]).

Remarkably, our approach performs almost at the level of the supervised models trained on fully-annotated datasets, beating the result of the unpaired SOTA. When the portion of available annotation drops (< 600 images), our method yields a greater gain and outperforms even those supervised models that have the exact same encoder within. We speculate that our unpaired method performs better because 1) supervised models depend on the amount and the quality of the annotation, and 2) even the raw images without the landmarks can contribute to the performance of our model. The proposed augmentation technique improves precision significantly on small subsets (e.g., 100 and 300) and this gain reduces as the dataset size grows (see plots in Fig. 4).

A. HYPERPARAMETERS

In the augmentation Algorithm (2), parameter $k = 15$ and $N_{\text{aug}} = 7000$, obtained by the grid search. As shown in Fig. 4 (top right), the method is robust to the choice of parameter $k$ and its optimal value does not significantly depend on the dataset size. We have confirmed this by splitting the landmarks data of 300-W in half. The first half was augmented via the manifold-barycenteric oversampling and then compared to proper sampling of the other half. The optimal parameters are drawn from the minimum of the OT distance ($W_2$) between the dataset parts.

The parameters in the adversarial model were: image loss coefficient $c_I = 150$, measure loss coefficient $c_L = 1.5$, style reconstruction coefficient $c_6 = 10$ (used in loss $L^A_{\text{rec}}$), coefficients of $L^A_{\text{rec}}$ are $c_1 = 0.8$, $c_2 = 0.1$, $c_3 = 1$, coefficients of $L^A_{\text{rec}}$ are $c_4 = 10^4$, $c_5 = 100$. The variance of the heatmaps $\sigma = 4$. The learning rates are: $2 \cdot 10^{-4}$ (image generator), $4 \cdot 10^{-4}$ (image and landmarks discriminators), $2 \cdot 10^{-5}$ (landmarks predictor), and $10^{-5}$ (style encoder). The batch size is 8 and accumulation of weights is used to make training more stable (moving average updates weights every 100 iterations). The complete training of the model requires 100-150k iterations, taking ~ 80 hours on one Nvidia Tesla V100 GPU. The augmentation typically takes several minutes, depending on the dataset.

VII. LAMBO VS. OTHER AUGMENTATION METHODS

For our comparison, we considered the popular methods from several major categories of the augmentation approaches, systematized in the review article [6]. The outcomes of these augmentations are summarized in Fig. 6.

To approximate the landmarks distribution, we use a simple fully connected network with three Linear + ReLU layers and internal dimension of 256. In all cases, we train the generation of landmarks coordinates $X^I$ based on the empirical distribution and compare the performance of all augmentation methods to the baseline (unpaired Cyclic GAN).

GANs [4] are natural synthesizing tools, entailing non-saturating logistic losses (ref. Section IV: $L_d$ and $L_g$). GAN-based augmentation does improve the baseline score (Table 2); however, as a parametric model, there is always uncertainty in selecting the model’s architecture and in tuning. Moreover, training a GAN model for a particular task could be unstable and time-consuming: small models are inaccurate, and the large ones overfit the data.

Variational autoencoder (VAE [42]) uses two FC networks as encoder and decoder. The latent vectors are sampled from the Normal distribution of size 100. In the augmentation task, VAE shows the worst performance and does not overcome the baseline. We considered several new types of VAEs (e.g., [43]); but they could not improve the target score and were often prone to the mode collapse (see Fig. 6).

Augmentation with Normalizing Flows (NF [44]) approximates parametric distribution by an invertible FC network. We observed that it requires relatively big landmarks sets to fit the data precisely, and it was not efficient in our weakly-supervised approbation (e.g., subsets of 100 landmarks).

A popular geometric transform (geometric [61]) augments data by random scaling up to 10% and by rotation up to $\pi/12$ with probability 0.3. Evidently, this procedure does not cover the manifold structure of any of the datasets and generates only some local perturbations to the landmarks.

