Landmark Estimation and Image Synthesis Guidance using Self-Supervised Networks

A Thesis

Submitted for the Degree of

Master of Technology (Research)

in the

Faculty of Engineering

by

Tejan Karmali

Department of Computational and Data Sciences
Indian Institute of Science
Bangalore – 560 012 (India)

February 2023

Declaration of Originality

I, **Tejan Karmali**, with SR No. **06-18-02-10-22-19-1-16613** hereby declare that the material presented in the thesis titled

Landmark Estimation and Image Synthesis Guidance using Self-Supervised Networks

represents original work carried out by me in the **Department of Computational and Data Sciences** at **Indian Institute of Science** during the years **2019-2022**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discussions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: R. Venkatesh Babu Advisor Signature

DEDICATED TO

my grandmother,
Late Smt. Sushilabai Karmali

Acknowledgements

First and foremost, I sincerely thank my advisor Dr. R. Venkatesh Babu for supporting me throughout my research and for always pushing to bring out the best in me. His continuous guidance on approaching problems from different perspectives and following first-principles has made me a better researcher. I am grateful to Dr. Varun Jampani for teaching me to ask the right questions in research and to choose the problems that are practically relevant. His critical feedback resulted in pushing me to think beyond the boundaries of my thought. I extend my gratitude to Dr. Maneesh Singh as an industrial mentor, inspiring me in conducting empirically driven R&D grounded in challenges faced in engineering practical solutions and in methodically and logically constructing and evaluating solutions. I extend my gratitude to Dr. Biplab Banerjee for the insightful comments which helped me further improve the thesis.

I am fortunate to have collaborated with Abhinav, Harsha, Susmit, Harsh, Rishubh, and Rahul for my research projects. I am highly indebted to Jogendra, Harsh, Kaushal, Naman, Gaurav, and Vikash for their constant support and motivation during my stay. Special thanks to Harsh for always being there in the times of need and supporting through the highs and lows of research. The discussions we have has over various ideas have helped me in shaping my research. I am lucky to have exemplary seniors like Jogendra and Sravanti who have inspired me to strive for the best. Good luck to Sunandini, Abhipsa, Ankit, and Rishubh in your PhD. Thanks to Rakshitha and Mary ma'am for their support in administrative matters.

Finally, without the love and support of my parents, Naresh and Seema, none of this would have been possible. For all the sacrifices they have made for me, I shall be eternally grateful.

Abstract

The exponential rise in the availability of data over the past decade has fuelled research in deep learning. While supervised deep learning models achieve near-human performance using annotated data, it comes with an additional cost of annotation. Additionally, there could be ambiguity in annotations due to human error. While an image classification task assigns one label to the whole image, as we increase the granularity of the task to landmark estimation, the annotator needs to pinpoint the landmark accurately. The self-supervised learning (SSL) paradigm overcomes these concerns by using pretext task based objectives to learn from large-scale unannotated data. In this work, we show how to extract relevant signals from pretrained self-supervised networks for a) a discriminative task of landmark estimation under limited annotations, and b) increasing perceptual quality of the images generated by generative adversarial network.

In this first part, we demonstrate the emergent correspondence tracking properties in the non-contrastive SSL framework. Using this as supervision, we propose LEAD which is an approach to discover landmarks from an unannotated collection of category-specific images. Existing works in self-supervised landmark detection are based on learning dense (pixel-level) feature representations from an image, which are further used to learn landmarks in a semi-supervised manner. While there have been advances in self-supervised learning of image features for instance-level tasks like classification, these methods do not ensure dense equivariant representations. The property of equivariance is of interest for dense prediction tasks like landmark estimation. In this work, we introduce an approach to enhance the learning of dense equivariant representations in a self-supervised fashion. We follow a two-stage training approach: first, we train a network using the BYOL [34] objective which operates at an instance level. The correspondences obtained through this network are further used to train a dense and compact representation of the image using a lightweight network. We show that having such a prior in the feature extractor helps in landmark detection, even under a drastically limited number of annotations while also improving generalization across scale variations.

Next, we utilize the rich feature space from the SSL framework as a "naturalness" prior

Abstract

to alleviate unnatural image generation from Generative Adversarial Networks (GAN), which is a popular class of generative models. Progress in GANs has enabled the generation of high-resolution photorealistic images of astonishing quality. StyleGANs allow for compelling attribute modification on such images via mathematical operations on the latent style vectors in the $\mathcal{W}/\mathcal{W}+$ space that effectively modulates the rich hierarchical representations of the generator. Such operations have recently been generalized beyond mere attribute swapping in the original StyleGAN paper to include interpolations. In spite of many significant improvements in StyleGANs, they are still seen to generate unnatural images. The quality of the generated images is a function of, (a) richness of the hierarchical representations learned by the generator, and, (b) linearity and smoothness of the style spaces. In this work, we propose Hierarchical Semantic Regularizer (HSR) which aligns the hierarchical representations learnt by the generator to corresponding powerful features learned by pretrained networks on large amounts of data. HSR not only improves generator representations but also the linearity and smoothness of the latent style spaces, leading to the generation of more natural-looking style-edited images. To demonstrate improved linearity, we propose a novel metric - Attribute Linearity Score (ALS). A significant reduction in the generation of unnatural images is corroborated by improvement in the Perceptual Path Length (PPL) metric by 15% across different standard datasets while simultaneously improving the linearity of attribute-change in the attribute editing tasks.

Publications based on this Thesis

The following works were conducted during my M.Tech.(Res.) which form this thesis:

1. **Tejan Karmali***, Abhinav Atrishi*, Sai Sree Harsha, Susmit Agrawal, Varun Jampani, and R. Venkatesh Babu. "LEAD: Self-Supervised Landmark Estimation by Aligning Distributions of Feature Similarity", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022.

2. **Tejan Karmali**, Rishubh Parihar, Susmit Agrawal, Harsh Rangwani, Varun Jampani, Maneesh Singh, and R. Venkatesh Babu. "Hierarchical Semantic Regularization of Latent Spaces in StyleGANs", European Conference on Computer Vision (ECCV) 2022.

The following works were conducted during my M.Tech.(Res.) but are not part of this thesis:

1. Rahul M.V., **Tejan Karmali**, Sarthak Sharma, Aurobrata Ghosh, R. Venkatesh Babu, László A. Jeni†, and Maneesh Singh†. "Deep Implicit Surface Point Prediction Networks", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

2. Harsh Rangwani, Naman Jaswani, **Tejan Karmali**, Varun Jampani, and R. Venkatesh Babu. "Improving GANs for Long-Tailed Data through Group Spectral Regularization", European Conference on Computer Vision (ECCV) 2022.

*denotes equal contribution
†denotes equally advised

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The success of deep learning methods has led to deployable solutions in prominent computer vision tasks such as object recognition [42, 97, 28, 102], object detection [87, 32, 88], semantic segmentation [13, 14, 43]. This is driven by the data collected over the years at a massive scale. Methods for the supervised learning paradigm are reliant on data sources which also provide annotations. If the deep neural network trained in supervised fashion is able to comprehend the underlying semantic concepts in the data, then it should be possible to transfer the trained model to other tasks and datasets. However, the performance of the transferred model carries a bias towards the original task or the data on which it was trained. Hence, this restricts the usage of the models to the related task or data distributions.

Self-supervised learning (SSL) methods focus on designing tasks from unannotated data [53]. The principle behind designing such a task is to obtain supervision from unannotated data which could make the model learn the underlying semantic concepts. The model is then transferred to learn a useful downstream task, leading to competitive performance compared to the supervised learning approaches. Self-supervised learning reduces the dependency over the labels. Competitive performance of SSL methods hints that a large fraction of knowledge required to solve a task can be extracted from the data even without access to its labels.

The research in self-supervised learning tasks has focused on downstream task of classification. Annotation process for classification involves one label per image. Therefore, it is easier for humans to annotate for classification and obtain large amount of annotated data. There have been recent advances in adapting these techniques for detection and segmentation. Compared to recognition where 1 label per image is required, getting annotations for detection (bounding box per object in the image) and segmentation (pixel-wise annotations) is highly demanding in terms of human labor cost as well as hours spent to annotate. Therefore, self-supervised learning is useful to reduce the reliance on the annotations in such tasks.

The self-supervised objectives have shown to be generalizable to task transfer. While these are discriminative tasks, it is still an open question as to how can features learnt on such tasks be used to improve the generative models.

Landmark Estimation. Image landmarks are distinct locations in an image that can provide useful information about the object, like its shape and pose. For example, for face images, the landmarks points can be the pupil of eyes, tip of the nose, and lip corners. Knowing the locations of facial landmarks can help in tasks like head pose estimation [37]. In general, landmarks are used to predict camera pose using Structure-from-Motion [39]. Landmark detection is a well studied problem in computer vision [132, 119, 131, 130, 29, 66] that was initially accomplished using annotated data. Landmark annotation requires a person to accurately label the pixel location where the landmark is present. This makes annotation a laborious, biased, and ambiguous task, motivating the need for newer paradigms such as few-shot learning [133, 126, 100, 115] and self-supervised learning [27, 127, 75, 44, 34].

Prior works in self-supervised landmark detection rely on the principles of reconstruction [130, 71] and equivariance [104, 106]. These methods are trained using dense objectives that are satisfied by every pixel (or by every patch of pixels, due to downsampling). This tends to capture only local information around each pixel, and is unaffected by structural changes in the image (like patch shuffling). Recent progress in Self-supervised learning proposes pretext tasks of instance-discriminative methods [44, 17, 19, 21, 16], which are shown to be superior for the purpose of pre-training. In this work we seek to exploit the advantage offered by recent works in contrastive learning to estimate landmarks over them.

Smoothening GAN Latent Space. Generative Adversarial Networks are at the forefront of image synthesis models. There have been advances in the Generative Adversarial Networks (GAN) to improve the photorealism of the generated images. StyleGAN [54, 56] is a landmark work which innovates architectural aspects of GAN that improves the photorealism of the images. StyleGAN also introduces an interpretable latent space which can be used for downstream image manipulation tasks [93, 92]. However, there exist two problems with StyleGAN, a) the latent space of generator has a subspace that maps to unnatural images which do not resemble the training data, and b) the image manipulation models assumes a linear model w.r.t. the attribute, while there is no guarantee for that during the training of GAN. In this work, we alleviate the first issue with the natural prior learned by SSL networks. We then explore the relationship between naturalness of the image and smoothness of the latent space.

In summary, in this thesis we have looked into the problems of landmark estimation and improving the latent space of GAN. We derive solutions to these problems from self-supervised

pretraining. The solutions to the first problem enables landmark estimation with few annotated data. The solution to the second problem improves the GAN latent space which leads to improved photorealism of the generated images as well as a more control in image manipulation task. The rest of the chapters in this thesis are organized as follows: In Chapter 2, we review background on self-supervised learning, landmark estimation and generative adversarial networks. In Chapter 3, we present a our method LEAD, a self-supervised pretraining method for landmark estimation. In Chapter 4, we improve the photorealism of generated images and linearize the latent space of GAN by aligning the intermediate feature spaces of Generator and self-supervised network. Finally, Chapter 5 describes our key contributions and possible future directions.

Chapter 2

Related Work

While the research in self-supervised pretraining has mainly explored the downstream task of classification, its applications in other tasks landmark estimation, which incur more annotation cost and image generation are under-explored. In the rest of these chapters, we provide an overview of existing methods for landmark estimation and image generation which are most related to our work.

2.1 Landmark Estimation

Unsupervised Landmark prediction: The landmark prediction task has traditionally been studied in a supervised learning setting. Given the annotation-heavy nature of the problem, recent approaches have emphasized on unsupervised pretraining to learn information-rich features. These approaches can be divided based on two principles: equivariance and image generation.

Thewlis et al. [106] proposed an approach that uses equivariance of the feature descriptors across image warps as an objective for supervision. Suwajanakorn et al. [101] extended this idea for 3D landmark discovery from multi-view image pairs. This idea has also been used to model symmetrically deformable objects [107], and to learn object frames [105]. Further, Thewlis et al. [104] supplemented it using the principle of transitivity, which ensured that the descriptors learnt are robust across images.

Generative objective for landmark detection was initially used by Zhang et al. [130] and Lorenz et al. [71]. The main idea is to learn an image autoencoder with a landmark discovery bottleneck. Jakab et al. [49] coupled it with conditional image generation which could decouple the appearance and pose over an image pair. The key downside of these methods is that, the discovered landmarks are not interpretable. This was addressed by [50] where the landmark

bottleneck is interpretable, due to availability of unpaired poses. [72] detects more semantically meaningful landmarks using self-training and deep clustering.

Self-supervised learning: Self-supervised learning follows the paradigm of training a network using a pretext task on a large-scale unlabeled dataset, followed by training a shallow network using limited annotated data. Initial works explored pretext tasks like classification of image orientation [31], patch-location prediction [80, 25], image colorization [127, 128], and clustering [11, 8]. While transformation invariant representation learning of an image [134, 67, 99, 64] has been extensively studied in supervised learning, the idea has outperformed prior pretext tasks when modelled as a contrastive learning [38] problem [117, 44, 17, 19, 21, 16, 46, 122, 45, 9, 108] in the self-supervised learning setting. Here, the main idea is to push the embeddings of the *query* image and its augmentation ("positive" image) closer, while repelling it against the embeddings of the "negative" images (all other images). This is achieved using the InfoNCE [111] loss. A key disadvantage of these methods is the use of a large number of "negative" images which incurs high memory requirements. For eg., SimCLR [15] recommends using large batch size, which naturally contains a many negatives against a single positive example. Another prominent method, MoCo [44], uses a memory bank to store features across different batches.

The issue of large memory requirement was mitigated by methods like [34] and [18], which achieve competitive performance without "negative" images. While both of these seminal works concentrate on the classification task, there are some advances in adapting these techniques for dense prediction tasks like detection and segmentation [89, 81, 120, 114] as well. The only work that adopted the contrastive learning objective for the task of landmark prediction is ContrastLandmarks (CL) [22], where they train the network with the InfoNCE [111] objective. To adapt the output feature map to the resolution of the image, they use a hypercolumn representation from features across different layers. The key differences between this work and LEAD are: 1) We learn dense and compact descriptors via a novel correspondence matching guided dimensionality reduction objective while CL uses the objective proposed by Thewlis et al. [105], and 2) We do not use any "negative" images, as landmarks are ubiquitous in a category-specific dataset.

2.2 Smoothening Latent Space of GAN

Generative Adversarial Networks. GAN proposed by Goodfellow *et al*. [33] a combination of two neural networks, *i.e*. generator G and discriminator D. For image synthesis the goal of D is to differentiate between real and generated images, whereas the G tries to fool the discriminator into classifying generated images as real. In the recent years several improvements

in architecture [58, 59, 85, 36, 76], optimization objectives [5, 7, 73, 74] and regularization [77, 35] have made GANs an ubiquitous choice for image synthesis. It has been observed that GANs developed for large scale datasets, suffer mode collapse when trained on limited data. Augmentation methods like DiffAugment [135], ADA [55], ContraD [51] , APA [52] etc. mitigate the collapse by reducing the overfitting of discriminator on limited samples.

Hierarchical Representations. In classical vision, methods which decompose image into a hierarchy have been exploited for the tasks of image stitching, manipulation and fusion [2, 23]. Building on this motivation Shocher *et al.* [94] develop an image translation and manipulation method, which exploits hierarchical consistency of features of generator and a classifier. However this method is restricted to single image translation and manipulation. In contrast our work we aim to train a smooth and generalizable GAN which can simultaneously generate diverse images, by using semantic hierarchical consistency of features.

