LARGE: Latent-Based Regression through GAN Semantics

Yotam Nitzan* Tel-Aviv University
Rinon Gal* Tel-Aviv University
Ofir Brenner Tel-Aviv University
Daniel Cohen-Or Tel-Aviv University

Abstract

We propose a novel method for solving regression tasks using few-shot or weak supervision. At the core of our method is the fundamental observation that GANs are incredibly successful at encoding semantic information within their latent space, even in a completely unsupervised setting. For modern generative frameworks, this semantic encoding manifests as smooth, linear directions which affect image attributes in a disentangled manner. These directions have been widely used in GAN-based image editing. In this work, we leverage them for few-shot regression. Specifically, we make the simple observation that distances traversed along such directions are good features for downstream tasks – reliably gauging the magnitude of a property in an image. In the absence of explicit supervision, we use these distances to solve tasks such as sorting a collection of images, and ordinal regression. With a few labels — as little as two — we calibrate these distances to real-world values and convert a pre-trained GAN into a state-of-the-art few-shot regression model. This enables solving regression tasks on datasets and attributes which are difficult to produce quality supervision for. Extensive experimental evaluations demonstrate that our method can be applied across a wide range of domains, leverage multiple latent direction discovery frameworks, and achieve state-of-the-art results in few-shot and low-supervision settings, even when compared to methods designed to tackle a single task.

Code is available on our project website.

1. Introduction

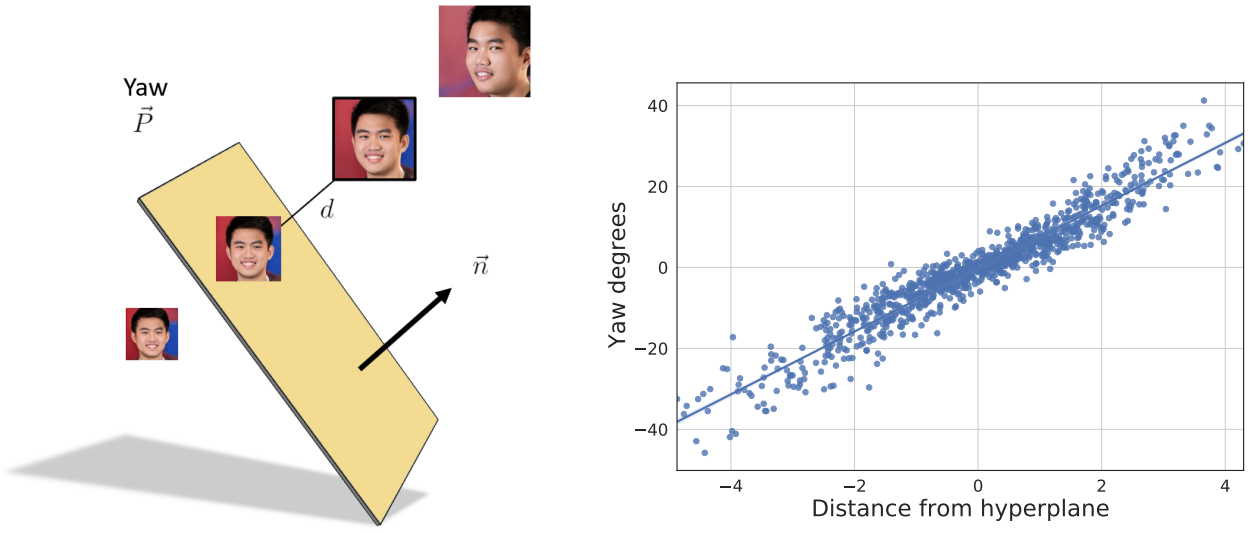
In recent years, Generative Adversarial Networks (GANs) [17] have been at the forefront of deep learning research. GANs revolutionized countless generative tasks, such as unconditional image synthesis [6, 28], cross-domain image-to-image translation [24, 72] and super-resolution [30]. Beyond generative tasks, numerous works have proposed to use GANs for downstream discriminative objectives, such as classification. Their shared premise is that a generator can synthesize novel samples - often in a controllable manner. These generated samples can then serve as a dataset for training models for downstream tasks. While this approach appears promising on paper [4, 33, 43, 53], this simple idea has enjoyed fairly limited success [12, 40].

We propose an alternative approach to harnessing the rapid advancement of GANs for downstream tasks. Specifically, we deviate from previous attempts to generate or augment training data. Instead, we focus on extracting information from the incredibly well-behaved latent space of modern GAN architectures, and specifically StyleGAN [27, 28]. The latent spaces of StyleGAN have been studied extensively [63], and were shown to be highly semantic and disentangled, properties which led to their wide use across a range of generative tasks. In this work, we leverage these properties in order to train few-shot regression models.

Specifically, we consider distances traversed along the normal vector of semantic hyperplanes (see fig. 1(a) for an illustration) and demonstrate that they are incredibly discriminative features for the task of regression. These normal vectors are commonly referred to as linear editing directions and many previous works proposed methods to identify them [20, 37, 45, 46].