Gaussian mixture model (GMM [62]) is a simple but efficient classical method of distribution approximation. We have set the number of components to 20 (count of Normal distributions). It improves the baseline but the performance decreases when the landmarks reach high dimensions.

FIGURE 5. Top row: the real images and the predicted landmarks. Bottom row: sampled barycenters (landmarks) and the generated images (fakes). Having a very realistic look, these fakes are proposed for the data augmentation in the corresponding tasks.
TABLE 2. Comparing various augmentation methods. Supervised: hourglass model (HG2 [17]) trained with loss $\mathcal{L}_{\text{rec}}$ and paired data; baseline: Cyclic GAN model, unpaired training on the full image dataset and a partial subset of landmarks; baseline + [geometric (Albumentations), GAN, VAE, GMM, NF, LAMBO]: the same unpaired Cyclic GAN model, trained with specific augmentation method. The number beneath the dataset name is the number of annotated images (ground truth landmarks) used in the unpaired training. In all tests on 300-W, the empirical standard deviation is 0.12, computed on 20 runs. IOD and %−MSE are the metrics on 300-W and the other two datasets, respectively.

| Method + Augmentation | 300-W 100 | 300-W 300 | 300-W 600 | 300-W full (3k) | Cardio 100 | Cardio 300 | Cardio 701 | Human 13,412 |
|------------------------|-----------|-----------|-----------|----------------|------------|-----------|------------|-------------|
| supervised (paired)    | 11.79     | 8.76      | 7.28      | 4.51           | 5.68       | 4.28      | 2.14       | 2.34        |
| baseline (unpaired)    | 8.68      | 7.88      | 6.95      | 5.10           | 4.39       | 4.02      | 2.70       | 2.61        |
| baseline + Albumentations [61] | 7.84 | 7.00      | 6.30      | 5.08           | 4.02       | 3.84      | 2.56       | 2.60        |
| baseline + GAN [4]     | 7.77      | 7.04      | 6.42      | 5.19           | 4.36       | 3.91      | 2.68       | 2.54        |
| baseline + VAE [42]    | 9.86      | 8.59      | 8.05      | 7.16           | 9.85       | 6.71      | 5.83       | 7.15        |
| baseline + GMM [62]    | 7.15      | 6.69      | 6.17      | 5.27           | 4.00       | 3.11      | 2.67       | 2.55        |
| baseline + NF [44]     | 11.6      | 7.87      | 7.52      | 5.75           | 4.88       | 3.82      | 3.03       | 2.62        |
| baseline + LAMBO (ours) | 7.34     | 6.48      | 5.83      | 4.96           | 3.42       | 2.86      | 2.34       | 2.47        |

The method also loses efficiency when the landmarks are unordered (like the cardiac contours).

We observe consistent behavior of all augmentation methods, with the empirical standard deviation of 0.12, computed on 20 runs.

VIII. DISCUSSION

Table 2 gauges the mapping model’s performance on all datasets for different augmentation types, but it also studies ablation of the effect of the augmentation by providing extra annotations to the model. We made subsets of original data to contain all the raw images but only a portion of available annotations (e.g., 100, 300, 600, and 3,148 landmarks for 300-W dataset). When we add the barycenter augmentation, we readily generate new 7,000 landmarks that, according to the visual inspection of Fig.3, clearly belong to the same manifold as the original data. Classical augmentation methods (such as rotations, linear stretches, reflections) are often “damaging” to the dataset even if they help improve the performance of a target task. We tuned its parameters on the validation set so that the deformation is small but sufficient to reduce the overfit.

To the contrary, the barycentric manifold oversampling preserves the consistency of the generated augmentation data, and as we observed, also outperforms these classical and modern deep learning based augmentation methods in the target tasks. Additional experiments with a combination of classical augmentation and LAMBO yielded similar outcomes, from which one may conclude that LAMBO inherently includes the small geometric augmentations. This is a valid hypothesis given that OT inherently encodes the ambient geometry of the manifold.

