High-Dimensional Bayesian Optimization with Invariance

Ekansh Verma * VERMAEKANSH55@GMAIL.COM
Indian Institute of Technology Madras

Souradip Chakraborty* SCHAKRA3@UMD.EDU
University of Maryland, College Park

Ryan-Rhys Griffiths RRG27@CAM.AC.UK
University of Cambridge

Abstract

High-dimensional black-box optimization problems are of critical importance in many areas of science and engineering. Bayesian optimisation (BO) has emerged as one of the most successful techniques for optimising expensive black-box objectives. However, efficient scaling of BO to high dimensional settings has proven to be extremely challenging. Strategies based on projecting high dimensional input data to a lower-dimensional manifold, such as variational autoencoders (VAEs) have recently grown in popularity, but there remains much scope to improve performance by structuring the VAE latent space in a manner that makes it amenable to BO. In this work, we improve the quality of the VAE latent space through the use of invariant augmentations learned using the marginal likelihood objective of a Gaussian process (GP). As an ablative feature, we show that our method improves the performance of a popular VAE-BO architecture.

Keywords: Bayesian Optimisation, Gaussian Processes, Invariance

1. Introduction

Black-box optimization problems are characterized by the absence of the analytic form and gradients of the function to be optimized and arise in many areas of science and engineering including drug discovery (Gómez-Bombarelli et al., 2018), chemical reaction optimisation (Shields et al., 2021) and artificial intelligence (Chen et al., 2018). In recent years, Bayesian optimization (BO) has emerged as one of the most powerful solution methods for black-box optimization (Turner et al., 2021). Efficient scaling of BO to high dimensional settings however, has proven to be challenging due to the poor scalability of Gaussian processes (GPs) (Rasmussen and Williams, 2006) which are a core component in the majority of effective BO algorithms. A recent approach to high-dimensional BO leverages the power of variational autoencoders (VAEs) (Kingma and Welling, 2014; Gómez-Bombarelli et al., 2018; Griffiths and Hernández-Lobato, 2020; Lu et al., 2018; Eissman et al., 2018; Moriconi et al., 2020; Siivola et al., 2020; Antonova et al., 2020) to project high dimensional input data to a lower dimensional manifold on which BO is subsequently performed, a methodology termed VAE-BO (Grosnit et al., 2021). Two key performance issues in VAE-BO however are:

1. The data acquired during BO execution is not used to re-train the VAE and update the structure of the latent space.

*. equal contribution

2. The unsupervised training of the VAE means the latent space cannot be structured according to the function values of the encoded latent points. As such, it becomes challenging to fit a surrogate model such as a GP and BO performance suffers.

In Tripp et al. (2020) the first issue is addressed through a weighted retraining mechanism which assigns more influence in subsequent retraining of the VAE to regions of the latent space with favourable black-box function values. Building on Tripp et al. (2020), Grosnit et al. (2021) additionally address the second issue through the application of metric learning (Xing et al., 2002), an approach which requires far fewer labelled data i.e. is more sample efficient, compared to previous approaches. The performance of metric learning-based approaches are known to be heavily dependent on the incorporation of known invariances and informed data augmentations (Ko and Gu, 2020; Chen et al., 2020) and it is from this fact that we take our inspiration to propose a general method to improve sample efficiency across all current VAE-BO architectures.

Our solution method aims to learn black-box specific data augmentations with minimal supervision. In our problem setting, the useful invariant augmentations relevant for the optimization of the black box function are unknown which is in contrast to self-supervised learning methods where the useful invariant augmentations are typically specified upfront using domain knowledge (Von Kügelgen et al., 2021). Our research presents a surprisingly simple yet effective methodology to identify useful augmentations by learning invariances through the GP marginal likelihood (van der Wilk et al., 2018). We specify the set of possible invariant augmentations for our tasks upfront (notably some of these invariant augmentations are damaging to task performance) and learn the most favourable augmentations for VAE-BO through the marginal likelihood. We then use the selected augmentations in addition to the points queried by BO to enhance the training set of the VAE. The approach yields marked performance gains when affixed to baseline VAE-BO architectures on the MNIST-Norm and classification tasks.

2. Background

We consider the global maximization problem

$$\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \tag{1}$$

where $\mathbf{x}^\star$ is the global optimizer of a black box function $f(\cdot) : \mathcal{X} \to \mathbb{R}$ over a high dimensional and structured input domain $\mathcal{X}$. The analytic form and derivatives of f are unavailable and f is expensive to evaluate.