What makes these latent-space distances useful features? First, the distances are globally consistent: all latent codes at a distance d from a semantic hyperplane resolve into images with similar attribute strengths (*e.g.* the same age). Hence, the link between distance and the corresponding attribute can be described with a function f. Second, distance is a scalar, allowing us to work with lower dimensional functions $f : \mathbb{R} \rightarrow \mathbb{R}$. Third, the relationship between the distances and attribute strength is monotone. Therefore, a total order of the distances corresponds to a total order on attribute strength. These attributes are already sufficient to make latent-space distances applicable as direct regression scores for applications where no conventional units are required or exist, *e.g.* ordinal regression.

To produce results with conventional units, such as head pose in degrees, we need to find an explicit approximation for f. Surprisingly, we find that a simple linear function: $f(d) = a \cdot d + b$ produces the best results for all attributes tested (see Figure 1(b)). Thus, we are able to fit an incredibly



(a) (b)

Figure 1. Latent distances from semantic hyperplanes serve as descriptive features. (a) An illustration of a hyperplane $\vec{P}$ and its normal $\vec{n}$ which defines a latent editing direction [45]. The latent code for the original image (black frame) is located at a distance of d from the hyperplane. (b) Scatter plot depicting the relationship between latent-space distances and the yaw angle [70] of real images. As can be seen, there is an approximately linear correlation between the two, $R^2 = 0.92$ (over the entire set).

simple function with only two parameters to regress complex semantic attributes in images. This allows us to perform regression in data domains and for semantic attributes where quality supervision is prohibitively difficult to acquire.

In order to perform regression on real images, the latent code corresponding to a given image is required. For this end, we use an off-the-shelf *GAN Inversion* encoder. We show that our method is compatible with several different encoders, namely pSp [41], e4e [54] and ReStyle [3].

Our model, outlined in Figure 2, is thus composed of several steps. First, we invert an image into the latent space. Next, we calculate the distance of the resulting latent code from a given semantic hyperplane. And last, we either use the distance as an uncalibrated score or, in the presence of a few labeled samples, use a simple regression model to convert it to real-world valued predictions. By following these steps, our method distills the strength of any semantic property into a single scalar, using only a pretrained generator and weak supervision.

In practice, we observe that the optimal latent space for *GAN inversion* may differ from the optimal space for finding semantic latent directions. In this case, a simple distance from the hyperplane cannot be calculated. To bridge the gap and derive a latent distance, we learn a task-specific mapping between distances in the two spaces using only latent-space considerations, with no additional supervision.

Through extensive evaluation, we show that our model can produce state-of-the-art results on few-shot learning tasks such as pose and age estimation, without *any* direct supervision on other domains, and that it can even match or outperform fully-supervised methods designed for specific tasks and trained on tens of thousands of samples. Where no supervision is available, we show that our model produces meaningful scores by demonstrating its applicability to the tasks of ordinal regression and sorting collections of images by the strength of a semantic property.

In summary, our contributions are:

- The observation that latent-space distances are highly semantic features, useful for downstream tasks.
- A scheme for converting a pretrained generator and a semantic latent-direction into a state-of-the-art few-shot regressive model.
- A new approach to analyzing layer-importance and mapping semantic distances between the latent spaces of a GAN.

2. Related Work

Latent Space of GANs: Recently, understanding and controlling the latent representation of pretrained GANs has attracted considerable attention. Notably, it has been shown that StyleGAN [26–28] creates a disentangled, smooth and semantically rich latent space. Many recent works have proposed methods to interpret the semantics encoded in this latent space and apply them to image editing [2, 20, 37, 45, 50, 57, 62]. Such methods typically identify a **linear** latent direction, which when traversed along, starting from an initial latent code, causes a gradual change in a single semantic property of the corresponding generated image.

To edit real images, one must first obtain the latent code from which the pretrained GAN can most accurately reconstruct the original input image. This task is commonly

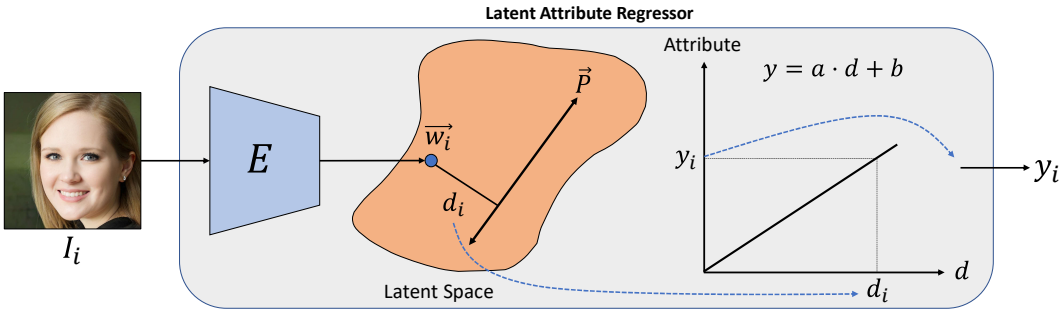


Figure 2. Outline of our proposed regression pipeline. An image I_i is inverted into a latent-code $\vec{w}_i$. The distance d_i of the code $\vec{w}_i$ from a semantic hyperplane $\vec{P}$ is calculated. Finally, d_i is input to the simple regression model which outputs the magnitude of the semantic attribute y_i of the image.

referred to as *GAN Inversion* and has been tackled by numerous recent works [1, 3, 28, 41, 54, 71]. For a thorough introduction, we refer the reader to a recent survey [63].