Knowledge Transfer Using Pre-Trained Features. Using pre-trained features trained on large scale datasets (*e.g.* ImageNet etc.) [41, 103] have been useful for various downstream tasks across applications [26, 47, 110, 123]. The recent development of the self-supervised approaches for representation learning [15, 85, 34] have further immensely improved the quality of features learnt. These features are being used in various applications like part segmentation, localization etc. without being explicitly trained on such tasks [12], which motivates our work which aims to transfer these semantic properties to G's feature space. Currently much work for transfer learning for GANs has focused on the fine-tuning large GANs using a few images for adapting it to a different domain [69, 78, 82, 79]. Recently a concurrent work [62] also aims to use pre-trained features to improve GANs. However their goal is to improve discriminator. On contrary we aim to enrich GAN feature space by imparting it with semantic properties, leading to a disentangled and smooth latent space.

Image Editing Using Latent Space Interpolations. Latent space of pre-trained StyleGAN models is highly structured [93] and is popularly used to perform realistic image edits in the generated images [1, 93, 118, 48, 92, 125, 4]. The primary idea in most of these approaches is to find a direction in the the extended latent space $\mathcal{W}+$ for editing attributes and transforming a latent code by moving in that direction to perform edits. StyleCLIP [84] learns the directions for attribute editing by getting the guidance from pretrained CLIP [86]. On the contrary, our work imposes constraints so that latent space has more naturally interpretable directions when used by the GAN-based image editing methods.

Chapter 3

LEAD: Self-Supervised Landmark Estimation by Aligning Distributions of Feature Similarity

3.1 Introduction

Most of the existing research in the field of self-supervised learning is focused towards the task of instance-level classification. Amongst the proposed pretext tasks for self-supervision, instance-discriminative methods [44, 17, 19, 21, 16], are known to be superior for the purpose of pre-training. Recent methods utilize these objectives for dense prediction tasks as well, where a distinct label is predicted either for every pixel (segmentation, landmark detection) or patch of pixels (detection) [89, 81, 120, 114]. The power of contrastive training is leveraged for landmark detection by Cheng et al. [22] to achieve state-of-the-art performance using Momentum Contrast (MoCo)-style [44] pre-training. This work demonstrates equivariant properties in the network when trained with a contrastive objective. This property is realised by extracting a hypercolumn-style feature map from the image. But using such a high-dimensional feature map (3840d for ResNet50 due to stacking up of features), which is 60× larger than existing approaches, to represent an image is not scalable to large images.

Our key insight is based on the observation that self-supervised training on category-specific datasets (dataset that consists of images that belong to only single category) leads to meaningful part-clustering in feature space. We further utilize this finding to propose a dense self-supervised objective for landmark prediction. Specifically, LEAD involves two stages: (1) Global representation learning, and (2) Correspondence-guided dense and compact representation learning. The network from stage 1 leads to meaningful part clustering in the feature space, and hence

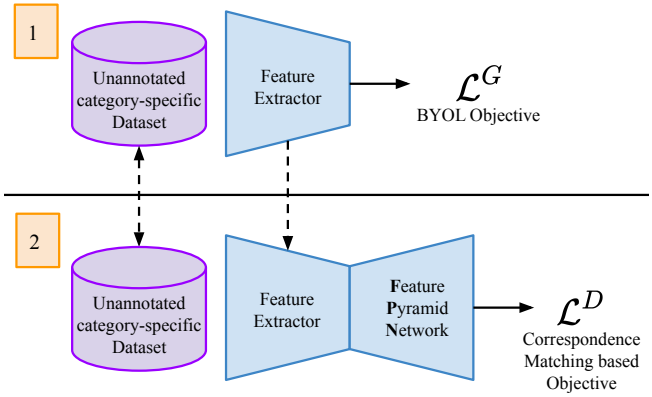


Figure 3.1: **LEAD Framework overview.** Two stage process for self-supervised landmark detection. **First**, an *instance-level* feature extractor is trained on a large Unannotated category-specific Dataset with the BYOL [34] objective. **Second**, using the correspondence matching property of the instance-level feature extractor, a *pixel-level* FPN [68] based feature extractor is trained on the same dataset. Finally, the pixel-level feature extractor is used to train a supervised regressor on limited data of landmark annotations.

can be used to draw correspondences between two images. This can be used for pixel/patch level training to learn compact descriptors that represent the spatial information of the image. We illustrate the high-level idea in Fig. 3.1, and include a detailed architecture in Fig. 3.2.

We measure the performance of LEAD using percentage of inter-ocular distance (IOD). Landmarks estimated using our feature extractor show ∼10% improvement over prior art on facial landmark estimation, along with a boost in performance in the setting of severely limited annotations. We further obtain improved generalization to alignment and scale changes in the input images.

In summary, our contributions are:

- We show the emergence of high-fidelity landmarks in Bootstrap-Your-Own-Latent (BYOL) [34] style instance-level feature learning framework. (Sec. 3.2.2)
- We utilise this property to guide the learning of dense and compact feature maps of the image via a novel dimensionality reduction objective. (Sec. 3.2.3)
- Our evaluations show significant improvements over prior art on challenging datasets and across degrees of annotations, both qualitatively and quantitatively. (Sec. 4.3)

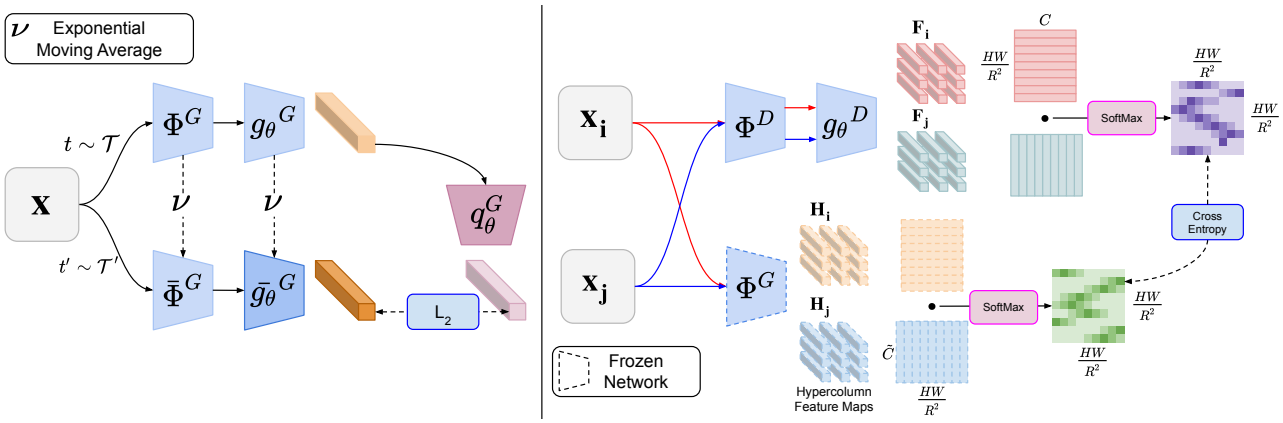


Figure 3.2: **LEAD training overview. Left:** Stage 1 of the training feature extractor Φ^G with BYOL objective, where the representation of key augmentation is predicted from query augmentation. **Right:** Stage 2 involves using frozen Φ^G to obtain dense correspondences, which are used to guide trainable network Φ^D to obtain dense and compact image representation. The correspondences, which also describe similarity between features, are converted to a probability distribution over spatial grid, by using a softmax (ref. Fig. 3.3). Distribution of Feature Similarity from Φ^D is guided by that from Φ^G using a cross-entropy loss.

3.2 Method

3.2.1 Background

Let $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}\}$ be a large-scale unannotated category-specific dataset. Our goal is to learn a feature extractor Φ, which, given $\boldsymbol{x} \in \mathcal{X}$ as input gives a feature map as output. As a pretext task, prior works have attempted to enforce instance-level representations to be invariant to transformations [22], and impose consistency on the dense pixel-level representations. In our approach LEAD, we use two stages. First, we learn a global representation of the image that leads to its part-wise clustering as described in Sec. 3.2.2. Then, we make use of this prior to guide the learning of a dense and compact representation of the image by a novel dimensionality reduction objective, which matches the distributions of feature similarity across two images, as described in Sec. 3.2.3.

3.2.2 Global Representation Learning

We follow the algorithm proposed in BYOL [34] to learn an instance-level representation of the image. BYOL uses an online network Φ^G and a target network $\bar{\Phi}^G$. Φ^G and $\bar{\Phi}^G$ share the same architecture, but the weights of $\bar{\Phi}^G$ are obtained using a momentum average of weights of Φ^G across multiple training iterations. These backbone networks are followed by projection heads g_θ^G and $\bar{g}_\theta{}^G$. Similar to the weights of the backbone, the weights of $\bar{g}_\theta{}^G$ are obtained using a

momentum average. The necessity for the projection heads in self-supervised training has been discussed extensively in SimCLR [16], where the authors find the representations of last layer before the projection head to be most useful. Additionally, the online network has a prediction head q_θ^G (Fig. 3.2).

The training objective is to predict the representation of one view of the image from another using q_θ^G. Given an image $\boldsymbol{x}$, its two views $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are generated by applying augmentations. We refer to $\boldsymbol{x_1}$ as the *query image* and $\boldsymbol{x_2}$ as the *key image*. Φ^G and $\bar{\Phi}^G$ generate features corresponding to $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ respectively. These feature maps are then projected using g_θ^G and $\bar{g_\theta}^G$ respectively to obtain the instance-level representations $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$. Since both the views belong to the same instance, the predictor q_θ^G is trained to predict $\boldsymbol{z_2}$ given $\boldsymbol{z_1}$. The squared L_2 loss shown below is minimized for training:

$$\mathcal{L}^G = \|q_\theta^G(\boldsymbol{z_1}) - \boldsymbol{z_2}\|_2^2 \tag{3.1}$$

As shown in CL [22], the self-supervised contrastive objective produces hypercolumn based feature maps that have semantic understanding of the correspondences at pixel level between two images. In addition, we find that the BYOL objective gives significantly better correspondences than the MoCo objective, as shown in the Fig. 3.3. Hypercolumns are used here, since the self-supervised networks downsample the input image largely to obtain an instance-level representation. Creating a hypercolumn based feature map involves concatenating the intermediate feature maps along the channel dimension. Since the intermediate feature maps have lower spatial resolution than the original input image, they are upsampled to match the resolution of the input image. This has been illustrated in Fig. 3.3. However, hypercolumns incur a large cost in terms of memory. In the next section, we improve upon this by injecting pixel-level information into the network, thereby learning a dense and compact representation of the image.

3.2.3 Dense and Compact Representation Learning

The bottleneck in framing the dense feature map learning problem is pixel-level correspondences. In the case of global feature vector learning, the image to form the positive pair is drawn by applying augmentation to the input image. But in the case of dense feature map learning, the correspondences between points in the query and the key images are not known. But since we have a trained BYOL network that can find *reasonable* (ref. Fig. 3.3) correspondences across images, we use it to guide the learning of dense and compact feature maps of

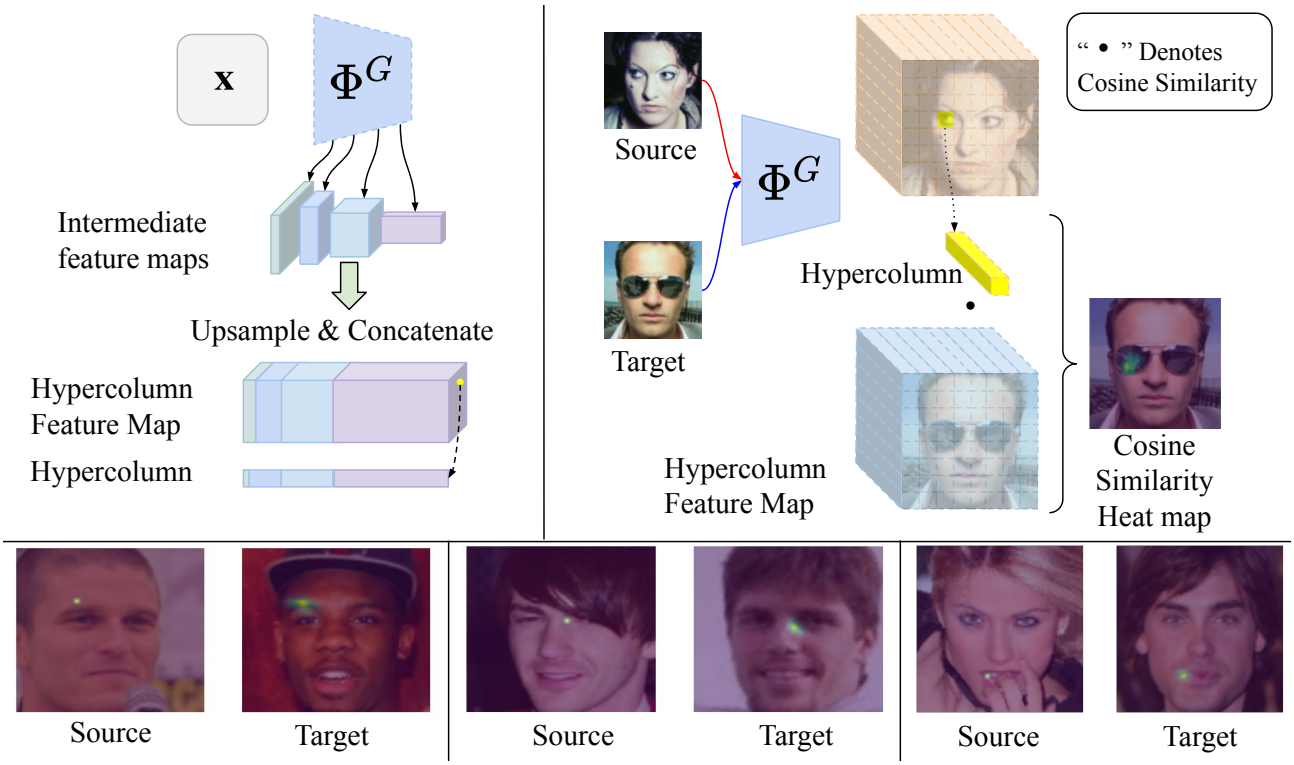


Figure 3.3: **Correspondence matching** performance using the hypercolumn representation. **Top Left:** Procedure to create hypercolumn from intermediate feature maps, by upsampling and concatenating them. Each feature vector across the spatial dimension denotes a hypercolumn. **Top Right:** Correspondence matching is performed from a point in the source image to the target image by taking cosine similarity of the hypercolumn corresponding to the source point and target's hypercolumn feature map, followed by softmax to obtain a heat map. **Bottom:** Examples of correspondence matching. Note that the resultant distribution peaks around the tracked point.

images.

For the hypercolumn feature vector (or hypercolumn, for short), the ability to track a semantic point across two image depends on the distance between them in the $\tilde{C}$-d feature space. In this space, the features are clustered according to their semantic meaning. We aim to learn a compact feature space which has this property.

We now elaborate on the training method followed to learn such a low-dimensional feature space (Fig. 3.2). We train an encoder-decoder network $\Phi^D : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times C}$. The encoder is initialised with Φ^G trained in Sec. 3.2.2. The output of the encoder goes to the projection head g_θ^D. We aim to retain the relationship defined by the cosine similarity between the hypercolumn feature maps from two images in their compact feature maps which are to be learnt. Let $\boldsymbol{x}_i$, $\boldsymbol{x}_j \in \mathcal{X}$ be two images, whose hypercolumn feature maps are $\boldsymbol{H}_i$,

$\boldsymbol{H}_j \in \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times \tilde{C}}$ respectively. Note that, $\tilde{C} \gg C$, which makes the hypercolumn representation memory-intensive during inference. Let $\boldsymbol{F}_i = g_\theta^D(\Phi^D(\boldsymbol{x}_i))$ be compact feature maps of the respective images. Let $\boldsymbol{f}_i^{uv} \in \boldsymbol{F}_i$ be a feature vector at spatial location (u, v) in feature map $\boldsymbol{F}_i$. Similarly, let $\boldsymbol{h}_i^{uv} \in \boldsymbol{H}_i$ be a feature vector at spatial location (u, v) in the hypercolumn feature map $\boldsymbol{H}_i$. Since the aim is to retain the inter-feature relationship, we use cosine similarity as the measure of relationship between two feature vectors. To cover the whole feature space, we take cosine similarity with all the feature vectors. This relationship between the feature vector and the feature space as a probability distribution indicates which subspace of the feature space the feature vector is most similar to:

$$q_{ij}^{uv}[k,l] = \frac{\exp(\boldsymbol{f}_i^{uvT} \boldsymbol{f}_j^{kl}/\tau)}{\sum_{m,n=0}^{\frac{H}{R},\frac{W}{R}} \exp(\boldsymbol{f}_i^{uvT} \boldsymbol{f}_j^{mn}/\tau)} \tag{3.2}$$

where τ is temperature, which is a hyperparameter controlling the concentration level of the probability distribution q_{ij}^{uv} [117].