Another notable advantage of LAMBO is its non-parametric nature. In practice, it is more preferable over the parametric models considered in Table 2, because it is asymptotically accurate and is sample-efficient in estimating the data density. In all of the experiments reported above, we have needed just a single parameter to initiate the graph representation in the barycenter domain and to, consequently, enlarge the annotation set (namely, the desired number of adjacent neighbors for computing the local barycenters).

We compare the unpaired training to the supervised case, but do not show the results of the semi-supervised scenario explicitly. The amount of augmented data is much larger than that of the paired data. Given that most of the generated landmarks are not paired, it is somewhat irrational to process the paired items separately. Actually, while trying to find the correspondences in the unpaired data, Cyclic GAN model learns the task compatibly to the analogous supervised variants (ref. Table 2), hence, also covering the semi-supervised setting [5] by the unpaired setup.

Also instrumental is LAMBO’s computation time (compared to that of the Cyclic GAN training). Computing the matrix of Wasserstein pairwise distances between 3148 landmarks in the 300-W dataset took about 10 minutes, and given that our method typically needs 100 – 1000 landmarks, the augmentation can occur relatively fast as is. Nevertheless, it could be further accelerated for the large datasets by fast kNN search [49].

Overall, the classical augmentation methods (geometric and GMM) are somewhat faster and the deep learning based models (like GAN, NF, VAE) usually take more time than LAMBO. Recent advances in OT, such as approximate nearest neighbor search under the Wasserstein distance in linear time [63] and the computation of Wasserstein barycenters in polynomial time, either exact or approximate [64], will be the subjects of future work.

IX. CONCLUSION

The problem of learning from unlabeled data is of utmost importance in data science, because the unlabeled data largely prevail over the annotated ones. Even if the annotations are available, they are oftentimes unpaired with the image data and require cleaning, coming from different sources and being rarely sufficient. Our data augmentation solution to this problem relies on a new knowledge representation approach, where we considered the available annotations in the space of barycentric manifold, computed with the help of Wasserstein distance and represented as a graph. Just like the classical augmentation methods, this new augmentation method inflates the size of the set of available labels; however, it does it entirely within the original data annotation manifold.
FIGURE 6. Visualization of different augmentation methods compared in this manuscript, applied to specific data and their partially annotated subsets. Models: variational autoencoder (VAE [42]), Gaussian mixture (GMM [62]), Normalization Flow (NF [44]), our method (LAMBO), and the original data landmarks w.o. augmentation (REAL). Notice how fake LAMBO landmarks resemble the REAL ones even in the scenarios with limited annotation (# of annotated images is shown in parenthesis).

APPENDIX

The appendix covers formal definitions of the Wasserstein space (IX-A) and the graph-based representation in this space (IX-B) that allow for the barycentric oversampling, while controlling the distance within the manifold (Theorem 1 in IX-C).

A. WASSERSTEIN SPACE

Given a pair (X, d) of a metric space X and a ground (distance) metric d, the p-th Wasserstein distance is defined as

\[ W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} d^p(x, y) d\pi(x, y), \]  

(8)

where \( \pi \in \Pi(\mu, \nu) \) is the set of all joint probability measures on \( X \times X \) whose marginals are \( \mu \) and \( \nu \), i.e., for all measurable sets \( A, B \subset X \), \( \pi(A \times X) = \mu(A) \), \( \pi(X \times B) = \nu(B) \).

[65, Def. 1.1]. If restricted to the subspace of the probability measures with a finite p-th moment,

\[ \mathcal{P}_p(X) = \left\{ \mu \in \mathcal{P}(X) \left| \int_X d^p(x', x) \mu(dx) < +\infty \right. \right\}, \]  

(9)

where \( x' \in X \) is arbitrary, the p-Wasserstein distance is indeed a valid metric [65, Def. 6.1]. Then, the pair \((\mathcal{P}_p(X), W_p)\) formally constitutes the Wasserstein space \( W_p(X) \), suitable for performing the data augmentation that we are after.