GANs for Discriminative Tasks: Several works sought to leverage the advancement of GANs for discriminative tasks. Arguably, the most straight-forward approach is to simply train a generator and use it to synthesize labeled samples. Such samples are inherently labeled when the generator performs either image-to-image translation between domains or class-conditioned synthesis. Following this approach, some works achieved competitive results using completely generated datasets [34, 53]. Others enriched a real dataset to improve performance in the low-data domain [4, 16, 59, 73] or on biased or unbalanced data [18, 22, 33, 39, 44]. When applied to the ImageNet classification task, some works [40, 47] demonstrated that performance sees only modest improvement when enriching the real set, and becomes poor when replacing it.

While these works cover a diverse collection of settings and approaches, they all used GANs in order to eventually generate more data **for training**. In the context of discriminative tasks, several recent methods have proposed to utilize GANs for additional purposes. Lang *et al*. [29] used StyleGAN to visualize counterfactual examples for explaining a pretrained classifier's predictions. Chai *et al*. [9] used style-mixing in the fine-layers of StyleGAN to generate augmentations that are ensembled together at test-time. Most related to our work is the representation learning framework GHFeat [65]. In their work, Xu *et al*. train an encoder for *GAN Inversion* into the latent space of a pretrained StyleGAN and demonstrate that the visual features learned by this encoder can be used to train a variety of models for downstream tasks in a fully supervised manner.

In contrast, we build on recent progress in the study of GANs' latent space and observe that *distances* within this space can already serve as one-dimensional discriminative features. Our method can leverage these features for simple discriminative tasks, such as sorting, using only weak supervision. With just a few labeled samples, as little as two, we are able to regress real world values. In settings where both our models are applicable, we compare our method with GHFeat and find that we can obtain more accurate results using significantly less data.

Few-Shot Regression: In the context of visual media, few-shot methods have been widely studied for classification [48, 51, 56], object detection [15, 60] and segmentation [14, 38]. Relatively few image-related few-shot models have stepped beyond these bounds. They tackle image-to-image translation [31], super resolution [49], motion prediction [19], and re-identification [61].

In particular, this scarcity holds for regression tasks, where a majority of research focused on pose estimation using strong task-specific priors [64] or applying meta-learning ideas to keypoint extraction and gaze estimation frameworks [36, 55]. For an overview, we refer the reader to a recent survey [58]. These works, however, all deal with specific tasks and adapt them across domains in a few-shot manner. Their use of task-specific supervision, priors and architectures prevents them from easily being extended to new objectives. In contrast, our method can be easily generalized across domains and different regression goals, including cases where collecting task-specific supervision from other domains would be difficult or outright impossible.

3. Method

A shared premise in GAN-based editing works is that single, semantic image properties can be manipulated through modifications of the latent codes used to generate the image. These modifications are conducted by discovering appropriate **global**, **linear** steering directions within the latent space of the generator (Figure 1(a)). Such directions take the form of vectors, $\vec{n}$, which can be used to induce a semantic change in *any* code $\vec{w}$ (*i.e*. they are **global**) through **linear** addition: $\vec{w}' = \vec{w} + \alpha\vec{n}$, where α is a scalar that controls the strength of the modification.

We propose to flip this idea on its head. If moving a fixed distance away from a latent hyperplane causes an attribute change that is similar for all images, then by determining

these distances we can gauge the strengths of these attributes. These distances can thus serve as features for regression.

To regress a single property in a real image I_i, we thus require two components: a semantic latent direction corresponding to the property in question $\vec{n}$, and the representation of the image in the latent-space of a pretrained GAN: $\vec{w}_i$. As previously discussed, semantic directions can be discovered using an array of editing methods [37,45,46]. Such direction can be alternatively viewed them as normals to a hyperplane $\vec{P}$ partitioning the latent space: $\vec{n} \cdot \vec{w} + b = 0$. Now, the magnitude of the property in the image is the distance of its latent representation to the hyperplane:

$$d = dist(\vec{w}_i, \vec{P}) = \vec{w}_i \cdot \vec{n} + b. \quad (1)$$

Some editing methods produce a value for the intercept b, while others do not. When b is unknown, we set it arbitrarily to $b = 0$. As b modifies all distances by a constant factor, dropping it has no effect on any sorting applications, and it can be effectively determined during calibration when training a regressor on real data. When b is known, the distance of 0 carries a special semantic meaning. For example, when considering head pose it corresponds to a frontal face.