Bayesian Optimization (BO): BO is an algorithm for solving Equation 1 (Shahriari et al., 2016; Frazier, 2018). BO operates in two stages: (1) A probabilistic surrogate model, typically a GP, is used to infer the black-box function, maintaining uncertainty estimates; (2) An acquisition function is maximised to acquire new points $\mathbf{x}$, trading off exploration and exploitation using the uncertainty estimates of the surrogate. The surrogate and acquisition operate in an iterative fashion, wherby the surrogate is retrained with acquired points *ad libitum* until a solution to the optimization problem is found.

VAE-BO: In VAE-BO (also termed latent space optimization (LSO) (Tripp et al., 2020)), the VAE maps a high dimensional input $\mathcal{X}$ to a low dimensional latent space $\mathcal{Z} \subseteq \mathbb{R}^d$. BO is performed in the latent space of the VAE outputting, $\mathbf{z}^* \in \mathcal{Z}$. The resultant $\mathbf{z}^*$ is projected back to the original input space $\mathbf{x}$ by the VAE decoder to evaluate the black-box f.

VAE-BO with weighted retraining: In VAE-BO the VAE is learned upfront without the possibility to update the latent space using newly acquired points. As such, the VAE struggles to decode points found in regions of the data space that it has not seen in training (Griffiths and Hernández-Lobato, 2020). To address this challenge, Tripp et al. (2020) propose weighted retraining of the VAE. The VAE is retrained on acquired points with a weight biasing the VAE towards regions with favourable black-box function values. There remains however, much room for improvement on weighted retraining VAE-BO through orthogonal methods to improve the structure of the latent space. Our contribution is to introduce one such method, which we discuss next.

3. Methodology

We enhance the sample efficiency of VAE-BO with weighted retraining by learning uncertainty-aware, black-box specific data augmentations. In standard settings, data augmentation refers to creating additional training examples by transforming the input data such that the ground truth labels remain unchanged i.e. are invariant to the transformation (Mikołajczyk and Grochowski, 2018). In black-box problems, we do not have prior knowledge about which invariances apply in the task. Hence, generating relevant data augmentations becomes nontrivial, meaning we must attempt to learn invariances from the limited available data. Given that we are in the small data regime we adopt the method of (van der Wilk et al., 2018) as opposed to more data-hungry methods for learning invariances (Benton et al., 2020).

In our method, referred to as GP_LCB, we specify the set of possible invariant augmentations upfront. In the tasks we examine, our inputs are high dimensional images and so our augmentation set A, contains translations $T(\mathbf{x})$, rotations $R(\mathbf{x})$ and smoothing $S(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$ is a high dimensional image. In our task setting, invariance for certain augmentations cannot be assumed as we are unaware of the difference in the ground truth labels, $|f(\mathbf{x}) - f(a(\mathbf{x}))|$, where $a \in A$. We leverage the posterior predictive distribution of the trained GP surrogate to understand the effect of different augmentations on task performance

$$Y = f(\mathbf{x}), Z = q(\mathbf{x}), Z_{aug} = q(a(\mathbf{x})), \tag{2}$$

$$\mathbb{P}(f^*|Z, Y, Z_{aug}) = \mathcal{N}(\mathbb{E}[f^*|Z, Y, Z_{aug}], \mathrm{Var}[f^*|Z, Y, Z_{aug}), \tag{3}$$

$$\mathbb{E}[f^*|Z, Y, Z_{aug}] = m(Z_{aug}) + k(Z_{aug}, Z)(K + \sigma_n^2 I)^{-1}(Y - m(Z)), \tag{4}$$

$$\mathrm{Var}[f^*|Z, Y, Z_{aug}] = k(Z_{aug}, Z_{aug}) - k(Z_{aug}, Z)(K + \sigma_n^2 I)^{-1}k(Z, Z_{aug}), \tag{5}$$

where $\mathbb{E}[f^*|Z, Y, Z_{aug}]$ represents the predictive mean over the latent representation of the augmented images Z_{aug} and $\mathrm{Var}[f^*|Z, Y, Z_{aug}]$ represents the predictive uncertainty.