Similarly, such a relationship can be defined for $\boldsymbol{h}_i^{uv}$ with $\boldsymbol{H}_j$ as well. We denote this probability distribution as p_{ij}^{uv}. This ultimately leads us to optimize q_{ij}^{uv} to mimic p_{ij}^{uv}. We use cross-entropy between the both of them to achieve this objective:

$$\mathcal{L}^D = \sum_{u,v=0}^{\frac{H}{R},\frac{W}{R}} \sum_{k,l=0}^{\frac{H}{R},\frac{W}{R}} -p_{ij}^{uv}[k,l] \cdot \log(q_{ij}^{uv}[k,l]) \tag{3.3}$$

3.2.4 Landmark Detection

At this stage we have a feature extractor that is learned in a self-supervised fashion. To obtain the final landmark prediction, a limited amount of annotated data is used. Feature extractor is frozen and a lightweight predictor Ψ is trained over it. Ψ gives landmark heatmaps as output ($\in \mathbb{R}^{H \times W \times K}$) where K is the number of landmarks present). Expected location of the landmark k, weighed by the heatmap gives its final position $(\hat{x}^k, \hat{y}^k)$. It is supervised by the annotated location of the landmark (x^k, y^k) with an l_2 loss.

3.3 Experiments

Dataset: We evaluate LEAD on human faces. Following prior works, we use the CelebA [70] dataset containing 162,770 images for pretraining the network. To evaluate the learnt representation, four datasets are used. We firstly use MAFL which is a subset of CelebA. Two variants of AFLW [63] are used: the first being AFLW_M which is the partition of AFLW with crops from

Figure 3.4: **Landmark Matching:** We observe that LEAD is able to predict the landmarks in the Query images (middle rows) using reference annotated image (first row). We compare our performance against DVE [104] and ContrastLandmarks [22] on a spectrum of head rotations.

MTFL [131]. It contains 10,122 training images and 2,995 test images. The second variant is AFLW$_R$, in which tighter crops of the face are used. This comprises of 10,122 training images and 2,991 testing images. We further use the 300-W [91] dataset which has 68 annotated landmarks, with 3148 training and 689 testing images. All the datasets are publicly available.

Implementation Details: We use a ResNet50 [40] backbone to train instance-level BYOL representation in stage 1. In stage 2, the feature extractor of the trained ResNet in stage 1 is used as weight initialization for the encoder. The decoder is made up of FPN [68]. It is a lightweight network following the encoder which incorporates features from multiple scales of the encoder to create the final dense feature representation. This idea is similar to the creation of a hypercolumn feature map. FPN builds the final representation from features at 1/4, 1/8, 1/16 and 1/32 scales, using upsampling blocks as proposed in [81]. The final dense representation has a feature dimension of 64 and spatial downscaling of 1/4. The feature projection head is composed of 2 linear layers with BatchNorm and ReLU.

We use BYOL for stage 1 training with a batch size of 256 for 200 epochs using the SGD optimizer. The learning rate is set to 3×10^{-2} with a cosine decay for stage 1 training. For stage 2, we train with a batch size of 256 for 20 epochs on the CelebA dataset. We set the temperature τ to be 0.05. For a fair comparison we train the supervised regressors with frozen feature extractor as proposed in [22]. The regressor initially comprises of 50 filters (to keep evaluations consistent with [22, 104]) of dimension $1 \times 1 \times K$ which transforms the input feature maps to heatmaps of intermediate virtual keypoints. These heatmaps are converted to $2K$ x-

Table 3.1: **Landmark matching** performance comparison against prior art on MAFL dataset. The error is reported as a percentage of inter-ocular distance.

Method	Feat. dim.	Same	Different
DVE [104]	64	0.92	**2.38**
CL [22]	64	0.92	2.62
BYOL + NMF	64	0.84	5.74
LEAD (ours)	64	**0.51**	2.60
CL [22]	256	0.71	**2.06**
BYOL + NMF	256	0.91	4.26
LEAD (ours)	256	**0.48**	2.50
CL [22]	3840	0.73	6.16
LEAD (ours)	3840	**0.49**	**3.06**

y pairs using a $softargmax$ layer, which are further linearly regressed to estimate manually annotated landmarks. Here K represents the number of annotated keypoints in the dataset. Following DVE, we resize the input image to (136×136) and then take a (96×96) central crop for performing the evaluations. For stage 1 training we take two (96×96) sized random crops. We perform all of our experiments on 2 Tesla V100 GPUs.

Evaluation: Following prior works, we use percentage of inter-ocular distance (IOD) as the error. We evaluate on two tasks, landmark matching and landmark regression. We describe each of the evaluation tasks next.

3.3.1 Landmark Matching

In the landmark matching task, we are given two images. One is a reference image for which the landmarks are known and the other is a query image, for which the landmarks are to be predicted. Prediction is done by choosing the feature descriptor of a landmark in the reference image, and finding the location of the most similar feature descriptor to it in the feature map of the query image using cosine similarity. In line with DVE [104], we evaluate on a dataset consisting of 500 same identity and 500 different identity pairs taken from MAFL. Qualitative results of matching are shown in Fig. 3.4, while quantitative results are presented in Table 3.1. Also shown in Table 3.1 is the Non-negative Matrix Factorization (NMF [65], which gives low-rank approximation of non-negative matrix) baseline, wherein we apply NMF over the learned hypercolumn thereby showing that our dimensionality reduction objective is superior to naively applying NMF over the learned hypercolumn. Similar to the trends from correspondence matching using hypercolumn in Fig. 3.3, the final dense model with 64 dimensional features is

Table 3.2: **Landmark regression** performance comparison against prior art. The error is reported as a percentage of inter-ocular distance. We achieve state-of-the-art result on the challenging AFLW datasets with ∼10% relative gain, while obtaining competitive results on MAFL and 300W.

Method	Unsupervised	MAFL	AFLW_M	AFLW_R	300W
TCDCN [132]	✘	7.95	7.65	-	5.54
RAR [119]	✘	-	7.23	-	4.94
MTCNN [131, 130]	✘	5.39	6.90	-	-
Wing Loss [29]	✘	-	-	-	4.04
Dense objective based					
Sparse [106]	✓	6.67	10.53	-	7.97
Structural Repr. [130]	✓	3.15	-	6.58	-
FAb-Net [116]	✓	3.44	-	-	5.71
Def. AE [95]	✓	5.45	-	-	-
Cond. Im. Gen [49]	✓	2.86	-	6.31	-
Int. KP. [50]	✓	-	-	-	5.12
Dense3D [105]	✓	4.02	10.99	10.14	8.23
DVE SmallNet [104]	✓	3.42	8.60	7.79	5.75
DVE Hourglass [104]	✓	2.86	7.53	6.54	**4.65**
Global Objective based					
ContrastLandmarks [22]	✓	<u>2.44</u>	6.99	6.27	5.22
LEAD (ours)	✓	**2.39**	**6.23**	**5.65**	<u>4.66</u>

able to meaningfully match the landmarks from reference image to query image. This is verified across a head rotation ranging from left-facing to frontal faces and right-facing images. The matching is consistent across genders, showing no bias for any gender.

3.3.2 Landmark Regression

In the task of landmark regression, a lightweight regressor is trained on top of the features extracted by the pretrained network. This is done using supervised learning on the evaluation dataset. We report the inter-ocular distance on landmark regression in Table 3.2. Our model trained using the BYOL objective achieves results which are ∼10% better than the prior-art on a relative scale, on 2 out of 4 evaluation datasets, while maintaining a competitive performance on the 300-W dataset. Regression performance is qualitatively verified in the Fig. 3.7.

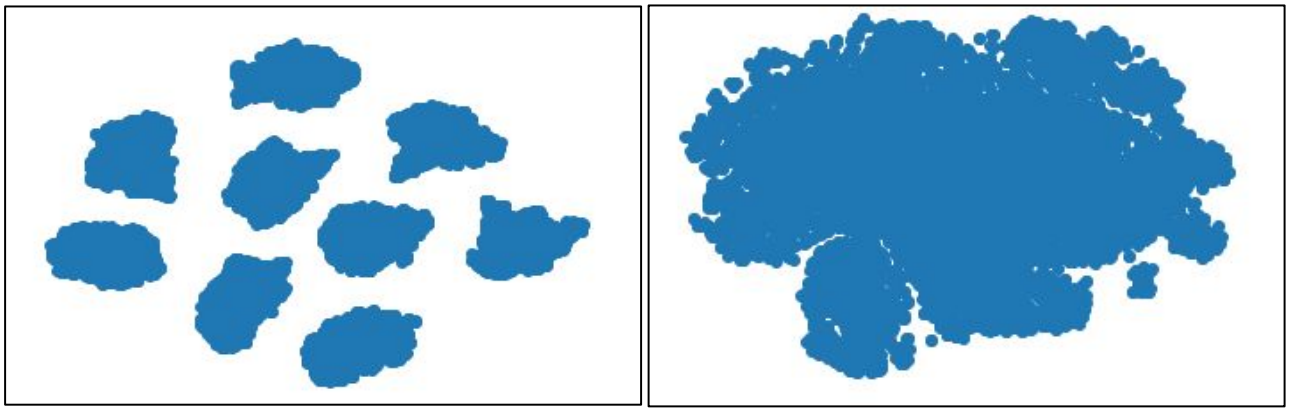


Figure 3.5: **t-SNE plots of output feature maps. Left:** LEAD stage 1 features **Right:** CL stage 1 features

3.3.3 Interpretability

We perform t-SNE clustering on MAFL test split, which contains 1000 images. We observe that the t-SNE embeddings obtained from our model trained with BYOL objective are interpretable. It divides the face spatially into 9 parts, where each clusters corresponds to one of the 9 parts. t-SNE clustering is visualized in Fig. 3.5 and interpretability of the clusters is verified in Fig. 3.6. We also compare our t-SNE plots against that of CL [22], wherein we see that CL embeddings are not well clustered when compared to LEAD which shows distinct clusters.

3.3.4 Ablation Studies

We ablate LEAD on factors like feature dimension, contribution from each stage, projection head, degree of annotation availability, and sensitivity to scale variations.

Feature Dimensions. Feature dimension plays a significant role in the landmark regression task. Since the regressor takes features as input, its capacity depends on the dimensions of the feature, i.e. a higher dimensional feature implies that the regressor has more capacity to learn, resulting in better predictions. Our experiments in Table 3.4 indicate a superior performance on the challenging AFLW$_M$ dataset, while achieving competitive performance on MAFL and AFLW$_R$. Surprisingly, we find a large deviation in performance trends on the 300W dataset compared to the results obtained using hypercolumn feature maps (ref. Tab. 3.2) as guidance for the compact feature maps.

How much does stage 2 objective contribute? To answer this question, we run experiments on 2 different pretraining (stage 1) objectives, followed by 2 different dimensionality reduction (stage 2) objectives. To compare directly, we take CL's [22] pretraining and dimensionality reduction objectives and our objectives for the same. We keep the architectures same

Table 3.3: **Effect of projection head on landmark matching**. Projection head affects the matching on different identity. On increasing the dimension of the projection head's output, improvement is observed. Further gains are observed on increasing the final representation's (Φ^D's output) dimension.

Feat. dim.	Proj. dim.	Same	Different
64	✘	**0.48**	2.79
64	64	0.51	2.64
64	256	0.51	**2.60**
128	256	0.47	2.58
256	256	**0.48**	**2.50**

Table 3.4: Effect of feature dimension on landmark regression task

Method	Feat. dim.	MAFL	AFLW_M	AFLW_R	300W
DVE [104]	64	2.86	7.53	6.54	**4.65**
CL [22]	64	**2.77**	7.21	**6.22**	5.19
LEAD (ours)	64	2.93	**6.61**	6.32	5.32
CL [22]	128	**2.71**	7.14	**6.14**	**5.09**
LEAD (ours)	128	2.91	**6.60**	6.21	5.41
CL [22]	256	**2.64**	7.17	6.14	**4.99**
LEAD (ours)	256	2.87	**6.51**	**6.12**	5.37

Table 3.5: **Number of annotations:** LEAD consistently produces the lowest inter-ocular distance under the presence of different levels of annotations on the AFLW_M dataset. The relative improvement is as high as **45%** over previous best (in case of '5 annotations' training setting)

Method	Feat. dim.	Number of annotations					
		1	5	10	20	50	100
DVE [104]	64	**14.23 ± 1.45**	**12.04 ± 2.03**	12.25 ± 2.42	11.46 ± 0.83	12.76 ± 0.53	11.88 ± 0.16
CL [22]	64	24.87 ± 2.67	15.15 ± 0.53	13.62 ± 1.08	11.77 ± 0.68	11.57 ± 0.03	10.06 ± 0.45
LEAD (Ours)	64	21.8 ± 2.54	13.34 ± 0.43	**11.50 ± 0.34**	**10.13 ± 0.45**	**9.29 ± 0.45**	**9.11 ± 0.25**
CL [22]	128	27.31 ± 1.39	18.66 ± 4.59	13.39 ± 0.30	11.77 ± 0.85	10.25 ± 0.22	9.46 ± 0.05
LEAD (ours)	128	**21.20 ± 1.67**	**13.22 ± 1.43**	**10.83 ± 0.65**	**9.69 ± 0.41**	**8.89 ± 0.2**	**8.83 ± 0.33**
CL [22]	256	28.00 ± 1.39	15.85 ± 0.86	12.98 ± 0.16	11.18 ± 0.19	9.56 ± 0.44	9.30 ± 0.20
LEAD (ours)	256	**21.39 ± 0.74**	**12.38 ± 1.28**	**11.01 ± 0.48**	**10.06 ± 0.59**	**8.51 ± 0.09**	**8.56 ± 0.21**
CL [22]	3840	42.69 ± 5.10	25.74 ± 2.33	17.61 ± 0.75	13.35 ± 0.33	10.67 ± 0.35	9.24 ± 0.35
LEAD (ours)	3840	**24.41 ± 1.38**	**14.11 ± 1.30**	**11.45 ± 0.88**	**10.21 ± 0.44**	**8.43 ± 0.25**	**8.09 ± 0.28**

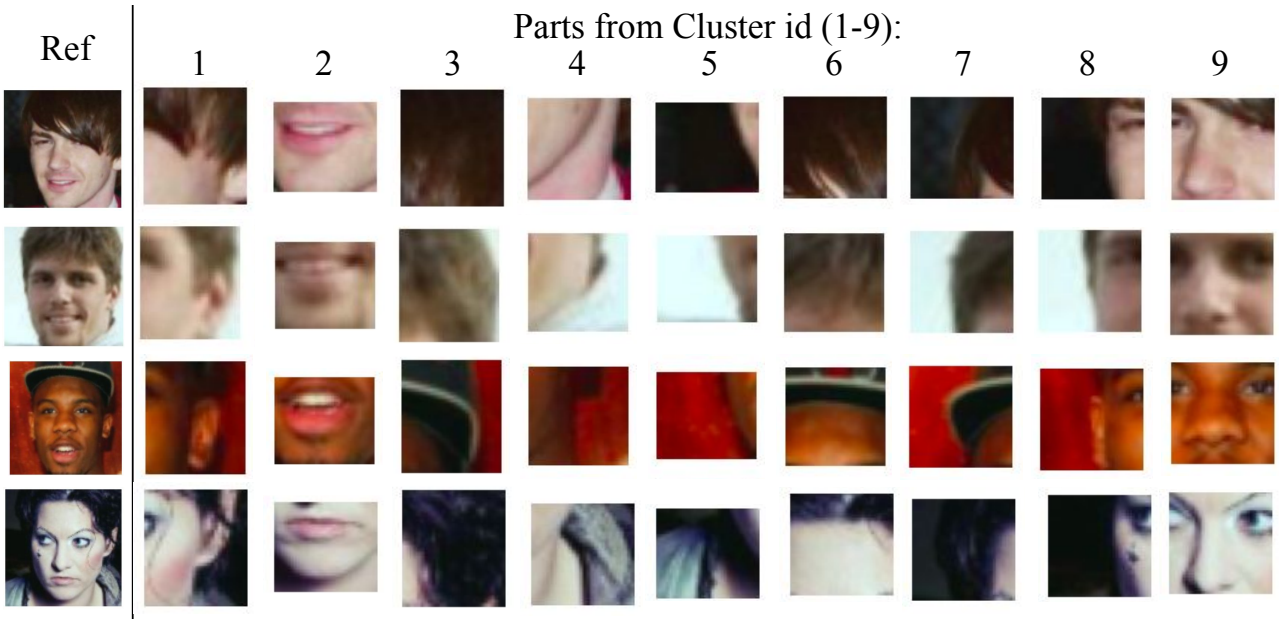


Figure 3.6: **t-SNE embeddings tend to cluster part-wise**. The 9 parts (along row) shown for each reference figure (along column) here belong to the 9 clusters in Fig. 3.5. Each cluster denotes a semantic part of the face.

as LEAD and only vary the objective function for a fair comparison. We report our findings in Table 3.6. Irrespective of the stage 1 training, LEAD's dimensionality reduction procedure improves the IOD.