While calculating the distance is a simple algebraic operation and requires no supervision, this does not imply that the method is unsupervised. Supervision is dictated by the methods used for finding the semantic latent direction and for performing GAN Inversion. These methods typically require weak or indirect supervision. InterFaceGAN [45], for example, requires only binary labeling of an attribute. In the head pose example, such annotation would amount to left/right pose information, rather than an explicit yaw angle, which is significantly harder to annotate. The GAN itself is trained in a completely unsupervised manner.

3.1. Inversion

A requirement of our approach is the ability to invert an image into the latent space of the GAN. Specifically, with StyleGAN, there exist numerous latent spaces that have been considered by previous works. We follow the common approach of inverting an image into the $\mathcal{W}+$ space, first introduced by Abdal *et al*. [1]. For an inversion method, we follow the encoder-based scheme. Our method works seamlessly with multiple off-the-shelf encoders, but shows improved performance when utilizing e4e [54], which was designed to produce latent codes that are highly editable. We provide a comparison of inversion methods in the appendix.

3.2. Bridging the Gap Between Spaces

A topic of ongoing research in the field of GAN-based image manipulation is the choice of optimal latent spaces for inversion [54] and editing [62]. In StyleGAN, the generator supports four latent spaces: $\mathcal{Z}$, $\mathcal{W}$, $\mathcal{W}+$ and $\mathcal{S}$. $\mathcal{Z}$ denotes the usual Gaussian prior, while the others denote increasingly complex spaces obtained by passing through StyleGAN's mapping network ($\mathcal{W}$), by assigning different $\mathcal{W}$ codes to different layers of the GAN ($\mathcal{W}+$) or by applying affine transformations to the codes ($\mathcal{S}$).

A common challenge in StyleGAN editing tasks is that the latent space most often used to identify semantic latent directions, $\mathcal{W}$, is not expressive enough to support accurate reconstructions of images. Conversely, the $\mathcal{W}+$ space, allows for accurate reconstructions - but behaves poorly under $\mathcal{W}$-based transformations [54]. One reason for this behaviour is that in the $\mathcal{W}+$ space, different codes affect different layers which in turn affect different semantic properties of the generated image. For example, pose is controlled by early layers of the network while colors are largely controlled by later layers [27, 66]. While applying a $\mathcal{W}$ space editing direction equally to all layer codes may still modify a desired property, not all layer modifications are required, or even affect the property at all. For such layers, the distance from the hyperplane is *irrelevant* to the magnitude of the semantic property. A naïve sum over all layer distances would therefore correlate poorly with the strength of the property. A natural question arising is then - which layers *are* relevant? We answer this question by learning an importance score for each layer in an unsupervised manner.

To do so, we sample a random latent code $\vec{w} \in \mathcal{W}$ in the same space as the semantic hyperplane. We edit the sampled code and obtain $\vec{w}_e = \vec{w} + \alpha\vec{n}$ which is then used to generate an edited image $I_e = G(\vec{w}_e)$. We map the original $\vec{w}$ to a code $\vec{w}^+ \in \mathcal{W}+$ by duplicating it, once for each layer. Finally, we set up a direct optimization scheme where we attempt to modify the mapped code, $\vec{w}^+$ such that it can be used to generate the edited image, i.e. we aim to solve:

$$\vec{w}^* = argmin_{\vec{w}^+} \left\| G\left(\vec{w}^+\right) - I_e \right\|^2 . \quad (2)$$

Note that $\vec{w}_e$ is an optimal solution for Eqn. (2). However, it is not the only solution. When optimizing $\vec{w}^+$ to solve the equation, layers which are *irrelevant* to the property will hardly change from their initialization. The per-layer magnitude of change in the optimized code is therefore an intuitive measure of the importance of each layer towards a semantic attribute. We solve Eqn. (2) for multiple initial codes using gradient descent and track the mean magnitude of the gradients along different layers. In order to avoid spurious results due to different layers having different gradient scales, we normalize gradients by their values when optimizing between unrelated-images. The normalized gradient magnitudes serve as per-layer importance scores, $\{S_i\}_{i=1}^{L}$, and are used to calculate a weighted sum of hyperplane distances:

$$d_{\mathcal{W}+} = dist_{\mathcal{W}+}(\vec{w}^+, \vec{P}_{\mathcal{W}}) = \sum_{i=1}^{L} S_i dist_{\mathcal{W}}(\vec{w}_i^+, \vec{P}_{\mathcal{W}}) \quad (3)$$

We use $d_{\mathcal{W}+}$ as an effective distance between $\vec{w}^+ \in \mathcal{W}+$ and a hyperplane $\vec{P}_{\mathcal{W}}$ given in $\mathcal{W}$.

3.3. Calibration

In many cases, one expects a regression model to output a prediction in "real-world" values, such as head pose in degrees. As presented, the latent-distances are uncalibrated. One can ask, given a sample for which $d = 3$, what is the actual head pose in degrees?

To answer this question, we need to train a function f to map from uncalibrated latent-space distances to actual real-world values. In general, f could be a complex function. However, we find that StyleGAN's latent space is sufficiently well structured that a linear function provides the best results. We therefore fit a simple linear regression model with one feature per sample - the distance to the semantic hyperplane. Such a linear model simply takes the form $y = a \cdot d + b$, where y is the calibrated property prediction, d the latent-space distance, and a, b are learned parameters which can be determined with as few as two sampled points.