We evaluate our entire augmentation set A under the GP posterior and implement a selection strategy to enhance the dataset for weighted retraining of the VAE. Tripp et al. (2020) demonstrate that improved BO performance is obtained by placing more probability

Algorithm 1 VAE-BO with weighted retraining and uncertainty-aware augmentations

Input: $\mathcal{D} = \{\mathbf{x_i}, f(\mathbf{x_i})\}_{i=1}^{N}$, query budget M, objective function $f(\mathbf{x})$, surrogate model GP($\mathbf{z}$), decoder $g(\mathbf{z})$, encoder $q(\mathbf{x})$, retrain frequency r, weighting function $w(\mathbf{x})$
Candidate Augmentations: $\mathcal{D}_{aug} = T(\mathbf{x}) \cup R(\mathbf{x}) \cup S(\mathbf{x})$; $\mathcal{D}_{VAE} = \mathcal{D}_{GP} = \mathcal{D}$
for $1, \ldots, M/r$ **do**
  Train VAE model with encoder g, decoder q on dataset $\mathcal{D}_{VAE}$ weighted by $\mathcal{W} = \{w(\mathbf{x})\}_{\mathbf{x} \in D_{VAE}}$
  for $1, \ldots, r$ **do**
    Compute latent variables for dataset $\mathcal{D}_{GP}$; $\mathcal{Z} = \{\mathbf{z} = q(\mathbf{x})\}_{\mathbf{x} \in \mathcal{D}}$
    Compute latent variables for augmentations $\mathcal{D}_{aug}$; $\mathcal{Z}_{aug} = \{\mathbf{z} = q(\mathbf{x})\}_{\mathbf{x} \in \mathcal{D}_{aug}}$
    Fit surrogate GP to $\mathcal{Z}$ and $\mathcal{D}$, and optimize to obtain new latent query point $\mathbf{z}^*$
    Predictive mean and variance from GP for augmentations, based on uncertainty select $\mathbf{z}^*_{aug}$
    Obtain corresponding input $\mathbf{x}^* = \mathrm{g}(\mathbf{z}^*)$, $\mathbf{x}^*_{aug} = \mathrm{g}(\mathbf{z}^*_{aug})$
    Evaluate $f(\mathbf{x}^*)$ and augment dataset $\mathcal{D}_{GP}$
    Evaluate $f(\mathbf{x}^*_{aug})$ and augment dataset $\mathcal{D}_{VAE}$
  end for
end for
Output: Augmented dataset $\mathcal{D}_{gp}$

mass on high-scoring points. Hence, we sample the augmentations with higher value of the posterior expectation as shown in Equation 4 for the optimization problem. Our selection strategy also penalizes higher GP posterior variance, as we want to include the augmentation-label pairs which exhibit higher confidence within the augmentation set, A together with high objective values.

To achieve this, we incorporate a simple strategy similar to the Lower Confidence Bound (LCB) acquisition function by computing the value of $\mathbb{E}[f^*|Z, Y, Z_{aug}] - \mathrm{Var}[f^*|Z, Y, Z_{aug}]$ for $Z_{aug} = q(a(\mathbf{x}))$ which ensures the selection of augmentations $a(\mathbf{x}), a \in A$ that maximize the objective subject to minimizing the variance. This turns out to be a significant component in enhancing the performance of the VAE-BO with weighted retraining method. The full algorithm for our approach is given in Algorithm 1.

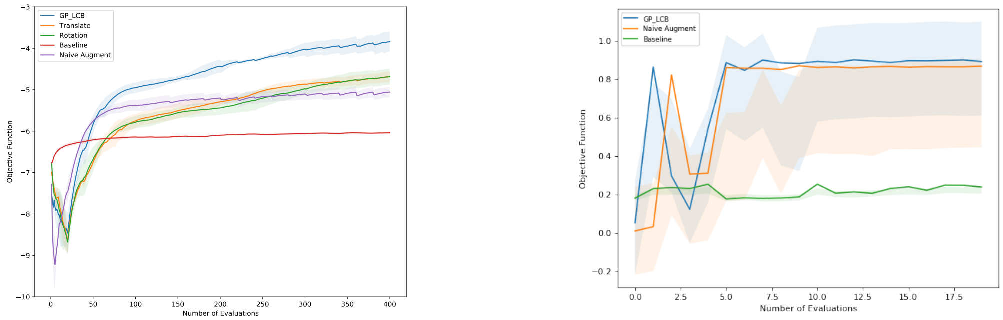
4. Experiments

We consider two tasks, both on the MNIST dataset (Deng, 2012) with four baselines.

MNIST-Norm: In this setting, we consider global minimization of the black box objective function defined as

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathcal{X}} |\mathbf{x} - \mathbf{x}_{ref}| \tag{6}$$

where $\mathbf{x}_{ref}$ is the chosen reference image. This task corresponds to generating the digits with a minimum norm objective with respect to the reference image and the data consists of the



(a) MNIST-Norm.

(b) MNIST-Classifier.

Figure 1: BO performance on the MNIST tasks.

normalized images from the MNIST dataset.

MNIST-Classifier: We use an image classifier to obtain the probability distribution over the labels. A scoring function, modeled as a squared exponential centered at the digit “3”, returns a score for each digit. The objective function is the expected score based on the image classifier and scoring function.