Is the projection head necessary in stage 2? Necessity of the projection head in self-supervised learning has been empirically shown to lead to meaningful representations [16]. We use it in our stage 1 training. However in stage 2, where we aim to get higher resolution feature maps as output, is the projection head still required? We use a projection head g_θ^D on the final feature map as given by Φ^D to apply the loss during training. Eventually the g_θ^D is discarded

Table 3.6: **Dimensionality reduction objective.** LEAD's proposed dimensionality reduction objective significantly improves the performance irrespective of the global representation learning objective. Results are reported on AFLW$_M$ dataset.

Global Rep. Obj. (Stage 1)	**Dim. Red. Obj. (Stage 2)**	**Feat. dim.**		
		64	**128**	**256**
CL	CL	7.86	7.81	7.31
CL	LEAD	**6.66**	**6.58**	**6.69**
LEAD	CL	7.89	7.86	7.41
LEAD	LEAD	**6.61**	**6.60**	**6.51**

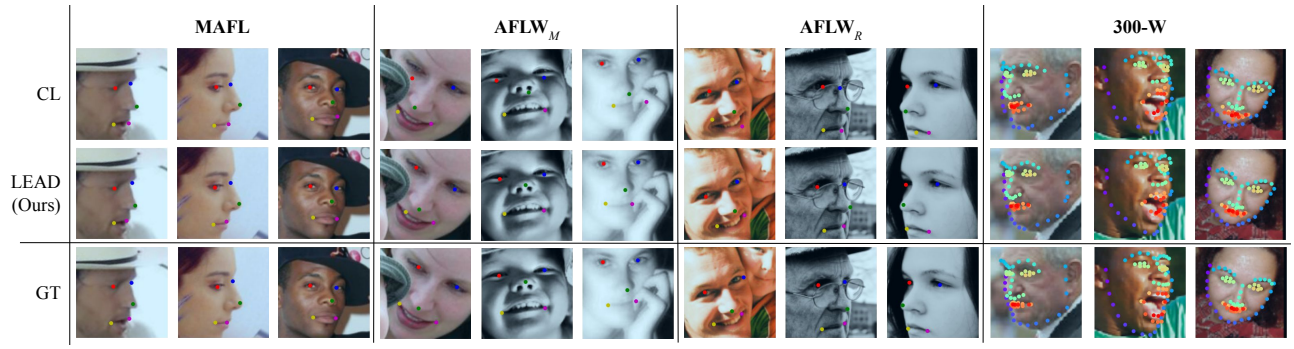


Figure 3.7: **Landmark regression:** We observe that features generated by pretraining using LEAD can easily be used to train a lightweight regressor to predict landmarks with high precision. Furthermore, the model is robust to aspects such as face orientation, lighting and minor occlusions.

and only Φ^D is utilised. Here, we ablate on the performance shown by Φ^D in the absence of projection head as well as the on the output dimension of the g_θ^D. Since we discard g_θ^D, we are allowed to keep its output's dimension as high as required. In our ablation (ref. Table 3.3), it is observed that for landmark matching on the same identity, there are marginal changes upon having g_θ^D as well as varying its output dimension. But the projection head emerges as a distinguishing component in case of matching on different identity. Consistent improvements are observed on increasing the feature dimension of the projection head. It can be seen that this leads to slight degradation of performance on the same identity. We also observe the effect of increasing the feature dimension by keeping the projection dimension fixed where we note a further improvement on matching.

How sensitive is it to the alignment and scale variations? At inference stage, the landmark regressor can encounter images which may have different alignments or scales when compared to the data it was trained on. To check the sensitivity of LEAD to these changes we use features from CelebA trained LEAD to train a landmark regressor on an unaligned-MAFL dataset. We create this dataset by taking images from MAFL subset of CelebA-in-the-wild [70] dataset cropped by the bounding box annotations. Furthermore, before taking a crop, we also randomly scale up the side length of the bounding box a factor uniformly randomly sampled between 1-1.5×. This results in zooming out of the image (ref. Fig. 3.11). We refer to this factor as "Zoom-out factor" We evaluate the regressor on the test split which is created by scaling up the side length of the bounding box by a zoom-out factor of 1-2× before cropping. We use 64d feature for this experiment. In Fig. 3.9, we observe that across the range of evaluated scales, LEAD outperforms CL [22]*. The gap between the two methods widens for larger zoom-out

*Same training and evaluation protocol was followed for both.

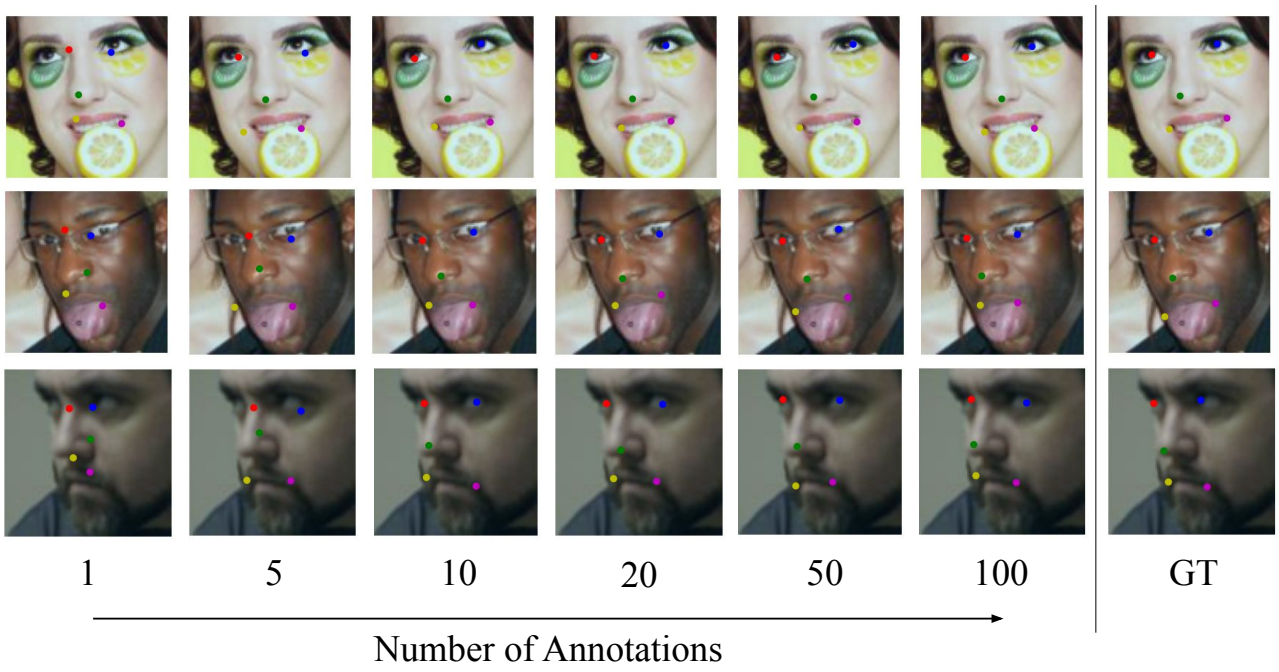


Figure 3.8: **Number of annotations.** Landmark prediction under different number of annotated images used for supervised training (mentioned below every column) on AFLW$_M$.

factors, which are unseen during training. We visualize the landmark regression against scale changes in Fig. 3.11.

How many annotated images are required for supervised training during evaluation? Since the evaluation of our method depends on the annotated samples, we run an ablation on the number of annotations required. We report the quantitative results in the Table 3.5, along with qualitative annotation-wise comparisons in Fig. 3.8. We test by varying the number of annotations to 1, 5, 10, 20, 50, and 100. We observe a consistent and significant gain in the performance with increasing number of annotations over the competent methods, a trend which even continues at different dimensions of features.

Table 3.7: Comparison of supervised training speeds at differ feature dimensions. Note that hypercolumn features (3840 feat. dim.) are 55× slower.

Feat. Dim.	FLOPS
3840	2.21
256	0.16
128	0.08
64	0.04

Complexity Analysis. The computational complexity during pretraining stages (Stage 1 and Stage 2) is $O((\frac{HW}{R^2})^2 \times C)$, similar to the prior art [104, 22]. Additionally, we have also reported the computational cost incurred by our method at the evaluation stage (training on a small annotated dataset) at different levels of feature dimensions in the Table 3.7, where we find

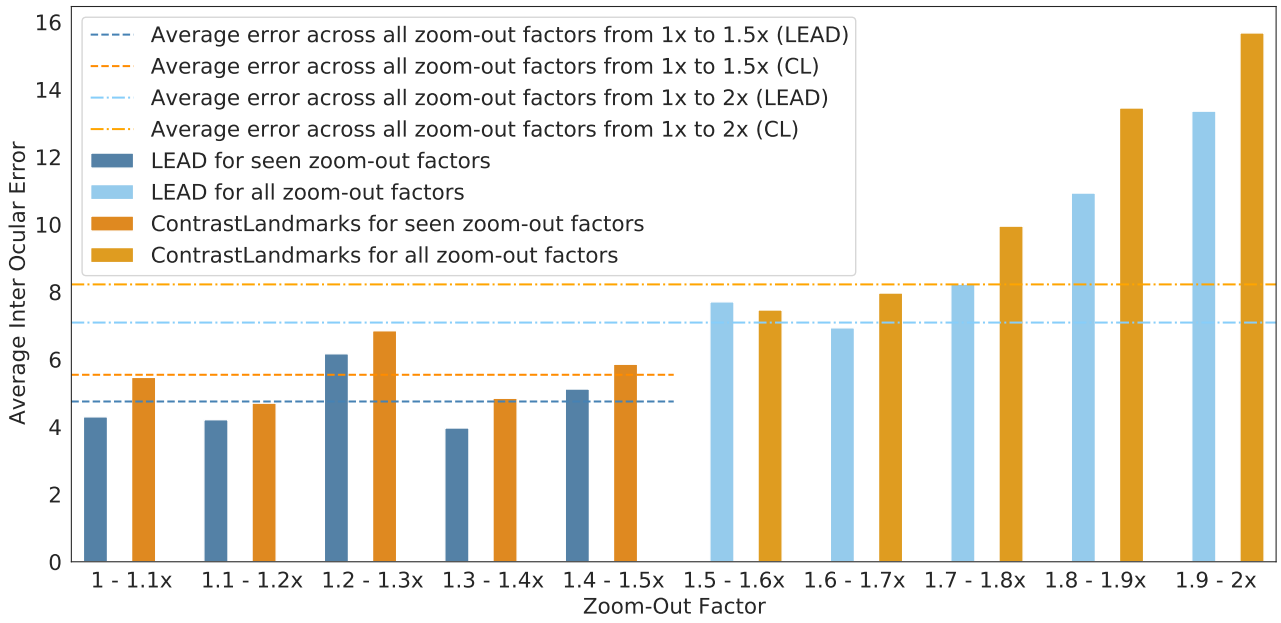


Figure 3.9: **Sensitivity to scale variations:** Sensitivity to seen scales (Zoom-out factor $\in$ 1-1.5x) vs unseen scales (Zoom-out factor $\in$ 1.5-2x) on unaligned-MAFL. LEAD performs better across scale changes, and is also less sensitive to *unseen* scales of face.

that training stage 2 at 64d is 55$\times$ faster compared to the hypercolumn which is 3840d.

3.4 Chapter Summary

In this work, we demonstrate the superiority of the LEAD framework to learn representation at instance level from a category specific dataset. We further utilize this prior to train a dense and compact representation of the image, guided by the correspondence matching property of the learnt representation. Our experiments demonstrate the superiority of the BYOL objective over contrastive tasks like MoCo on category specific data for landmark detection. Our proposed dimensionality reduction method improves the results on both feature extractors. A future research direction could be the usage of this correspondence matching property to learn a variety of dense prediction tasks.

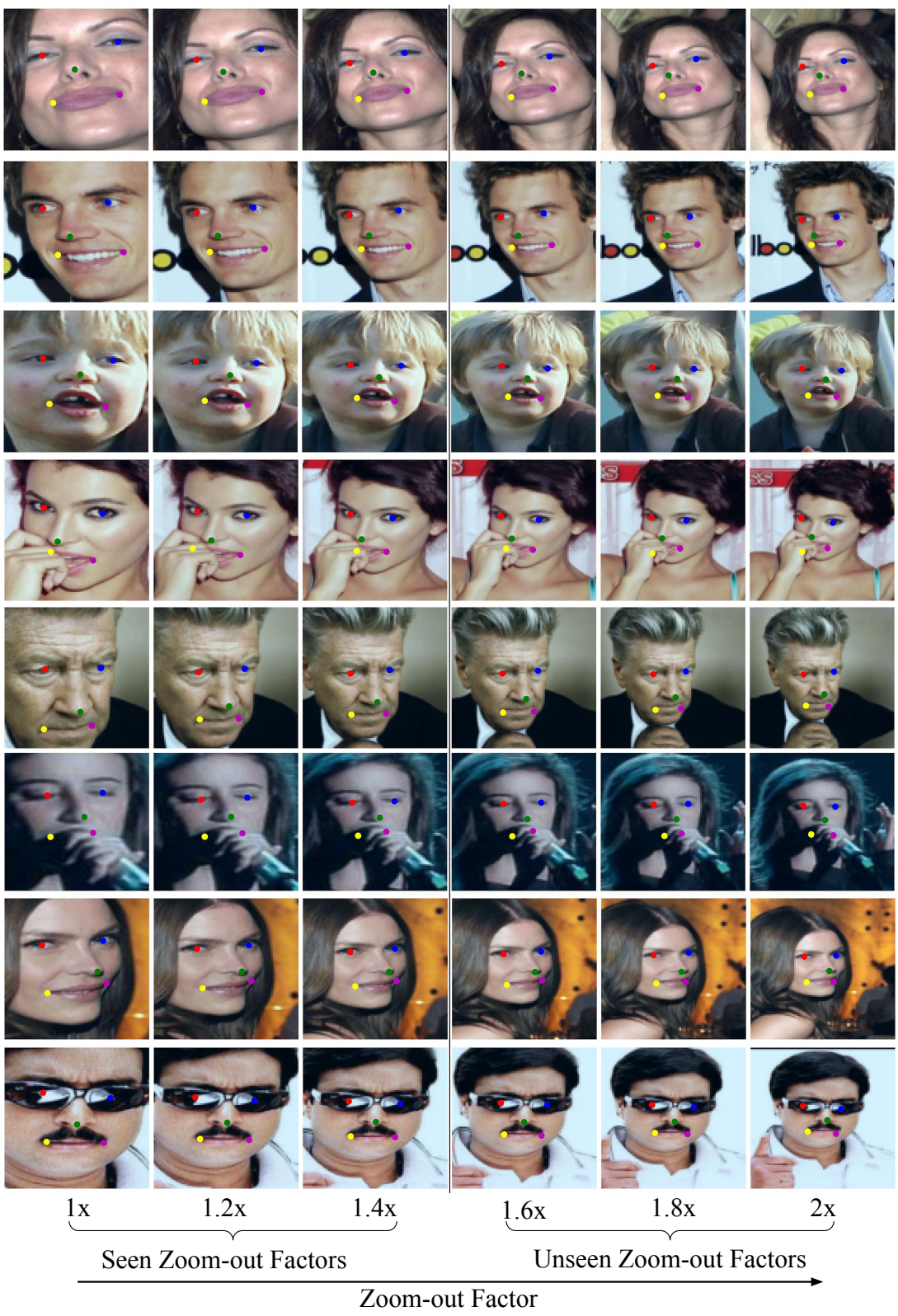


Figure 3.10: **Scale variation.** LEAD Landmark regression visualization across differently scaled (seen and unseen) images of unaligned-MAFL.