Once trained, the model gets the distance from the hyperplane as input and predicts the real-world interpretable value. Finally, the entire pipeline produces a few-shot image-property regression model.

4. Experiments

We demonstrate our method on several domains and properties. When performing quantitative evaluation, we annotate unlabeled datasets using pretrained networks (specifically WHENet [70] and DEX [42]). Human face image experiments use the official StyleGAN2 pretrained on FFHQ [27]. Age experiments use the CACD [10] and CelebA-HQ [25] datasets, while all other evaluations are conducted on CelebA-HQ. Leaf image experiments use the Plant-Village dataset [23]. Cat image experiments use the official StyleGAN-ADA [26] model trained on AFHQ [11], and are evaluated on the test-split of the same set. Experiments on additional domains are provided in the appendix. In all cases, the GAN was trained in a completely unsupervised manner, without any labels. All results are shown on real images. Where feasible, we report the mean and standard deviation over a thousand repetitions using randomly samples subsets of the data.

4.1. Feature Space Comparisons

We start by demonstrating that distances in the latent space of the GAN are more semantically meaningful and better behaved than equivalent distances in alternative feature spaces. In particular, we compare our method with multiple baselines operating in a similar manner to the InterFaceGAN [45] approach. First, a large collection of binary-tagged images are acquired, along with their representation in each chosen feature space. Then, we train an SVM in the given feature space using the binary labels, providing us with a separating hyperplane matching the semantic attribute described by the labels. Finally, the distance to the hyperplane within the given feature space is used as a discriminative feature for training a regressor.

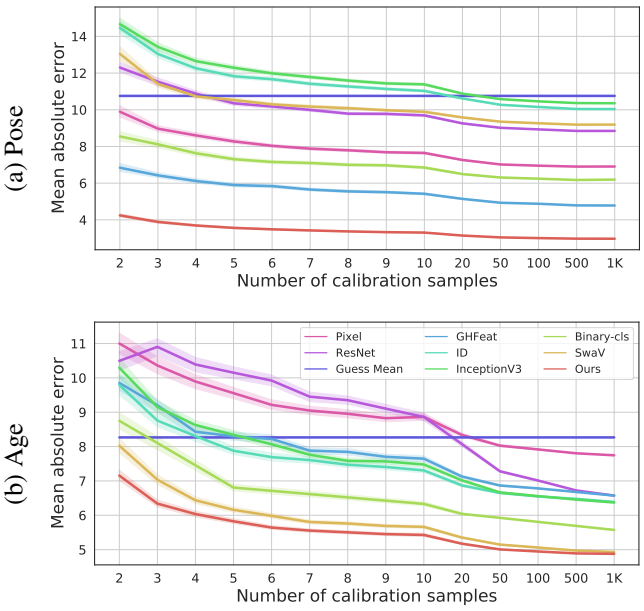


Figure 3. Comparisons to hyperplane-distance baselines operating in different feature spaces. Our method outperforms all baselines, indicating GAN-space distances are more semantically meaningful.

The simplest baseline considers the pixel space of the image as a feature space. Other baselines use feature spaces learned by deep neural networks, varying in architectures and tasks. Specifically, we consider the feature spaces learned by: GHFeat [65], SwaV [8] - an unsupervised representation learning approach, ResNet101-2W [68] and InceptionV3 [52] ImageNet classifiers, and a face recognition network [13] ("ID"). Lastly, we compare to the feature space learned by a ResNet18 [21] model trained to provide binary classification of the images for the *same* semantic attribute ("Binary-cls").

As can be seen in Figure 3, for all experiments, our model outperforms the baselines. These results demonstrate that distances to semantic boundaries within the latent space of a pretrained generator are more semantically meaningful, and serve as better discriminative features for *linear* regression than distances in alternative feature spaces. Furthermore, these results demonstrate that the idea of utilizing a one-dimensional distance metric in some learned feature space is not universal, but relies upon the extensive semantic knowledge encapsulated by the GAN. Indeed, for some configurations, the baselines perform worse than simply returning the mean of the training set's distribution.

4.2. Calibrated Results

Having established that the latent space of StyleGAN is a source of semantically meaningful feature representations,

we turn to evaluating the strength of these features with respect to prior works. Towards this end, we demonstrate the performance of our model on a set of regression tasks: pose and age estimation for human faces. Additionally, pose estimation for cars is provided in the appendix. We train and compare models over a wide range of supervision settings, showing that our approach achieves state-of-the-art performance in the few-shot domain as well as surprisingly competitive results on larger datasets, even when compared to methods designed for a single, specific task.

For each attribute we evaluate against a set of methods designed for the specific task, often utilizing a task specific architecture. Furthermore, we compare against GHFeat [65], which proposed the use of a different set of GAN-inspired features, and is thus comparable in all scenarios. Comparisons against GHFeat are conducted by training a linear regression model directly on their features. Additionally, we compare to the distance-based GHFeat variant introduced in Subsection 4.1 ("GHFeat SVM"). We find that this variant performs better than the original in few-shot scenarios.