We compare our proposed approach with the following baselines:

Naive Augment: A direct extension of self-supervised augmentation methods for our black-box optimization setting. Here all augmentations are considered. Notably, translational augmentations are detrimental to performance because the norm objective value is sensitive to translations. Our method learns not to append translational augmentations to the VAE training set and as such, achieves improved performance.
Translate: Consider only translational augmentations $T(\mathbf{x})$.

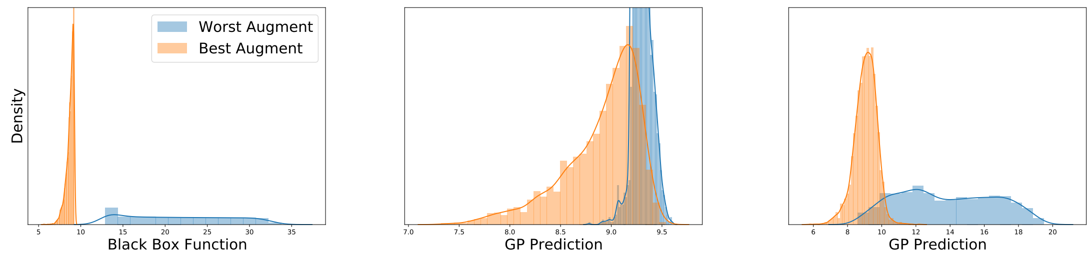
Rotate: Consider only rotational augmentations $R(\mathbf{x})$.

Baseline: VAE-BO with weighted retraining as described in (Tripp et al., 2020).

To evaluate the efficacy of our proposed method, we used 4 random splits for the initial dataset and further, ran all experiments for 3 different seeds. The results are given in Figure 1. For the MNIST-Norm task, GP_LCB method clearly outperforms other approaches considered for high dimensional BO. Notably, in the MNIST-Classifier task, Naive Augment method performs comparably as image classification is invariant to all transformations in the set A.

5. Discussion

For the MNIST-Norm task, we plot a visualization in Figure 2 to understand the effectiveness of learning augmentations. We are interested in minimizing equation 6. Thus, high-scoring points will result in low objective function output. The visualization indicates:



(a) True Objective Function (b) GP Augmentation: Start (c) GP Augmentation: End

Figure 2: Density plot for showing that the GP learns to suggest better augmentations as optimization proceeds.

1. **The methodology learns to identify high-scoring augmentations:** We plot the learned augmentation objective function values for the high-scoring and the low-scoring image augmentations. As BO progresses, our method successfully discriminates between the high and low-scoring image augmentations.

2. **Inspecting high-scoring augmentations:** We plot the top-scoring augmentation from our method for three iterations in Figure 3 of Appendix A at the start, middle and the end of optimization. We note that our method augments the VAE training set with high-scoring augmentations as the optimization progresses. This causes the feasible region to extend to high-scoring points, enabling the optimization to find better points as optimization proceeds.

6. Conclusion

Tackling high dimensional black-box optimization problems with VAE-BO has proven challenging due to the absence of a latent space structured to facilitate BO. We propose a simple method to tackle this problem using black-box specific augmentations, leveraging the marginal likelihood to learn invariant augmentations and enhance the sample-efficiency of the VAE-BO with weighted retraining method. We achieve superior performance to naive augmentation strategies on the MNIST-norm task. The experimental results indicate that incorporating invariances without understanding their effect can hurt BO performance.

Our work also highlights the importance of uncertainty estimation when performing informed data augmentation. In future work, we plan to incorporate VAE uncertainty (Notin et al., 2021) and metric learning VAE-BO (Grosnit et al., 2021) to improve the quality of augmentations, as well as consider a broader range of VAE-BO architectures (Maus et al., 2022). We would also like to extend our method to further real-world tasks with non image-based inputs, namely chemical (Jin et al., 2018; Moss and Griffiths, 2020) and materials design (Griffiths et al., 2021).

References

Rika Antonova, Akshara Rai, Tianyu Li, and Danica Kragic. Bayesian optimization in variational latent spaces with dynamic compression. In *Conference on Robot Learning*, pages 456–465, 2020.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in Neural Information Processing Systems*, 33:17605–17616, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Yutian Chen, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver, and Nando de Freitas. Bayesian optimization in AlphaGo. *arXiv preprint arXiv:1812.06855*, 2018.

Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Stephan Eissman, Daniel Levy, Rui Shu, Stefan Bartzsch, and Stefano Ermon. Bayesian optimization and attribute adjustment. In *Proc. 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Peter I Frazier. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 2020.

Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. *arXiv preprint arXiv:2105.02637*, 2021.