Any given point cloud could then be seen as the probability distribution living in a space of probability measures over \( \mathcal{P}(\mathbb{R}^d) \). The landmark data point is a point cloud, i.e., not a single, but a set of multiple points on an Euclidean plane. A point cloud in \( \mathbb{R}^2 \) could be modeled as a discrete probability measure \( \mu \) with the weights \( a \in \mathbb{R}_{\geq 0}^n \) and the locations \( x \in \mathbb{R}^{n \times 2} \), \( \mu = \sum_{i=1}^{n} a_i \delta_{x_i} \), where \( \delta_{x_i} \) is the Dirac measure at the position \( x_i \in \mathbb{R}^2 \) and \( \sum_{i=1}^{n} a_i = 1 \).

The distinctive feature of the Wasserstein distance is that it takes the geometry of the probability measures into account by using the ground distance of the base space. Because
The existence of barycenters is proven for the Wasserstein spaces [9], and given that the computational tools for computing them exist [10], we propose to characterize the space of landmarks with the Wasserstein distance. Such a knowledge representation allows us to compute weighted means of the existing annotations, ultimately leading to an efficient arrangement for the augmentation.

**B. BARYCENTRIC OVERSAMPLING**

SMOTE [33] is an established geometric approach to random oversampling and data augmentation, followed by many extensions [37], [38], with the original motivation to balance classes in the imbalanced learning problem [66]. Its main idea is to introduce synthetic data points with each new point being the convex combination of an existing data point and one of its $k$-nearest neighbors.

We generalize the SMOTE algorithm to introduce synthetic points as convex combinations of arbitrary number of points, instead of just pairs, thus sampling from the simplices of the clique complex of a neighborhood graph. That is, a position of a new point is defined by the barycentric coordinates with respect to a simplex spanned by an arbitrary number of data points being sufficiently close. Each simplex is seen as a dense local region, thus interpolating within a simplex is assumed to result in the probable data points close to the data submanifold.

Let $\Lambda_k$ be the set of all vectors of $k + 1$ elements, such that $\lambda_i \geq 0$ and $\sum_{i=0}^{k} \lambda_i = 1$. Given a set of point clouds $\{\mu_i\}_{i=0}^{k} \in \mathcal{W}_2(\mathbb{R}^2)$ a Wasserstein $k$-simplex $\sigma(\{\mu_i\}_{i=0}^{k})$ is the set of all convex combinations of its vertices $\{\mu_i\}_{i=0}^{k}$:

$$\sigma(\{\mu_i\}_{i=0}^{k}) = \left\{ \bar{\mu}(\lambda) \mid \lambda \in \Lambda_k \right\},$$

where a given vector $\lambda$ of barycentric coordinates uniquely determines any point cloud $\bar{\mu}(\lambda) \in \sigma \subset \mathcal{W}_2(\mathbb{R}^2)$ w.r.t. the simplex $\sigma$ spanned by $\{\mu_i\}_{i=0}^{k}$.

Simplices could be “glued” together to comprise a simplicial complex, formally a collection of simplices. In particular, given a set of points $X$ in a metric space $X$ and a threshold parameter $\varepsilon \geq 0$, the Vietoris-Rips (VR) complex [67] on $X$ is defined as

$$\mathcal{R}_\varepsilon(X) = \{\sigma \subseteq X \mid d(x, x') \leq \varepsilon, \forall (x, x') \in \sigma \}.$$  

The expression above is a collection of $k$-simplices $\sigma$, such that $d(x, x') \leq \varepsilon$ for all pairs of points $(x, x') \in \sigma$ for some distance function $d(\cdot, \cdot)$ on $X$. The VR complex is equivalent to the clique complex of the $\varepsilon$-neighborhood graph, that is a $k$-simplex $\sigma$ corresponds a clique in the $\varepsilon$-neighborhood graph. In practice, we consider the clique complex of the $k$-NN neighborhood graph.