Figure 3.11: **Low lighting conditions.** LEAD Landmark regression performs well despite having low lighting conditions.

Chapter 4

Hierarchical Semantic Regularization of Latent Spaces in StyleGANs

4.1 Introduction

Having shown how pretrained features from self-supervised learning can benefit discriminative task of landmark estimation, we now investigate how can they be leveraged to improve the realism in the image synthesis models. Recent years have seen tremendous advances in Generative Adversarial Network (GAN) [33] architectures and their training methods to produce highly photorealistic images [10, 57]. Progress in the StyleGAN family of GAN architectures has shown promise by improving both the image quality, as well as the quality of latent space representations which enables controlled image generation. This is achieved by transforming an input noise space $\mathcal{Z}$ to a latent style space $\mathcal{W}$ which modulates a synthesis network at various levels of representation hierarchies to generate an image with that style. This enables generation of compelling synthetic images with novel styles as well as practically useful applications such as GAN-based image attribute editing, style mixing, etc. [93, 92, 48, 83, 84, 3]. Nonetheless, such networks still often produce unrealistic images (ref. Fig. 4.1). Note for examples, the (residual glass) artifact in the red circular inset produced by StyleGAN2-ADA (kindly zoom for details).

These quality issues in StyleGANs can have the following sources: (a) the hierarchical representation spaces in the synthesis network, (b) the latent style space, in particular the linearity and smoothness of such spaces, and (c) the functions used to transform the representation spaces in (a) using the corresponding hierarchical style vectors in (b). Our work seeks to address some of these issues.

We take inspiration from the recent advances in self-supervised and supervised learning [42,

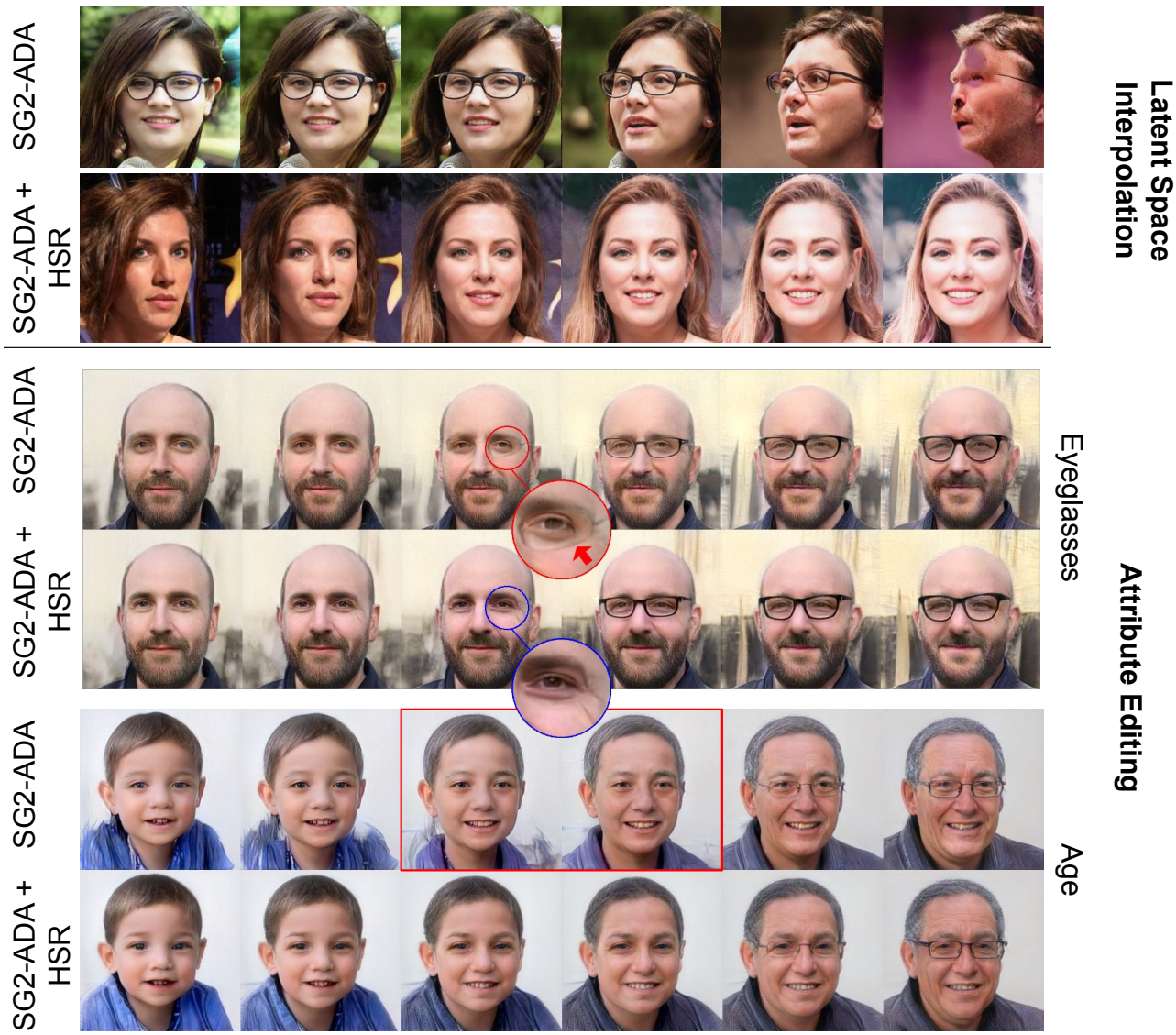


Figure 4.1: Hierarchical Semantic Regularizer (HSR) improves the latent space to semantic image mapping to produce more natural-looking images. **Top:** We show latent interpolation for images from bottom 10%-ile image pairs ranked by PPL, a metric to measure smoothness of latent space. **Bottom:** Latent space using HSR mitigates artefacts in images during attribute edit transition (as seen in binary attributed like "Eyeglasses") and can transition smoothly (young to old (SG2-ADA) vs. young to middle-age to old (SG2-ADA+HSR), in continuous attributes like "Age"). Zoom in to observe the effects.

12, 20, 98] which have allowed for the learning of powerful image representations translating into significant performance improves on image classification and other vision tasks [61, 98, 30]. Training on large datasets of natural images, like ImageNet [24], allows these techniques to learn hierarchically organized feature spaces capturing richer statistical patterns in natural images:

shallower layer capturing low-level image features and the deeper layers abstract features highly correlated with visual semantics. Such pre-trained representations can be harnessed to enhance the representational power of StyleGANs.

In fact, we demonstrate that transferring powerful pretrained representations mentioned above allow us to mitigate simultaneously the challenges associated with both the representation spaces in the synthesis network as well as the latent style spaces modulating these representations. To allow for such a transfer, we propose to use a regularization mechanism, called the Hierarchical Semantic Regularizer (HSR) which aligns the generator's features to those from an appropriate, state of the art, pretrained feature extractor at several corresponding scales (levels) of the generator network. The architecture is shown in Fig. 4.3.

Karras *et al.* [54] introduced the Perceptual Path Length (PPL) metric to measure the smoothness of mapping from a latent space to the output image and showed its correlation with the generated image quality. We demonstrate that HSR regularization in StyleGAN training leads to $\sim 15\%$ relative improvement in PPL over StyleGAN2, leading to more realistic interpolations. Please refer to the circular insets in Fig. 4.1.

A power approach for controlled synthesis of novel images is via linear (convex) interpolation between attributes* corresponding to real images. Applications such as image editing utilize such capabilities under the presumption that style spaces are both linear as well as decorrelated allowing for desired controlled edits. Since, PPL does not measure linearity, we propose a novel metric, Attribute Linearity Score (ALS), to measure linearity in the attibute space. We demonstrate that HSR simultaneously improves linearity leading to smoother edits with significantly reduced editing artifacts (Fig. 4.1). A mean relative improvement of 15.5% over StyleGAN2-ADA is achieved on the ALS metric.

Our contributions are:

- A novel Hierarchical Semantic Regularizer (HSR) improving the generation of natural-looking synthetic images from StyleGANs. HSR is presented in (Sec. 4.2 with an analysis of design choices 4.2.3).
- Extensive bench-marking of improvements by HSR regularization on popular datasets, especially when utilizing linear interpolations for attribute editing (Sec. 4.3).
- Since linearity of the latent attribute space is very important for performing controlled edits, we propose a new metric, Attribute Linearity Score (ALS), in (Sec. 4.3.3) and demonstrate improved linearity over the baselines.

*We use style and attributes interchangeably.

4.2 Approach

In this section, we first describe the objective of GAN framework, properties of StyleGAN, and its evaluation in Sec. 4.2.1. Then, we describe Hierarchical Semantic Regularizer (HSR) (Sec. 4.2.2) and discuss its design in Sec. 4.2.3.

4.2.1 Preliminaries

Generative Adversarial Networks. GAN involves two competing networks, namely a Generator G and a Discriminator D. Taking a noise $\mathbf{z}$ sampled from a distribution P_z as input, G generates an image $G(\mathbf{z}) \in \mathbb{R}^{3\times H\times W}$. Whereas, D takes an input image $x \in \mathbb{R}^{3\times H\times W}$, and tries to classify it as real or generated. The objective of G is to fool D into making it classify the generated image as a real one. Formally, the learning objective can be written as:

$$\begin{aligned} \max_D \mathcal{L}_D &= \mathop{\mathbb{E}}_{x\sim P_r}[log(D(x))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] \\ \min_G \mathcal{L}_G &= \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] \end{aligned} \tag{4.1}$$

StyleGAN. In StyleGAN, an architectural modification is introduced where $\mathbf{z}$ is transformed into a semantic latent space through a sequence of linear layers called Mapping Network G_m, before generating the image I through a Synthesis Network G_s as $I = G_s(\mathbf{w})$. Hence, $G = G_s \circ G_m$. The space learnt by G_m is known as $\mathcal{W}+$-space. It is observed that $\mathcal{W}+$ is more meaningful in terms of attributes learned from the training data as compared to noise space $\mathcal{Z}$. Several methods [93, 48, 92] propose ways to find attribute-specific directions in $\mathcal{W}+$ latent space.

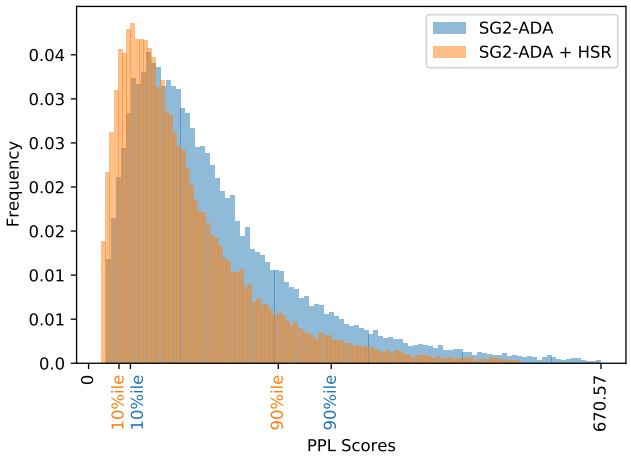


Figure 4.2: Distribution of PPL over 50k images from SG2-ADA and SG2-ADA+HSR. HSR improves the perceptual quality of top and bottom 10%-ile images, thus leading to more natural-looking images.

Perceptual Path Length. To measure the smoothness of the mapping from a latent space to the output image, Karras *et al.* [57] proposed Perceptual Path Length (PPL). The requirement for this metric arises due to generation of unnatural images by GAN despite having low FID [57]. PPL aims to quantify the smoothness of latent space to output space mapping by measuring average of LPIPS [129] dis-

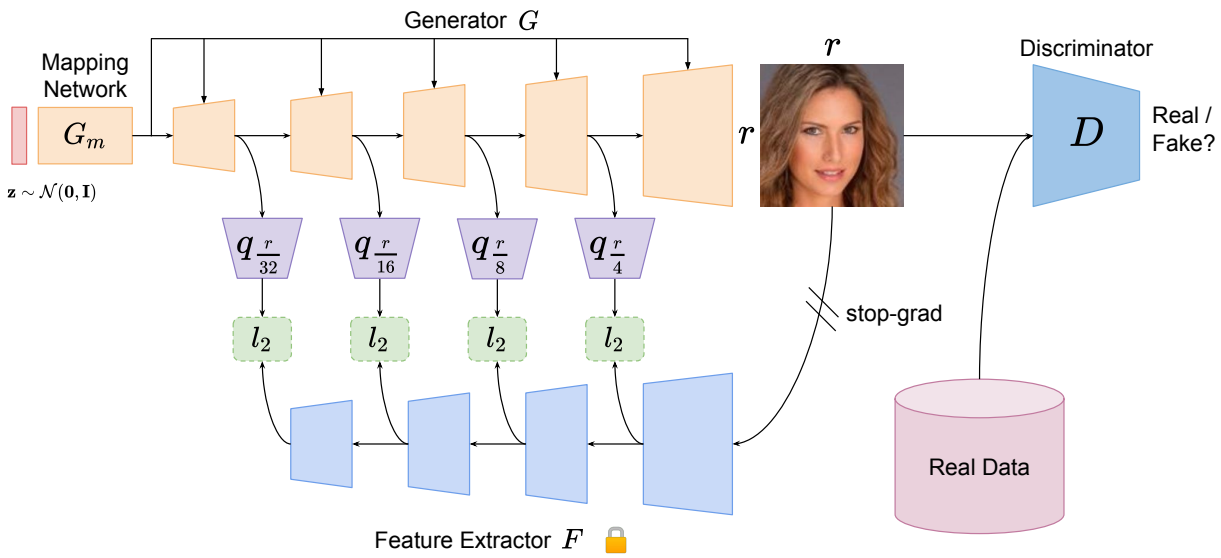


Figure 4.3: **Hierarchical Semantic Regularizer:** We use a pre-trained network to extract features at various resolution hierarchically. We then train linear predictors over generator features to predict the pre-trained features hierarchically. This transfers the semantic knowledge to generator feature space, making it's latent space meaningful, disentangled and editable.

tances between two generated images under small perturbations in the latent space. A smoother latent space should have lesser PPL when compared to an uneven latent space. It is shown [56] that PPL correlates well with image quality, *i.e.* good quality images pairs will have less PPL, while if any one of the image is of bad quality, the PPL would be high. The images are sampled randomly without any truncation trick [60, 10] to compute PPL. As observed in Fig. 4.5, the bottom 10%-ile by PPL (sorted in increasing order) among the generated images appear as out-of-distribution images. Hence, the mean PPL score can be used to quantify the extent of non-smooth regions of latent space which produce unnatural images. Hence, we will be using this metric as a primary metric for comparison of the smoothness of latent space learnt by the models.