Antoine Grosnit, Rasul Tutunov, Alexandre Max Maraval, Ryan-Rhys Griffiths, Alexander I Cowen-Rivers, Lin Yang, Lin Zhu, Wenlong Lyu, Zhitang Chen, Jun Wang, et al. High-dimensional Bayesian optimisation with variational autoencoders and deep metric learning. *arXiv preprint arXiv:2106.03609*, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018. URL `http://proceedings.mlr.press/v80/jin18a.html`.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014.

Byungsoo Ko and Geonmo Gu. Embedding expansion: Augmentation in embedding space for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7255–7264, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Xiaoyu Lu, Javier Gonzalez, Zhenwen Dai, and Neil Lawrence. Structured variationally auto-encoded optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3267–3275. PMLR, 10–15 Jul 2018. URL `http://proceedings.mlr.press/v80/lu18c.html`.

Natalie Maus, Haydn T Jones, Juston S Moore, Matt J Kusner, John Bradshaw, and Jacob R Gardner. Local latent space Bayesian optimization over structured inputs. *arXiv preprint arXiv:2201.11872*, 2022.

Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

Riccardo Moriconi, Marc Peter Deisenroth, and KS Sesh Kumar. High-dimensional Bayesian optimization using low-dimensional feature spaces. *Machine Learning*, 109(9):1925–1943, 2020.

Henry B Moss and Ryan-Rhys Griffiths. Gaussian process molecule property prediction with flowmo. *arXiv preprint arXiv:2010.01118*, 2020.

Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in VAE latent space using decoder uncertainty. *CoRR*, abs/2107.00096, 2021. URL `https://arxiv.org/abs/2107.00096`.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 2016.

Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 2021.

Eero Siivola, Javier Gonzalez, Andrei Paleyes, and Aki Vehtari. Good practices for Bayesian optimization of high dimensional structured spaces, 12 2020.

Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *Advances in Neural Information Processing Systems*, 33, 2020.

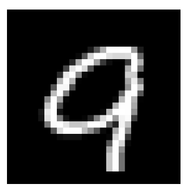
Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, 2021.

Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. *Advances in Neural Information Processing Systems*, 31, 2018.

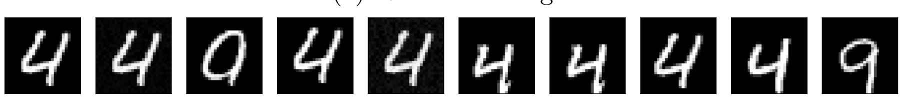
Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34, 2021.

Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 521–528, 2002.

Appendix A.



(a) Reference Image



(b) GP Augmentation : Start

(c) GP Augmentation : Mid

(d) GP Augmentation : End

Figure 3: Top scoring image augmentations from GP predictions

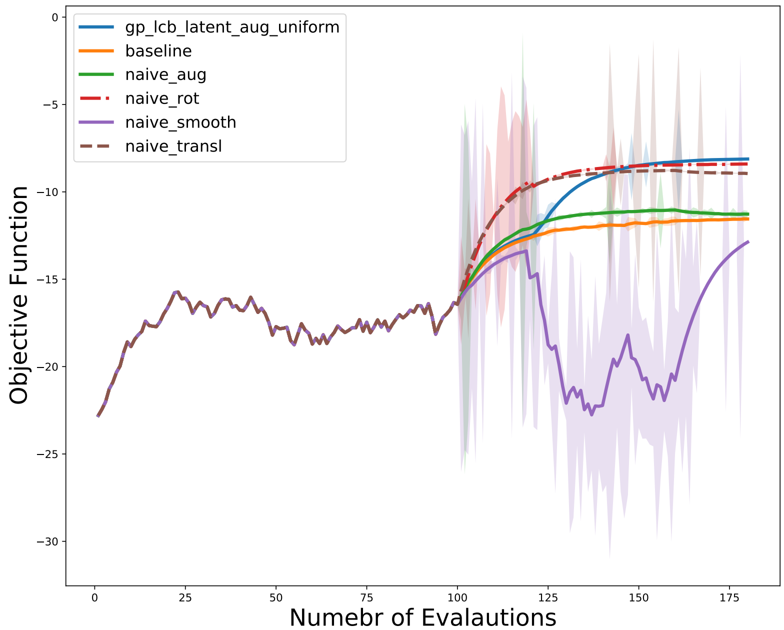


Figure 4: Performance of BO on CelebA-Norm task (Liu et al., 2015). CelebA-Norm task is invariant to the rotation and translation baselines. However, we can observe that our proposed approach learns the augmentations suitable for the task and outperforms the baselines.