Human Pose: We evaluate the performance of our model on the task of predicting human head poses, and specifically yaw. The semantic direction used by our method was extracted using InterFaceGAN [45]. In addition to GHFeat, we compare to two dedicated pose estimation methods: FSA [67], a fully supervised method and SSV [35], a method which learns pose estimations in a self-supervised manner and then calibrates them to dataset-specific values with few labeled samples. In this sense, their method can also be regarded as a few-shot approach. In Figure 4(a) we show the mean absolute errors (MAE) of yaw estimation using the outlined approaches. Our method consistently outperforms all methods when presented with limited supervision, and remains competitive up to a thousand labeled samples. Of particular note is the fact that our model displays better performance than SSV, indicating that the latent space of a pretrained GAN encodes more meaningful pose information than comparable self-supervised methods which were designed and trained specifically to extract pose.

Human Age: We evaluate our performance on the task of human age estimation. The age editing direction used by our method was discovered through natural language descriptions using StyleCLIP [37]. Specifically, the boundary was extracted with the prompts "old face" and "young face". In addition to GHFeat, we compare to CORAL [7]. Figure 4(b) shows the MAE on age estimation in years. Our method outperforms the alternatives when presented with limited data, and is only outpaced by CORAL when provided with tens of thousands of samples. Specifically, our 20-sample model obtains a MAE of 7.46, in line with CORAL's 20K-sample result – 7.59. Note that, unlike pose, age is a property that does not manifest linearly in pixel-space. The visual difference between a person at ages 8 and 13 is much greater

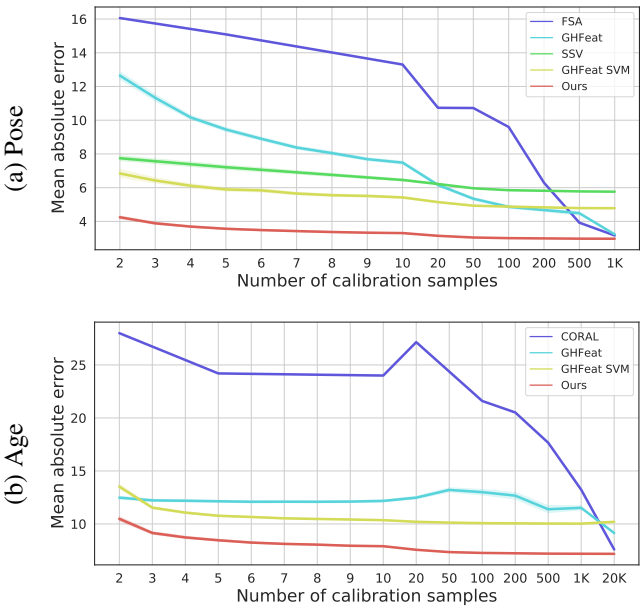


Figure 4. Comparisons to alternative models as a function of the number of labeled images used in training. In (a) we compare to GHFeat, FSA, and SSV on the CelebA-HQ dataset. In (b) we compare to GHFeat and CORAL on the CACD dataset.

than the difference between 30 and 35. However, we find that a linear regression model still performs well in this case. Furthermore, this experiment demonstrates that textually described boundaries are suitable for regression - opening a path to few-shot regression of many properties that can be reasonably described through natural language.

As verified by the experiments, our method achieves state-of-the-art results in the few-shot domain, even when compared to models designed for specific regression tasks.

4.3. Uncalibrated Results

We next discuss domains and attributes for which continuous supervision is not available. Even in such cases, our method produces meaningful scores which describe the magnitude of a semantic property in the image. Without supervision, the scores cannot be calibrated to any real-world human-interpretable value. Nevertheless, uncalibrated scores are still useful for applications such as sorting and ordinal regression. Here, we demonstrate their applicability to the task of sorting a collection of images by a given property – *e.g.* according to how happy the person in the image is. An ordinal regression experiment is provided in the appendix.

Figure 5 shows results for sorting a set of face images according to properties described through textual prompts, using distances to boundaries discovered through StyleCLIP's global mapping approach. We sort the same randomly sampled collection according to four distinct attributes: expression, hair color, hair length, and amount of makeup.