In our work, we chose the density-adaptive continuous $k$NN (ckNN) neighborhood graph [68] $G_{\delta,k} = (X, E)$, where

---

*2i.e., the weighted means of an arbitrary number of points.*
an edge \((x, x')\) is present in the graph, if and only if,
\[
(x, x') \in E \iff d(x, x') < \delta \sqrt{d(x, x_k) d(x, x'_k)},
\]
where \(x_k\) and \(x_k'\) are the \(k\)-nearest neighbors of points \(x\) and \(x'\) respectively. Intuitively, the parameter \(k \in \mathbb{N} \setminus \{0\}\) controls the radius of a sphere around each specific point, and \(\delta \in \mathbb{R}_{>0}\) affects their radii, all at once. By that, a sphere around a point adapts to the local density of the data and the \(k\)NN graph has an edge if a pair of spheres intersects. Examples of the manyfold-barycentric data points on such graph are illustrated in Figure 7.

C. DISTANCE BETWEEN SIMPLEX AND ITS VERTICES

We sample new landmarks from the convex hulls with adjacent basis elements. So, a natural question is: what is the distance between the new data and the basis elements? How does the performance depend on the convex hull structure? The following theorem addresses this question and provides the upper bound approximation.

**Theorem 1:** Consider a convex polyhedron with \(k\) vertices \(\{x_1, \ldots, x_k\}\) in some metric space with a linear distance function \(\rho(x, y)\). Let \(x^*\) be an internal point of the polyhedron. For our purposes, this point could be taken as a barycenter.

Then, there is a set of spheres
\[
\{B_1(x_1, r_1), \ldots, B_k(x_k, r_k)\}, \quad \forall i : r_i = \rho(x_i, x^*),
\]
that covers the entire polyhedron.

And, moreover, let \(\mu_1\) be a uniform measure in the polyhedron and \(\mu_2\) be a discrete uniform measure on \(\{x_1, \ldots, x_k\}\), then
\[
W_2^2(\mu_1, \mu_2) \leq \frac{1}{2k} \sum_{i=1}^{k} r_i^2.
\]

**Proof:** The first statement can be proved by induction by varying the number of vertices \(k\) in the polyhedron. The base of induction for \(k = 1, 2\) is evident. Assuming that the statement is true for \(k\), we will prove it for \(k + 1\). Let’s delete one vertex from the polyhedron. Consider two cases. In the first one, the internal point is inside of the polyhedron of \(k\) points. Then, by the induction assumption, it is covered by the circles. In the second case, the internal point is outside of the polyhedron of \(k\) points. Then, one projects the internal point into the nearest point of the \(k\)-polyhedron. With this projection, all distances to the projection point are reduced. For any two points and a plane that separates them, we may project one of the two points and reduce the distance between them. This statement holds for all points of the \(k\)-polyhedron and all the projections of the internal point.\(^3\)

We, thus, obtained that all \(k\)-polyhedrons are covered.

Now, consider an arbitrary point of the \(k + 1\)-polyhedron and draw a line from the internal point to this point. This line intersects some \(k\)-polyhedron and belongs to some circle. We found two points (the internal one and the point of intersection) that belong to one circle and our arbitrary point, located on the interval between them. Hence, the first statement is proven.

We have covered the polyhedron by the circles. Consider an intersection of one circle and the polyhedron. We map all points of this sector to the center of the circle (the polyhedron’s vertex). The transport distance of such mapping is \(r^2/2\) because the angle is independent of the radius. A uniform measure in this sector covers the uniform measure of the corresponding part of the polyhedron. Consider two uniform distributions, when one distribution is “inside” the other. The Wasserstein distance from the outer measure to any common point of the two measures is greater than the distance from the inner measure. Finally, the factor \(1/k\) comes from the uniformity of the measures on the vertices, yielding the final formula.

**REFERENCES**

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” 2017, arXiv:1703.10593.

[2] Z. Yi, H. Zhang, P. Tan, and M. Gong, “DualGAN: Unsupervised dual learning for image-to-image translation,” 2017, arXiv:1704.02510.