4.2.2 Hierarchical Semantic Regularizer

Feature extractors of networks pretrained on large datasets (*e.g.* ImageNet etc.) of natural images using classification or self-supervised losses store strong priors about the data, that are organized hierarchically. Each level of hierarchy captures a different semantic feature of data. The statistics of wide variety of natural images are captured by these networks [41, 6, 109]. Due to the inherent differences in the nature of tasks, discriminative models capture different kinds of features compared to generative ones. Therefore, we seek to enrich the G's intermediate feature space with guidance from a pretrained feature extractor.

We first give a general idea of the proposed regularizer and then dive into various design choices made in it's formulation. Given an image $\mathbf{x}$ as input, the feature extractor F returns semantically meaningful features from it. We attempt to make the generator aware of this explicit semantic feature space. To this end, we freeze the feature extractor and treat it as a fixed function that maps from image space to a semantically meaningful feature space.

Given such a mapping of the generated image, we try to align the Generator's features of this image through a set of feature predictors. This alignment is inspired by BYOL [34]. As illustrated in Fig. 4.3, we attach a predictor branch q to the Generator G. The objective of q is to learn a mapping from generator's intermediate feature map $G^{\pi_G^i}(\mathbf{z})$ to pretrained feature extractor's intermediate feature $F^{\pi_F^i}(G(\mathbf{z}))$, where π_G in π_F denote the ordered set of layer numbers in the G and F at which we attach the predictors (ref. Eq. 4.2). We attach multiple such predictor networks q_i at different scales of generator.

$$\mathcal{L}_G = \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z}[log(1 - D(G(\mathbf{z})))] + \sum_{i=0}^{|\pi_G^i|} \mathop{\mathbb{E}}_{\mathbf{z}\sim P_z} \|q(G^{\pi_G^i}(\mathbf{z})) - F^{\pi_F^i}(G(\mathbf{z}))\|_2^2 \tag{4.2}$$

4.2.3 Design Choices

We analyse the effect of our Hierarchical Semantic Regularizer (HSR) against different design choices. For this purpose, we choose AnimalFace-Dog dataset which consists of 389 images. Since this is a low-shot dataset, we use StyleGAN2-ADA as our baseline. We perform all our experiments on 256×256 resolution.

What should be the choice of feature extractor? For this analysis, we choose 5 different feature extractors. We take combinations of CNN or transformer based networks trained using either self-supervised or supervised classification objective. We take ResNet-50 as the CNN backbone for both self-supervised (DINO) and supervised networks. For transformer-based networks, we use ViT-DINO and DeiT. Apart from trained networks, we also consider a randomly initialized ViT for baseline comparison.

We find that all pretrained feature extractors when used through HSR loss lead to introduction of meaningful semantic features in the intermediate latent spaces of the Generator. This is evidenced by reduction of PPL Score in Table 4.1, which signifies reduction in non-meaningful generations from the GAN. The reduction in PPL also implies improved disentanglement [57] and linearity in the $\mathcal{W}$ space of the Generator, which is a desired property for many applications. We get $\geq 6.2\%$ improvement in the PPL score when guided by these networks. ViT

Table 4.1: **Feature space ablation**: Ablating over different feature extractors for usage in HSR. Regularizing using ViT DINO's features gives best results.

	FID ↓	PPL↓
StyleGAN2-ADA	53.28	59.27
+ ViT (RandInit)	53.65	56.97
+ ResNet50 DINO [12]	54.33	55.6
+ DeiT [109]	53.22	54.71
+ ResNet50 [42]	52.88	52.23
+ ViT DINO [12]	**51.58**	**48.02**

Table 4.2: **Level of semantics**: A gradation in the improvement over the baseline is observed as we supervise from high-level semantics to low-level semantics. Best results are obtained when all the levels are supervised.

	FID↓	PPL↓
StyleGAN2-ADA	53.28	59.27
+ High-level ($\frac{r}{32}$, $\frac{r}{16}$)	53.15	57.73
+ Mid-level ($\frac{r}{16}$, $\frac{r}{8}$)	52.91	54.18
+ Low-level ($\frac{r}{8}$, $\frac{r}{4}$)	53.66	51.77
+ All levels	**51.58**	**48.02**

DINO's features stand apart, by improving the PPL score by 19% over the baseline. This is also supported by recent findings of Amir *et al.* [6], where they show several inherent properties of features from ViT-DINO, that are useful for computer vision tasks. With these results, we fix ViT DINO as the choice of the feature extractor for the rest of the experiments.

Which layers of Generator are more important? The StyleGAN generator G generates images using 7 synthesis blocks: starting from 4×4, up to full resolution of 256×256. Of these, we consider synthesis blocks having features of resolution 8, 16, 32, 64. This corresponds to scaling down of resolution r to $\frac{r}{32}$, $\frac{r}{16}$, $\frac{r}{8}$, and $\frac{r}{4}$. We choose these scales as it largely corresponds to the scales of downsampling by each block in SoTA CNN architectures [42, 98]. The first block of G (which have low resolution, but are responsible for high-level semantics) are supervised by the last block of the feature extractor (as they also are responsible for high-level semantics). Similarly, the next three blocks of G are supervised by the respective blocks of the feature extractor that bring out similar level of semantics.

To decide which layers contribute the most to the improvement in PPL, we divide the 4 blocks into 3 groups. The 3 groups specialize in high ($\frac{r}{32}$, $\frac{r}{16}$), mid ($\frac{r}{16}$, $\frac{r}{8}$), and low ($\frac{r}{8}$, $\frac{r}{4}$) level of semantics. We observe, in Table 4.2, that it is the supervision at low-level semantics which is most useful for the G. We observe a gradation in the improvement over the baseline, as high-level semantic supervision is least useful, followed by middle, and low. Overall, supervision at all levels turns out to cause the highest improvement.

Does Path Length Regularizer (PLR) complement HSR? Path Length Regularizer (PLR) was introduced in StyleGAN2 [57]. The intuition behind PLR is to promote fixed magnitude non-zero change in the resulting image when moving by a fixed step size in the $\mathcal{W}+$-space. As reported in Table 4.4, we find that HSR itself gives slightly better improvement than

Table 4.3: **Full data Results**: We report FID, Precision, Recall and PPL for different methods. With full data our method (SG2+HSR) produces better results across all the evaluation metrics.

FFHQ-140k	FID↓	Precision↑	Recall↑	PPL↓
SG2	3.92	**0.68**	0.45	175.09
+ HSR	**3.72**	**0.68**	**0.48**	**144.59**
LSUN-Church				
SG2	4.08	**0.60**	0.34	916.15
+ HSR	**3.82**	**0.60**	**0.41**	**678.55**

Table 4.4: **Performance wrt PLR.** PLR and HSR complement each other, while being equally effective individually.

PLR	HSR	FID↓	PPL↓
✘	✘	57.97	75.63
✓	✘	53.28	59.27
✘	✓	52.98	58.60
✓	✓	**51.58**	**48.02**

the PLR over the baseline. While the best effect is noted when both, PLR and HSR, are applied together. **Insight.** PLR's objective is to improve latent space smoothness, which leads to better PPL. Since PPL and image quality (natural-ness of image) are correlated, applying PLR improves the image quality. Whereas in HSR, we enforce the generator to predict in a feature space learnt from natural images using a pretrained feature extractor as prior. We observe that this objective, which targets bringing feature space of generator closer to a "natural" feature space also leads to improvement in the smoothness of latent space, as measured by PPL. This shows that image quality and latent space smoothness are complementary and related concepts. Therefore, optimizing for both gives better PPL score.

4.3 Experiments

In this section, we demonstrate the effectiveness of HSR experimentally. We first describe the experimental setup for all our experiments. Then, we evaluate the quantitative performance on several real-world datasets of varying sizes. Finally, we show improved linearity of latent space through attribute editing.

4.3.1 Experimental Setup

Datasets. We run our experiments on FFHQ [54] (70k images), LSUN-Church [124] (1.2M images), AnimalFace-Dog (389 images), AnimalFace-Cat [96] (160 images), and CUB200 [113] (12k images) datasets. We augment the datasets by taking the horizontal flip of every image, doubling the number of images in the original dataset. We resize the data to spatial dimensions of 256×256.

Implementation Details. We use StyleGAN2-ADA (SG2-ADA) as the baseline GAN, with its architecture for 256×256 images, with batch size of 16. Predictors q contain `Conv1x1-LeakyReLU-`

`Conv1x1`, with hidden dimension of 4096. We make use of 2 A6000 GPUs for training.

In order to leverage the rich feature space of pretrained networks using the generated images, we turn on the HSR regularizer after training the GAN for 500kimgs. By this point in the training, the GAN learns to generate images that start to look like the real images. We use the feature extractor of ViT-DINO [12] to extract features. We resize the image to 224×224 before feeding into ViT-DINO. We use the intermediate output (after disarding CLS token) of its 3, 6, 9, and 12th transformer blocks, to supervise the generator's output at the 64, 32, 16, and 8 resolution respectively to align the semantics at various hierarchical levels, since ViT-DINO has been to shown to have a high-to-low level semantics emerging in its stack of transformer blocks [6]. We resize the generator's intermediate outputs to 14×14 before applying the loss function (l_2).

4.3.2 Results

On standard full datasets of FFHQ and LSUN-Church, we compare over StyleGAN2 [57]. We also evaluate our method for limited data sizes. Traditionally, GANs have shown to perform poorly on smaller datasets, until recently several approaches [55, 52, 121] have been proposed which enables GANs to learn well on limited data. We observe that irrespective of dataset size, asking the generator to be predictive of semantic features of rich feature extractors via HSR improves the smoothness of the latent space, as it is evident by an average relative improvement in PPL scores of about 14.2% on average in Table 4.5, while that of 17.4% in case of full FFHQ and 25.93% in LSUN-Church(ref. Table 4.3). This is also evident qualitatively in Fig. 4.4 and 4.5, where we observe an improved latent-to-image mapping even in bottom 10%-ile images, when ranked by PPL scores. Thus, HSR raises the lower bound for the natural-ness of the images produced by a generator (also ref. Fig. 4.2).

Improvement in Worst Images. We have quantitatively shown that applying the HSR regularization improves the quality of worst images that the generator can produce. We also demonstrate this qualitatively in 2 ways. First, we compare the Mahalanobis distance between the generated images and the moments of the real data from a set of 5000 generated images. We present the results of 30 farthest images in Fig. 4.6. It can be seen that unnatural, non-face images are being generated by the baseline, which are virtually absent when HSR is applied. In the case of the LSUN-Church dataset, the images lack in structural aspects related to churches and shows presence of unnatural colors in the image. While after applying HSR, the images reproduce the structure faithfully, for e.g., in the edifices.

Secondly, we present the results of images sampled from the bottom 10% according to the PPL score. We present these results in Fig. 4.7. A similar trend is observed with the presence

Table 4.5: **Results on Limited Data** We present results on different limited data cases for FFHQ (left) dataset and on real-world datasets (right). We apply our regularizer on the strong baseline of StyleGAN2+ADA which is designed for limited data. We observe a significant decrease in PPL over baselines which implies a smooth, disentangled and meaningful latent space, while preserving photorealism (comparable FID).

Dataset	Method	FID↓	PPL↓	Dataset	Method	FID↓	PPL↓
FFHQ-1k	StyleGAN2-ADA	**19.14**	98.79	AnimalFace Dog	StyleGAN2-ADA	53.28	59.27
	+ HSR	21.76	**90.39**		+ HSR	**51.58**	**48.02**
FFHQ-2k	StyleGAN2-ADA	**14.74**	136.14	AnimalFace Cat	StyleGAN2-ADA	**39.50**	50.76
	+ HSR	15.53	**115.38**		+ HSR	40.25	**40.75**
FFHQ-10k	StyleGAN2-ADA	**7.16**	164.21	CUB	StyleGAN2-ADA	**5.78**	265.46
	+ HSR	8.08	**126.35**		+ HSR	6.15	**237.81**

of artefacts in and around the facial regions. In both cases (faces and churches), the artefacts are greatly reduced after the application of the HSR regularizer, making the images look more natural.

4.3.3 Analysis of Linearity of Latent Space

Motivation. Latent space of a pre-trained StyleGAN has meaningful directions embedded in it. Shen *et al*. [93] shows that $\mathcal{W}+$ latent space is disentangled with respect to image semantics and there exist linear directions $\mathbf{d}$ in this space that control specific semantic attributes in the generated images. This is an important property of the latent space which is commonly used in controlled image synthesis [84] and image editing [1], as it leads to smooth interpolation between any two generated images. Furthermore, it is observed that the magnitude of latent transformations linearly correlates with the magnitude of the attribute changes in generated images [136]. Although, multiple works [1, 48, 93, 118, 84] are built upon this property to generate desired image transformations, there is no established metric to evaluate the extent of this linear correlation in the latent space. To this end, we propose a new metric called *Attribute Linearity Score (ALS)* for quantifying this linear correlation between the extent of latent transformations and the attribute changes.