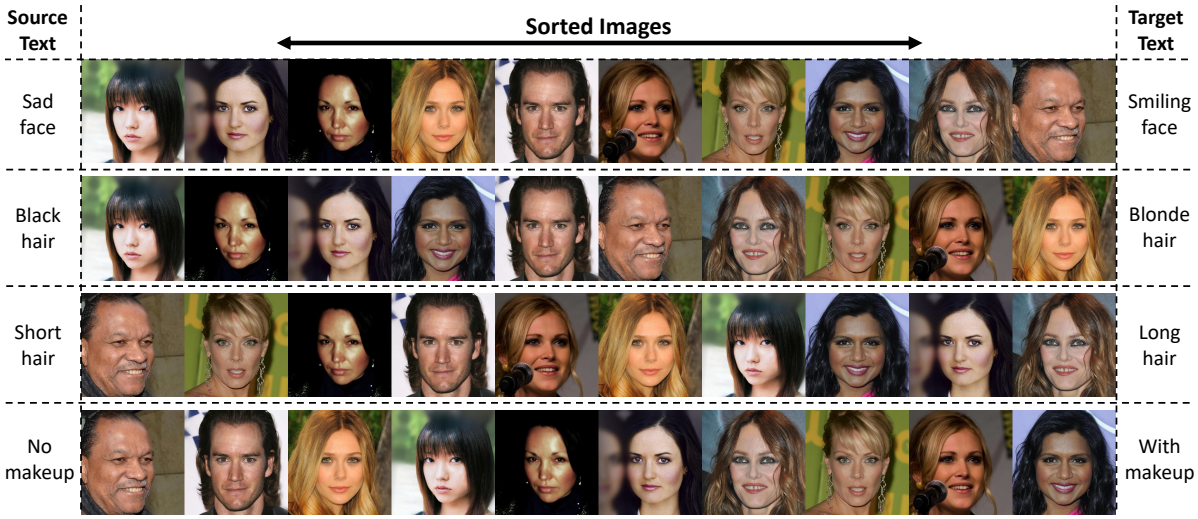


Figure 5. Sorting images according to textual descriptions of semantic properties. In each row, we sort the same set of randomly sampled CelebA-HQ images according to their distance from a text-based editing boundary extracted by StyleCLIP. Each row's editing direction is induced by the source text (left) and the target text (right).

Figure 6. Sorting images from the Plant-Village dataset using a semantic direction extracted by InterFaceGAN. In each row, we sort randomly sampled sets containing images labeled as either healthy or sick. We separate rows by type of disease to facilitate visual comparisons.

In Figure 6 we show the results of sorting a collection of leaf images corresponding to a 'sick or healthy' direction, using a boundary extracted with InterFaceGAN [45] based on binary annotations. Our method successfully turns these binary annotations into continuous values which allow us to determine which leaves are sicker than others. For example, in the last row, the number of "black spots" decreases gradually while moving from the most-sick leave towards the healthy leaves.

In Figure 7 we show results on sorting collections of cats randomly sampled from the AFHQ dataset [11], using semantic directions discovered in an unsupervised manner with SeFA [46]. Our method extends seamlessly to these additional domains and semantic direction discovery approaches. Note that this result is obtained with nearly no supervision. Both StyleGAN and SeFA operate without any labels, while the encoder relied only on a deep perceptual metric [69].

As can be seen, the ordered results largely align with human expectation. In order to verify this claim we conduct a user study on the human and cat face domains. The full details of the baselines and the experimental setup are provided in the appendix. Results are summarized in Table 1. As verified by the study, our model learns to regress more consistent scores for images across both domains.

As demonstrated through our experiments, our approach learns to regress meaningful, uncalibrated values across multiple domains and using a wide range of latent-direction discovery methods.

Figure 7. Sorting images from AFHQ-cat using semantic directions extracted by SeFA. The sorting attribute is shown to the left of each row.

Table 1. User study results for sorting quality. For each attribute we report the percent of responders which preferred the sorting induced by each of three sorting methods. Our method consistently provides an order which is more consistent with human preference. See the appendix for more details.

(a) Human faces

Attribute	Ours	CLIP	Random
Hair color	**73.55%**	25.81%	0.65%
Makeup	**70.53%**	18.25%	11.23%
Expression	**53.45%**	43.27%	3.27%
Hair length	**84.73%**	2.18%	13.09%
Average	**70.56%**	22.38%	7.06%

(a) Cats

Attribute	Ours	SSV	Random
Yaw	**66.67%**	28.99%	4.35%
Pitch	**82.50%**	7.50%	10.00%
Average	**75.71%**	16.71%	7.58%

5. Discussion

We have presented a method for leveraging the semantic structure of a pre-trained GAN for weakly-supervised and few-shot regression tasks. Our method builds upon extensive prior works which explored the latent space of StyleGAN. Some of these works found semantic linear latent directions and applied them to image editing. Others directly utilized latent codes as features for regression tasks. However, we are the first to note that insights from both approaches can be combined. We have shown that distances to semantic hyperplanes can serve as simple and incredibly discriminative features. In the absence of explicit supervision, these latent space distances can be used for applications such as sorting or ordinal regression. With as few as two labeled samples, they can be calibrated to real-world values, producing state-of-the-art few-shot regression models.

While we have observed linearity in all semantic attributes which we tested, this property is unlikely to hold universally. In some cases, the best calibration function might be non-linear, and indeed for some attributes even finding a disentangled latent direction may be infeasible in the first place.

Lastly, our method relies on the ability to invert images into the latent space of the GAN. For rare attributes or images outside the generator's domain, this can prove challenging. As inversion is significantly more difficult for multi-class GANs [5, 9, 32] such as BigGAN [6], we follow prior work [9, 29, 65] and focus our attention on the state-of-the-art single-object GAN - StyleGAN. However, improved inversion methods are a topic of ongoing research [63]. As such, we are looking forward to see these barriers overcome.