[3] K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka, “WaveCycleGAN: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks,” 2018, arXiv:1809.10288.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” 2019, arXiv:1912.04958.

[5] I. Bespalov, N. Buzun, and D. Dylv, “BRULE: Barycenter-regularized unsupervised landmark extraction,” 2021, arXiv:2006.11643.

[6] C. Shorten and T. M. Khooshgoftaar, “A survey on image data augmentation for deep learning,” J. Big Data, vol. 6, no. 1, pp. 1–5, Dec. 2019.

[7] T. DeVries and G. W. Taylor, “Dataset augmentation in feature space,” 2017, arXiv:1702.05538.

[8] N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, and D. V. Dylv, “Anomaly detection in medical imaging with deep perceptual autoencoders,” IEEE Access, vol. 9, pp. 118571–118583, 2021.

[9] M. Agueh and G. Carlier, “Barycenters in the Wasserstein space,” SIAM J. Math. Anal., vol. 43, no. 2, pp. 904–924, 2011.

[10] M. Cuturi and A. Doucet, “Fast computation of Wasserstein barycenters,” in Proc. ICML, 2014, pp. 685–693.

[11] Y.-H. Kim and B. Pass, “Wasserstein barycenters over Riemannian manifolds,” Adv. Math., vol. 307, pp. 640–683, Feb. 2017.

[12] N. Bonneel, G. Peyré, and M. Cuturi, “Wasserstein barycentric coordinates: Histogram regression using optimal transport,” ACM Trans. Graph., vol. 35, no. 4, p. 71, Apr. 2016.

[13] Y. Sun, X. Wang, and X. Tang, “Deep convolutional network cascade for facial point detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, p. 3476.

[14] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Learning and transferring multi-task deep representation for face alignment,” 2014, arXiv:1408.3967.

[15] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, “Robust facial landmark detection via recurrent attentive-refinement networks,” in Proc. ECCV, vol. 9905, Oct. 2016, pp. 57–72.

[16] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” 2017, arXiv:1711.06753.

[17] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” 2016, arXiv:1605.06214.

[18] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent., May 2015, pp. 145–155.

[19] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, “Unsupervised learning of landmarks by exchanging descriptor vectors,” in Proc. ICCV, 2019, pp. 1–10.

[20] J. Thewlis, H. Bilen, and A. Vedaldi, “Unsupervised object learning from dense equivariant image labelling,” 2017, arXiv:1706.02932.
NAZAR BUZUN was born in Baturyn, USSR, in 1991. He received the B.Sc. and M.Sc. degrees in applied mathematics and physics from the Moscow Institute of Physics and Technology (MIPT), in 2014, and the Ph.D. degree in mathematics from the Humboldt University of Berlin, in 2020.

He was a Research Scientist at the Skoltech Computational Imaging Laboratory. Since 2020, he has been with the Artificial Intelligence Research Institute, investigating the long time dependencies in Reinforcement Learning. His research interests include math statistics, optimal transport theory, generative adversarial networks, and long-term memory modeling.

OLEG KACHAN received the M.Sc. degree in applied informatics and mathematics from the HSE University, Moscow, Russia, in 2017. He is currently pursuing the Ph.D. degree with the Skolkovo Institute of Science and Technology, under the supervision of Prof. Alexander Bernstein.

He is a Visiting Graduate Researcher with the Institute for Pure and Applied Mathematics, University of California, Los Angeles (UCLA), studying the topology-preserving dimensionality reduction. He is also with the HSE University. His research interests include computer vision, dimensionality reduction, topological data analysis, and learning from non-Euclidean data, such as graphs, simplicial complexes, and manifolds, and devoted to modeling and learning from the higher-order interactions in neuroimaging data, such as EEG and fMRI.

DMITRY V. DYLOV (Member, IEEE) received the M.Sc. degree in applied physics and mathematics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 2006, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2010.

He is an Associate Professor and the Head of Computational Imaging Laboratory, Skoltech. His group specializes on computational imaging, computer/medical vision, and fundamental aspects of image formation.

* * *