Attribute Linearity Score (ALS). Let the attribute strength be given by attribute score (logit value) from a pretrained attribute classifier C [54]. Consider two latent codes $\mathbf{w_0}$ and $\mathbf{w_1} \in \mathcal{W}+$ and their corresponding generated images $G(\mathbf{w_0})$ and $G(\mathbf{w_1})$ (using the generator G). Convex combinations of $\mathbf{w_0}$ and $\mathbf{w_1}$ generate interpolated latent codes $\mathbf{w_t}$ (Eq. 4.3) on the line segment joining the two latent codes $\mathbf{w_0}$ and $\mathbf{w_1}$. Let the corresponding generated images

Figure 4.4: Latent space interpolation of top 10-%ile images, ranked by PPL score. SG2-ADA images show traces of artifacts which are absent after applying HSR.

be denoted by $G(\mathbf{w_t})$. Linearity of the latent space with respect to the attribute strength C implies that the attribute score for the image $G(\mathbf{w_t})$ should be the same convex combination of the attribute strengths of $G(\mathbf{w_0})$ and $G(\mathbf{w_1})$ (Eq. 4.4). This expectation of linearity is supported by a (statistical) significant analysis in [136].

$$\mathbf{w_t} = \mathbf{w_0} + t * (\mathbf{w_1} - \mathbf{w_0}),\ t \in (0, 1) \tag{4.3}$$

$$C(G(\mathbf{w_t})) \approx C(G(\mathbf{w_0})) + t * (C(G(\mathbf{w_1})) - C(G(\mathbf{w_0}))),\ t \in (0, 1) \tag{4.4}$$

Consider the example shown in Fig. 4.9a, where we depict the transformation of the smile attribute. On the left, we show the plot of attribute scores with the interpolation parameter t using smile classifier C_s and on the right we show the image samples $G(\mathbf{w_t})$ for $t \in (0, 1)$. A model with a linear latent space structure should have this plot close to the "ideal" (shown in dotted) straight line between the two end points. Similar plots are shown for the "smile" and

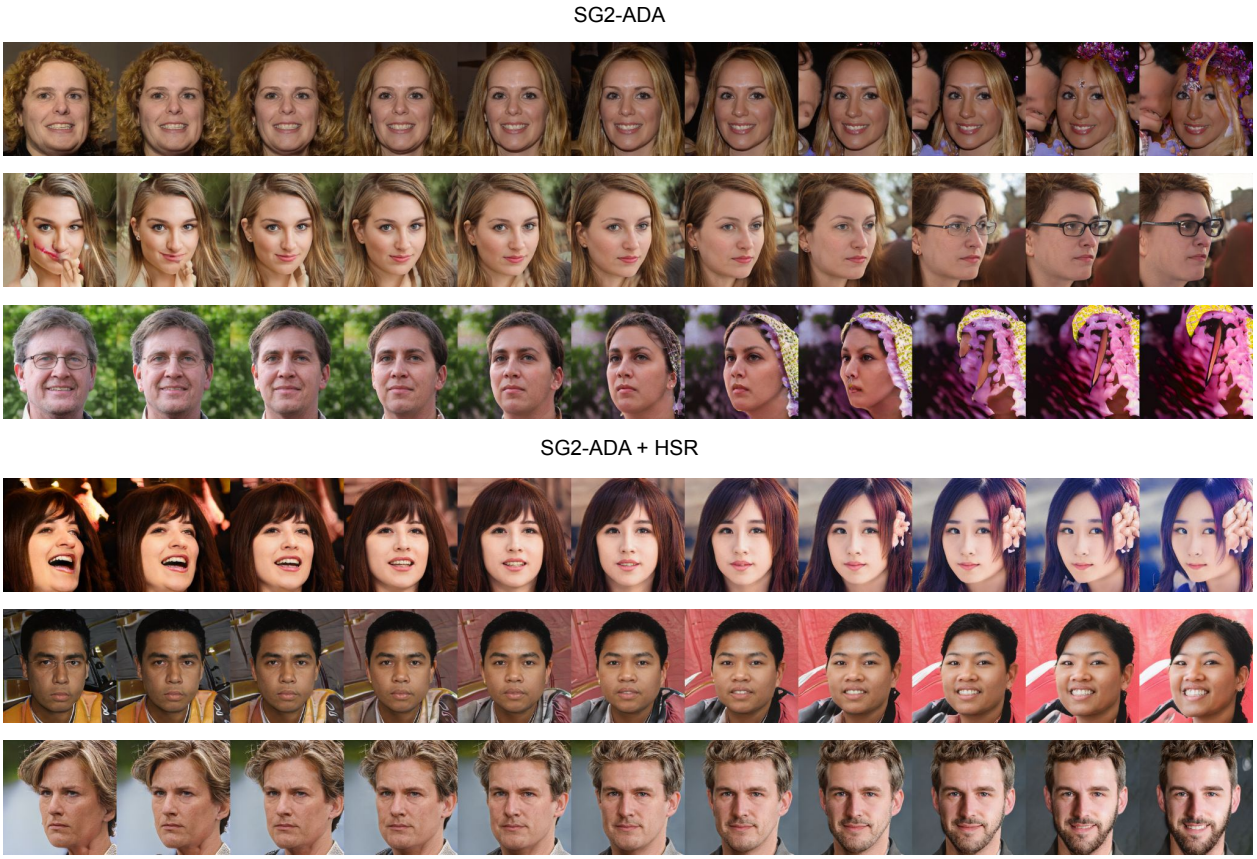


Figure 4.5: Latent space interpolation of bottom 10-%ile images, ranked by PPL score. SG2-ADA latent space accommodates more unnatural images, while leads to increase in PPL score. Upon adding HSR, the latent space maps to more natural face-like images.

"male" attributes in 4.9b. In both cases, we observe a significant departure from linearity.

The ALS score quantifies the deviation from the line segment defined in Eq. 4.4 using the mean squared error metric. To compute this, we first define a set of equally-spaced interpolation points $t \in \{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$. For each attribute $j \in \{1, \ldots, M\}$), the squared difference (Δ_{tj}) is computed using Eq. 4.5. The ALS score (Δ_T) is defined as the mean of Δ_{tj} over all M attributes and N interpolation points (*i.e.* $\Delta_T = \frac{1}{NM} \sum_{t=1}^{N} \sum_{j=1}^{M} \Delta_{tj}$).

$$\Delta_{tj} = ||C_j(G(\mathbf{w_t})) - C_j(G(\mathbf{w_0})) - t * (C_j(G(\mathbf{w_1})) - C_j(G(\mathbf{w_0})))||^2 \tag{4.5}$$

In the following sections, we first evaluate effect of linearity on appplying HSR, by measuring ALS. Then we show it's application in measuring edits in images. We use StyleGAN2-ADA model as the baseline trained on FFHQ-10k for results in the rest of this section.

ALS Evaluation. Our proposed HSR is able to provide a smooth structure to the latent

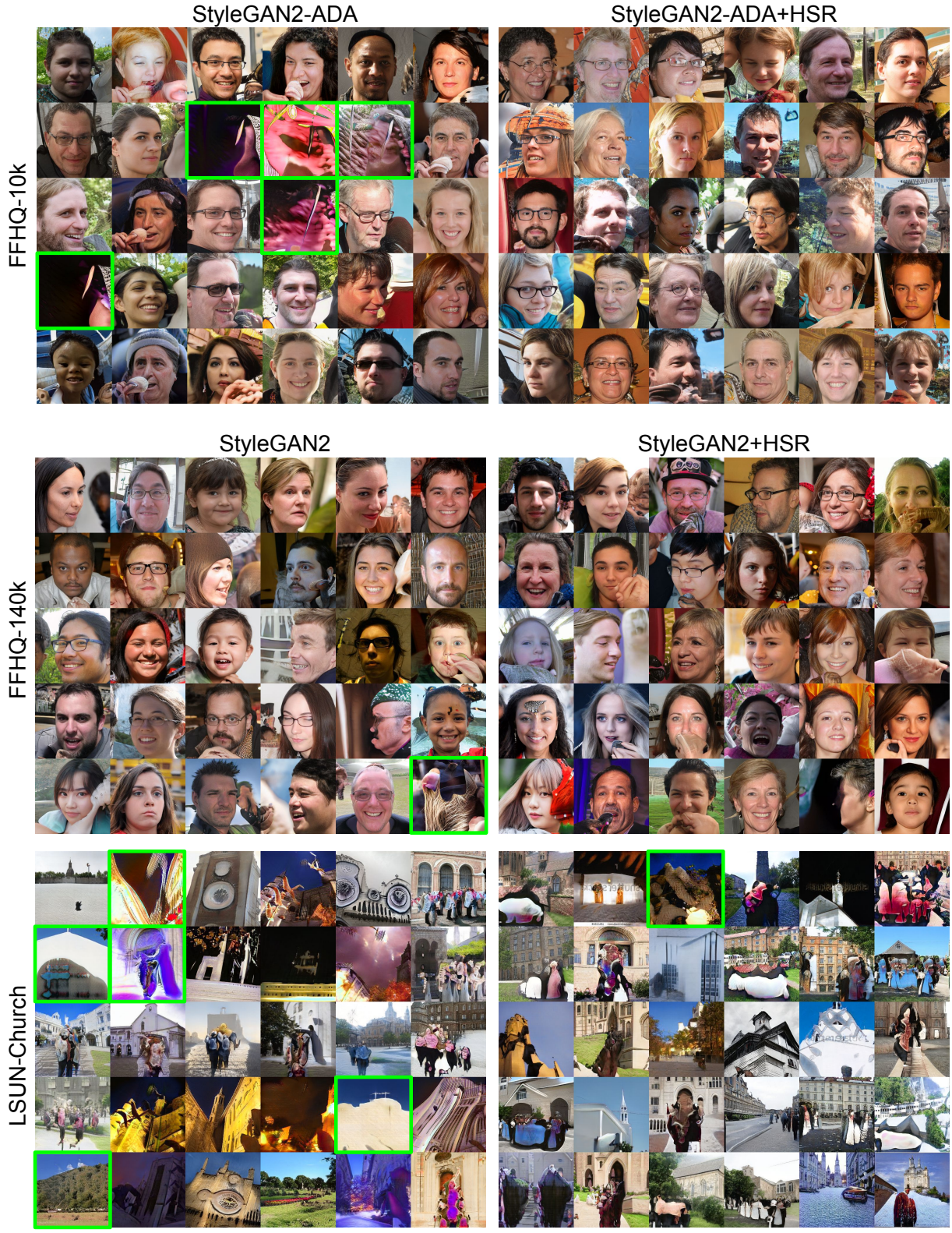


Figure 4.6: Worst 30 Images according to the Mahalanobis distance to Inception moments of respective datasets. Highlighted images show structural irregularities in the respective image category (face/church).

Figure 4.7: Worst Images according to the PPL scores. Highlighted images have high degree of artefacts.

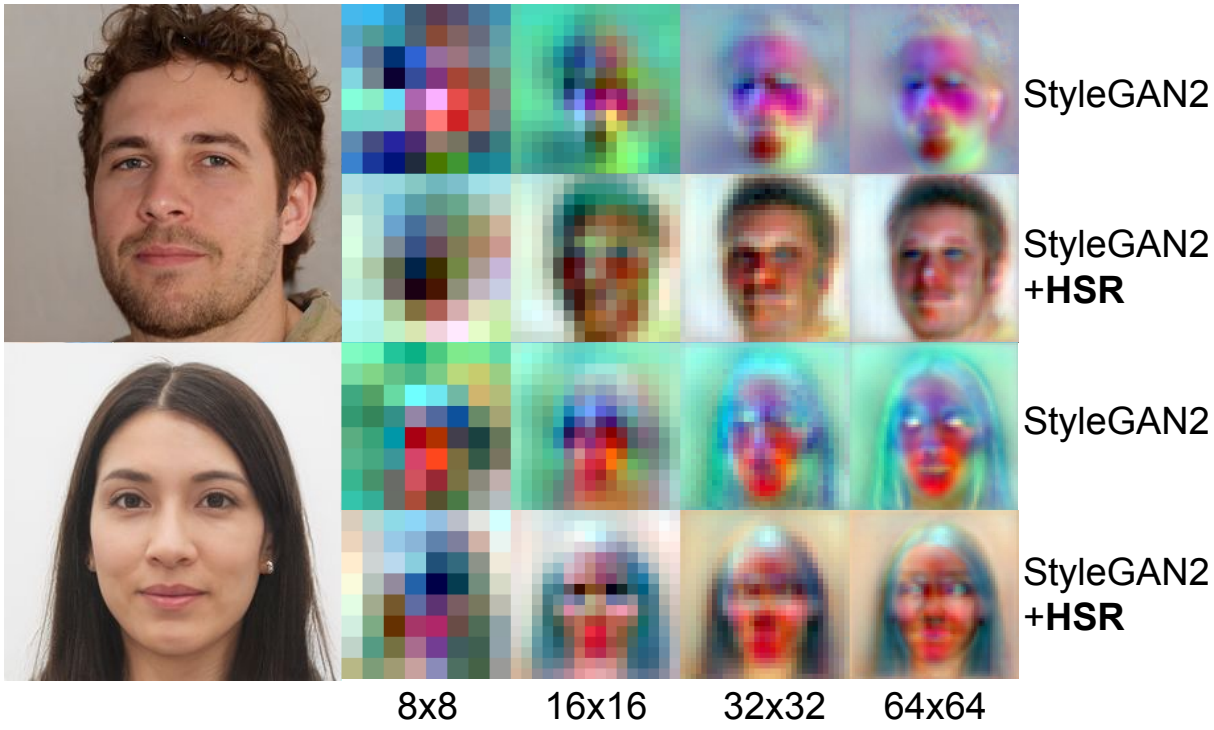


Figure 4.8: Comparison of Intermediate RGB outputs from the Generator. Upon adding HSR, the intermediate RGB outputs are more similar to final images in terms of color as well as structure.

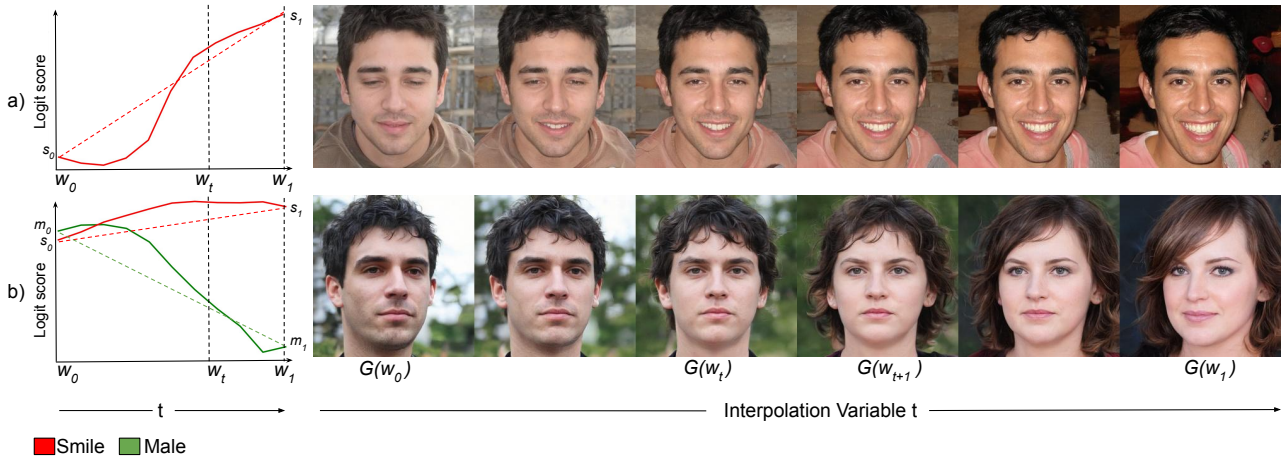
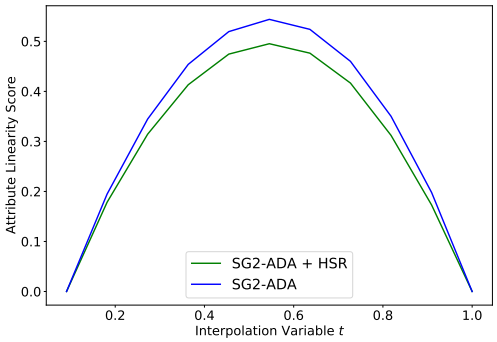


Figure 4.9: **Linearity of the latent space:** Here we show the transition images generated by the intermediate latent code $\mathbf{w_t}$ in the right and the corresponding attribute scores s_t for for smile (row 1 and 2) and m_t for male attribute (row 2). For brevity we have written $s_t = C_s(G(\mathbf{w_t}))$ and $m_t = C_m(G(\mathbf{w_t}))$.

Figure 4.10: ALS score comparison upon adding HSR. *(Right)*: Mean ALS computed for each value of the interpolation variable t. HSR is able to achieve a lower value of ALS supporting the linearity induced by ALS. *(Bottom)*: ALS score computed for all the face attributes separately.



	Gender	Smile	Age	Hair	Bangs	Beard	Mean
SG2-ADA	1.38	1.48	1.18	1.96	1.95	1.60	1.59
+HSR	**1.12**	**0.99**	**1.15**	**1.87**	**1.62**	**1.16**	**1.32**

space which is evident by the lower ALS scores of our model. To further analyse the structure of the latent space we perform latent space interpolations and generate a sequence of images. To quantitatively evaluate the interpolation results, we used the proposed ALS scores for the interpolations. The lower ALS score represent the latent space is well structured and the magnitude of the attributes are linearly correlated with the latent transformation. The ALS scores for our model and baseline model in Table. 4.10 for following set of popular attributes {gender, smile, age, hair, bangs, beard} [93, 92, 48]. Additionally, Fig. 4.10 (right) shows the variation of the mean attribute delta ($\Delta_{t,.}$) with the interpolation parameter t. We can observe that in the middle region $t \in [0.4, 0.8]$ the baseline model has high deviation from linear behaviour, which is significantly less in our HSR regularized model. This is also seen quantitatively through proposed ALS-attribute score, in which our model outperforms baseline by **15%** of relative improvement. We can observe that the interpolations generated using the HSR results in smooth transitions and has high visual quality throughout the interpolation. The StyleGAN2+ADA model without HSR has sudden transitions in between and has some artifacts present (ref. Fig. 4.1).