We hope that our work can inspire others to consider the latent space of GANs as a source of semantically-rich supervision which can be leveraged to tackle a wide range of downstream tasks.

6. Broader Impact

Our model consists of a general method for performing regression on images. As such, its impact is dependent on the tasks for which it is used. Such tasks could have a wide range of positive benefits in numerous computer vision related fields. For example, the ability to quantify levels of disease in plants, and perhaps in other domains, may be of benefit in the agricultural and healthcare fields. Enabling few-shot regression may further assist with 'democratizing' neural networks, in the sense that the method could enable smaller businesses or research groups in performing regression in scenarios where labeled data may be out of their means.

On the other hand, our model could be used in applications which violate privacy, for example by facilitating the collection of information such as age or sentiments from individual photographs.

Furthermore, our model is sensitive to the same biases found in the data used to train the GAN - and in the case of natural-language based regression, also to the biases present in CLIP. As such, it may assist in perpetuating biases such as gender norms (as seen in the makeup sorting experiment) or racial discrimination (*e.g.* Asian descent is correlated with age in the data set, and this is reflected in model predictions).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 3, 4

[2] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows, 2020. 2

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. *arXiv preprint arXiv:2104.02699*, 2021. 2, 3

[4] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 1, 3

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019. 8

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1, 8

[7] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. 6

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 5

[9] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *CVPR*, 2021. 3, 8

[10] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 5

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5, 7

[12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 1

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 5

[14] Yuki Endo and Yoshihiro Kanamori. Few-shot semantic image synthesis using stylegan prior. *arXiv preprint arXiv:2103.14877*, 2021. 3

[15] Qi Fan, W. Zhuo, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4012–4021, 2020. 3

[16] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. 3

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 1

[18] Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. 2019. 3

[19] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018. 3

[20] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 1, 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Dana Cohen Hochberg, Raja Giryes, and Hayit Greenspan. Style encoding for class-specific image generation. In *Medical Imaging 2021: Image Processing*, volume 11596, page 1159631. International Society for Optics and Photonics, 2021. 3

[23] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. 5

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 2, 5

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2, 4, 5

[28] 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*, pages 8110–8119, 2020. 1, 2, 3

[29] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. *arXiv preprint arXiv:2104.13369*, 2021. 3, 8

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[31] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019. 3

[32] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021. 8

[33] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018. 1, 3

[34] Richard T Marriott, Safa Madiouni, Sami Romdhani, Stéphane Gentric, and Liming Chen. An assessment of gans for identity-related applications. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 3

[35] Siva Karthik Mustikovela, Varun Jampani, Shalini De Mello, Sifei Liu, Umar Iqbal, Carsten Rother, and Jan Kautz. Self-supervised viewpoint learning from image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3981, 2020. 6

[36] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377, 2019. 3

[37] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 1, 2, 4, 6

[38] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018. 3

[39] Vikram V Ramaswamy, Sunnis SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. *arXiv preprint arXiv:2012.01469*, 2020. 3

[40] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *arXiv preprint arXiv:1905.10887*, 2019. 1, 3

[41] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020. 2, 3

[42] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015. 5

[43] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 1

[44] Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020. 3

[45] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 1, 2, 4, 5, 6, 7

[46] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 1, 4, 7

[47] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my gan? In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229, 2018. 3

[48] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. 3

[49] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020. 3

[50] Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan steerability without optimization. *arXiv preprint arXiv:2012.05328*, 2020. 2

[51] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018. 3

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[53] Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019. 1, 3

[54] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 2, 3, 4

[55] Hung-Yu Tseng, Shalini De Mello, Jonathan Tremblay, Sifei Liu, Stan Birchfield, Ming-Hsuan Yang, and Jan Kautz. Few-shot viewpoint estimation. *arXiv preprint arXiv:1905.04957*, 2019. 3

[56] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016. 3

[57] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *arXiv preprint arXiv:2002.03754*, 2020. 2

[58] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-

shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 3

[59] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7278–7286, 2018. 3

[60] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020. 3

[61] Lin Wu, Yang Wang, Hongzhi Yin, Meng Wang, and Ling Shao. Few-shot deep adversarial learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 29:1233–1245, 2019. 3

[62] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation, 2020. 2, 4

[63] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 1, 3, 8

[64] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision*, pages 192–210. Springer, 2020. 3

[65] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. *arXiv e-prints*, pages arXiv–2007, 2020. 3, 5, 6, 8

[66] Ceyuan Yang, Yujun Shen, and Bolei Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129(5):1451–1466, 2021. 4

[67] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019. 6

[68] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[70] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 2, 5

[71] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 3

[72] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

[73] Xinyue Zhu, Yifan Liu, Zengchang Qin, and Jiahong Li. Data augmentation in emotion classification using generative adversarial networks. *arXiv preprint arXiv:1711.00648*, 2017. 3