Editability. The semantically rich structure of the latent space is widely used for performing semantic edits on the generated images [93, 1, 118, 84, 125, 4]. For instance, if we have to add the attribute smile to the generated face image, one can edit the latent code as $\mathbf{w_{edit}} = \mathbf{w} + \alpha\mathbf{d}$ where α is edit strength and $\mathbf{d}$ is the direction for the smile attribute edit operation. However, often, the attribute scores of the edits performed by such methods does not change linearly with the edit strength parameter α as observed in Fig. 4.11. To this end, we perform the following experiment: Given an input source image I_0, we first perform attribute edit on it using latent space transformation to obtain I_1 using an existing approach [93]. Then, we use the latent code optimization to find the corresponding latent codes $\mathbf{w_0}$ and $\mathbf{w_1}$ in the latent space. Finally, we followed the same approach explained in Sec. 4.3.3 to generate intermediate images I_t using w_t

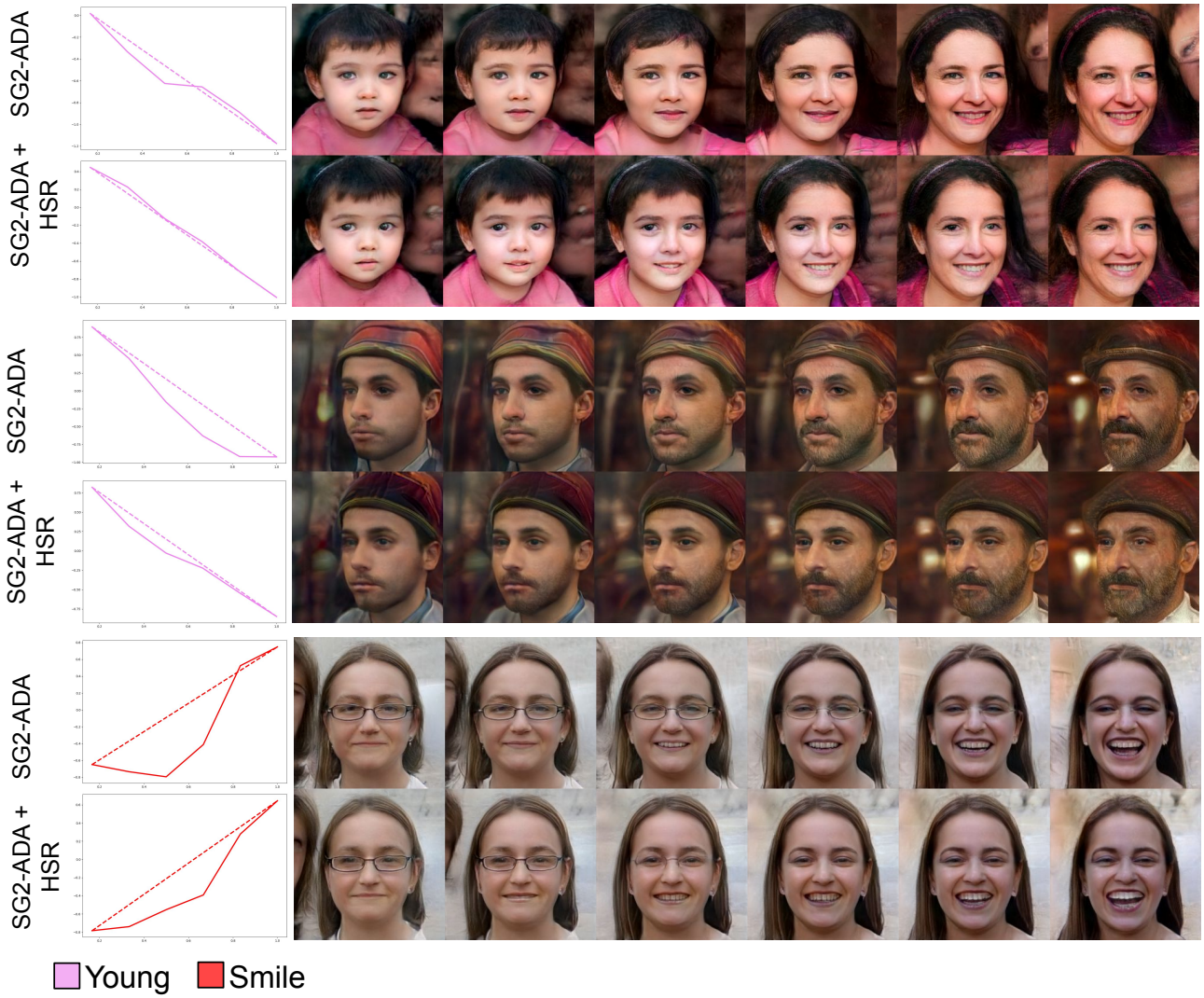


Figure 4.11: **Applying HSR improves the linearity of change in attributes.** Here we show improved linearity for "Young" and "Smile" attributes. Plots show attribute score on Y-axis, interpolation variable t on X-axis.

for $t\{\in 0, \frac{1}{N}, \frac{2}{N}, ...1\}$. The results of the interpolation for edits are shown in Fig. 4.11. We compared StyleGAN2-ADA with and without HSR in this experiment. One can observe that in all the cases, adding HSR resulted in added linearity in the attribute scores plots. This property is highly desired in editing methods as it provides a fine-grained control over the attributes in the generated images. Also, observe that both the models evaluated are following the linear line closely in the first two examples. This suggests that the transitions along the age attribute is much more interpretable as it follows linearity. In all the three examples the model with HSR is able to approximate the linear line the better than the baseline without HSR. From the

images, we can visually observe that the interpolations produced smooth transitions and their is no sudden jump in the attribute when using HSR. Also note that, the first and last images from both the models do not match "pixel perfectly", as they is generated by optimization of latent code by different models (with and without HSR).

4.4 Chapter Summary

We proposed a novel, hierarchical semantic regularizer called HSR which allows us to regularize the latent representations in StyleGANs by aligning them to powerful ones learnt by state of the art classifiers trained on large datasets. HSR is shown to significantly improve the quality of the generated images, especially those created via linear interpolation between attributes corresponding to real images. It further has a desirable property that the latent attribute space becomes more linear. To measure linearity, a novel metric *Attribute Linearity Score (ALS)* was introduced. Copious experiments on standard benchmarks validate the benefits of HSR and demonstrate statistically significant improvement in the quality of synthesized images. This leads us to some interesting avenues for the future work: Enforcing structural priors (*e.g.* linear) in the latent space manifold while training a GAN, which can lead to easier and fine-grained attribute editing.

Chapter 5

Conclusion

In this thesis, we explored properties of self-supervised pretraining to solve the problems of landmark estimation and improving image generation in GANs.

For the task of landmark estimation, we rely on the emergence of equivariance in the SSL pretraining methods. We find that SSL methods which do not use negative images have an emergent dense correspondence tracking ability. Such a setup saves the cost of compute, incurred by other SSL pretraining methods [19]. Having established this property, we propose a method to have a finer control over the feature dimensions of the output feature map of the pretraining network, thus enabling the training of the (few-shot) supervised training for landmark estimation with lesser compute cost. We demonstrate the superiority of our method on the 4 different challenging datasets. Our methods generalizes across scale variations of the face.

Next, we tackle the problem unnatural image generation from generative adversarial networks (GAN). For this task, we utilize the natural-image prior learnt by SSL. We propose a regularizer to align the intermediate feature spaces of the generator and the pretrained feature extractor at different levels of feature hierarchy. This not only improves the image generation, but also smoothens the latent space of the generator, which has applications in controlled attribute editing of the images. We show this by proposing a metric, Attribute Linearity Score (ALS) which measures the linearity of the latent space w.r.t. image attributes.

Future Directions. Having shown the applications of SSL pretrained networks to solve both discriminative and generative tasks, the next property to explore in SSL pretrained network is that of its relation with human visual system. This opens the doors to have them as proxy for humans to evaluate similarities between images which leads to design of newer evaluation metrics to analyze images. Current metrics to analyze generated images (for eg. FID) use

networks trained with object classification as task. This leads to task bias and to favor certain classes in the generated images. The task bias can be corrected by employing pretrained networks from SSL which follow generic objectives.

The recent advancements by the Diffusion Models (DM) spark up the question of combining HSR with Diffusion Models. While DM having diverse and realistic outputs, we have not yet seen unsupervised attribute specific interpretations out of its latent space. As compared to that, GAN has seen more progress on the nature of its latent space and how can it be manipulated to extract attribute-specific edits out of it [92, 84, 112]. Hence, it is an important open problem to investigate the latent space of large DMs, such as Stable Diffusion [90], to search for the attribute-specific interpretations. As there is "linearity of attribute-space" interpretation in StyleGAN, a parallel interpretation of DM's latent space can lead to a latent-space smoothening technique to obtain consistent outputs.

Bibliography

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 6, 33, 39

[2] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 6

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Trans. Graph.*, 40(4), 2021. 24

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 40 (4):1–12, 2021. 6, 39

[5] Isabela Albuquerque, João Monteiro, Thang Doan, Breandan Considine, Tiago Falk, and Ioannis Mitliagkas. Multi-objective training of generative adversarial networks with multiple discriminators. In *ICML*, 2019. 6

[6] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *CoRR*, abs/2112.05814, 2021. URL `https://arxiv.org/abs/2112.05814`. 28, 30, 32

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *ICML*, 2017. 6

[8] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[9] Philip Bachman, R Devon Hjelm, and William Buchwalter. *Learning Representations by Maximizing Mutual Information across Views*. 2019. 5

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 24, 28

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021. 6, 25, 30, 32

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017. 1

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 5, 6

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2, 5, 7, 10, 18

[17] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2, 5, 7

[18] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020. 5

[19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5, 7, 42

[20] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 25

[21] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 2, 5, 7

[22] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. Unsupervised discovery of object landmarks via contrastive learning. *arXiv preprint arXiv:2006.14787*, 2020. viii, 5, 7, 9, 10, 13, 14, 15, 16, 17, 19, 20

[23] Binod Kumar Choudhary, Navin Kumar Sinha, and Prem Shanker. Pyramid method in image processing. *Journal of Information Systems and Communication*, 3(1):269, 2012. 6

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 25

[25] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. 5

[26] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 6

[27] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2014. 2

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 1

[29] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2, 15

[30] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. doi: 10.1109/CVPR.2016.265. 25

[31] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 5

[32] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014. 5, 24

[34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. ii, vii, 2, 5, 6, 8, 9, 29

[35] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017. 6

[36] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 6

[37] Aryaman Gupta, Kalpit C. Thakkar, Vineet Gandhi, and P. J. Narayanan. Nose, eyes and ears: Head pose estimation by locating facial keypoints. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1977–1981, 2019. 2

[38] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 2006. 5

[39] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. 2 edition, 2004. 2

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 13

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 28

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 24, 30

[43] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1

[44] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 5, 7

[45] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron Van Den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 5

[46] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 5

[47] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016. 6

[48] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 6, 24, 27, 33, 39

[49] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, 2018. 4, 15

[50] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 15

[51] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eo6U4CAwVmg. 6

[52] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Deceive D: Adaptive Pseudo Augmentation for GAN training with limited data. In *NeurIPS*, 2021. 6, 32

[53] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(11):4037–4058, nov 2021. 1

[54] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 26, 31, 33

[55] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 6, 32

[56] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 2, 28

[57] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 24, 27, 29, 30, 32

[58] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 6

[59] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 6

[60] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 28

[61] Simon Kornblith, Jon Shlens, and Quoc V. Le. Do better imagenet models transfer better? 2019. URL `https://arxiv.org/pdf/1805.08974.pdf`. 25

[62] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. *arXiv preprint arXiv:2112.09130*, 2021. 6

[63] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 12

[64] Dmitry Laptev, Nikolay Savinov, Joachim M. Buhmann, and Marc Pollefeys. Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[65] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 14

[66] Jianshu Li, Pan Zhou, Y. Chen, Jian Zhao, S. Roy, Shuicheng Yan, Jiashi Feng, and T. Sim. Task relation networks. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2

[67] Jianshu Li, Jian Zhao, Fang Zhao, Hao Liu, Jing Li, Shengmei Shen, Jiashi Feng, and Terence Sim. Robust face recognition with deep multi-view representation learning. In *Proceedings of the 24th ACM International Conference on Multimedia*, 2016. 5

[68] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. vii, 8, 13

[69] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations*, 2020. 6

[70] 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)*, 2015. 12, 19

[71] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 4

[72] Dimitrios Mallis, Enrique Sanchez, Matt Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 5

[73] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *CVPR*, 2017. 6

[74] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, 2018. 6

[75] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020. 2

[76] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *ICLR*, 2018. 6

[77] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 6

[78] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPRW*, 2020. 6

[79] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *ICCV*, 2019. 6

[80] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016. 5

[81] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *Advances in Neural Information Processing Systems*, 2020. 5, 7, 13

[82] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021. 6

[83] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 24

[84] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 6, 24, 33, 39, 43

[85] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 6

[86] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[87] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 1

[88] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[89] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatilly consistent representation learning. In *CVPR*, 2021. 5, 7

[90] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 43

[91] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*. 13

[92] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 2, 6, 24, 27, 39, 43

[93] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, 2020. 2, 6, 24, 27, 33, 39

[94] Assaf Shocher, Yossi Gandelsman, Inbar Mosseri, Michal Yarom, Michal Irani, William T Freeman, and Tali Dekel. Semantic pyramid for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7457–7466, 2020. 6

[95] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and

appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 15

[96] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1354–1367, 2012. doi: 10.1109/TPAMI.2011.227. 31

[97] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[98] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 25, 30

[99] Kihyuk Sohn and Honglak Lee. Learning invariant representations with local transformations. In *ICML*, 2012. 5

[100] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020. 2

[101] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in Neural Information Processing Systems*, 2018. 4

[102] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 1

[103] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 6

[104] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *International Conference on Computer Vision*. viii, 2, 4, 13, 14, 15, 17, 20

[105] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in Neural Information Processing Systems*, 2017. 4, 5, 15

[106] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 4, 15

[107] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Modelling and unsupervised learning of symmetric deformable object categories. In *Advances in Neural Information Processing Systems*, 2018. 4

[108] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, 2020. 5

[109] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 28, 30

[110] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 6

[111] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 5

[112] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 43

[113] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 31

[114] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 5, 7

[115] Zhecan Wang and Jian Zhao. Conditional dual-agent gans for photorealistic and annotation preserving image synthesis. 2017. 2

[116] O. Wiles, A.S. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *British Machine Vision Conference*, 2018. 15

[117] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 12

[118] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 6, 33, 39

[119] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Computer Vision – ECCV 2016*. 2, 15

[120] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. 2021. 5, 7

[121] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=9BpjtPMyDQ. 32

[122] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[123] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 6

[124] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 31

[125] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14263–14272, 2021. 6, 39

[126] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[127] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2, 5

[128] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 5

[129] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 27

[130] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . 2, 4, 15

[131] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Computer Vision – ECCV 2014*, . 2, 13, 15

[132] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:918–930, 2016. 2, 15

[133] F. Zhao, Jian Zhao, Shuicheng Yan, and Jiashi Feng. Dynamic conditional networks for few-shot learning. In *ECCV*, 2018. 2

[134] Jian Zhao, Jianshu Li, F. Zhao, Xuecheng Nie, Y. Chen, Shuicheng Yan, and Jiashi Feng. Marginalized cnn: Learning deep invariant representations. In *BMVC*, 2017. 5

[135] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 6

[136] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2021. 33, 34